STAT 520

FORECASTING AND TIME SERIES

Fall, 2013

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

1 Introduction and Examples

Complementary reading: Chapter 1 (CC).

TERMINOLOGY: A **time series** is a sequence of ordered data. The "ordering" refers generally to time, but other orderings could be envisioned (e.g., over space, etc.). In this class, we will be concerned exclusively with time series that are

- measured on a single **continuous** random variable Y
- equally spaced in **discrete time**; that is, we will have a single realization of Y at each second, hour, day, month, year, etc.

UBIQUITY: Time series data arise in a variety of fields. Here are just a few examples.

- In **business**, we observe daily stock prices, weekly interest rates, quarterly sales, monthly supply figures, annual earnings, etc.
- In **agriculture**, we observe annual yields (e.g., crop production), daily crop prices, annual herd sizes, etc.
- In **engineering**, we observe electric signals, voltage measurements, etc.
- In **natural sciences**, we observe chemical yields, turbulence in ocean waves, earth tectonic plate positions, etc.
- In **medicine**, we observe EKG measurements on patients, drug concentrations, blood pressure readings, etc.
- In **epidemiology**, we observe the number of flu cases per day, the number of health-care clinic visits per week, annual tuberculosis counts, etc.
- In **meteorology**, we observe daily high temperatures, annual rainfall, hourly wind speeds, earthquake frequency, etc.
- In **social sciences**, we observe annual birth and death rates, accident frequencies, crime rates, school enrollments, etc.

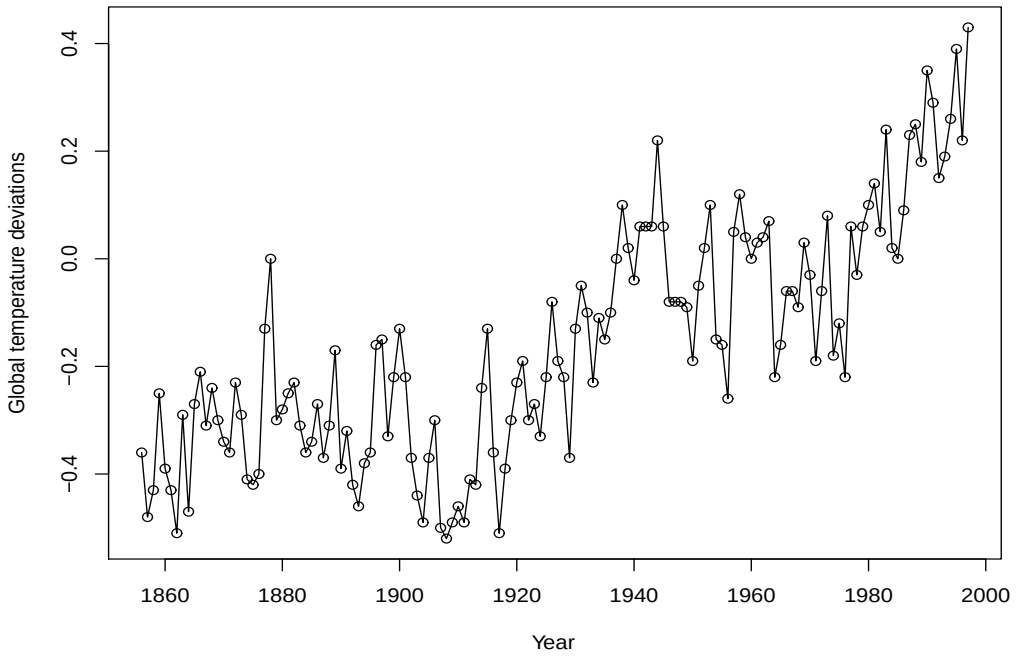


Figure 1.1: Global temperature data. The data are a combination of land-air average temperature anomalies, measured in degrees Centigrade.

Example 1.1. *Global temperature data.* "Global warming" refers to an increase in the average temperature of the Earth's near-surface air and oceans since the mid-20th century and its projected continuation. The data in Figure 1.1 are annual temperature deviations (1856-1997) in deg C, measured from a baseline average.

- Data file: `globaltemps`
- There are $n = 142$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions? (Are we doomed?)

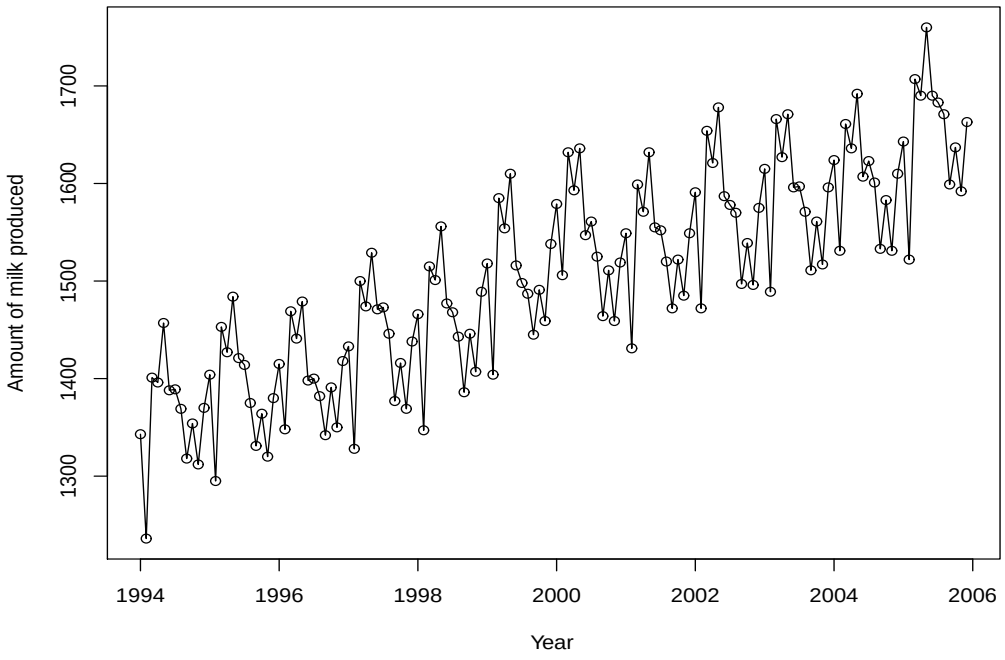


Figure 1.2: United States milk production data. Monthly production figures, measured in millions of pounds, from January, 1994 to December, 2005.

Example 1.2. *Milk production data.* Commercial dairy farming produces the vast majority of milk in the United States. The data in Figure 1.2 are the monthly U.S. milk production (in millions of pounds) from January, 1994 to December, 2005.

- Data file: `milk` (TSA)
- There are $n = 144$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns?
- Predictions?

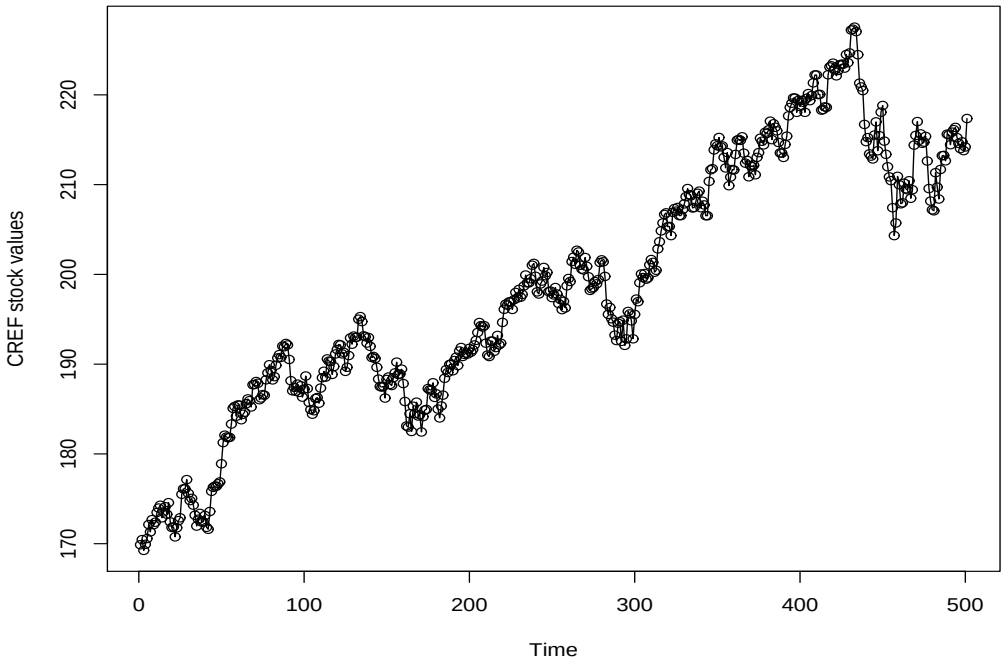


Figure 1.3: CREF stock data. Daily values of one unit of CREF stock values: August 26, 2004 to August 15, 2006.

Example 1.3. *CREF stock data.* TIAA-CREF is the leading provider of retirement accounts and products to employees in academic, research, medical, and cultural institutions. The data in Figure 1.3 are daily values of one unit of the CREF (College Retirement Equity Fund) stock fund from 8/26/04 to 8/15/06.

- Data file: `CREF` (TSA)
- There are $n = 501$ observations.
- Measurements are taken **each trading day**.
- What are the noticeable patterns?
- Predictions? (My retirement depends on these!)

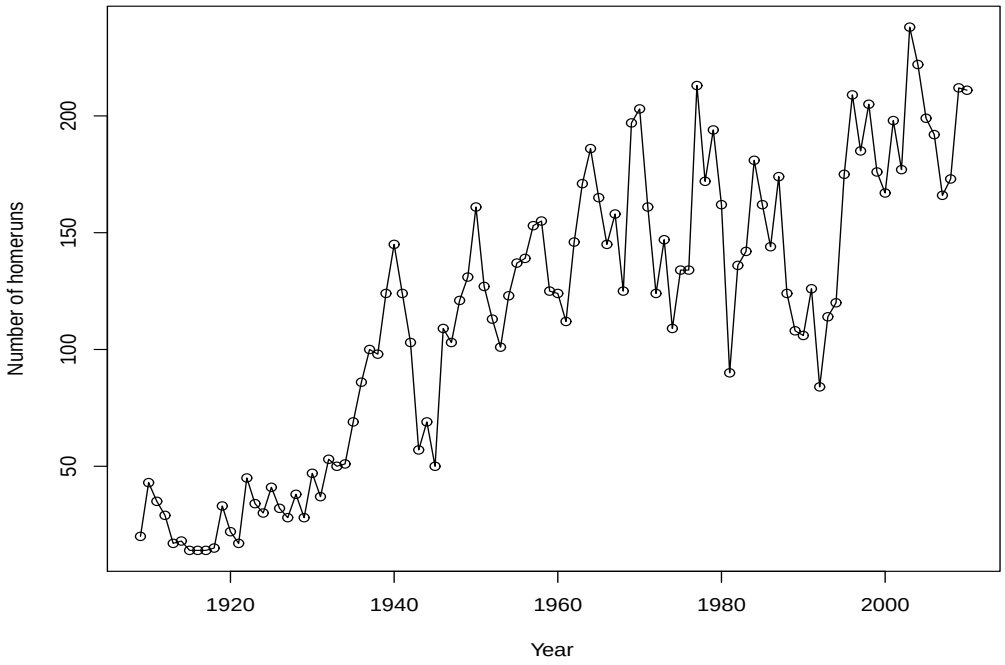


Figure 1.4: Homerun data. Number of homeruns hit by the Boston Red Sox each year during 1909-2010.

Example 1.4. *Homerun data.* The Boston Red Sox are a professional baseball team based in Boston, Massachusetts, and a member of the Major League Baseball's American League Eastern Division. The data in Figure 1.4 are the number of homeruns hit by the team each year from 1909 to 2010. **Source:** Ted Hornback (Spring, 2010).

- Data file: `homeruns`
- There are $n = 102$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

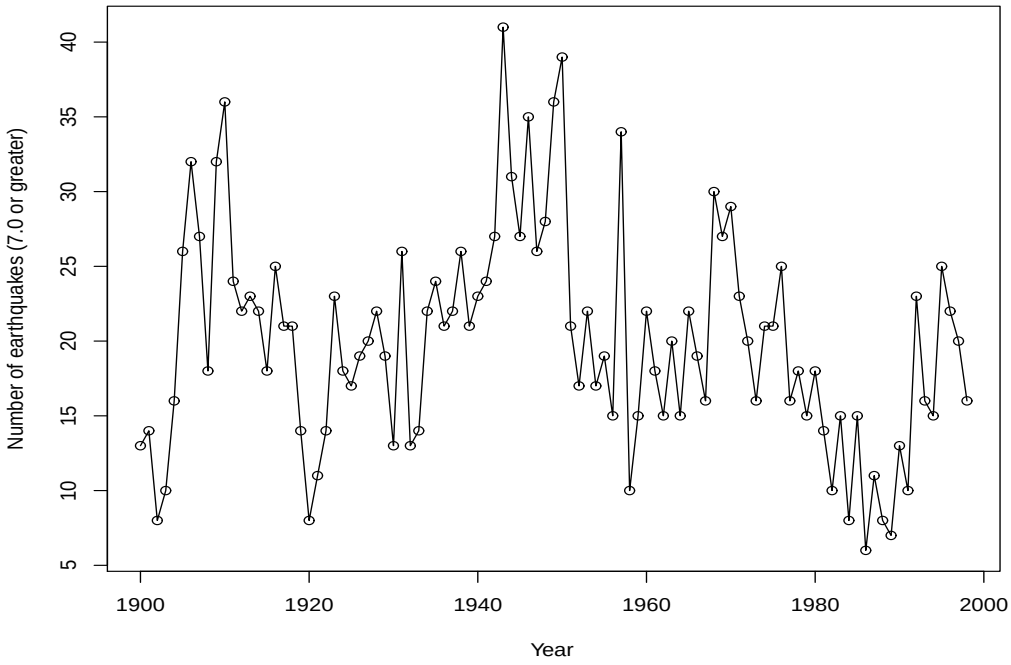


Figure 1.5: Earthquake data. Number of "large" earthquakes per year from 1900-1998.

Example 1.5. *Earthquake data.* An earthquake occurs when there is a sudden release of energy in the Earth's crust. Earthquakes are caused mostly by rupture of geological faults, but also by other events such as volcanic activity, landslides, mine blasts, and nuclear tests. The data in Figure 1.5 are the number of global earthquakes annually (with intensities of 7.0 or greater) during 1900-1998. **Source:** Craig Whitlow (Spring, 2010).

- Data file: `earthquake`
- There are $n = 99$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

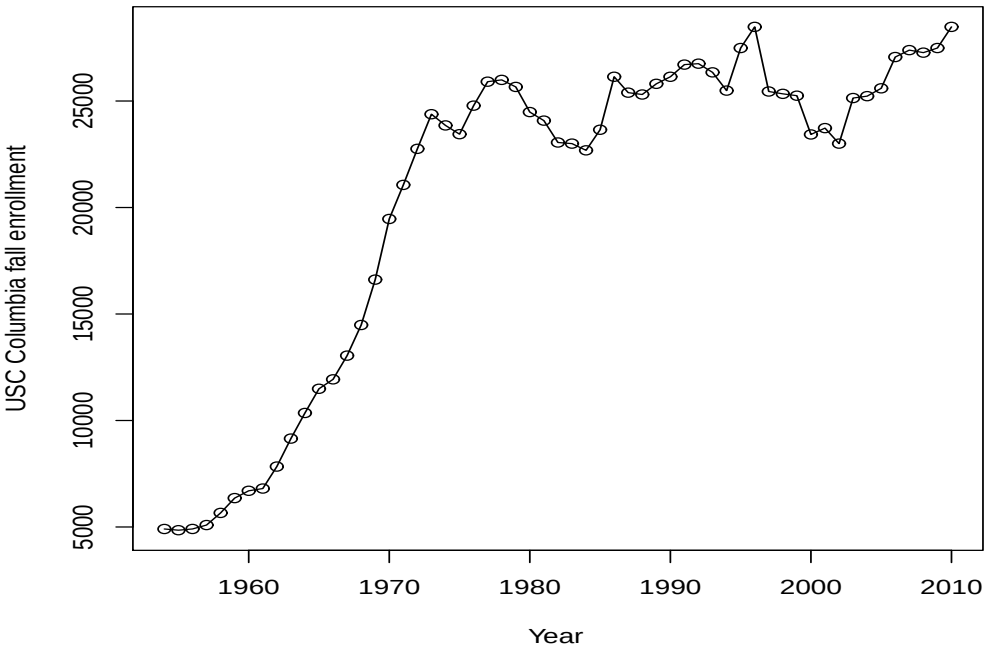


Figure 1.6: University of South Carolina fall enrollment data. Number of students registered for classes on the Columbia campus during 1954-2010.

Example 1.6. *Enrollment data.* The data in Figure 1.6 are the annual fall enrollment counts for USC (Columbia campus only, 1954-2010). The data were obtained from the USC website `http://www.ipr.sc.edu/enrollment/`, which contains the enrollment counts for all campuses in the USC system.

- Data file: `enrollment`
- There are $n = 57$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

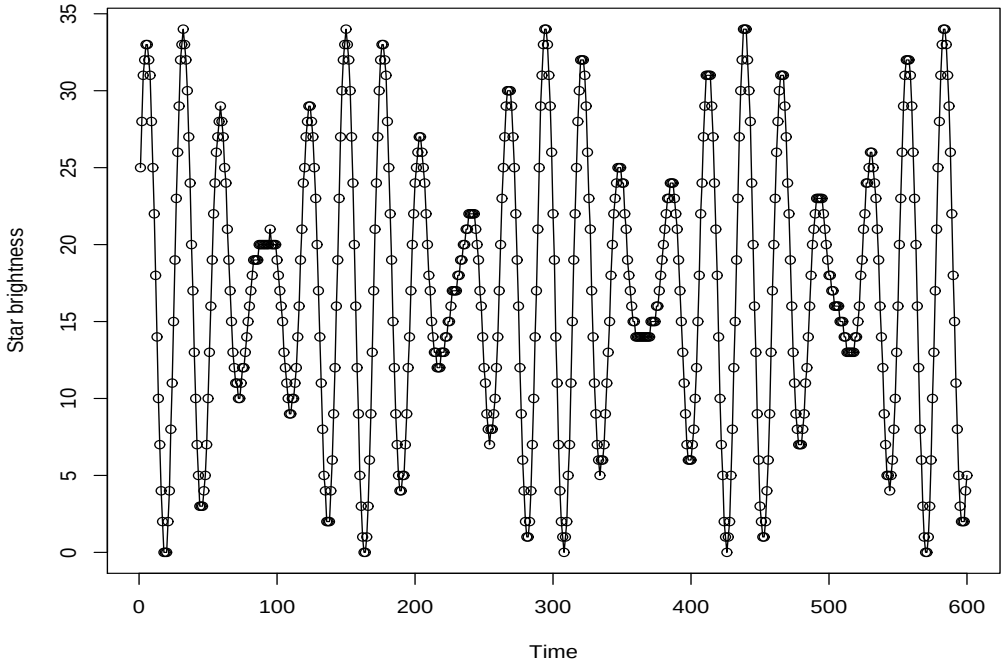


Figure 1.7: Star brightness data. Measurements for a single star taken over 600 consecutive nights.

Example 1.7. *Star brightness data.* Two factors determine the brightness of a star: its luminosity (how much energy it puts out in a given time) and its distance from the Earth. The data in Figure 1.7 are nightly brightness measurements (in magnitude) of a single star over a period of 600 nights.

- Data file: `star` (TSA)
- There are $n = 600$ observations.
- Measurements are taken **each night**.
- What are the noticeable patterns?
- Predictions?

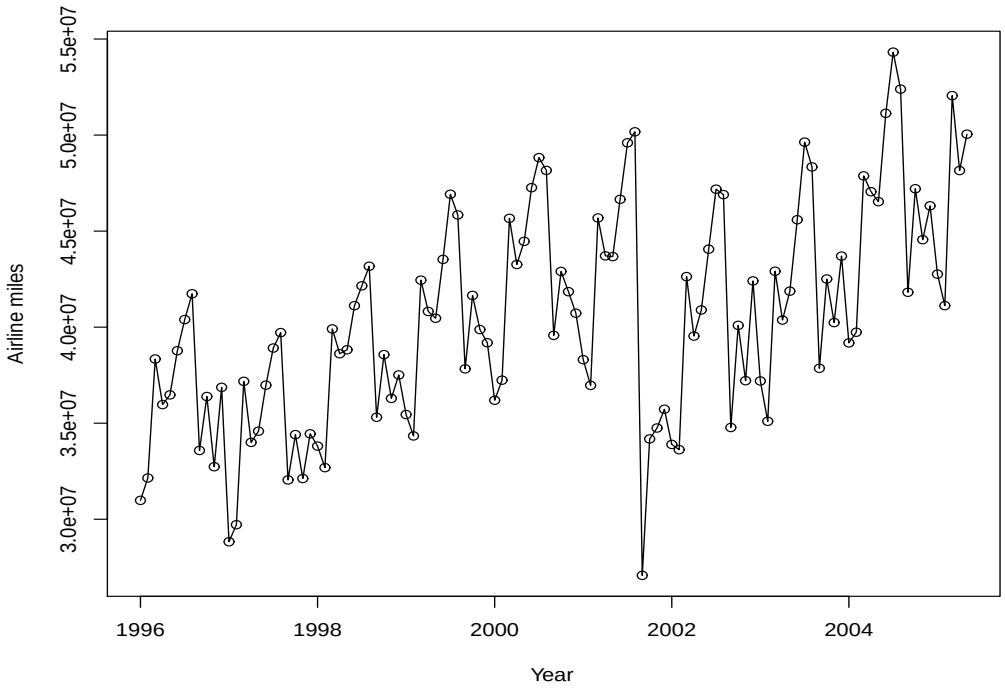


Figure 1.8: Airline passenger mile data. The number of miles, in thousands, traveled by passengers in the United States from January, 1996 to May, 2005.

Example 1.8. *Airline mile data.* The Bureau of Transportation Statistics publishes monthly passenger traffic data reflecting 100 percent of scheduled operations for airlines in the United States. The data in Figure 1.8 are monthly U.S. airline passenger miles traveled from 1/1996 to 5/2005.

- Data file: `airmiles` (TSA)
- There are $n = 113$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns?
- Predictions?

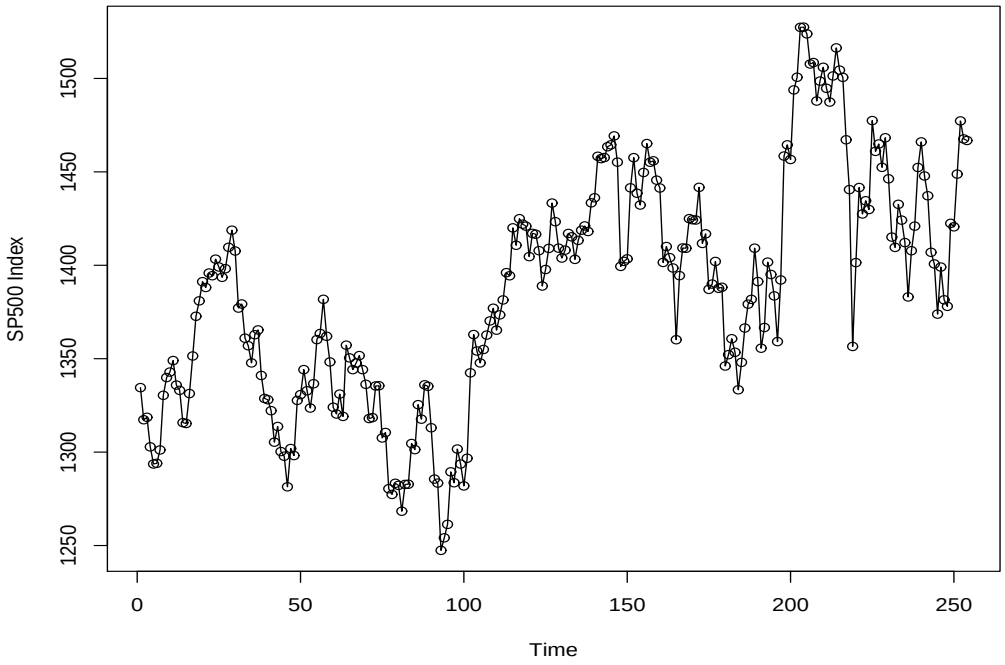


Figure 1.9: S&P Index price data. Daily values of the index from June 6, 1999 to June 5, 2000.

Example 1.9. *S&P*500 *Index data.* The S&P500 is a capitalization-weighted index (published since 1957) of the prices of 500 large-cap common stocks actively traded in the United States. The data in Figure 1.9 are the daily S&P500 Index prices measured during June 6, 1999 to June 5, 2000.

- Data file: `sp500`
- There are $n = 254$ observations.
- Measurements are taken **each trading day**.
- What are the noticeable patterns?
- Predictions?

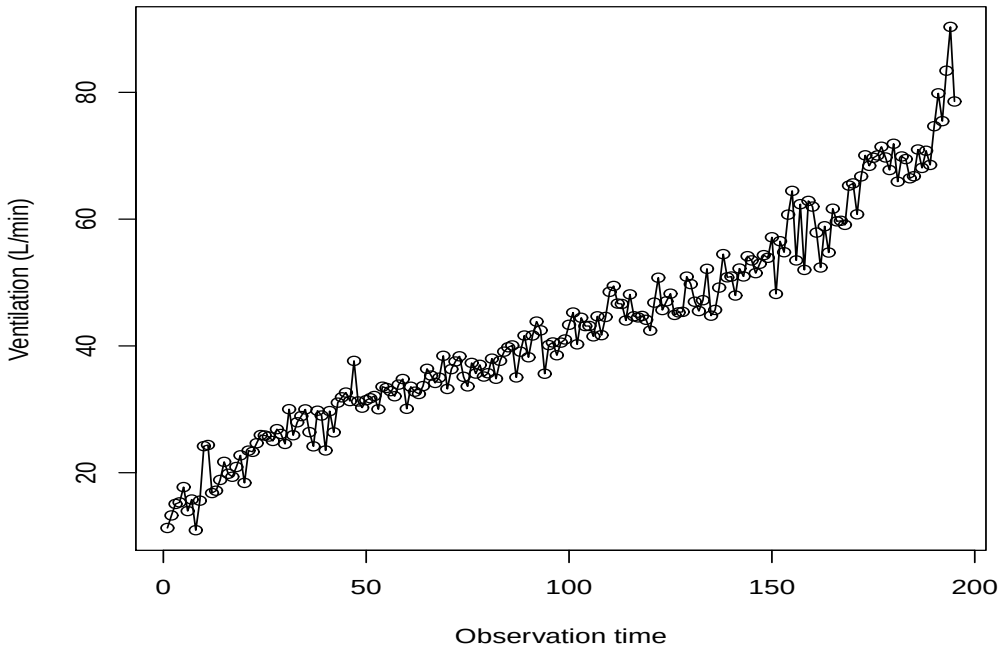


Figure 1.10: Ventilation data. Ventilation measurements on a single cyclist at 15 second intervals.

Example 1.10. *Ventilation data.* Collecting expired gases during exercise allows one to quantify many outcomes during an exercise test. One such outcome is the ventilatory threshold; i.e., the point at which lactate begins to accumulate in the blood. The data in Figure 1.10 are ventilation observations (L/min) on a single cyclist during exercise. Observations are recorded every 15 seconds. **Source:** Joe Alemany (Spring, 2010).

- Data file: `ventilation`
- There are $n = 195$ observations.
- Measurements are taken **each 15 seconds**.
- What are the noticeable patterns?
- Predictions?

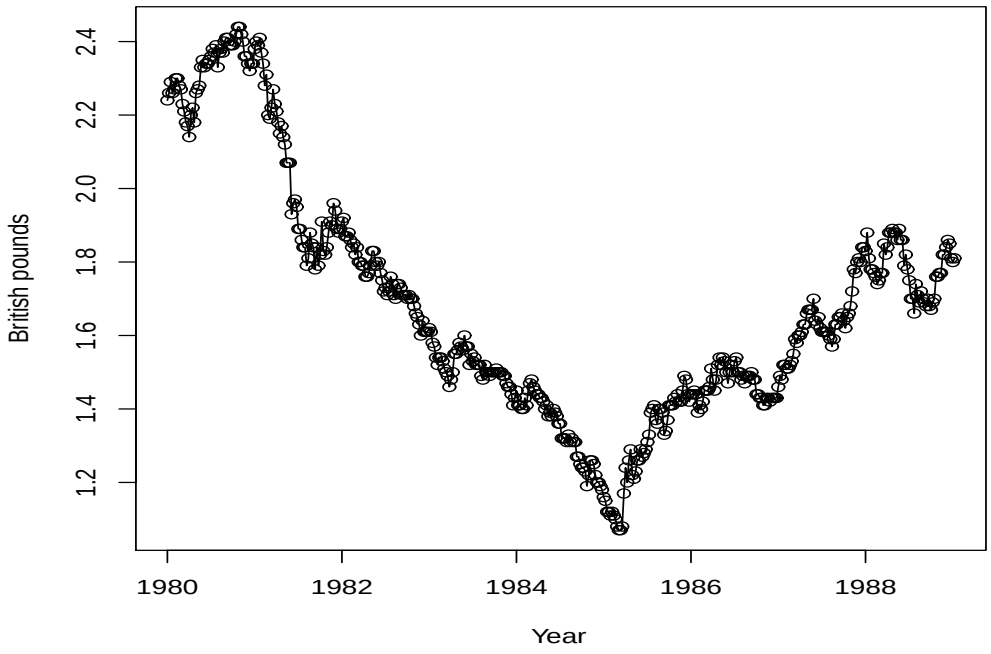


Figure 1.11: Exchange rate data. Weekly exchange rate of US dollar compared to the British pound, from 1980-1988.

Example 1.11. *Exchange rate data.* The pound sterling, often simply called "the pound," is the currency of the United Kingdom and many of its territories. The data in Figure 1.11 are weekly exchange rates of the US dollar and the British pound between the years 1980 and 1988.

- Data file: `exchangerate`
- There are $n = 470$ observations.
- Measurements are taken **each week**.
- What are the noticeable patterns?
- Predictions?

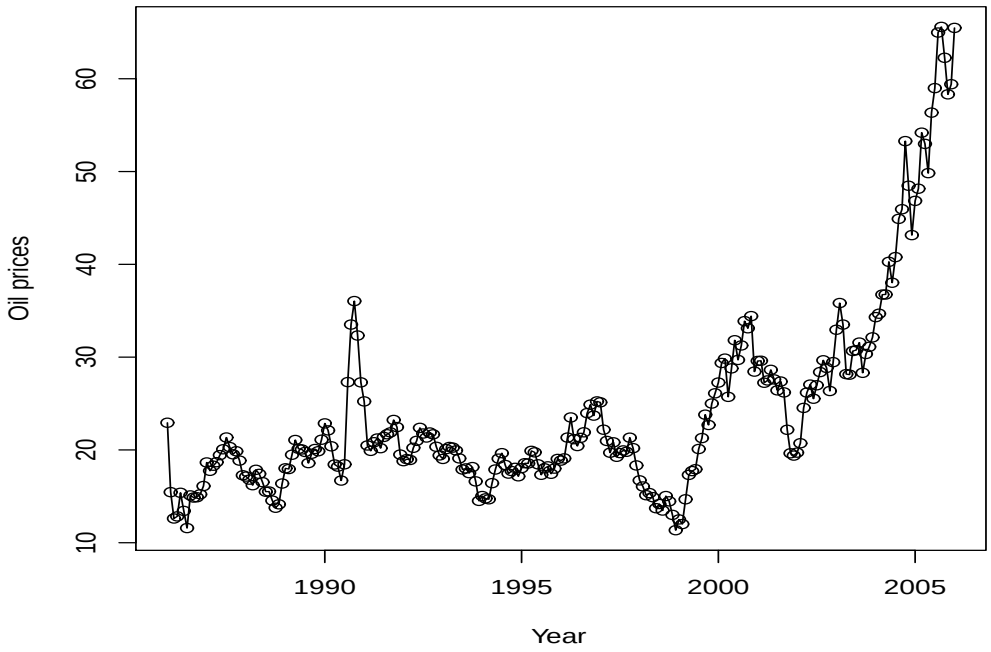


Figure 1.12: Crude oil price data. Monthly spot prices in dollars from Cushing, OK, from 1/1986 to 1/2006.

Example 1.12. *Oil price data.* Crude oil prices behave much as any other commodity with wide price swings in times of shortage or oversupply. The crude oil price cycle may extend over several years responding to changes in demand. The data in Figure 1.12 are monthly spot prices for crude oil (measured in U.S. dollars per barrel) from Cushing, OK.

- Data file: `oil.price` (TSA)
- There are $n = 241$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns?
- Predictions?

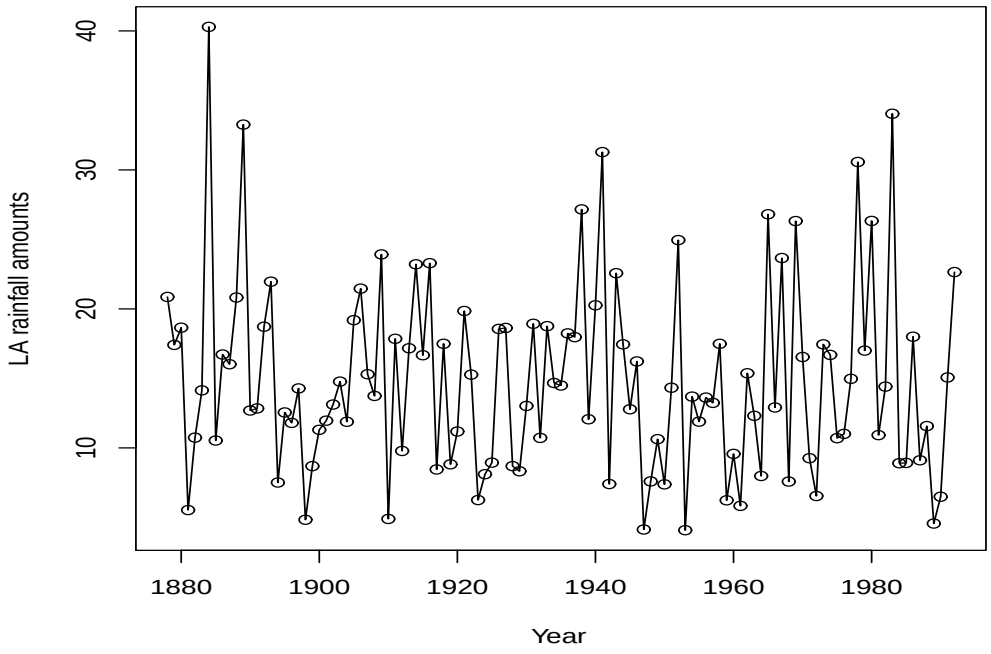


Figure 1.13: Los Angeles rainfall data. Annual precipitation measurements, in inches, during 1878-1992.

Example 1.13. *Annual rainfall data.* Los Angeles averages 15 inches of precipitation annually, which mainly occurs during the winter and spring (November through April) with generally light rain showers, but sometimes as heavy rainfall and thunderstorms. The data in Figure 1.13 are annual rainfall totals for Los Angeles during 1878-1992.

- Data file: `larain` (TSA)
- There are $n = 115$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

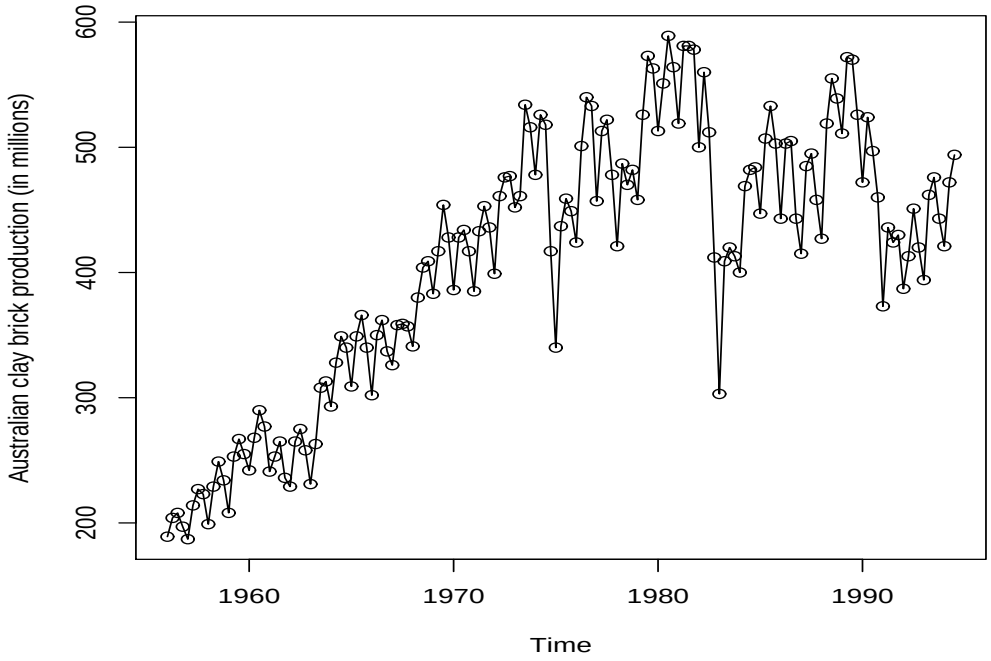


Figure 1.14: Australian clay brick production data. Number of bricks (in millions) produced from 1956-1994.

Example 1.14. *Brick production data.* Clay bricks remain extremely popular for the cladding of houses and small commercial buildings throughout Australia due to their versatility of use, tensile strength, thermal properties and attractive appearance. The data in Figure 1.14 represent the number of bricks produced in Australia (in millions) during 1956-1994. The data are quarterly.

- Data file: `brick`
- There are $n = 155$ observations.
- Measurements are taken **each quarter**.
- What are the noticeable patterns?
- Predictions?

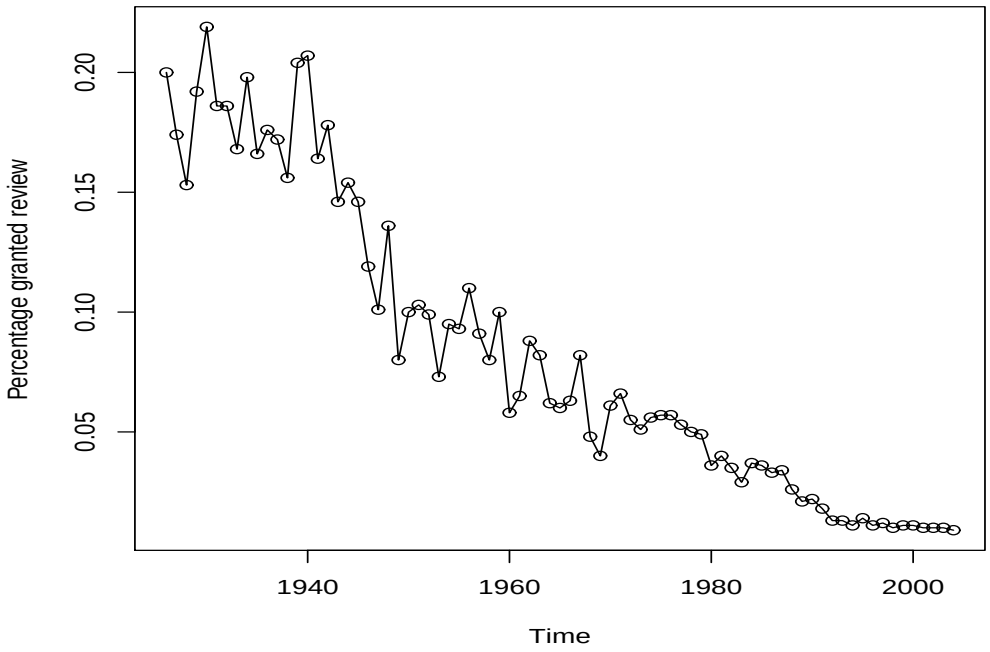


Figure 1.15: United States Supreme Court data. Percent of cases granted review during 1926-2004.

Example 1.15. *Supreme Court data.* The Supreme Court of the United States has ultimate (but largely discretionary) appellate jurisdiction over all state and federal courts, and original jurisdiction over a small range of cases. The data in Figure 1.15 represent the acceptance rate of cases appealed to the Supreme Court during 1926-2004. **Source:** Jim Manning (Spring, 2010).

- Data file: `supremecourt`
- There are $n = 79$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

IMPORTANCE: The purpose of time series analysis is twofold:

1. to **model** the stochastic (random) mechanism that gives rise to the series of data
2. to **predict** (forecast) the future values of the series based on the previous history.

NOTES: The analysis of time series data calls for a "new way of thinking" when compared to other statistical methods courses. Essentially, we get to see only a single measurement from a population (at time t) instead of a sample of measurements at a fixed point in time (**cross-sectional data**).

- The special feature of time series data is that they are not independent! Instead, observations are **correlated** through time.
  - Correlated data are generally more difficult to analyze.
  - Statistical theory in the absence of independence becomes markedly more difficult.
- Most classical statistical methods (e.g., regression, analysis of variance, etc.) assume that observations are statistically **independent**. For example, in the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

  or an ANOVA model like

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

  we typically assume that the ϵ error terms are independent and identically distributed (**iid**) normal random variables with mean 0 and constant variance.
- There can be additional **trends** or seasonal variation patterns (**seasonality**) that may be difficult to identify and model.
- The data may be highly non-normal in appearance and be possibly contaminated by **outliers**.

MODELING: Our overarching goal in this course is to build (and use) time series models for data. This breaks down into different parts.

1. Model specification (identification)
    - Consider different classes of time series models for **stationary processes**.
    - Use descriptive statistics, graphical displays, subject matter knowledge, etc. to make sensible candidate selections.
    - Abide by the **Principle of Parsimony**.
2. Model fitting
    - Once a candidate model is chosen, estimate the parameters in the model.
    - We will use **least squares** and/or **maximum likelihood** to do this.
3. Model diagnostics
    - Use statistical inference and graphical displays to check how well the model fits the data.
    - This part of the analysis may suggest the candidate model is inadequate and may point to more appropriate models.

TIME SERIES PLOT: The **time series plot** is the most basic graphical display in the analysis of time series data. The plot is a basically a scatterplot of Y_t versus t, with straight lines connecting the points. Notationally,

$$Y_t = \text{value of the variable } Y \text{ at time } t, \text{ for } t = 1, 2, ..., n.$$

The subscript t tells us to which time point the measurement Y_t corresponds. Note that in the sequence $Y_1, Y_2, ..., Y_n$, the subscripts are very important because they correspond to a particular ordering of the data. This is perhaps a change in mind set from other methods courses where the time element is ignored.

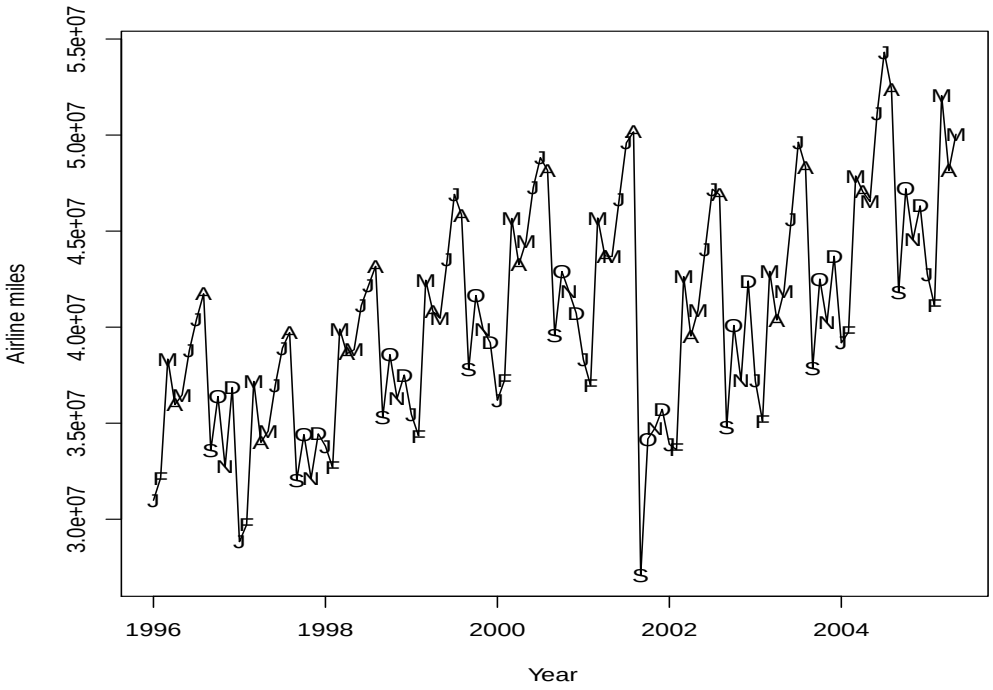


Figure 1.16: Airline passenger mile data. The number of miles, in thousands, traveled by passengers in the United States from January, 1996 to May, 2005. Monthly plotting symbols have been added.

GRAPHICS: The time series plot is vital, both to describe the data and to help formulating a sensible model. Here are some simple, but important, guidelines when constructing these plots.

- Give a clear, self-explanatory title or figure caption.
- State the units of measurement in the axis labels or figure caption.
- Choose the scales carefully (including the size of the intercept). Default settings from software may be sufficient.
- Label axes clearly.
- Use special plotting symbols where appropriate; e.g., months of the year, days of the week, actual numerical values for outlying values, etc.

2 Fundamental Concepts

Complementary reading: Chapter 2 (CC).

2.1 Summary of important distribution theory

DISCLAIMER: Going forward, we must be familiar with the following results from probability and distribution theory (e.g., STAT 511, etc.). If you have not had this material, you should find a suitable reference and study up on your own. See also pp 24-26 (CC).

REVIEW: Informally, a **random variable** Y is a variable whose value can not be predicted with certainty. Instead, the variable is said to vary according to a **probability distribution** which describes which values Y can assume and with what probability it assumes those values. There are basically two types of random variables. **Discrete** random variables take on specific values with positive probability. **Continuous** random variables have positive probability assigned to intervals of possible values. In this course, we will restrict attention to random variables Y which are best viewed as continuous (or at least quantitative).

2.1.1 Univariate random variables

DEFINITION: The **(cumulative) distribution function (cdf)** of a random variable Y, denoted by $F_Y(y)$, is a function that gives the probability $F_Y(y) = P(Y \leq y)$, for all $-\infty < y < \infty$. Mathematically, a random variable Y is said to be continuous if its cdf $F_Y(y)$ is a continuous function of y.

TERMINOLOGY: Let Y be a continuous random variable with cdf $F_Y(y)$. The **probability density function** (**pdf**) for Y, denoted by $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy} F_Y(y),$$

provided that this derivative exists.

PROPERTIES: Suppose that Y is a **continuous** random variable with pdf $f_Y(y)$ and support R (that is, the set of all values that Y can assume). Then

(1) $f_Y(y) > 0$, for all $y \in R$,

(2) the function $f_Y(y)$ satisfies $\int_R f_Y(y)dy = 1$.

RESULT: Suppose Y is a continuous random variable with pdf $f_Y(y)$ and cdf $F_Y(y)$. Then

$$P(a < Y < b) = \int_a^b f_Y(y)dy = F_Y(b) - F_Y(a).$$

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R. The **expected value** (or **mean**) of Y is given by

$$E(Y) = \int_R y f_Y(y)dy.$$

Mathematically, we require that

$$\int_R |y| f_Y(y)dy < \infty.$$

If this is not true, then we say that $E(Y)$ does not exist. If g is a real-valued function, then $g(Y)$ is a random variable and

$$E[g(Y)] = \int_R g(y) f_Y(y)dy,$$

provided that this integral exists.

PROPERTIES OF EXPECTATIONS: Let Y be a random variable with pdf $f_Y(y)$ and support R, suppose that $g, g_1, g_2, ..., g_k$ are real-valued functions, and let a be any real constant. Then

(a) $E(a) = a$

(b) $E[ag(Y)] = aE[g(Y)]$

(c) $E[\sum_{j=1}^k g_j(Y)] = \sum_{j=1}^k E[g_j(Y)]$.

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$, support R, and mean $E(Y) = \mu$. The **variance** of Y is given by

$$\text{var}(Y) = E[(Y-\mu)^2] = \int_R (y-\mu)^2 f_Y(y)dy.$$

In general, it will be easier to use the **variance computing formula**

$$\text{var}(Y) = E(Y^2) - [E(Y)]^2.$$

We will often use the statistical symbol σ^2 or σ^2_Y to denote var(Y).

FACTS:

(a) var$(Y) \geq 0$. var$(Y) = 0$ if and only if the random variable Y has a **degenerate distribution**; i.e., all the probability mass is located at one support point.

(b) The larger (smaller) var(Y) is, the more (less) spread in the possible values of Y about the mean $\mu = E(Y)$.

(c) var(Y) is measured in (units)^2. The standard deviation of Y is $\sigma = \sqrt{\sigma^2} = \sqrt{\text{var}(Y)}$ and is measured in the original units of Y.

IMPORTANT RESULT: Let Y be a random variable, and suppose that a and b are fixed constants. Then

$$\text{var}(a + bY) = b^2\text{var}(Y).$$

2.1.2 Bivariate random vectors

TERMINOLOGY: Let X and Y be continuous random variables. (X, Y) is called a **continuous random vector**, and the **joint probability density function (pdf)** of X and Y is denoted by $f_{X,Y}(x, y)$.

PROPERTIES: The function $f_{X,Y}(x, y)$ has the following properties:

(1) $f_{X,Y}(x, y) > 0$, for all $(x, y) \in R \subseteq \mathcal{R}^2$

(2) The function $f_{X,Y}(x, y)$ satisfies $\int\int_R f_{X,Y}(x, y)dxdy = 1$.

RESULT: Suppose (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$. Then

$$P[(X, Y) \in B] = \int \int_B f_{X,Y}(x, y)dxdy,$$

for any set $B \subset \mathcal{R}^2$.

TERMINOLOGY: Suppose that (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$. The **joint cumulative distribution function (cdf)** for (X, Y) is given by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t, s)dtds,$$

for all $(x, y) \in \mathcal{R}^2$. It follows upon differentiation that the joint pdf is given by

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y),$$

wherever this mixed partial derivative is defined.

RESULT: Suppose that (X, Y) has joint pdf $f_{X,Y}(x, y)$ and support R. Let $g(X, Y)$ be a real vector valued function of (X, Y); i.e., $g : \mathcal{R}^2 \to \mathcal{R}$. Then

$$E[g(X, Y)] = \int \int_R g(x, y) f_{X,Y}(x, y)dxdy.$$

If this quantity is not finite, then we say that $E[g(X, Y)]$ does not exist.

PROPERTIES OF EXPECTATIONS: Let (X, Y) be a random vector, suppose that $g, g_1, g_2, ..., g_k$ are real vector valued functions from $\mathcal{R}^2 \to \mathcal{R}$, and let a be any real constant. Then

(a) $E(a) = a$

(b) $E[ag(X, Y)] = aE[g(X, Y)]$

(c) $E[\sum_{j=1}^{k} g_j(X, Y)] = \sum_{j=1}^{k} E[g_j(X, Y)]$.

TERMINOLOGY: Suppose that (X, Y) is a continuous random vector with joint cdf $F_{X,Y}(x, y)$, and denote the marginal cdfs of X and Y by $F_X(x)$ and $F_Y(y)$, respectively. The random variables X and Y are **independent** if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

for all values of x and y. It can hence be shown that X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

for all values of x and y. That is, the joint pdf $f_{X,Y}(x,y)$ factors into the product the marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively.

RESULT: Suppose that X and Y are **independent** random variables. Let $g(X)$ be a function of X only, and let $h(Y)$ be a function of Y only. Then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)],$$

provided that all expectations exist. Taking $g(X) = X$ and $h(Y) = Y$, we get

$$E(XY) = E(X)E(Y).$$

TERMINOLOGY: Suppose that X and Y are random variables with means $E(X) = \mu_X$ and $E(Y) = \mu_Y$, respectively. The **covariance** between X and Y is

$$\begin{aligned} \text{cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

The latter expression is called the **covariance computing formula**. The covariance is a numerical measure that describes how two variables are linearly related.

- If $\text{cov}(X,Y) > 0$, then X and Y are positively linearly related.
- If $\text{cov}(X,Y) < 0$, then X and Y are negatively linearly related.
- If $\text{cov}(X,Y) = 0$, then X and Y are not linearly related.

RESULT: If X and Y are independent, then $\text{cov}(X,Y) = 0$. The converse is not necessarily true.

RESULT: Suppose that X and Y are random variables.

$$\begin{aligned} \text{var}(X+Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) \\ \text{var}(X-Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y). \end{aligned}$$

RESULT: Suppose that X and Y are **independent** random variables.

$$\begin{aligned} \text{var}(X+Y) &= \text{var}(X) + \text{var}(Y) \\ \text{var}(X-Y) &= \text{var}(X) + \text{var}(Y). \end{aligned}$$

RESULTS: Suppose that X and Y are random variables. The covariance operator satisfies the following:

(a) $\text{cov}(X,Y) = \text{cov}(Y,X)$

(b) $\text{cov}(X,X) = \text{var}(X)$.

(c) $\text{cov}(a+bX, c+dY) = bd\text{cov}(X,Y)$, for any constants a, b, c, and d.

DEFINITION: Suppose that X and Y are random variables. The **correlation** between X and Y is defined by

$$\rho = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

NOTES:

(1) $-1 \leq \rho \leq 1$.

(2) If $\rho = 1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 > 0$. That is, X and Y are perfectly positively linearly related; i.e., the bivariate probability distribution of (X,Y) lies entirely on a straight line with positive slope.

(3) If $\rho = -1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 < 0$. That is, X and Y are perfectly negatively linearly related; i.e., the bivariate probability distribution of (X,Y) lies entirely on a straight line with negative slope.

(4) If $\rho = 0$, then X and Y are not linearly related.

RESULT: If X and Y are independent, then $\rho = \rho_{X,Y} = 0$. The converse is not true in general. However,

$$\rho = \text{corr}(X,Y) = 0 \Longrightarrow X \text{ and } Y \text{ independent}$$

when (X,Y) has a **bivariate normal distribution**.

2.1.3 Multivariate extensions and linear combinations

EXTENSION: We use the notation $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ and $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. The joint cdf of $\boldsymbol{Y}$ is

$$\begin{aligned} F_{\boldsymbol{Y}}(\boldsymbol{y}) &= P(Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_n \leq y_n) \\ &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \cdots \int_{-\infty}^{y_n} f_{\boldsymbol{Y}}(\boldsymbol{t}) dt_1 dt_2 \cdots dt_n, \end{aligned}$$

where $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ and $f_{\boldsymbol{Y}}(\boldsymbol{y})$ denotes the joint pdf of $\boldsymbol{Y}$.

EXTENSION: Suppose that the random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ has joint cdf $F_{\boldsymbol{Y}}(\boldsymbol{y})$, and suppose that the random variable Y_i has cdf $F_{Y_i}(y_i)$, for $i = 1, 2, ..., n$. Then, $Y_1, Y_2, ..., Y_n$ are independent random variables if and only if

$$F_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} F_{Y_i}(y_i);$$

that is, the joint cdf can be factored into the product of the marginal cdfs. Alternatively, $Y_1, Y_2, ..., Y_n$ are independent random variables if and only if

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i);$$

that is, the joint pdf can be factored into the product of the marginal pdfs.

MATHEMATICAL EXPECTATION: Suppose that $Y_1, Y_2, ..., Y_n$ are (mutually) **independent** random variables. For real valued functions $g_1, g_2, ..., g_n$,

$$E[g_1(Y_1)g_2(Y_2)\cdots g_n(Y_n)] = E[g_1(Y_1)]E[g_2(Y_2)]\cdots E[g_n(Y_n)],$$

provided that each expectation exists.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ are random variables and that $a_1, a_2, ..., a_n$ are constants. The function

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is called a **linear combination** of the random variables $Y_1, Y_2, ..., Y_n$.

REMARK: Linear combinations are commonly seen in the theoretical development of time series models. We therefore must be familiar with the following results.

EXPECTED VALUE OF A LINEAR COMBINATION:

$$E(U) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i)$$

VARIANCE OF A LINEAR COMBINATION:

$$\begin{aligned} \text{var}(U) = \text{var}\left(\sum_{i=1}^{n} a_i Y_i\right) &= \sum_{i=1}^{n} a_i^2 \text{var}(Y_i) + 2\sum_{i<j} a_i a_j \text{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2 \text{var}(Y_i) + \sum_{i \neq j} a_i a_j \text{cov}(Y_i, Y_j) \end{aligned}$$

COVARIANCE BETWEEN TWO LINEAR COMBINATIONS: Suppose that

$$\begin{aligned} U_1 &= \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n \\ U_2 &= \sum_{j=1}^{m} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m. \end{aligned}$$

Then,

$$\text{cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{cov}(Y_i, X_j).$$

2.1.4 Miscellaneous

GEOMETRIC SUMS: Suppose that a is any real number and that $|r| < 1$. Then, the finite geometric sum

$$\sum_{j=0}^{n} ar^j = \frac{a(1 - r^{n+1})}{1 - r}.$$

Taking limits of both sides, we get

$$\sum_{j=0}^{\infty} ar^j = \frac{a}{1 - r}.$$

These formulas should be committed to memory.

2.2 Time series and stochastic processes

TERMINOLOGY: The sequence of random variables $\{Y_t : t = 0, 1, 2, ...,\}$, or more simply denoted by $\{Y_t\}$, is called a **stochastic process**. It is a collection of random variables indexed by time t; that is,

$$\begin{aligned} Y_0 &= \text{value of the process at time } t = 0 \\ Y_1 &= \text{value of the process at time } t = 1 \\ Y_2 &= \text{value of the process at time } t = 2 \\ &\vdots \\ Y_n &= \text{value of the process at time } t = n. \end{aligned}$$

The subscripts are important because they refer to which time period the value of Y is being measured. A stochastic process can be described as "a statistical phenomenon that evolves through time according to a set of probabilistic laws."

- A complete probabilistic time series model for $\{Y_t\}$, in fact, would specify all of the **joint distributions** of random vectors $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$, for all $n = 1, 2, ...,$ or, equivalently, specify the joint probabilities

$$P(Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_n \leq y_n),$$

  for all $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ and $n = 1, 2, ...,$.

- This specification is not generally needed in practice. In this course, we specify only the first and second-order moments; i.e., expectations of the form $E(Y_t)$ and $E(Y_t Y_{t-k})$, for $k = 0, 1, 2, ...,$ and $t = 0, 1, 2, ...$

- Much of the important information in most time series processes is captured in these first and second moments (or, equivalently, in the means, variances, and covariances).

2.3 Means, variances, and covariances

TERMINOLOGY: For the stochastic process $\{Y_t : t = 0, 1, 2, ..., \}$, the **mean function** is defined as

$$\mu_t = E(Y_t),$$

for $t = 0, 1, 2, ...,$. That is, μ_t is the theoretical (or population) mean for the series at time t. The **autocovariance function** is defined as

$$\gamma_{t,s} = \text{cov}(Y_t, Y_s),$$

for $t, s = 0, 1, 2, ...,$ where $\text{cov}(Y_t, Y_s) = E(Y_t Y_s) - E(Y_t)E(Y_s)$. The **autocorrelation function** is given by

$$\rho_{t,s} = \text{corr}(Y_t, Y_s),$$

where

$$\text{corr}(Y_t, Y_s) = \frac{\text{cov}(Y_t, Y_s)}{\sqrt{\text{var}(Y_t)\text{var}(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}.$$

- Values of $\rho_{t,s}$ near $\pm 1 \Longrightarrow$ strong linear dependence between Y_t and Y_s.
- Values of $\rho_{t,s}$ near $0 \Longrightarrow$ weak linear dependence between Y_t and Y_s
- Values of $\rho_{t,s} = 0 \Longrightarrow$ Y_t and Y_s are **uncorrelated**.

2.4 Some (named) stochastic processes

Example 2.1. A stochastic process $\{e_t : t = 0, 1, 2, ..., \}$ is called a **white noise process** if it is a sequence of independent and identically distributed (iid) random variables with

$$\begin{aligned} E(e_t) &= \mu_e \\ \text{var}(e_t) &= \sigma_e^2. \end{aligned}$$

- Both μ_e and σ_e^2 are constant (free of t).

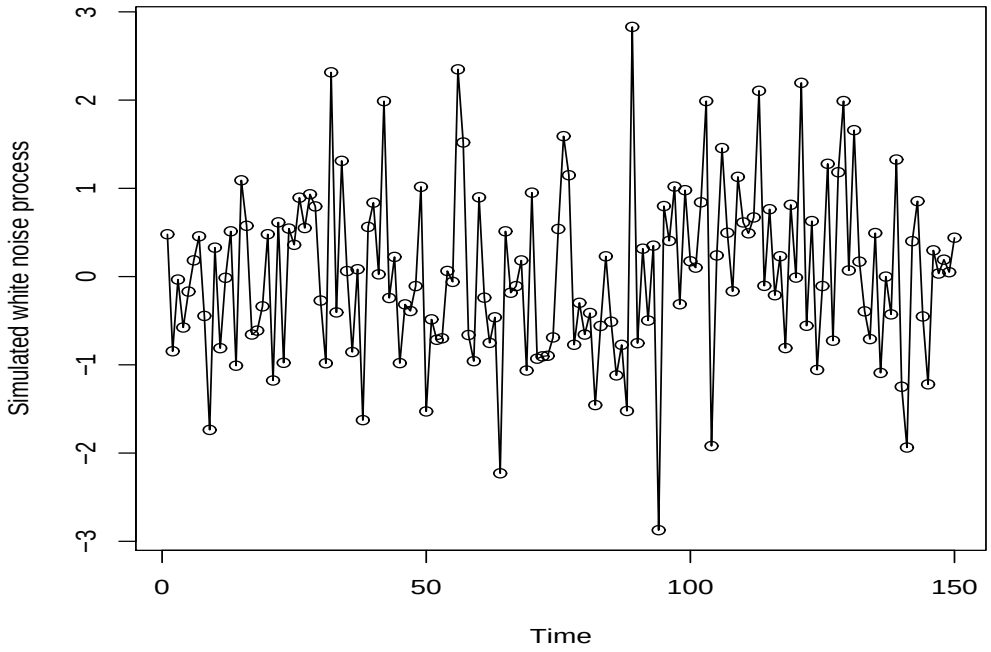


Figure 2.1: A simulated white noise process $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, where $n = 150$ and $\sigma_e^2 = 1$.

- It is often assumed that $\mu_e = 0$; that is, $\{e_t\}$ is a zero mean process.
- A slightly less restrictive definition would require that the e_t's are uncorrelated (not independent). However, under normality; i.e., $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, this distinction becomes vacuous (for linear time series models).

AUTOCOVARIANCE FUNCTION: For $t = s$,

$$\text{cov}(e_t, e_s) = \text{cov}(e_t, e_t) = \text{var}(e_t) = \sigma_e^2.$$

For $t \neq s$,

$$\text{cov}(e_t, e_s) = 0,$$

because the e_t's are independent. Thus, the autocovariance function of $\{e_t\}$ is

$$\gamma_{t,s} = \begin{cases} \sigma_e^2, & |t-s| = 0 \\ 0, & |t-s| \neq 0. \end{cases}$$

AUTOCORRELATION FUNCTION: For $t = s$,

$$\rho_{t,s} = \text{corr}(e_t, e_s) = \text{corr}(e_t, e_t) = \frac{\gamma_{t,t}}{\sqrt{\gamma_{t,t}\gamma_{t,t}}} = 1.$$

For $t \neq s$,

$$\rho_{t,s} = \text{corr}(e_t, e_s) = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = 0.$$

Thus, the autocorrelation function is

$$\rho_{t,s} = \begin{cases} 1, & |t-s| = 0 \\ 0, & |t-s| \neq 0. \end{cases}$$

REMARK: A white noise process, by itself, is rather uninteresting for modeling real data. However, white noise processes still play a crucial role in the analysis of time series data! Time series processes $\{Y_t\}$ generally contain two different types of variation:

- **systematic** variation (that we would like to capture and model; e.g., trends, seasonal components, etc.)
- **random** variation (that is just inherent background noise in the process).

Our goal as data analysts is to extract the systematic part of the variation in the data (and incorporate this into our model). If we do an adequate job of extracting the systematic part, then the only part "left over" should be random variation, which can be modeled as white noise.

Example 2.2. Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Define

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_n &= e_1 + e_2 + \cdots + e_n. \end{aligned}$$

By this definition, note that we can write, for $t > 1$,

$$Y_t = Y_{t-1} + e_t,$$

where $E(e_t) = 0$ and $\text{var}(e_t) = \sigma_e^2$. The process $\{Y_t\}$ is called a **random walk process**. Random walk processes are used to model stock prices, movements of molecules in gases and liquids, animal locations, etc.

MEAN FUNCTION: The mean of Y_t is

$$\begin{aligned} \mu_t &= E(Y_t) \\ &= E(e_1 + e_2 + \cdots + e_t) \\ &= E(e_1) + E(e_2) + \cdots + E(e_t) = 0. \end{aligned}$$

That is, $\{Y_t\}$ is a zero mean process.

VARIANCE FUNCTION: The variance of Y_t is

$$\begin{aligned} \text{var}(Y_t) &= \text{var}(e_1 + e_2 + \cdots + e_t) \\ &= \text{var}(e_1) + \text{var}(e_2) + \cdots + \text{var}(e_t) = t\sigma_e^2, \end{aligned}$$

because $\text{var}(e_1) = \text{var}(e_2) = \cdots = \text{var}(e_t) = \sigma_e^2$ and $\text{cov}(e_t, e_s) = 0$ for all $t \neq s$.

AUTOCOVARIANCE FUNCTION: For $t \leq s$, the autocovariance of Y_t and Y_s is

$$\begin{aligned} \gamma_{t,s} = \text{cov}(Y_t, Y_s) &= \text{cov}(e_1 + e_2 + \cdots + e_t,\ e_1 + e_2 + \cdots + e_t + e_{t+1} + \cdots + e_s) \\ &= \text{cov}(e_1 + e_2 + \cdots + e_t,\ e_1 + e_2 + \cdots + e_t) \\ &\qquad + \text{cov}(e_1 + e_2 + \cdots + e_t,\ e_{t+1} + \cdots + e_s) \\ &= \sum_{i=1}^{t} \text{cov}(e_i, e_i) + \sum\sum_{1 \leq i \neq j \leq t} \text{cov}(e_i, e_j) \\ &= \sum_{i=1}^{t} \text{var}(e_i) = \sigma_e^2 + \sigma_e^2 + \cdots + \sigma_e^2 = t\sigma_e^2. \end{aligned}$$

Because $\gamma_{t,s} = \gamma_{s,t}$, the autocovariance function for a random walk process is

$$\gamma_{t,s} = t\sigma_e^2, \quad \text{for } 1 \leq t \leq s.$$

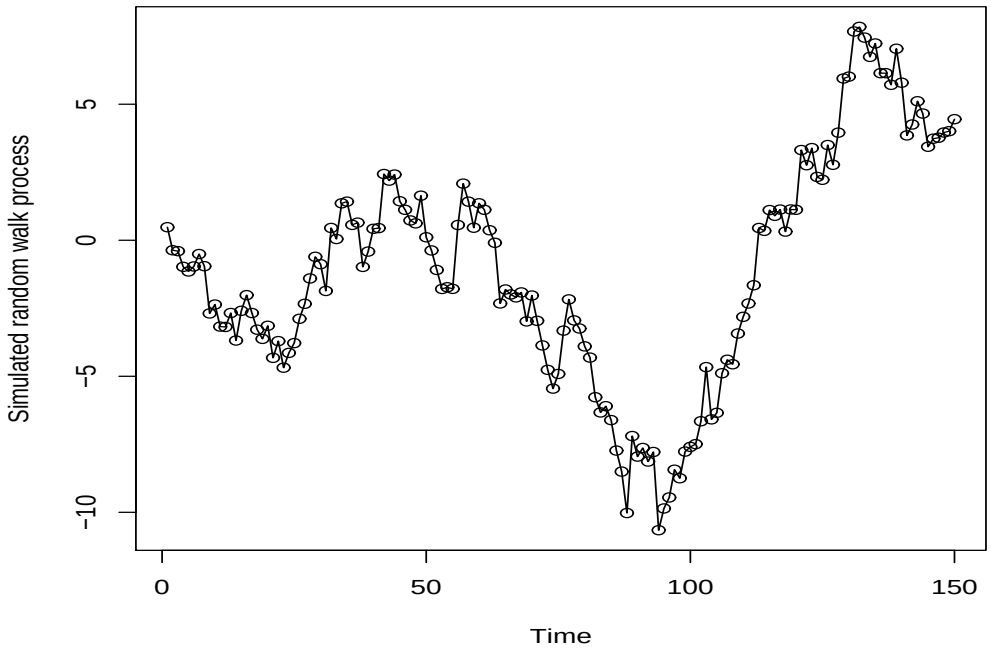


Figure 2.2: A simulated random walk process $Y_t = Y_{t-1} + e_t$, where $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, $n = 150$, and $\sigma_e^2 = 1$. This process has been constructed from the simulated white noise process $\{e_t\}$ in Figure 2.1.

AUTOCORRELATION FUNCTION: For $1 \leq t \leq s$, the autocorrelation function for a random walk process is

$$\rho_{t,s} = \text{corr}(Y_t, Y_s) = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \frac{t\sigma_e^2}{\sqrt{t\sigma_e^2 s\sigma_e^2}} = \sqrt{\frac{t}{s}}.$$

- Note that when t is closer to s, the autocorrelation $\rho_{t,s}$ is closer to 1. That is, two observations Y_t and Y_s close together in time are likely to be close together, especially when t and s are both large (later on in the series).
- On the other hand, when t is far away from s (that is, for two points Y_t and Y_s far apart in time), the autocorrelation is closer to 0.

Example 2.3. Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Define

$$Y_t = \frac{1}{3}(e_t + e_{t-1} + e_{t-2}),$$

that is, Y_t is a running (or **moving**) **average** of the white noise process (averaged across the most recent 3 time periods). Note that this example is slightly different than that on pp 14-15 (CC).

MEAN FUNCTION: The mean of Y_t is

$$\begin{aligned} \mu_t = E(Y_t) &= E\left[\frac{1}{3}(e_t + e_{t-1} + e_{t-2})\right] \\ &= \frac{1}{3}[E(e_t) + E(e_{t-1}) + E(e_{t-2})] = 0, \end{aligned}$$

because $\{e_t\}$ is a zero-mean process. $\{Y_t\}$ is a zero mean process.

VARIANCE FUNCTION: The variance of Y_t is

$$\begin{aligned} \text{var}(Y_t) &= \text{var}\left[\frac{1}{3}(e_t + e_{t-1} + e_{t-2})\right] \\ &= \frac{1}{9}\text{var}(e_t + e_{t-1} + e_{t-2}) \\ &= \frac{1}{9}[\text{var}(e_t) + \text{var}(e_{t-1}) + \text{var}(e_{t-2})] = \frac{3\sigma_e^2}{9} = \frac{\sigma_e^2}{3}, \end{aligned}$$

because $\text{var}(e_t) = \sigma_e^2$ for all t and because e_t, e_{t-1}, and e_{t-2} are independent (all covariance terms are zero).

AUTOCOVARIANCE FUNCTION: We need to consider different cases.

Case 1: If $s = t$, then

$$\gamma_{t,s} = \gamma_{t,t} = \text{cov}(Y_t, Y_t) = \text{var}(Y_t) = \frac{\sigma_e^2}{3}.$$

Case 2: If $s = t + 1$, then

$$\begin{aligned} \gamma_{t,s} = \gamma_{t,t+1} &= \text{cov}(Y_t, Y_{t+1}) \\ &= \text{cov}\left[\frac{1}{3}(e_t + e_{t-1} + e_{t-2}), \frac{1}{3}(e_{t+1} + e_t + e_{t-1})\right] \\ &= \frac{1}{9}[\text{cov}(e_t, e_t) + \text{cov}(e_{t-1}, e_{t-1})] \\ &= \frac{1}{9}[\text{var}(e_t) + \text{var}(e_{t-1})] = \frac{2\sigma_e^2}{9}. \end{aligned}$$

Case 3: If $s = t + 2$, then

$$\begin{aligned}\gamma_{t,s} = \gamma_{t,t+2} &= \text{cov}(Y_t, Y_{t+2}) \\ &= \text{cov}\left[\frac{1}{3}(e_t + e_{t-1} + e_{t-2}), \frac{1}{3}(e_{t+2} + e_{t+1} + e_t)\right] \\ &= \frac{1}{9}\text{cov}(e_t, e_t) \\ &= \frac{1}{9}\text{var}(e_t) = \frac{\sigma_e^2}{9}.\end{aligned}$$

Case 4: If $s > t + 2$, then $\gamma_{t,s} = 0$ because Y_t and Y_s will have no common white noise error terms.

Because $\gamma_{t,s} = \gamma_{s,t}$, the autocovariance function can be written as

$$\gamma_{t,s} = \begin{cases} \sigma_e^2/3, & |t-s| = 0 \\ 2\sigma_e^2/9, & |t-s| = 1 \\ \sigma_e^2/9, & |t-s| = 2 \\ 0, & |t-s| > 2. \end{cases}$$

AUTOCORRELATION FUNCTION: Recall that the autocorrelation function is

$$\rho_{t,s} = \text{corr}(Y_t, Y_s) = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}.$$

Because $\gamma_{t,t} = \gamma_{s,s} = \sigma_e^2/3$, the autocorrelation function for this process is

$$\rho_{t,s} = \begin{cases} 1, & |t-s| = 0 \\ 2/3, & |t-s| = 1 \\ 1/3, & |t-s| = 2 \\ 0, & |t-s| > 2. \end{cases}$$

- Observations Y_t and Y_s that are 1 unit apart in time have the same autocorrelation regardless of the values of t and s.
- Observations Y_t and Y_s that are 2 units apart in time have the same autocorrelation regardless of the values of t and s.
- Observations Y_t and Y_s that are more than 2 units apart in time are uncorrelated.

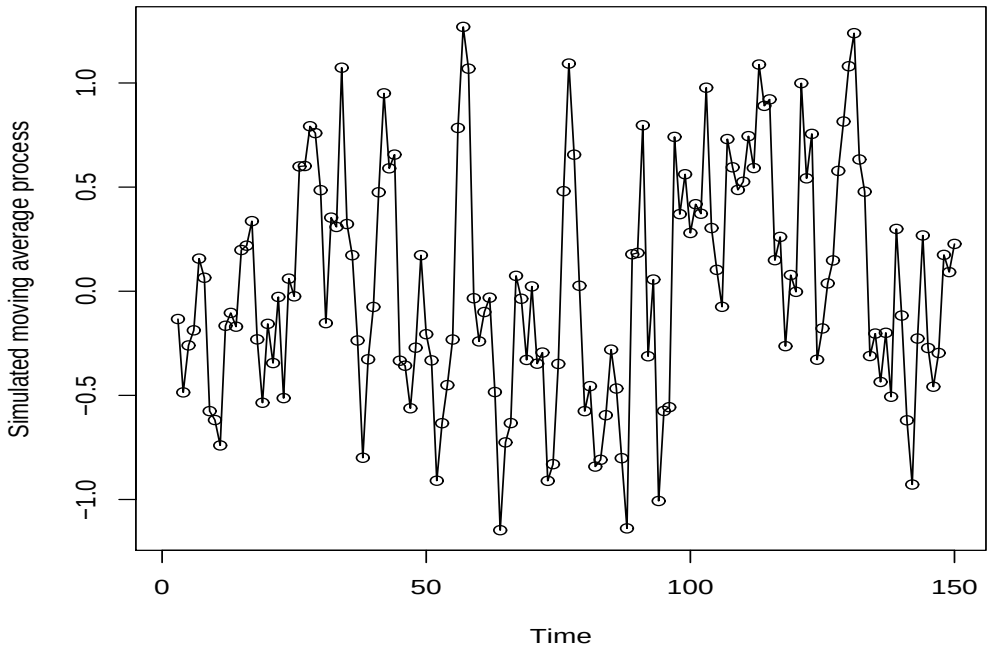


Figure 2.3: A simulated moving average process $Y_t = \frac{1}{3}(e_t + e_{t-1} + e_{t-2})$, where $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, $n = 150$, and $\sigma_e^2 = 1$. This process has been constructed from the simulated white noise process $\{e_t\}$ in Figure 2.1.

Example 2.4. Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Consider the stochastic process defined by

$$Y_t = 0.75Y_{t-1} + e_t,$$

that is, Y_t is directly related to the (downweighted) previous value of the process Y_{t-1} and the random error e_t (a "shock" or "innovation" that occurs at time t). This is called an **autoregressive model**. Autoregression means "regression on itself." Essentially, we can envision "regressing" Y_t on Y_{t-1}.

NOTE: We will postpone mean, variance, autocovariance, and autocorrelation calculations for this process until Chapter 4 when we discuss autoregressive models in more detail. A simulated realization of this process appears in Figure 2.4.

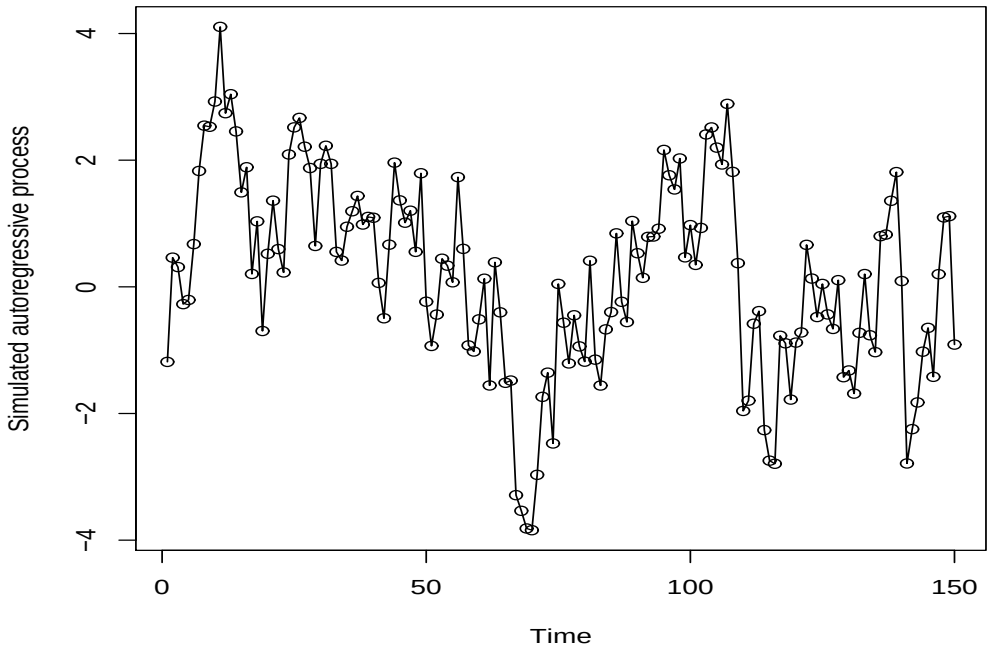


Figure 2.4: A simulated autoregressive process $Y_t = 0.75Y_{t-1}+e_t$, where $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, $n = 150$, and $\sigma_e^2 = 1$.

Example 2.5. Many time series exhibit seasonal patterns that correspond to different weeks, months, years, etc. One way to describe seasonal patterns is to use models with deterministic parts which are trigonometric in nature. Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Consider the process defined by

$$Y_t = a\sin(2\pi\omega t + \phi) + e_t.$$

In this model, a is the amplitude, ω is the frequency of oscillation, and ϕ controls the phase shift. With $a = 2$, $\omega = 1/52$ (one cycle/52 time points), and $\phi = 0.6\pi$, note that

$$E(Y_t) = 2\sin(2\pi t/52 + 0.6\pi),$$

since $E(e_t) = 0$. Also, $\text{var}(Y_t) = \text{var}(e_t) = \sigma_e^2$. The mean function, and three realizations of this process (one realization corresponding to $\sigma_e^2 = 1$, $\sigma_e^2 = 4$, and $\sigma_e^2 = 16$) are depicted in Figure 2.5.

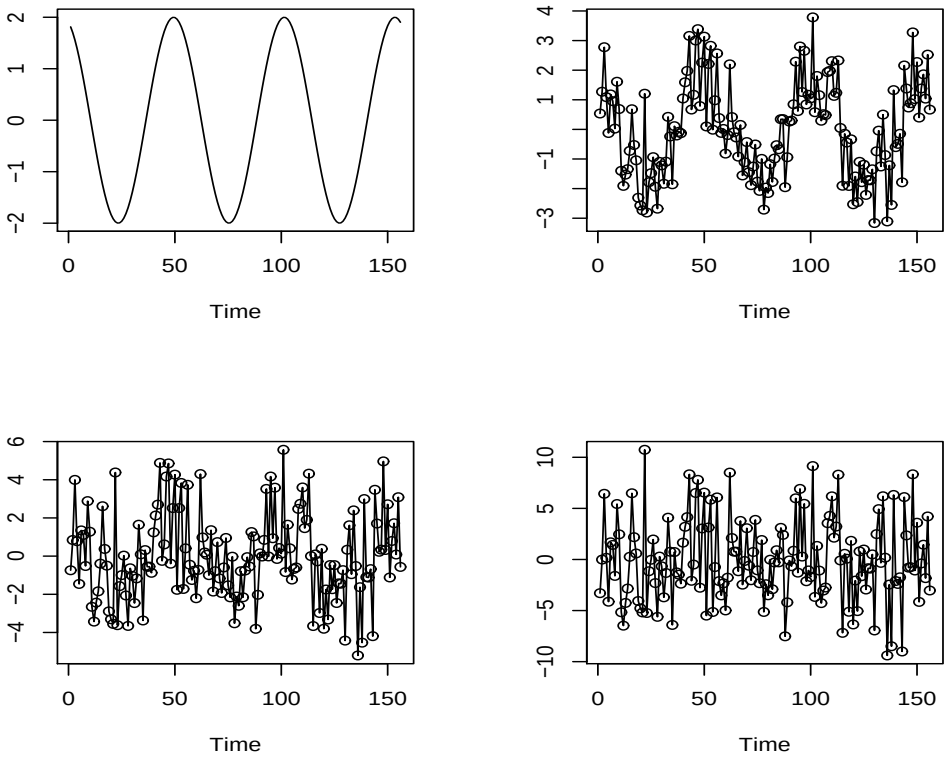


Figure 2.5: Sinusoidal model illustration. Top left: $E(Y_t) = 2\sin(2\pi t/52 + 0.6\pi)$. The other plots are simulated realizations of this process with $\sigma_e^2 = 1$ (top right), $\sigma_e^2 = 4$ (bottom left), and $\sigma_e^2 = 16$ (bottom right). In each simulated realization, $n = 156$.

2.5 Stationarity

NOTE: **Stationarity** is a very important concept in the analysis of time series data. Broadly speaking, a time series is said to be stationary if there is no systematic change in mean (no trend), if there is no systematic change in variance, and if strictly periodic variations have been removed. In other words, the properties of one section of the data are much like those of any other section.

IMPORTANCE: Much of the theory of time series is concerned with stationary time series. For this reason, time series analysis often requires one to transform a nonstationary

time series into a stationary one to use this theory. For example, it may be of interest to remove the trend and seasonal variation from a set of data and then try to model the variation in the residuals (the pieces "left over" after this removal) by means of a stationary stochastic process.

STATIONARITY: The stochastic process $\{Y_t : t = 0, 1, 2, ..., n\}$ is said to be **strictly stationary** if the joint distribution of

$$Y_{t_1}, Y_{t_2}, ..., Y_{t_n}$$

is the same as

$$Y_{t_1-k}, Y_{t_2-k}, ..., Y_{t_n-k}$$

for all time points $t_1, t_2, ..., t_n$ and for all time lags k. In other words, shifting the time origin by an amount k has no effect on the joint distributions, which must therefore depend only on the intervals between $t_1, t_2, ..., t_n$. **This is a very strong condition.**

IMPLICATION: Since the above condition holds for all sets of time points $t_1, t_2, ..., t_n$, it must hold when $n = 1$; i.e., there is only one time point.

- This implies Y_t and Y_{t-k} have the same **marginal distribution** for all t and k.
- Because these marginal distributions are the same,

$$\begin{aligned} E(Y_t) &= E(Y_{t-k}) \\ \text{var}(Y_t) &= \text{var}(Y_{t-k}), \end{aligned}$$

  for all t and k.
- Therefore, for a strictly stationary process, both $\mu_t = E(Y_t)$ and $\gamma_{t,t} = \text{var}(Y_t)$ are **constant** over time.

ADDITIONAL IMPLICATION: Since the above condition holds for all sets of time points $t_1, t_2, ..., t_n$, it must hold when $n = 2$; i.e., there are only two time points.

- This implies (Y_t, Y_s) and (Y_{t-k}, Y_{s-k}) have the same **joint distribution** for all t, s, and k.

- Because these joint distributions are the same,

$$\text{cov}(Y_t, Y_s) = \text{cov}(Y_{t-k}, Y_{s-k}),$$

for all t, s, and k.

- Therefore, for a strictly stationary process, for $k = s$,

$$\gamma_{t,s} = \text{cov}(Y_t, Y_s) = \text{cov}(Y_{t-s}, Y_0) = \text{cov}(Y_0, Y_{t-s}).$$

But, also, for $k = t$, we have

$$\text{cov}(Y_t, Y_s) = \text{cov}(Y_0, Y_{s-t}).$$

Putting the last two results together, we have

$$\gamma_{t,s} = \text{cov}(Y_t, Y_s) = \text{cov}(Y_0, Y_{|t-s|}) = \gamma_{0,|t-s|}.$$

This means that the covariance between Y_t and Y_s does not depend on the actual values of t and s; it only depends on the time difference $|t - s|$.

NEW NOTATION: For a (strictly) stationary process, the covariance $\gamma_{t,s}$ depends only on the time difference $|t - s|$. The quantity $|t - s|$ is the distance between time points Y_t and Y_s. In other words, the covariance between Y_t and any observation $k = |t - s|$ time points from it only depends on the **lag** k. Therefore, we write

$$\begin{aligned} \gamma_k &= \text{cov}(Y_t, Y_{t-k}) \\ \rho_k &= \text{corr}(Y_t, Y_{t-k}). \end{aligned}$$

We use this simpler notation only when we refer to a process which is stationary. Note that by taking $k = 0$, we have

$$\gamma_0 = \text{cov}(Y_t, Y_t) = \text{var}(Y_t).$$

Also,

$$\rho_k = \text{corr}(Y_t, Y_{t-k}) = \frac{\gamma_k}{\gamma_0}.$$

SUMMARY: For a process which is (strictly) stationary,

1. The mean function $\mu_t = E(Y_t)$ is **constant** throughout time; i.e., μ_t is free of t.
2. The covariance between any two observations **depends only the time lag** between them; i.e., $\gamma_{t,t-k}$ depends only on k (not on t).

REMARK: Strict stationarity is a condition that is much too restrictive for most applications. Moreover, it is difficult to assess the validity of this assumption in practice. Rather than impose conditions on all possible (marginal and joint) distributions of a process, we will use a milder form of stationarity that only deals with the first two moments.

DEFINITION: The stochastic process $\{Y_t : t = 0, 1, 2, ..., n\}$ is said to be **weakly stationary** (or **second-order stationary**) if

1. The mean function $\mu_t = E(Y_t)$ is **constant** throughout time; i.e., μ_t is free of t.
2. The covariance between any two observations **depends only the time lag** between them; i.e., $\gamma_{t,t-k}$ depends only on k (not on t).

Nothing is assumed about the collection of joint distributions of the process. Instead, we only are specifying the characteristics of the first two moments of the process.

REALIZATION: Clearly, strict stationarity implies weak stationarity. It is also clear that the converse to statement is not true, in general. However, if we append the additional assumption of multivariate normality (for the Y_t process), then the two definitions do coincide; that is,

$$\text{weak stationarity} + \text{multivariate normality} \implies \text{strict stationarity}.$$

CONVENTION: For the purpose of modeling time series data in this course, we will rarely (if ever) make the distinction between strict stationarity and weak stationarity. When we use the term "stationary process," this is understood to mean that the process is weakly stationary.

EXAMPLES: We now reexamine the time series models introduced in the last section.

- Suppose that $\{e_t\}$ is a **white noise process**. That is, $\{e_t\}$ consists of iid random variables with $E(e_t) = \mu_e$ and $\text{var}(e_t) = \sigma_e^2$, both constant (free of t). In addition, the autocovariance function $\gamma_k = \text{cov}(Y_t, Y_{t-k})$ is given by

$$\gamma_k = \begin{cases} \sigma_e^2, & k = 0 \\ 0, & k \neq 0, \end{cases}$$

  which is free of time t (i.e., γ_k depends only on k). Thus, a white noise process is stationary.

- Suppose that $\{Y_t\}$ is a **random walk process**. That is,

$$Y_t = Y_{t-1} + e_t,$$

  where $\{e_t\}$ is white noise with $E(e_t) = 0$ and $\text{var}(e_t) = \sigma_e^2$. We calculated $\mu_t = E(Y_t) = 0$, for all t, which is free of t. However,

$$\text{cov}(Y_t, Y_{t-k}) = \text{cov}(Y_{t-k}, Y_t) = (t-k)\sigma_e^2,$$

  which clearly depends on time t. Thus, a random walk process is not stationary.

- Suppose that $\{Y_t\}$ is a **moving average process** given by

$$Y_t = \frac{1}{3}(e_t + e_{t-1} + e_{t-2}),$$

  where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. We calculated $\mu_t = E(Y_t) = 0$ (which is free of t) and $\gamma_k = \text{cov}(Y_t, Y_{t-k})$ to be

$$\gamma_k = \begin{cases} \sigma_e^2/3, & k = 0 \\ 2\sigma_e^2/9, & k = 1 \\ \sigma_e^2/9, & k = 2 \\ 0, & k > 2. \end{cases}$$

  Because $\text{cov}(Y_t, Y_{t-k})$ is free of time t, this moving average process is stationary.

- Suppose that $\{Y_t\}$ is the **autoregressive process**

$$Y_t = 0.75Y_{t-1} + e_t,$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. We avoided the calculation of $\mu_t = E(Y_t)$ and $\text{cov}(Y_t, Y_{t-k})$ for this process, so we will not make a definite determination here. However, it turns out that if e_t is independent of Y_{t-1}, Y_{t-2}, ..., and if $\sigma_e^2 > 0$, then this autoregressive process is stationary (details coming later).

- Suppose that $\{Y_t\}$ is the **sinusoidal process** defined by

$$Y_t = a\sin(2\pi\omega t + \phi) + e_t,$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. Clearly $\mu_t = E(Y_t) = a\sin(2\pi\omega t + \phi)$ is not free of t, so this sinusoidal process is not stationary.

- Consider the **random cosine wave process**

$$Y_t = \cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right],$$

where Φ is a uniform random variable from 0 to 1; i.e., $\Phi \sim \mathcal{U}(0,1)$. The calculations on pp 18-19 (CC) show that this process is (perhaps unexpectedly) stationary.

IMPORTANT: In order to start thinking about viable stationary time series models for real data, we need to have a stationary process. However, as we have just seen, many data sets exhibit nonstationary behavior. A simple, but effective, technique to convert a nonstationary process into a stationary one is to examine data differences.

DEFINITION: Consider the process $\{Y_t : t = 0, 1, 2, ..., n\}$. The **(first) difference process** of $\{Y_t\}$ is defined by

$$\nabla Y_t = Y_t - Y_{t-1},$$

for $t = 1, 2, ..., n$. In many situations, a nonstationary process $\{Y_t\}$ can be "transformed" into a stationary process by taking (first) differences. For example, the random walk $Y_t = Y_{t-1} + e_t$, where $e_t \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, is not stationary. However, the first difference process $\nabla Y_t = Y_t - Y_{t-1} = e_t$ is zero mean white noise, which is stationary!

3 Modeling Deterministic Trends

Complementary reading: Chapter 3 (CC).

3.1 Introduction

DISCUSSION: In this course, we consider time series models for realizations of a stochastic process $\{Y_t : t = 0, 1, ..., n\}$. This will largely center around models for **stationary processes**. However, as we have seen, many time series data sets exhibit a **trend**; i.e., a long-term change in the mean level. We know that such series are not stationary because the mean changes with time.

- An obvious difficulty with the definition of a trend is deciding what is meant by the phrase "long-term." For example, climatic processes can display cyclical variation over a long period of time, say, 1000 years. However, if one has just 40-50 years of data, this long-term cyclical pattern might be missed and be interpreted as a trend which is linear.

- Trends can be "elusive," and an analyst may mistakenly conjecture that a trend exists when it really does not. For example, in Figure 2.2 (page 33), we have a realization of a random walk process

$$Y_t = Y_{t-1} + e_t,$$

  where $e_t \sim$ iid $\mathcal{N}(0, 1)$. There is no trend in the mean of this random walk process. Recall that $\mu_t = E(Y_t) = 0$, for all t. However, it would be easy to incorrectly assert that true downward and upward trends are present.

- On the other hand, it may be hard to detect trends if the data are very noisy. For example, the lower right plot in Figure 2.5 (page 38) is a noisy realization of a sinusoidal process considered in the last chapter. It is easy to miss the true cyclical structure from looking at the plot.

DETERMINISTIC TREND MODELS: In this chapter, we consider models of the form

$$Y_t = \mu_t + X_t,$$

where μ_t is a **deterministic function** that describes the trend and X_t is **random error**. Note that if, in addition, $E(X_t) = 0$ for all t (a common assumption), then

$$E(Y_t) = \mu_t$$

is the **mean function** for the process $\{Y_t\}$. In practice, different deterministic trend functions could be considered. One popular choice is

$$\mu_t = \beta_0 + \beta_1 t,$$

which says that the mean function increases (decreases) **linearly** with time. The function

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

is appropriate if there is a **quadratic** trend present. More generally, if the deterministic trend can be described by a kth order polynomial in time, we can consider

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k.$$

If the deterministic trend is cyclical, we could consider functions of the form

$$\mu_t = \beta_0 + \sum_{j=1}^{m} (\alpha_j \cos \omega_j t + \beta_j \sin \omega_j t),$$

where the α_j's and β_j's are regression parameters and the ω_j's are related to frequencies of the trigonometric functions $\cos \omega_j t$ and $\sin \omega_j t$. Fitting these and other deterministic trend models (and even combinations of them) can be accomplished using the **method of least squares**, as we will demonstrate later in this chapter.

LOOKING AHEAD: In this course, we want to deal with **stationary** time series models for data. Therefore, if there is a deterministic trend present in the process, we want to remove it. There are two general ways to do this.

1. **Estimate** the trend and then subtract the estimated trend from the data (perhaps after transforming the data). Specifically, estimate μ_t with $\widehat{\mu}_t$ and then model the **residuals**

$$\widehat{X}_t = Y_t - \widehat{\mu}_t$$

as a stationary process. We can use regression methods to estimate μ_t and then implement standard diagnostics on the residuals $\widehat{X}_t$ to check for violations of stationarity and other assumptions.

   - If the residuals are stationary, we can use a stationary time series model (Chapter 4) to describe their behavior.
   - Forecasting takes place by first forecasting the residual process $\{\widehat{X}_t\}$ and then inverting the transformations described above to arrive back at forecasts for the original series $\{Y_t\}$. We will pursue forecasting techniques in Chapter 9.

   IMPORTANT: If we assert that a trend exists and we fit a deterministic model that incorporates it, we are implicitly assuming that the trend lasts "forever." In some applications, this might be reasonable, but probably not in most.

2. Another approach, developed extensively by Box and Jenkins, is to apply **differencing** repeatedly to the series $\{Y_t\}$ until the differenced observations resemble a realization of a stationary time series. We can then use the theory of stationary processes for the modeling, analysis, and prediction of the stationary series and then transform this analysis back in terms of the original series $\{Y_t\}$. This approach is studied in Chapter 5.

3.2 Estimation of a constant mean

A CONSTANT "TREND": We first consider the most elementary type of trend, namely, a **constant** trend. Specifically, we consider the model

$$Y_t = \mu + X_t,$$

where μ is constant (free of t) and where $E(X_t) = 0$. Note that, under this zero mean error assumption, we have

$$E(Y_t) = \mu.$$

That is, the process $\{Y_t\}$ has an overall population mean function $\mu_t = \mu$, for all t. The most common estimate of μ is

$$\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t,$$

the **sample mean**. It is easy to check that $\overline{Y}$ is an **unbiased estimator** of μ; i.e., $E(\overline{Y}) = \mu$. This is true because

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{t=1}^{n} Y_t\right) = \frac{1}{n}\sum_{t=1}^{n} E(Y_t) = \frac{1}{n}\sum_{t=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

Therefore, under the minimal assumption that $E(X_t) = 0$, we see that $\overline{Y}$ is an unbiased estimator of μ. To assess the precision of $\overline{Y}$ as an estimator of μ, we examine $\text{var}(\overline{Y})$.

RESULT: If $\{Y_t\}$ is a **stationary** process with autocorrelation function ρ_k, then

$$\text{var}(\overline{Y}) = \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right],$$

where $\text{var}(Y_t) = \gamma_0$.

RECALL: If $\{Y_t\}$ is an iid process, that is, $Y_1, Y_2, ..., Y_n$ is an iid (random) sample, then

$$\text{var}(\overline{Y}) = \frac{\gamma_0}{n}.$$

Therefore, $\text{var}(\overline{Y})$, in general, can be larger than or smaller than γ_0/n depending on the values of ρ_k through

$$\frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right] - \frac{\gamma_0}{n} = \frac{2\gamma_0}{n}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k.$$

- If this quantity is smaller than zero, then $\overline{Y}$ is a better estimator of μ than $\overline{Y}$ is in an iid sampling context; that is, $\text{var}(\overline{Y}) < \gamma_0/n$.
- If this quantity is larger than zero, then $\overline{Y}$ is a worse estimator of μ than $\overline{Y}$ is in an iid sampling context; that is, $\text{var}(\overline{Y}) > \gamma_0/n$.

Example 3.1. Suppose that $\{Y_t\}$ is a **moving average process** given by

$$Y_t = \frac{1}{3}(e_t + e_{t-1} + e_{t-2}),$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In the last chapter, we calculated

$$\gamma_k = \begin{cases} \sigma_e^2/3, & k = 0 \\ 2\sigma_e^2/9, & k = 1 \\ \sigma_e^2/9, & k = 2 \\ 0, & k > 2. \end{cases}$$

The lag 1 autocorrelation for this process is

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{2\sigma_e^2/9}{\sigma_e^2/3} = 2/3.$$

The lag 2 autocorrelation for this process is

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = \frac{\sigma_e^2/9}{\sigma_e^2/3} = 1/3.$$

Also, $\rho_k = 0$ for all $k > 2$. Therefore,

$$\begin{aligned} \text{var}(\overline{Y}) &= \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right] \\ &= \frac{\gamma_0}{n} + \frac{4(n-1)\gamma_0 + 2(n-2)\gamma_0}{3n^2} > \frac{\gamma_0}{n}. \end{aligned}$$

Therefore, we lose efficiency in estimating μ with $\overline{Y}$ when compared to using $\overline{Y}$ in an iid sampling context. The positive autocorrelations make estimation of μ less precise.

Example 3.2. Suppose that $\{Y_t\}$ is a stationary process with autocorrelation function $\rho_k = \phi^k$, where $-1 < \phi < 1$. For this process, the autocorrelation decays exponentially as the lag k increases. As we will see in Chapter 4, the **autoregressive of order 1**, AR(1), **process** possesses this autocorrelation function. To examine the effect of estimating μ with $\overline{Y}$ in this situation, we use an approximation for $\text{var}(\overline{Y})$ for large n, specifically,

$$\text{var}(\overline{Y}) = \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right] \approx \frac{\gamma_0}{n}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right),$$

where we have taken $(1 - k/n) \approx 1$ for n large. Therefore, with $\rho_k = \phi^k$, we have

$$\begin{aligned}
\text{var}(\overline{Y}) &\approx \frac{\gamma_0}{n}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right) \\
&= \frac{\gamma_0}{n}\left[1 + 2\left(\sum_{k=0}^{\infty}\phi^k - 1\right)\right] \\
&= \frac{\gamma_0}{n}\left[1 + 2\left(\frac{1}{1-\phi} - 1\right)\right] = \left(\frac{1+\phi}{1-\phi}\right)\frac{\gamma_0}{n}.
\end{aligned}$$

For example, if $\phi = -0.6$, then

$$\text{var}(\overline{Y}) \approx 0.25\left(\frac{\gamma_0}{n}\right).$$

Using $\overline{Y}$ produces a more precise estimate of μ than in an iid (random) sampling context. The negative autocorrelations $\rho_1 = -0.6$, $\rho_3 = (-0.6)^3$, etc., "outweigh" the positive ones $\rho_2 = (-0.6)^2$, $\rho_4 = (-0.6)^4$, etc., making $\text{var}(\overline{Y})$ smaller than γ_0/n.

Example 3.3. In Examples 3.1 and 3.2, we considered stationary processes in examining the precision of $\overline{Y}$ as an estimator for μ. In this example, we have the same goal, but we consider the **random walk process** $Y_t = Y_{t-1} + e_t$, where $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. As we already know, this process is not stationary, so we can not use the $\text{var}(\overline{Y})$ formula presented earlier. However, recall that this process can be written out as

$$Y_1 = e_1, \qquad Y_2 = e_1 + e_2, \qquad ..., \qquad Y_n = e_1 + e_2 + \cdots + e_n,$$

so that

$$\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t = \frac{1}{n}\left[ne_1 + (n-1)e_2 + (n-2)e_3 + \cdots + 2e_{n-1} + 1e_n\right].$$

Therefore, we can derive an expression for $\text{var}(\overline{Y})$ directly:

$$\begin{aligned}
\text{var}(\overline{Y}) &= \frac{1}{n^2}\left[n^2\text{var}(e_1) + (n-1)^2\text{var}(e_2) + \cdots + 2^2\text{var}(e_{n-1}) + 1^2\text{var}(e_n)\right] \\
&= \frac{\sigma_e^2}{n^2}[1^2 + 2^2 + \cdots + (n-1)^2 + n^2] \\
&= \frac{\sigma_e^2}{n^2}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{\sigma_e^2}{n}\left[\frac{(n+1)(2n+1)}{6}\right].
\end{aligned}$$

- This result is surprising! Note that as n increases, so does $\text{var}(\overline{Y})$. That is, averaging a larger sample produces a worse (i.e., more variable) estimate of μ than averaging a smaller one!!

- This is quite different than the results obtained for stationary processes. The nonstationarity in the data causes very bad things to happen, even in the relatively simple task of estimating an overall process mean.

RESULT: Suppose that $Y_t = \mu + X_t$, where μ is constant, $X_t \sim \mathcal{N}(0, \gamma_0)$, and $\{X_t\}$ is a stationary process. Under these assumptions, $Y_t \sim \mathcal{N}(\mu, \gamma_0)$ and $\{Y_t\}$ is stationary. Therefore,

$$\overline{Y} \sim \mathcal{N}\left\{\mu, \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right]\right\}$$

so that

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right]}} \sim \mathcal{N}(0, 1).$$

Since the sampling distribution of Z does not depend on any unknown parameters, we say that Z is a **pivotal quantity** (or, more simply, a **pivot**). If γ_0 and the ρ_k's are known, then a $100(1-\alpha)$ percent **confidence interval** for μ is

$$\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right]},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the standard normal distribution.

REMARK: Note that if $\rho_k = 0$, for all k, then $\overline{Y} \sim \mathcal{N}(\mu, \gamma_0/n)$, and the confidence interval formula just presented reduces to

$$\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\gamma_0}{n}},$$

which we recognize as the confidence interval for μ when random sampling is used. The impact of the autocorrelations ρ_k will be the same on the confidence interval. That is, more negative autocorrelations ρ_k will make the standard error

$$\text{se}(\overline{Y}) = \sqrt{\frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right]}$$

smaller, which will make the confidence interval more precise (i.e., shorter). On the other hand, positive autocorrelations will make this quantity larger, thereby lengthening the interval, making it less informative.

REMARK: Of course, in real life, rarely will anyone tell us the values of γ_0 and the ρ_k's. These are model (population) parameters. However, if the sample size n is large and "good" (large-sample) estimates of these quantities can be calculated, we would expect this interval to be approximately valid when the estimates are substituted in for the true values. We will talk about estimation of γ_0 and the autocorrelations later.

3.3 Regression methods

3.3.1 Straight line regression

STRAIGHT LINE MODEL: We now consider the deterministic time trend model

$$\begin{aligned} Y_t &= \mu_t + X_t \\ &= \beta_0 + \beta_1 t + X_t, \end{aligned}$$

where $\mu_t = \beta_0 + \beta_1 t$ and where $E(X_t) = 0$. We are considering a **simple linear regression model** for the process $\{Y_t\}$, where time t is the predictor. By "fitting this model," we mean that we would like to estimate the **regression parameters** β_0 and β_1 (the intercept and slope, respectively) using the observed data $Y_1, Y_2, ..., Y_n$. The X_t's are random errors and are not observed.

LEAST SQUARES ESTIMATION: To estimate β_0 and β_1, we will use the **method of least squares**. Specifically, we find the values of β_0 and β_1 that minimize the objective function

$$\begin{aligned} Q(\beta_0, \beta_1) &= \sum_{t=1}^{n} (Y_t - \mu_t)^2 \\ &= \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t)]^2 = \sum_{t=1}^{n} (Y_t - \beta_0 - \beta_1 t)^2. \end{aligned}$$

This can be done using a multivariable calculus argument. Specifically, the partial derivatives of $Q(\beta_0, \beta_1)$ are given by

$$\begin{aligned}\frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0} &= -2\sum_{t=1}^{n}(Y_t - \beta_0 - \beta_1 t) \\ \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1} &= -2\sum_{t=1}^{n} t(Y_t - \beta_0 - \beta_1 t).\end{aligned}$$

Setting these derivatives equal to zero and jointly solving for β_0 and β_1, we get

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{t}. \\ \widehat{\beta}_1 &= \frac{\sum_{t=1}^{n}(t-\overline{t})Y_t}{\sum_{t=1}^{n}(t-\overline{t})^2}.\end{aligned}$$

These are the **least squares estimators** of β_0 and β_1.

PROPERTIES: The following results can be established algebraically. *Note carefully which statistical assumptions are needed for each result.*

- Under just the mild assumption of $E(X_t) = 0$, for all t, the least squares estimators are **unbiased**. That is, $E(\widehat{\beta}_0) = \beta_0$ and $E(\widehat{\beta}_1) = \beta_1$.

- Under the assumptions that $E(X_t) = 0$, $\{X_t\}$ independent, and $\text{var}(X_t) = \gamma_0$ (a constant, free of t), then

$$\begin{aligned}\text{var}(\widehat{\beta}_0) &= \gamma_0\left[\frac{1}{n} + \frac{\overline{t}^2}{\sum_{t=1}^{n}(t-\overline{t})^2}\right] \\ \text{var}(\widehat{\beta}_1) &= \frac{\gamma_0}{\sum_{t=1}^{n}(t-\overline{t})^2}.\end{aligned}$$

  Note that a zero mean white noise process $\{X_t\}$ satisfies these assumptions.

- In addition to the assumptions $E(X_t) = 0$, $\{X_t\}$ independent, and $\text{var}(X_t) = \gamma_0$, if we also assume that the X_t's are **normally distributed**, then

$$\widehat{\beta}_0 \sim \mathcal{N}\left\{\beta_0, \gamma_0\left[\frac{1}{n} + \frac{\overline{t}^2}{\sum_{t=1}^{n}(t-\overline{t})^2}\right]\right\}$$

  and

$$\widehat{\beta}_1 \sim \mathcal{N}\left[\beta_1, \frac{\gamma_0}{\sum_{t=1}^{n}(t-\overline{t})^2}\right].$$

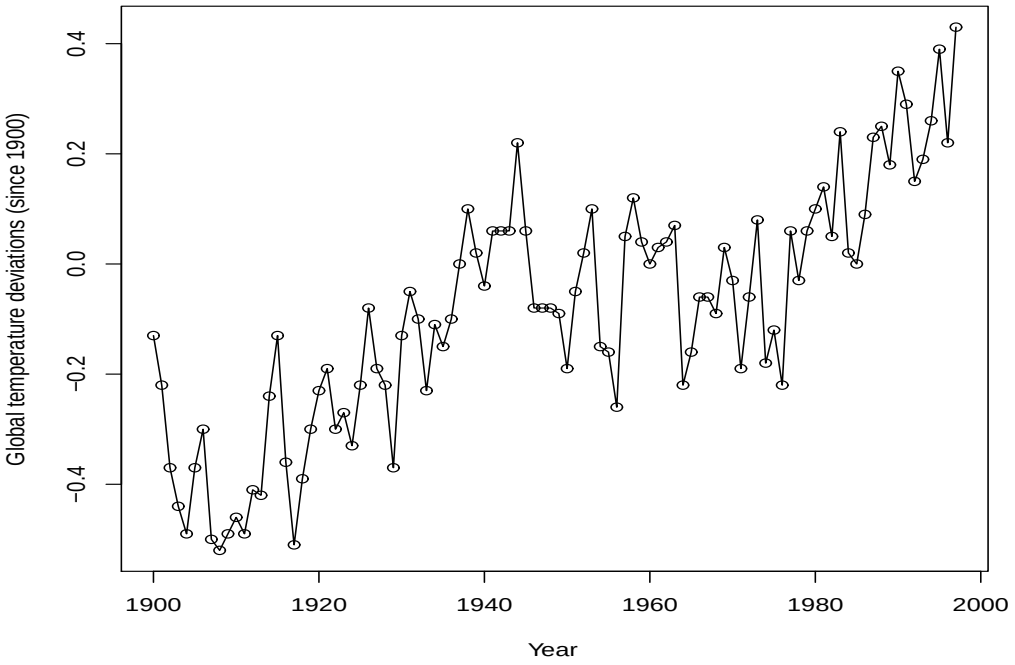


Figure 3.1: Global temperature data. The data are a combination of land-air average temperature anomalies, measured in degrees Centigrade. Time period: 1900-1997.

IMPORTANT: You should recall that these four assumptions on the errors X_t, that is, **zero mean**, **independence**, **homoscedasticity**, and **normality**, are the usual assumptions on the errors in a standard regression setting. However, with most time series data sets, at least one of these assumptions will be violated. The implication, then, is that standard errors of the estimators, confidence intervals, t tests, probability values, etc., quantities that are often provided in computing packages (e.g., R, etc.), will not be meaningful. Proper usage of this output requires the four assumptions mentioned above to hold. The only instance in which these are exactly true is if $\{X_t\}$ is a zero-mean normal white noise process (an assumption you likely made in your previous methods courses where regression was discussed).

Example 3.4. Consider the global temperature data from Example 1.1 (notes), but let's restrict attention to the time period 1900-1997. These data are depicted in Figure 3.1.

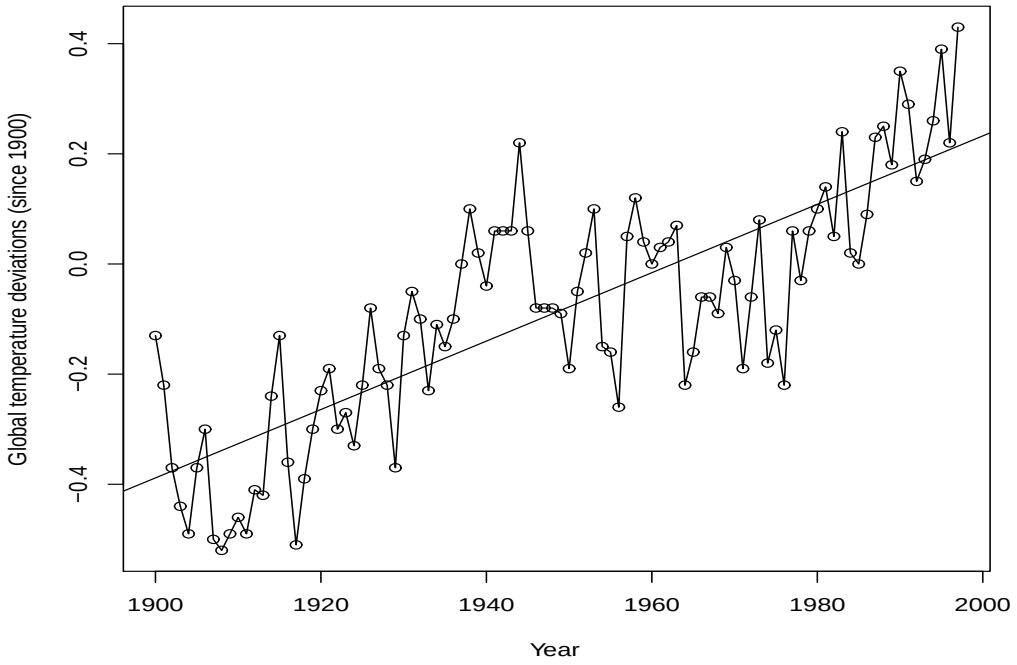


Figure 3.2: Global temperature data (1900-1997) with a straight line trend fit.

Over this time period, there is an apparent upward trend in the series. Suppose that we estimate this trend by fitting the straight line regression model

$$Y_t = \beta_0 + \beta_1 t + X_t,$$

for $t = 1900, 1901, ..., 1997$, where $E(X_t) = 0$. Here is the output from fitting this model in R.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.219e+01  9.032e-01  -13.49   <2e-16 ***
time(globaltemps.1900)  6.209e-03  4.635e-04   13.40   <2e-16 ***

Residual standard error: 0.1298 on 96 degrees of freedom
Multiple R-squared: 0.6515,     Adjusted R-squared: 0.6479
F-statistic: 179.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

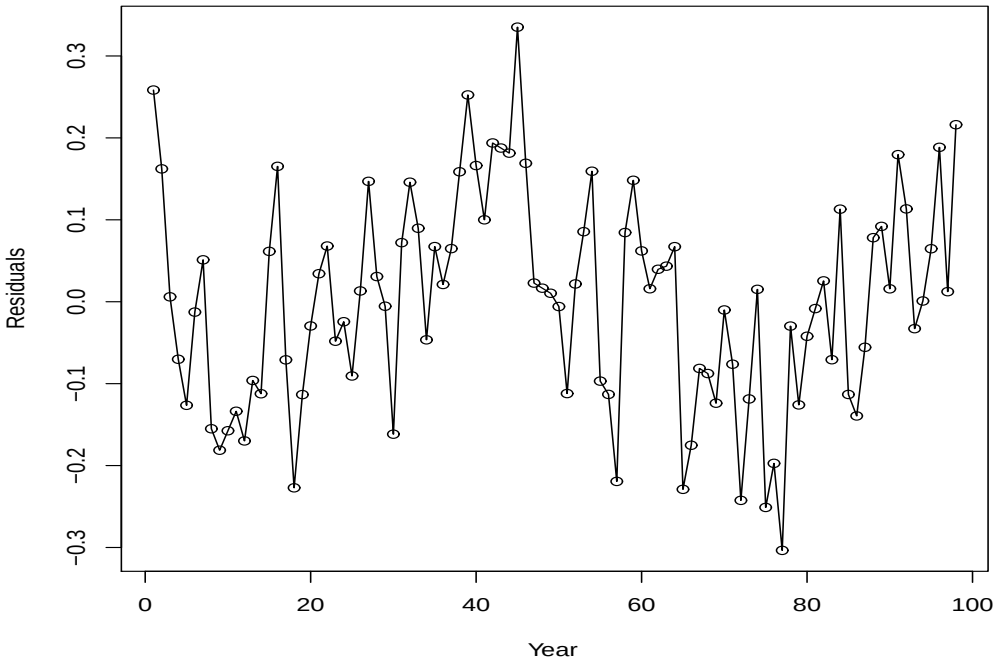


Figure 3.3: Global temperature data (1990-1997). Residuals from the straight line trend model fit.

ANALYSIS: We interpret the regression coefficient output only. As we have learned, standard errors, t tests, and probability values may not be meaningful! The least squares estimates are $\widehat{\beta}_0 = -12.19$ and $\widehat{\beta}_1 = 0.0062$ so that the fitted regression model is

$$\widehat{Y}_t = -12.19 + 0.0062t.$$

This is the equation of the line superimposed over the series in Figure 3.2.

RESIDUALS: The **residuals** from the least squares fit are given by

$$\widehat{X}_t = Y_t - \widehat{Y}_t,$$

that is, the observed data Y_t minus the **fitted values** given by the equation in $\widehat{Y}_t$. In this example (with the straight line model fit), the residuals are given by

$$\begin{aligned} \widehat{X}_t &= Y_t - \widehat{Y}_t \\ &= Y_t + 12.19 - 0.0062t, \end{aligned}$$

for $t = 1900, 1901, ..., 1997$. Remember that one of the main reasons for fitting the straight line model was to capture the linear trend. Now that we have done this, the **residual process** defined by

$$\widehat{X}_t = Y_t + 12.19 - 0.0062t$$

contains information in the data that is not accounted for in the straight line trend model. For this reason, it is called the **detrended series**. This series is plotted in Figure 3.3. Essentially, this is a time series plot of the residuals from the straight line fit versus time, the predictor variable in the model. This detrended series does appear to be somewhat stationary, at least much more so than the original series $\{Y_t\}$. However, just from looking at the plot, it is a safe bet that the residuals are not white noise.

DIFFERENCING: Instead of fitting the deterministic model to the global temperature series to remove the linear trend, suppose that we had examined the **first difference process** $\{\nabla Y_t\}$, where

$$\nabla Y_t = Y_t - Y_{t-1}.$$

We have learned that taking differences can be an effective means to remove non-stationary patterns. Doing so here, as evidenced in Figure 3.4, produces a new process that does appear to be somewhat stationary.

DISCUSSION: We have just seen, by means of an example, that both **detrending** (using regression to fit the trend) and **differencing** can be helpful in transforming a nonstationary process into one which is (or at least appears) stationary.

- One advantage of differencing over detrending to remove trend is that no parameters are estimated in taking differences.
- One disadvantage of differencing is that it does not provide an "estimate" of the error process X_t.
- If an estimate of the error process is crucial, detrending may be more appropriate. If the goal is only to coerce the data to stationarity, differencing may be preferred.

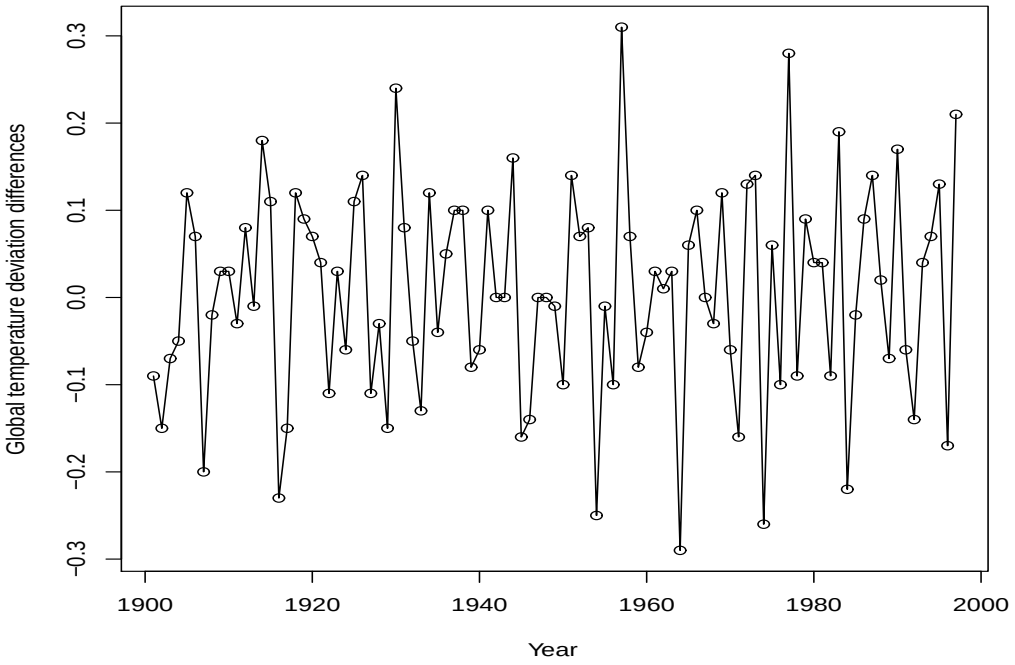


Figure 3.4: Global temperature first data differences (1900-1997).

3.3.2 Polynomial regression

POLYNOMIAL REGRESSION: We now consider the deterministic time trend model

$$\begin{aligned} Y_t &= \mu_t + X_t \\ &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k + X_t, \end{aligned}$$

where $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k$ and where $E(X_t) = 0$. The mean function μ_t is a polynomial function with degree $k \geq 1$.

- If $k = 1$, $\mu_t = \beta_0 + \beta_1 t$ is a **linear** trend function.
- If $k = 2$, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$ is a **quadratic** trend function.
- If $k = 3$, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$ is a **cubic** trend function, and so on.

LEAST SQUARES ESTIMATION: The least squares estimates of β_0, β_1, β_2, ..., β_k are obtained in the same way as in the $k = 1$ case; namely, the estimates are obtained by minimizing the objective function

$$\begin{aligned} Q(\beta_0, \beta_1, \beta_2, ..., \beta_k) &= \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k)]^2 \\ &= \sum_{t=1}^{n} (Y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \cdots - \beta_k t^k)^2 \end{aligned}$$

with respect to β_0, β_1, β_2, ..., β_k. Here, there are $k + 1$ partial derivatives and $k + 1$ equations to solve (in simple linear regression, $k = 1$, so there were 2 equations to solve).

- Unfortunately (without the use of more advanced notation), there are no convenient, closed-form expressions for the least squares estimators when $k > 1$. This turns out not to be a major distraction, because we use computing to fit the model anyway.
- Under the mild assumption that the errors have zero mean; i.e., that $E(X_t) = 0$, it follows that the least squares estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, ..., $\widehat{\beta}_k$ are **unbiased estimators** of their population analogues; i.e., $E(\widehat{\beta}_i) = \beta_i$, for $i = 0, 1, 2, ..., k$.
- As in the simple linear regression case ($k = 1$), additional assumptions on the errors X_t are needed to derive the sampling distribution of the least squares estimators, namely, independence, constant variance, and normality.
- Regression output (e.g., in R, etc.) is correct only under these additional assumptions. Thee analyst must keep this in mind.

Example 3.5. Data file: `gold` (TSA). Of all the precious metals, gold is the most popular as an investment. Like most commodities, the price of gold is driven by supply and demand as well as speculation. Figure 3.5 contains a time series of $n = 254$ daily observations on the price of gold (per troy ounce) in US dollars during the year 2005. There is a clear nonlinear trend in the data, so a straight-line model would not be appropriate.

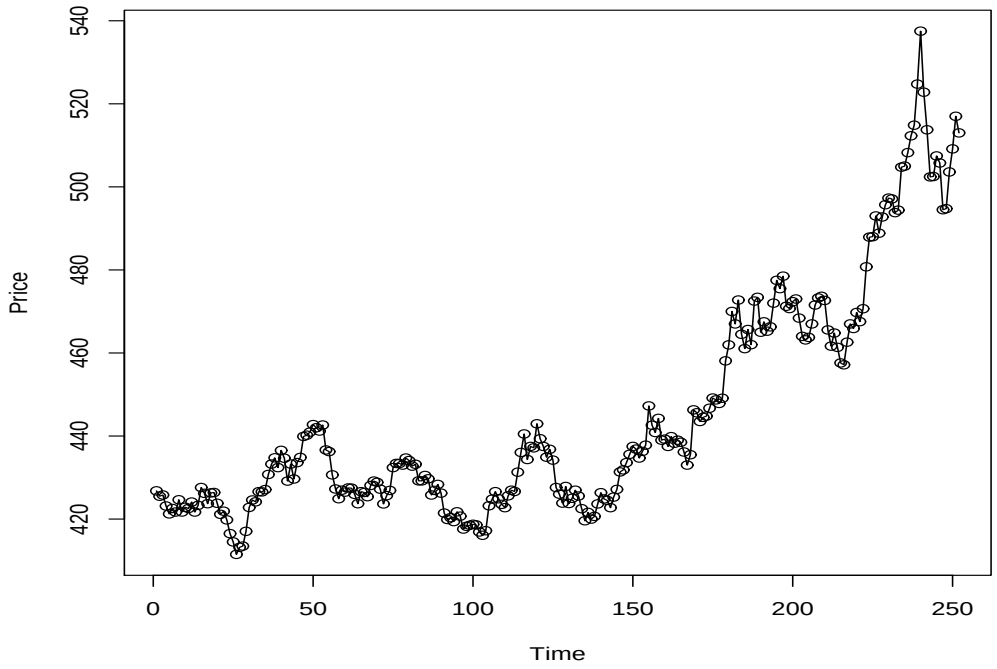


Figure 3.5: Gold price data. Daily price in US dollars per troy ounce: 1/4/05-12/30/05.

In this example, we use R to detrend the data by fitting the **quadratic regression model**

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + X_t,$$

for $t = 1, 2, ..., 254$, where $E(X_t) = 0$. Here is the output from fitting this model in R.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.346e+02  1.771e+00  245.38   <2e-16 ***
t           -3.618e-01  3.233e-02  -11.19   <2e-16 ***
t.sq         2.637e-03  1.237e-04   21.31   <2e-16 ***

Residual standard error: 9.298 on 249 degrees of freedom
Multiple R-squared: 0.8838,     Adjusted R-squared: 0.8828
F-statistic: 946.6 on 2 and 249 DF,  p-value: < 2.2e-16
```

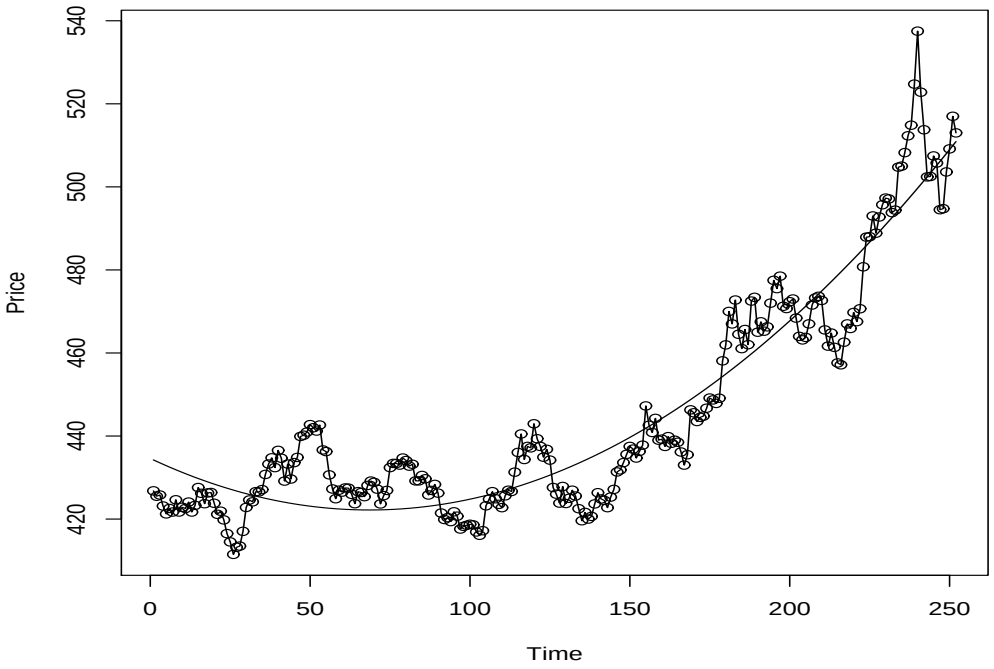


Figure 3.6: Gold price data with a quadratic trend fit.

ANALYSIS: Again, we focus only on the values of the least squares estimates. The fitted regression equation is

$$\widehat{Y}_t = 434.6 - 0.362t + 0.00264t^2,$$

for $t = 1, 2, ..., 254$. This fitted model is superimposed over the time series in Figure 3.6.

RESIDUALS: The residual process is

$$\begin{aligned} \widehat{X}_t &= Y_t - \widehat{Y}_t \\ &= Y_t - 434.6 + 0.362t - 0.00264t^2, \end{aligned}$$

for $t = 1, 2, ..., 254$, and is depicted in Figure 3.7. This detrended series appears to be somewhat stationary, at least, much more so than the original time series. However, it should be obvious that the detrended (residual) process is not white noise. There is still an enormous amount of momentum left in the residuals. Of course, we know that this renders most of the R output on the previous page meaningless.

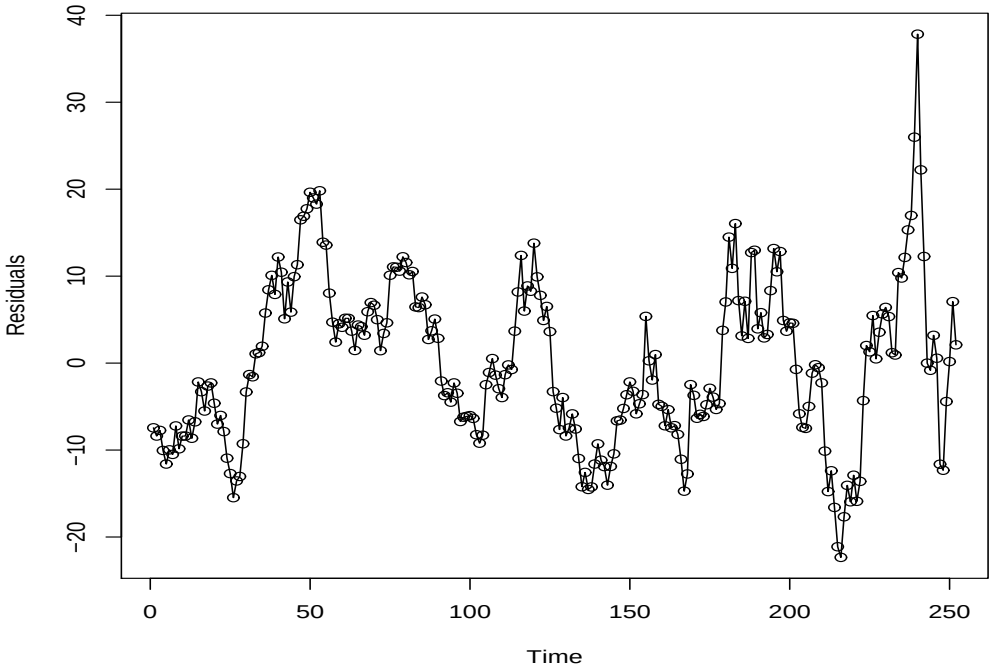


Figure 3.7: Gold price data. Residuals from the quadratic trend fit.

3.3.3 Seasonal means model

SEASONAL MEANS MODEL: Consider the deterministic trend model

$$Y_t = \mu_t + X_t,$$

where $E(X_t) = 0$ and where the mean function

$$\mu_t = \begin{cases} \beta_1, & t = 1, 13, 25, ... \\ \beta_2, & t = 2, 14, 26, ... \\ \vdots & \\ \beta_{12}, & t = 12, 24, 36, ... \end{cases}$$

The regression parameters β_1, β_2, ..., β_{12} are fixed constants. This is called a **seasonal means model**. This model does not take the shape of the seasonal trend into account; instead, it merely says that observations 12 months apart have the same mean, and

this mean does not change through time. Other seasonal means models with a different number of parameters could be specified. For instance, for **quarterly data**, we could use a mean function with 4 regression parameters β_1, β_2, β_3, and β_4.

FITTING THE MODEL: We can still use least squares to fit the seasonal means model. The least squares estimates of the regression parameters are simple to compute, but difficult to write mathematically. In particular,

$$\widehat{\beta}_1 = \frac{1}{n_1} \sum_{t \in \mathcal{A}_1} Y_t,$$

where the set $\mathcal{A}_1 = \{t : t = 1 + 12j,\ j = 0, 1, 2, ..., \}$. In essence, to compute $\widehat{\beta}_1$, we sum the values $Y_1, Y_{13}, Y_{25}, ...$, and then divide by n_1, the number of observations in month 1 (e.g., January). Similarly,

$$\widehat{\beta}_2 = \frac{1}{n_2} \sum_{t \in \mathcal{A}_2} Y_t,$$

where the set $\mathcal{A}_2 = \{t : t = 2 + 12j,\ j = 0, 1, 2, ..., \}$. Again, we sum the values $Y_2, Y_{14}, Y_{26}, ...$, and then divide by n_2, the number of observations in month 2 (e.g., February). In general,

$$\widehat{\beta}_i = \frac{1}{n_i} \sum_{t \in \mathcal{A}_i} Y_t,$$

where the set $\mathcal{A}_i = \{t : t = i + 12j,\ j = 0, 1, 2, ..., \}$, for $i = 1, 2, ..., 12$, where n_i is the number of observations in month i.

Example 3.6. Data file: `beersales` (TSA). The data in Figure 3.8 are monthly beer sales (in millions of barrels) in the United States from 1/80 through 12/90. This time series has a relatively constant mean overall (i.e., there are no apparent linear trends and the repeating patterns are relatively constant over time), so a seasonal means model may be appropriate. Fitting the model can be done in R; here are the results.

```
Coefficients:
          Estimate  Std. Error t value   Pr(>|t|)
January    13.1608     0.1647   79.90   <2e-16 ***
February   13.0176     0.1647   79.03   <2e-16 ***
March      15.1058     0.1647   91.71   <2e-16 ***
```

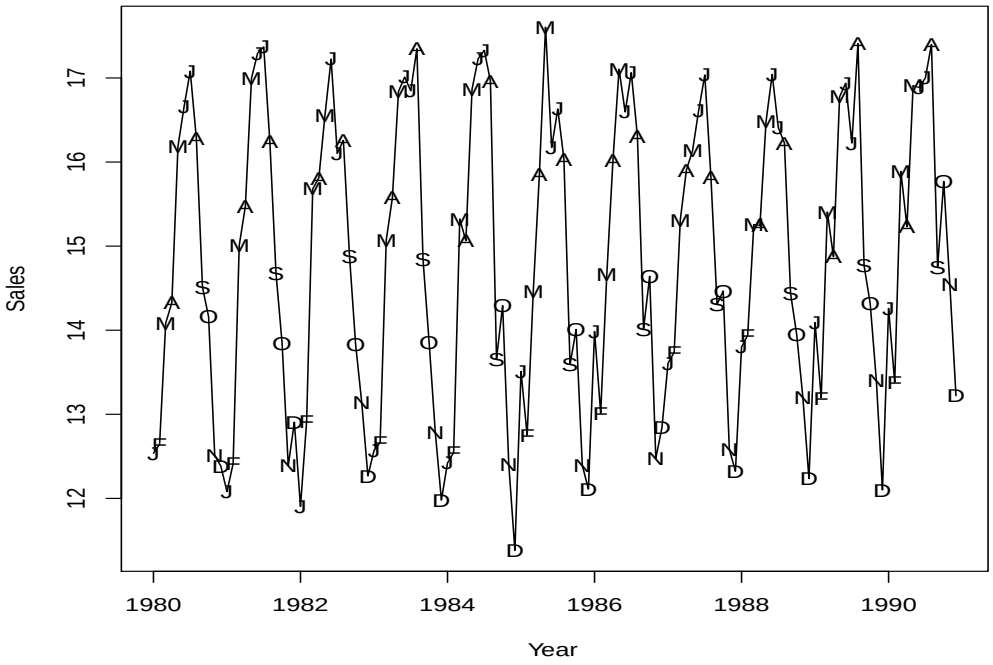


Figure 3.8: Monthly US beer sales from 1980-1990. The data are measured in millions of barrels.

```
April       15.3981     0.1647   93.48   <2e-16 ***
May         16.7695     0.1647  101.81   <2e-16 ***
June        16.8792     0.1647  102.47   <2e-16 ***
July        16.8270     0.1647  102.16   <2e-16 ***
August      16.5716     0.1647  100.61   <2e-16 ***
September   14.4045     0.1647   87.45   <2e-16 ***
October     14.2848     0.1647   86.72   <2e-16 ***
November    12.8943     0.1647   78.28   <2e-16 ***
December    12.3404     0.1647   74.92   <2e-16 ***
```

DISCUSSION: The only quantities that have relevance are the least squares estimates. The estimate $\widehat{\beta}_i$ is simply the sample mean of the observations for month i; thus, $\widehat{\beta}_i$ is an unbiased estimate of the ith (population) mean monthly sales β_i. The test statistics and p-values are used to test $H_0 : \beta_i = 0$, a largely nonsensical hypothesis in this example.

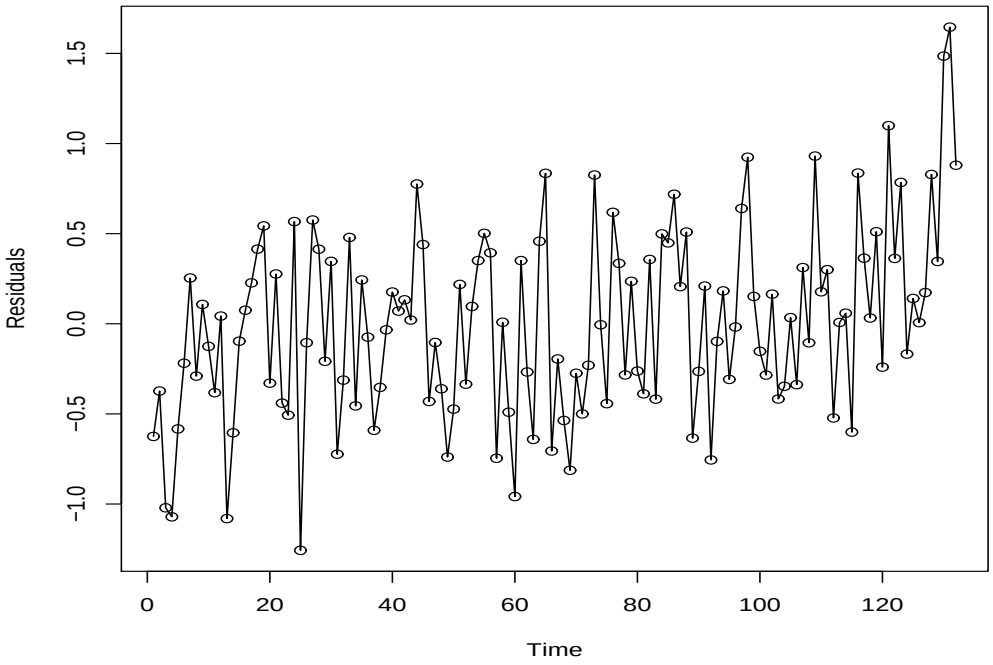


Figure 3.9: Beer sales data. Residuals from the seasonal means model fit.

RESIDUALS: A plot of the residuals from the seasonal means model fit, that is,

$$\begin{aligned} \widehat{X}_t &= Y_t - \widehat{Y}_t \\ &= Y_t - \sum_{i=1}^{12} \widehat{\beta}_i I_{\mathcal{A}_i}(t) \end{aligned}$$

is in Figure 3.9. The expression $\sum_{i=1}^{12} \widehat{\beta}_i I_{\mathcal{A}_i}(t)$, where $I(\cdot)$ is the **indicator function**, is simply the sample mean for the set of observations at time t. This residual process looks somewhat stationary, although I can detect a slightly increasing trend.

3.3.4 Cosine trend model

REMARK: The seasonal means model is somewhat simplistic in that it does not take the shape of the seasonal trend into account. We now consider a more elaborate regression equation that can be used to model data with seasonal trends.

COSINE TREND MODEL: Consider the deterministic time trend model

$$\begin{aligned} Y_t &= \mu_t + X_t \\ &= \beta\cos(2\pi ft + \Phi) + X_t, \end{aligned}$$

where $\mu_t = \beta\cos(2\pi ft + \Phi)$ and where $E(X_t) = 0$. The trigonometric mean function μ_t consists of different parts:

- β is the **amplitude**. The function μ_t oscillates between $-\beta$ and β.
- f is the **frequency** $\Longrightarrow$ $1/f$ is the **period** (the time it takes to complete one full cycle of the function). For monthly data, the period is 12 months; i.e., the frequency is $f = 1/12$.
- Φ controls the **phase shift**. This represents a horizontal shift in the mean function.

MODEL FITTING: Fitting this model is difficult unless we transform the mean function into a simpler expression. We use the trigonometric identity

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$$

to write

$$\begin{aligned} \beta\cos(2\pi ft + \Phi) &= \beta\cos(2\pi ft)\cos(\Phi) - \beta\sin(2\pi ft)\sin(\Phi) \\ &= \beta_1\cos(2\pi ft) + \beta_2\sin(2\pi ft), \end{aligned}$$

where $\beta_1 = \beta\cos\Phi$ and $\beta_2 = -\beta\sin\Phi$, so that the phase shift parameter

$$\Phi = \tan^{-1}\left(-\frac{\beta_2}{\beta_1}\right)$$

and the amplitude $\beta = \sqrt{\beta_1^2 + \beta_2^2}$. The rewritten expression,

$$\mu_t = \beta_1\cos(2\pi ft) + \beta_2\sin(2\pi ft),$$

is a linear function of β_1 and β_2, where $\cos(2\pi ft)$ and $\sin(2\pi ft)$ play the roles of predictor variables. Adding an intercept term for flexibility, say β_0, we get

$$Y_t = \beta_0 + \beta_1\cos(2\pi ft) + \beta_2\sin(2\pi ft) + X_t.$$

REMARK: When we fit this model, we must be aware of the values used for the time t, as it has a direct impact on how we specify the frequency f. For example,

- if we have monthly data and use the generic time specification $t = 1, 2, ..., 12, 13, ...$, then we specify $f = 1/12$.
- if we have monthly data, but we use the years themselves as predictors; i.e., $t = 1990, 1991, 1992$, etc., we use $f = 1$, because 12 observations arrive each year.

Example 3.6 (continued). We now use R to fit the cosine trend model to the beer sales data. Because the predictor variable t is measured in years 1980, 1981, ..., 1990 (with 12 observations each year), we use $f = 1$. Here is the output:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     14.80446    0.05624  263.25   <2e-16 ***
har.cos(2*pi*t) -2.04362    0.07953  -25.70   <2e-16 ***
har.sin(2*pi*t)  0.92820    0.07953   11.67   <2e-16 ***

Residual standard error: 0.6461 on 129 degrees of freedom
Multiple R-squared: 0.8606,     Adjusted R-squared: 0.8584
F-statistic: 398.2 on 2 and 129 DF,  p-value: < 2.2e-16
```

ANALYSIS: The fitted model

$$\widehat{Y}_t = 14.8 - 2.04\cos(2\pi t) + 0.93\sin(2\pi t),$$

is superimposed over the data in Figure 3.10. The least squares estimates $\widehat{\beta}_0 = 14.8$, $\widehat{\beta}_1 = -2.04$, and $\widehat{\beta}_2 = 0.93$ are the only useful pieces of information in the output.

RESIDUALS: The (detrended) residual process is

$$\begin{aligned} \widehat{X}_t &= Y_t - \widehat{Y}_t \\ &= Y_t - 14.8 + 2.04\cos(2\pi t) - 0.93\sin(2\pi t), \end{aligned}$$

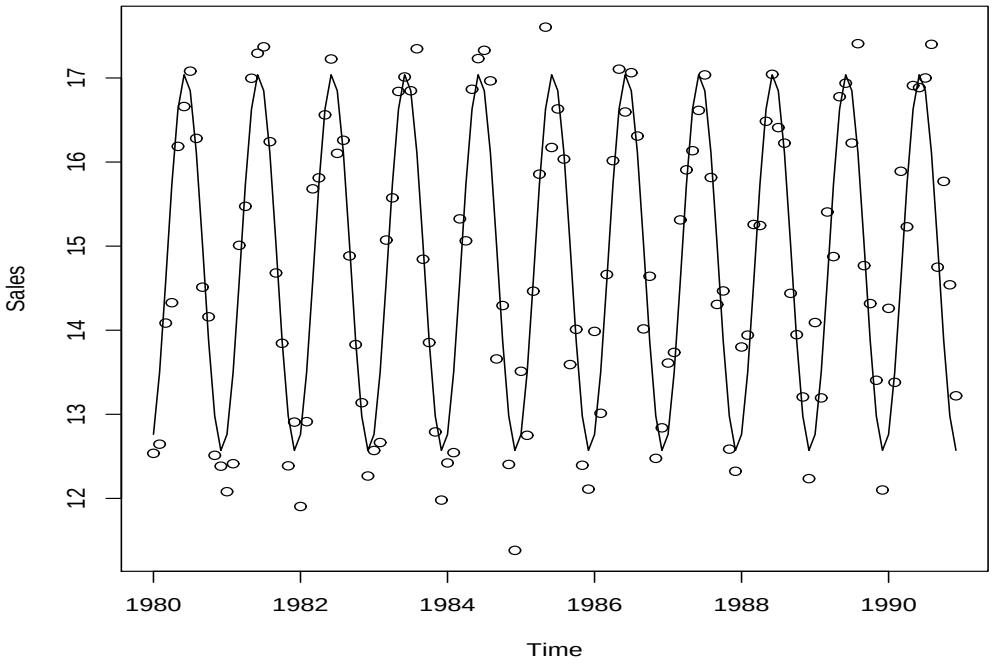


Figure 3.10: Beer sales data with a cosine trend model fit.

which is depicted in Figure 3.11. The residuals from the cosine trend fit appear to be somewhat stationary, but are probably not white noise.

REMARK: The seasonal means and cosine trend models are competing models; that is, both models are useful for seasonal data.

- The cosine trend model is more **parsimonious**; i.e., it is a simpler model because there are 3 regression parameters to estimate. On the other hand, the (monthly) seasonal means model has 12 parameters that need to be estimated!
- Remember, regression parameters (in any model) are estimated with the data. The more parameters we have in a model, the more data we need to use to estimate them. This leaves us with less information to estimate other quantities (e.g., residual variance, etc.). In the end, we have regression estimates that are less precise.
- The mathematical argument on pp 36-39 (CC) should convince you of this result.

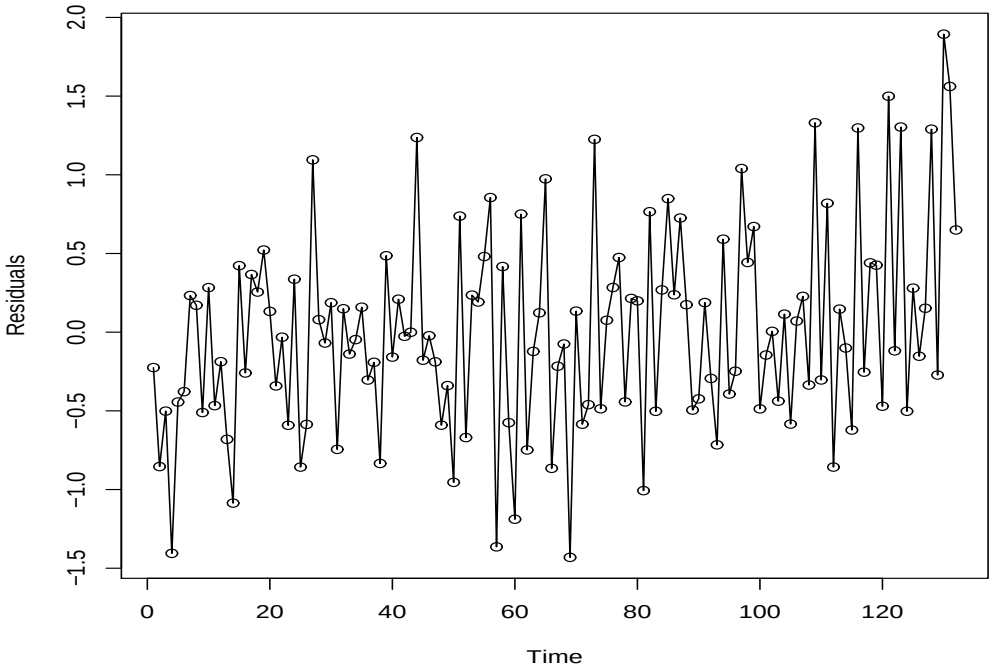


Figure 3.11: Beer sales data. Residuals from the cosine trend model fit.

3.4 Interpreting regression output

RECALL: In fitting the deterministic model

$$Y_t = \mu_t + X_t,$$

we have learned the following:

- for least squares estimates to be unbiased, all we need is $E(X_t) = 0$, for all t.
- for the variances of the least squares estimates (and standard errors) seen in R output to be meaningful, we need $E(X_t) = 0$, $\{X_t\}$ independent, and $\text{var}(X_t) = \gamma_0$ (constant). These assumptions are met if $\{X_t\}$ is a white noise process.
- for t tests and probability values to be valid, we need the last three assumptions to hold; in addition, normality is needed on the error process $\{X_t\}$.

NEW RESULT: If $\text{var}(X_t) = \gamma_0$ is constant, an estimate of γ_0 is given by

$$S^2 = \frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2,$$

where $\widehat{\mu}_t$ is the least squares estimate of μ_t and p is the number of regression parameters in μ_t. The term $n-p$ is called the **error degrees of freedom**. If $\{X_t\}$ is independent, then $E(S^2) = \gamma_0$; i.e., S^2 is an **unbiased estimator** of γ_0. The **residual standard deviation** is defined by,

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2},$$

the (positive) square root of S^2.

- The smaller S is, the better fit of the model. Therefore, in comparing two model fits (for two different models), we can look at the value of S in each model to judge which model may be preferred (caution is needed in doing this).
- The larger S is, the noisier the error process likely is. This makes the least squares estimates more variable and predictions less precise.

RESULT: For any data set $\{Y_t : t = 1, 2, ..., n\}$, we can write algebraically

$$\underbrace{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}_{\text{SST}} = \underbrace{\sum_{t=1}^{n}(\widehat{Y}_t - \overline{Y})^2}_{\text{SSR}} + \underbrace{\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2}_{\text{SSE}}.$$

These quantities are called **sums of squares** and form the basis for the following **analysis of variance** (ANOVA) **table**.

Source	df	SS	MS	F
Model	$p-1$	SSR	$\text{MSR} = \frac{\text{SSR}}{p-1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n-p$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p}$	
Total	$n-1$	SST		

COEFFICIENT OF DETERMINATION: Since SST = SSR + SSE, it follows that the proportion of the total variation in the data explained by the deterministic model is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

the **coefficient of determination**. The larger R^2 is, the better the deterministic part of the model explains the variability in the data. Clearly, $0 \leq R^2 \leq 1$.

IMPORTANT: It is critical to understand what R^2 does and does not measure. Its value is computed under the assumption that the deterministic trend model is correct and assesses how much of the variation in the data may be attributed to that relationship rather than just to inherent variation.

- If R^2 is small, it may be that there is a lot of random inherent variation in the data, so that, although the deterministic trend model is reasonable, it can only explain so much of the observed overall variation.
- Alternatively, R^2 may be close to 1, but a particular model may not be the best model. In fact, R^2 could be very "high," but not relevant because a better model may exist.

ADJUSTED R^2: A slight variant of the coefficient of determination is

$$\overline{R}^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}.$$

This is called the **adjusted R^2 statistic**. It is useful for comparing models with different numbers of parameters.

3.5 Residual analysis (model diagnostics)

RESIDUALS: Consider the deterministic trend model

$$Y_t = \mu_t + X_t,$$

where $E(X_t) = 0$. In this chapter, we have talked about using the method of least squares to fit models of this type (e.g., straight line regression, polynomial regression, seasonal means, cosine trends, etc.). The **fitted model** is $\widehat{Y}_t = \widehat{\mu}_t$ and the **residual process** is

$$\widehat{X}_t = Y_t - \widehat{Y}_t.$$

The residuals from the model fit are important. In essence, they serve as proxies (predictions) for the true errors X_t, which are not observed. The residuals can help us learn about the validity of the assumptions made in our model.

STANDARDIZED RESIDUALS: If the model above is fit using least squares (and there is an intercept term in the model), then algebraically,

$$\sum_{t=1}^{n}(Y_t - \widehat{Y}_t) = 0,$$

that is, the sum of the residuals is equal to zero. Thus, the residuals have mean zero and the **standardized residuals**, defined by

$$\widehat{X}_t^* = \frac{\widehat{X}_t}{S},$$

are unitless quantities. If desired, we can use the standardized residuals for model diagnostic purposes. The standardized residuals defined here are not exactly zero mean, unit variance quantities, but they are approximately so. Thus, if the model is adequate, we would expect most standardized residuals to fall between -3 and 3.

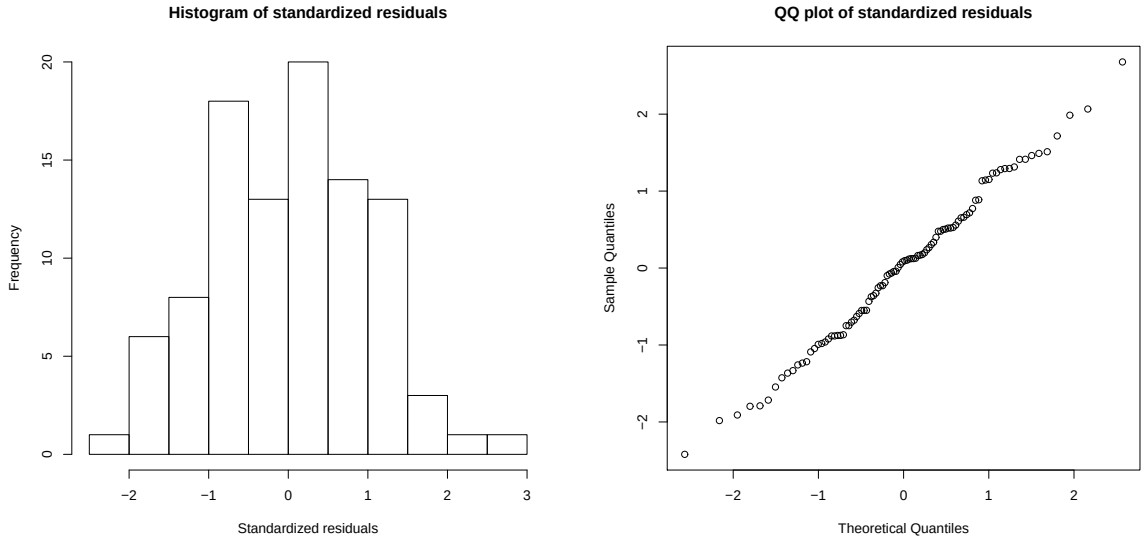
3.5.1 Assessing normality

NORMALITY: If the error process $\{X_t\}$ is normally distributed, then we would expect the residuals to also be approximately normally distributed. We can therefore diagnose this assumption by examining the (standardized) residuals and looking for evidence of normality. We can use histograms and **normal probability plots** (also known as quantile-quantile, or **qq plots**) to do this.

- Histograms which resemble heavily skewed empirical distributions are evidence against normality.

- A normal probability plot is a scatterplot of ordered residuals $\widehat{X}_t$ (or standardized residuals $\widehat{X}_t^*$) versus the ordered theoretical normal quantiles (or **normal scores**). The idea behind this plot is simple. If the residuals are normally distributed, then plotting them versus the corresponding normal quantiles (i.e., values from a normal distribution) should produce a straight line (or at least close).

Example 3.4 (continued). In Example 3.4, we fit a straight line trend model to the global temperature data. Below are the histogram and qq plot for the standardized residuals. Does normality seem to be supported?



SHAPIRO-WILK TEST: Histograms and qq plots provide only visual evidence of normality. The **Shapiro-Wilk test** is a formal hypothesis test that can be used to test

$$H_0 : \text{the (standardized) residuals are normally distributed}$$

versus

$$H_1 : \text{the (standardized) residuals are not normally distributed.}$$

The test is carried out by calculating a statistic W approximately equal to the **sample correlation** between the ordered (standardized) residuals and the normal scores. The

higher this correlation, the higher the value of W. Therefore, small values of W are evidence against H_0. The null distribution of W is very complicated, but probability values (p-values) are produced in R automatically. If the p-value is smaller than the significance level for the test (e.g., $\alpha = 0.05$, etc.), then we reject H_0 and conclude that there is a violation in the normality assumption. Otherwise, we do not reject H_0.

Example 3.4 (continued). In Example 3.4, we fit a straight line trend model to the global temperature data. The Shapiro-Wilk test on the standardized residuals produces the following output:

```
> shapiro.test(rstudent(fit))
Shapiro-Wilk normality test
data: rstudent(fit)
W = 0.9934, p-value = 0.915
```

Because the p-value for the test is not small, we do not reject H_0. This test does not provide evidence of non-normality for the standardized residuals.

3.5.2 Assessing independence

INDEPENDENCE: Plotting the residuals versus time can provide visual insight on whether or not the (standardized) residuals exhibit **independence** (although it is often easier to detect gross violations of independence). Residuals that "hang together" are not what we would expect to see from a sequence of independent random variables. Similarly, residuals that oscillate back and forth too notably also do not resemble this sequence.

RUNS TEST: A **runs test** is a nonparametric test which calculates the number of runs in the (standardized) residuals. The formal test is

$$H_0 : \text{the (standardized) residuals are independent}$$

versus

$$H_1 : \text{the (standardized) residuals are not independent.}$$

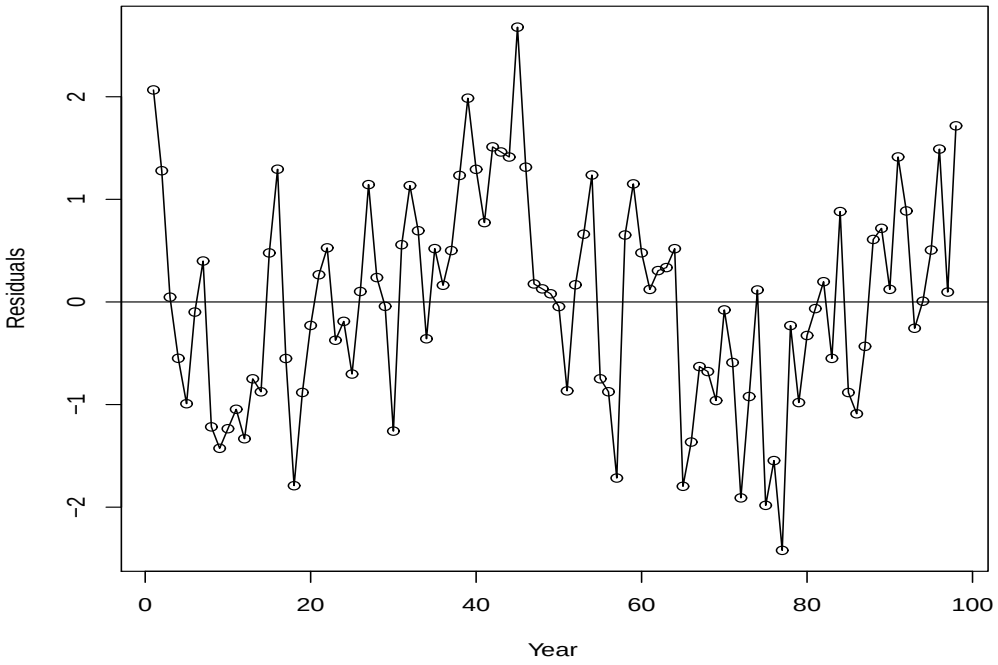


Figure 3.12: Standardized residuals from the straight line trend model fit for the global temperature data. A horizontal line at zero has been added.

In particular, the test examines the (standardized) residuals in sequence to look for patterns that would give evidence against independence. Runs above or below 0 (the approximate median of the residuals) are counted.

- A small number of runs would indicate that neighboring values are **positively dependent** and tend to hang together over time.
- Too many runs would indicate that the data oscillate back and forth across their median. This suggests that neighboring residuals are **negatively dependent**.
- Therefore, either too few or too many runs lead us to reject independence.

Example 3.4 (continued). In Example 3.4, we fit a straight line trend model to the global temperature data. A runs test on the standardized residuals produces the following output:

```
> runs(rstudent(fit))
$pvalue
[1] 3.65e-06
$observed.runs
[1] 27
$expected.runs
[1] 49.81633
```

The p-value for the test is extremely small, so we would reject H_0. The evidence points to the standardized residuals being not independent. The R output also produces the expected number of runs (computed under the assumption of independence). The observed number of runs is too much lower than the expected number to support independence.

BACKGROUND: If the (standardized) residuals are truly independent, it is possible to write out the probability mass function of R, the number of runs. This mass function is

$$f_R(r) = \begin{cases} \binom{n_1-1}{(r/2)-1}\binom{n_2-1}{(r/2)-1} \Big/ \binom{n_1+n_2}{n_1}, & \text{if } r \text{ is even} \\ \left[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\right] \Big/ \binom{n_1+n_2}{n_1}, & \text{if } r \text{ is odd,} \end{cases}$$

where

- n_1 = the number of residuals less than zero
- n_2 = the number of residuals greater than zero
- r_1 = the number of runs less than zero
- r_2 = the number of runs less than zero
- $r = r_1 + r_2$.

IMPLEMENTATION: When n_1 and n_2 are large, the number of runs R is approximately normally distribution with mean

$$\mu_R = 1 + \frac{2n_1n_2}{n}$$

and variance

$$\sigma^2_R = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}.$$

Therefore, values of

$$Z = \frac{|R - \mu_R|}{\sigma_R} > z_{\alpha/2}$$

lead to the rejection of H_0. The notation $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution.

3.5.3 Sample autocorrelation function

RECALL: Consider the stationary stochastic process $\{Y_t : t = 1, 2, ..., n\}$. In Chapter 2, we defined the autocorrelation function to be

$$\rho_k = \text{corr}(Y_t, Y_{t-k}) = \frac{\gamma_k}{\gamma_0},$$

where $\gamma_k = \text{cov}(Y_t, Y_{t-k})$ and $\gamma_0 = \text{var}(Y_t)$. Perhaps more aptly named, ρ_k is the **population autocorrelation function** because it depends on the true parameters for the process $\{Y_t\}$. In real life (that is, with real data) these population parameters are unknown, so we don't get to know the true ρ_k. However, we can estimate it. This leads to the definition of the sample autocorrelation function.

TERMINOLOGY: For a set of time series data $Y_1, Y_2, ..., Y_n$, we define the **sample autocorrelation function**, at lag k, by

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

where $\overline{Y}$ is the sample mean of $Y_1, Y_2, ..., Y_n$ (i.e., all the data are used to compute $\overline{Y}$). The sample version r_k is a point estimate of the true (population) autocorrelation ρ_k.

USAGE WITH STANDARDIZED RESIDUALS: Because we are talking about using standardized residuals to check regression model assumptions, we can examine the sample autocorrelation function of the standardized residual process $\{\widehat{X}_t^*\}$. Replacing Y_t with $\widehat{X}_t^*$ and $\overline{Y}$ with $\overline{\widehat{X}^*}$ in the above definition, we get

$$r_k^* = \frac{\sum_{t=k+1}^{n}(\widehat{X}_t^* - \overline{\widehat{X}^*})(\widehat{X}_{t-k}^* - \overline{\widehat{X}^*})}{\sum_{t=1}^{n}(\widehat{X}_t^* - \overline{\widehat{X}^*})^2}.$$

Note that when the sum of the standardized residuals equals zero (which occurs when least squares is used and when an intercept is included in the model), we also have $\overline{\widehat{X}^*} = 0$. Therefore, the formula above reduces to

$$r_k^* = \frac{\sum_{t=k+1}^{n} \widehat{X}_t^* \widehat{X}_{t-k}^*}{\sum_{t=1}^{n} (\widehat{X}_t^*)^2}.$$

IMPORTANT: If the standardized residual process $\{\widehat{X}_t^*\}$ is white noise, then

$$r_k^* \sim \mathcal{AN}\left(0, \frac{1}{n}\right),$$

for n large. The notation $\mathcal{AN}$ is read "approximately normal." For $k \neq l$, it also turns out that $\text{cov}(r_k^*, r_l^*) \approx 0$. These facts are established in Chapter 6.

- If the standardized residuals are truly white noise, then we would expect r_k^* to fall within 2 standard errors of 0. That is, values of r_k^* within $\pm 2/\sqrt{n}$ are within the margin of error under the white noise assumption.
- Values of r_k^* larger than $\pm 2/\sqrt{n}$ (in absolute value) are outside the margin of error, and, thus, are not consistent with what we would see from a white noise process. More specifically, this would suggest that there is dependence (autocorrelation) at lag k in the standardized residual process.

GRAPHICAL TOOL: The plot of r_k (or r_k^* if we are examining standardized residuals) versus k is called a **correlogram**. If we are assessing whether or not the process is white noise, it is helpful to put horizontal dashed lines at $\pm 2/\sqrt{n}$ so we can easily see if the sample autocorrelations fall outside the margin of error.

Example 3.4 (continued). In Example 3.4, we fit a straight line trend model to the global temperature data. In Figure 3.13, we display the correlogram for the standardized residuals $\{\widehat{X}_t^*\}$ from the straight line fit.

- Note that many of the sample estimates r_k^* fall outside the $\pm 2/\sqrt{n}$ margin of error cutoff. These residuals likely do not resemble a white noise process.

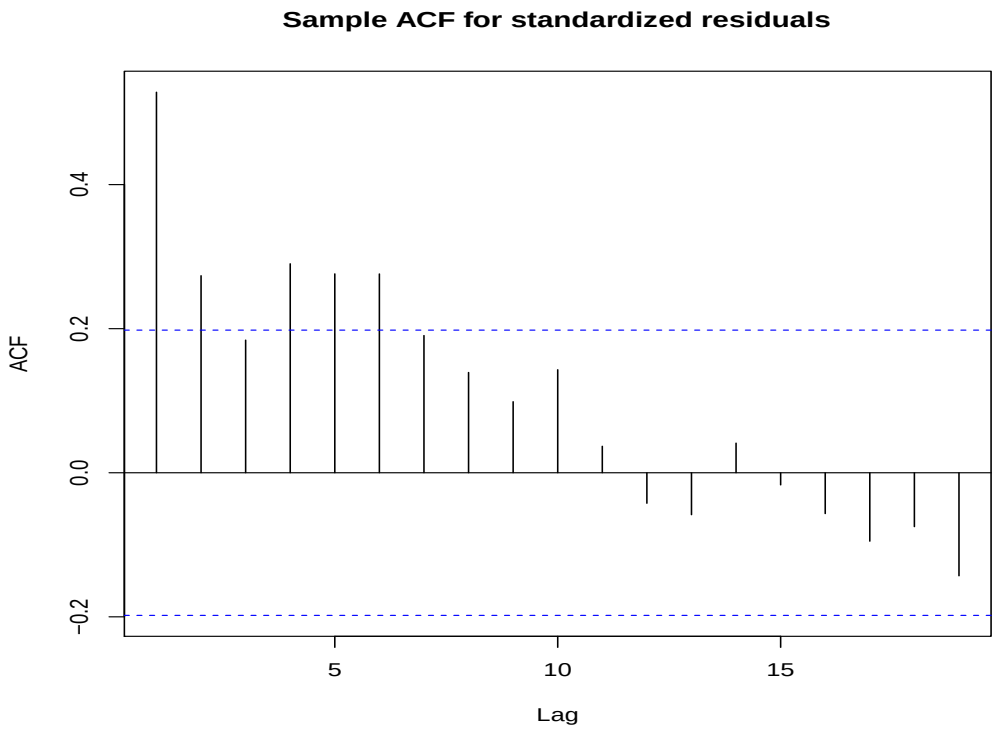


Figure 3.13: Global temperature data. Sample autocorrelation function for the standardized residuals from the straight line model fit.

- There is still a substantial amount of structure left in the residuals. In particular, there is strong positive autocorrelation at early lags and the sample ACF tends to decay somewhat as k increases.

SIMULATION EXERCISE: Let's generate some white noise processes and examine their sample autocorrelation functions! Figure 3.14 (left) displays two simulated white noise processes $e_t \sim$ iid $\mathcal{N}(0, 1)$, where $n = 100$. With $n = 100$, the margin of error for each sample autocorrelation r_k is

$$\text{margin of error } = \pm 2/\sqrt{100} = \pm 0.2.$$

Figure 3.14 (right) displays the sample correlograms (one for each simulated white noise series) with horizontal lines at the ± 0.2 margin of error cutoffs. Even though the generated data are truly white noise, we still do see some values of r_k (one for each realization)

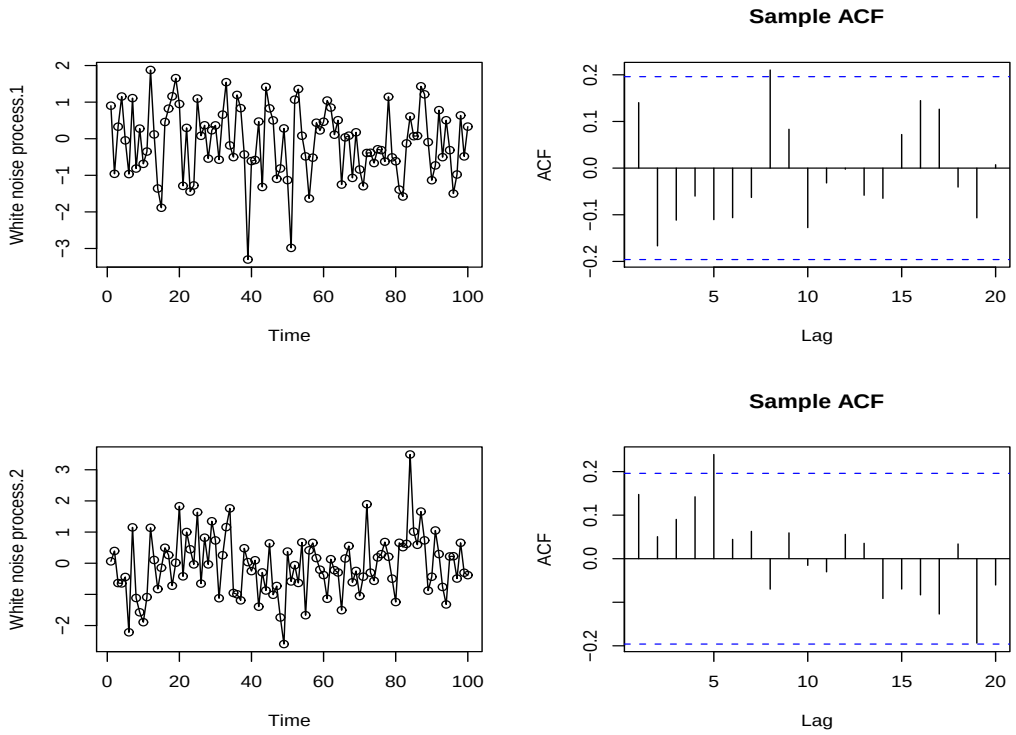


Figure 3.14: Two simulated standard normal white noise processes with their associated sample autocorrelation functions.

that fall outside the margin of error cutoffs. Why does this happen?

- In essence, every time we compare r_k to its margin of error cutoffs $\pm 2/\sqrt{n}$, we are performing a **hypothesis test**, namely, we are testing $H_0 : \rho_k = 0$ at a significance level of approximately $\alpha = 0.05$.
- Therefore, 5 percent of the time on average, we will observe a significant result which is really a "false alarm" (i.e., a Type I Error).
- When you are interpreting correlograms, keep this in mind. If there are patterns in the values of r_k and many which extend beyond the margin of error (especially at early lags), the series is probably not white noise. On the other hand, a stray statistically significant value of r_k at, say, lag $k = 17$ is likely just a false alarm.

4 Models for Stationary Time Series

Complementary reading: Chapter 4 (CC).

4.1 Introduction

RECALL: In the last chapter, we used regression to "detrend" time series data with the hope of removing non-stationary patterns and producing residuals that resembled a stationary process. We also learned that differencing can be an effective technique to transform a non-stationary process into one which is stationary. In this chapter, we consider (linear) time series models for stationary processes. Recall that stationary time series are those whose statistical properties do not change over time.

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. A **general linear process** is defined by

$$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \Psi_3 e_{t-3} + \cdots .$$

That is, Y_t, the value of the process at time t, is a weighted linear combination of white noise terms at the current and past times. *The processes that we examine in this chapter are special cases of this general linear process.* In general, $E(Y_t) = 0$ and

$$\gamma_k = \text{cov}(Y_t, Y_{t-k}) = \sigma_e^2 \sum_{i=0}^{\infty} \Psi_i \Psi_{i+k},$$

for $k \geq 0$, where we set $\Psi_0 = 1$.

- For mathematical reasons (to ensure stationarity), we will assume that the Ψ_i's are **square summable**, that is,
$$\sum_{i=1}^{\infty} \Psi_i^2 < \infty.$$
- A nonzero mean μ could be added to the right-hand side of the general linear process above; this would not affect the stationarity properties of $\{Y_t\}$. Therefore, there is no harm in assuming that the process $\{Y_t\}$ has zero mean.

4.2 Moving average processes

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The process

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

is called a **moving average process of order** q, denoted by **MA**(q). Note that this is a special case of the general linear process with $\Psi_0 = 1$, $\Psi_1 = -\theta_1$, $\Psi_2 = -\theta_2$, ..., $\Psi_q = -\theta_q$, and $\Psi_{q^*} = 0$ for all $q^* > q$.

4.2.1 MA(1) process

TERMINOLOGY: With $q = 1$, the moving average process defined above becomes

$$Y_t = e_t - \theta e_{t-1}.$$

This is called an **MA(1) process**. For this process, the mean is

$$E(Y_t) = E(e_t - \theta e_{t-1}) = E(e_t) - \theta E(e_{t-1}) = 0.$$

The variance is

$$\begin{aligned} \gamma_0 = \text{var}(Y_t) &= \text{var}(e_t - \theta e_{t-1}) \\ &= \text{var}(e_t) + \theta^2 \text{var}(e_{t-1}) - 2\theta \text{cov}(e_t, e_{t-1}) \\ &= \sigma_e^2 + \theta^2 \sigma_e^2 = \sigma_e^2(1 + \theta^2). \end{aligned}$$

The autocovariance at lag 1 is given by

$$\begin{aligned} \gamma_1 = \text{cov}(Y_t, Y_{t-1}) &= \text{cov}(e_t - \theta e_{t-1}, e_{t-1} - \theta e_{t-2}) \\ &= \text{cov}(e_t, e_{t-1}) - \theta \text{cov}(e_t, e_{t-2}) - \theta \text{cov}(e_{t-1}, e_{t-1}) + \theta^2 \text{cov}(e_{t-1}, e_{t-2}) \\ &= -\theta \text{var}(e_{t-1}) = -\theta \sigma_e^2. \end{aligned}$$

For any lag $k > 1$, $\gamma_k = \text{cov}(Y_t, Y_{t-k}) = 0$, because no white noise subscripts in Y_t and Y_{t-k} will overlap.

AUTOCOVARIANCE FUNCTION: For an MA(1) process,

$$\gamma_k = \begin{cases} \sigma_e^2(1+\theta^2), & k=0 \\ -\theta\sigma_e^2, & k=1 \\ 0, & k>1. \end{cases}$$

AUTOCORRELATION FUNCTION: For an MA(1) process,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k=0 \\ -\dfrac{\theta}{1+\theta^2}, & k=1 \\ 0, & k>1. \end{cases}$$

IMPORTANT: **The MA(1) process has zero correlation beyond lag $k = 1$!** Observations one time unit apart are correlated, but observations more than one time unit apart are not. This is important to keep in mind when we entertain models for real data using empirical evidence (e.g., sample autocorrelations r_k, etc.).

FACTS: The following theoretical results hold for an MA(1) process.

- When $\theta = 0$, the MA(1) process reduces to a white noise process.
- As θ ranges from -1 to 1, the (**population**) lag 1 autocorrelation ρ_1 ranges from 0.5 to -0.5; see pp 58 (CC).
- The largest ρ_1 can be is 0.5 (when $\theta = -1$) and the smallest ρ_1 can be is -0.5 (when $\theta = 1$). Therefore, if we were to observe a **sample** lag 1 autocorrelation r_1 that was well outside $[-0.5, 0.5]$, this would be inconsistent with the MA(1) model.
- The population lag 1 autocorrelation

  $$\rho_1 = -\frac{\theta}{1+\theta^2}$$

  remains the same if θ is replaced by $1/\theta$. Therefore, if someone told you the value of ρ_1 for an MA(1) process, you could not identify the corresponding value of θ uniquely. This is somewhat problematic and will have consequences in due course (e.g., when we discuss invertibility).

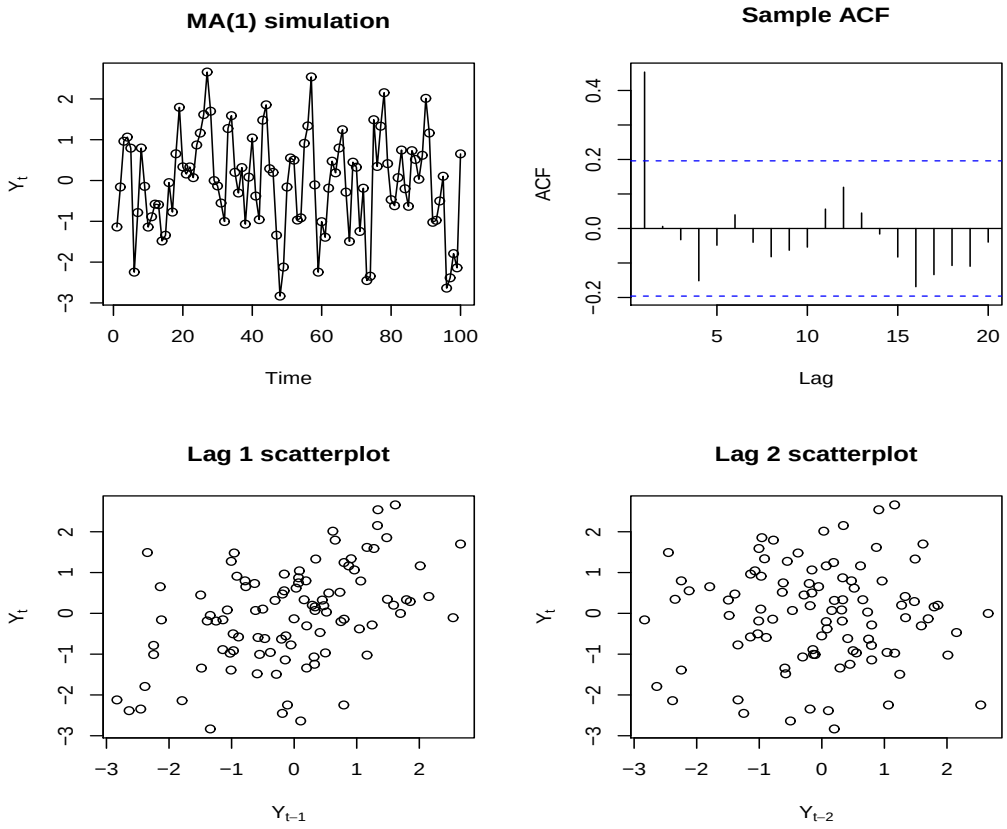


Figure 4.1: Upper left: MA(1) simulation with $\theta = -0.9$, $n = 100$, and $\sigma_e^2 = 1$. Upper right: Sample autocorrelation function r_k. Lower left: Scatterplot of Y_t versus Y_{t-1}. Lower right: Scatterplot of Y_t versus Y_{t-2}.

Example 4.1. We use R to simulate the MA(1) process $Y_t = e_t - \theta e_{t-1}$, where $\theta = -0.9$, $n = 100$, and $e_t \sim$ iid $\mathcal{N}(0, 1)$.

- Note that
$$\theta = -0.9 \implies \rho_1 = \frac{-(-0.9)}{1 + (-0.9)^2} \approx 0.497.$$
- There is a moderately strong positive autocorrelation at lag 1. Of course, $\rho_k = 0$, for all $k > 1$.
- The sample ACF in Figure 4.1 (upper right) looks like what we would expect from the MA(1) theory. There is a pronounced "spike" at $k = 1$ in the sample ACF and little action elsewhere (for $k > 1$). The error bounds at $\pm 2/\sqrt{100} = 0.2$ correspond

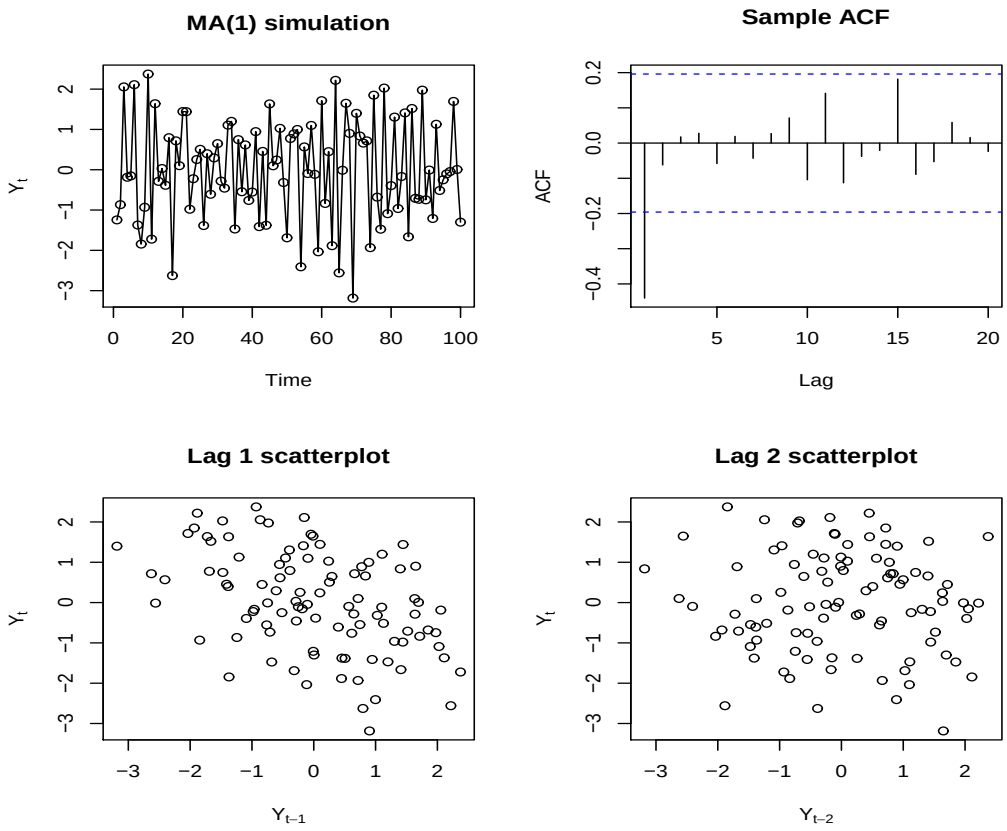


Figure 4.2: Upper left: MA(1) simulation with $\theta = 0.9$, $n = 100$, and $\sigma_e^2 = 1$. Upper right: Sample autocorrelation function r_k. Lower left: Scatterplot of Y_t versus Y_{t-1}. Lower right: Scatterplot of Y_t versus Y_{t-2}.

to those for a **white noise process**; not an MA(1) process.

- The lag 1 scatterplot; i.e., the scatterplot of Y_t versus Y_{t-1}, shows a moderate increasing linear relationship. This is expected because of the moderately strong positive lag 1 autorcorrelation.

- The lag 2 scatterplot; i.e., the scatterplot of Y_t versus Y_{t-2}, shows no linear relationship. This is expected because $\rho_2 = 0$ for an MA(1) process.

- Figure 4.2 displays a second MA(1) simulation, except with $\theta = 0.9$. In this model, $\rho_1 \approx -0.497$ and $\rho_k = 0$, for all $k > 1$. Compare Figure 4.2 with Figure 4.1.

4.2.2 MA(2) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The process

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

is a **moving average process of order 2**, denoted by **MA(2)**. For this process, the mean is

$$E(Y_t) = E(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) = E(e_t) - \theta_1 E(e_{t-1}) - \theta_2 E(e_{t-2}) = 0.$$

The variance is

$$\begin{aligned}
\gamma_0 = \text{var}(Y_t) &= \text{var}(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) \\
&= \text{var}(e_t) + \theta_1^2 \text{var}(e_{t-1}) + \theta_2^2 \text{var}(e_{t-2}) + \underbrace{6 \text{ covariance terms}}_{\text{all } = 0} \\
&= \sigma_e^2 + \theta_1^2 \sigma_e^2 + \theta_2^2 \sigma_e^2 = \sigma_e^2(1 + \theta_1^2 + \theta_2^2).
\end{aligned}$$

The autocovariance at lag 1 is given by

$$\begin{aligned}
\gamma_1 = \text{cov}(Y_t, Y_{t-1}) &= \text{cov}(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) \\
&= \text{cov}(-\theta_1 e_{t-1}, e_{t-1}) + \text{cov}(-\theta_2 e_{t-2}, -\theta_1 e_{t-2}) \\
&= -\theta_1 \text{var}(e_{t-1}) + (-\theta_2)(-\theta_1)\text{var}(e_{t-2}) \\
&= -\theta_1 \sigma_e^2 + \theta_1 \theta_2 \sigma_e^2 = (-\theta_1 + \theta_1 \theta_2)\sigma_e^2.
\end{aligned}$$

The autocovariance at lag 2 is given by

$$\begin{aligned}
\gamma_2 = \text{cov}(Y_t, Y_{t-2}) &= \text{cov}(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) \\
&= \text{cov}(-\theta_2 e_{t-2}, e_{t-2}) \\
&= -\theta_2 \text{var}(e_{t-2}) = -\theta_2 \sigma_e^2.
\end{aligned}$$

For any lag $k > 2$,

$$\gamma_k = \text{cov}(Y_t, Y_{t-k}) = 0,$$

because no white noise subscripts in Y_t and Y_{t-k} will overlap.

AUTOCOVARIANCE FUNCTION: For an MA(2) process,

$$\gamma_k = \begin{cases} \sigma_e^2(1+\theta_1^2+\theta_2^2), & k=0 \\ (-\theta_1+\theta_1\theta_2)\sigma_e^2, & k=1 \\ -\theta_2\sigma_e^2, & k=2 \\ 0 & k>2. \end{cases}$$

AUTOCORRELATION FUNCTION: For an MA(2) process,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k=0 \\ \dfrac{-\theta_1+\theta_1\theta_2}{1+\theta_1^2+\theta_2^2}, & k=1 \\ \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}, & k=2 \\ 0 & k>2. \end{cases}$$

IMPORTANT: **The MA(2) process has zero correlation beyond lag $k = 2$!** Observations 1 or 2 time units apart are correlated. Observations more than two time units apart are not correlated.

Example 4.2. We use R to simulate the MA(2) process

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2},$$

where $\theta_1 = 0.9$, $\theta_2 = -0.7$, $n = 100$, and $e_t \sim$ iid $\mathcal{N}(0, 1)$. For this process,

$$\rho_1 = \frac{-\theta_1+\theta_1\theta_2}{1+\theta_1^2+\theta_2^2} = \frac{-0.9+(0.9)(-0.7)}{1+(0.9)^2+(-0.7)^2} \approx -0.665$$

and

$$\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2} = \frac{-(-0.7)}{1+(0.9)^2+(-0.7)^2} \approx 0.304.$$

Figure 4.3 displays the simulated MA(2) time series, the sample ACF, and the lag 1 and 2 scatterplots. There are pronounced "spikes" at $k = 1$ and $k = 2$ in the sample ACF and little action elsewhere (for $k > 2$). The lagged scatterplots display negative (positive) autocorrelation at lag 1 (2). All of these observations are consistent with the MA(2) theory. Note that the error bounds at $\pm 2/\sqrt{100} = 0.2$ correspond to those for a **white noise process**; not an MA(2) process.

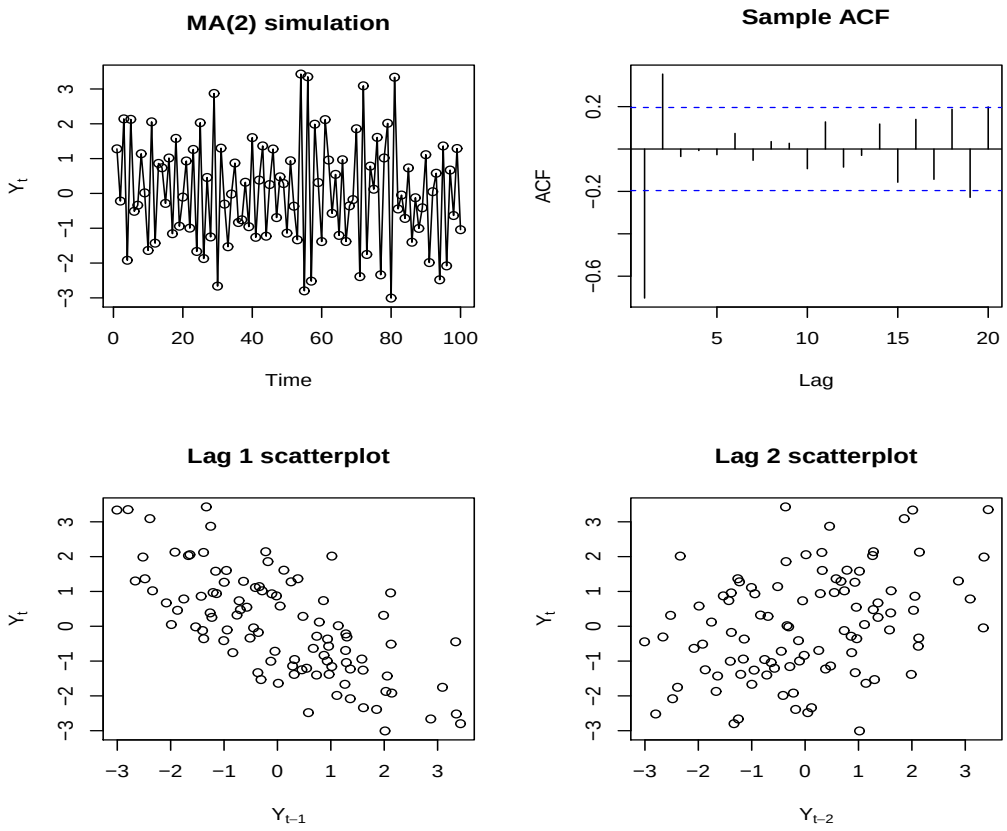


Figure 4.3: Upper left: MA(2) simulation with $\theta_1 = 0.9$, $\theta_2 = -0.7$, $n = 100$, and $\sigma_e^2 = 1$. Upper right: Sample autocorrelation function r_k. Lower left: Scatterplot of Y_t versus Y_{t-1}. Lower right: Scatterplot of Y_t versus Y_{t-2}.

4.2.3 MA(q) process

MODEL: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The MA($q$) process is

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}.$$

Standard calculations show that

$$E(Y_t) = 0$$

and

$$\gamma_0 = \text{var}(Y_t) = \sigma_e^2(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2).$$

AUTOCORRELATION FUNCTION: For an MA(q) process,

$$\rho_k = \begin{cases} 1, & k = 0 \\ \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}, & k = 1, 2, ..., q-1 \\ \dfrac{-\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}, & k = q \\ 0 & k > q. \end{cases}$$

The salient feature is that the (population) ACF ρ_k is **nonzero** for lags $k = 1, 2, ..., q$. For all lags $k > q$, the ACF $\rho_k = 0$.

4.3 Autoregressive processes

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

is called an **autoregressive process of order** p, denoted by **AR**(p).

- In this model, the value of the process at time t, Y_t, is a weighted linear combination of the values of the process from the previous p time points plus a "shock" or "innovation" term e_t at time t.
- We assume that e_t, the innovation at time t, is **independent** of all previous process values $Y_{t-1}, Y_{t-2}, ...,$.
- We continue to assume that $E(Y_t) = 0$. A nonzero mean could be added to the model by replacing Y_t with $Y_t - \mu$ (for all t). This would not affect the stationarity properties.
- This process (assuming that is stationary) is a special case of the general linear process defined at the beginning of this chapter.

4.3.1 AR(1) process

TERMINOLOGY: Take $p = 1$ in the general AR(p) process and we get

$$Y_t = \phi Y_{t-1} + e_t.$$

This is an **AR(1) process**. Note that if $\phi = 1$, this process reduces to a random walk. If $\phi = 0$, this process reduces to white noise.

VARIANCE: Assuming that this process is stationary (it isn't always), the variance of Y_t can be obtained in the following way. In the AR(1) equation, take variances of both sides to get

$$\begin{aligned} \text{var}(Y_t) &= \text{var}(\phi Y_{t-1} + e_t) \\ &= \phi^2 \text{var}(Y_{t-1}) + \text{var}(e_t) + \underbrace{2\phi \text{cov}(Y_{t-1}, e_t)}_{=\,0} \\ &= \phi^2 \text{var}(Y_{t-1}) + \sigma_e^2. \end{aligned}$$

Assuming stationarity, $\text{var}(Y_t) = \text{var}(Y_{t-1}) = \gamma_0$. Therefore, we have

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_e^2 \quad \Longrightarrow \quad \gamma_0 = \frac{\sigma_e^2}{1 - \phi^2}.$$

Because $\gamma_0 > 0$, this equation implies that $0 < \phi^2 < 1$, that is, $-1 < \phi < 1$.

AUTOCOVARIANCE: To find the autocovariance function γ_k, multiply both sides of the AR(1) equation by Y_{t-k} to get

$$Y_t Y_{t-k} = \phi Y_{t-1} Y_{t-k} + e_t Y_{t-k}.$$

Taking expectations of both sides, we have

$$E(Y_t Y_{t-k}) = \phi E(Y_{t-1} Y_{t-k}) + E(e_t Y_{t-k}).$$

We now make the following observations:

- Because e_t is independent of Y_{t-k} (by assumption), we have

$$E(e_t Y_{t-k}) = E(e_t) E(Y_{t-k}) = 0.$$

- Because $\{Y_t\}$ is a zero mean process (by assumption), we have

$$\begin{aligned}\gamma_k = \text{cov}(Y_t, Y_{t-k}) &= E(Y_t Y_{t-k}) - E(Y_t)E(Y_{t-k}) = E(Y_t Y_{t-k}) \\ \gamma_{k-1} = \text{cov}(Y_{t-1}, Y_{t-k}) &= E(Y_{t-1} Y_{t-k}) - E(Y_{t-1})E(Y_{t-k}) = E(Y_{t-1} Y_{t-k}).\end{aligned}$$

From these two observations, we have established the following (recursive) relationship for an AR(1) process:

$$\gamma_k = \phi\gamma_{k-1}.$$

When $k = 1$,

$$\gamma_1 = \phi\gamma_0 = \phi\left(\frac{\sigma_e^2}{1-\phi^2}\right).$$

When $k = 2$,

$$\gamma_2 = \phi\gamma_1 = \phi^2\left(\frac{\sigma_e^2}{1-\phi^2}\right).$$

This pattern continues for larger k. In general, the autocovariance function for an AR(1) process is

$$\gamma_k = \phi^k\left(\frac{\sigma_e^2}{1-\phi^2}\right), \quad \text{for } k = 0, 1, 2, ..., .$$

AUTOCORRELATION: For an AR(1) process,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\phi^k\left(\frac{\sigma_e^2}{1-\phi^2}\right)}{\frac{\sigma_e^2}{1-\phi^2}} = \phi^k, \quad \text{for } k = 0, 1, 2, ..., .$$

IMPORTANT: For an AR(1) process, because $-1 < \phi < 1$, the (population) ACF $\rho_k = \phi_k$ decays **exponentially** as k increases.

- If ϕ is close to ± 1, then the decay will be more slowly.
- If ϕ is not close to ± 1, then the decay will take place rapidly.
- If $\phi > 0$, then all of the autocorrelations will be positive.
- If $\phi < 0$, then the autocorrelations will alternate from negative ($k = 1$), to positive ($k = 2$), to negative ($k = 3$), to positive ($k = 4$), and so on.
- Remember these theoretical patterns so that when we see sample ACFs (from real data!), we can make sensible decisions about potential model selection.

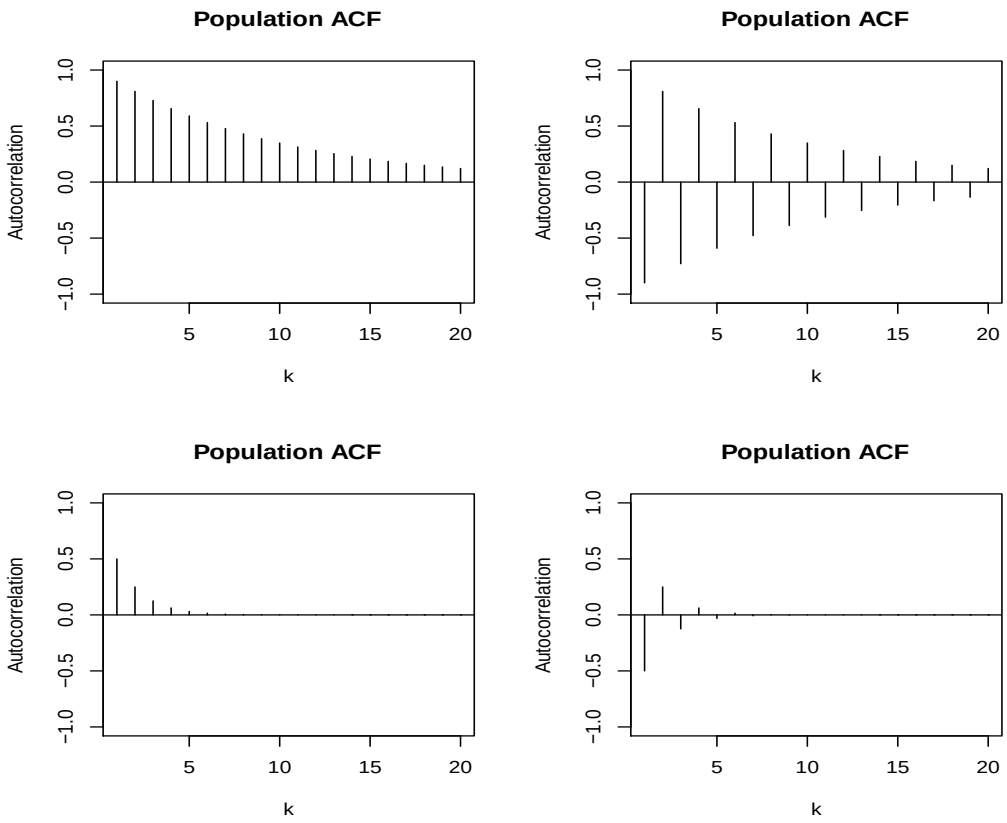


Figure 4.4: Population ACFs for AR(1) processes. Upper left: $\phi = 0.9$. Upper right: $\phi = -0.9$. Lower left: $\phi = 0.5$. Lower right: $\phi = -0.5$.

Example 4.3. We use R to simulate four different AR(1) processes

$$Y_t = \phi Y_{t-1} + e_t,$$

with $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 100$. We choose

- $\phi = 0.9$ (large ρ_1, ACF should decay slowly, all ρ_k positive)
- $\phi = -0.9$ (large ρ_1, ACF should decay slowly, ρ_k alternating)
- $\phi = 0.5$ (moderate ρ_1, ACF should decay more quickly, all ρ_k positive)
- $\phi = -0.5$ (moderate ρ_1, ACF should decay more quickly, ρ_k alternating).

These choices of ϕ are consistent with those in Figure 4.4, which depicts the true (population) AR(1) autocorrelation functions.

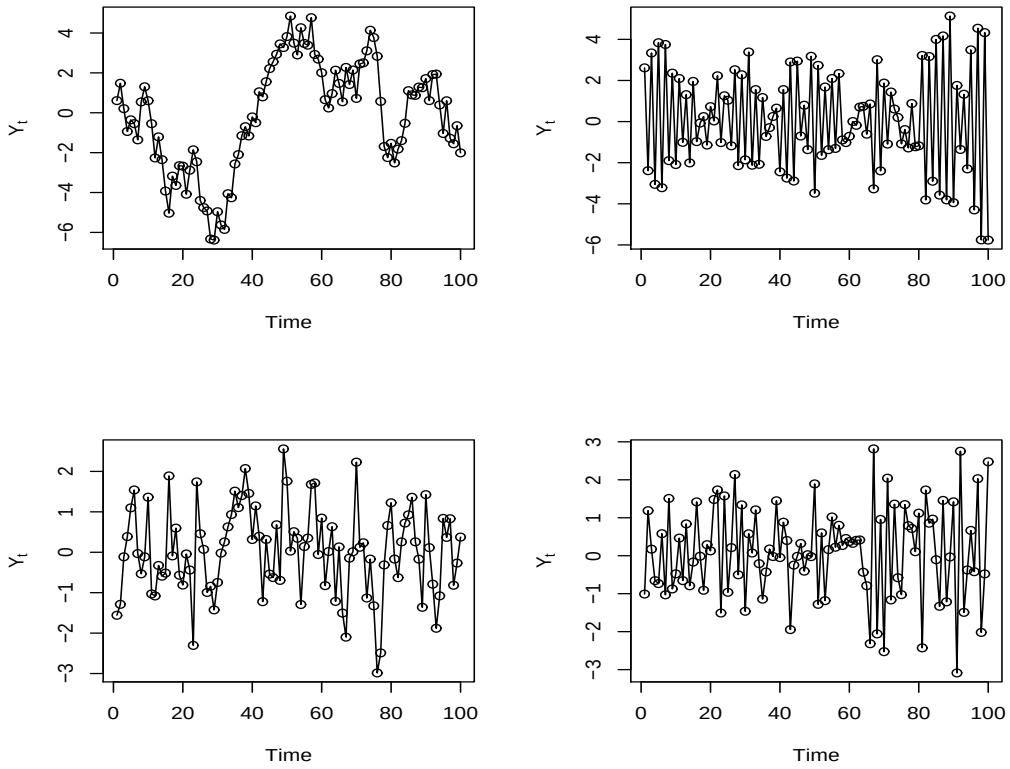


Figure 4.5: AR(1) simulations with $n = 100$ and $\sigma_e^2 = 1$. Upper left: $\phi = 0.9$. Upper right: $\phi = -0.9$. Lower left: $\phi = 0.5$. Lower right: $\phi = -0.5$.

- In Figure 4.5, note the differences between the series on the left ($\phi > 0$) and the series on the right ($\phi < 0$).
  - When $\phi > 0$, the series tends to "hang together" (since $\rho_1 > 0$).
  - When $\phi < 0$, there is more oscillation (since $\rho_1 < 0$).
- In Figure 4.6, we display the sample autocorrelation functions. Compare the sample ACFs to the theoretical ACFs in Figure 4.4. The fact that these figures do not agree completely is a byproduct of the sample autocorrelations r_k exhibiting sampling variability. The error bounds at $\pm 2/\sqrt{100} = 0.2$ correspond to those for a **white noise process**; not an AR(1) process.

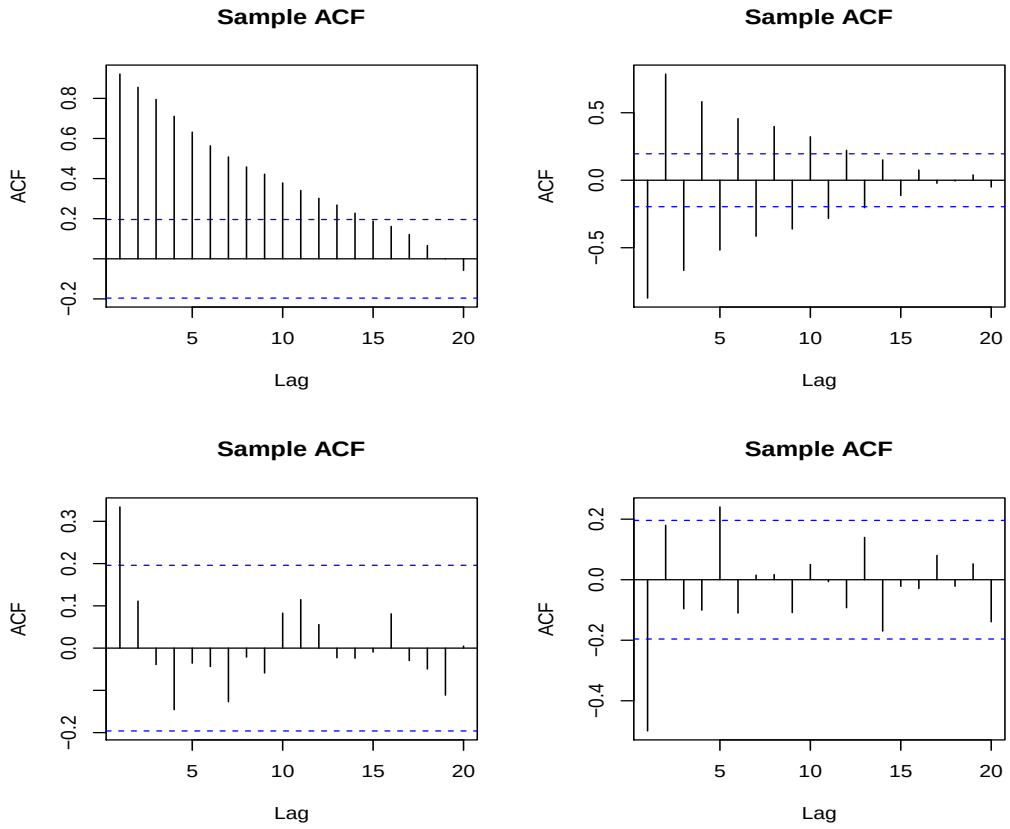


Figure 4.6: Sample ACFs for AR(1) simulations with $n = 100$ and $\sigma_e^2 = 1$. Upper left: $\phi = 0.9$. Upper right: $\phi = -0.9$. Lower left: $\phi = 0.5$. Lower right: $\phi = -0.5$.

OBSERVATION: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. We now show that the AR(1) process

$$Y_t = \phi Y_{t-1} + e_t$$

can be written in the form of a general linear process

$$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \Psi_3 e_{t-3} + \cdots.$$

To show this, write $Y_{t-1} = \phi Y_{t-2} + e_{t-1}$ so that

$$\begin{aligned} Y_t &= \phi Y_{t-1} + e_t \\ &= \phi(\phi Y_{t-2} + e_{t-1}) + e_t \\ &= e_t + \phi e_{t-1} + \phi^2 Y_{t-2}. \end{aligned}$$

Substituting in $Y_{t-2} = \phi Y_{t-3} + e_{t-2}$, we get

$$\begin{aligned} Y_t &= e_t + \phi e_{t-1} + \phi^2(\phi Y_{t-3} + e_{t-2}) \\ &= e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 Y_{t-3}. \end{aligned}$$

Continuing this type of substitution indefinitely, we get

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots .$$

Therefore, the AR(1) process is a special case of the general linear process with $\Psi_j = \phi^j$, for $j = 0, 1, 2, ...,$.

STATIONARITY CONDITION: The AR(1) process

$$Y_t = \phi Y_{t-1} + e_t$$

is **stationary** if and only if $|\phi| < 1$, that is, if $-1 < \phi < 1$. If $|\phi| \geq 1$, then the AR(1) process is not stationary.

4.3.2 AR(2) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The **AR(2) process** is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t.$$

- The current value of the process, Y_t, is a weighted linear combination of the values of the process from the previous **two** time periods, plus a random innovation (error) at the current time.
- We continue to assume that $E(Y_t) = 0$. A nonzero mean μ could be added to model by replacing Y_t with $Y_t - \mu$ for all t.
- We continue to assume that e_t is independent of Y_{t-k}, for all $k = 1, 2, ...,$.

- Just as the AR(1) model requires certain conditions for stationarity, the AR(2) model does too. A thorough discussion of stationarity for the AR(2) model, and higher order AR models, becomes very theoretical. We highlight only the basic points.

TERMINOLOGY: First, we define the operator B to satisfy

$$BY_t = Y_{t-1},$$

that is, B "backs up" the current value Y_t one time unit to Y_{t-1}. For this reason, we call B the **backshift operator**. Similarly,

$$B^2Y_t = BBY_t = BY_{t-1} = Y_{t-2}.$$

In general, $B^kY_t = Y_{t-k}$. Using this new notation, we can rewrite the AR(2) model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$$

in the following way:

$$Y_t = \phi_1 BY_t + \phi_2 B^2 Y_t + e_t.$$

Rewriting this equation, we get

$$Y_t - \phi_1 BY_t - \phi_2 B^2 Y_t = e_t \iff (1 - \phi_1 B - \phi_2 B^2)Y_t = e_t.$$

Finally, treating the B as a dummy variable for algebraic reasons (and using the more conventional algebraic symbol x), we define the **AR(2) characteristic polynomial** as

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2$$

and the corresponding **AR(2) characteristic equation** to be

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 = 0.$$

IMPORTANT: Characterizing the stationarity conditions for the AR(2) model is done by examining this equation and the solutions to it; i.e., the **roots** of $\phi(x) = 1 - \phi_1 x - \phi_2 x^2$.

NOTE: Applying the quadratic formula to the AR(2) characteristic equation, we see that the roots of $\phi(x) = 1 - \phi_1 x - \phi_2 x^2$ are

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}.$$

- The roots are both real if $\phi_1^2 + 4\phi_2 > 0$.
- The roots are both complex if $\phi_1^2 + 4\phi_2 < 0$
- There is a single real root with multiplicity 2 if $\phi_1^2 + 4\phi_2 = 0$.

STATIONARITY CONDITIONS: The AR(2) process is **stationary** when the roots of $\phi(x) = 1 - \phi_1 x - \phi_2 x^2$ both exceed 1 in absolute value (or in modulus if the roots are complex). This occurs if and only if

$$\phi_1 + \phi_2 < 1 \qquad \phi_2 - \phi_1 < 1 \qquad |\phi_2| < 1$$

(see Appendix B, pp 84, CC). These are the **stationarity conditions** for the AR(2) model. A sketch of this stationarity region (in the ϕ_1-ϕ_2 plane) appears in Figure 4.7.

RECALL: Define $i = \sqrt{-1}$ so that $z = a + bi$ is a complex number. The **modulus** of $z = a + bi$ is

$$|z| = \sqrt{a^2 + b^2}.$$

AUTOCORRELATION FUNCTION: To derive the population ACF for an AR(2) process, start with the AR(2) model equation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$$

and multiply both sides by Y_{t-k} to get

$$Y_t Y_{t-k} = \phi_1 Y_{t-1} Y_{t-k} + \phi_2 Y_{t-2} Y_{t-k} + e_t Y_{t-k}.$$

Taking expectations of both sides gives

$$E(Y_t Y_{t-k}) = \phi_1 E(Y_{t-1} Y_{t-k}) + \phi_2 E(Y_{t-2} Y_{t-k}) + E(e_t Y_{t-k}).$$

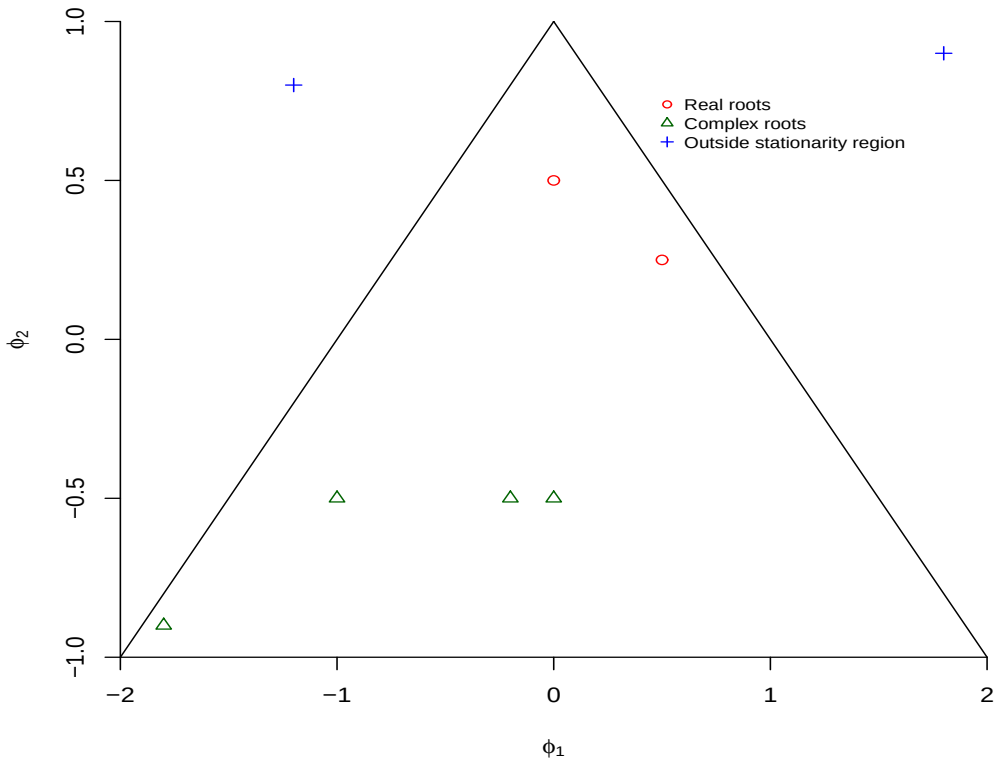


Figure 4.7: Stationarity region for the AR(2) model. The point (ϕ_1, ϕ_2) must fall inside the triangular region to satisfy the stationarity conditions. Points falling below the curve $\phi_1^2 + 4\phi_2 = 0$ are complex solutions. Those falling above $\phi_1^2 + 4\phi_2 = 0$ are real solutions.

Because $\{Y_t\}$ is a zero mean process, $E(Y_tY_{t-k}) = \gamma_k$, $E(Y_{t-1}Y_{t-k}) = \gamma_{k-1}$, and $E(Y_{t-2}Y_{t-k}) = \gamma_{k-2}$. Because e_t is independent of Y_{t-k}, $E(e_tY_{t-k}) = E(e_t)E(Y_{t-k}) = 0$. This proves that

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2}.$$

Dividing through by $\gamma_0 = \text{var}(Y_t)$ gives

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2}.$$

These are called the **Yule-Walker equations** for the AR(2) process.

NOTE: For $k = 1$ and $k = 2$, the Yule-Walker equations provide

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 \\ \rho_2 &= \phi_1\rho_1 + \phi_2, \end{aligned}$$

where $\rho_0 = 1$. Solving this system for ρ_1 and ρ_2, we get

$$\rho_1 = \frac{\phi_1}{1-\phi_2} \qquad \text{and} \qquad \rho_2 = \frac{\phi_1^2 + \phi_2 - \phi_2^2}{1-\phi_2}.$$

- Therefore, we have closed-form expressions for ρ_1 and ρ_2 in terms of ϕ_1 and ϕ_2.
- If we want to find higher lag autocorrelations, we can use the (recursive) relation

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2}.$$

  For example, $\rho_3 = \phi_1\rho_2 + \phi_2\rho_1$, $\rho_4 = \phi_1\rho_3 + \phi_2\rho_2$, and so on.

REMARK: For those of you that like formulas, it is possible to write out closed-form expressions for the autocorrelations in an AR(2) process. Denote the roots of the AR(2) characteristic polynomial by $1/G_1$ and $1/G_2$ and assume that these roots both exceed 1 in absolute value (or modulus). Straightforward algebra shows that

$$\begin{aligned} G_1 &= \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} \\ G_2 &= \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2}. \end{aligned}$$

- If $G_1 \neq G_2$, then

$$\rho_k = \frac{(1-G_2^2)G_1^{k+1} - (1-G_1^2)G_2^{k+1}}{(G_1 - G_2)(1 + G_1G_2)}.$$

- If $1/G_1$ and $1/G_2$ are complex (i.e., when $\phi_1^2 + 4\phi_2 < 0$), then

$$\rho_k = R^k \frac{\sin(\Theta k + \Phi)}{\sin(\Phi)},$$

  where $R = \sqrt{-\phi_2}$, $\Theta = \cos^{-1}(\phi_1/2\sqrt{-\phi_2})$, and $\Phi = \tan^{-1}[(1-\phi_2)/(1+\phi_2)]$.

- If $G_1 = G_2$ (i.e., when $\phi_1^2 + 4\phi_2 = 0$), then

$$\rho_k = \left[1 + k\left(\frac{1+\phi_2}{1-\phi_2}\right)\right]\left(\frac{\phi_1}{2}\right)^k.$$

DISCUSSION: Personally, I don't think these formulas are all that helpful for computation purposes. So, why present them? After all, we could use the Yule-Walker equations for computation.

- The formulas are helpful in that they reveal typical shapes of the AR(2) population ACFs. This is important because when we see these shapes with real data (through the sample ACFs), this will aid us in model selection/identification.

- Denote the roots of the AR(2) characteristic polynomial by $1/G_1$ and $1/G_2$. If the AR(2) process is stationary, then both of these roots are larger than 1 (in absolute value or modulus). However,

$$|1/G_1| > 1,\ |1/G_2| > 1 \quad \Longrightarrow \quad |G_1| < 1,\ |G_2| < 1.$$

  Therefore, each of

$$\begin{aligned}
\rho_k &= \frac{(1-G_2^2)G_1^{k+1} - (1-G_1^2)G_2^{k+1}}{(G_1 - G_2)(1+G_1G_2)} \\
\rho_k &= R^k \frac{\sin(\Theta k + \Phi)}{\sin(\Phi)} \\
\rho_k &= \left[1 + k\left(\frac{1+\phi_2}{1-\phi_2}\right)\right]\left(\frac{\phi_1}{2}\right)^k
\end{aligned}$$

  satisfies the following:

$$\rho_k \to 0, \quad \text{as } k \to \infty.$$

- Therefore, in an AR(2) process, the population autocorrelations ρ_k (in magnitude) decay towards zero as k increases. Further inspection reveals that the decay is exponential in nature. In addition, when the roots are complex, the values of ρ_k resemble a sinusoidal pattern that dampens out as k increases.

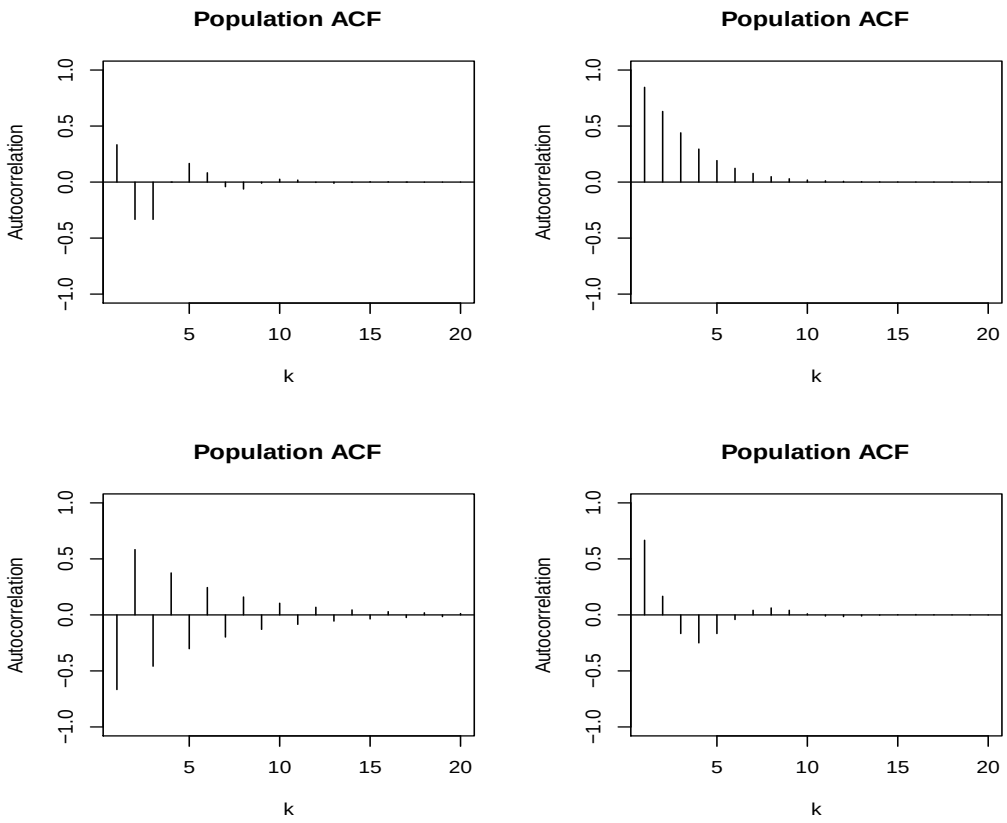


Figure 4.8: Population ACFs for AR(2) processes. Upper left: $(\phi_1, \phi_2) = (0.5, -0.5)$. Upper right: $(\phi_1, \phi_2) = (1.1, -0.3)$. Lower left: $(\phi_1, \phi_2) = (-0.5, 0.25)$. Lower right: $(\phi_1, \phi_2) = (1, -0.5)$.

Example 4.4. We use R to simulate four AR(2) processes $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$, with $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 100$. We choose

- $(\phi_1, \phi_2) = (0.5, -0.5)$. CP: $\phi(x) = 1 - 0.5x + 0.5x^2$. Complex roots.
- $(\phi_1, \phi_2) = (1.1, -0.3)$. CP: $\phi(x) = 1 - 1.1x + 0.3x^2$. Two distinct (real) roots.
- $(\phi_1, \phi_2) = (-0.5, 0.25)$. CP: $\phi(x) = 1 + 0.5x - 0.25x^2$. Two distinct (real) roots.
- $(\phi_1, \phi_2) = (1, -0.5)$. CP: $\phi(x) = 1 - x + 0.5x^2$. Complex roots.

These choices of (ϕ_1, ϕ_2) are consistent with those in Figure 4.8 that depict the true (population) AR(2) autocorrelation functions.

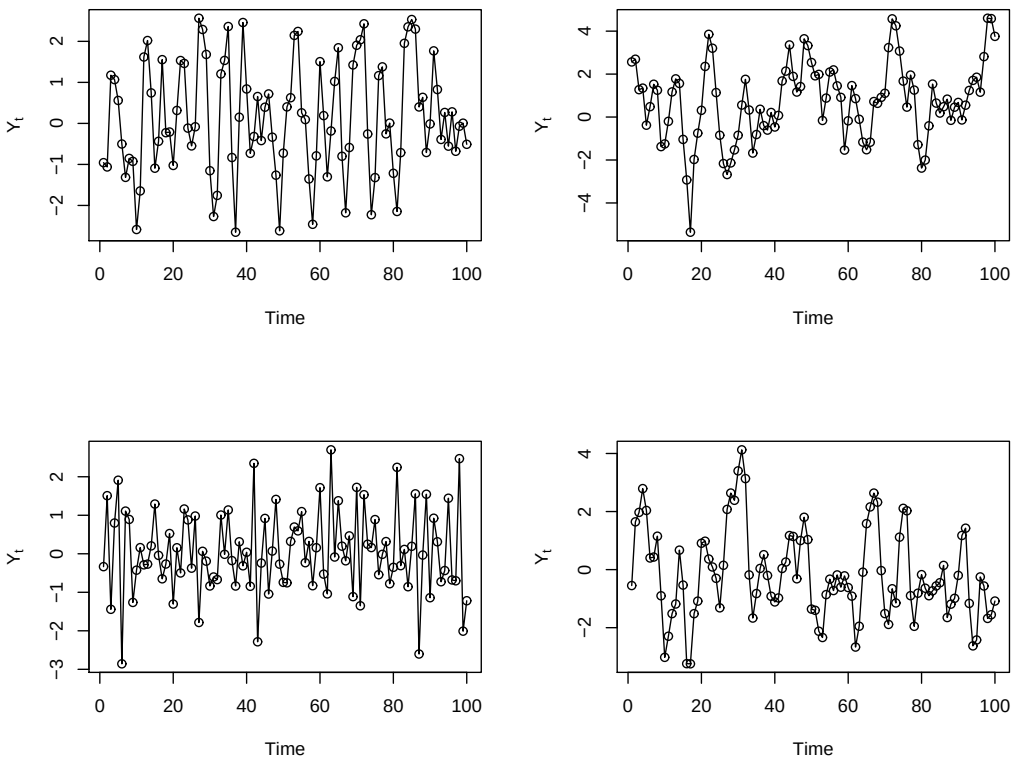


Figure 4.9: AR(2) simulations with $n = 100$ and $\sigma_e^2 = 1$. Upper left: $(\phi_1, \phi_2) = (0.5, -0.5)$. Upper right: $(\phi_1, \phi_2) = (1.1, -0.3)$. Lower left: $(\phi_1, \phi_2) = (-0.5, 0.25)$. Lower right: $(\phi_1, \phi_2) = (1, -0.5)$.

- Consistent with the theory (see the population ACFs in Figure 4.8), the first (upper left), second (upper right), and the fourth (lower right) series do "hang together;" this is because of the positive lag 1 autocorrelation. The third series (lower left) tends to oscillate, as we would expect since $\rho_1 < 0$.

- The sample ACFs in Figure 4.10 resemble somewhat their theoretical counterparts (at least at the first lag). Later lags generally deviate from the known theoretical autocorrelations (there is a good reason for this). The error bounds at $\pm 2/\sqrt{100} = 0.2$ correspond to those for a **white noise process**; not an AR(1) process.

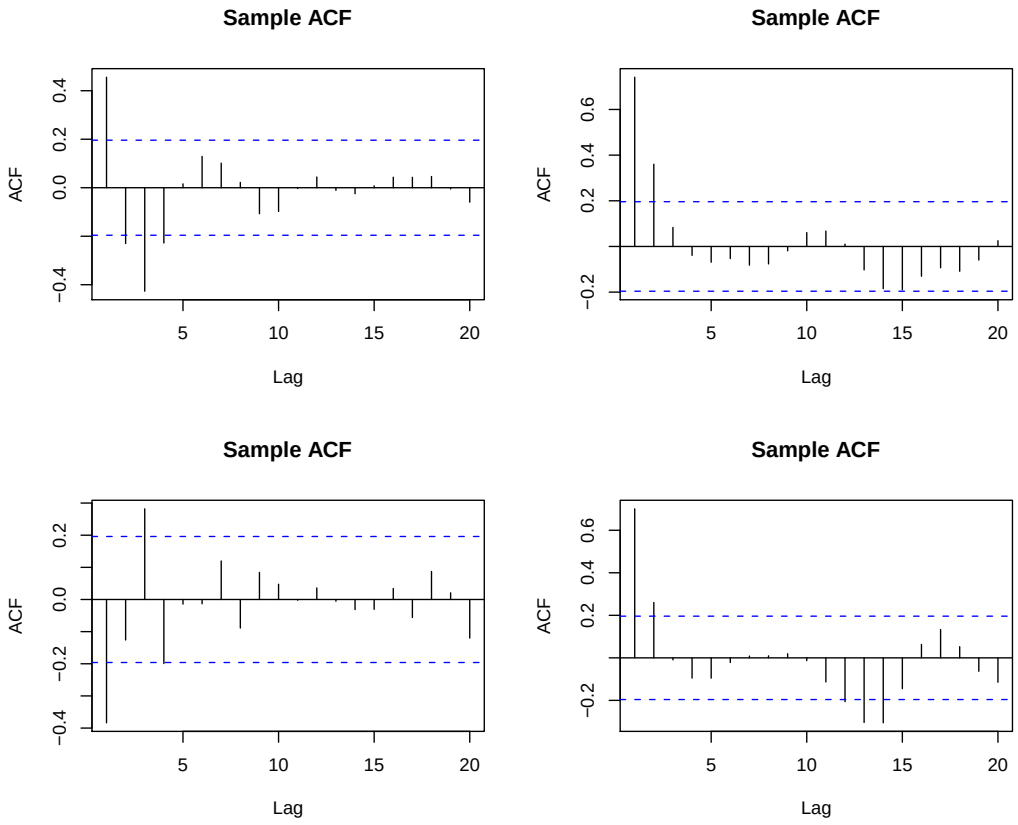


Figure 4.10: Sample ACFs for AR(2) simulations with $n = 100$ and $\sigma_e^2 = 1$. Upper left: $(\phi_1, \phi_2) = (0.5, -0.5)$. Upper right: $(\phi_1, \phi_2) = (1.1, -0.3)$. Lower left: $(\phi_1, \phi_2) = (-0.5, 0.25)$. Lower right: $(\phi_1, \phi_2) = (1, -0.5)$.

VARIANCE: For the AR(2) process,

$$\gamma_0 = \text{var}(Y_t) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_e^2}{(1-\phi_2)^2 - \phi_1^2}.$$

NOTE: The AR(2) model can be expressed as a **general linear process**

$$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \Psi_3 e_{t-3} + \cdots.$$

If $1/G_1$ and $1/G_2$ are the roots of the AR(2) characteristic polynomial, then

$$\Psi_j = \frac{G_1^{j+1} - G_2^{j+1}}{G_1 - G_2}, \qquad \Psi_j = R^j \frac{\sin[(j+1)\Theta]}{\sin(\Theta)}, \qquad \Psi_j = (1+j)\phi_1^j,$$

depending on if $G_1 \neq G_2$, G_1 and G_2 are complex, or $G_1 = G_2$, respectively.

4.3.3 AR(p) process

RECALL: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The general autoregressive process of order p, denoted **AR**(p), is

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t.$$

In backshift operator notation, we can write the model as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Y_t = e_t,$$

yielding the **AR**(p) **characteristic equation**

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0.$$

IMPORTANT: An AR(p) process is stationary if and only if the p roots of $\phi(x)$ each exceed 1 in absolute value (or in modulus if the roots are complex).

- Consider an AR(1) process

$$Y_t = \phi Y_{t-1} + e_t \iff (1 - \phi B) Y_t = e_t.$$

  The AR(1) characteristic polynomial is $\phi(x) = 1 - \phi x$. Therefore,

$$\phi(x) = 1 - \phi x = 0 \implies x = \frac{1}{\phi}.$$

  Clearly,

$$|x| = \left|\frac{1}{\phi}\right| > 1 \iff |\phi| < 1,$$

  which was the stated stationarity condition for the AR(1) process.

- Consider an AR(2) process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \iff (1 - \phi_1 B - \phi_2 B^2) Y_t = e_t.$$

  The AR(2) characteristic polynomial is $\phi(x) = 1 - \phi_1 x - \phi_2 x^2$ whose two roots are

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}.$$

The AR(2) process is stationary if and only if both roots are larger than 1 in absolute value (or in modulus if complex). That is, both roots must lie outside the unit circle.

- The same condition on the roots of $\phi(x)$ is needed for stationarity with any AR(p) process.

YULE-WALKER EQUATIONS: Assuming stationarity and zero means, consider the AR(p) process equation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

and multiply both sides by Y_{t-k} to get

$$Y_t Y_{t-k} = \phi_1 Y_{t-1} Y_{t-k} + \phi_2 Y_{t-2} Y_{t-k} + \cdots + \phi_p Y_{t-p} Y_{t-k} + e_t Y_{t-k}.$$

Taking expectations gives

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p}$$

and dividing through by the process variance γ_0, we get

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}.$$

Plugging in $k = 1, 2, ..., p$, and using the fact that $\rho_k = \rho_{-k}$, we get

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 + \cdots + \phi_p \rho_{p-1} \\
\rho_2 &= \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 + \cdots + \phi_p \rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \phi_3 \rho_{p-3} + \cdots + \phi_p.
\end{aligned}$$

These are the **Yule-Walker equations**. For known values of $\phi_1, \phi_2, ..., \phi_p$, we can compute the first lag p autocorrelations $\rho_1, \rho_2, ..., \rho_p$. Values of ρ_k, for $k > p$, can be obtained by using the recursive relation above. The AR(p) ACF tails off as k gets larger. It does so as a mixture of exponential decays and/or damped sine waves, depending on if roots are real or complex.

4.4 Invertibility

TERMINOLOGY: We define a process $\{Y_t\}$ to be **invertible** if it can written as a "mathematically meaningful" autoregressive process (possibly of infinite order). Invertibility is an important theoretical property. For prediction purposes, it is important to restrict our attention to the class of invertible models.

ILLUSTRATION: From the definition, we see that stationary autoregressive models are automatically invertible. However, moving average models may not be. For example, consider the MA(1) model

$$Y_t = e_t - \theta e_{t-1},$$

or, slightly rewritten,

$$e_t = Y_t + \theta e_{t-1}.$$

Note that we can write

$$e_t = Y_t + \theta \underbrace{(Y_{t-1} + \theta e_{t-2})}_{= e_{t-1}} = Y_t + \theta Y_{t-1} + \theta^2 e_{t-2}.$$

Repeated similar substitution reveals that

$$e_t = Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \cdots,$$

or slightly rewritten

$$Y_t = \underbrace{-\theta Y_{t-1} - \theta^2 Y_{t-2} - \theta^3 Y_{t-3} - \cdots + e_t}_{\text{"AR}(\infty)\text{"}}.$$

- For this autoregressive representation to be "mathematically meaningful," we need the infinite series of θ coefficients to be finite; that is, we need $\sum_{j=1}^{\infty} \theta^j < \infty$. This occurs if and only if $|\theta| < 1$.
- We have expressed an MA(1) as an infinite-order AR model. The MA(1) process is **invertible** if and only if $|\theta| < 1$.
- Compare this MA(1) "invertibility condition" with the stationarity condition of $|\phi| < 1$ for the AR(1) model.

IMPORTANCE: A model must be invertible for us to be able to identify the model parameters associated with it. For example, for an MA(1) model, it is straightforward to show that both of the following processes have the same autocorrelation function:

$$\begin{aligned} Y_t &= e_t - \theta e_{t-1} \\ Y_t &= e_t - \frac{1}{\theta} e_{t-1}. \end{aligned}$$

Put another way, if we knew the common ACF, we could not say if the MA(1) model parameter was θ or $1/\theta$. Thus, we impose the condition that $|\theta| < 1$ to ensure invertibility (identifiability). Note that under this condition, the second MA(1) model, rewritten

$$Y_t = -\left(\frac{1}{\theta}\right) Y_{t-1} - \left(\frac{1}{\theta}\right)^2 Y_{t-2} - \left(\frac{1}{\theta}\right)^3 Y_{t-3} - \cdots + e_t,$$

is no longer meaningful because the series $\sum_{j=1}^{\infty} \left(\frac{1}{\theta}\right)^j$ diverges.

NOTE: Rewriting the MA(1) model using backshift notation, we see that

$$Y_t = (1 - \theta B) e_t.$$

The function $\theta(x) = 1 - \theta x$ is called the **MA(1) characteristic polynomial** and

$$\theta(x) = 1 - \theta x = 0$$

is called the **MA(1) characteristic equation**. The root of this equation is

$$x = \frac{1}{\theta}.$$

For this process to be invertible, we require the root of the characteristic equation to exceed 1 (in absolute value). Doing so implies that $|\theta| < 1$.

GENERALIZATION: The MA(q) process

$$\begin{aligned} Y_t &= e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \\ &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) e_t \end{aligned}$$

is **invertible** if and only if the roots of the **MA**(q) **characteristic polynomial** $\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q$ all exceed 1 in absolute value (or modulus).

SUMMARY: We have discussed two important theoretical properties of autoregressive (AR) and moving average (MA) models, namely, **stationarity** and **invertibility**. Here is a summary of the important findings.

- For an AR(p) process to be **stationary**, we need the roots of the AR characteristic polynomial
$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$$
to all exceed 1 in absolute value (or modulus).

- For an MA(q) process to be **invertible**, we need the roots of the MA characteristic polynomial
$$\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q$$
to all exceed 1 in absolute value (or modulus).

- All invertible MA processes are stationary.

- All stationary AR processes are invertible.

- Any invertible MA(q) process corresponds to an infinite order AR process.

- Any stationary AR(p) process corresponds to an infinite order MA process.

4.5 Autoregressive moving average (ARMA) processes

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The process
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$
is an **autoregressive moving average process** of orders p and q, written **ARMA**(p, q). AR(p) and MA(q) processes are each special cases of the ARMA(p, q) process.

- An ARMA$(p, 0)$ process is the same as an AR(p) process.

- An ARMA$(0, q)$ process is the same as an MA(q) process.

REMARK: A stationary time series may often be adequately modeled by an ARMA model involving fewer parameters than a pure MA or AR process by itself. This is an example of the **Principle of Parsimony**; i.e., finding a model with as few parameters as possible, but which gives an adequate representation of the data.

BACKSHIFT NOTATION: The ARMA(p, q) process, expressed using backshift notation, is

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)Y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)e_t$$

or, more succinctly, as

$$\phi(B)Y_t = \theta(B)e_t,$$

where

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q. \end{aligned}$$

- For the ARMA(p, q) process to be **stationary**, we need the roots of the AR characteristic polynomial $\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$ to all exceed 1 in absolute value (or modulus).
- For the ARMA(p, q) process to be **invertible**, we need the roots of the MA characteristic polynomial $\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q$ to all exceed 1 in absolute value (or modulus).

Example 4.5. Write each of the models

(i) $Y_t = 0.3Y_{t-1} + e_t$

(ii) $Y_t = e_t - 1.3e_{t-1} + 0.4e_{t-2}$

(iii) $Y_t = 0.5Y_{t-1} + e_t - 0.3e_{t-1} + 1.2e_{t-2}$

(iv) $Y_t = 0.4Y_{t-1} + 0.45Y_{t-2} + e_t + e_{t-1} + 0.25e_{t-2}$

using backshift notation and determine whether the model is stationary and/or invertible.

SOLUTIONS.

(i) The model in (i) is an **AR(1)** with $\phi = 0.3$. In backshift notation, this model is $(1 - 0.3B)Y_t = e_t$. The characteristic polynomial is

$$\phi(x) = 1 - 0.3x,$$

which has the root $x = 10/3$. Because this root exceeds 1 in absolute value, this process is stationary. The process is also invertible since it is a stationary AR process.

(ii) The model in (ii) is an **MA(2)** with $\theta_1 = 1.3$ and $\theta_2 = -0.4$. In backshift notation, this model is $Y_t = (1 - 1.3B + 0.4B^2)e_t$. The characteristic polynomial is

$$\theta(x) = 1 - 1.3x + 0.4x^2,$$

which has roots $x = 2$ and $x = 1.25$. Because these roots both exceed 1 in absolute value, this process is invertible. The process is also stationary since it is an invertible MA process.

(iii) The model in (iii) is an **ARMA(1,2)** with $\phi_1 = 0.5$, $\theta_1 = 0.3$ and $\theta_2 = -1.2$. In backshift notation, this model is $(1 - 0.5B)Y_t = (1 - 0.3B + 1.2B^2)e_t$. The AR characteristic polynomial is

$$\phi(x) = 1 - 0.5x,$$

which has the root $x = 2$. Because this root is greater than 1, this process is stationary. The MA characteristic polynomial is

$$\theta(x) = 1 - 0.3x + 1.2x^2,$$

which has roots $x \approx 0.125 \pm 0.904i$. The modulus of each root is

$$|x| \approx \sqrt{(0.125)^2 + (0.904)^2} \approx 0.913,$$

which is less than 1. Therefore, this process is not invertible.

(iv) The model in (iv), at first glance, appears to be an ARMA(2,2) with $\phi_1 = 0.4$, $\phi_2 = 0.45$, $\theta_1 = -1$, and $\theta_2 = -0.25$. In backshift notation, this model is written as

$$(1 - 0.4B - 0.45B^2)Y_t = (1 + B + 0.25B^2)e_t.$$

However, the AR and MA characteristic polynomials in this instance factor as

$$(1 + 0.5B)(1 - 0.9B)Y_t = (1 + 0.5B)(1 + 0.5B)e_t.$$

In (mixed) ARMA models, the AR and MA characteristic polynomials can not share any common factors. Here, they do; namely, $(1 + 0.5B)$. Canceling, we have

$$(1 - 0.9B)Y_t = (1 + 0.5B)e_t,$$

which we identify as an **ARMA(1,1)** model with $\phi_1 = 0.9$ and $\theta_1 = -0.5$. This process is stationary since the root of $\phi(x) = 1 - 0.9x$ is $x = 10/9 > 1$. This process is invertible since the root of $\theta(x) = 1 + 0.5x$ is $x = -2$, which exceeds 1 in absolute value.

AUTOCORRELATION FUNCTION: Take the ARMA(p, q) model equation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

and multiply both sides by Y_{t-k} to get

$$\begin{aligned} Y_t Y_{t-k} &= \phi_1 Y_{t-1} Y_{t-k} + \phi_2 Y_{t-2} Y_{t-k} + \cdots + \phi_p Y_{t-p} Y_{t-k} \\ &\quad + e_t Y_{t-k} - \theta_1 e_{t-1} Y_{t-k} - \theta_2 e_{t-2} Y_{t-k} - \cdots - \theta_q e_{t-q} Y_{t-k}. \end{aligned}$$

For $k > q$, we have $E(e_t Y_{t-k}) = E(e_{t-1} Y_{t-k}) = \cdots = E(e_{t-q} Y_{t-k}) = 0$ so that

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p}.$$

Dividing through by the process variance γ_0, we get, for $k > q$,

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}.$$

Plugging in $k = 1, 2, ..., p$, and using the fact that $\rho_k = \rho_{-k}$, we arrive again at the **Yule Walker equations**:

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 + \cdots + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 + \cdots + \phi_p\rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \phi_3\rho_{p-3} + \cdots + \phi_p.
\end{aligned}$$

A similar system can be derived which involves $\theta_1, \theta_2, ..., \theta_q$.

- The R function `ARMAacf` can compute autocorrelations numerically for any stationary ARMA(p, q) process (including those that are purely AR or MA).
- The ACF for the ARMA(p, q) process tails off after lag q in a manner similar to the AR(p) process.
- However, unlike the AR(p) process, the first q autocorrelations depend on both $\theta_1, \theta_2, ..., \theta_q$ and $\phi_1, \phi_2, ..., \phi_p$.

SPECIAL CASE: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. The process

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$$

is called an **ARMA**$(1, 1)$ **process**. This is a special case of the ARMA(p, q) process with $p = q = 1$. In backshift notation, the process can be written as

$$(1 - \phi B)Y_t = (1 - \theta B)e_t$$

yielding $\phi(x) = 1 - \phi x$ and $\theta(x) = 1 - \theta x$ as the AR and MA characteristic polynomials, respectively. As usual, the conditions for stationarity and invertibility are that the roots of both polynomials exceed 1 in absolute value.

MOMENTS: The calculations on pp 78-79 (CC) show that

$$\gamma_0 = \left(\frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2}\right)\sigma_e^2,$$

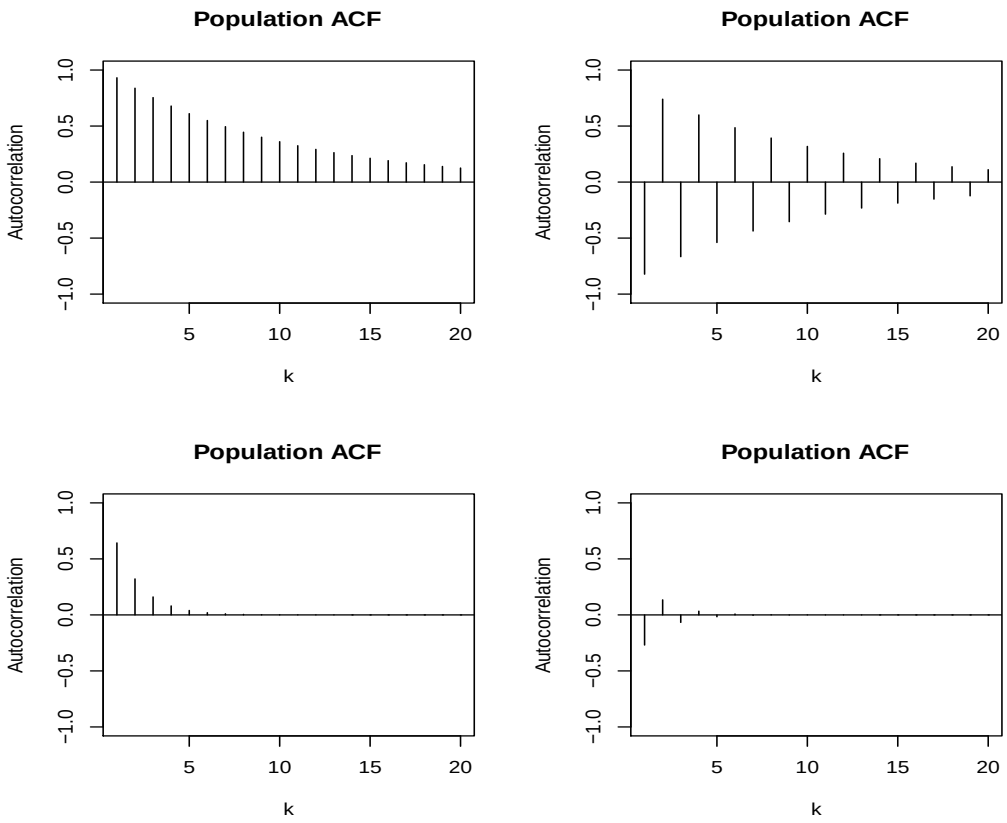


Figure 4.11: Population ACFs for ARMA(1,1) processes. Upper left: $(\phi, \theta) = (0.9, -0.25)$. Upper right: $(\phi, \theta) = (-0.9, -0.25)$. Lower left: $(\phi, \theta) = (0.5, -0.25)$. Lower right: $(\phi, \theta) = (-0.5, -0.25)$.

$\gamma_1 = \phi\gamma_0 - \theta\sigma_e^2$, and $\gamma_k = \phi\gamma_{k-1}$, for $k \geq 2$. The autocorrelation function is shown to satisfy

$$\rho_k = \left[\frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\right]\phi^{k-1}.$$

Note that when $k = 1$, ρ_1 is equal to a quantity that depends on ϕ and θ. This is different than the AR(1) process where ρ_1 depends on ϕ only. However, as k gets larger, the autocorrelation ρ_k decays in a manner similar to the AR(1) process. Figure 4.11 displays some different ARMA(1,1) ACFs.

REMARK: That the ARMA(1,1) model can be written in the general linear process form defined at the beginning of the chapter is shown on pp 78-79 (CC).

5 Models for Nonstationary Time Series

Complementary reading: Chapter 5 (CC).

5.1 Introduction

RECALL: Suppose $\{e_t\}$ is a zero mean white noise process with variance $\text{var}(e_t) = \sigma_e^2$. In the last chapter, we considered the class of ARMA models

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

or, expressed more succinctly,

$$\phi(B) Y_t = \theta(B) e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned} \phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q). \end{aligned}$$

- We learned that a process $\{Y_t\}$ in this class is **stationary** if and only if the roots of the AR characteristic polynomial $\phi(x)$ all exceed 1 in absolute value (or modulus).
- We learned that a process $\{Y_t\}$ in this class is **invertible** if and only if the roots of the MA characteristic polynomial $\theta(x)$ all exceed 1 in absolute value (or modulus).
- In this chapter, we extend this class of models to handle processes which are **nonstationary**. We accomplish this by generalizing the class of ARMA models to include differencing.
- Doing so gives rise to a much larger class of models, the **autoregressive integrated moving average** (**ARIMA**) class. This class incorporates a wide range of nonstationary time series processes.

TERMINOLOGY: Suppose that $\{Y_t\}$ is a stochastic process. The **first difference** process $\{\nabla Y_t\}$ consists of

$$\nabla Y_t = Y_t - Y_{t-1}.$$

The **second difference** process $\{\nabla^2 Y_t\}$ consists of

$$\begin{aligned}\nabla^2 Y_t = \nabla(\nabla Y_t) &= \nabla Y_t - \nabla Y_{t-1} \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2}.\end{aligned}$$

In general, the d**th difference** process $\{\nabla^d Y_t\}$ consists of

$$\nabla^d Y_t = \nabla(\nabla^{d-1} Y_t) = \nabla^{d-1} Y_t - \nabla^{d-1} Y_{t-1},$$

for $d = 1, 2, ...,$. We take $\nabla^0 Y_t = Y_t$ by convention.

Example 5.1. Suppose that $\{Y_t\}$ is a random walk process

$$Y_t = Y_{t-1} + e_t,$$

where $\{e_t\}$ is zero mean white noise with variance $\text{var}(e_t) = \sigma_e^2$. We know that $\{Y_t\}$ is not stationary because its autocovariance function depends on t (see Chapter 2). However, the first difference process

$$\nabla Y_t = Y_t - Y_{t-1} = e_t$$

is white noise, which is stationary.

- In Figure 5.1 (top), we display a simulated random walk process with $n = 150$ and $\sigma_e^2 = 1$. Note how the sample ACF of the series decays very, very slowly over time. *This is typical of a nonstationary series.*
- The first difference (white noise) process also appears in Figure 5.1 (bottom), along with its sample ACF. As we would expect from a white noise process, nearly all of the sample autocorrelations r_k are within the $\pm 2/\sqrt{n}$ bounds.
- As this simple example shows, it is possible to "transform" a nonstationary process into one that is stationary by taking differences.

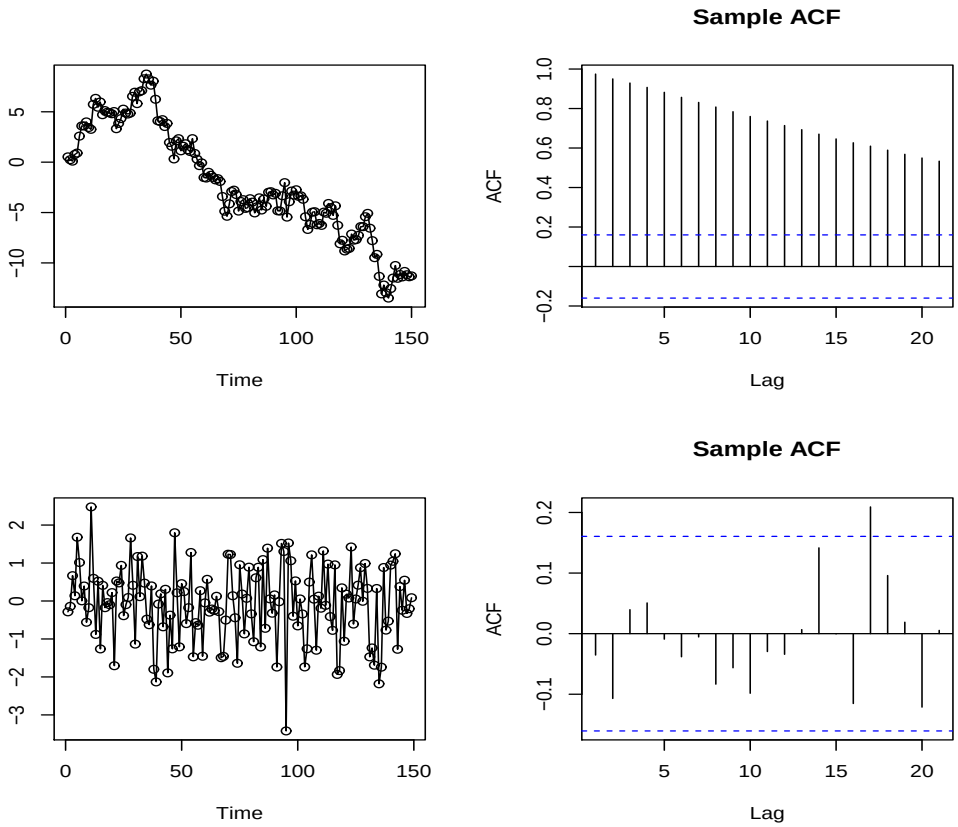


Figure 5.1: Top: A simulated random walk process $\{Y_t\}$ and its sample ACF, with $n = 150$ and $\sigma_e^2 = 1$. Bottom: The first difference process $\{\nabla Y_t\}$ and its sample ACF.

LINEAR TREND MODELS: In Chapter 3, we talked about how to use regression methods to fit models of the form

$$Y_t = \mu_t + X_t,$$

where μ_t is a deterministic trend function and where $\{X_t\}$ is a stochastic process with $E(X_t) = 0$. Suppose that $\{X_t\}$ is **stationary** and that the true trend function is

$$\mu_t = \beta_0 + \beta_1 t,$$

a **linear** function of time. Clearly, $\{Y_t\}$ is not a stationary process because

$$\begin{aligned} E(Y_t) &= E(\beta_0 + \beta_1 t + X_t) \\ &= \beta_0 + \beta_1 t + E(X_t) = \beta_0 + \beta_1 t, \end{aligned}$$

which depends on t. The first differences are given by

$$\nabla Y_t = Y_t - Y_{t-1} = (\beta_0 + \beta_1 t + X_t) - [\beta_0 + \beta_1(t-1) + X_{t-1}] = \beta_1 + X_t - X_{t-1}.$$

Note that

$$E(\nabla Y_t) = E(\beta_1 + X_t - X_{t-1}) = \beta_1 + E(X_t) - E(X_{t-1}) = \beta_1.$$

Also,

$$\begin{aligned}\text{cov}(\nabla Y_t, \nabla Y_{t-k}) &= \text{cov}(\beta_1 + X_t - X_{t-1}, \beta_1 + X_{t-k} - X_{t-k-1}) \\ &= \text{cov}(X_t, X_{t-k}) - \text{cov}(X_t, X_{t-k-1}) \\ &\quad - \text{cov}(X_{t-1}, X_{t-k}) + \text{cov}(X_{t-1}, X_{t-k-1}).\end{aligned}$$

Because $\{X_t\}$ is stationary, each of these covariance terms does not depend on t. Therefore, both $E(\nabla Y_t)$ and $\text{cov}(\nabla Y_t, \nabla Y_{t-k})$ are free of t; i.e., $\{\nabla Y_t\}$ is a stationary process. *Taking first differences removes a linear determinstic trend.*

QUADRATIC TRENDS: Suppose that the true deterministic trend model is

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2,$$

a **quadratic** function of time. Clearly, $\{Y_t\}$ is not a stationary process since $E(Y_t) = \mu_t$. The first difference process consists of

$$\begin{aligned}\nabla Y_t = Y_t - Y_{t-1} &= (\beta_0 + \beta_1 t + \beta_2 t^2 + X_t) - [\beta_0 + \beta_1(t-1) + \beta_2(t-1)^2 + X_{t-1}] \\ &= (\beta_1 - \beta_2) + 2\beta_2 t + X_t - X_{t-1}\end{aligned}$$

and $E(\nabla Y_t) = \beta_1 - \beta_2 + 2\beta_2 t$, which depends on t. Therefore, $\{\nabla Y_t\}$ is not a stationary process. The second difference process consists of

$$\begin{aligned}\nabla^2 Y_t &= \nabla Y_t - \nabla Y_{t-1} \\ &= [(\beta_1 - \beta_2) + 2\beta_2 t + X_t - X_{t-1}] - [(\beta_1 - \beta_2) + 2\beta_2(t-1) + X_{t-1} - X_{t-2}] \\ &= 2\beta_2 + X_t - 2X_{t-1} + X_{t-2}.\end{aligned}$$

Therefore, $E(\nabla^2 Y_t) = 2\beta_2$ and $\text{cov}(\nabla^2 Y_t, \nabla^2 Y_{t-k})$ are free of t. This shows that $\{\nabla^2 Y_t\}$ is stationary. *Taking second differences removes a quadratic deterministic trend.*

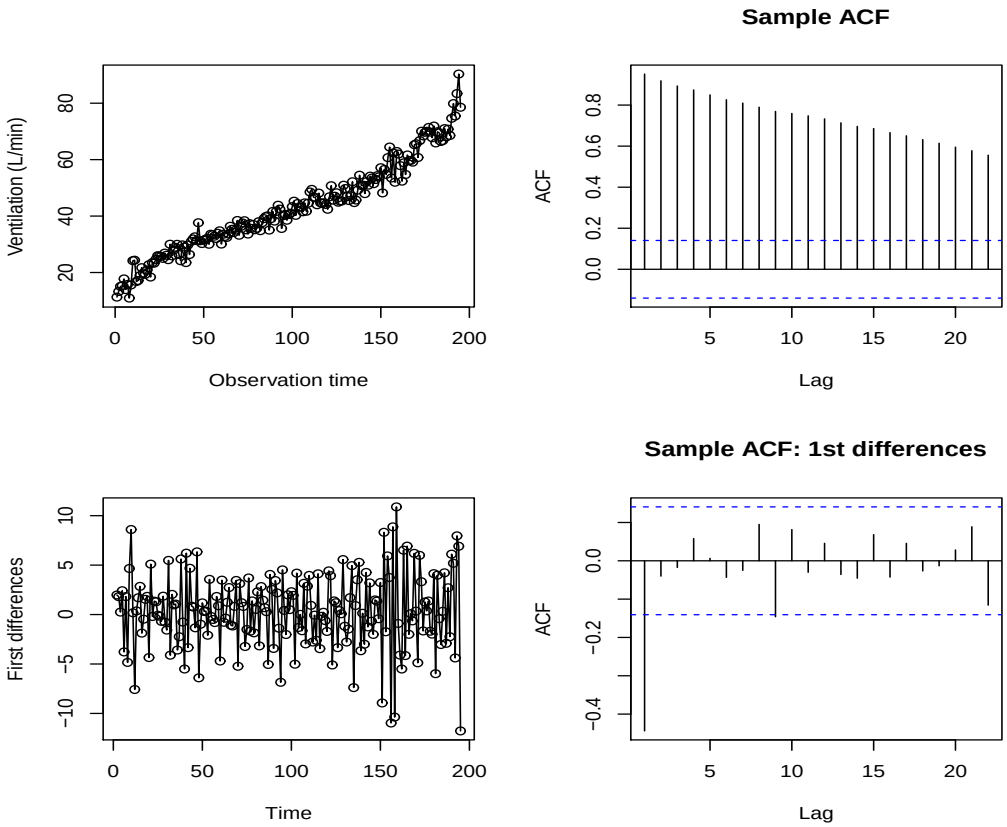


Figure 5.2: Ventilation measurements at 15 second intervals. Top: Ventilation series $\{Y_t\}$ with sample ACF. Bottom: First difference process $\{\nabla Y_t\}$ with sample ACF.

GENERALIZATION: Suppose that $Y_t = \mu_t + X_t$, where μ_t is a deterministic trend function and $\{X_t\}$ is a stationary process with $E(X_t) = 0$. In general, if

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d$$

is a polynomial in t of degree d, then the dth difference process $\{\nabla^d Y_t\}$ is stationary.

Example 5.2. The data in Figure 5.2 are ventilation observations (L/min) on a single cyclist recorded every 15 seconds during exercise. **Source:** Joe Alemany (Spring, 2010).

- The ventilation time series $\{Y_t\}$ does not resemble a stationary process. There is a pronounced increasing **linear trend** over time. Nonstationarity is also reinforced by examining the sample ACF for the series. In particular, the sample ACF decays very, very slowly (a sure sign of nonstationarity).

- The first difference series $\{\nabla Y_t\}$ does resemble a process with a constant mean. In fact, the sample ACF for $\{\nabla Y_t\}$ looks like what we would expect from an MA(1) process (i.e., a pronounced spike at $k = 1$ and little action elsewhere).
- To summarize, the evidence in Figure 5.2 suggests an MA(1) model for the difference process $\{\nabla Y_t\}$.

5.2 Autoregressive integrated moving average (ARIMA) models

TERMINOLOGY: A stochastic process $\{Y_t\}$ is said to follow an **autoregressive integrated moving average** (**ARIMA**) model if the dth differences $W_t = \nabla^d Y_t$ follow a stationary ARMA model. There are three important values which characterize an ARIMA process:

- p, the order of the autoregressive component
- d, the number of differences needed to arrive at a stationary ARMA(p, q) process
- q, the order of the moving average component.

In particular, we have the general relationship:

$$Y_t \text{ is ARIMA}(p, d, q) \iff W_t = \nabla^d Y_t \text{ is ARMA}(p, q).$$

RECALL: A stationary ARMA(p, q) process can be represented as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) e_t$$

or, more succinctly, as

$$\phi(B) Y_t = \theta(B) e_t,$$

where $\{e_t\}$ is zero mean white noise with variance $\text{var}(e_t) = \sigma_e^2$. In the ARIMA$(p, d, q)$ family, take $d = 1$ so that

$$W_t = \nabla Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B) Y_t$$

follows an ARMA(p, q) model. Therefore, an ARIMA$(p, 1, q)$ process can be written succinctly as

$$\phi(B)(1-B)Y_t = \theta(B)e_t.$$

Similarly, take $d = 2$ so that

$$\begin{aligned} W_t = \nabla^2 Y_t &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= Y_t - 2BY_t + B^2 Y_t \\ &= (1 - 2B + B^2)Y_t = (1-B)^2 Y_t \end{aligned}$$

follows an ARMA(p, q) model. Therefore, an ARIMA$(p, 2, q)$ process can be written as

$$\phi(B)(1-B)^2 Y_t = \theta(B)e_t.$$

In general, an **ARIMA**(p, d, q) **process** can be written as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t.$$

IMPORTANT: In practice (with real data), there will rarely be a need to consider values of the differencing order $d > 2$. Most real time series data can be coerced into a stationarity ARMA process by taking one difference or occasionally two differences (perhaps after transforming the series initially).

REMARK: Autoregressive (AR) models, moving average (MA) models, and autoregressive moving average (ARMA) models are all members of the ARIMA(p, d, q) family. In particular,

- AR$(p) \longleftrightarrow$ ARIMA$(p, 0, 0)$
- MA$(q) \longleftrightarrow$ ARIMA$(0, 0, q)$
- ARMA$(p, q) \longleftrightarrow$ ARIMA$(p, 0, q)$
- ARI$(p, d) \longleftrightarrow$ ARIMA$(p, d, 0)$
- IMA$(d, q) \longleftrightarrow$ ARIMA$(0, d, q)$.

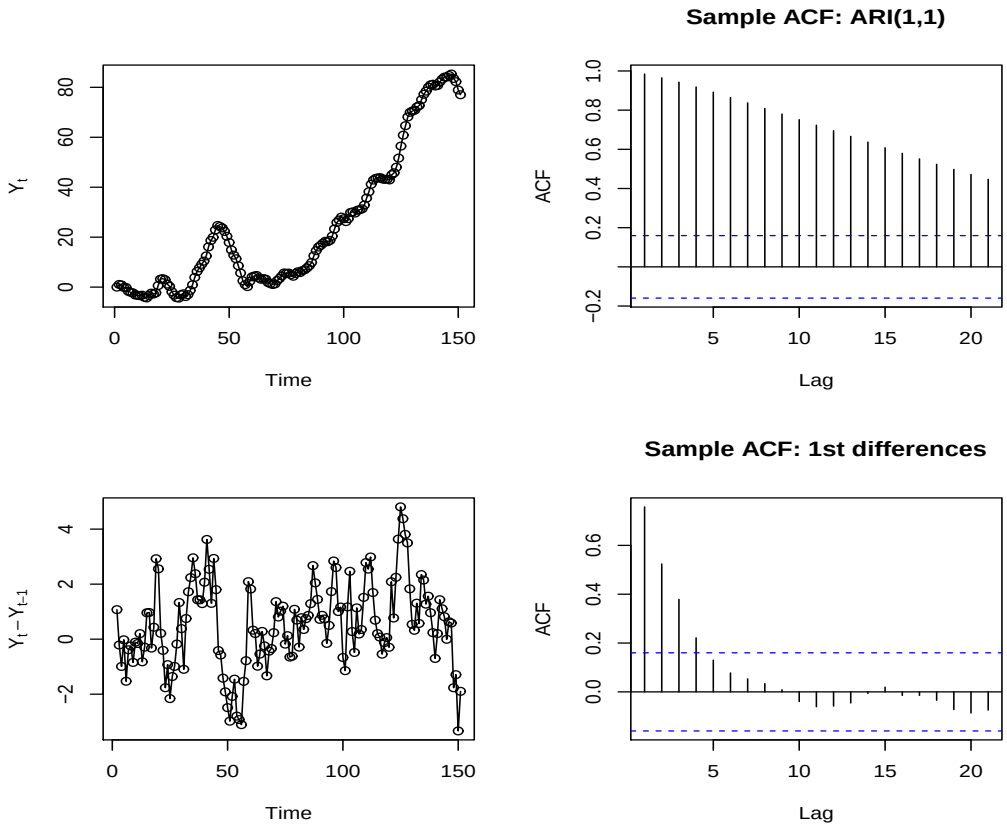


Figure 5.3: Top: ARI(1,1) simulation, with $\phi = 0.7$, $n = 150$, and $\sigma_e^2 = 1$, and the sample ACF. Bottom: First difference process with sample ACF.

Example 5.3. Suppose $\{e_t\}$ is a zero mean white noise process. Identify each model

(a) $Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + e_t$

(b) $Y_t = 1.5Y_{t-1} - 0.5Y_{t-2} + e_t - e_{t-1} + 0.25e_{t-2}$

as an ARIMA(p, d, q) process. That is, specify the values of p, d, and q.

SOLUTIONS.

(a) Upon first glance,

$$Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + e_t$$

looks like an AR(2) process with $\phi_1 = 1.7$ and $\phi_2 = -0.7$. However, upon closer inspection, we see this process is not stationary because the AR(2) stationary con-

ditions

$$\phi_1 + \phi_2 < 1 \qquad \phi_2 - \phi_1 < 1 \qquad |\phi_2| < 1$$

are not met with $\phi_1 = 1.7$ and $\phi_2 = -0.7$ (in particular, the first condition is not met). However, note that we can write this process as

$$\begin{aligned} Y_t - 1.7Y_{t-1} + 0.7Y_{t-2} = e_t &\iff Y_t - 1.7BY_t + 0.7B^2Y_t = e_t \\ &\iff (1 - 1.7B + 0.7B^2)Y_t = e_t \\ &\iff (1 - 0.7B)(1 - B)Y_t = e_t \\ &\iff (1 - 0.7B)W_t = e_t, \end{aligned}$$

where

$$W_t = (1 - B)Y_t = Y_t - Y_{t-1}$$

are the first differences. We identify $\{W_t\}$ as a stationary AR(1) process with $\phi = 0.7$. Therefore, $\{Y_t\}$ is an ARIMA(1,1,0) $\iff$ ARI(1,1) process with $\phi = 0.7$. This ARI(1,1) process is simulated in Figure 5.3.

(b) Upon first glance,

$$Y_t = 1.5Y_{t-1} - 0.5Y_{t-2} + e_t - e_{t-1} + 0.25e_{t-2}$$

looks like an ARMA(2,2) process, but this process is not stationary either. To see why, note that we can write this process as

$$\begin{aligned} & Y_t - 1.5Y_{t-1} + 0.5Y_{t-2} = e_t - e_{t-1} + 0.25e_{t-2} \\ \iff\ & (1 - 1.5B + 0.5B^2)Y_t = (1 - B + 0.25B^2)e_t \\ \iff\ & (1 - 0.5B)(1 - B)Y_t = (1 - 0.5B)^2 e_t \\ \iff\ & (1 - B)Y_t = (1 - 0.5B)e_t \\ \iff\ & W_t = (1 - 0.5B)e_t, \end{aligned}$$

where $W_t = (1-B)Y_t = Y_t - Y_{t-1}$. Here, the first differences $\{W_t\}$ follow an MA(1) model with $\theta = 0.5$. Therefore, $\{Y_t\}$ is an ARIMA(0,1,1) $\iff$ IMA(1,1) process with $\theta = 0.5$. A realization of this IMA(1,1) process is shown in Figure 5.4.

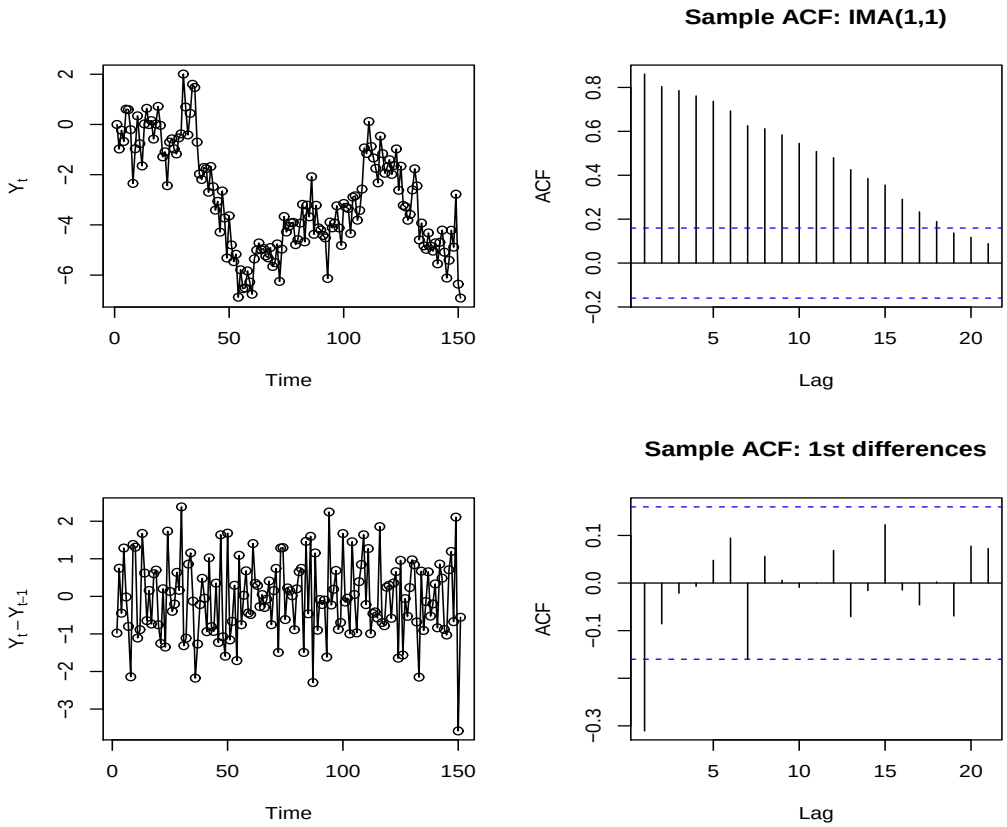


Figure 5.4: Top: IMA(1,1) simulation, with $\theta = 0.5$, $n = 150$, and $\sigma_e^2 = 1$, and the sample ACF. Bottom: First difference process with sample ACF.

5.2.1 IMA(1,1) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. An ARIMA$(p, d, q)$ process with $p = 0$, $d = 1$, and $q = 1$ is called an **IMA(1,1) process** and is given by

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1}.$$

This model is very popular in economics applications. Note that if $\theta = 0$, the IMA(1,1) process reduces to a random walk.

REMARK: We first note that an IMA(1,1) process can be written as

$$(1 - B)Y_t = (1 - \theta B)e_t.$$

If we (mistakenly) treated this as an ARMA(1,1) process with characteristic operators

$$\begin{aligned}\phi(B) &= 1 - B \\ \theta(B) &= 1 - \theta B,\end{aligned}$$

it would be clear that this process is not stationary since the AR characteristic polynomial $\phi(x) = 1 - x$ has a **unit root**, that is, the root of $\phi(x)$ is $x = 1$. More appropriately, we write

$$(1 - B)Y_t = (1 - \theta B)e_t \quad \Longleftrightarrow \quad W_t = (1 - \theta B)e_t,$$

and note that the first differences

$$W_t = (1 - B)Y_t = Y_t - Y_{t-1}$$

follow an MA(1) model with parameter θ. From Chapter 4, we know that the first difference process $\{W_t\}$ is invertible if and only if $|\theta| < 1$. To summarize,

$$\{Y_t\} \text{ follows an IMA(1,1)} \quad \Longleftrightarrow \quad \{W_t\} \text{ follows an MA(1).}$$

5.2.2 IMA(2,2) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. An ARIMA$(p, d, q)$ process with $p = 0$, $d = 2$, and $q = 2$ is called an **IMA(2,2) process** and can be expressed as

$$(1 - B)^2 Y_t = (1 - \theta_1 B - \theta_2 B^2)e_t,$$

or, equivalently,

$$\nabla^2 Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}.$$

In an IMA(2,2) process, the second differences

$$W_t = \nabla^2 Y_t = (1 - B)^2 Y_t$$

follow an MA(2) model. Invertibility is assessed by examining the MA characteristic operator $\theta(B) = 1 - \theta_1 B - \theta_2 B^2$. An IMA(2,2) process is simulated in Figure 5.5.

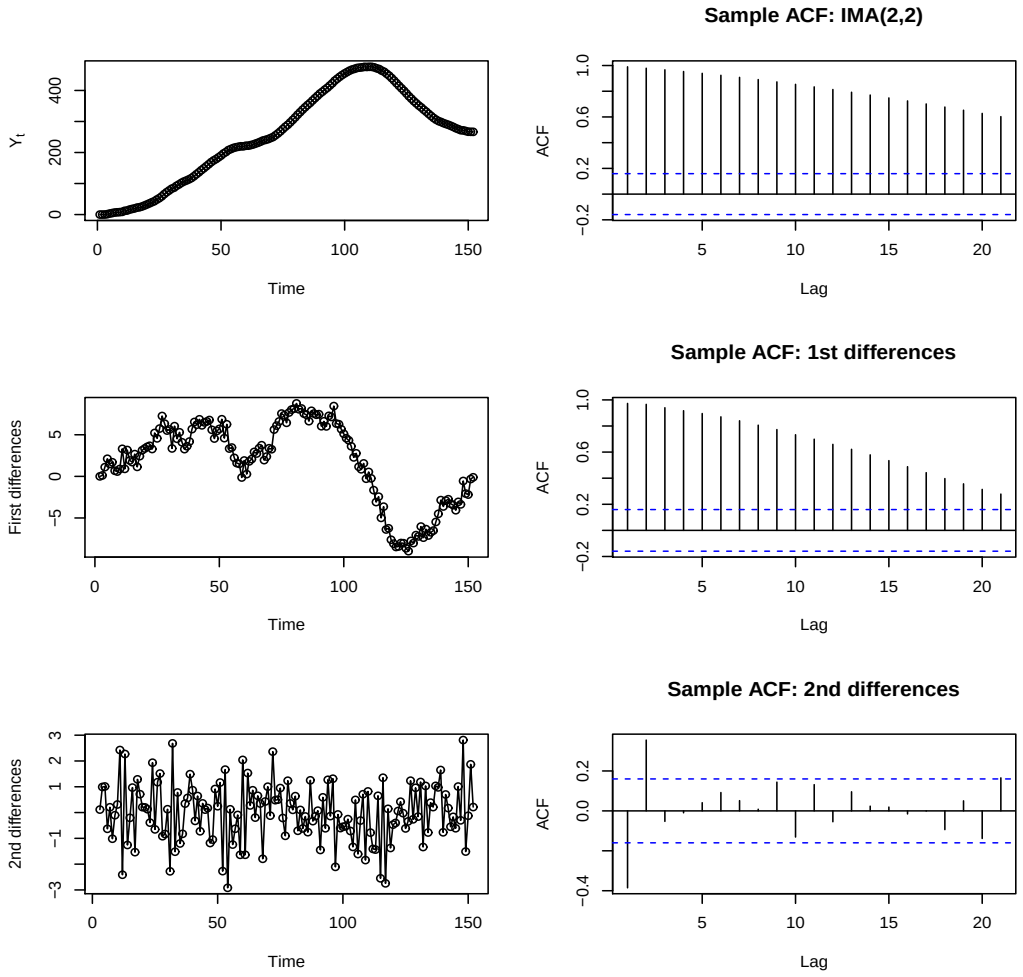


Figure 5.5: Top: IMA(2,2) simulation with $n = 150$, $\theta_1 = 0.3$, $\theta_2 = -0.3$, and $\sigma_e^2 = 1$. Middle: First difference process. Bottom: Second difference process.

- The defining characteristic of an IMA(2,2) process is its very strong autocorrelation at all lags. This is also seen in the sample ACF.
- The **first** difference process $\{\nabla Y_t\}$, which is that of an IMA(1,2), is also clearly nonstationary to the naked eye. This is also seen in the sample ACF.
- The **second** difference process $\{\nabla^2 Y_t\}$ is an (invertible) MA(2) process. This is suggested in the sample ACF for the second differences. Note how there are clear spikes in the ACF at lags $k = 1$ and $k = 2$.

5.2.3 ARI(1,1) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. An ARIMA$(p, d, q)$ process with $p = 1$, $d = 1$, and $q = 0$ is called an **ARI(1,1) process** and can be expressed as

$$(1 - \phi B)(1 - B)Y_t = e_t,$$

or, equivalently,

$$Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + e_t.$$

Note that the first differences $W_t = (1 - B)Y_t$ satisfy the model

$$(1 - \phi B)W_t = e_t,$$

which we recognize as an AR(1) process with parameter ϕ. The first difference process $\{W_t\}$ is stationary if and only if $|\phi| < 1$.

REMARK: Upon first glance, the process

$$Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + e_t$$

looks like an AR(2) model. However this process is not stationary since the coefficients satisfy $(1 + \phi) - \phi = 1$; this violates the stationarity requirements for the AR(2) model. An ARI(1,1) process is simulated in Figure 5.3.

5.2.4 ARIMA(1,1,1) process

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. An ARIMA$(p, d, q)$ process with $p = 1$, $d = 1$, and $q = 1$ is called an **ARIMA(1,1,1) process** and can be expressed as

$$(1 - \phi B)(1 - B)Y_t = (1 - \theta B)e_t,$$

or, equivalently,

$$Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + e_t - \theta e_{t-1}.$$

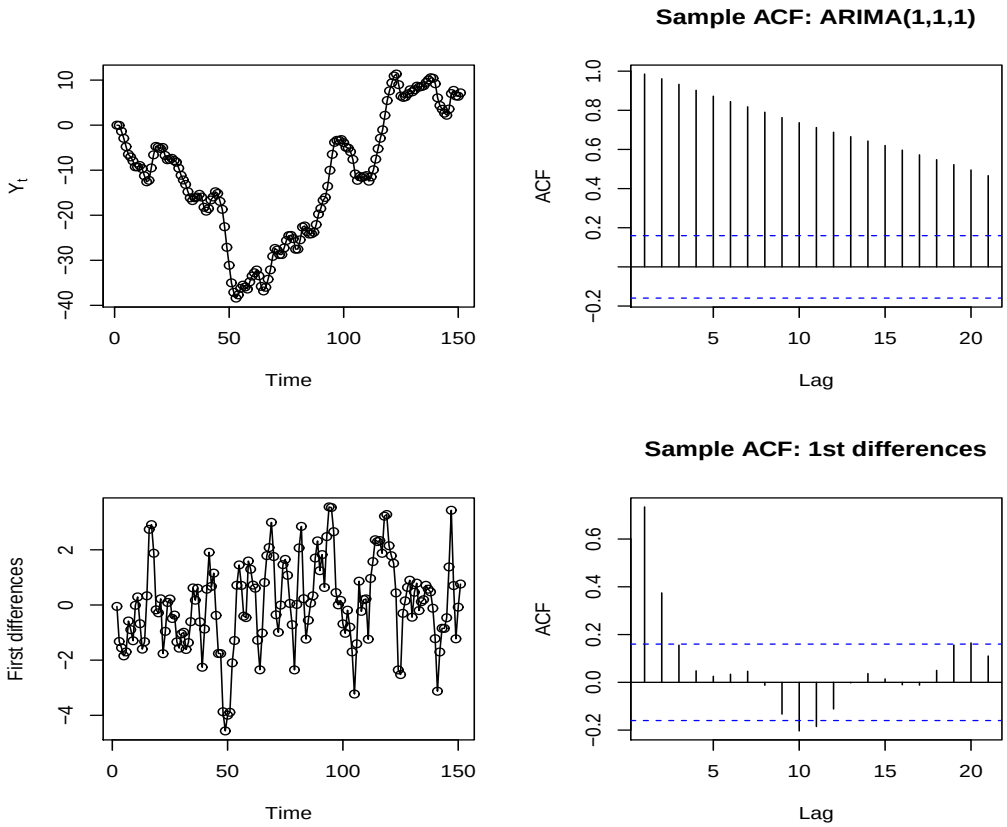


Figure 5.6: Top: ARIMA(1,1,1) simulation, with $n = 150$, $\phi = 0.5$, $\theta = -0.5$, and $\sigma_e^2 = 1$, and the sample ACF. Bottom: First difference process with sample ACF.

Note that the first differences $W_t = (1 - B)Y_t$ satisfy the model

$$(1 - \phi B)W_t = (1 - \theta B)e_t,$$

which we recognize as an ARMA(1,1) process with parameters ϕ and θ.

- The first difference process $\{W_t\}$ is **stationary** if and only if $|\phi| < 1$. The first difference process $\{W_t\}$ is **invertible** if and only if $|\theta| < 1$.
- A simulated ARIMA(1,1,1) process appears in Figure 5.6. The ARIMA(1,1,1) simulated series Y_t is clearly nonstationary. The first difference series $W_t = \nabla Y_t$ appears to have a constant mean, and its sample ACF resembles that of a stationary ARMA(1,1) process (as it should).

5.3 Constant terms in ARIMA models

RECALL: An ARIMA(p, d, q) process can be written as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t,$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. An extension of this model is

$$\phi(B)(1-B)^d Y_t = \theta_0 + \theta(B)e_t,$$

where the parameter θ_0 is a **constant term**.

IMPORTANT: The parameter θ_0 plays very different roles when

- $d = 0$ (a stationary ARMA model)
- $d > 0$ (a nonstationary model).

STATIONARY CASE: Suppose that $d = 0$, in which case the no-constant model becomes

$$\phi(B)Y_t = \theta(B)e_t,$$

a stationary ARMA process, where the AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q).\end{aligned}$$

To examine the effects of adding a constant term, suppose that we replace Y_t with $Y_t - \mu$, where $\mu = E(Y_t)$. The model becomes

$$\begin{aligned}\phi(B)(Y_t - \mu) = \theta(B)e_t &\implies \phi(B)Y_t - \phi(B)\mu = \theta(B)e_t \\ &\implies \phi(B)Y_t - (1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu = \theta(B)e_t \\ &\implies \phi(B)Y_t = \underbrace{(1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu}_{=\,\theta_0} + \theta(B)e_t,\end{aligned}$$

so that

$$\theta_0 = (1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu \iff \mu = \frac{\theta_0}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}.$$

IMPORTANT: In a stationary ARMA process $\{Y_t\}$, adding a constant term θ_0 to the model does not affect the stationarity properties of $\{Y_t\}$.

NONSTATIONARY CASE: The impact of adding a constant term θ_0 to the model when $d > 0$ is quite different. As the simplest example in the ARIMA(p, d, q) family, take $p = q = 0$ and $d = 1$ so that

$$(1 - B)Y_t = \theta_0 + e_t \quad \Longleftrightarrow \quad Y_t = \theta_0 + Y_{t-1} + e_t.$$

This model is called a **random walk with drift**; see pp 22 (CC). Note that we can write via successive substitution

$$\begin{aligned} Y_t &= \theta_0 + Y_{t-1} + e_t \\ &= \theta_0 + \underbrace{\theta_0 + Y_{t-2} + e_{t-1}}_{= Y_{t-1}} + e_t \\ &= 2\theta_0 + Y_{t-2} + e_t + e_{t-1} \\ &\vdots \\ &= (t-k)\theta_0 + Y_k + e_t + e_{t-1} + \cdots + e_{t-k+1}. \end{aligned}$$

Therefore, the process $\{Y_t\}$ contains a **linear deterministic trend** with slope θ_0.

IMPORTANT: The previous finding holds for any (nonstationary) ARIMA$(p, 1, q)$ model, that is, adding a constant term θ_0 induces a **linear** deterministic trend. Also,

- adding a constant term θ_0 to an ARIMA$(p, 2, q)$ model induces a **quadratic** deterministic trend,
- adding a constant term θ_0 to an ARIMA$(p, 3, q)$ model induces a **cubic** deterministic trend, and so on.

Note that for very large t, the constant (deterministic trend) term can become very dominating so that it forces the time series to follow a nearly deterministic pattern. Therefore, a constant term should be added to a nonstationary ARIMA model (i.e., $d > 0$) only if it is strongly warranted.

5.4 Transformations

REVIEW: If we are trying to model a nonstationary time series, it may be helpful to transform the data first before we examine any data differences (or before "detrending" the data if we use regression methods from Chapter 3).

- For example, if there is clear evidence of nonconstant variance over time (e.g., the variance increases over time, etc.), then a suitable **transformation** to the data might remove (or lessen the impact of) the nonconstant variance pattern.
- Applying a transformation to address nonconstant variance is regarded as a "first step." This is done before using differencing as a means to achieve stationarity.

Example 5.4. Data file: `electricity` (TSA). Figure 5.7 displays monthly electricity usage in the United States (usage from coal, natural gas, nuclear, petroleum, and wind) between January, 1973 and December, 2005.

- From the plot, we can see that there is increasing variance over time; e.g., the series is much more variable at later years than it is in earlier years.
- Time series that exhibit this "fanning out" shape are not stationary because the variance changes over time.
- Before we try to model these data, we should first apply a transformation to make the variance constant (that is, we would like to first "stabilize" the variance).

THEORY: Suppose that the variance of nonstationary process $\{Y_t\}$ can be written as

$$\text{var}(Y_t) = c_0 f(\mu_t),$$

where $\mu_t = E(Y_t)$ and c_0 is a positive constant free of μ_t. Therefore, the variance is not constant because it is a function of μ_t, which is changing over time. Our goal is to find a function T so that the transformed series $T(Y_t)$ has constant variance. Such a function is

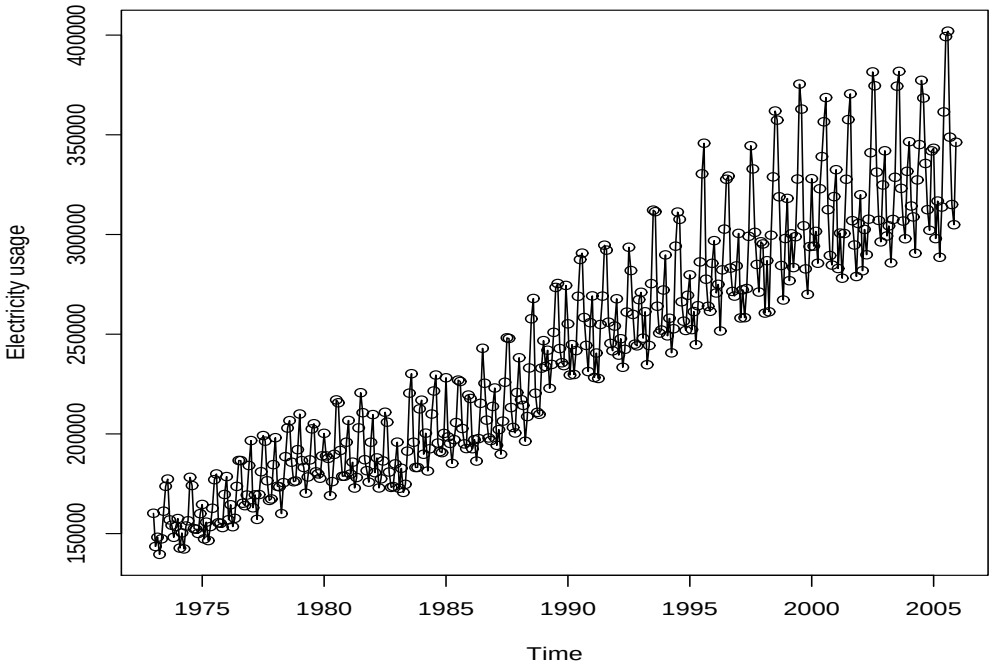


Figure 5.7: Electricity data. Monthly U.S. electricity generation, measured in millions of kilowatt hours, from 1/1973 to 12/2005.

called a **variance stabilizing transformation** function. Consider approximating the function T by a first-order Taylor-series expansion about the point μ_t, that is,

$$T(Y_t) \approx T(\mu_t) + T'(\mu_t)(Y_t - \mu_t),$$

where $T'(\mu_t)$ is the first derivative of $T(Y_t)$, evaluated at μ_t. Now, note that

$$\begin{aligned} \text{var}[T(Y_t)] &\approx \text{var}[T(\mu_t) + T'(\mu_t)(Y_t - \mu_t)] \\ &= c_0[T'(\mu_t)]^2 f(\mu_t). \end{aligned}$$

Therefore, we want to find the function T which satisfies

$$\text{var}[T(Y_t)] \approx c_0[T'(\mu_t)]^2 f(\mu_t) \stackrel{\text{set}}{=} c_1,$$

where c_1 is a constant free of μ_t. Solving this expression for $T'(\mu_t)$, we get the differential equation

$$T'(\mu_t) = \sqrt{\frac{c_1}{c_0 f(\mu_t)}} = \frac{c_2}{\sqrt{f(\mu_t)}},$$

where $c_2 = \sqrt{c_1/c_0}$ is free of μ_t. Integrating both sides, we get

$$T(\mu_t) = \int \frac{c_2}{\sqrt{f(\mu_t)}}\, d\mu_t + c_3,$$

where c_3 is a constant free of μ_t. In the calculations below, the values of c_2 and c_3 can be taken to be anything, as long as they are free of μ_t.

- If $\text{var}(Y_t) = c_0\mu_t$, so that the variance of the series is proportional to the mean, then
$$T(\mu_t) = \int \frac{c_2}{\sqrt{\mu_t}}\, d\mu_t = 2c_2\sqrt{\mu_t} + c_3,$$
where c_3 is a constant free of μ_t. If we take $c_2 = 1/2$ and $c_3 = 0$, we see that the **square root** of the series, $T(Y_t) = \sqrt{Y_t}$, will provide a constant variance.

- If $\text{var}(Y_t) = c_0\mu_t^2$, so that the standard deviation of the series is proportional to the mean, then
$$T(\mu_t) = \int \frac{c_2}{\sqrt{\mu_t^2}}\, d\mu_t = c_2 \ln(\mu_t) + c_3,$$
where c_3 is a constant free of μ_t. If we take $c_2 = 1$ and $c_3 = 0$, we see that the **logarithm** of the series, $T(Y_t) = \ln(Y_t)$, will provide a constant variance.

- If $\text{var}(Y_t) = c_0\mu_t^4$, so that the standard deviation of the series is proportional to the square of the mean, then
$$T(\mu_t) = \int \frac{c_2}{\sqrt{\mu_t^4}}\, d\mu_t = c_2\left(-\frac{1}{\mu_t}\right) + c_3,$$
where c_3 is a constant free of μ_t. If we take $c_2 = -1$ and $c_3 = 0$, we see that the **reciprocal** of the series, $T(Y_t) = 1/Y_t$, will provide a constant variance.

BOX-COX TRANSFORMATIONS: More generally, we can use a **power transformation** introduced by Box and Cox (1964). The transformation is defined by

$$T(Y_t) = \begin{cases} \dfrac{Y_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Y_t), & \lambda = 0, \end{cases}$$

Table 5.1: Box-Cox transformation parameters λ and their associated transformations.

λ	$T(Y_t)$	Description
-2.0	$1/Y_t^2$	Inverse square
-1.0	$1/Y_t$	Reciprocal
-0.5	$1/\sqrt{Y_t}$	Inverse square root
0.0	$\ln(Y_t)$	Logarithm
0.5	$\sqrt{Y_t}$	Square root
1.0	Y_t	Identity (no transformation)
2.0	Y_t^2	Square

where λ is called the **transformation parameter**. Some common values of λ, and their implied transformations are given in Table 5.1.

NOTE: To see why the logarithm transformation $T(Y_t) = \ln(Y_t)$ is used when $\lambda = 0$, note that by L'Höptial's Rule (from calculus),

$$\lim_{\lambda \to 0} \frac{Y_t^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{Y_t^\lambda \ln(Y_t)}{1} = \ln(Y_t).$$

- A variance stabilizing transformation can only be performed on a **positive** series, that is, when $Y_t > 0$, for all t. This turns out not to be prohibitive, because if some or all of the series Y_t is negative, we can simply add (the same) positive constant c to each observation, where c is chosen so that everything becomes positive. Adding c will not affect the (non)stationarity properties of $\{Y_t\}$.

- Remember, a variance stabilizing transformation, if needed, should be performed **before** taking any data differences.

- Frequently, a transformation performed to stabilize the variance will also improve an approximation of normality. We will discuss the normality assumption later (Chapters 7-8) when we address issues in statistical inference.

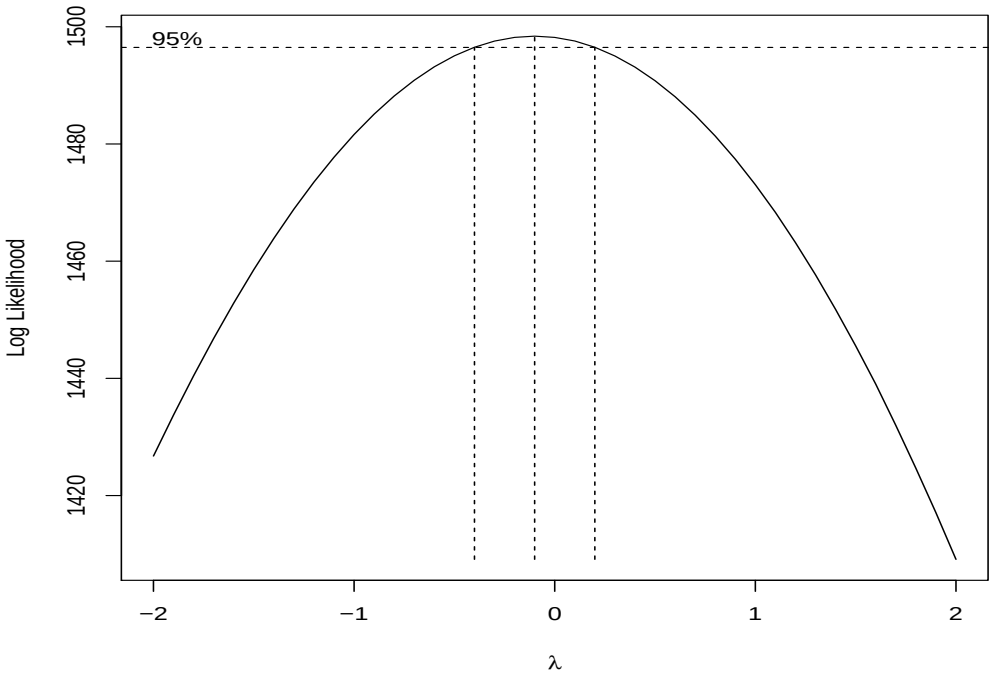


Figure 5.8: Electricity data. Log-likelihood function versus λ. Note that λ is on the horizontal axis. A 95 percent confidence interval for λ is also depicted.

DETERMINING λ: We can let the data "suggest" a suitable transformation in the Box-Cox power family.

- We do this by treating λ as a parameter, writing the log-likelihood function of the data (under the normality assumption), and finding the value of λ which maximizes the log-likelihood function; i.e., the maximum likelihood estimate (MLE) of λ.
- There is an R function `BoxCox.ar` that does all of the calculations. The function also provides an approximate 95 percent confidence interval for λ, which is constructed using the large sample properties of MLEs.
- The computations needed to produce a figure like the one in Figure 5.8 can be time consuming if the series is long (i.e., n is large). Also, the profile log-likelihood is not always as "smooth" as that seen in Figure 5.8.

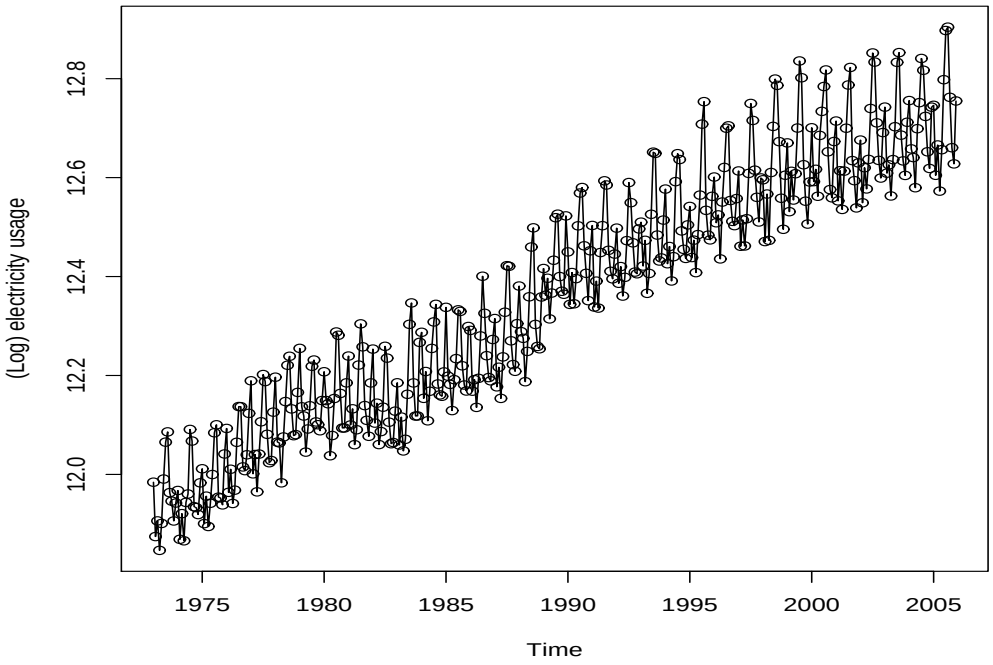


Figure 5.9: Electricity data (transformed). Monthly U.S. electricity generation measured on the log scale.

Example 5.4 (continued). Figure 5.8 displays the profile log-likelihood of λ for the electricity data. The value of λ (on the horizontal axis) that maximizes the log-likelihood function looks to be $\lambda \approx -0.1$, suggesting the transformation

$$T(Y_t) = Y_t^{-0.1}.$$

However, this transformation makes little practical sense. An approximate 95 percent confidence interval for λ looks to be about $(-0.4, 0.2)$. Because $\lambda = 0$ is in this interval, a log transformation $T(Y_t) = \ln(Y_t)$ is not unreasonable.

- The log-transformed series $\{\ln Y_t\}$ is displayed in Figure 5.9. We see that applying the log transformation has notably lessened the nonconstant variance (although there still is a mild increase in the variance over time).
- Now that we have applied the transformation, we can now return to our previous

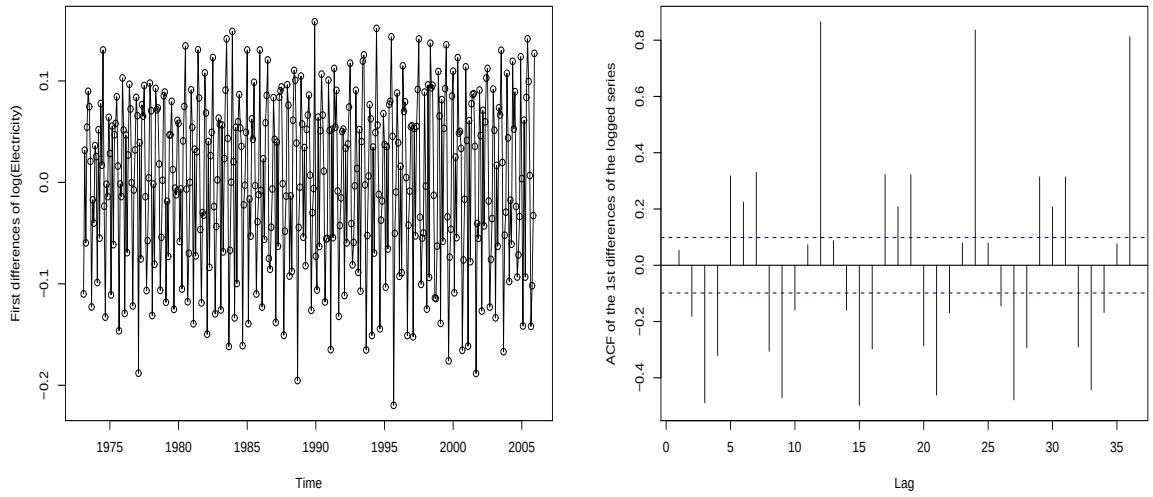


Figure 5.10: Electricity data. Left: $W_t = \log Y_t - \log Y_{t-1}$, the first differences of the log-transformed data. Right: The sample autocorrelation function of the $\{W_t\}$ data.

modeling techniques. For the log-transformed series, there is still a pronounced linear trend over time. Therefore, we consider the first difference process (on the log scale), given by

$$W_t = \log Y_t - \log Y_{t-1} = \nabla \log Y_t.$$

- The $\{W_t\}$ series is plotted in Figure 5.10 (left) along with the sample ACF of the $\{W_t\}$ series (right). The $\{W_t\}$ series appears to have a constant mean.
- However, the sample ACF suggests that there is still a large amount of structure in the data that remains after differencing the log-transformed series.
- In particular, there looks to be significant autocorrelations that arise according to a seasonal pattern. We will consider seasonal processes that model this type of variability in Chapter 10.

REMARK: Taking the differences of a log-transformed series, as we have done in this example, often arises in financial applications where Y_t (e.g., stock price, portfolio return, etc.) tends to have stable **percentage changes** over time. See pp 99 (CC).

6 Model Specification

Complementary reading: Chapter 6 (CC).

6.1 Introduction

RECALL: Suppose that $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In general, an ARIMA$(p, d, q)$ process can be written as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\end{aligned}$$

and

$$(1-B)^d Y_t = \nabla^d Y_t$$

is the series of dth differences. In this chapter, we discuss techniques on how to choose suitable values of p, d, and q for an observed (or transformed) time series. We want our choices to be consistent with the underlying structure of the observed data. Bad choices of p, d, and q lead to bad models, which, in turn, lead to bad predictions (forecasts) of future values.

6.2 The sample autocorrelation function

RECALL: For time series data $Y_1, Y_2, ..., Y_n$, the **sample autocorrelation function** (ACF), at lag k, is given by

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

where $\overline{Y}$ is the sample mean of $Y_1, Y_2, ..., Y_n$.

IMPORTANT: The sample autocorrelation r_k is an **estimate** of the true (population) autocorrelation ρ_k. As with any statistic, r_k has a **sampling distribution** which describes how it varies from sample to sample. We would like to know this distribution so we can quantify the uncertainty in values of r_k that we might see in practice.

THEORY: For a stationary ARMA(p, q) process,

$$\sqrt{n}(r_k - \rho_k) \xrightarrow{d} \mathcal{N}(0, c_{kk}),$$

as $n \to \infty$, where

$$c_{kk} = \sum_{l=-\infty}^{\infty} (\rho_l^2 + \rho_{l-k}\rho_{l+k} - 4\rho_k\rho_l\rho_{l-k} + 2\rho_k^2\rho_l^2).$$

In other words, when the sample size n is large, the sample autocorrelation r_k is **approximately normally distributed** with mean ρ_k and variance c_{kk}/n; i.e.,

$$r_k \sim \mathcal{AN}\left(\rho_k, \frac{c_{kk}}{n}\right).$$

We now examine some specific models and specialize this general result to those models.

1. **WHITE NOISE**: For a white noise process, the formula for c_{kk} simplifies considerably because nearly all the terms in the sum above are zero. For large n,

   $$r_k \sim \mathcal{AN}\left(0, \frac{1}{n}\right),$$

   for $k = 1, 2, ...,$. This explains why $\pm 2/\sqrt{n}$ serve as approximate margin of error bounds for r_k. Values of r_k outside these bounds would be "unusual" under the white noise model assumption.

2. **AR(1)**: For a stationary AR(1) process $Y_t = \phi Y_{t-1} + e_t$, the formula for c_{kk} also reduces considerably. For large n,

   $$r_k \sim \mathcal{AN}(\rho_k, \sigma^2_{r_k}),$$

   where $\rho_k = \phi^k$ and

   $$\sigma^2_{r_k} = \frac{1}{n}\left[\frac{(1+\phi^2)(1-\phi^{2k})}{1-\phi^2} - 2k\phi^{2k}\right].$$

3. **MA(1)**: For an invertible MA(1) process $Y_t = e_t - \theta e_{t-1}$, we treat the $k = 1$ and $k > 1$ cases separately.

   - **Case 1:** Lag $k = 1$. For large n,

   $$r_1 \sim \mathcal{AN}(\rho_1, \sigma^2_{r_1}),$$

   where $\rho_1 = -\theta/(1+\theta^2)$ and

   $$\sigma^2_{r_1} = \frac{1 - 3\rho_1^2 + 4\rho_1^4}{n}.$$

   - **Case 2:** Lag $k > 1$. For large n,

   $$r_k \sim \mathcal{AN}(0, \sigma^2_{r_k}),$$

   where

   $$\sigma^2_{r_k} = \frac{1 + 2\rho_1^2}{n}.$$

4. **MA**(q): For an invertible MA(q) process,

   $$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

   the sample autocorrelation r_k, **for all** $k > q$**,** satisfies

   $$r_k \sim \mathcal{AN}\left[0, \frac{1}{n}\left(1 + 2\sum_{j=1}^{q} \rho_j^2\right)\right],$$

   when n is large.

REMARK: The MA(q) result above suggests a natural **large-sample** test for

$$H_0 : \text{MA}(q) \text{ process is appropriate}$$

versus

$$H_1 : \text{MA}(q) \text{ process is not appropriate.}$$

If H_0 is true, then the sample autocorrelation

$$r_{q+1} \sim \mathcal{AN}\left[0, \frac{1}{n}\left(1 + 2\sum_{j=1}^{q} \rho_j^2\right)\right].$$

Therefore, the random variable

$$Z = \frac{r_{q+1}}{\sqrt{\frac{1}{n}\left(1 + 2\sum_{j=1}^{q}\rho_j^2\right)}} \sim \mathcal{AN}(0,1).$$

We can not use Z as a test statistic to test H_0 versus H_1 because Z depends on $\rho_1, \rho_2, ..., \rho_q$ which, in practice, are unknown. However, when n is large, we can use r_j as an estimate for ρ_j. This should not severely impact the large sample distribution of Z because r_j should be "close" to ρ_j when n is large. Making this substitution gives the large-sample **test statistic**

$$Z^* = \frac{r_{q+1}}{\sqrt{\frac{1}{n}\left(1 + 2\sum_{j=1}^{q} r_j^2\right)}}.$$

When H_0 is true, $Z^* \sim \mathcal{AN}(0,1)$. Therefore, a level α decision rule is to reject H_0 in favor of H_1 when

$$|Z^*| > z_{\alpha/2},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the $\mathcal{N}(0,1)$ distribution. This is a **two-sided** test. Of course, an equivalent decision rule is to reject H_0 when the (two-sided) probability value is less than α.

Example 6.1. From a time series of $n = 200$ observations, we calculate $r_1 = -0.49$, $r_2 = 0.31$, $r_3 = -0.13$, $r_4 = 0.07$, and $|r_k| < 0.09$ for $k > 4$. Which moving average (MA) model is most consistent with these sample autocorrelations?

SOLUTION. To test

$$H_0 : \text{MA(1) process is appropriate}$$

versus

$$H_1 : \text{MA(1) process is not appropriate}$$

we compute

$$z^* = \frac{r_2}{\sqrt{\frac{1}{n}\left(1 + 2r_1^2\right)}} = \frac{0.31}{\sqrt{\frac{1}{200}\left[1 + 2(-0.49)^2\right]}} \approx 3.60.$$

This is not a reasonable value of Z^* under H_0; e.g., the p-value is

$$\text{pr}(|Z^*| > 3.60) \approx 0.0003.$$

Therefore, we would reject H_0 and conclude that the MA(1) model is not appropriate.

To test

$$H_0 : \text{MA(2) process is appropriate}$$

versus

$$H_1 : \text{MA(2) process is not appropriate}$$

we compute

$$z^* = \frac{r_3}{\sqrt{\frac{1}{n}(1 + 2r_1^2 + 2r_2^2)}} = \frac{-0.13}{\sqrt{\frac{1}{200}[1 + 2(-0.49)^2 + 2(0.31)^2]}} \approx -1.42.$$

This is not an unreasonable value of Z^* under H_0; e.g., the p-value is

$$\text{pr}(|Z^*| > 1.42) \approx 0.16.$$

Therefore, we would not reject H_0. An MA(2) model is not inconsistent with these sample autocorrelations.

Example 6.2. *Monte Carlo simulation.* Consider the model

$$Y_t = e_t + 0.7e_{t-1},$$

an MA(1) process with $\theta = -0.7$, where $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 200$. In this example, we use a technique known as **Monte Carlo simulation** to simulate the sampling distributions of the sample autocorrelations r_1, r_2, r_5, and r_{10}. Here is how this is done:

- We simulate an MA(1) process with $\theta = -0.7$ and compute r_1 with the simulated data. Note that the R function `arima.sim` can be used to simulate this process.
- We repeat this simulation exercise a large number of times, say, M times. With each simulated series, we compute r_1.
- If we simulate M different series, we will have M corresponding values of r_1.
- We can then plot the M values of r_1 in a histogram. This histogram represents the **Monte Carlo sampling distribution** of r_1.
- For each simulation, we can also record the values of r_2, r_5, and r_{10}. We can then construct their corresponding histograms.

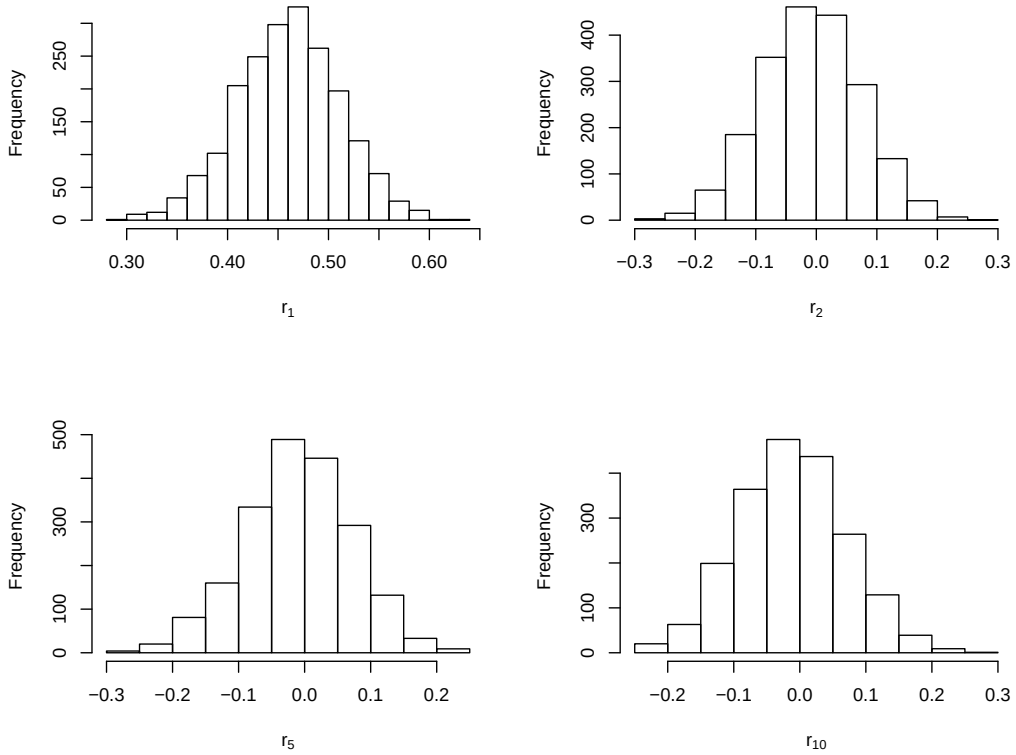


Figure 6.1: Monte Carlo simulation. Histograms of sample autocorrelations based on $M = 2000$ Monte Carlo samples of size $n = 200$ taken from an MA(1) process with $\theta = -0.7$. Upper left: r_1. Upper right: r_2. Lower left: r_5. Lower right: r_{10}. The histograms are approximations to the true sampling distributions when $n = 200$.

- Note that the approximate sampling distribution of r_1 is centered around

$$\rho_1 = \frac{-(-0.7)}{1 + (-0.7)^2} \approx 0.47.$$

  The other sampling distributions are centered around $\rho_2 = 0$, $\rho_5 = 0$, and $\rho_{10} = 0$, as expected. All distributions take on a normal shape, also as expected.

- IMPORTANT: The true large-sample distribution result

$$\sqrt{n}(r_k - \rho_k) \xrightarrow{d} \mathcal{N}(0, c_{kk})$$

  is a result that requires the sample size $n \to \infty$. With $n = 200$, we see that the normal distribution (large-sample) property has largely taken shape.

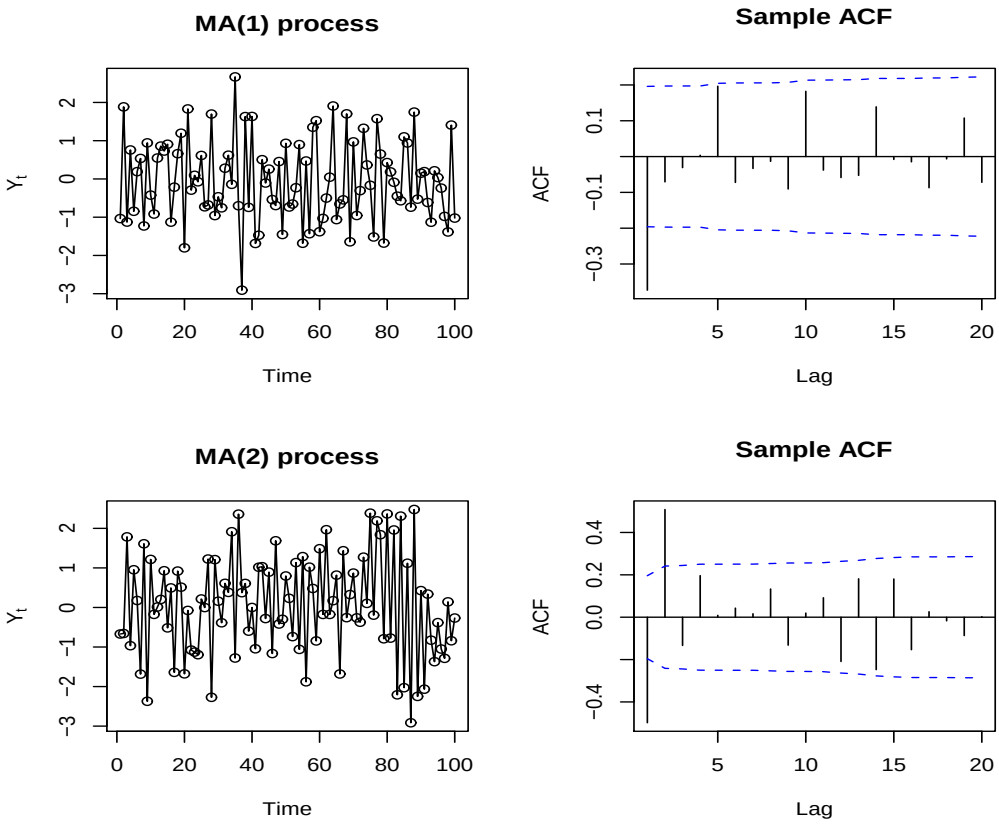


Figure 6.2: Simulated MA(1) and MA(2) processes with $n = 100$ and $\sigma_e^2 = 1$. Moving average error bounds are used in the corresponding sample ACFs; not the white noise error bounds $\pm 2/\sqrt{n}$.

Example 6.3. We use R to generate data from two moving average processes:

1. $Y_t = e_t - 0.5e_{t-1} \iff \mathbf{MA(1)}$, with $\theta = 0.5$
2. $Y_t = e_t - 0.5e_{t-1} + 0.5e_{t-2} \iff \mathbf{MA(2)}$, with $\theta_1 = 0.5$ and $\theta_2 = -0.5$.

We take $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 100$. In Figure 6.2, we display the realized time series and the corresponding sample autocorrelation functions (ACFs).

- However, instead of using the white noise margin of error bounds, that is,

$$\pm\frac{2}{\sqrt{n}} = \pm\frac{2}{\sqrt{100}} = \pm 0.2,$$

we use the more precise error bounds from the large sample distribution

$$r_k \sim \mathcal{AN}\left[0, \frac{1}{n}\left(1 + 2\sum_{j=1}^{q}\rho_j^2\right)\right].$$

- In particular, for each lag k, the (estimated) standard error bounds are placed at

$$\pm 1.96\sqrt{\frac{1}{100}\left(1 + 2\sum_{j=1}^{k-1} r_j^2\right)}.$$

- That is, error bounds at lag k are computed assuming that the MA$(k-1)$ model is appropriate. Values of r_k which exceed these bounds are deemed to be statistically significant. Note that the MA error bounds are not constant, unlike those computed under the white noise assumption.

6.3 The partial autocorrelation function

RECALL: We have seen that for MA(q) models, the population ACF ρ_k is nonzero for lags $k \leq q$ and $\rho_k = 0$ for lags greater than q. That is, the ACF for an MA(q) process "drops off" to zero after lag q.

- Therefore, the ACF provides a considerable amount of information about the order of the dependence when the process is truly a moving average.
- On the other hand, if the process is autoregressive (AR), then the ACF may not tell us much about the order of the dependence.
- It is therefore worthwhile to develop a function that will behave like the ACF for MA models, but for use with AR models instead. This function is called the **partial autocorrelation function** (**PACF**).

MOTIVATION: To set our ideas, consider a stationary, zero mean **AR(1)** process

$$Y_t = \phi Y_{t-1} + e_t,$$

where $\{e_t\}$ is zero mean white noise. The autocovariance between Y_t and Y_{t-2} is

$$\begin{aligned}
\gamma_2 &= \text{cov}(Y_t, Y_{t-2}) \\
&= \text{cov}(\phi Y_{t-1} + e_t, Y_{t-2}) \\
&= \text{cov}[\phi(\phi Y_{t-2} + e_{t-1}) + e_t, Y_{t-2}] \\
&= \text{cov}(\phi^2 Y_{t-2} + \phi e_{t-1} + e_t, Y_{t-2}) \\
&= \phi^2 \text{cov}(Y_{t-2}, Y_{t-2}) + \phi \text{cov}(e_{t-1}, Y_{t-2}) + \text{cov}(e_t, Y_{t-2}) \\
&= \phi^2 \text{var}(Y_{t-2}) + 0 + 0 = \phi^2 \gamma_0,
\end{aligned}$$

where $\gamma_0 = \text{var}(Y_t) = \text{var}(Y_{t-2})$. Recall that e_{t-1} and e_t are independent of Y_{t-2}.

- Note that if Y_t followed an MA(1) process, then $\gamma_2 = \text{cov}(Y_t, Y_{t-2}) = 0$.
- This not true for an AR(1) process because Y_t depends on Y_{t-2} through Y_{t-1}.

STRATEGY: Suppose that we "break" the dependence between Y_t and Y_{t-2} in an AR(1) process by removing (or partialing out) the effect of Y_{t-1}. To do this, consider the quantities $Y_t - \phi Y_{t-1}$ and $Y_{t-2} - \phi Y_{t-1}$. Note that

$$\text{cov}(Y_t - \phi Y_{t-1}, Y_{t-2} - \phi Y_{t-1}) = \text{cov}(e_t, Y_{t-2} - \phi Y_{t-1}) = 0,$$

because e_t is independent of Y_{t-1} and Y_{t-2}. Now, we make the following observations.

- In the AR(1) model, if ϕ is known, we can think of

  $$Y_t - \phi Y_{t-1}$$

  as the **prediction error** from regressing Y_t on Y_{t-1} (with no intercept; this is not needed because we are assuming a zero mean process).

- Similarly, the quantity

  $$Y_{t-2} - \phi Y_{t-1}$$

  can be thought of as the prediction error from regressing Y_{t-2} on Y_{t-1}, again with no intercept.

- Both of these prediction errors are **uncorrelated** with the intervening variable Y_{t-1}. To see why, note that

$$\begin{aligned}\text{cov}(Y_t - \phi Y_{t-1}, Y_{t-1}) &= \text{cov}(Y_t, Y_{t-1}) - \phi\text{cov}(Y_{t-1}, Y_{t-1}) \\ &= \gamma_1 - \phi\gamma_0 = 0,\end{aligned}$$

because $\gamma_1 = \phi\gamma_0$ in the AR(1) model. An identical argument shows that

$$\text{cov}(Y_{t-2} - \phi Y_{t-1}, Y_{t-1}) = \gamma_1 - \phi\gamma_0 = 0.$$

AR(2): Consider a stationary, zero mean AR(2) process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t,$$

where $\{e_t\}$ is zero mean white noise. Suppose that we "break" the dependence between Y_t and Y_{t-3} in the AR(2) process by removing the effects of **both** Y_{t-1} and Y_{t-2}. That is, consider the quantities

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}$$

and

$$Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}.$$

Note that

$$\text{cov}(Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}, Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}) = \text{cov}(e_t, Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}) = 0,$$

because e_t is independent of Y_{t-1}, Y_{t-2}, and Y_{t-3}. Again, we note the following:

- In the AR(2) case, if ϕ_1 and ϕ_2 are known, then the quantity

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}$$

can be thought of as the **prediction error** from regressing Y_t on Y_{t-1} and Y_{t-2} (with no intercept).

- Similarly, the quantity

$$Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}$$

can be thought of as the prediction error from regressing Y_{t-3} on Y_{t-1} and Y_{t-2}, again with no intercept.

- Both of these prediction errors are **uncorrelated** with the intervening variables Y_{t-1} and Y_{t-2}.

TERMINOLOGY: For a zero mean time series, let $\widehat{Y}_t^{(k-1)}$ denote the population regression of Y_t on the variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}$, that is,

$$\widehat{Y}_t^{(k-1)} = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_{k-1} Y_{t-(k-1)}.$$

Let $\widehat{Y}_{t-k}^{(k-1)}$ denote the population regression of Y_{t-k} on the variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}$, that is,

$$\widehat{Y}_{t-k}^{(k-1)} = \beta_1 Y_{t-(k-1)} + \beta_2 Y_{t-(k-2)} + \cdots + \beta_{k-1} Y_{t-1}.$$

The **partial autocorrelation function** (**PACF**) of a stationary process $\{Y_t\}$, denoted by ϕ_{kk}, satisfies $\phi_{11} = \rho_1$ and

$$\phi_{kk} = \text{corr}(Y_t - \widehat{Y}_t^{(k-1)}, Y_{t-k} - \widehat{Y}_{t-k}^{(k-1)}),$$

for $k = 2, 3, ...,$.

- With regards to Y_t and Y_{t-k}, the quantities $\widehat{Y}_t^{(k-1)}$ and $\widehat{Y}_{t-k}^{(k-1)}$ are linear functions of the **intervening variables** $Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}$.
- The quantities $Y_t - \widehat{Y}_t^{(k-1)}$ and $Y_{t-k} - \widehat{Y}_{t-k}^{(k-1)}$ are called the **prediction errors**. The PACF at lag k is defined to be the correlation between these errors.
- If the underlying process $\{Y_t\}$ is normal, then an equivalent definition is

  $$\phi_{kk} = \text{corr}(Y_t, Y_{t-k} | Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}),$$

  the correlation between Y_t and Y_{t-k}, **conditional** on the intervening variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}$.
- That is, ϕ_{kk} measures the correlation between Y_t and Y_{t-k} after removing the linear effects of $Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}$.

RECALL: We now revisit our AR(1) calculations. Consider the model

$$Y_t = \phi Y_{t-1} + e_t.$$

We showed that

$$\text{cov}(Y_t - \phi Y_{t-1}, Y_{t-2} - \phi Y_{t-1}) = \text{cov}(e_t, Y_{t-2} - \phi Y_{t-1}) = 0.$$

In this example, the quantities $Y_t - \phi Y_{t-1}$ and $Y_{t-2} - \phi Y_{t-1}$ are the prediction errors from regressing Y_t on Y_{t-1} and Y_{t-2} on Y_{t-1}, respectively. That is, with $k = 2$, the general expressions

$$\begin{aligned} \widehat{Y}_t^{(k-1)} &= \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_{k-1} Y_{t-(k-1)} \\ \widehat{Y}_{t-k}^{(k-1)} &= \beta_1 Y_{t-(k-1)} + \beta_2 Y_{t-(k-2)} + \cdots + \beta_{k-1} Y_{t-1} \end{aligned}$$

become

$$\begin{aligned} \widehat{Y}_t^{(2-1)} &= \phi Y_{t-1} \\ \widehat{Y}_{t-2}^{(2-1)} &= \phi Y_{t-1}. \end{aligned}$$

Therefore, we have shown that for the AR(1) model,

$$\phi_{22} = \text{corr}(Y_t - \widehat{Y}_t^{(2-1)}, Y_{t-2} - \widehat{Y}_{t-2}^{(2-1)}) = 0$$

because

$$\text{cov}(Y_t - \widehat{Y}_t^{(2-1)}, Y_{t-2} - \widehat{Y}_{t-2}^{(2-1)}) = \text{cov}(Y_t - \phi Y_{t-1}, Y_{t-2} - \phi Y_{t-1}) = 0.$$

IMPORTANT: **For the AR(1) model**, it follows that $\phi_{11} \neq 0$ ($\phi_{11} = \rho_1$) and

$$\phi_{22} = \phi_{33} = \phi_{44} = \cdots = 0.$$

That is, $\phi_{kk} = 0$, for all $k > 1$.

RECALL: We now revisit our AR(2) calculations. Consider the model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t.$$

We showed that

$$\text{cov}(Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}, Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}) = 0.$$

Note that in this example, the quantities $Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}$ and $Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}$ are the prediction errors from regressing Y_t on Y_{t-1} and Y_{t-2} and Y_{t-3} on Y_{t-1} and Y_{t-2}, respectively. That is, with $k = 3$, the general expressions

$$\begin{aligned} \widehat{Y}_t^{(k-1)} &= \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_{k-1} Y_{t-(k-1)} \\ \widehat{Y}_{t-k}^{(k-1)} &= \beta_1 Y_{t-(k-1)} + \beta_2 Y_{t-(k-2)} + \cdots + \beta_{k-1} Y_{t-1} \end{aligned}$$

become

$$\begin{aligned} \widehat{Y}_t^{(3-1)} &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} \\ \widehat{Y}_{t-3}^{(3-1)} &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2}. \end{aligned}$$

Therefore, we have shown that for the AR(2) model,

$$\phi_{33} = \text{corr}(Y_t - \widehat{Y}_t^{(3-1)}, Y_{t-3} - \widehat{Y}_{t-3}^{(3-1)}) = 0$$

because

$$\text{cov}(Y_t - \widehat{Y}_t^{(3-1)}, Y_{t-3} - \widehat{Y}_{t-3}^{(3-1)}) = \text{cov}(Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}, Y_{t-3} - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}) = 0.$$

IMPORTANT: **For the AR(2) model**, it follows that $\phi_{11} \neq 0$, $\phi_{22} \neq 0$, and

$$\phi_{33} = \phi_{44} = \phi_{55} = \cdots = 0.$$

That is, $\phi_{kk} = 0$, for all $k > 2$.

GENERAL RESULT: For an **AR**(p) **process**, we have the following results:

- $\phi_{11} \neq 0$, $\phi_{22} \neq 0$, ..., $\phi_{pp} \neq 0$; i.e., the first p partial autocorrelations are nonzero
- $\phi_{kk} = 0$, for all $k > p$.

For an AR(p) model, the PACF "drops off" to zero after the pth lag. Therefore, the PACF can help to determine the order of an AR(p) process just like the ACF helps to determine the order of an MA(q) process!

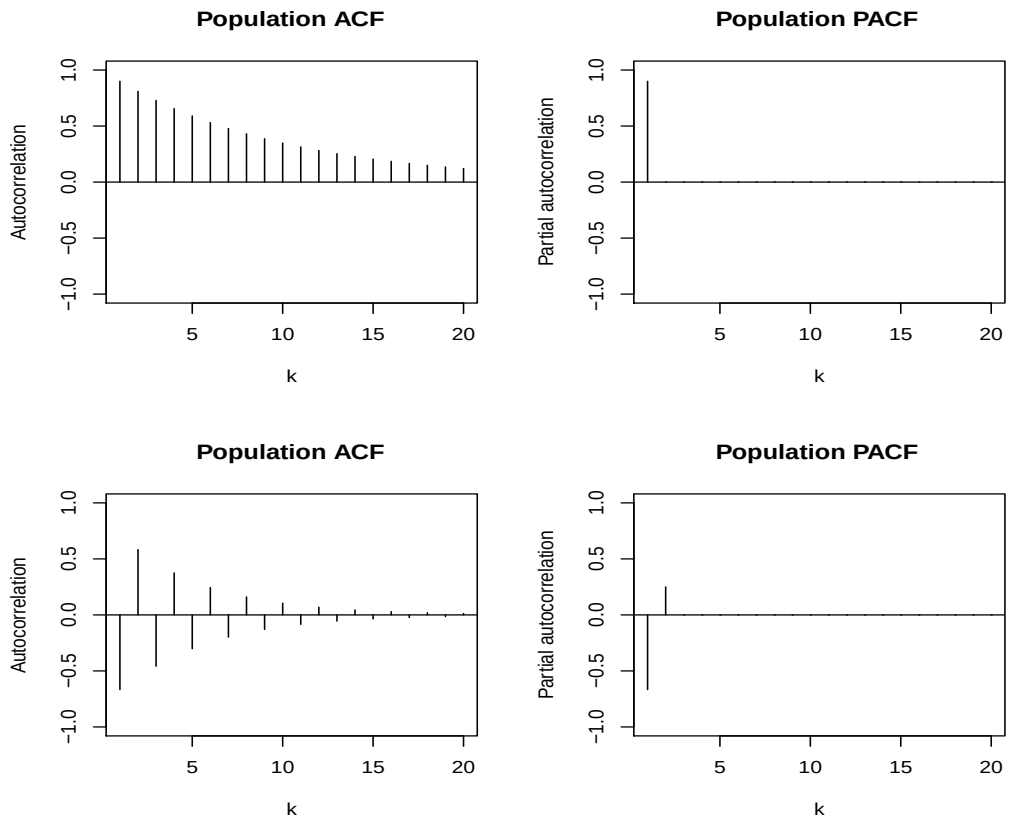


Figure 6.3: Top: AR(1) model with $\phi = 0.9$; population ACF (left) and population PACF (right). Bottom: AR(2) model with $\phi_1 = -0.5$ and $\phi_2 = 0.25$; population ACF (left) and population PACF (right).

Example 6.4. We use R to generate observations from two autoregressive processes:

(i) $Y_t = 0.9Y_{t-1} + e_t \iff$ **AR(1)**, with $\phi = 0.9$

(ii) $Y_t = -0.5Y_{t-1} + 0.25Y_{t-2} + e_t \iff$ **AR(2)**, with $\phi_1 = -0.5$ and $\phi_2 = 0.25$.

We take $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 150$. Figure 6.3 displays the true (**population**) ACF and PACF for these processes. Figure 6.4 displays the simulated time series from each AR model and the **sample** ACF/PACF.

- The population PACFs in Figure 6.3 display the characteristics that we have just derived; that is, the AR(1) PACF drops off to zero when the lag $k > 1$. The AR(2) PACF drops off to zero when the lag $k > 2$.

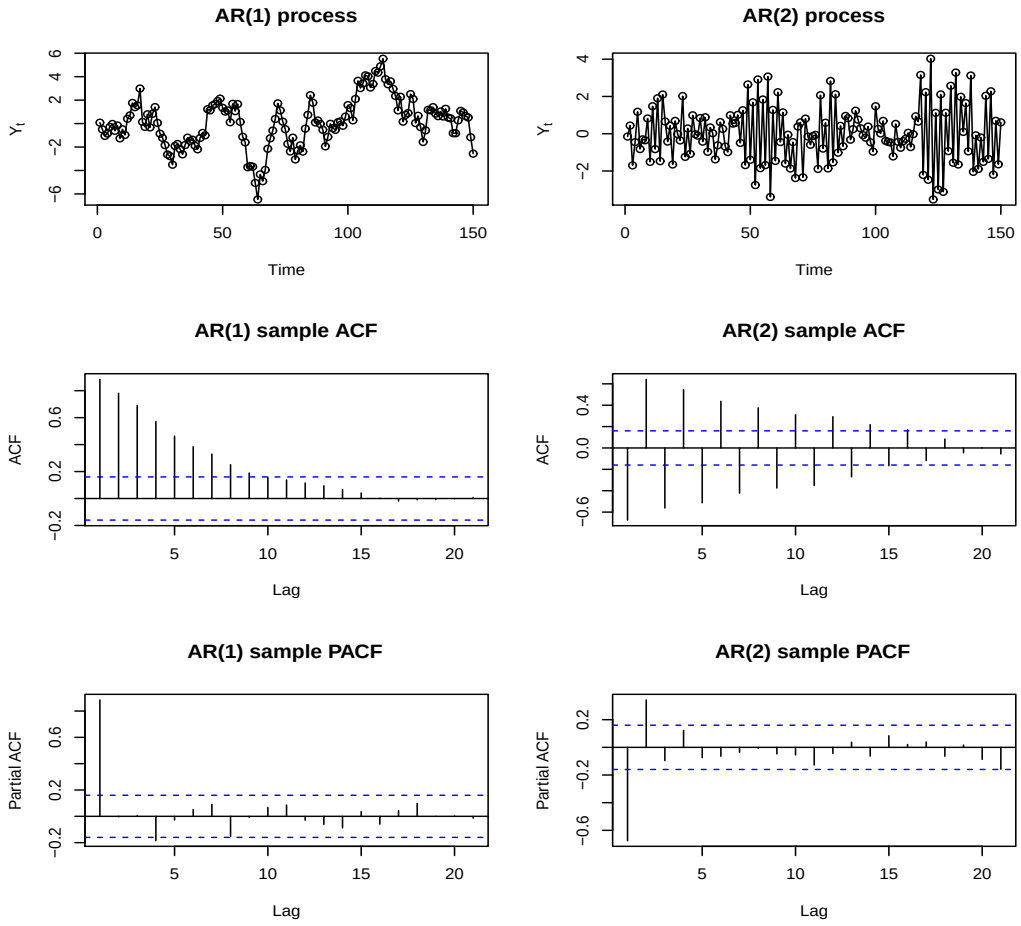


Figure 6.4: Left: AR(1) simulation with $e_t \sim$ iid $\mathcal{N}(0,1)$ and $n = 150$; sample ACF (middle), and sample PACF (bottom). Right: AR(2) simulation with $e_t \sim$ iid $\mathcal{N}(0,1)$ and $n = 150$; sample ACF (middle), and sample PACF (bottom).

- Figure 6.4 displays the sample ACF/PACFs. Just as the sample ACF is an estimate of the true (population) ACF, the sample PACF is an **estimate** of the true (population) PACF.
- Note that the sample PACF for the AR(1) simulation declares $\widehat{\phi}_{kk}$ insignificant for $k > 1$. The estimates of ϕ_{kk}, for $k > 1$, are all within the margin of error bounds. The sample PACF for the AR(2) simulation declares $\widehat{\phi}_{kk}$ insignificant for $k > 2$.
- We will soon discuss why the **PACF error bounds** here are correct.

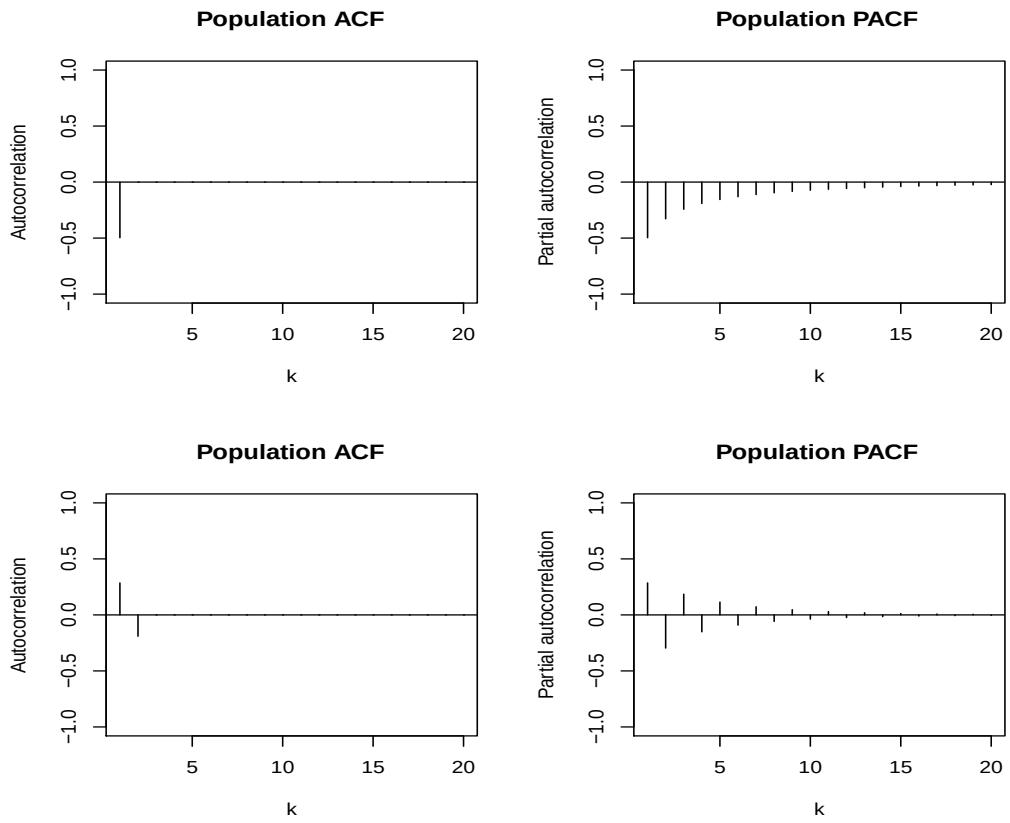


Figure 6.5: Top: MA(1) model with $\theta = 0.9$; population ACF (left) and population PACF (right). Bottom: MA(2) model with $\theta_1 = -0.5$ and $\theta_2 = 0.25$; population ACF (left) and population PACF (right).

CURIOSITY: How does the PACF behave for a **moving average** process? To answer this, consider the invertible MA(1) model, $Y_t = e_t - \theta e_{t-1}$. For this process, it can be shown that

$$\phi_{kk} = \frac{\theta^k(\theta^2 - 1)}{1 - \theta^{2(k+1)}},$$

for $k \geq 1$. Because $|\theta| < 1$ (invertibility requirement), note that

$$\lim_{k\to\infty} \phi_{kk} = \lim_{k\to\infty} \frac{\theta^k(\theta^2 - 1)}{1 - \theta^{2(k+1)}} = 0.$$

That is, the PACF for the MA(1) process decays to zero as the lag k increases, much like the ACF decays to zero for the AR(1). The same happens in higher order MA models.

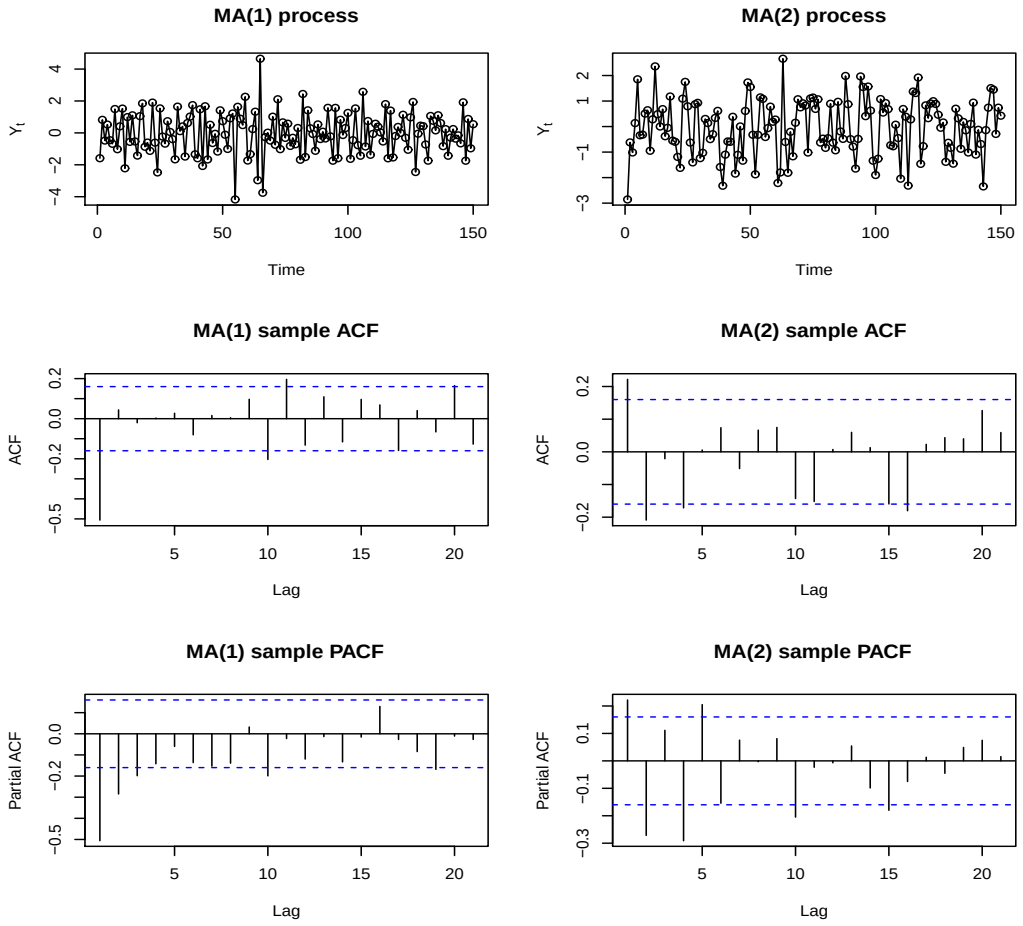


Figure 6.6: Left: MA(1) simulation with $e_t \sim$ iid $\mathcal{N}(0,1)$ and $n = 150$; sample ACF (middle), and sample PACF (bottom). Right: MA(2) simulation with $e_t \sim$ iid $\mathcal{N}(0,1)$ and $n = 150$; sample ACF (middle), and sample PACF (bottom).

IMPORTANT: **The PACF for an MA process behaves much like the ACF for an AR process of the same order.**

Example 6.5. We use R to generate observations from two moving average processes:

(i) $Y_t = e_t - 0.9e_{t-1} \iff$ **MA(1)**, with $\theta = 0.9$

(ii) $Y_t = e_t + 0.5e_{t-1} - 0.25e_{t-2} \iff$ **MA(2)**, with $\theta_1 = -0.5$ and $\theta_2 = 0.25$.

We take $e_t \sim$ iid $\mathcal{N}(0,1)$ and $n = 150$. Figure 6.5 displays the true (**population**) ACF and PACF for these processes. Figure 6.6 displays the simulated time series from each

MA model and the **sample** ACF/PACF.

- The population ACFs in Figure 6.5 display the well-known characteristics; that is, the MA(1) ACF drops off to zero when the lag $k > 1$. The MA(2) ACF drops off to zero when the lag $k > 2$.
- The population PACF in Figure 6.5 for both the MA(1) and MA(2) decays to zero as the lag k increases. This is the theoretical behavior exhibited in the ACF for an AR process.
- The sample versions in Figure 6.6 largely agree with what we know to be true theoretically.

COMPARISON: The following table succinctly summarizes the behavior of the ACF and PACF for moving average and autoregressive processes.

	AR(p)	MA(q)
ACF	Tails off	Cuts off after lag q
PACF	Cuts off after lag p	Tails off

Therefore, the ACF is the key tool to help determine the order of a MA process. The PACF is the key tool to help determine the order of an AR process. For mixed ARMA processes, we need a different tool (coming up).

COMPUTATION: For any stationary ARMA process, it is possible to compute the theoretical PACF values ϕ_{kk}, for $k = 1, 2, ...,$. For a fixed k, we have the following **Yule-Walker equations**:

$$\begin{aligned}
\rho_1 &= \phi_{k,1} + \rho_1\phi_{k,2} + \rho_2\phi_{k,3} + \cdots + \rho_{k-1}\phi_{kk} \\
\rho_2 &= \rho_1\phi_{k,1} + \phi_{k,2} + \rho_1\phi_{k,3} + \cdots + \rho_{k-2}\phi_{kk} \\
&\vdots \\
\rho_k &= \rho_{k-1}\phi_{k,1} + \rho_{k-2}\phi_{k,2} + \rho_{k-3}\phi_{k,3} + \cdots + \phi_{kk},
\end{aligned}$$

where

$$\begin{aligned}\rho_j &= \text{corr}(Y_t, Y_{t-j}) \\ \phi_{k,j} &= \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}, \quad j = 1, 2, ..., k-1 \\ \phi_{kk} &= \text{corr}(Y_t, Y_{t-k}|Y_{t-1}, Y_{t-2}, ..., Y_{t-(k-1)}).\end{aligned}$$

For known $\rho_1, \rho_2, ..., \rho_k$, we can solve this system for $\phi_{k,1}, \phi_{k,2}, ..., \phi_{k,k-1}, \phi_{kk}$, and keep the value of ϕ_{kk}.

Example 6.6. The `ARMAacf` function in R will compute partial autocorrelations for any stationary ARMA model. For example, for the AR(2) model

$$Y_t = 0.6Y_{t-1} - 0.4Y_{t-2} + e_t,$$

we compute the first ten (theoretical) autocorrelations ρ_k and partial autocorrelations ϕ_{kk}. Note that I use the `round` function for aesthetic reasons.

```
> round(ARMAacf(ar = c(0.6,-0.4), lag.max = 10),digits=3)
     0      1      2      3      4      5      6      7      8      9     10
 1.000  0.429 -0.143 -0.257 -0.097  0.045  0.066  0.022 -0.013 -0.017 -0.005

> round(ARMAacf(ar = c(0.6,-0.4), lag.max = 10, pacf=TRUE),digits=3)
 [1]  0.429 -0.400  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
```

Similarly, for the MA(2) model

$$Y_t = e_t + 0.6e_{t-1} - 0.4e_{t-2} + e_t,$$

we compute the first ten (theoretical) autocorrelations and partial autocorrelations.

```
> round(ARMAacf(ma = c(0.6,-0.4), lag.max = 10),digits=3)
     0      1      2      3      4      5      6      7      8      9     10
 1.000  0.237 -0.263  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000

> round(ARMAacf(ma = c(0.6,-0.4), lag.max = 10, pacf=TRUE),digits=3)
 [1]  0.237 -0.338  0.196 -0.189  0.149 -0.134  0.116 -0.105  0.095 -0.086
```

ESTIMATION: The partial autocorrelation ϕ_{kk} can be estimated by taking the Yule-Walker equations and substituting r_k in for the true autocorrelations ρ_k, that is,

$$\begin{aligned} r_1 &= \phi_{k,1} + r_1\phi_{k,2} + r_2\phi_{k,3} + \cdots + r_{k-1}\phi_{kk} \\ r_2 &= r_1\phi_{k,1} + \phi_{k,2} + r_1\phi_{k,3} + \cdots + r_{k-2}\phi_{kk} \\ &\vdots \\ r_k &= r_{k-1}\phi_{k,1} + r_{k-2}\phi_{k,2} + r_{k-3}\phi_{k,3} + \cdots + \phi_{kk}. \end{aligned}$$

This system can then be solved for $\phi_{k,1}, \phi_{k,2}, ..., \phi_{k,k-1}, \phi_{kk}$ as before, but now the solutions are estimates $\widehat{\phi}_{k,1}, \widehat{\phi}_{k,2}, ..., \widehat{\phi}_{k,k-1}, \widehat{\phi}_{kk}$. This can be done for each $k = 1, 2, ...,$.

RESULT: When the AR(p) model is correct, then for large n,

$$\widehat{\phi}_{kk} \sim \mathcal{AN}\left(0, \frac{1}{n}\right),$$

for all $k > p$. Therefore, we can use $\pm z_{\alpha/2}/\sqrt{n}$ as "critical points" to test, at level α,

$$H_0 : \text{AR}(p) \text{ model is appropriate}$$
$$\text{versus}$$
$$H_1 : \text{AR}(p) \text{ model is not appropriate}$$

in the same way that we tested whether or not a specific MA model was appropriate using the sample autocorrelations r_k. See Example 6.1 (notes).

6.4 The extended autocorrelation function

REMARK: We have learned that the autocorrelation function (ACF) can help us determine the order of an MA(q) process because $\rho_k = 0$, for all lags $k > q$. Similarly, the partial autocorrelation function (PACF) can help us determine the order of an AR(p) process because $\phi_{kk} = 0$, for all lags $k > p$. Therefore, in the sample versions of the ACF and PACF, we can look for values of r_k and $\widehat{\phi}_{kk}$, respectively, that are consistent with this theory. We have also discussed formal testing procedures that can be used to

determine if a given MA(q) or AR(p) model is appropriate. A problem, however, is that neither the sample ACF nor sample PACF is all that helpful if the underlying process is a mixture of autoregressive and moving average parts, that is, an **ARMA** process. Therefore, we introduce a new function to help us identify the orders of an ARMA(p, q) process, the **extended autocorrelation function**.

MOTIVATION: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Recall that a stationary ARMA(p, q) process can be expressed as

$$\phi(B)Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned} \phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q). \end{aligned}$$

To start our discussion, note that

$$\begin{aligned} W_t \equiv \phi(B)Y_t &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)Y_t \\ &= Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} \end{aligned}$$

follows an MA(q) model, that is,

$$W_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)e_t.$$

Of course, the $\{W_t\}$ process is not observed because W_t depends on $\phi_1, \phi_2, ..., \phi_p$, which are unknown parameters.

STRATEGY: Suppose that we regress Y_t on $Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ (that is, use the p lagged versions of Y_t as independent variables in a multiple linear regression) and use ordinary least squares to fit the no-intercept model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t,$$

where ϵ_t denotes a generic error term (not the white noise term in the MA process). This would produce estimates $\widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p$ from which we could compute

$$\begin{aligned} \widehat{W}_t &= (1 - \widehat{\phi}_1 B - \widehat{\phi}_2 B^2 - \cdots - \widehat{\phi}_p B^p)Y_t \\ &= Y_t - \widehat{\phi}_1 Y_{t-1} - \widehat{\phi}_2 Y_{t-2} - \cdots - \widehat{\phi}_p Y_{t-p}. \end{aligned}$$

These values (which are merely the residuals from the regression) serve as proxies for the true $\{W_t\}$ process, and we could now treat these residuals as our "data."

- In particular, we could construct the sample ACF for the $\widehat{W}_t$ data so that we can learn about the order q of the MA part of the process.
- For example, if we fit an AR(2) model $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$ and the residuals $\widehat{W}_t$ look to follow an MA(2) process, then this would suggest that a mixed ARMA(2,2) model is worthy of consideration.

PROBLEM: We have just laid out a sensible strategy on how to select candidate ARMA models; i.e., choosing values for p and q. The problem is that ordinary least squares regression estimates $\widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p$ are **inconsistent** estimates of $\phi_1, \phi_2, ..., \phi_p$ when the underlying process is ARMA(p, q). Inconsistency means that the estimates $\widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p$ estimate the wrong things (in a large-sample sense). Therefore, the strategy that we have just described could lead to **incorrect identification** of p and q.

ADJUSTMENT: We now describe an "algorithm" to repair the approach just outlined.

0. Consider using ordinary least squares to fit the same no-intercept AR(p) model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t,$$

where ϵ_t denotes the error term (not the white noise term in an MA process). If the true process is an ARMA(p, q), then the least squares estimates from the regression, say,

$$\widehat{\phi}_1^{(0)}, \widehat{\phi}_2^{(0)}, ..., \widehat{\phi}_p^{(0)}$$

will be inconsistent and the least squares residuals

$$\widehat{\epsilon}_t^{(0)} = Y_t - \widehat{\phi}_1^{(0)} Y_{t-1} - \widehat{\phi}_2^{(0)} Y_{t-2} - \cdots - \widehat{\phi}_p^{(0)} Y_{t-p}$$

will not be white noise. In fact, if $q \geq 1$ (so that the true process is ARMA), then the residuals $\widehat{\epsilon}_t^{(0)}$ and lagged versions of them will contain information about the process $\{Y_t\}$.

1. Because the residuals $\widehat{\epsilon}_t^{(0)}$ contain information about the value of q, we first fit the model

$$Y_t = \phi_1^{(1)} Y_{t-1} + \phi_2^{(1)} Y_{t-2} + \cdots + \phi_p^{(1)} Y_{t-p} + \beta_1^{(1)} \widehat{\epsilon}_{t-1}^{(0)} + \epsilon_t^{(1)},$$

   Note that we have added the lag 1 residuals $\widehat{\epsilon}_{t-1}^{(0)}$ from the initial model fit as a predictor in the regression.

   - If the order of the MA part of the ARMA process is truly $q = 1$, then the least squares estimates

$$\widehat{\phi}_1^{(1)}, \widehat{\phi}_2^{(1)}, ..., \widehat{\phi}_p^{(1)}$$

     will be consistent; i.e., they will estimate the true AR parameters in large samples.

   - If $q > 1$, then the estimates will be inconsistent and the residual process $\{\widehat{\epsilon}_t^{(1)}\}$ will not be white noise.

2. If $q > 1$, then the residuals from the most recent regression $\widehat{\epsilon}_t^{(1)}$ still contain information about the value of q, so we next fit the model

$$Y_t = \phi_1^{(2)} Y_{t-1} + \phi_2^{(2)} Y_{t-2} + \cdots + \phi_p^{(2)} Y_{t-p} + \beta_1^{(2)} \widehat{\epsilon}_{t-1}^{(1)} + \beta_2^{(2)} \widehat{\epsilon}_{t-2}^{(0)} + \epsilon_t^{(2)}.$$

   Note that in this model, we have added the lag 2 residuals $\widehat{\epsilon}_{t-2}^{(0)}$ from the initial model fit as well as the lag 1 residuals $\widehat{\epsilon}_{t-1}^{(1)}$ from the most recent fit.

   - If the order of the MA part of the ARMA process is truly $q = 2$, then the least squares estimates

$$\widehat{\phi}_1^{(2)}, \widehat{\phi}_2^{(2)}, ..., \widehat{\phi}_p^{(2)}$$

     will be consistent; i.e., they will estimate the true AR parameters in large samples.

   - If $q > 2$, then the estimates will be inconsistent and the residual process $\{\widehat{\epsilon}_t^{(2)}\}$ will not be white noise.

3. We continue this iterative process, at each step, adding the residuals from the most recent fit in the same fashion. For example, at the next step, we would fit

$$Y_t = \phi_1^{(3)} Y_{t-1} + \phi_2^{(3)} Y_{t-2} + \cdots + \phi_p^{(3)} Y_{t-p} + \beta_1^{(3)} \widehat{\epsilon}_{t-1}^{(2)} + \beta_2^{(3)} \widehat{\epsilon}_{t-2}^{(1)} + \beta_3^{(3)} \widehat{\epsilon}_{t-3}^{(0)} + \epsilon_t^{(3)}.$$

We continue fitting higher order models until residuals (from the most recent fit) resemble a white noise process.

EXTENDED ACF: In practice, the true orders p and q of the ARMA(p, q) model are unknown and have to be estimated. Based on the strategy outlined, however, we can estimate p and q using a new type of function. For an AR(m) model fit, define the mth **sample extended autocorrelation function** (**EACF**) $\widehat{\rho}_j^{(m)}$ as the sample ACF for the residual process

$$\begin{aligned}\widehat{W}_t^{(j)} &= (1 - \widehat{\phi}_1^{(j)}B - \widehat{\phi}_2^{(j)}B^2 - \cdots - \widehat{\phi}_m^{(j)}B^m)Y_t \\ &= Y_t - \widehat{\phi}_1^{(j)}Y_{t-1} - \widehat{\phi}_2^{(j)}Y_{t-2} - \cdots - \widehat{\phi}_m^{(j)}Y_{t-m},\end{aligned}$$

for $m = 0, 1, 2, ...$, and $j = 0, 1, 2, ...,$. Here, the subscript j refers to the iteration number in the aforementioned sequential fitting process (hence, j refers to the order the MA part). The value m refers to the AR part of the process. Usually the maximum values of m and j are taken to be 10 or so.

	MA					
AR	0	1	2	3	4	$\cdots$
0	$\widehat{\rho}_1^{(0)}$	$\widehat{\rho}_2^{(0)}$	$\widehat{\rho}_3^{(0)}$	$\widehat{\rho}_4^{(0)}$	$\widehat{\rho}_5^{(0)}$	$\cdots$
1	$\widehat{\rho}_1^{(1)}$	$\widehat{\rho}_2^{(1)}$	$\widehat{\rho}_3^{(1)}$	$\widehat{\rho}_4^{(1)}$	$\widehat{\rho}_5^{(1)}$	$\cdots$
2	$\widehat{\rho}_1^{(2)}$	$\widehat{\rho}_2^{(2)}$	$\widehat{\rho}_3^{(2)}$	$\widehat{\rho}_4^{(2)}$	$\widehat{\rho}_5^{(2)}$	$\cdots$
3	$\widehat{\rho}_1^{(3)}$	$\widehat{\rho}_2^{(3)}$	$\widehat{\rho}_3^{(3)}$	$\widehat{\rho}_4^{(3)}$	$\widehat{\rho}_5^{(3)}$	$\cdots$
4	$\widehat{\rho}_1^{(4)}$	$\widehat{\rho}_2^{(4)}$	$\widehat{\rho}_3^{(4)}$	$\widehat{\rho}_4^{(4)}$	$\widehat{\rho}_5^{(4)}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

REPRESENTATION: It is useful to arrange the estimates $\widehat{\rho}_j^{(m)}$ in a **two-way table** where one direction corresponds to the AR part and the other direction corresponds to the MA part. Mathematical arguments show that, as $n \to \infty$,

$$\begin{aligned}\widehat{\rho}_j^{(m)} &\longrightarrow 0, \quad \text{for } 0 \leq m - p < j - q \\ \widehat{\rho}_j^{(m)} &\longrightarrow c \neq 0, \quad \text{otherwise.}\end{aligned}$$

Therefore, the true large-sample extended autocorrelation function (EACF) table for an **ARMA**$(1,1)$ process, for example, looks like

	MA						
AR	0	1	2	3	4	5	$\cdots$
0	x	x	x	x	x	x	$\cdots$
1	x	**0**	0	0	0	0	$\cdots$
2	x	x	0	0	0	0	$\cdots$
3	x	x	x	0	0	0	$\cdots$
4	x	x	x	x	0	0	$\cdots$
5	x	x	x	x	x	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

In this table, the "0" entries correspond to the zero limits of $\widehat{\rho}_j^{(m)}$. The "x" entries correspond to limits of $\widehat{\rho}_j^{(m)}$ which are nonzero. Therefore, the geometric pattern formed by the zeros is a "wedge" with a tip at (1,1). This tip corresponds to the values of $p = 1$ and $q = 1$ in the ARMA model.

The true large-sample EACF table for an **ARMA**$(2,2)$ process looks like

	MA						
AR	0	1	2	3	4	5	$\cdots$
0	x	x	x	x	x	x	$\cdots$
1	x	x	x	x	x	x	$\cdots$
2	x	x	**0**	0	0	0	$\cdots$
3	x	x	x	0	0	0	$\cdots$
4	x	x	x	x	0	0	$\cdots$
5	x	x	x	x	x	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

In this table, we see that the tip of the wedge is at the point (2,2). This tip corresponds to the values of $p = 2$ and $q = 2$ in the ARMA model.

The true large-sample EACF table for an **ARMA**$(2,1)$ process looks like

	MA						
AR	0	1	2	3	4	5	$\cdots$
0	x	x	x	x	x	x	$\cdots$
1	x	x	x	x	x	x	$\cdots$
2	x	**0**	0	0	0	0	$\cdots$
3	x	x	0	0	0	0	$\cdots$
4	x	x	x	0	0	0	$\cdots$
5	x	x	x	x	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

In this table, we see that the tip of the wedge is at the point (2,1). This tip corresponds to the values of $p = 2$ and $q = 1$ in the ARMA model.

DISCLAIMER: The tables shown above represent **theoretical results** for infinitely large sample sizes. Of course, with real data, we would not expect the tables to follow such a clear cut pattern. Remember, the sample EACF values $\widehat{\rho}_j^{(m)}$ are estimates, so they have inherent sampling variation! This is important to keep in mind. For some data sets, the sample EACF table may reveal 2 or 3 models which are consistent with the estimates. In other situations, the sample EACF may be completely ambiguous and give little or no information, especially if the sample size n is small.

SAMPLING DISTRIBUTION: When the residual process

$$\widehat{W}_t^{(j)} = (1 - \widehat{\phi}_1^{(j)}B - \widehat{\phi}_2^{(j)}B^2 - \cdots - \widehat{\phi}_m^{(j)}B^m)Y_t$$

is truly white noise, then the sample extended autocorrelation function estimator

$$\widehat{\rho}_j^{(m)} \sim \mathcal{AN}\left(0, \frac{1}{n-m-j}\right),$$

when n is large. Therefore, we would expect 95 percent of the estimates $\widehat{\rho}_j^{(m)}$ to fall within $\pm 1.96/\sqrt{n-m-j}$. Values outside these cutoffs are classified with an "x" in the **sample EACF**. Values within these bounds are classified with a "0."

Example 6.7. We use R to simulate data from three different ARMA(p, q) processes and examine the sample EACF produced in R. The first simulation is an

- **ARMA(1,1)**, with $n = 200$, $\phi = 0.6$, $\theta = -0.8$, and $e_t \sim$ iid $\mathcal{N}(0, 1)$.

The sample EACF produced from the simulation was

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	o	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	x	o	o	o	o	o	o	o	o	o
3	x	x	x	o	o	o	o	o	o	o	o	o	o	o
4	x	o	x	o	x	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	o	o	o	o	o
6	x	x	o	x	x	o	o	o	o	o	o	o	o	o
7	x	x	o	x	o	x	o	o	o	o	o	o	o	o

INTERPRETATION: This sample EACF agrees largely with the theory, which says that there should be a wedge of zeros with tip at (1,1); the "x"s at (2,4) and (4,4) may be false positives. If one is willing to additionally assume that the "x" at (3,2) is a false positive, then an ARMA(2,1) model would also be deemed consistent with these estimates.

The second simulation is an

- **ARMA(2,2)**, with $n = 200$, $\phi_1 = 0.5$, $\phi_2 = -0.5$, $\theta_1 = -0.8$, $\theta_2 = 0.2$, and $e_t \sim$ iid $\mathcal{N}(0, 1)$.

The sample EACF produced from the simulation was

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	o	x	o	o	o	o	o	x	o	o	o
1	x	x	x	o	x	o	o	o	o	o	x	o	x	o
2	x	o	o	o	o	o	o	o	o	o	x	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	x	o	x	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	o	o	o	o	o
6	x	x	x	x	o	o	o	o	x	o	o	o	o	o
7	x	o	x	x	x	o	o	o	o	o	o	o	o	o

INTERPRETATION: This sample EACF also agrees largely with the theory, which says that there should be a wedge of zeros with tip at (2,2). If one is willing to additionally assume that the "x" at (4,3) is a false positive, then an ARMA(2,1) model would also be deemed consistent with these estimates.

Finally, we use an

- **ARMA(3,3)**, with $n = 200$, $\phi_1 = 0.8$, $\phi_2 = 0.8$, $\phi_3 = -0.9$, $\theta_1 = 0.9$, $\theta_2 = -0.8$, $\theta_3 = 0.2$, and $e_t \sim$ iid $\mathcal{N}(0, 1)$.

The sample EACF produced from the simulation was

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	x	x	o	x	x	x	x	x	x	x	o	x	x
2	x	x	x	o	x	x	x	x	x	x	x	o	x	x
3	x	x	o	x	x	x	x	x	o	o	o	o	o	o
4	x	x	o	x	o	o	o	o	o	o	o	o	o	o
5	x	o	o	x	o	o	o	o	o	o	o	o	o	o
6	x	o	o	x	o	x	o	o	o	o	o	o	o	o
7	x	o	o	x	o	o	o	o	o	o	o	o	o	o

INTERPRETATION: This sample EACF does not agree with the theory, which says that there should be a wedge of zeros with tip at (3,3). There is more of a "block" of zeros; not a wedge. If we saw this EACF in practice, it would not be all that helpful in model selection.

6.5 Nonstationarity

REVIEW: In general, an ARIMA(p, d, q) process can be written as

$$\phi(B)(1 - B)^d Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\end{aligned}$$

and

$$(1 - B)^d Y_t = \nabla^d Y_t.$$

Up until now, we have discussed three functions to help us identify possible values for p and q in stationary ARMA processes.

- The sample **ACF** can be used to determine the order q of a purely MA process.
- The sample **PACF** can be used to determine the order p of a purely AR process.
- The sample **EACF** can be used to determine the orders p and q of a mixed ARMA process.

DIFFERENCING: For a series of data, a clear indicator of nonstationarity is that the sample ACF exhibits a **very slow decay** across lags. This occurs because in a nonstationary process, the series tends to "hang together" and displays "trends."

- When there is a clear trend in the data (e.g., linear) and the sample ACF for a series decays very slowly, take first differences.
- If the sample ACF for the first differences resembles that a stationary ARMA process (the ACF decays quickly), then take $d = 1$ in the ARIMA(p, d, q) family and use the ACF, PACF, and EACF (on the first differences) to identify plausible values of p and q.
- If the sample ACF for the first differences still exhibits a slow decay across lags, take second differences and use $d = 2$. One can then use the ACF, PACF, and EACF (on the second differences) to identify plausible values of p and q. There should rarely be a need to consider values of $d > 2$. In fact, I have found that it is not all that often that even second differences ($d = 2$) are needed.

- If a transformation is warranted (e.g., because of clear evidence of heteroscedasticity), implement it up front before taking any differences. Then, use these guidelines to choose p, d, and q for the transformed series.

TERMINOLOGY: **Overdifferencing** occurs when we choose d to be too large. For example, suppose that the correct model for a process $\{Y_t\}$ is an IMA(1,1), that is,

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1},$$

where $|\theta| < 1$ and $\{e_t\}$ is zero mean white noise. The first differences are given by

$$\nabla Y_t = Y_t - Y_{t-1} = e_t - \theta e_{t-1},$$

which is a stationary and invertible MA(1) process. The second differences are given by

$$\begin{aligned}
\nabla^2 Y_t &= \nabla Y_t - \nabla Y_{t-1} \\
&= (e_t - \theta e_{t-1}) - (e_{t-1} - \theta e_{t-2}) \\
&= e_t - (1+\theta)e_{t-1} + \theta e_{t-2} \\
&= [1 - (1+\theta)B + \theta B^2]e_t.
\end{aligned}$$

The second difference process is not invertible because

$$\theta(x) = 1 - (1+\theta)x + \theta x^2$$

has a unit root $x = 1$. Therefore, by unnecessarily taking second differences, we have created a problem. Namely, we have differenced an invertible MA(1) process (for first differences) into one which is not invertible. Recall that if a process is not invertible (here, the second differences), then the parameters in the model can not be estimated uniquely. In this example, the correct value of d is $d = 1$. Taking $d = 2$ would be an example of overdifferencing.

INFERENCE: Instead of relying on the sample ACF, which may be subjective in "borderline cases," we can formally test whether or not an observed time series is stationary using the methodology proposed by Dickey and Fuller (1979).

DEVELOPMENT: To set our ideas, consider the model

$$Y_t = \alpha Y_{t-1} + X_t,$$

where $\{X_t\}$ is a stationary AR(k) process, that is,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_k X_{t-k} + e_t,$$

where $\{e_t\}$ is zero mean white noise. Therefore,

$$\begin{aligned} Y_t &= \alpha Y_{t-1} + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_k X_{t-k} + e_t \\ &= \alpha Y_{t-1} + \phi_1(Y_{t-1} - \alpha Y_{t-2}) + \phi_2(Y_{t-2} - \alpha Y_{t-3}) + \cdots + \phi_k(Y_{t-k} - \alpha Y_{t-k-1}) + e_t. \end{aligned}$$

After some algebra, we can rewrite this model for Y_t as

$$\phi^*(B) Y_t = e_t,$$

where

$$\phi^*(B) = \phi(B)(1 - \alpha B)$$

and where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_k B^k)$ is the usual AR characteristic operator of order k. Note that

- if $\alpha = 1$, then $\phi^*(B) = \phi(B)(1 - B)$, that is, $\phi^*(x)$, a polynomial of degree $k + 1$, has a unit root and $\{Y_t\}$ is not stationary.
- if $-1 < \alpha < 1$, then $\phi^*(x)$ does not have a unit root, and $\{Y_t\}$ is a stationary AR($k + 1$) process.

The **augmented Dickey-Fuller** (ADF) **unit root test** therefore tests

$$H_0 : \alpha = 1 \text{ (nonstationarity)}$$
$$\text{versus}$$
$$H_1 : \alpha < 1 \text{ (stationarity)}.$$

IMPLEMENTATION: Dickey and Fuller advocated that this test could be carried out using least squares regression. To see how, note that when $H_0 : \alpha = 1$ is true (i.e., the process is nonstationary), the model for Y_t can be written as

$$\begin{aligned} Y_t - Y_{t-1} &= \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \cdots + \phi_k(Y_{t-k} - Y_{t-k-1}) + e_t \\ &= aY_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \cdots + \phi_k(Y_{t-k} - Y_{t-k-1}) + e_t, \end{aligned}$$

where $a = \alpha - 1$. **Note that $a = 0$ when $\alpha = 1$.** That is,

$$H_0 : \alpha = 1 \text{ is true} \quad \Longleftrightarrow \quad H_0 : a = 0 \text{ is true.}$$

Using difference notation, the model under $H_0 : \alpha = 1$ is

$$\nabla Y_t = aY_{t-1} + \phi_1 \nabla Y_{t-1} + \phi_2 \nabla Y_{t-2} + \cdots + \phi_k \nabla Y_{t-k} + e_t.$$

Therefore, we carry out the test by regressing ∇Y_t on Y_{t-1}, ∇Y_{t-1}, ∇Y_{t-2}, ..., ∇Y_{t-k}. We can then decide between H_0 and H_1 by examining the size of the least-squares estimate of a. In particular,

- if the least squares regression estimate of a is significantly different from 0, we reject H_0 and conclude that the process is stationary.
- if the least squares regression estimate of a is not significantly different from 0, we do not reject H_0. This decision would suggest the process $\{Y_t\}$ is nonstationary.

REMARK: The test statistic needed to test H_0 versus H_1, and its large-sample distribution, are complicated (the test statistic is similar to the t test statistic from ordinary least squares regression; however, the large-sample distribution is not t). Fortunately, there is an R function to implement the test automatically. The only thing we need to do is choose a value of k in the model

$$\nabla Y_t = aY_{t-1} + \phi_1 \nabla Y_{t-1} + \phi_2 \nabla Y_{t-2} + \cdots + \phi_k \nabla Y_{t-k} + e_t,$$

that is, the value k is the order of the AR process for ∇Y_t. Of course, the true value of k is unknown. However, we can have R determine the "best value" of k using model selection criteria that we will discuss in the next subsection.

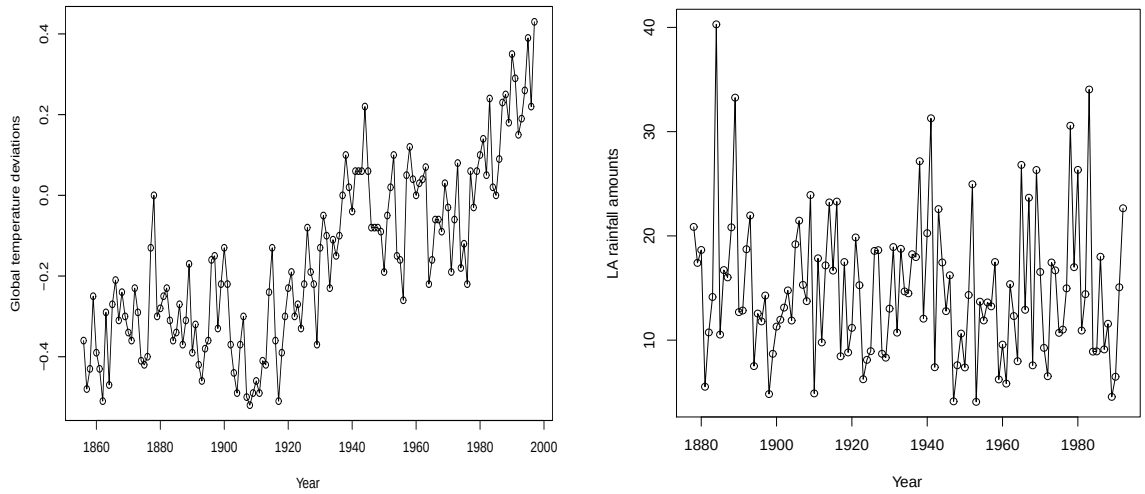


Figure 6.7: Left: Global temperature data. Right: Los Angeles annual rainfall data.

Example 6.8. We illustrate the ADF test using two data sets from Chapter 1, the global temperature data set (Example 1.1, pp 2, notes) and the Los Angeles annual rainfall data set (Example 1.13, pp 14, notes). For the global temperature data, the command `ar(diff(globtemp))` is used to determine the "best" value of k for the differences. Here, it is $k = 3$. The ADF test output is

```
Null hypothesis: Unit root.
Alternative hypothesis: Stationarity.
ADF statistic:
        Estimate Std. Error t value Pr(>|t|)
adf.reg   -0.031      0.049  -0.636      0.1
  Lag orders: 1 2 3
  Number of available observations: 138
```

In particular, the output automatically produces the p-value for the test

$$H_0 : \alpha = 1 \text{ (nonstationarity)}$$

versus

$$H_1 : \alpha < 1 \text{ (stationarity)}.$$

The large p-value here (> 0.10) does not refute $H_0 : \alpha = 1$. There is insufficient evidence to conclude that the global temperature process is stationary. For the LA rainfall data, the command `ar(diff(larain))` is used to determine the best value of k, which is $k = 4$.

```
Null hypothesis: Unit root.
Alternative hypothesis: Stationarity.
ADF statistic:
        Estimate Std. Error t value Pr(>|t|)
adf.reg   -0.702      0.207  -3.385    0.015
  Lag orders: 1 2 3 4
  Number of available observations: 110
```

The small p-value here ($p = 0.015$) indicates strong evidence against $H_0 : \alpha = 1$. There is sufficient evidence to conclude that the LA rainfall process is stationary.

DISCUSSION: When performing the ADF test, some words of caution are in order.

- When $H_0 : \alpha = 1$ is true, the AR characteristic polynomial $\phi^*(B) = \phi(B)(1 - \alpha B)$ contains a unit root. In other words, $\{Y_t\}$ is nonstationary, but $\{\nabla Y_t\}$ is stationary. This is called **difference nonstationarity**. The ADF procedure we have described, more precisely, tests for difference nonstationarity.

- Because of this, the ADF test outlined here may not have sufficient power to reject H_0 when the process is truly stationary. In addition, the test may reject H_0 incorrectly because a different form of nonstationarity is present (one that can not be overcome merely by taking first differences).

- The ADF test outcome must be interpreted with these points in mind, especially when the sample size n is small. In other words, do not blindly interpret the ADF test outcome as a yes/no indicator of nonstationarity.

IMPORTANT: To implement the ADF test in R, we need to install the `uroot` package. Installing this package has to be done manually.

6.6 Other model selection methods

TERMINOLOGY: The **Akaike's Information Criterion** (**AIC**) says to select the ARMA(p, q) model which minimizes

$$\text{AIC} = -2 \ln L + 2k,$$

where $\ln L$ is the natural logarithm of the maximized likelihood function (computed under a distributional assumption for $Y_1, Y_2, ..., Y_n$) and k is the number of parameters in the model (excluding the white noise variance). In a stationary no-intercept ARMA(p, q) model, there are $k = p + q$ parameters.

- The likelihood function gives (loosely speaking) the "probability of the data," so we would like for it to be as large as possible. This is equivalent to wanting $-2 \ln L$ to be as small as possible.
- The $2k$ term serves as a penalty, namely, we do not want models with too many parameters (adhering to the Principle of Parsimony).
- The AIC is an estimator of the expected **Kullback-Leibler divergence**, which measures the closeness of a candidate model to the truth. The smaller this divergence, the better the model. See pp 130 (CC).
- The AIC is used more generally for model selection in statistics (not just in the analysis of time series data). Herein, we restrict attention to its use in selecting candidate stationary ARMA(p, q) models.

TERMINOLOGY: The **Bayesian Information Criterion** (**BIC**) says to select the ARMA(p, q) model which minimizes

$$\text{BIC} = -2 \ln L + k \ln n,$$

where $\ln L$ is the natural logarithm of the maximized likelihood function and k is the number of parameters in the model (excluding the white noise variance). In a stationary no-intercept ARMA(p, q) model, there are $k = p + q$ parameters.

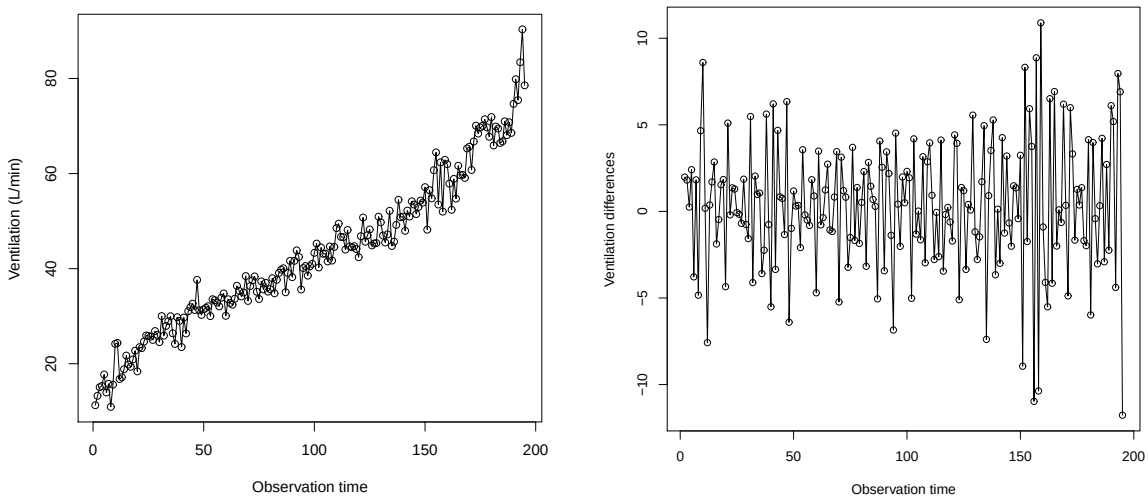


Figure 6.8: Ventilation measurements at 15 second intervals. Left: Ventilation data. Right: First differences.

- Both AIC and BIC require the maximization of a log likelihood function (we assume normality). When compared to AIC, BIC offers a stiffer penalty for overparameterized models since $\ln n$ will often exceed 2.

Example 6.9. We use the BIC as a means for model selection with the ventilation data in Example 1.10 (pp 11, notes); see also Example 5.2 (pp 117, notes). Figure 6.8 shows the original series (left) and the first difference process (right). The BIC output (next page) is provided by R. Remember that the smaller the BIC, the better the model.

- The original ventilation series displays a clear linear trend. The ADF test (results not shown) provides a p-value of $p > 0.10$, indicating that the series is difference nonstationary.
- We therefore find the "best" ARMA(p, q) model for the first differences; that is, we are taking $d = 1$, so we are essentially finding the "best" ARIMA$(p, 1, q)$ model.
- The BIC output in Figure 6.8 shows that the best model (smallest BIC) for the differences contains a lag 1 error component; i.e., $q = 1$.

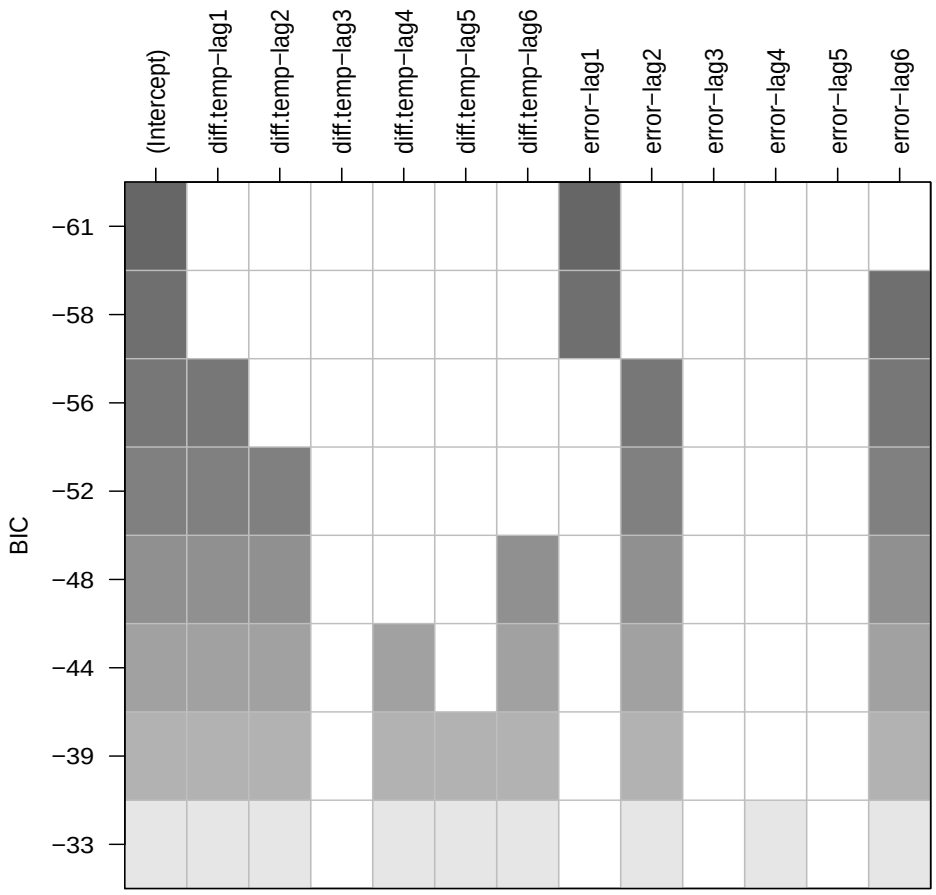


Figure 6.9: Ventilation data. ARMA best subsets output for the first difference process $\{\nabla Y_t\}$ using the BIC.

- Therefore, the model that provides the smallest BIC for $\{\nabla Y_t\}$ is an MA(1).
- In other words, the "best" model for the original ventilation series, as judged by the BIC, is an ARIMA(0,1,1); i.e., an IMA(1,1).

DISCLAIMER: Model selection according to BIC (or AIC) does not always provide "selected" models that are easily interpretable. Therefore, while AIC and BIC are model selection tools, they are not the only tools available to us. The ACF, PACF, and EACF may direct us to models that are different than those deemed "best" by the AIC/BIC.

6.7 Summary

SUMMARY: Here is a summary of the techniques that we have reviewed this chapter. This summary is presented in an "algorithm" format to help guide the data analyst through the ARIMA model selection phase. Advice is interspersed throughout.

1. Plot the data and identify an appropriate transformation if needed.
    - Examining the time series plot, we can get an idea about whether the series contains a trend, seasonality, outliers, nonconstant variance, etc. This understanding often provides a basis for postulating a possible data transformation.
    - Examine the time series plot for nonconstant variance and perform a suitable transformation (from the Box-Cox family); see Chapter 5. Alternatively, the data analyst can try several transformations and choose the one that does the best at stabilizing the variance.
    - Always implement a transformation before taking any data differences.
2. Compute the sample ACF and the sample PACF of the original series (or transformed series) and further confirm the need for differencing.
    - If the sample ACF decays very, very slowly, this usually indicates that it is a good idea to take first differences.
    - Tests for stationarity (ADF test) can also be implemented at this point on the original or transformed series. In a borderline case, differencing is generally recommended.
    - Higher order differencing may be needed (however, I have found that it generally is not). One can perform an ADF test for stationarity of the first differences to see if taking second differences is warranted. In nearly all cases, d is not larger than 2 (i.e., taking second differences).
    - Some authors argue that the consequences of overdifferencing are much less serious than those of underdifferencing. However, overdifferencing can create model identifiability problems.

3. Compute the sample ACF, the sample PACF, and the sample EACF of the original, properly transformed, properly differenced, or properly transformed/differenced series to identify the orders of p and q.

    - Usually, p and q are not larger than 4 (excluding seasonal models, which we have yet to discuss).
    - Use knowledge of the patterns for theoretical versions of these functions; i.e.,
        - the ACF for an MA(q) drops off after lag q
        - the PACF for an AR(p) drops off after lag p
        - the "tip" in the EACF identifies the proper ARMA(p, q) model.
    - We identify the orders p and q by matching the patterns in the sample ACF/PACF/EACF with the theoretical patterns of known models.
    - To build a reasonable model, ideally, we need a minimum of about $n = 50$ observations, and the number of sample ACF and PACF to be calculated should be about $n/4$ (a rough guideline). It might be hard to identify an adequate model with smaller data sets.
    - "The art of model selection is very much like the method of an FBI's agent criminal search. Most criminals disguise themselves to avoid being recognized." This is also true of the ACF, PACF, and EACF. Sampling variation can disguise the theoretical ACF/PACF/EACF patterns.
    - BIC and AIC can also be used to identify models consistent with the data.

REMARK: It is rare, after going through all of this, that the analyst will be able to identify a single model that is a "clear-cut" choice. It is more likely that a small number of **candidate models** have been identified from the steps above.

NEXT STEP: With our (hopefully small) set of candidate models, we then move forward to parameter estimation and model diagnostics (model checking). These topics are the subjects of Chapter 7 and Chapter 8, respectively. Once a final model has been chosen, fit, and diagnosed, forecasting then becomes the central focus (Chapter 9).

7 Estimation

Complementary reading: Chapter 7 (CC).

7.1 Introduction

RECALL: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. In general, an ARIMA$(p, d, q)$ process can be written as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\end{aligned}$$

and

$$(1-B)^d Y_t = \nabla^d Y_t$$

is the series of dth differences. In the last chapter, we were primarily concerned with selecting values of p, d, and q which were consistent with the observed (or suitably transformed) data, that is, we were concerned with **model selection**.

PREVIEW: In this chapter, our efforts are directed towards **estimating** parameters in this class of models. In doing so, it suffices to restrict attention to stationary ARMA(p, q) models. If $d > 0$ (which corresponds to a nonstationary process), the methodology described herein can be applied to the suitably differenced process $(1-B)^d Y_t = \nabla^d Y_t$. Therefore, when we write $Y_1, Y_2, ..., Y_n$ to represent our "data" in this chapter, it is understood that $Y_1, Y_2, ..., Y_n$ may denote the original data, the differenced data, transformed data (e.g., log-transformed, etc.), or possibly data that have been transformed and differenced.

PREVIEW: We will discuss three estimation techniques: **method of moments**, **least squares**, and **maximum likelihood**.

7.2 Method of moments

TERMINOLOGY: The **method of moments** (**MOM**) approach to estimation consists of equating sample moments to the corresponding population (theoretical) moments and solving the resulting system of equations for the model parameters.

7.2.1 Autoregressive models

AR(1): Consider the stationary AR(1) model

$$Y_t = \phi Y_{t-1} + e_t,$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In this model, there are two parameters: ϕ and σ_e^2. The MOM estimator of ϕ is obtained by setting the population lag one autocorrelation ρ_1 equal to the sample lag one autocorrelation r_1 and solving for ϕ, that is,

$$\rho_1 \stackrel{\text{set}}{=} r_1.$$

For this model, we know $\rho_1 = \phi$ (see Chapter 4). Therefore, the MOM estimator of ϕ is

$$\widehat{\phi} = r_1.$$

AR(2): For the AR(2) model,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t,$$

there are three parameters: ϕ_1, ϕ_2, and σ_e^2. To find the MOM estimators of ϕ_1 and ϕ_2, recall the **Yule-Walker equations** (derived in Chapter 4) for the AR(2):

$$\begin{aligned} \rho_1 &= \phi_1 + \rho_1\phi_2 \\ \rho_2 &= \rho_1\phi_1 + \phi_2. \end{aligned}$$

Setting $\rho_1 = r_1$ and $\rho_2 = r_2$, we have

$$\begin{aligned} r_1 &= \phi_1 + r_1\phi_2 \\ r_2 &= r_1\phi_1 + \phi_2. \end{aligned}$$

Solving this system for ϕ_1 and ϕ_2 produces the MOM estimators

$$\begin{aligned}\widehat{\phi}_1 &= \frac{r_1(1-r_2)}{1-r_1^2} \\ \widehat{\phi}_2 &= \frac{r_2-r_1^2}{1-r_1^2}.\end{aligned}$$

AR(p): For the general AR(p) process,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t,$$

there are $p+1$ parameters: ϕ_1, ϕ_2, ..., ϕ_p and σ_e^2. We again recall the Yule-Walker equations from Chapter 4:

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 + \cdots + \phi_p\rho_{p-1} \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 + \cdots + \phi_p\rho_{p-2} \\ &\vdots \\ \rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \phi_3\rho_{p-3} + \cdots + \phi_p.\end{aligned}$$

Just as in the AR(2) case, we set $\rho_1 = r_1$, $\rho_2 = r_2$, ..., $\rho_p = r_p$ to obtain

$$\begin{aligned}r_1 &= \phi_1 + \phi_2 r_1 + \phi_3 r_2 + \cdots + \phi_p r_{p-1} \\ r_2 &= \phi_1 r_1 + \phi_2 + \phi_3 r_1 + \cdots + \phi_p r_{p-2} \\ &\vdots \\ r_p &= \phi_1 r_{p-1} + \phi_2 r_{p-2} + \phi_3 r_{p-3} + \cdots + \phi_p.\end{aligned}$$

The MOM estimators $\widehat{\phi}_1$, $\widehat{\phi}_2$, ..., $\widehat{\phi}_p$ solve this system of equations.

REMARK: Calculating MOM estimates (or any estimates) in practice should be done using software. The MOM approach may produce estimates $\widehat{\phi}_1$, $\widehat{\phi}_2$, ..., $\widehat{\phi}_p$ that fall "outside" the **stationarity region**, even if the process is truly stationary! That is, the estimated AR(p) polynomial, say,

$$\widehat{\phi}_{\text{MOM}}(x) = 1 - \widehat{\phi}_1 x - \widehat{\phi}_2 x^2 - \cdots - \widehat{\phi}_p x^p$$

may possess roots which do not exceed 1 in absolute value (or modulus).

7.2.2 Moving average models

MA(1): Consider the invertible MA(1) process

$$Y_t = e_t - \theta e_{t-1},$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In this model, there are two parameters: θ and σ_e^2. To find the MOM estimator of θ, we solve

$$\rho_1 = \frac{-\theta}{1+\theta^2} \stackrel{\text{set}}{=} r_1 \quad \Longleftrightarrow \quad r_1\theta^2 + \theta + r_1 = 0$$

for θ. Using the quadratic formula, we find that the solutions to this equation are

$$\theta = \frac{-1 \pm \sqrt{1-4r_1^2}}{2r_1}.$$

- If $|r_1| > 0.5$, then no real solutions for θ exist.
- If $|r_1| = 0.5$, then the solutions for θ are ± 1, which corresponds to an MA(1) model that is not invertible.
- If $|r_1| < 0.5$, the invertible solution for θ is the MOM estimator

$$\widehat{\theta} = \frac{-1 + \sqrt{1-4r_1^2}}{2r_1}.$$

NOTE: For higher order MA models, the difficulties become more pronounced. For the general MA(q) case, we are left to solve the highly nonlinear system

$$\begin{aligned} \rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} &\stackrel{\text{set}}{=} r_k, \quad k = 1, 2, ..., q-1 \\ \rho_q = \frac{-\theta_q}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} &\stackrel{\text{set}}{=} r_q, \end{aligned}$$

for θ_1, θ_2, ..., θ_q. Just as in the MA(1) case, there will likely be multiple solutions, only of which at most one will correspond to a fitted invertible model.

IMPORTANT: MOM estimates are not recommended for use with MA models. They are hard to obtain and (as we will see) they are not necessarily "good" estimates.

7.2.3 Mixed ARMA models

ARMA(1,1): Consider the ARMA(1,1) process

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1},$$

where $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In this model, there are three parameters: ϕ, θ, and σ_e^2. Recall from Chapter 4 that

$$\rho_k = \left[\frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\right]\phi^{k-1}.$$

It follows directly that

$$\frac{\rho_2}{\rho_1} = \phi.$$

Setting $\rho_1 = r_1$ and $\rho_2 = r_2$, the MOM estimator of ϕ is given by

$$\widehat{\phi} = \frac{r_2}{r_1}.$$

The MOM estimator of θ then solves

$$r_1 = \frac{(1-\theta\widehat{\phi})(\widehat{\phi}-\theta)}{1-2\theta\widehat{\phi}+\theta^2}.$$

This is a quadratic equation in θ, so there are two solutions. The invertible solution $\widehat{\theta}$ (if any) is kept; i.e., $\widehat{\theta}_{\text{MOM}} = 1 - \widehat{\theta}x$ has root x larger than 1 in absolute value.

7.2.4 White noise variance

GOAL: We now wish to estimate the white noise variance σ_e^2. To do this, we first note that for any stationary ARMA model, the process variance $\gamma_0 = \text{var}(Y_t)$ can be estimated by the **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \overline{Y})^2.$$

- For a general **AR**(p) process, we recall from Chapter 4 that

$$\gamma_0 = \frac{\sigma_e^2}{1-\phi_1\rho_1-\phi_2\rho_2-\cdots-\phi_p\rho_p} \implies \sigma_e^2 = (1-\phi_1\rho_1-\phi_2\rho_2-\cdots-\phi_p\rho_p)\gamma_0.$$

Therefore, the MOM estimator of σ_e^2 is obtained by substituting in $\widehat{\phi}_k$ for ϕ_k, r_k for ρ_k, and S^2 for γ_0. We obtain

$$\widehat{\sigma}_e^2 = (1 - \widehat{\phi}_1 r_1 - \widehat{\phi}_2 r_2 - \cdots - \widehat{\phi}_p r_p) S^2.$$

- For a general **MA**(q) process, we recall from Chapter 4 that

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_e^2 \quad \Longrightarrow \quad \sigma_e^2 = \frac{\gamma_0}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}.$$

Therefore, the MOM estimator of σ_e^2 is obtained by substituting in $\widehat{\theta}_k$ for θ_k and S^2 for γ_0. We obtain

$$\widehat{\sigma}_e^2 = \frac{S^2}{1 + \widehat{\theta}_1^2 + \widehat{\theta}_2^2 + \cdots + \widehat{\theta}_q^2}.$$

- For an **ARMA(1,1)** process,

$$\gamma_0 = \left(\frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2}\right)\sigma_e^2 \quad \Longrightarrow \quad \sigma_e^2 = \left(\frac{1 - \phi^2}{1 - 2\phi\theta + \theta^2}\right)\gamma_0.$$

Therefore, the MOM estimator of σ_e^2 is obtained by substituting in $\widehat{\theta}$ for θ, $\widehat{\phi}$ for ϕ, and S^2 for γ_0. We obtain

$$\widehat{\sigma}_e^2 = \left(\frac{1 - \widehat{\phi}^2}{1 - 2\widehat{\phi}\widehat{\theta} + \widehat{\theta}^2}\right) S^2.$$

7.2.5 Examples

Example 7.1. Suppose $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. In this example, we use Monte Carlo simulation to approximate the sampling distributions of the MOM estimators of θ and σ_e^2 in the MA(1) model

$$Y_t = e_t - \theta e_{t-1}.$$

We take $\theta = 0.7$, $\sigma_e^2 = 1$, and $n = 100$. Recall that the MOM approach is generally not recommended for use with MA models. We will now see why this is true.

- We simulate $M = 2000$ MA(1) time series, each of length $n = 100$, with $\theta = 0.7$ and $\sigma_e^2 = 1$.

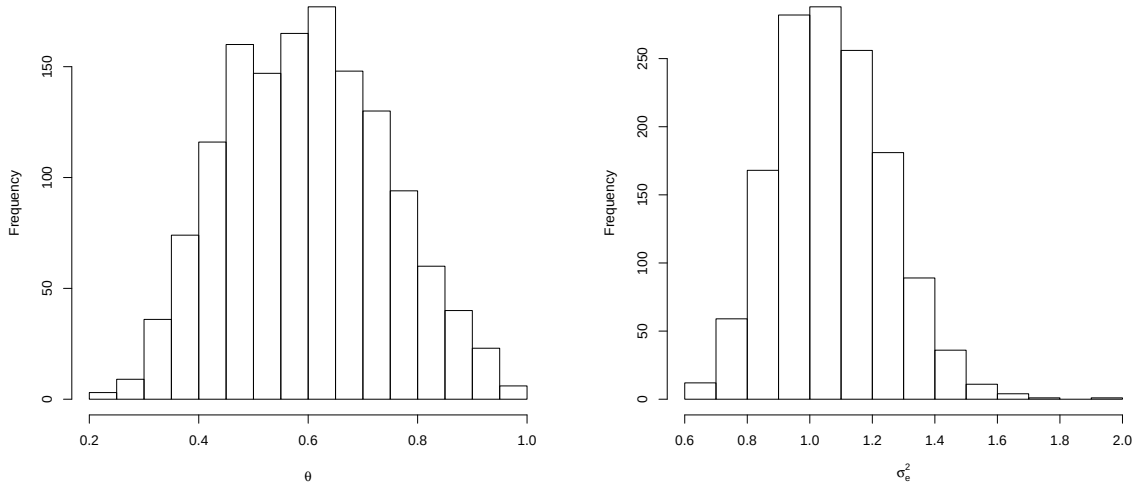


Figure 7.1: Monte Carlo simulation. Left: Histogram of MOM estimates of θ in the MA(1) model. Right: Histogram of MOM estimates of σ_e^2. The true values are $\theta = 0.7$ and $\sigma_e^2 = 1$. The sample size is $n = 100$.

- For each simulated series, we compute the MA(1) MOM estimates

$$\begin{aligned} \widehat{\theta} &= \frac{-1 + \sqrt{1 - 4r_1^2}}{2r_1} \\ \widehat{\sigma}_e^2 &= \frac{S^2}{1 + \widehat{\theta}^2}, \end{aligned}$$

  if they exist. Recall the formula for $\widehat{\theta}$ only makes sense when $|r_1| < 0.5$.

- Of the $M = 2000$ simulated series, only 1388 produced a value of $|r_1| < 0.5$. For the other 612 simulated series, the MOM estimates do not exist (therefore, the histograms in Figure 7.1 contain only 1388 estimates).

- The Monte Carlo distribution of $\widehat{\theta}$ illustrates why MOM estimation is not recommended for MA models. The sampling distribution is not even centered at the true value of $\theta = 0.7$. The MOM estimator $\widehat{\theta}$ is negatively biased.

- The Monte Carlo distribution of $\widehat{\sigma}_e^2$ is slightly skewed to the right with mean larger than $\sigma_e^2 = 1$. The MOM estimator $\widehat{\sigma}_e^2$ looks to be slightly positively biased.

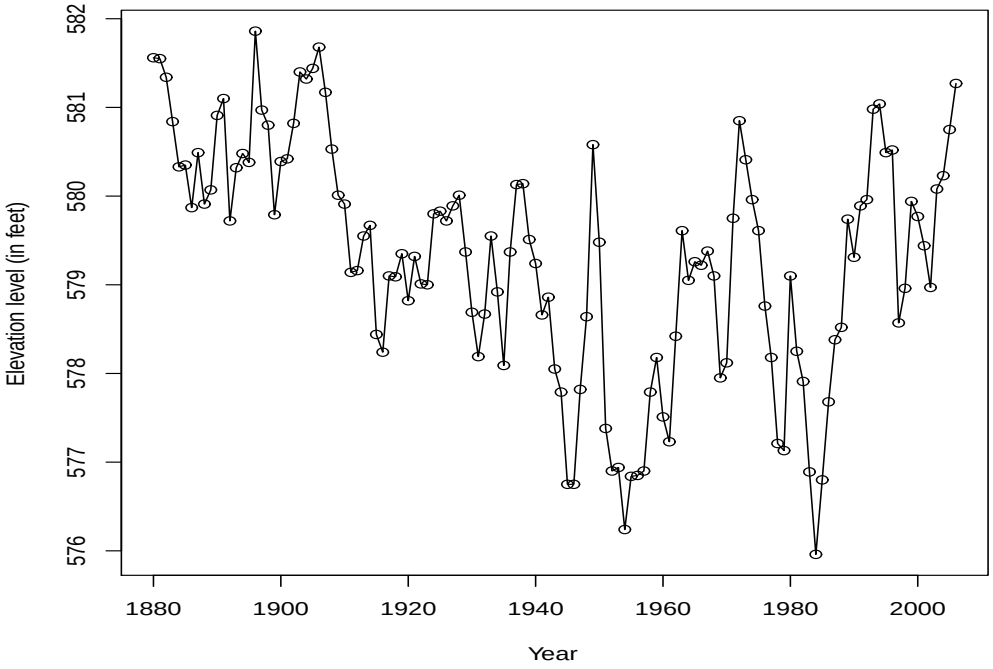


Figure 7.2: Lake Huron data. Average July water surface elevation (measured in feet) during 1880-2006.

Example 7.2. Data file: `huron`. Figure 7.2 displays the average July water surface elevation (measured in feet) from 1880-2006 at Harbor Beach, Michigan, on Lake Huron. The sample ACF and PACF for the series, both given in Figure 7.3, suggest that an AR(1) model or possibly an AR(2) model may be appropriate.

AR(1): First, we consider the AR(1) model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t.$$

Note that this model includes a parameter μ for the overall mean. By inspection, it is clear that $\{Y_t\}$ is not a zero mean process. I used R to compute the sample statistics

$$r_1 = 0.831 \qquad r_2 = 0.643 \qquad \overline{y} = 579.309 \qquad s^2 = 1.783978.$$

For these data, the AR(1) MOM estimate of ϕ is

$$\widehat{\phi} = r_1 = 0.831$$

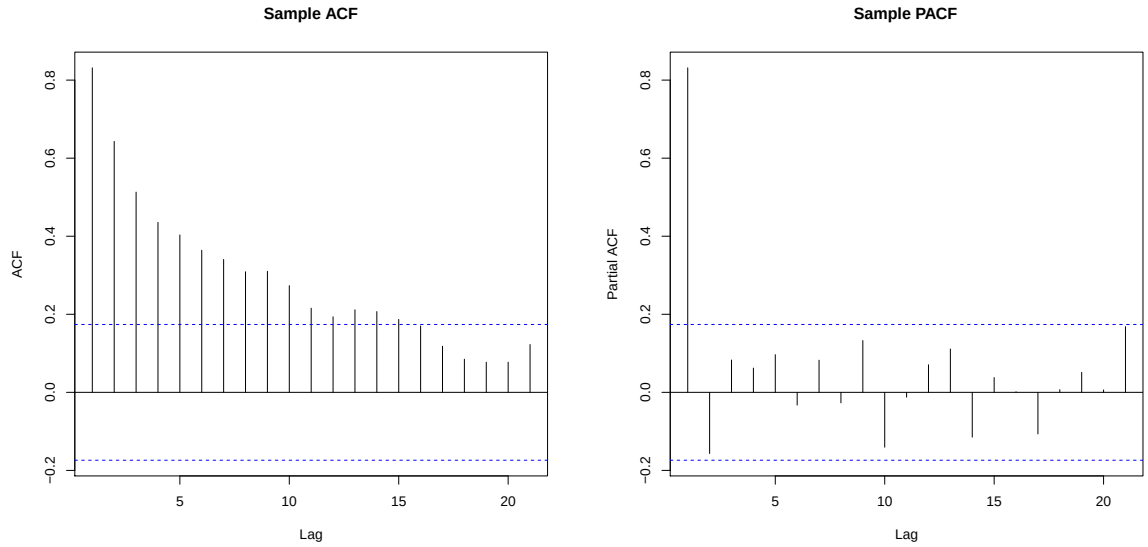


Figure 7.3: Lake Huron data. Left: Sample ACF. Right: Sample PACF.

Using $\overline{y}$ as an (unbiased) estimate of μ, the fitted AR(1) model is

$$Y_t - 579.309 = 0.831(Y_{t-1} - 579.309) + e_t,$$

or, equivalently (after simplifying),

$$Y_t = 97.903 + 0.831Y_{t-1} + e_t.$$

The AR(1) MOM estimate of the white noise variance is

$$\begin{aligned} \widehat{\sigma}_e^2 &= (1 - \widehat{\phi} r_1)s^2 \\ &= [1 - (0.831)(0.831)](1.783978) \approx 0.552. \end{aligned}$$

We can have R automate the estimation process. Here is the output:

```
> ar(huron,order.max=1,AIC=F,method='yw') # method of moments

Coefficients:
     1
0.8315

Order selected 1  sigma^2 estimated as  0.5551
```

AR(2): Consider the AR(2) model

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t.$$

For these data, the AR(2) MOM estimates of ϕ_1 and ϕ_2 are

$$\begin{aligned}\widehat{\phi}_1 &= \frac{r_1(1-r_2)}{1-r_1^2} = \frac{0.831(1-0.643)}{1-(0.831)^2} \approx 0.959 \\ \widehat{\phi}_2 &= \frac{r_2-r_1^2}{1-r_1^2} = \frac{0.643-(0.831)^2}{1-(0.831)^2} \approx -0.154\end{aligned}$$

so the fitted AR(2) model is

$$Y_t - 579.309 = 0.959(Y_{t-1} - 579.309) - 0.154(Y_{t-2} - 579.309) + e_t,$$

or, equivalently (after simplifying),

$$Y_t = 112.965 + 0.959Y_{t-1} - 0.154Y_{t-2} + e_t.$$

The AR(2) MOM estimate of the white noise variance is

$$\begin{aligned}\widehat{\sigma}_e^2 &= (1 - \widehat{\phi}_1 r_1 - \widehat{\phi}_2 r_2)s^2 \\ &= [1 - (0.959)(0.831) - (-0.154)(0.643)](1.783978) \approx 0.539.\end{aligned}$$

In R, fitting the AR(2) model gives

```
> ar(huron,order.max=2,AIC=F,method='yw') # method of moments
Coefficients:
      1        2
 0.9617  -0.1567
Order selected 2  sigma^2 estimated as  0.5458
```

REMARK: Note that there are minor differences in the estimates obtained "by hand" and those from using R's automated procedure. These are likely due to rounding error and/or computational errors (e.g., in solving the Yule Walker equations, etc.). It should also be noted that the R command `ar(huron,order.max=1,AIC=F,method='yw')` fits the model (via MOM) by centering all observations first about an estimate of the overall mean. This is why no "intercept" output is given.

7.3 Least squares estimation

REMARK: The MOM approach to estimation in stationary ARMA models is not always satisfactory. In fact, your authors recommend to avoid MOM estimation in any model with moving average components. We therefore consider other estimation approaches, starting with **conditional least squares** (**CLS**).

7.3.1 Autoregressive models

AR(1): Consider the stationary AR(1) model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t,$$

where note that a nonzero mean $\mu = E(Y_t)$ has been added for flexibility. For this model, the **conditional sum of squares function** is

$$S_C(\phi, \mu) = \sum_{t=2}^{n} [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2.$$

- With a sample of time series data $Y_1, Y_2, ..., Y_n$, note that the $t = 1$ term does not make sense because there is no Y_0 observation.
- The principle of least squares says to choose the values of ϕ and μ that will minimize $S_C(\phi, \mu)$.

For the AR(1) model, this amounts to solving

$$\frac{\partial S_C(\phi, \mu)}{\partial \phi} \stackrel{\text{set}}{=} 0$$
$$\frac{\partial S_C(\phi, \mu)}{\partial \mu} \stackrel{\text{set}}{=} 0$$

for ϕ and μ. This is a multivariate calculus problem and the details of its solution are shown on pp 154-155 (CC).

- In the AR(1) model, the CLS estimators are

$$\begin{aligned}\widehat{\phi} &= \frac{\sum_{t=2}^{n}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y})}{\sum_{t=2}^{n}(Y_t - \overline{Y})^2} \\ \widehat{\mu} &\approx \overline{Y}.\end{aligned}$$

- For this AR(1) model, the CLS estimator $\widehat{\phi}$ is approximately equal to r_1, the lag one sample autocorrelation (the only difference is that the denominator does not include the $t = 1$ term). We would therefore expect the difference between $\widehat{\phi}$ and r_1 (the MOM estimator) to be negligible when the sample size n is large.

- The CLS estimator $\widehat{\mu}$ is only approximately equal to the sample mean $\overline{Y}$, but the approximation should be adequate when the sample size n is large.

AR(p): In the general AR(p) model, the conditional sum of squares function is

$$S_C(\phi_1, \phi_2, ..., \phi_p, \mu) = \sum_{t=p+1}^{n} [(Y_t - \mu) - \phi_1(Y_{t-1} - \mu) - \phi_2(Y_{t-2} - \mu) - \cdots - \phi_p(Y_{t-p} - \mu)]^2,$$

a function of $p + 1$ parameters. The sum starts at $t = p + 1$ because estimates are based on the sample $Y_1, Y_2, ..., Y_n$. Despite being more complex, the CLS estimators are found in the same way, that is, ϕ_1, ϕ_2, ..., ϕ_p and μ are chosen to minimize $S_C(\phi_1, \phi_2, ..., \phi_p, \mu)$. The CLS estimator of μ is

$$\widehat{\mu} \approx \overline{Y},$$

an approximation when n is large (i.e., much larger than p). The CLS estimators for ϕ_1, ϕ_2, ..., ϕ_p are well approximated by the solutions to the sample **Yule-Walker equations**:

$$\begin{aligned} r_1 &= \phi_1 + \phi_2 r_1 + \phi_3 r_2 + \cdots + \phi_p r_{p-1} \\ r_2 &= \phi_1 r_1 + \phi_2 + \phi_3 r_1 + \cdots + \phi_p r_{p-2} \\ &\vdots \\ r_p &= \phi_1 r_{p-1} + \phi_2 r_{p-2} + \phi_3 r_{p-3} + \cdots + \phi_p. \end{aligned}$$

Therefore, in stationary AR models, the MOM and CLS estimates should be approximately equal.

7.3.2 Moving average models

MA(1): We first consider the zero mean invertible MA(1) model

$$Y_t = e_t - \theta e_{t-1},$$

where $\{e_t\}$ is a zero mean white noise process. Recall from Chapter 4 that we can rewrite an invertible MA(1) model as an infinite-order AR model; i.e.,

$$Y_t = \underbrace{-\theta Y_{t-1} - \theta^2 Y_{t-2} - \theta^3 Y_{t-3} - \cdots + e_t}_{\text{“AR}(\infty)\text{”}}.$$

Therefore, the CLS estimator of θ is the value of θ which minimizes

$$S_C(\theta) = \sum e_t^2 = \sum (Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \cdots\ )^2.$$

Unfortunately, minimizing $S_C(\theta)$ as stated is not a practical exercise, because we have only the observed sample $Y_1, Y_2, ..., Y_n$. We therefore rewrite the MA(1) model as

$$e_t = Y_t + \theta e_{t-1},$$

and take $e_0 \equiv 0$. Then, conditional on $e_0 = 0$, we can write

$$\begin{aligned} e_1 &= Y_1 \\ e_2 &= Y_2 + \theta e_1 \\ e_3 &= Y_3 + \theta e_2 \\ &\vdots \\ e_n &= Y_n + \theta e_{n-1}. \end{aligned}$$

Using these expressions for $e_1, e_2, ..., e_n$, we can now find the value of θ that minimizes

$$S_C(\theta) = \sum_{t=1}^{n} e_t^2.$$

This minimization problem can be carried out numerically, searching over a grid of θ values in $(-1, 1)$ and selecting the value of θ that produces the smallest possible $S_C(\theta)$. This minimizer is the CLS estimator of θ in the MA(1) model.

MA(q): The technique just described for MA(1) estimation via CLS can be carried out for any higher-order MA(q) model in the same fashion. When $q > 1$, the problem becomes finding the values of θ_1, θ_2, ..., θ_q such that

$$\begin{aligned} S_C(\theta_1, \theta_2, ..., \theta_q) &= \sum_{t=1}^{n} e_t^2 \\ &= \sum_{t=1}^{n} (Y_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q})^2, \end{aligned}$$

is minimized, subject to the initial conditions that $e_0 = e_{-1} = \cdots = e_{-q} = 0$. This can be done numerically, searching over all possible values of θ_1, θ_2, ..., θ_q which yield an invertible solution.

7.3.3 Mixed ARMA models

ARMA(1,1): We again consider only the zero mean ARMA(1,1) process

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1},$$

where $\{e_t\}$ is zero mean white noise. We first rewrite the model as

$$e_t = Y_t - \phi Y_{t-1} + \theta e_{t-1},$$

with the goal of minimizing

$$S_C(\phi, \theta) = \sum_{t=1}^{n} e_t^2.$$

There are now two "startup" problems, namely, specifying values for e_0 and Y_0. The authors of your text recommend avoiding specifying Y_0, taking $e_1 = 0$, and minimizing

$$S_C^*(\phi, \theta) = \sum_{t=2}^{n} e_t^2$$

with respect to ϕ and θ instead. Similar modification is recommended for ARMA models when $p > 1$ and/or when $q > 1$. See pp 157-158 (CC).

7.3.4 White noise variance

NOTE: Nothing changes with our formulae for the white noise variance estimates that we saw previously when discussing the MOM approach. The only difference is that now CLS estimates for the ϕ's and θ's are used in place of MOM estimates.

- **AR**(p):
$$\widehat{\sigma}_e^2 = (1 - \widehat{\phi}_1 r_1 - \widehat{\phi}_2 r_2 - \cdots - \widehat{\phi}_p r_p) S^2.$$
- **MA**(q):
$$\widehat{\sigma}_e^2 = \frac{S^2}{1 + \widehat{\theta}_1^2 + \widehat{\theta}_2^2 + \cdots + \widehat{\theta}_q^2}.$$
- **ARMA(1,1)**:
$$\widehat{\sigma}_e^2 = \left(\frac{1 - \widehat{\phi}^2}{1 - 2\widehat{\phi}\widehat{\theta} + \widehat{\theta}^2} \right) S^2.$$

7.3.5 Examples

Example 7.3. Data file: `gota`. The Göta River is located in western Sweden near Göteburg. The annual discharge rates (volume, measured in m^3/s) from 1807-1956 are depicted in Figure 7.4. The sample ACF and PACF are given in Figure 7.5.

- The sample ACF suggests that an **MA(1)** model
$$Y_t = \mu + e_t - \theta e_{t-1}$$
is worth considering. Note that this model includes an intercept term μ for the overall mean. Clearly, $\{Y_t\}$ is not a zero mean process.
- The sample PACF suggests that an **AR(2)** model
$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t$$
is also worth considering.

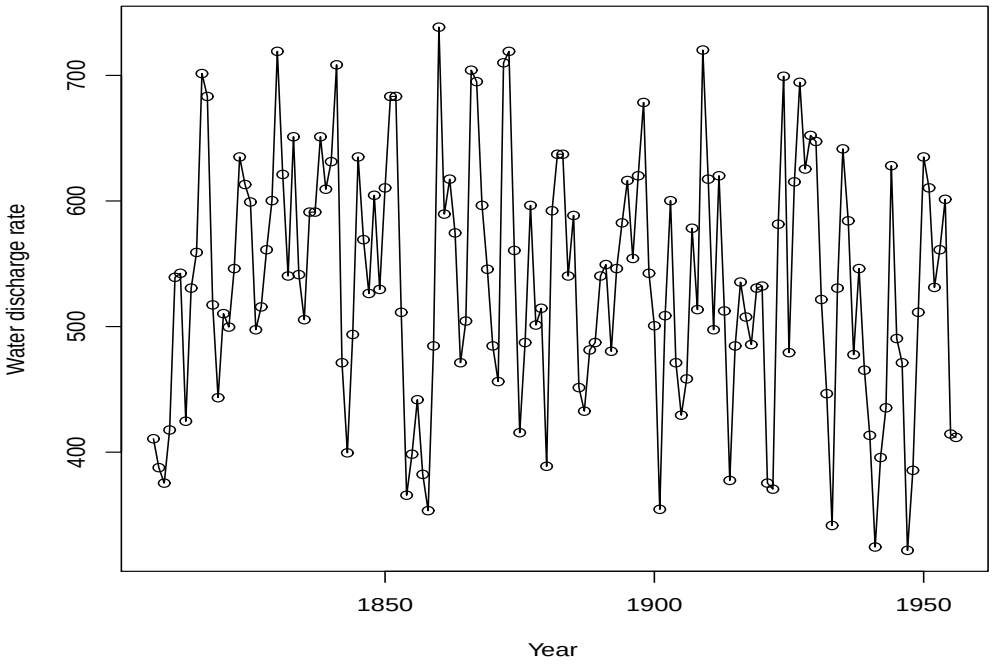


Figure 7.4: Göta River data. Water flow discharge rates (volume, measured in m^3/s) from 1807-1956.

- We will fit an MA(1) model in this example using both MOM and CLS.

MOM: I used R to compute the following: $r_1 = 0.458$, $\overline{y} = 535.4641$, and $s^2 = 9457.164$. For the Göta River discharge data, the MOM estimate of θ is

$$\widehat{\theta} = \frac{-1 + \sqrt{1 - 4r_1^2}}{2r_1} = \frac{-1 + \sqrt{1 - 4(0.458)^2}}{2(0.458)} \approx -0.654.$$

Therefore, the fitted MA(1) model for the discharge rate process is

$$Y_t = 535.4641 + e_t + 0.654e_{t-1}.$$

The white noise variance is estimated to be

$$\widehat{\sigma}_e^2 = \frac{s^2}{1 + \widehat{\theta}^2} = \frac{9457.164}{1 + (-0.654)^2} \approx 6624.$$

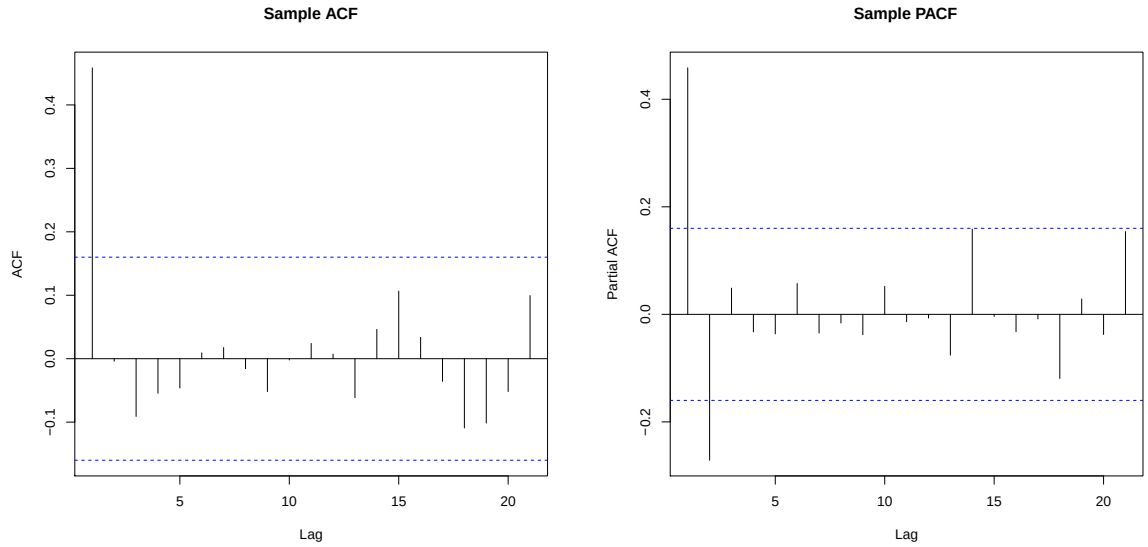


Figure 7.5: Göta River data. Left: Sample ACF. Right: Sample PACF.

CLS: Here is the R output summarizing the CLS fit:

```
> arima(gota,order=c(0,0,1),method='CSS') # conditional least squares
Coefficients:
         ma1  intercept
      0.5353   534.7199
s.e.  0.0593    10.4303
sigma^2 estimated as 6973:  part log likelihood = -876.57
```

The CLS estimates are $\widehat{\theta} = -0.5353$ (remember, R negates MA parameters/estimates) and $\widehat{\mu} = 534.7199$, which gives the fitted MA(1) model

$$Y_t = 534.7199 + e_t + 0.5353e_{t-1}.$$

The white noise variance estimate is $\widehat{\sigma}_e^2 \approx 6973$.

- The R output gives estimated **standard errors** of the CLS estimates, so we can assess their significance.
- We will learn later that CLS estimates are approximately normal in large samples.

- Therefore, an approximate 95 percent confidence interval for θ is

$$-0.5353 \pm 1.96(0.0593) \quad \Longrightarrow \quad (-0.652, -0.419).$$

We are 95 percent confident that θ is between -0.652 and -0.419. Note that this confidence interval does not include 0.

COMPARISON: It is instructive to compare the MOM and CLS estimates for the Göta River discharge data. This comparison (to 3 decimal places) is summarized below.

Method	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{\sigma}_e^2$
MOM	535.464	-0.654	6624
CLS	534.720	-0.535	6973

- The estimates for μ are very close. The MA(1) estimate is equal to $\overline{y}$ whereas the CLS estimate is only approximately equal to $\overline{y}$. See pp 155 (CC).
- The estimates for θ are not close. As previously mentioned, the MOM approach for MA models is generally not recommended.
- The estimates for σ_e^2 are notably different as well.

Example 7.4. We now revisit the Lake Huron water surface elevation data in Example 7.2 and use R to fit AR(1) and AR(2) models

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$$

and

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t,$$

respectively, using conditional least squares (CLS). Recall that in Example 7.2 we fit both the AR(1) and AR(2) models using MOM.

AR(1): Here is the R output summarizing the CLS fit:

```
> arima(huron,order=c(1,0,0),method='CSS') # conditional least squares
Coefficients:
         ar1  intercept
      0.8459   579.2788
s.e.  0.0469     0.4027
sigma^2 estimated as 0.489:  part log likelihood = -134.77
```

The fitted AR(1) model, using CLS, is

$$Y_t - 579.2788 = 0.8459(Y_{t-1} - 579.2788) + e_t,$$

or, equivalently (to 3 significant digits),

$$Y_t = 89.267 + 0.846Y_{t-1} + e_t.$$

The white noise variance estimate, using CLS, is $\widehat{\sigma}_e^2 \approx 0.489$.

AR(2): Here is the R output summarizing the CLS fit:

```
> arima(huron,order=c(2,0,0),method='CSS') # conditional least squares
Coefficients:
         ar1      ar2  intercept
      0.9874  -0.1702   579.2691
s.e.  0.0878   0.0871     0.3355
sigma^2 estimated as 0.4776:  part log likelihood = -133.27
```

The fitted AR(2) model, using CLS, is

$$Y_t - 579.2691 = 0.9874(Y_{t-1} - 579.2691) - 0.1702(Y_{t-2} - 579.2691) + e_t,$$

or, equivalently (to 3 significant digits),

$$Y_t = 105.890 + 0.987Y_{t-1} - 0.170Y_{t-2} + e_t.$$

The white noise variance estimate, using CLS, is $\widehat{\sigma}_e^2 \approx 0.4776$.

COMPARISON: It is instructive to compare the MOM and CLS estimates for the Lake Huron data. This comparison (to 3 decimal places) is summarized below.

	AR(1)			**AR(2)**			
Method	$\widehat{\mu}$	$\widehat{\phi}$	$\widehat{\sigma}^2_e$	$\widehat{\mu}$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\sigma}^2_e$
MOM	579.309	0.831	0.552	579.309	0.959	-0.154	0.539
CLS	579.279	0.846	0.489	579.269	0.987	-0.170	0.478

- Note that the MOM and CLS estimates for μ and the ϕ's are in large agreement. This is common in purely AR models (not in models with MA components).

QUESTION: For the Lake Huron data, which model is preferred: AR(1) or AR(2)?

- The $\widehat{\sigma}^2_e$ estimate is slightly smaller in the AR(2) fit, but only marginally.
- Using the CLS estimates, note that an approximate 95 percent confidence interval for ϕ_2 in the AR(2) model is

$$-0.1702 \pm 1.96(0.0871) \quad \Longrightarrow \quad (-0.341, 0.001).$$

  This interval does (barely) include 0, indicating that $\widehat{\phi}_2$ is not statistically different from 0.
- Note also that the estimated standard error of $\widehat{\phi}_1$ (in the CLS output) is almost twice as large in the AR(2) model as in the AR(1) model. **Reason:** When we fit a higher-order model, we lose precision in the other model estimates (especially if the higher-order terms are not needed).
- It is worth noting that the AR(1) model is the ARMA model identified as having the smallest BIC (using `armasubsets` in R; see Chapter 6).
- For the last three reasons, and with an interest in being parsimonious, I would pick the AR(1) if I had to choose between the two.

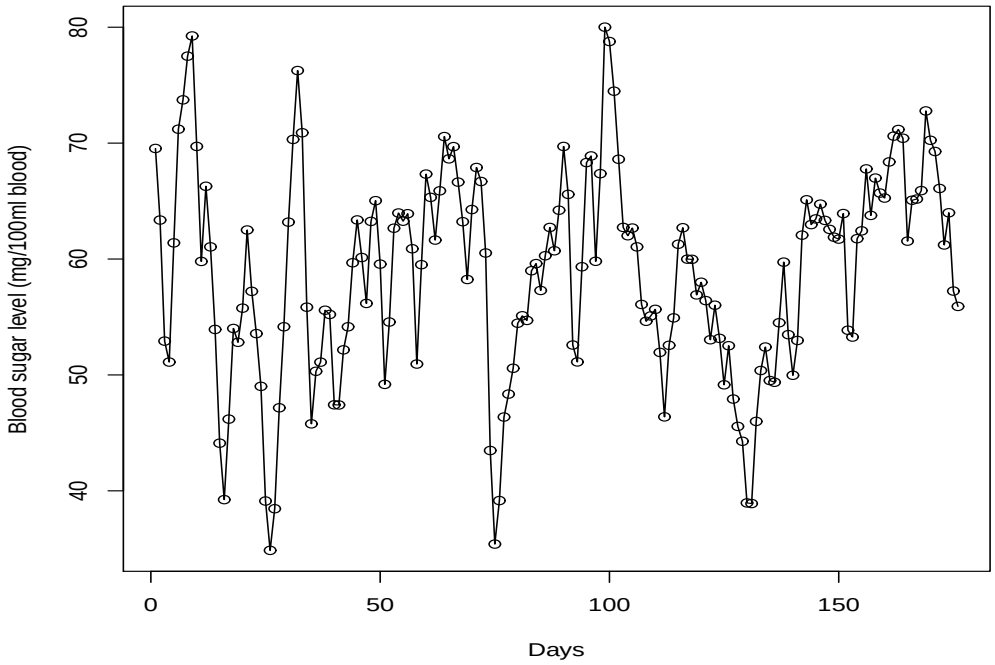


Figure 7.6: Bovine blood sugar data. Blood sugar levels (mg/100ml blood) for a single cow measured for $n = 176$ consecutive days.

Example 7.5. Data file: `cows`. The data in Figure 7.6 represent daily blood sugar concentrations (measured in mg/100ml of blood) on a single cow being dosed intermuscularly with 10 mg of dexamethasone (commonly given to increase milk production).

- The sample ACF in Figure 7.7 shows an AR-type decay, while the PACF in Figure 7.7 also shows an MA-type (oscillating) decay with "spikes" at the first three lags.
- ARMA(1,1) and AR(3) models are consistent with the sample ACF/PACF.

Consider using an ARMA(1,1) model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t - \theta e_{t-1}$$

to represent this process. Note that we have added an overall mean μ parameter in the model. Clearly, $\{Y_t\}$ is not a zero mean process. Therefore, there are three parameters to estimate and we do so using conditional least squares (CLS).

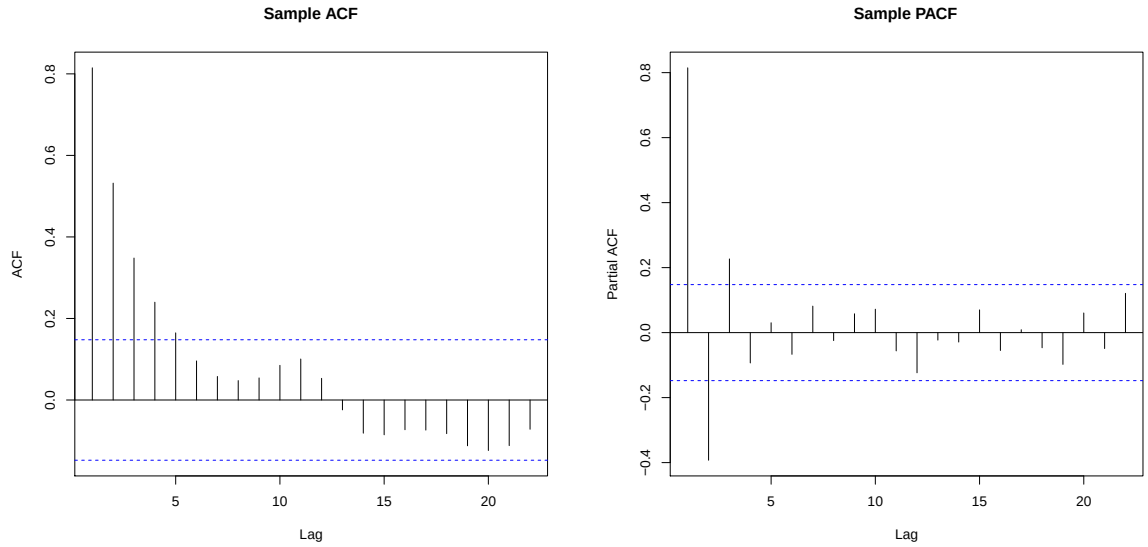


Figure 7.7: Bovine blood sugar data. Left: Sample ACF. Right: Sample PACF.

ARMA(1,1): Here is the R output summarizing the CLS fit:

```
> arima(cows,order=c(1,0,1),method='CSS') # conditional least squares
Coefficients:
         ar1     ma1  intercept
      0.6625  0.6111    58.7013
s.e.  0.0616  0.0670     1.6192
sigma^2 estimated as 20.38:  part log likelihood = -515.01
```

Therefore, the fitted ARMA(1,1) model is

$$Y_t - 58.7013 = 0.6625(Y_{t-1} - 58.7013) + e_t + 0.6111e_{t-1}$$

or, equivalently,

$$Y_t = 19.8117 + 0.6625Y_{t-1} + e_t + 0.6111e_{t-1}.$$

The white noise variance estimate, using CLS, is $\widehat{\sigma}_e^2 \approx 20.38$. From examining the (estimated) standard errors in the output, it is easy to see that both CLS estimates $\widehat{\phi} = 0.6625$ and $\widehat{\theta} = -0.6111$ are significantly different from 0.

7.4 Maximum likelihood estimation

TERMINOLOGY: The **method of maximum likelihood** is the most commonly-used technique to estimate unknown parameters (not just in time series models, but in nearly all statistical models).

- An advantage of maximum likelihood in fitting time series models is that parameter estimates are based on the entire observed sample $Y_1, Y_2, ..., Y_n$. There is no need to worry about "start up" values.
- Another advantage is that maximum likelihood estimators have very nice large-sample distributional properties. This makes statistical inference proceed in a straightforward manner.
- The main disadvantage is that we have to specify a joint probability distribution for the random variables in the sample. This makes the method more mathematical.

TERMINOLOGY: The **likelihood function** L is a function that describes the joint distribution of the data $Y_1, Y_2, ..., Y_n$. However, it is viewed as a function of the model parameters with the observed data being fixed.

- Therefore, when we maximize the likelihood function with respect to the model parameters, we are finding the values of the parameters (i.e., the estimates) that are most consistent with the observed data.

AR(1): To illustrate how maximum likelihood estimates are obtained, consider the AR(1) model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t,$$

where $\{e_t\}$ is a **normal** zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$ and where $\mu = E(Y_t)$ is the overall (process) mean. There are three parameters in this model: ϕ, μ, and σ_e^2. The probability density function (pdf) of $e_t \sim \mathcal{N}(0, \sigma_e^2)$ is

$$f(e_t) = \frac{1}{\sqrt{2\pi}\sigma_e} \exp(-e_t^2/2\sigma_e^2),$$

for all $-\infty < e_t < \infty$, where $\exp(\cdot)$ denotes the exponential function. Because $e_1, e_2, ..., e_n$ are independent, the joint pdf of $e_2, e_3, ..., e_n$ is given by

$$\begin{aligned} f(e_2, e_3, ..., e_n) = \prod_{t=2}^{n} f(e_t) &= \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}\sigma_e} \exp(-e_t^2/2\sigma_e^2) \\ &= (2\pi\sigma_e^2)^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma_e^2}\sum_{t=2}^{n} e_t^2\right). \end{aligned}$$

To write out the joint pdf of $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, we can first perform a multivariate transformation using

$$\begin{aligned} Y_2 &= \mu + \phi(Y_1 - \mu) + e_2 \\ Y_3 &= \mu + \phi(Y_2 - \mu) + e_3 \\ \vdots &= \quad \vdots \\ Y_n &= \mu + \phi(Y_{n-1} - \mu) + e_n, \end{aligned}$$

with $Y_1 = y_1$ (fixed). This will give us the (conditional) joint distribution of $Y_2, Y_3, ..., Y_n$, given $Y_1 = y_1$. Applying the laws of conditioning, the joint pdf of $\mathbf{Y}$; i.e., the likelihood function $L \equiv L(\phi, \mu, \sigma_e^2|\mathbf{y})$, is given by

$$L = L(\phi, \mu, \sigma_e^2|\mathbf{y}) = f(y_2, y_3, ..., y_n|y_1) f(y_1).$$

The details on pp 159 (CC) show that

$$\begin{aligned} f(y_2, y_3, ..., y_n|y_1) &= (2\pi\sigma_e^2)^{-(n-1)/2} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{t=2}^{n}[(y_t - \mu) - \phi(y_{t-1} - \mu)]^2\right\} \\ f(y_1) &= \left[\frac{1}{2\pi\sigma_e^2/(1-\phi^2)}\right]^{1/2} \exp\left[-\frac{(y_1 - \mu)^2}{2\sigma_e^2/(1-\phi^2)}\right]. \end{aligned}$$

Multiplying these pdfs and simplifying, we get

$$L = L(\phi, \mu, \sigma_e^2|\mathbf{y}) = (2\pi\sigma_e^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-\frac{S(\phi, \mu)}{2\sigma_e^2}\right],$$

where

$$S(\phi, \mu) = (1-\phi^2)(y_1 - \mu) + \sum_{t=2}^{n}[(y_t - \mu) - \phi(y_{t-1} - \mu)]^2.$$

For this AR(1) model, the maximum likelihood estimators (MLEs) of ϕ, μ, and σ_e^2 are the values which maximize $L(\phi, \mu, \sigma_e^2|\mathbf{y})$.

REMARK: In this AR(1) model, the function $S(\phi, \mu)$ is called the **unconditional sum-of-squares function**. Note that when $S(\phi, \mu)$ is viewed as random,

$$S(\phi, \mu) = (1 - \phi^2)(Y_1 - \mu) + S_C(\phi, \mu),$$

where $S_C(\phi, \mu)$ is the conditional sum of squares function defined in Section 7.3.1 (notes) for the same AR(1) model.

- We have already seen in Section 7.3.1 (notes) that the conditional least squares (CLS) estimates of ϕ and μ are found by minimizing $S_C(\phi, \mu)$.
- The **unconditional least squares** (**ULS**) estimates of ϕ and μ are found by minimizing $S(\phi, \mu)$. ULS is regarded as a "compromise" between CLS and the method of maximum likelihood.
- We will not pursue the ULS approach.

NOTE: The approach to finding MLEs in any stationary ARMA(p, q) model is the same as what we have just outlined in the special AR(1) case. The likelihood function L becomes more complex in larger models. However, this turns out not to be a big deal for us because we will use software to do the estimation. R can compute MLEs in any stationary ARMA(p, q) model using the `arima` function. This function also provides (estimated) standard errors of the MLEs.

DISCUSSION: We have talked about three methods of estimation: method of moments, least squares (conditional and unconditional), and maximum likelihood. Going forward, which procedure should we use? To answer this, Box, Jenkins, and Reinsel (1994) write

> "Generally, the conditional and unconditional least squares estimators serve as satisfactory approximations to the maximum likelihood estimator for large sample sizes. However, simulation evidence suggests a preference for the maximum likelihood estimator for small or moderate sample sizes, especially if the moving average operator has a root close to the boundary of the invertibility region."

7.4.1 Large-sample properties of MLEs

THEORY: Suppose that $\{e_t\}$ is a normal zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Consider a stationary ARMA(p, q) process

$$\phi(B)Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q).\end{aligned}$$

The maximum likelihood estimators $\widehat{\phi}_j$ and $\widehat{\theta}_k$ satisfy

$$\sqrt{n}(\widehat{\phi}_j - \phi_j) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\widehat{\phi}_j}), \quad \text{for } j = 1, 2, ..., p,$$

and

$$\sqrt{n}(\widehat{\theta}_k - \theta_k) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\widehat{\theta}_k}), \quad \text{for } k = 1, 2, ..., q,$$

respectively, as $n \to \infty$. In other words, for large n,

$$\begin{aligned}\widehat{\phi}_j &\sim \mathcal{AN}(\phi_j, \sigma^2_{\widehat{\phi}_j}/n) \\ \widehat{\theta}_k &\sim \mathcal{AN}(\theta_k, \sigma^2_{\widehat{\theta}_k}/n),\end{aligned}$$

for all $j = 1, 2, ..., p$ and $k = 1, 2, ..., q$. **Implication:** Maximum likelihood estimators are consistent and asymptotically normal.

SPECIFIC CASES:

- **AR(1)**.

$$\widehat{\phi} \sim \mathcal{AN}\left(\phi, \frac{1 - \phi^2}{n}\right)$$

- **AR(2)**.

$$\widehat{\phi}_1 \sim \mathcal{AN}\left(\phi_1, \frac{1 - \phi_2^2}{n}\right)$$

$$\widehat{\phi}_2 \sim \mathcal{AN}\left(\phi_2, \frac{1 - \phi_2^2}{n}\right)$$

- **MA(1)**.

$$\widehat{\theta} \sim \mathcal{AN}\left(\theta, \frac{1-\theta^2}{n}\right)$$

- **MA(2)**.

$$\widehat{\theta}_1 \sim \mathcal{AN}\left(\theta_1, \frac{1-\theta_2^2}{n}\right)$$
$$\widehat{\theta}_2 \sim \mathcal{AN}\left(\theta_2, \frac{1-\theta_2^2}{n}\right)$$

- **ARMA(1,1)**.

$$\widehat{\phi} \sim \mathcal{AN}\left[\phi, \frac{c(\phi,\theta)(1-\phi^2)}{n}\right]$$
$$\widehat{\theta} \sim \mathcal{AN}\left[\theta, \frac{c(\phi,\theta)(1-\theta^2)}{n}\right],$$

where $c(\phi,\theta) = [(1-\phi\theta)/(\phi-\theta)]^2$.

REMARK: In multi-parameter models; e.g., AR(2), MA(2), ARMA(1,1), etc., the MLEs are (asymptotically) correlated. This correlation can also be large; see pp 161 (CC) for further description.

IMPORTANT: The large-sample distributional results above make getting **large-sample confidence intervals** for ARMA model parameters easy. For example, an approximate $100(1-\alpha)$ percent confidence interval for ϕ in an **AR(1)** model is

$$\widehat{\phi} \pm z_{\alpha/2}\sqrt{\frac{1-\widehat{\phi}^2}{n}}.$$

An approximate $100(1-\alpha)$ percent confidence interval for θ in an **MA(1)** model is

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1-\widehat{\theta}^2}{n}}.$$

Note the form of these intervals. In words, the form is

"ML point estimate $\pm$ $z_{\alpha/2}$ (estimated standard error)."

- Approximate confidence intervals for the other ARMA model parameters are computed in the same way.

- The nice thing about R is that ML estimates and their (estimated) standard errors are given in the output (as they were for CLS estimates), so we have to do almost no calculation by hand.
- Furthermore, examining these confidence intervals can give us information about which estimates are statistically different from zero. This is a key part of assessing model adequacy.

NOTE: Maximum likelihood estimators (MLEs) and least-squares estimators (both CLS and ULS) have the same large-sample distributions. Large sample distributions of MOM estimators can be quite different for purely MA models (although they are the same for purely AR models). See pp 162 (CC).

7.4.2 Examples

Example 7.6. We revisit the Göta River discharge data in Example 7.3 (notes) and use R to fit an MA(1) model

$$Y_t = \mu + e_t - \theta e_{t-1},$$

using the method of maximum likelihood. Here is the output from R:

```
> arima(gota,order=c(0,0,1),method='ML') # maximum likelihood
Coefficients:
         ma1  intercept
      0.5350   535.0311
s.e.  0.0594    10.4300
sigma^2 estimated as 6957:  log likelihood = -876.58,  aic = 1757.15
```

ESTIMATES: The ML estimates are $\widehat{\theta} = -0.5350$ (remember, R negates the MA parameters/estimates) and $\widehat{\mu} = 535.0311$, which gives the fitted model

$$Y_t = 535.0311 + e_t + 0.5350 e_{t-1}.$$

The white noise variance estimate is $\widehat{\sigma}_e^2 \approx 6957$. An approximate 95 percent confidence interval for θ is

$$-0.5350 \pm 1.96(0.0594) \implies (-0.651, -0.419).$$

We are 95 percent confident that θ is between -0.651 and -0.419. This interval is almost identical to the one based on the CLS estimate; see Example 7.3.

COMPARISON: We compare the estimates from all three methods (MOM, CLS, and MLE) with the Göta River discharge data. This comparison (to 3 decimal places) is summarized below.

Method	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{\sigma}_e^2$
MOM	535.464	-0.654	6624
CLS	534.720	-0.535	6973
MLE	535.031	-0.535	6957

Note that the CLS and ML estimates of θ are identical (to three decimal places). The MOM estimate of θ is noticeably different. Recall that MOM estimation is not advised for models with MA components.

Example 7.7. The data in Figure 7.8 (left) are the number of global earthquakes annually (with intensities of 7.0 or greater) during 1900-1998. **Source:** Craig Whitlow (Spring, 2010). We examined these data in Chapter 1 (Example 1.5, pp 6).

- Because the data (number of earthquakes) are "counts," this suggests that a transformation is needed. The Box-Cox transformation output in Figure 7.8 (right) shows that $\lambda = 0.5$ resides in an approximate 95 percent confidence interval for λ. Recall that $\lambda = 0.5$ corresponds to the **square-root transformation**.
- R output for the square-root transformed series is given in Figure 7.9. The `armasubsets` output, which ranks competing ARMA models according to their BIC, selects an ARMA(1,1) model. This model is also consistent with the sample ACF and PACF.

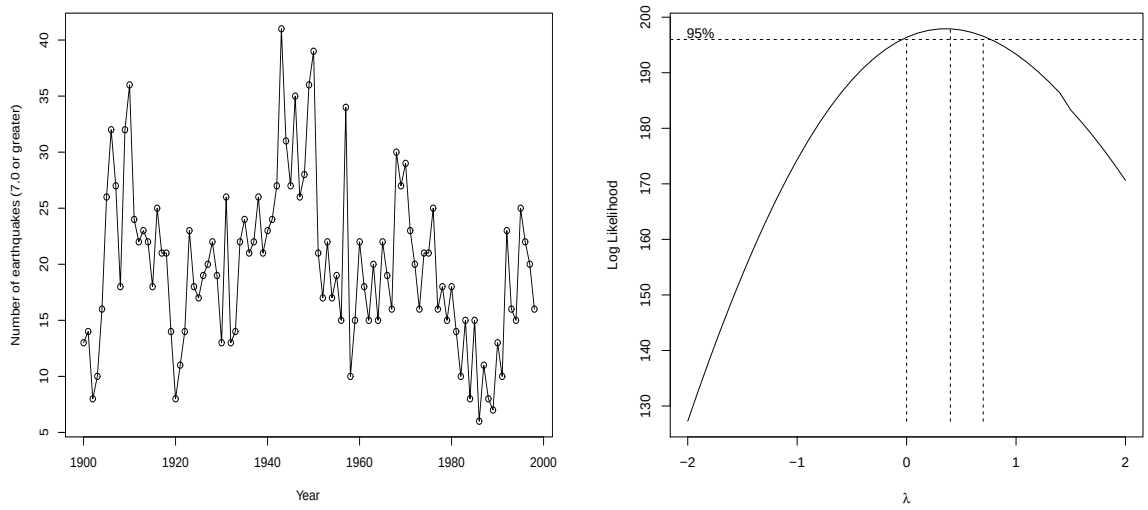


Figure 7.8: Earthquake data. Left: Number of "large" earthquakes per year from 1900-1998. Right: Box-Cox transformation output (profile log-likelihood function of λ).

- We therefore fit an ARMA(1,1) model to the $\{\sqrt{Y_t}\}$ process, that is,

$$\sqrt{Y_t} - \mu = \phi(\sqrt{Y_{t-1}} - \mu) + e_t - \theta e_{t-1}.$$

- We will use maximum likelihood. The R output is given below:

```
> arima(sqrt(earthquake),order=c(1,0,1),method='ML') # maximum likelihood
Coefficients:
         ar1      ma1  intercept
      0.8352  -0.4295     4.3591
s.e.  0.0811   0.1277     0.2196
sigma^2 estimated as 0.4294:  log likelihood = -98.88,  aic = 203.76
```

For this model, the maximum likelihood estimates based on these data are $\widehat{\phi} = 0.8352$, $\widehat{\theta} = 0.4295$, and $\widehat{\mu} = 4.3591$. The fitted model is

$$\sqrt{Y_t} - 4.3591 = 0.8352(\sqrt{Y_{t-1}} - 4.3591) + e_t - 0.4295e_{t-1}$$

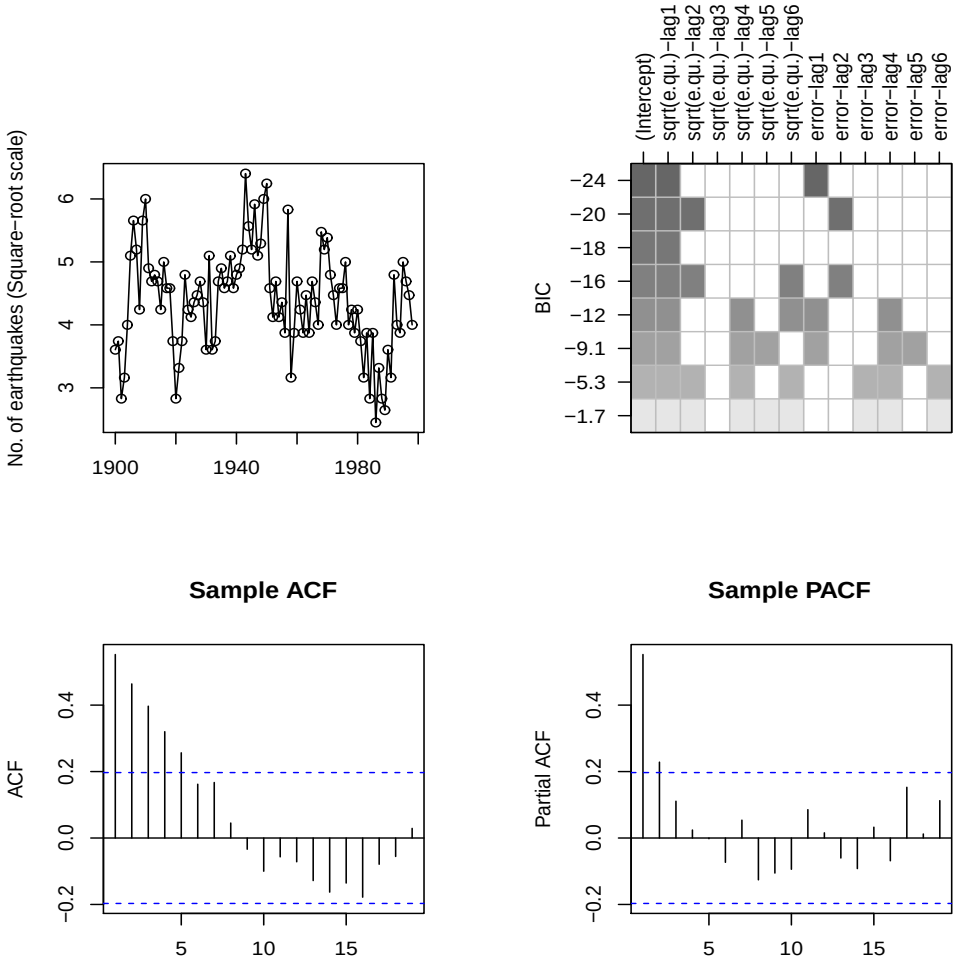


Figure 7.9: Earthquake data. Upper left: Time series plot of $\{\sqrt{Y_t}\}$ process. Upper right: `armasubsets` output (on square-root scale). Lower left: Sample ACF of $\{\sqrt{Y_t}\}$. Lower right: Sample PACF of $\{\sqrt{Y_t}\}$.

or, equivalently,

$$\sqrt{Y_t} = 0.7184 + 0.8352\sqrt{Y_{t-1}} + e_t - 0.4295e_{t-1}.$$

The white noise variance estimate, using maximum likelihood, is $\widehat{\sigma}_e^2 \approx 0.4294$. From examining the (estimated) standard errors in the output, it is easy to see that both ML estimates $\widehat{\phi} = 0.8352$ and $\widehat{\theta} = 0.4295$ are significantly different from 0. Approximate 95 percent confidence intervals for ϕ and θ, computed separately, are (0.676,0.994) and (0.179,0.680), respectively.

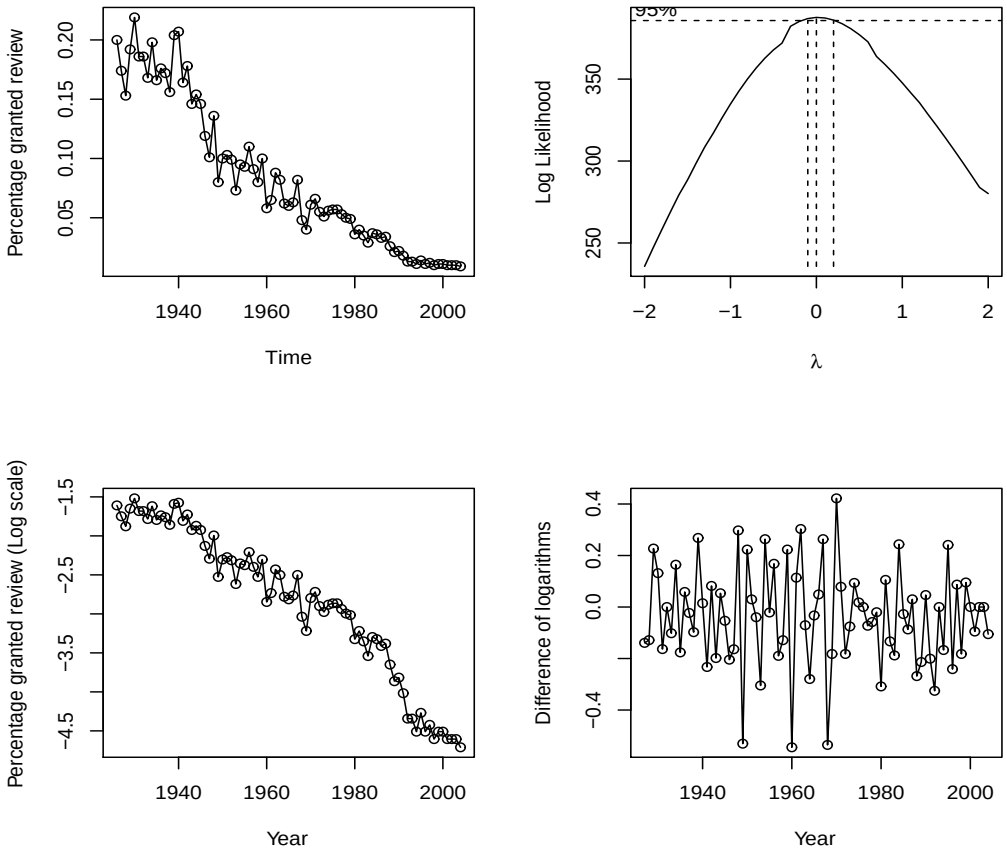


Figure 7.10: U.S. Supreme Court data. Upper left: Percent of cases granted review during 1926-2004. Upper right: Box-Cox transformation output. Lower left: Log-transformed data $\{\log Y_t\}$. Lower right: First differences of log-transformed data $\{\nabla \log Y_t\}$.

Example 7.8. The data in Figure 7.10 (upper left) represent the acceptance rate of cases appealed to the Supreme Court during 1926-2004. **Source:** Jim Manning (Spring, 2010). We examined these data in Chapter 1 (Example 1.15, pp 16).

- The time series plot suggests that this process $\{Y_t\}$ is not stationary. There is a clear linear downward trend. There is also a notable nonconstant variance problem.
- The `BoxCox.ar` transformation output in Figure 7.10 (upper right) suggests a **log-transformation** is appropriate; note that $\lambda \approx 0$.

- The log-transformed series $\{\log Y_t\}$ in Figure 7.10 (lower left) still displays the linear trend, as expected. However, the variance in the $\{\log Y_t\}$ process is more constant than in the original series. It looks like the log-transformation has "worked."

- The lower right plot in Figure 7.10 gives the first differences of the log-transformed process $\{\nabla \log Y_t\}$. This process appears to be stationary.

- The sample ACF, PACF, EACF, and `armasubsets` results (not shown) suggest that an MA(1) model for $\{\nabla \log Y_t\} \Longleftrightarrow$ an IMA(1,1) model for $\{\log Y_t\}$, that is,

$$\nabla \log Y_t = e_t - \theta e_{t-1},$$

  may be appropriate. Here is the R output from fitting this model:

```
> arima(log(supremecourt),order=c(0,1,1),method='ML') # ML
Coefficients:
          ma1
      -0.3556
s.e.   0.0941
sigma^2 estimated as 0.03408:  log likelihood = 21.04,  aic = -40.08
```

Therefore, the fitted model is

$$\nabla \log Y_t = e_t - 0.3556 e_{t-1},$$

or, equivalently,

$$\log Y_t = \log Y_{t-1} + e_t - 0.3556 e_{t-1}.$$

The white noise variance estimate, using maximum likelihood, is $\widehat{\sigma}_e^2 \approx 0.03408$. From examining the (estimated) standard error in the output, it is easy to see that the ML estimate $\widehat{\theta} = 0.3556$ is significantly different from 0.

COMMENT: Note that there is no estimated intercept term in the output above. Recall that in ARIMA(p, d, q) models with $d > 0$, intercept terms are generally not used.

8 Model Diagnostics

Complementary reading: Chapter 8 (CC).

8.1 Introduction

RECALL: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. In general, an ARIMA$(p, d, q)$ process can be written as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t,$$

where the AR and MA characteristic operators are

$$\begin{aligned} \phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \end{aligned}$$

and

$$(1-B)^d Y_t = \nabla^d Y_t$$

is the series of dth differences. Until now, we have discussed the following topics:

- **Model specification** (model selection). This deals with specifying the values of p, d, and q that are most consistent with the observed (or possibly transformed) data. This was the topic of Chapter 6.
- **Model fitting** (parameter estimation). This deals with estimating model parameters in the ARIMA(p, d, q) class. This was the topic of Chapter 7.

PREVIEW: In this chapter, we are now concerned with **model diagnostics**, which generally means that we are "checking the fit of the model." We were exposed to this topic in Chapter 3, where we encountered deterministic trend models of the form

$$Y_t = \mu_t + X_t,$$

where $E(X_t) = 0$. We apply many of the same techniques we used then to our situation now, that is, to diagnose the fit of ARIMA(p, d, q) models.

8.2 Residual analysis

TERMINOLOGY: **Residuals** are random quantities which describe the part of the variation in $\{Y_t\}$ that is not explained by the fitted model. In general, we have the general relationship (not just in time series models, but in nearly all statistical models):

$$\text{Residual}_t = \text{Observed } Y_t - \text{Predicted } Y_t.$$

Calculating residuals from an ARIMA(p, d, q) model fit based on an observed sample $Y_1, Y_2, ..., Y_n$ can be difficult. It is most straightforward with purely AR models, so we start there first.

AR(p): Consider the stationary AR(p) model

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + e_t,$$

where $\mu = E(Y_t)$ is the overall (process) mean and where $\{e_t\}$ is a zero mean white noise process. This model can be reparameterized as

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t,$$

where $\theta_0 = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$ is the **intercept term**. For this model, the residual at time t is

$$\begin{aligned} \widehat{e}_t &= Y_t - \widehat{Y}_t \\ &= Y_t - (\widehat{\theta}_0 + \widehat{\phi}_1 Y_{t-1} + \widehat{\phi}_2 Y_{t-2} + \cdots + \widehat{\phi}_p Y_{t-p}) \\ &= Y_t - \widehat{\theta}_0 - \widehat{\phi}_1 Y_{t-1} - \widehat{\phi}_2 Y_{t-2} - \cdots - \widehat{\phi}_p Y_{t-p}, \end{aligned}$$

where $\widehat{\phi}_j$ is an estimator of ϕ_j (e.g., ML, CLS, etc.), for $j = 1, 2, ..., p$, and where

$$\widehat{\theta}_0 = \widehat{\mu}(1 - \widehat{\phi}_1 - \widehat{\phi}_2 - \cdots - \widehat{\phi}_p)$$

is the estimated intercept. Therefore, once we observe the values of $Y_1, Y_2, ..., Y_n$ in our sample, we can compute the n residuals.

SUBTLETY: The first p residuals must be computed using **backcasting**, which is a mathematical technique used to "reverse predict" the unseen values of $Y_0, Y_{-1}, ..., Y_{1-p}$,

that is, the p values of the process $\{Y_t\}$ before time $t = 1$. We will not discuss backcasting in detail, but be aware that it is needed to compute early residuals in the process.

ARMA(p, q): To define residuals for an invertible ARMA model containing moving average terms, we exploit the fact that the model can be written as an inverted autoregressive process. To be specific, recall that any zero-mean invertible ARMA(p, q) model can be written as

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t,$$

where the π coefficients are functions of the ϕ and θ parameters in the specific ARMA(p, q) model. Residuals are of the form

$$\widehat{e}_t = Y_t - \widehat{\pi}_1 Y_{t-1} - \widehat{\pi}_2 Y_{t-2} - \widehat{\pi}_3 Y_{t-3} - \cdots,$$

where $\widehat{\pi}_j$ is an estimator for π_j, for $j = 1, 2, ...,$.

IMPORTANT: The observed residuals $\widehat{e}_t$ serve as "proxies" for the white noise terms e_t. We can therefore learn about the quality of the model fit by examining the residuals.

- If the model is correctly specified and our estimates are "reasonably close" to the true parameters, then the residuals should behave **roughly** like an iid normal white noise process, that is, a sequence of independent, normal random variables with zero mean and constant variance.
- If the model is not correctly specified, then the residuals will not behave roughly like an iid normal white noise process. Furthermore, examining the residuals carefully may help us identify a better model.

TERMINOLOGY: It is very common to instead work with residuals which have been standardized, that is,

$$\widehat{e}_t^* = \frac{\widehat{e}_t}{\widehat{\sigma}_e},$$

where $\widehat{\sigma}_e^2$ is an estimate of the white noise error variance σ_e^2. We call these **standardized residuals**.

- If the model is correctly specified, then the standardized residuals $\{\widehat{e}_t^*\}$, like their unstandardized counterparts, should behave **roughly** like an iid normal white noise process.
- From the standard normal distribution, we know then that most of the standardized residuals $\{\widehat{e}_t^*\}$ should fall between -3 and 3.
- Standardized residuals that fall outside this range could correspond to observations which are "outlying" in some sense; we'll make this more concrete later. If many standardized residuals fall outside $(-3, 3)$, this suggests that the error process $\{e_t\}$ has a heavy-tailed distribution (common in financial time series applications).

8.2.1 Normality and independence

DIAGNOSTICS: Histograms and qq plots of the residuals can be used to assess the normality assumption visually. Time series plots of the residuals can be helpful to detect "patterns" which violate the independence assumption.

- We can also apply the hypothesis tests for normality (Shapiro-Wilk) and independence (runs test) with the standardized residuals, just as we did in Chapter 3 with the deterministic trend models.
- The **Shapiro-Wilk test** formally tests

$$H_0 : \text{the (standardized) residuals are normally distributed}$$

versus

$$H_1 : \text{the (standardized) residuals are not normally distributed.}$$

- The **runs test** formally tests

$$H_0 : \text{the (standardized) residuals are independent}$$

versus

$$H_1 : \text{the (standardized) residuals are not independent.}$$

- For either test, small p-values lead to the rejection of H_0 in favor of H_1.

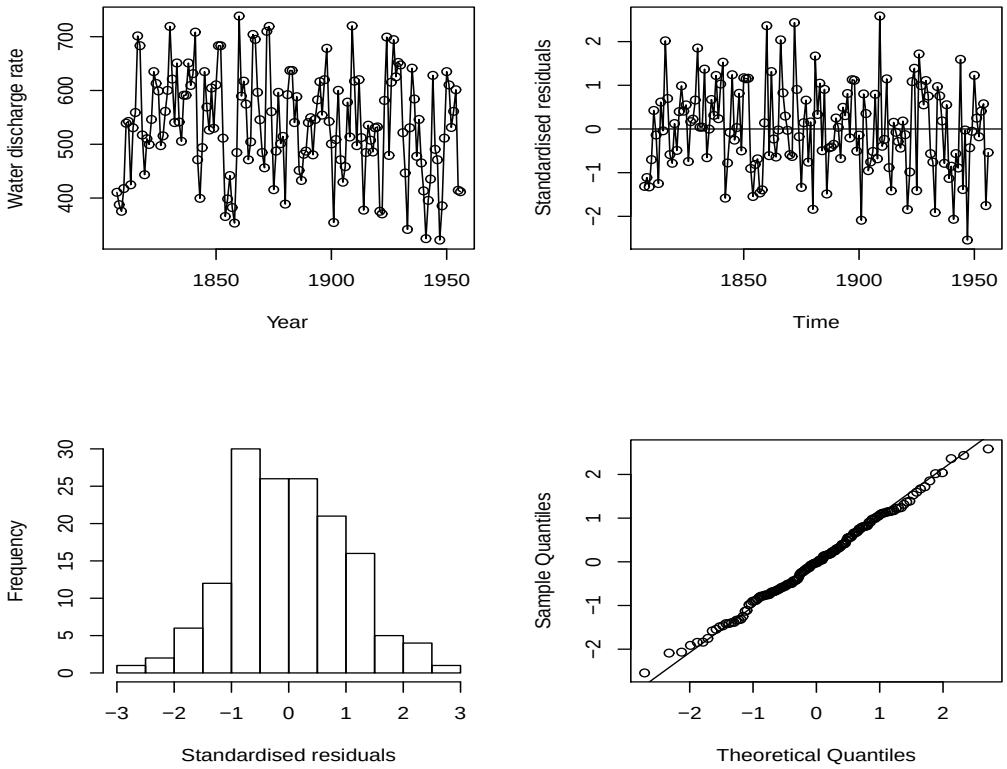


Figure 8.1: Göta River discharge data. Upper left: Discharge rate time series. Upper right: Standardized residuals from an MA(1) fit with zero line added. Lower left: Histogram of the standardized residuals from MA(1) fit. Lower right: QQ plot of the standardized residuals from MA(1) fit.

Example 8.1. In Example 7.3 (pp 189, notes), we examined the Göta River discharge rate data and used an MA(1) process to model them. The fit using maximum likelihood in Example 7.6 (pp 202, notes) was

$$Y_t = 535.0311 + e_t + 0.5350e_{t-1}.$$

Figure 8.1 displays the time series plot (upper right), the histogram (lower left), and the qq plot (lower right) of the standardized residuals. The histogram and the qq plot show no gross departures from **normality**. This observation is supported by the Shapiro-Wilk test for normality, which we perform in R. Here is the output:

```
> shapiro.test(rstandard(gota.ma1.fit))
Shapiro-Wilk normality test
W = 0.9951, p-value = 0.8975
```

The large p-value is not evidence against normality (i.e., we do not reject H_0). To examine the **independence** assumption, note that the time series of the residuals in Figure 8.1 (upper right) displays no discernible patterns and looks to be random in appearance. This observation is supported by the runs test for independence, which we also perform in R. Here is the output:

```
> runs(rstandard(gota.ma1.fit))
$pvalue
[1] 0.29
$observed.runs
[1] 69
$expected.runs
[1] 75.94667
```

Therefore, we do not have evidence against independence (i.e., we do not reject H_0).

CONCLUSION: For the Göta River discharge data, (standardized) residuals from a MA(1) fit look to reasonably satisfy the normality and independence assumptions.

Example 8.2. In Example 7.2 (pp 182, notes), we examined the Lake Huron elevation data and considered using an AR(1) process to model them. Here is the R output from fitting an AR(1) model via maximum likelihood:

```
> huron.ar1.fit = arima(huron,order=c(1,0,0),method='ML')
> huron.ar1.fit
Coefficients:
         ar1  intercept
      0.8586   579.4921
s.e.  0.0465     0.4268
sigma^2 estimated as 0.4951:  log likelihood = -136.24,  aic = 276.48
```

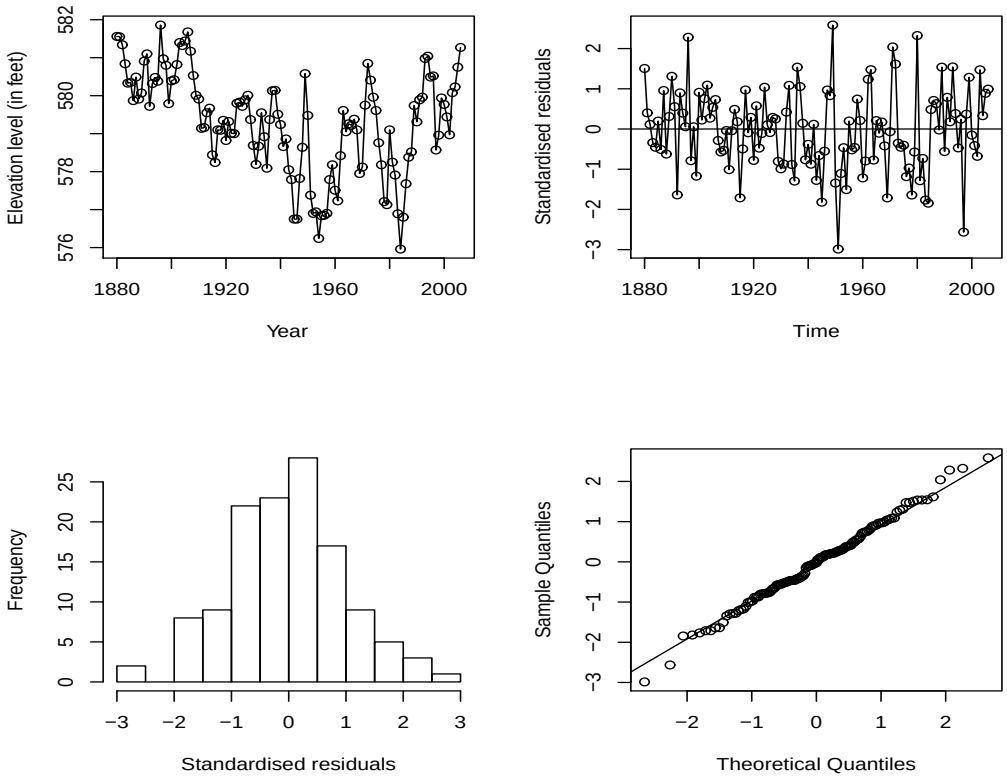


Figure 8.2: Lake Huron elevation data. Upper left: Elevation time series. Upper right: Standardized residuals from an AR(1) fit with zero line added. Lower left: Histogram of the standardized residuals from AR(1) fit. Lower right: QQ plot of the standardized residuals from AR(1) fit.

Therefore, the fitted AR(1) model is

$$Y_t - 579.4921 = 0.8586(Y_{t-1} - 579.4921) + e_t$$

or, equivalently,

$$Y_t = 81.9402 + 0.8586Y_{t-1} + e_t.$$

Figure 8.2 displays the time series plot (upper right), the histogram (lower left), and the qq plot (lower right) of the standardized residuals. The histogram and the qq plot show no gross departures from normality. The time series plot of the standardized residuals displays no noticeable patterns and looks like a stationary random process.

The R output for the Shapiro-Wilk and runs tests is given below:

```
> shapiro.test(rstandard(huron.ar1.fit))
Shapiro-Wilk normality test
W = 0.9946, p-value = 0.9156

> runs(rstandard(huron.ar1.fit))
$pvalue
[1] 0.373
$observed.runs
[1] 59
$expected.runs
[1] 64.49606
```

CONCLUSION: For the Lake Huron elevation data, (standardized) residuals from a AR(1) fit look to reasonably satisfy the normality and independence assumptions.

8.2.2 Residual ACF

RECALL: In Chapter 6, we discovered that for a **white noise process**, the sample autocorrelation satisfies

$$r_k \sim \mathcal{AN}\left(0, \frac{1}{n}\right),$$

for large n. Furthermore, the sample autocorrelations r_j and r_k, for $j \neq k$, are **approximately uncorrelated**.

- Therefore, to further check the adequacy of a fitted ARIMA(p, d, q) model, it is a good idea to examine the sample autocorrelation function (ACF) of the residuals.
- To separate our discussion in Chapter 6 from now, we will denote

  $$\widehat{r}_k = k\text{th sample autocorrelation of the residuals } \widehat{e}_t,$$

  for $k = 1, 2, ...,$. That is, the "hat" symbol in $\widehat{r}_k$ will remind us that we are now dealing with residuals.

- We remarked earlier in this chapter that

  > "If the model is correctly specified and our estimates are "reasonably close" to the true parameters, then the residuals should behave **roughly** like an iid normal white noise process."

- We say "roughly," because even if the correct model is fit, the sample autocorrelations of the residuals, $\widehat{r}_k$, have sampling distributions that are a little different than that of white noise (most prominently at early lags).

- In addition, $\widehat{r}_j$ and $\widehat{r}_k$, for $j \neq k$, are correlated, notably so at early lags and more weakly at later lags.

RESULTS: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. In addition, suppose that we have identified and fit the correct ARIMA(p, d, q) model

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t$$

using maximum likelihood. All of the following are large-sample results (i.e., they are approximate for large n).

- **MA(1).**

$$\begin{aligned}
\text{var}(\widehat{r}_1) &\approx \frac{\theta^2}{n} \\
\text{var}(\widehat{r}_k) &\approx \frac{1-(1-\theta^2)\theta^{2k-2}}{n}, \quad \text{for } k > 1 \\
\text{corr}(\widehat{r}_1, \widehat{r}_k) &\approx -\text{sign}(\theta)\left[\frac{(1-\theta^2)\theta^{k-2}}{1-(1-\theta^2)\theta^{2k-2}}\right], \quad \text{for } k > 1,
\end{aligned}$$

  where $\text{sign}(\theta) = 1$, if $\theta > 0$ and $\text{sign}(\theta) = -1$, if $\theta < 0$.

- **MA(2).**

$$\begin{aligned}
\text{var}(\widehat{r}_1) &\approx \frac{\theta_2^2}{n} \\
\text{var}(\widehat{r}_2) &\approx \frac{\theta_2^2 + \theta_1^2(1+\theta_2)^2}{n} \\
\text{var}(\widehat{r}_k) &\approx \frac{1}{n}, \quad \text{for } k > 2.
\end{aligned}$$

- **AR(1).**

$$\begin{aligned}
\text{var}(\widehat{r}_1) &\approx \frac{\phi^2}{n} \\
\text{var}(\widehat{r}_k) &\approx \frac{1-(1-\phi^2)\phi^{2k-2}}{n}, \quad \text{for } k > 1 \\
\text{corr}(\widehat{r}_1, \widehat{r}_k) &\approx -\text{sign}(\phi)\left[\frac{(1-\phi^2)\phi^{k-2}}{1-(1-\phi^2)\phi^{2k-2}}\right], \quad \text{for } k > 1.
\end{aligned}$$

- **AR(2).**

$$\begin{aligned}
\text{var}(\widehat{r}_1) &\approx \frac{\phi_2^2}{n} \\
\text{var}(\widehat{r}_2) &\approx \frac{\phi_2^2 + \phi_1^2(1+\phi_2)^2}{n} \\
\text{var}(\widehat{r}_k) &\approx \frac{1}{n}, \quad \text{for } k > 2.
\end{aligned}$$

NOTE: The MA(2) result that $\text{var}(\widehat{r}_k) \approx 1/n$, for $k > 2$, may not hold if (θ_1, θ_2) is "close" to the boundary of the invertibility region for the MA(2) model. The same is true for the AR(2) if (ϕ_1, ϕ_2) is "close" to the boundary of the stationarity region.

MAIN POINT: Even if we fit the correct ARIMA(p, d, q) model, the residuals from the fit will not follow a white noise process exactly. At very early lags, there are noticeable differences from a white noise process. For larger lags, the differences become negligible.

Example 8.3. In Example 8.1, we examined the residuals from an MA(1) fit to the Göta River discharge data (via ML).

- The sample ACF of the MA(1) residuals is depicted in Figure 8.3 with margin of error bounds at

$$\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{150}} \approx 0.163.$$

  That is, the margin of error bounds in Figure 8.3 are computed under the **white noise assumption**.

- In this example, we calculate estimates of $\text{var}(\widehat{r}_k)$, for $k = 1, 2, ..., 10$.

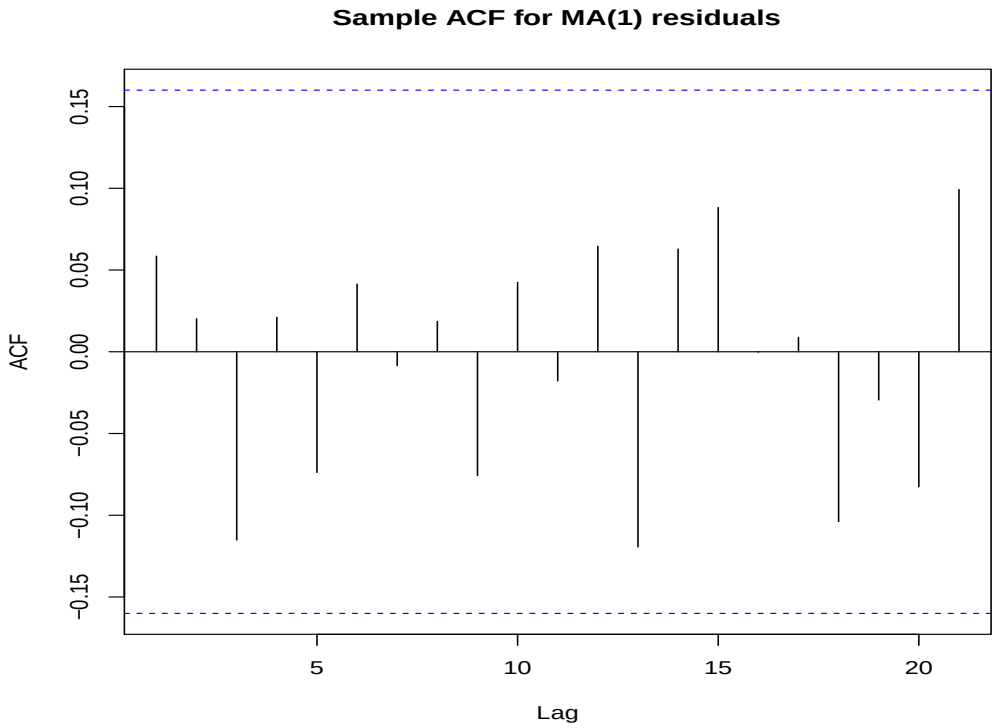


Figure 8.3: Göta River discharge data. Sample ACF of the residuals from an MA(1) model fit.

- For an MA(1) model fit,

$$\begin{aligned} \widehat{\text{var}}(\widehat{r}_1) &\approx \frac{\widehat{\theta}^2}{n} \\ \widehat{\text{var}}(\widehat{r}_k) &\approx \frac{1-(1-\widehat{\theta}^2)\widehat{\theta}^{2k-2}}{n}, \quad \text{for } k > 1. \end{aligned}$$

  Note that in these formulae, $\widehat{\theta}$ replaces θ making these **estimates** of the true variances stated earlier.

Recall that the MA(1) model fit to these data (via ML) was

$$Y_t = 535.0311 + e_t + 0.5350e_{t-1}.$$

so that $\widehat{\theta} = -0.5350$. Therefore,

$$\begin{aligned} \widehat{\text{var}}(\widehat{r}_1) &\approx \frac{(-0.5350)^2}{150} \approx 0.001908 \\ \widehat{\text{var}}(\widehat{r}_k) &\approx \frac{1-[1-(-0.5350)^2](-0.5350)^{2k-2}}{150}, \quad \text{for } k > 1. \end{aligned}$$

Here are first 10 sample autocorrelations for the residuals from the MA(1) fit:

```
> acf(residuals(gota.ma1.fit),plot=F,lag.max=10)
     1      2      3      4      5      6      7      8      9     10
 0.059  0.020 -0.115  0.021 -0.074  0.041 -0.009  0.019 -0.076  0.042
```

We now construct a table which displays these sample autocorrelations, along with their ± 2 estimated standard errors

$$\pm 2\widehat{\text{se}}(\widehat{r}_k) = \pm 2\sqrt{\widehat{\text{var}}(\widehat{r}_k)},$$

for $k = 1, 2, ..., 10$. Values of $\widehat{r}_k$ more than 2 (estimated) standard errors away from 0 would be considered inconsistent with the fitted model.

k	1	2	3	4	5	6	7	8	9	10
$\widehat{r}_k$	0.059	0.020	-0.115	0.021	-0.074	0.041	-0.009	0.019	-0.076	0.042
$2\widehat{\text{se}}(\widehat{r}_k)$	0.087	0.146	0.158	0.162	0.163	0.163	0.163	0.163	0.163	0.163

- Note that as k gets larger, $2\widehat{\text{se}}(\widehat{r}_k)$ approaches

$$\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{150}} \approx 0.163$$

  the white noise margin of error bounds.

- None of the sample autocorrelations fall outside the $\pm 2\widehat{\text{se}}(\widehat{r}_k)$ bounds.

- This finding further supports the MA(1) model choice for these data.

REMARK: In addition to examining the sample autocorrelations of the residuals individually, it is useful to consider them as a group.

- Although sample autocorrelations may be moderate individually; e.g., each within the $\pm 2\widehat{\text{se}}(\widehat{r}_k)$ bounds, it could be that as a group the sample autocorrelations are "excessive," and therefore inconsistent with the fitted model.

- To address this potential occurrence, Ljung and Box (1978) developed a procedure, based on the sample autocorrelations of the residuals, to test formally whether or not a certain model in the ARMA(p, q) family was appropriate.

LJUNG-BOX TEST: In particular, the **modified Ljung-Box** test statistic

$$Q_* = n(n+2)\sum_{k=1}^{K}\frac{\widehat{r}_k^2}{n-k}$$

can be used to test

$$H_0 : \text{the ARMA}(p, q) \text{ model is appropriate}$$

versus

$$H_1 : \text{the ARMA}(p, q) \text{ model is not appropriate.}$$

- The sample autocorrelations $\widehat{r}_k$, for $k = 1, 2, ..., K$, are computed under the ARMA(p, q) model assumption in H_0. If a nonstationary model is fit ($d > 0$), then the ARMA(p, q) model refers to the suitably differenced process.
- The value K is called the **maximum lag**; it's choice is somewhat arbitrary.
- Somewhat diaphanously, the authors of your text recommend that K be chosen so that the Ψ_j weights of the general linear process representation of the ARMA(p, q) model (under H_0) are negligible for all $j > K$. Recall that any stationary ARMA(p, q) process can be written as

  $$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \cdots,$$

  where $\{e_t\}$ is a zero mean white noise process.
- Typically one can simply compute Q_* for various choices of K and determine if the same decision is reached for all values of K.
- For a fixed K, a level α decision rule is to reject H_0 if the value of Q_* exceeds the upper α quantile of the χ^2 distribution with $K - p - q$ degrees of freedom, that is,

  $$\text{Reject } H_0 \text{ if } \; Q_* > \chi^2_{K-p-q,\alpha}.$$

- Fitting an erroneous model tends to inflate Q_*, so this is a one sided test. R produces p-values for this test automatically.
- The `tsdiag` function in R will compute Q_* at all lags specified by the user.

Example 8.4. In Example 8.1, we examined the residuals from an MA(1) fit to the Göta River discharge data (via ML). Here we illustrate the use of the modified Ljung-Box test for the MA(1) model. Recall that we computed the first 10 sample autocorrelations:

```
> acf(residuals(gota.ma1.fit),plot=F,lag.max=10)
     1      2      3      4      5      6      7      8      9     10
 0.059  0.020 -0.115  0.021 -0.074  0.041 -0.009  0.019 -0.076  0.042
```

Taking $K = 10$ and $n = 150$, the modified Ljung-Box statistic is

$$\begin{aligned} Q_* &= 150(150+2)\left[\frac{(0.059)^2}{150-1} + \frac{(0.020)^2}{150-2} + \cdots + \frac{(0.042)^2}{150-10}\right] \\ &\approx 5.13. \end{aligned}$$

To test MA(1) model adequacy, we compare Q_* to the upper α quantile of a χ^2 distribution with $K - p - q = 10 - 0 - 1 = 9$ degrees of freedom and reject the MA(1) model if Q_* exceeds this quantile. With $\alpha = 0.05$,

$$\chi^2_{9,0.05} = 16.91898,$$

which I found using the `qchisq(0.95,9)` command in R. Because the test statistic Q_* does not exceed this upper quantile, we do not reject H_0.

REMARK: Note that R can perform the modified Ljung-Box test automatically. Here is the output:

```
> Box.test(residuals(gota.ma1.fit),lag=10,type="Ljung-Box",fitdf=1)
Box-Ljung test
X-squared = 5.1305, df = 9, p-value = 0.8228
```

We do not have evidence against MA(1) model adequacy for these data when $K = 10$.

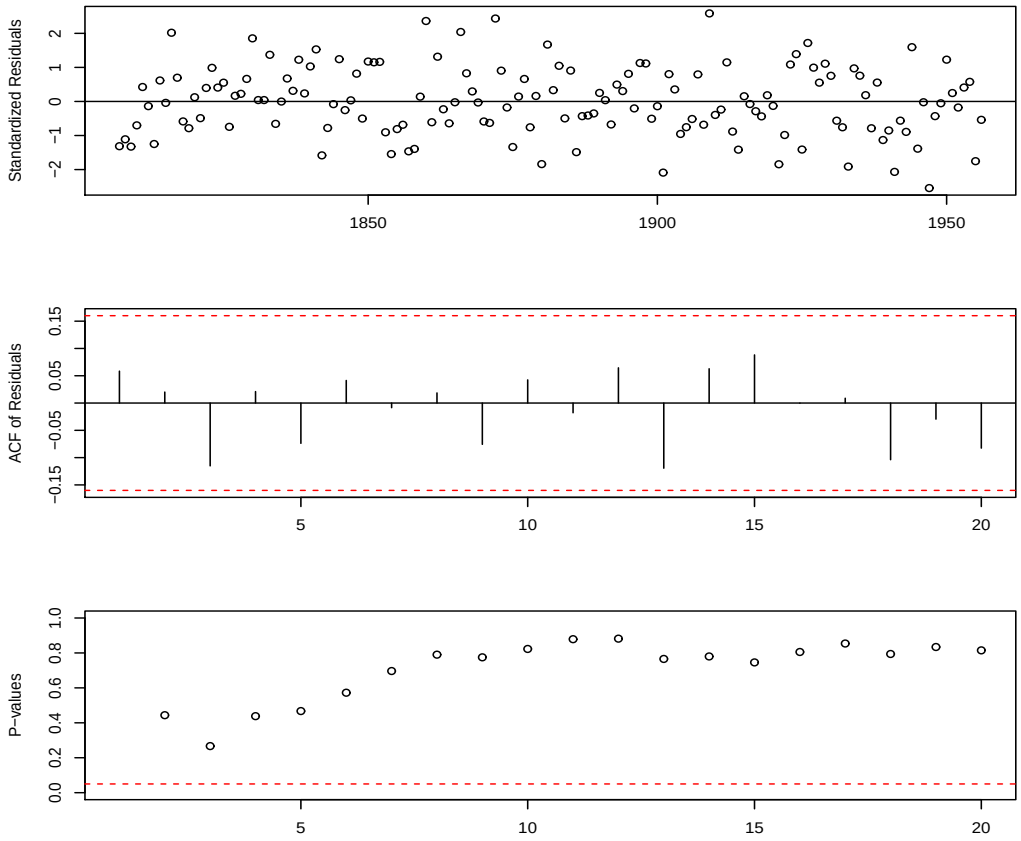


Figure 8.4: Göta River discharge data. Residual graphics and modified Ljung-Box p-values for MA(1) fit. This figure was created using the `tsdiag` function in R.

GRAPHICS: The R function `tsdiag` produces the plot in Figure 8.4.

- The top plot displays the residuals plotted through time (without connecting lines).
- The middle plot displays the sample ACF of the residuals.
- The bottom plot displays the p-values of the modified Ljung-Box test for various values of K. A horizontal line at $\alpha = 0.05$ is added.

For the Göta River discharge data, we see in Figure 8.4 that all of the modified Ljung-Box test p-values are larger than 0.05, lending further support of the MA(1) model.

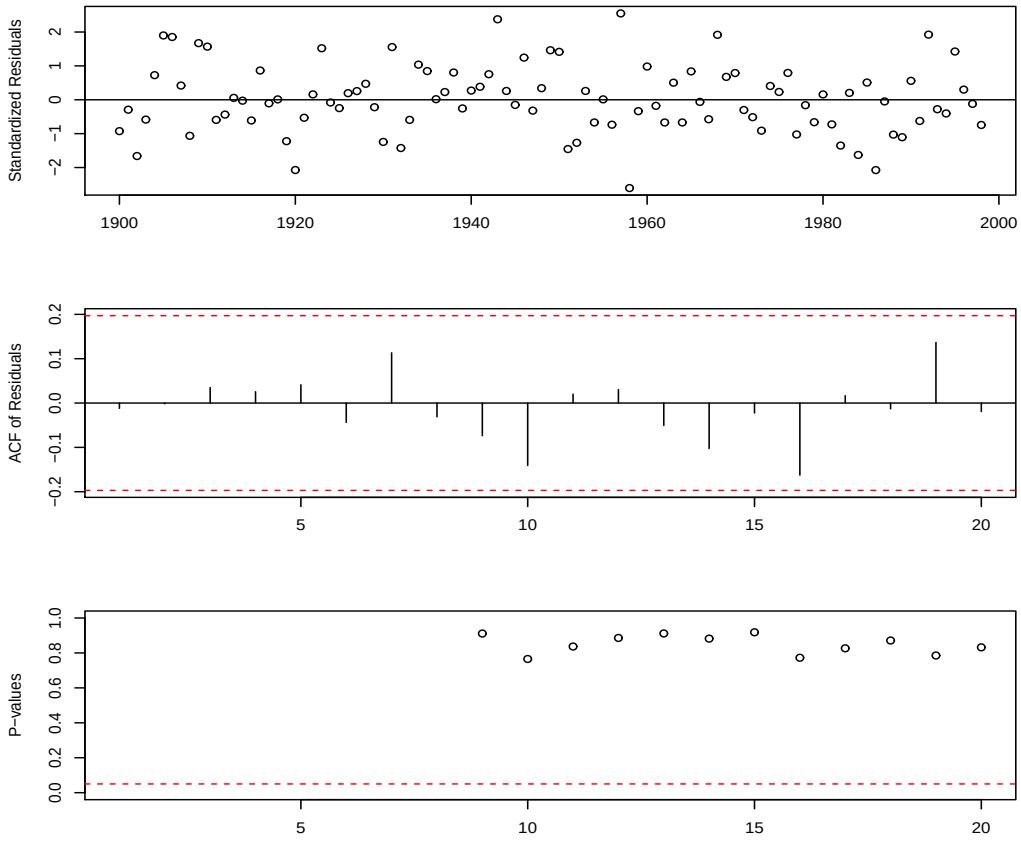


Figure 8.5: Earthquake data. Residual graphics and modified Ljung-Box p-values for ARMA(1,1) fit to the square-root transformed data.

Example 8.5. In Example 7.7 (pp 203, notes), we fit an ARMA(1,1) model to the (square-root transformed) earthquake data using maximum likelihood. Figure 8.5 displays the `tsdiag` output for the ARMA(1,1) model fit.

- The Shapiro-Wilk test does not reject normality (p-value = 0.7202). The runs test does not reject independence (p-value = 0.679). Both the Shapiro-Wilk and runs tests were applied to the standardized residuals.
- The residual output in Figure 8.5 fully supports the ARMA(1,1) model.

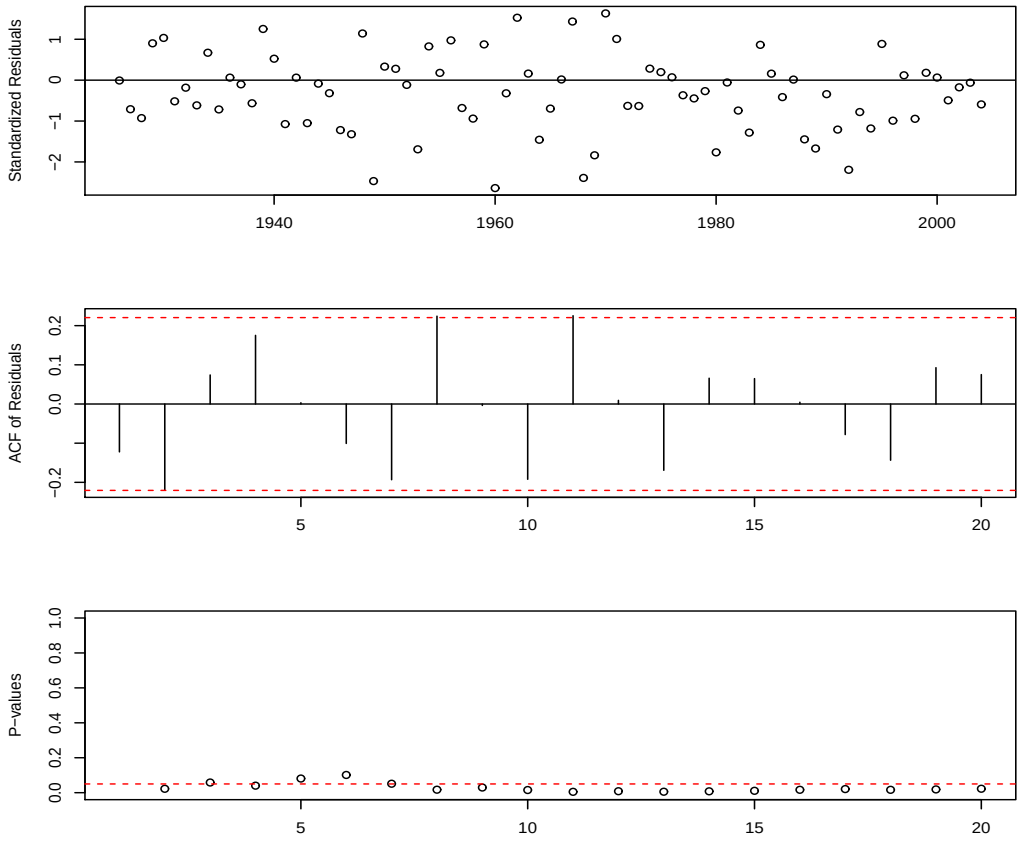


Figure 8.6: U.S. Supreme Court data. Residual graphics and modified Ljung-Box p-values for IMA(1,1) fit to the log transformed data.

Example 8.6. In Example 7.8 (pp 206, notes), we fit an IMA(1,1) model to the (log transformed) Supreme Court data using maximum likelihood. Figure 8.6 displays the `tsdiag` output for the IMA(1,1) model fit.

- The Shapiro-Wilk test does not reject normality (p-value = 0.5638). The runs test does not reject independence (p-value = 0.864). Both the Shapiro-Wilk and runs tests were applied to the standardized residuals.
- The modified Ljung-Box test p-values in Figure 8.6 raise serious concerns over the adequacy of the IMA(1,1) model fit.

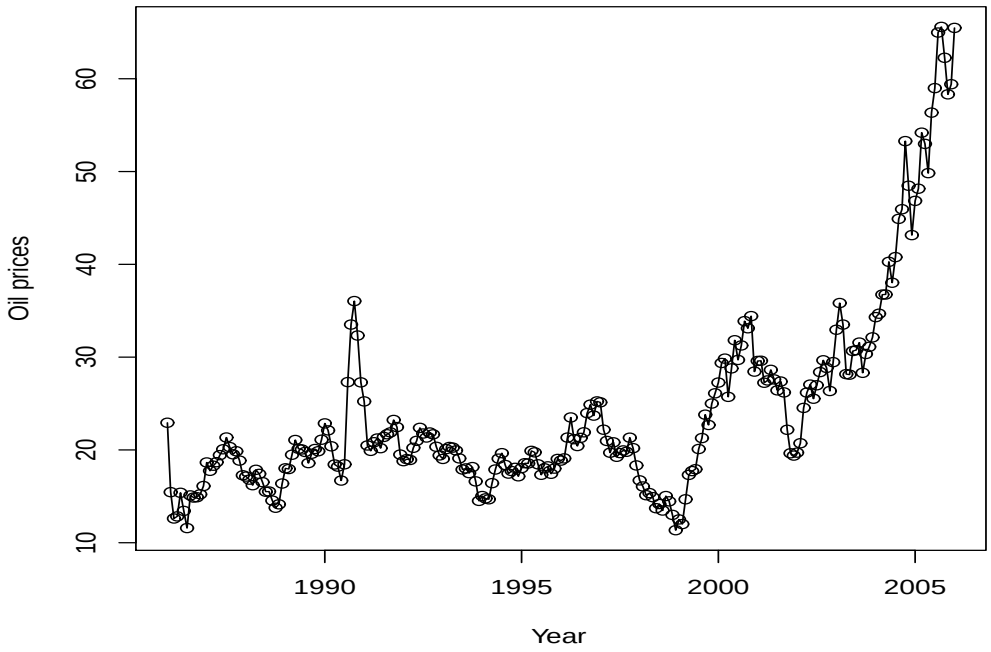


Figure 8.7: Crude oil price data. Monthly spot prices in dollars from Cushing, OK, from 1/1986 to 1/2006.

Example 8.7. The data in Figure 8.7 are monthly spot prices for crude oil (measured in U.S. dollars per barrel). We examined these data in Chapter 1 (Example 1.12, pp 13). In this example, we assess the fit of an IMA(1,1) model for $\{\log Y_t\}$; i.e.,

$$\nabla \log Y_t = e_t - \theta e_{t-1}.$$

I have arrived at this candidate model using our established techniques from Chapter 6; these details are omitted for brevity. I used maximum likelihood to fit the model.

- In Figure 8.8, we display the $\{\nabla \log Y_t\}$ process (upper left), along with plots of the standardized residuals from the IMA(1,1) fit.
- It is difficult to notice a pattern in the time series plot of the residuals, although there are notable **outliers** on the low and high sides.

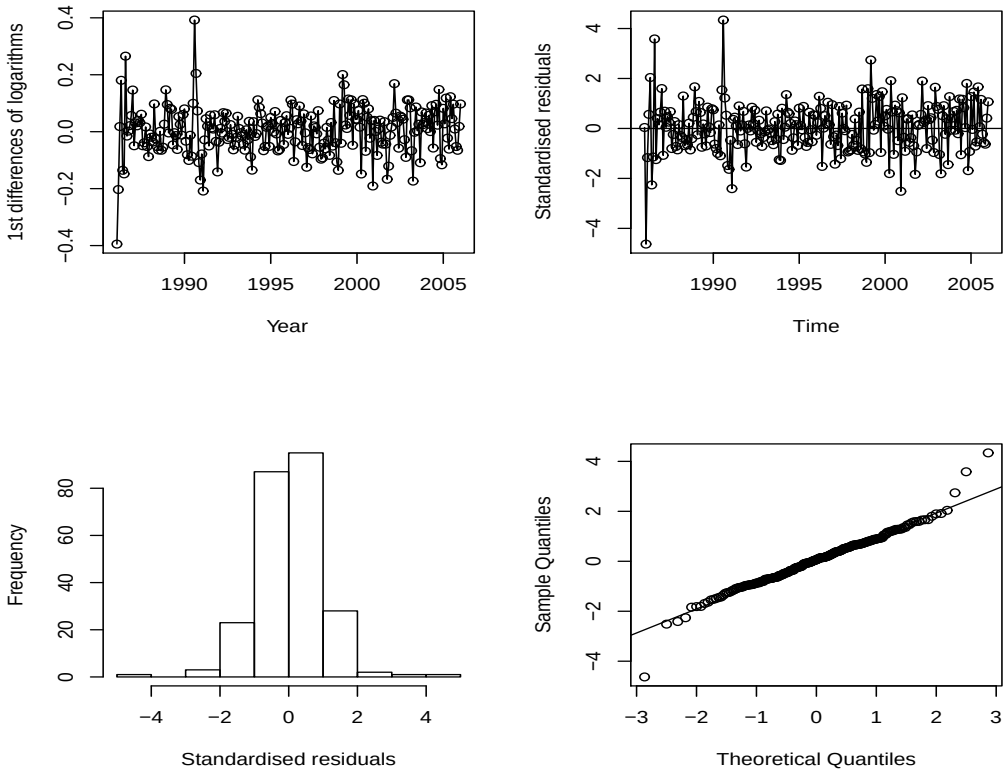


Figure 8.8: Oil price data with IMA(1,1) fit to $\{\log Y_t\}$. Upper left: $\{\nabla \log Y_t\}$ process. Upper right: Standardized residuals with zero line added. Lower left: Histogram of the standardized residuals. Lower right: QQ plot of the standardized residuals.

- The Shapiro-Wilk test strongly rejects normality of the residuals (p-value < 0.0001). This is likely due to the extreme outliers on each side, which are not "expected" under the assumption of normality. The runs test does not reject independence (p-value $= 0.341$).
- The `tsdiag` output for the IMA(1,1) residuals is given in Figure 8.9. The top plot displays the residuals from the IMA(1,1) fit with "outlier limits" at

$$z_{0.025/241} \approx 3.709744,$$

which is the upper $1 - 0.05/2(241)$ quantile of the $\mathcal{N}(0, 1)$ distribution.

- R is implementing a "Bonferroni" correction to test each residual as an **outlier**.

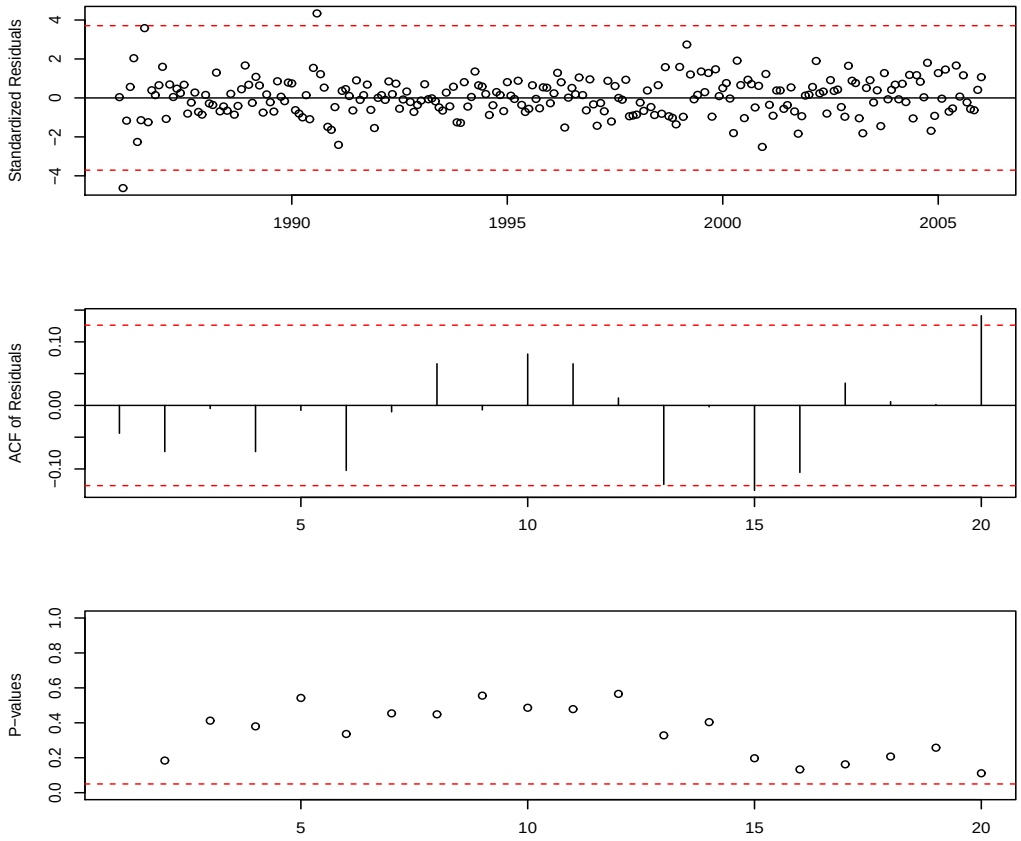


Figure 8.9: Oil price data. Residual graphics and modified Ljung-Box p-values for IMA(1,1) fit to the log transformed data.

- According to the Bonferroni criterion, residuals which exceed this value (3.709744) in absolute value would be classified as outliers. The one around 1991 likely corresponds to the U.S. invasion of Iraq (the first one).
- The sample ACF for the residuals raises concern, but the modified Ljung-Box p-values do not suggest lack of fit (although it becomes interesting for large K).
- The IMA(1,1) model for the log-transformed data appears to do a fairly good job. I am a little concerned about the outliers and the residual ACF. **Intervention analysis** (Chapter 11) may help to adjust for the outlying observations.

8.3 Overfitting

REMARK: In addition to performing a thorough residual analysis, **overfitting** can be a useful diagnostic technique to further assess the validity of an assumed model. Basically, "overfitting" refers to the process of a fitting a model more complicated than the one under investigation and then

(a) examining the **significance** of the additional parameter estimates

(b) examining the **change** in the estimates from the assumed model.

EXAMPLE: Suppose that, after the model specification phase and residual diagnostics, we are strongly considering an AR(2) model for our data, that is,

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t.$$

To perform an overfit, we would fit the following two models:

- **AR(3):**
$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + e_t$$
- **ARMA(2,1):**
$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t - \theta e_{t-1}.$$
- If the additional AR parameter estimate $\widehat{\phi}_3$ is significantly different than zero, then this would be evidence that an AR(3) model is worthy of investigation. If $\widehat{\phi}_3$ is **not** significantly different than zero and the estimates of ϕ_1 and ϕ_2 do not change much from their values in the AR(2) model fit, this would be evidence that the more complicated AR(3) model is not needed.
- If the additional MA parameter estimate $\widehat{\theta}$ is significantly different than zero, then this would be evidence that an ARMA(2,1) model is worthy of investigation. If $\widehat{\theta}$ is not significantly different than zero and the estimates of ϕ_1 and ϕ_2 do not change much from their values in the AR(2) model fit, this would be evidence that the more complicated ARMA(2,1) model is not needed.

IMPORTANT: When **overfitting** an ARIMA(p, d, q) model, we consider the following two models:

(a) ARIMA$(p+1, d, q)$

(b) ARIMA$(p, d, q+1)$.

That is, one overfit model increases p by 1. The other increases q by 1.

Example 8.8. Our residual analysis this chapter suggests that an MA(1) model for the Göta River discharge data is very reasonable. We now overfit using an MA(2) model and an ARMA(1,1) model. Here is the R output from all three model fits:

```
> gota.ma1.fit
Call: arima(x = gota, order = c(0, 0, 1), method = "ML")
Coefficients:
         ma1  intercept
      0.5350   535.0311
s.e.  0.0594    10.4300
sigma^2 estimated as 6957:  log likelihood = -876.58,  aic = 1757.15

> gota.ma2.overfit
Call: arima(x = gota, order = c(0, 0, 2), method = "ML")
Coefficients:
         ma1     ma2  intercept
      0.6153  0.1198   534.8117
s.e.  0.0861  0.0843    11.7000
sigma^2 estimated as 6864:  log likelihood = -875.59,  aic = 1757.18

> gota.arma11.overfit
Call: arima(x = gota, order = c(1, 0, 1), method = "ML")
Coefficients:
         ar1     ma1  intercept
      0.1574  0.4367   534.8004
s.e.  0.1292  0.1100    11.5217
sigma^2 estimated as 6891:  log likelihood = -875.87,  aic = 1757.74
```

ANALYSIS: In the MA(2) overfit, we see that a 95 percent confidence interval for θ_2, the additional MA model parameter, is

$$-0.1198 \pm 1.96(0.0843) \quad \Longrightarrow \quad (-0.285, 0.045),$$

which does include 0. Therefore, $\widehat{\theta}_2$ is not statistically different than zero, which suggests that the MA(2) model is not necessary. In the ARMA(1,1) overfit, we see that a 95 percent confidence interval for ϕ, the additional AR model parameter, is

$$0.1574 \pm 1.96(0.1292) \quad \Longrightarrow \quad (-0.096, 0.411),$$

which also includes 0. Therefore, $\widehat{\phi}$ is not statistically different than zero, which suggests that the ARMA(1,1) model is not necessary. The following table summarizes the output on the last page:

Model	$\widehat{\theta}$ ($\widehat{\text{se}}$)	Additional estimate	Significant?	$\widehat{\sigma}_e^2$	AIC
MA(1)	0.5350(0.0594)	$--$	$--$	6957	1757.15
MA(2)	0.6153(0.0861)	$\widehat{\theta}_2$	no	6864	1757.18
ARMA(1,1)	0.4367(0.1100)	$\widehat{\phi}$	no	6891	1757.74

Because the additional estimates in the overfit models are not statistically different from zero, there is no reason to further consider either model. Note also how the estimate of θ becomes less precise in the two larger models.

DISCUSSION: We have finished our discussions on model specification, model fitting, and model diagnostics. Having done so, you are now well-versed in modeling time series data in the ARIMA(p, d, q) family. Hopefully, you have realized that the process of building a model is not always clear cut and that some "give and take" is necessary. Remember, no model is perfect! Furthermore, model building takes creativity and patience; it is not a black box exercise. Overall, our goal as data analysts is to find the best possible model which explains the variation in the data in a clear and concise manner. Having done this, our task now moves to using the fitted model for forecasting.

9 Forecasting

Complementary reading: Chapter 9 (CC).

9.1 Introduction

RECALL: We have discussed two types of statistical models for time series data, namely, deterministic trend models (Chapter 3) of the form

$$Y_t = \mu_t + X_t,$$

where $\{X_t\}$ is a zero mean stochastic process, and ARIMA(p, d, q) models of the form

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t,$$

where $\{e_t\}$ is zero mean white noise. For both types of models, we have studied model specification, model fitting, and diagnostic procedures to assess model fit.

PREVIEW: We now switch our attention to **forecasting**.

- We start with a sample of process values up until time t, say, $Y_1, Y_2, ..., Y_t$. These are our **observed data**.
- Forecasting refers to the technique of predicting future values of the process, i.e.,

  $$Y_{t+1}, Y_{t+2}, Y_{t+3}, ..., .$$

  In general, Y_{t+l} is the value of the process at time $t+l$, where $l \geq 1$.
- We call t the **forecast origin** and l the **lead time**. The value Y_{t+l} is "l steps ahead" of the most recently observed value Y_t.

IMPORTANT: By "forecasting," we mean that we are trying to predict the value of a future random variable Y_{t+l}. In general, prediction is a more challenging problem

than, say, estimating a population (model) parameter. Model parameters are fixed (but unknown) values. Random variables are not fixed; they are random.

APPROACH: We need to adopt a formal mathematical criterion to calculate model forecasts. The criterion that we will use is based on the **mean squared error of prediction**

$$\text{MSEP} = E\{[Y_{t+l} - h(Y_1, Y_2, ..., Y_t)]^2\}.$$

- Suppose that we have a sample of observed data $Y_1, Y_2, ..., Y_t$ and that we would like to predict Y_{t+l}.
- The approach we take is to choose the function $h(Y_1, Y_2, ..., Y_t)$ that minimizes MSEP. This function will be our forecasted value of Y_{t+l}.
- The general solution to this minimization problem is

  $$h(Y_1, Y_2, ..., Y_t) = E(Y_{t+l}|Y_1, Y_2, ..., Y_t),$$

  the **conditional expectation** of Y_{t+l}, given the observed data $Y_1, Y_2, ..., Y_t$ (see Appendices E and F, CC).
- Adopting conventional notation, we write

  $$\widehat{Y}_t(l) = E(Y_{t+l}|Y_1, Y_2, ..., Y_t).$$

  This is called the **minimum mean squared error (MMSE) forecast**. That is, $\widehat{Y}_t(l)$ is the MMSE forecast of Y_{t+l}.

Conditional Expectation rules:

- The conditional expectation $E(Z|Y_1, Y_2, ..., Y_t)$ is a function of $Y_1, Y_2, ..., Y_t$.
- If c is a constant, then $E(c|Y_1, Y_2, ..., Y_t) = c$.
- If Z_1 and Z_2 are random variables, then

  $$E(Z_1 + Z_2|Y_1, Y_2, ..., Y_t) = E(Z_1|Y_1, Y_2, ..., Y_t) + E(Z_2|Y_1, Y_2, ..., Y_t);$$

  i.e., conditional expectation is **additive** (just like unconditional expectation).

- If Z is a function of $Y_1, Y_2, ..., Y_t$, say, $Z = f(Y_1, Y_2, ..., Y_t)$, then

$$E(Z|Y_1, Y_2, ..., Y_t) = E[f(Y_1, Y_2, ..., Y_t)|Y_1, Y_2, ..., Y_t] = f(Y_1, Y_2, ..., Y_t).$$

  In other words, once you condition on $Y_1, Y_2, ..., Y_t$, any function of $Y_1, Y_2, ..., Y_t$ acts as a constant.

- If Z is **independent** of $Y_1, Y_2, ..., Y_t$, then

$$E(Z|Y_1, Y_2, ..., Y_t) = E(Z).$$

9.2 Deterministic trend models

RECALL: Consider the model

$$Y_t = \mu_t + X_t,$$

where μ_t is a deterministic (non-random) trend function and where $\{X_t\}$ is assumed to be a white noise process with $E(X_t) = 0$ and $\text{var}(X_t) = \gamma_0$ (constant). By direct calculation, the l-step ahead forecast is

$$\begin{aligned}
\widehat{Y}_t(l) &= E(Y_{t+l}|Y_1, Y_2, ..., Y_t) \\
&= E(\mu_{t+l} + X_{t+l}|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(\mu_{t+l}|Y_1, Y_2, ..., Y_t)}_{=\ \mu_{t+l}} + \underbrace{E(X_{t+l}|Y_1, Y_2, ..., Y_t)}_{=\ E(X_{t+l})=0} = \mu_{t+l},
\end{aligned}$$

because μ_{t+l} is constant and because X_{t+l} is a zero mean random variable independent of $Y_1, Y_2, ..., Y_t$. Therefore,

$$\widehat{Y}_t(l) = \mu_{t+l}$$

is the MMSE forecast.

- For example, if $\mu_t = \beta_0 + \beta_1 t$, a **linear trend** model, then

$$\widehat{Y}_t(l) = \mu_{t+l} = \beta_0 + \beta_1(t + l).$$

- If $\mu_t = \beta_0 + \beta_1 \cos(2\pi f t) + \beta_2 \sin(2\pi f t)$, a **cosine trend** model, then

$$\widehat{Y}_t(l) = \mu_{t+l} = \beta_0 + \beta_1 \cos[2\pi f(t+l)] + \beta_2 \sin[2\pi f(t+l)].$$

ESTIMATION: Of course, MMSE forecasts must be estimated! For example, in the linear trend model, $\widehat{Y}_t(l)$ is estimated by

$$\widehat{\mu}_{t+l} = \widehat{\beta}_0 + \widehat{\beta}_1(t+l).$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimates of β_0 and β_1, respectively. In the cosine trend model, $\widehat{Y}_t(l)$ is estimated by

$$\widehat{\mu}_{t+l} = \widehat{\beta}_0 + \widehat{\beta}_1 \cos[2\pi f(t+l)] + \widehat{\beta}_2 \sin[2\pi f(t+l)],$$

where $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ are the least squares estimates.

Example 9.1. In Example 3.4 (pp 53, notes), we fit a straight line trend model to the global temperature deviation data. The fitted model is

$$\widehat{Y}_t = -12.19 + 0.0062t,$$

where $t = 1900, 1991, ..., 1997$, depicted visually in Figure 9.1. Here are examples of forecasting with this estimated trend model:

- In 1997, we could have used the model to predict for 1998,

$$\widehat{\mu}_{1998} = \widehat{\mu}_{1997+1} = -12.19 + 0.0062(1997+1) \approx 0.198.$$

- For 2005 (8 steps ahead of 1997),

$$\widehat{\mu}_{2005} = \widehat{\mu}_{1997+8} = -12.19 + 0.0062(1997+8) \approx 0.241.$$

- For 2020 (23 steps ahead of 1997),

$$\widehat{\mu}_{2020} = \widehat{\mu}_{1997+23} = -12.19 + 0.0062(1997+23) \approx 0.334.$$

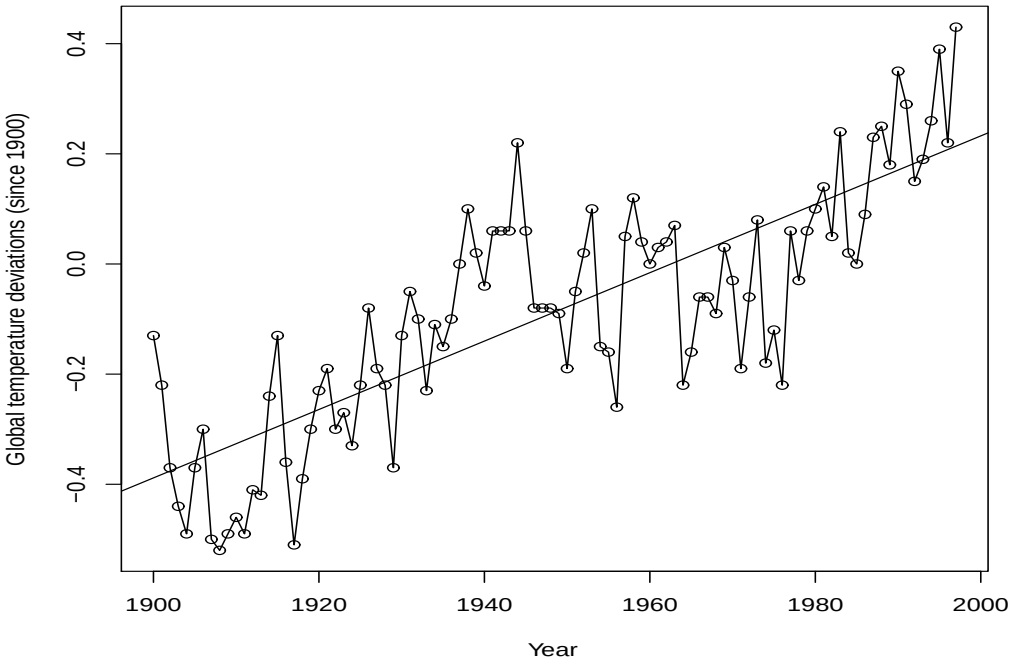


Figure 9.1: Global temperature data. The least squares straight line fit is superimposed.

Example 9.2. In Example 3.6 (pp 66, notes), we fit a cosine trend model to the monthly US beer sales data (in millions of barrels), which produced the fitted model

$$\widehat{Y}_t = 14.8 - 2.04\cos(2\pi t) + 0.93\sin(2\pi t),$$

where $t = 1980, 1980.083, 1980.166, ..., 1990.916$. Note that

- $t = 1980$ refers to January, 1980,
- $t = 1980.083$ refers to February, 1980,
- $t = 1980.166$ refers to March, 1980, and so on.
- These values for t are used because data arrive monthly and "year" is used as a predictor in the regression.
- This fitted model is depicted in Figure 9.2.

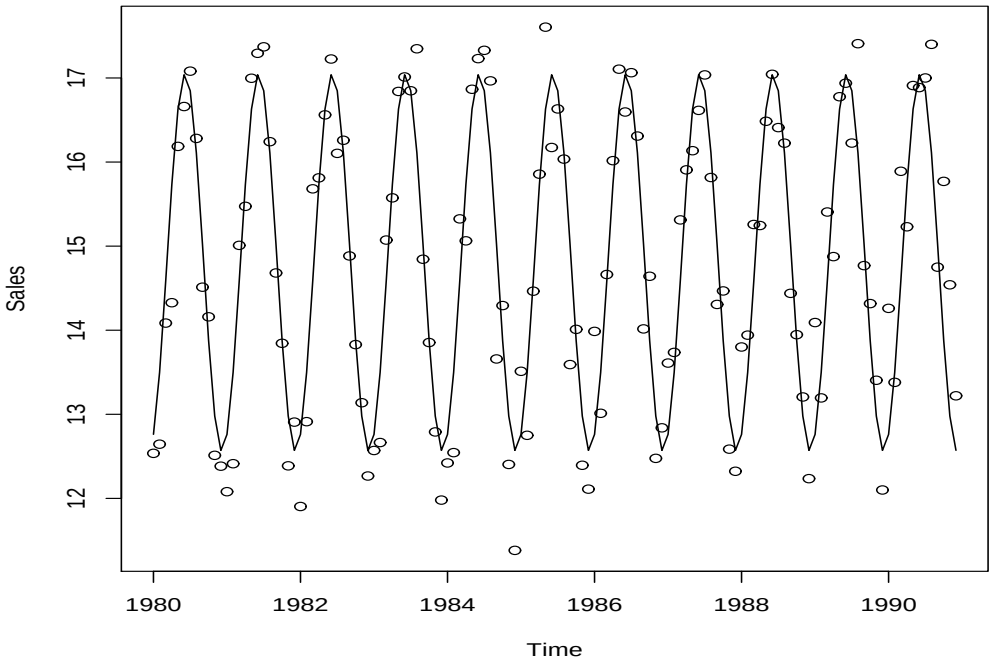


Figure 9.2: Beer sales data. The least squares cosine trend fit is superimposed.

- In December, 1990, we could have used the model to predict for January, 1991,

$$\widehat{\mu}_{1991} = 14.8 - 2.04\cos[2\pi(1991)] + 0.93\sin[2\pi(1991)] \approx 12.76.$$

- For June, 1992,

$$\widehat{\mu}_{1992.416} = 14.8 - 2.04\cos[2\pi(1992.416)] + 0.93\sin[2\pi(1992.416)] \approx 17.03.$$

  Note that the beginning of June, 1992 corresponds to $t = 1992.416$.

REMARK: One major drawback with predictions made from deterministic trend models is that they are based only on the least squares model fit, that is, the forecast for Y_{t+l} ignores the correlation between Y_{t+l} and $Y_1, Y_2, ..., Y_t$. Therefore, the analyst who makes these predictions is ignoring this correlation and, in addition, is assuming that the fitted trend is applicable indefinitely into the future; i.e., "the trend lasts forever."

TERMINOLOGY: For deterministic trend models of the form

$$Y_t = \mu_t + X_t,$$

where $E(X_t) = 0$ and $\text{var}(X_t) = \gamma_0$ (constant), the **forecast error** at lead time l, denoted by $e_t(l)$, is the difference between the value of the process at time $t + l$ and the MMSE forecast at this time. Mathematically,

$$\begin{aligned} e_t(l) &= Y_{t+l} - \widehat{Y}_t(l) \\ &= \mu_{t+l} + X_{t+l} - \mu_{t+l} = X_{t+l}. \end{aligned}$$

For all $l \geq 1$,

$$\begin{aligned} E[e_t(l)] &= E(X_{t+l}) = 0 \\ \text{var}[e_t(l)] &= \text{var}(X_{t+l}) = \gamma_0. \end{aligned}$$

- The first equation implies that forecasts are **unbiased** because the forecast error is an unbiased estimator of 0.
- The second equation implies that the forecast error variance is **constant** for all lead times l.
- These facts will be useful in deriving **prediction intervals** for future values.

9.3 ARIMA models

GOAL: We now discuss forecasting methods with ARIMA models. Recall that an ARIMA(p, d, q) process can be written generally as

$$\phi(B)(1 - B)^d Y_t = \theta_0 + \theta(B)e_t,$$

where θ_0 is an intercept term. We first focus on **stationary** ARMA(p, q) models, that is, ARIMA(p, d, q) models with $d = 0$. Special cases are treated in detail.

9.3.1 AR(1)

AR(1): Suppose that $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. Consider the AR(1) model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t,$$

where the overall (process) mean $\mu = E(Y_t)$.

1-step ahead forecast: The MMSE forecast of Y_{t+1}, the 1-step ahead forecast, is

$$\begin{aligned}
\widehat{Y}_t(1) &= E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= E[\underbrace{\mu + \phi(Y_t - \mu) + e_{t+1}}_{= Y_{t+1}} |Y_1, Y_2, ..., Y_t] \\
&= E(\mu|Y_1, Y_2, ..., Y_t) + E[\phi(Y_t - \mu)|Y_1, Y_2, ..., Y_t] + E(e_{t+1}|Y_1, Y_2, ..., Y_t).
\end{aligned}$$

From the properties of conditional expectation, we note the following:

- $E(\mu|Y_1, Y_2, ..., Y_t) = \mu$, because μ is a constant.
- $E[\phi(Y_t - \mu)|Y_1, Y_2, ..., Y_t] = \phi(Y_t - \mu)$, because $\phi(Y_t - \mu)$ is a function of $Y_1, Y_2, ..., Y_t$.
- $E(e_{t+1}|Y_1, Y_2, ..., Y_t) = E(e_{t+1}) = 0$, because e_{t+1} is independent of $Y_1, Y_2, ..., Y_t$.

Therefore, the MMSE forecast of Y_{t+1} is

$$\widehat{Y}_t(1) = \mu + \phi(Y_t - \mu).$$

2-step ahead forecast: The MMSE forecast of Y_{t+2}, the 2-step ahead forecast, is

$$\begin{aligned}
\widehat{Y}_t(2) &= E(Y_{t+2}|Y_1, Y_2, ..., Y_t) \\
&= E[\underbrace{\mu + \phi(Y_{t+1} - \mu) + e_{t+2}}_{= Y_{t+2}} |Y_1, Y_2, ..., Y_t] \\
&= \underbrace{E(\mu|Y_1, Y_2, ..., Y_t)}_{= \mu} + \underbrace{E[\phi(Y_{t+1} - \mu)|Y_1, Y_2, ..., Y_t]}_{= (**)} + \underbrace{E(e_{t+2}|Y_1, Y_2, ..., Y_t)}_{= E(e_{t+2}) = 0}.
\end{aligned}$$

Now, the expression in $(**)$ is equal to

$$\begin{aligned} E[\phi(Y_{t+1}-\mu)|Y_1, Y_2, ..., Y_t] &= E\{\phi[\underbrace{\mu + \phi(Y_t - \mu) + e_{t+1}}_{=\, Y_{t+1}} - \mu]|Y_1, Y_2, ..., Y_t\} \\ &= E\{\phi[\phi(Y_t - \mu) + e_{t+1}]|Y_1, Y_2, ..., Y_t\} \\ &= E[\phi^2(Y_t - \mu)|Y_1, Y_2, ..., Y_t] + E(\phi e_{t+1}|Y_1, Y_2, ..., Y_t). \end{aligned}$$

From the properties of conditional expectation, we again note the following:

- $E[\phi^2(Y_t - \mu)|Y_1, Y_2, ..., Y_t] = \phi^2(Y_t - \mu)$, because $\phi^2(Y_t - \mu)$ is a function of $Y_1, Y_2, ..., Y_t$.
- $E(\phi e_{t+1}|Y_1, Y_2, ..., Y_t) = \phi E(e_{t+1}|Y_1, Y_2, ..., Y_t) = \phi E(e_{t+1}) = 0$, because e_{t+1} is independent of $Y_1, Y_2, ..., Y_t$.

Therefore, the MMSE forecast of Y_{t+2} is

$$\widehat{Y}_t(2) = \mu + \phi^2(Y_t - \mu).$$

l-step ahead forecast: For larger lead times, this pattern continues. In general, the MMSE forecast of Y_{t+l}, for all $l \geq 1$, is

$$\widehat{Y}_t(l) = \mu + \phi^l(Y_t - \mu).$$

- When $-1 < \phi < 1$ (stationarity condition), note that $\phi^l \approx 0$ when l is large.
- Therefore, as l increases without bound, the l-step ahead MMSE forecast

$$\widehat{Y}_t(l) \to \mu.$$

  In other words, MMSE forecasts will "converge" to the overall process mean μ as the lead time l increases.

IMPORTANT: That the MMSE forecast $\widehat{Y}_t(l) \to \mu$ as $l \to \infty$ is a characteristic of all **stationary** ARMA(p, q) models.

FORECAST ERROR: In the AR(1) model, the **1-step ahead forecast error** is

$$\begin{aligned} e_t(1) &= Y_{t+1} - \widehat{Y}_t(1) \\ &= \underbrace{\mu + \phi(Y_t - \mu) + e_{t+1}}_{=\, Y_{t+1}} - [\underbrace{\mu + \phi(Y_t - \mu)}_{=\, \widehat{Y}_t(1)}] \\ &= e_{t+1}. \end{aligned}$$

Therefore,

$$\begin{aligned} E[e_t(1)] &= E(e_{t+1}) = 0 \\ \text{var}[e_t(1)] &= \text{var}(e_{t+1}) = \sigma_e^2. \end{aligned}$$

Because the 1-step ahead forecast error $e_t(1)$ is an unbiased estimator of 0, we say that the 1-step ahead forecast $\widehat{Y}_t(1)$ is **unbiased**. The second equation says that the 1-step ahead forecast error $e_t(1)$ has constant variance. To find the l**-step ahead forecast error**, $e_t(l)$, we first remind ourselves (pp 94, notes) that a zero mean AR(1) process can be written as an infinite order MA model, that is,

$$Y_t - \mu = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots .$$

Therefore,

$$Y_{t+l} - \mu = e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1} + \phi^l e_t + \cdots .$$

The l-step ahead forecast error is

$$\begin{aligned} e_t(l) &= Y_{t+l} - \widehat{Y}_t(l) \\ &= Y_{t+l} - \mu + \mu - \widehat{Y}_t(l) \\ &= \underbrace{e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1} + \phi^l e_t + \cdots}_{=\, Y_{t+l} - \mu} \underbrace{-\phi^l (Y_t - \mu)}_{=\, \mu - \widehat{Y}_t(l)} \\ &= e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1} + \phi^l e_t + \cdots \\ &\qquad -\phi^l (\underbrace{e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots}_{=\, Y_t - \mu}) \\ &= e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1}. \end{aligned}$$

Therefore, the l-step ahead forecast error has mean

$$E[e_t(l)] = E(e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1}) = 0,$$

i.e., forecasts are **unbiased**. The variance of the l-step ahead forecast error is

$$\begin{aligned}
\text{var}[e_t(l)] &= \text{var}(e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + \cdots + \phi^{l-1} e_{t+1}) \\
&= \text{var}(e_{t+l}) + \phi^2 \text{var}(e_{t+l-1}) + \phi^4 \text{var}(e_{t+l-2}) + \cdots + \phi^{2(l-1)} \text{var}(e_{t+1}) \\
&= \sigma_e^2 + \phi^2 \sigma_e^2 + \phi^4 \sigma_e^2 + \cdots + \phi^{2(l-1)} \sigma_e^2 \\
&= \sigma_e^2 \sum_{k=0}^{l-1} \phi^{2k} = \sigma_e^2 \left(\frac{1 - \phi^{2l}}{1 - \phi^2} \right).
\end{aligned}$$

Assuming stationarity, note that $\phi^{2l} \to 0$ as $l \to \infty$ (because $-1 < \phi < 1$) and

$$\text{var}[e_t(l)] \to \frac{\sigma_e^2}{1 - \phi^2} = \gamma_0 = \text{var}(Y_t).$$

IMPORTANT: That $\text{var}[e_t(l)] \to \gamma_0 = \text{var}(Y_t)$ as $l \to \infty$ is a characteristic of all **stationary** ARMA(p, q) models.

Example 9.3. In Example 8.2 (pp 213, notes), we examined the Lake Huron elevation data (from 1880-2006) and we used an AR(1) process to model them.

- The fit using maximum likelihood is

$$Y_t - 579.4921 = 0.8586(Y_{t-1} - 579.4921) + e_t,$$

  so that $\widehat{\mu} = 579.4921$, $\widehat{\phi} = 0.8586$, and the white noise error variance estimate $\widehat{\sigma}_e^2 = 0.4951$. The last value observed was $Y_t = 581.27$ (the elevation for **2006**).

- With $l = 1$, the (estimated) MMSE forecast for Y_{t+1} (for 2007) is

$$\widehat{Y}_t(1) = 579.4921 + 0.8586(581.27 - 579.4921) \approx 581.02.$$

- With $l = 2$, the (estimated) MMSE forecast for Y_{t+2} (for 2008) is

$$\widehat{Y}_t(2) = 579.4921 + (0.8586)^2(581.27 - 579.4921) \approx 580.80.$$

- With $l = 10$, the (estimated) MMSE forecast for Y_{t+10} (for 2016) is

$$\widehat{Y}_t(10) = 579.4921 + (0.8586)^{10}(581.27 - 579.4921) \approx 579.88.$$

NOTE: The R function `predict` provides (estimated) MMSE forecasts and (estimated) standard errors of the forecast error for any ARIMA(p, d, q) model fit. For example, consider the Lake Huron data with lead times $l = 1, 2, ..., 20$ (which corresponds to years 2007, 2008, ..., 2026). R produces the following output:

```
huron.ar1.predict <- predict(huron.ar1.fit,n.ahead=20)
> round(huron.ar1.predict$pred,3)
Start = 2007
End = 2026
 [1] 581.019 580.803 580.618 580.458 580.322 580.205 580.104 580.017 579.943 579.879
[11] 579.825 579.778 579.737 579.703 579.673 579.647 579.625 579.607 579.590 579.576

> round(huron.ar1.predict$se,3)
Start = 2007
End = 2026
 [1] 0.704 0.927 1.063 1.152 1.214 1.258 1.289 1.311 1.328 1.340 1.349 1.355 1.360
[14] 1.363 1.366 1.367 1.369 1.370 1.371 1.371
```

- In Figure 9.3, we display the Lake Huron data. The full data set is from 1880-2006 (one elevation reading per year).
- However, for aesthetic reasons (to emphasize the MMSE forecasts), we start the series in the plot at year 1940.
- The **estimated** MMSE forecasts in the R `predict` output are computed using

$$\widehat{Y}_t(l) = \widehat{\mu} + \widehat{\phi}^l(Y_t - \widehat{\mu}),$$

  for $l = 1, 2, ..., 20$, starting with $Y_t = 581.27$, the observed elevation in 2006. Therefore, the forecasts in Figure 9.3 start at 2007 and end in 2026.
- In the output above, note how MMSE forecasts $\widehat{Y}_t(l)$ approach the estimated mean $\widehat{\mu} = 579.492$, as l increases. This can also be clearly seen in Figure 9.3.
- The (estimated) standard errors of the forecast error (in the `predict` output above) are used to construct **prediction intervals**. We will discuss their construction in due course.

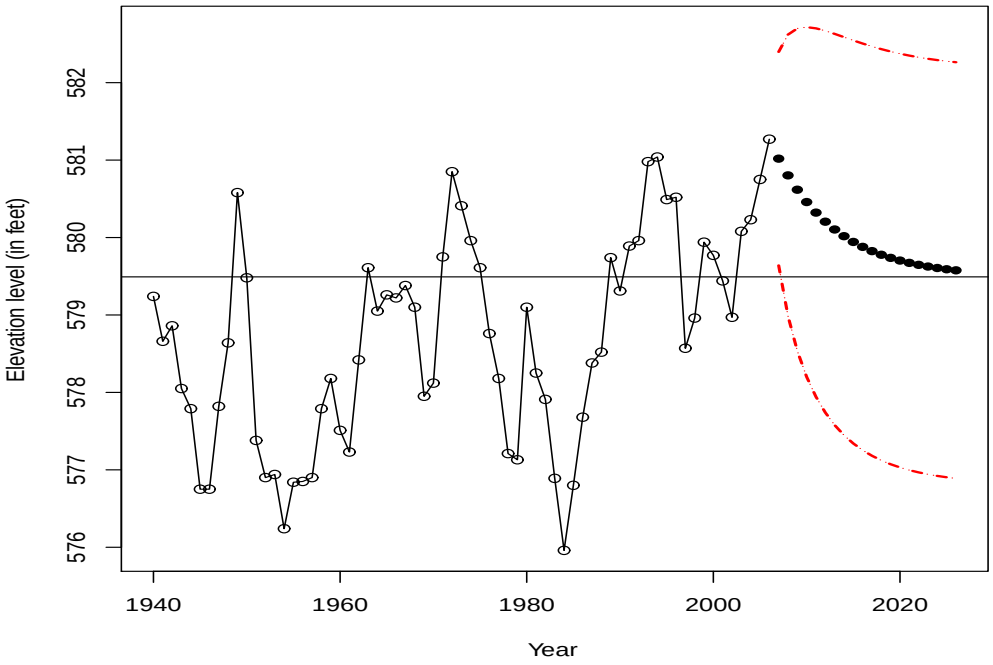


Figure 9.3: Lake Huron elevation data. The full data set is from 1880-2006. This figure starts the series at 1940. AR(1) estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 20$. These lead times correspond to years 2007-2026.

- Specifically, the (estimated) standard errors of the forecast error (in the `predict` output above) are given by

$$\widehat{\text{se}}[e_t(l)] = \sqrt{\widehat{\text{var}}[e_t(l)]} = \sqrt{\widehat{\sigma}_e^2 \left(\frac{1 - \widehat{\phi}^{2l}}{1 - \widehat{\phi}^2} \right)},$$

  where $\widehat{\sigma}_e^2 = 0.4951$ and $\widehat{\phi} = 0.8586$.

- Note how these (estimated) standard errors approach

$$\lim_{l \to \infty} \widehat{\text{se}}[e_t(l)] = \sqrt{\frac{\widehat{\sigma}_e^2}{1 - \widehat{\phi}^2}} = \sqrt{\frac{0.4951}{1 - (0.8586)^2}} \approx 1.373.$$

  This value (1.373) is the square root of the estimated AR(1) process variance $\widehat{\gamma}_0$.

9.3.2 MA(1)

MA(1): Suppose that $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. Consider the invertible MA(1) process

$$Y_t = \mu + e_t - \theta e_{t-1},$$

where the overall (process) mean $\mu = E(Y_t)$.

1-step ahead forecast: The MMSE forecast of Y_{t+1}, the 1-step ahead forecast, is

$$\begin{aligned}
\widehat{Y}_t(1) &= E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= E(\mu + e_{t+1} - \theta e_t|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(\mu|Y_1, Y_2, ..., Y_t)}_{=\ \mu} + \underbrace{E(e_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\ E(e_{t+1})=0} - \underbrace{E(\theta e_t|Y_1, Y_2, ..., Y_t)}_{=\ (**)}.
\end{aligned}$$

To compute $(**)$, recall (pp 105, notes) that a zero mean invertible MA(1) process can be written in its "AR(∞)" expansion

$$e_t = (Y_t - \mu) + \theta(Y_{t-1} - \mu) + \theta^2(Y_{t-2} - \mu) + \theta^3(Y_{t-3} - \mu) + \cdots,$$

a weighted (theoretically infinite) linear combination of $Y_{t-j} - \mu$, for $j = 0, 1, 2, ...,$. That is, e_t can be expressed as a function of $Y_1, Y_2, ..., Y_t$, and hence

$$E(\theta e_t|Y_1, Y_2, ..., Y_t) = \theta e_t.$$

Therefore, the 1-step ahead forecast

$$\widehat{Y}_t(1) = \mu - \theta e_t.$$

From the representation above, note that the white noise term e_t can be "computed" in the 1-step ahead forecast as a byproduct of estimating θ and μ in the MA(1) fit.

***l*-step ahead forecast:** The MMSE prediction for Y_{t+l}, $l > 1$, is given by

$$\begin{aligned}
\widehat{Y}_t(l) &= E(Y_{t+l}|Y_1, Y_2, ..., Y_t) \\
&= E(\mu + e_{t+l} - \theta e_{t+l-1}|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(\mu|Y_1, Y_2, ..., Y_t)}_{=\ \mu} + \underbrace{E(e_{t+l}|Y_1, Y_2, ..., Y_t)}_{=\ E(e_{t+l})=0} - \underbrace{E(\theta e_{t+l-1}|Y_1, Y_2, ..., Y_t)}_{=\ \theta E(e_{t+l-1})=0},
\end{aligned}$$

because e_{t+l} and e_{t+l-1} are both independent of $Y_1, Y_2, ..., Y_t$, when $l > 1$. Therefore, we have shown that for the MA(1) model, MMSE forecasts are

$$\widehat{Y}_t(l) = \begin{cases} \mu - \theta e_t, & l = 1 \\ \mu, & l > 1. \end{cases}$$

The key feature of an MA(1) process is that observations one unit apart in time are correlated, whereas observations $l > 1$ units apart in time are not. For $l > 1$, there is no autocorrelation to exploit in making a prediction; this is why a constant mean prediction is made. **Note:** More generally, for any purely MA(q) process, the MMSE forecast is $\widehat{Y}_t(l) = \mu$ at all lead times $l > q$.

REMARK: Just as we saw in the AR(1) model case, note that $\widehat{Y}_t(l) \to \mu$ as $l \to \infty$. This is a characteristic of $\widehat{Y}_t(l)$ in all **stationary** ARMA(p, q) models.

FORECAST ERROR: In the MA(1) model, the **1-step ahead forecast error** is

$$\begin{aligned} e_t(1) &= Y_{t+1} - \widehat{Y}_t(1) \\ &= \underbrace{\mu + e_{t+1} - \theta e_t}_{= Y_{t+1}} - \underbrace{(\mu - \theta e_t)}_{= \widehat{Y}_t(1)} \\ &= e_{t+1}. \end{aligned}$$

Therefore,

$$\begin{aligned} E[e_t(1)] &= E(e_{t+1}) = 0 \\ \text{var}[e_t(1)] &= \text{var}(e_{t+1}) = \sigma_e^2. \end{aligned}$$

As in the AR(1) model, 1-step ahead forecasts are unbiased and the variance of the 1-step ahead forecast error is constant. The variance of the l**-step ahead prediction error** $e_t(l)$, for $l > 1$, is given by

$$\begin{aligned} \text{var}[e_t(l)] &= \text{var}[Y_{t+l} - \widehat{Y}_t(l)] \\ &= \text{var}(\underbrace{\mu + e_{t+l} - \theta e_{t+l-1}}_{= Y_{t+l}} - \mu) \\ &= \text{var}(e_{t+l} - \theta e_{t+l-1}) \\ &= \text{var}(e_{t+l}) + \theta^2 \text{var}(e_{t+l-1}) - 2\theta \underbrace{\text{cov}(e_{t+l}, e_{t+l-1})}_{= 0} \\ &= \sigma_e^2 + \theta^2 \sigma_e^2 = \sigma_e^2(1 + \theta^2). \end{aligned}$$

Summarizing,

$$\text{var}[e_t(l)] = \begin{cases} \sigma_e^2, & l = 1 \\ \sigma_e^2(1+\theta^2), & l > 1. \end{cases}$$

REMARK: In the MA(1) model, note that as $l \to \infty$,

$$\text{var}[e_t(l)] \to \gamma_0 = \text{var}(Y_t).$$

This is a characteristic of $\text{var}[e_t(l)]$ in all **stationary** ARMA(p, q) models.

Example 9.4. In Example 7.6 (pp 202, notes), we examined the Göta River discharge rate data (1807-1956) and used an MA(1) process to model them. The fitted model (using ML) is

$$Y_t = 535.0311 + e_t + 0.5350e_{t-1}.$$

so that $\widehat{\mu} = 535.0311$, $\widehat{\theta} = -0.5350$ and $\widehat{\sigma}_e^2 = 6957$. Here are the MA(1) forecasts for lead times $l = 1, 2, ..., 10$, computed using the `predict` function in R:

```
> gota.ma1.predict <- predict(gota.ma1.fit,n.ahead=10)
> round(gota.ma1.predict$pred,3)
Start = 1957
End = 1966
[1] 510.960 535.031 535.031 535.031 535.031 535.031 535.031 535.031 535.031 535.031

> round(gota.ma1.predict$se,3)
Start = 1957
End = 1966
[1] 83.411 94.599 94.599 94.599 94.599 94.599 94.599 94.599 94.599 94.599
```

- In Figure 9.4, we display the Göta River data. The full data set is from 1807-1956 (one discharge reading per year). However, to emphasize the MMSE forecasts in the plot, we start the series at year 1890.
- With $l = 1, 2, ..., 10$, the MMSE forecasts in the `predict` output and in Figure 9.4 start at 1957 and end in 1966.
- From the `predict` output, note that $\widehat{Y}_t(1) = 510.960$, the **1-step ahead forecast**, is the only "informative" one. Forecasts for $l > 1$ are $\widehat{Y}_t(l) = \widehat{\mu} \approx 535.0311$.

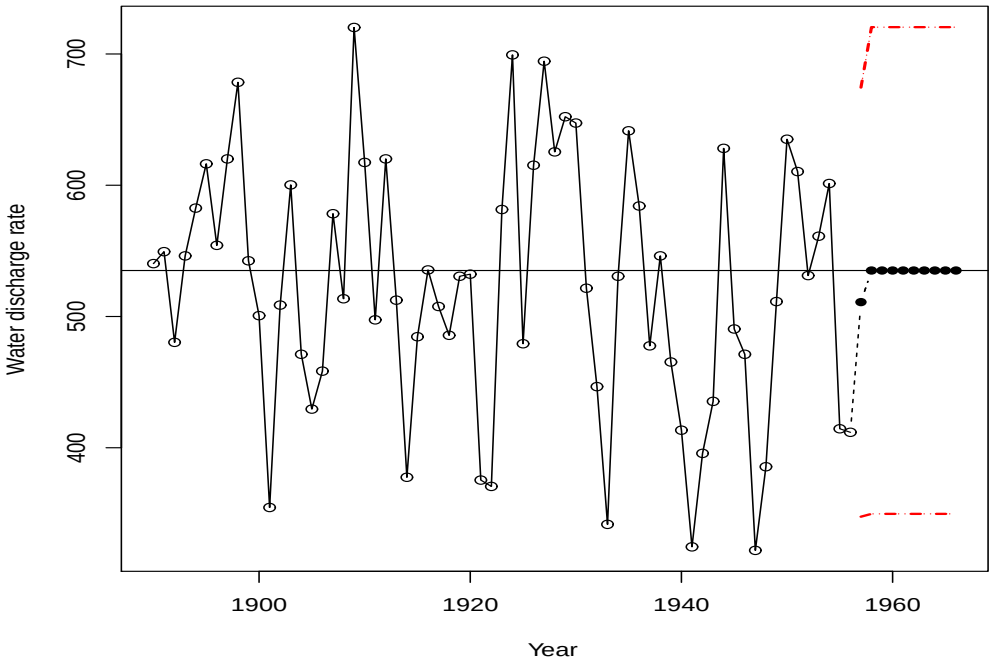


Figure 9.4: Göta River discharge data. The full data set is from 1807-1956. This figure starts the series at 1890. MA(1) estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 10$. These lead times correspond to years 1957-1966.

- Recall that MA(1) forecasts only exploit the autocorrelation at the $l = 1$ lead time! In the MA(1) process, there is no autocorrelation after the first lag. All future forecasts (after the first) will revert to the process mean estimate.

- For lead time $l = 1$,

$$\widehat{\text{se}}[e_t(1)] = \sqrt{\widehat{\text{var}}[e_t(1)]} = \sqrt{\widehat{\sigma}_e^2} = \sqrt{6957} \approx 83.411$$

- For any lead time $l > 1$,

$$\widehat{\text{se}}[e_t(l)] = \sqrt{\widehat{\text{var}}[e_t(l)]} = \sqrt{\widehat{\sigma}_e^2(1 + \widehat{\theta}^2)} \approx \sqrt{6957[1 + (-0.5350)^2]} \approx 94.599.$$

This value (94.599) is the square root of the estimated MA(1) process variance $\widehat{\gamma}_0$.

9.3.3 ARMA(p, q)

ARMA(p, q): Suppose that $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$ and consider the ARMA(p, q) process

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}.$$

To calculate the l-step ahead MMSE forecast, replace the time index t with $t+l$ and take conditional expectations of both sides (given the process history $Y_1, Y_2, ..., Y_t$). Doing this leads directly to the following **difference equation**:

$$\begin{aligned}\widehat{Y}_t(l) &= \theta_0 + \phi_1 \widehat{Y}_t(l-1) + \phi_2 \widehat{Y}_t(l-2) + \cdots + \phi_p \widehat{Y}_t(l-p) \\ &\quad - \theta_1 E(e_{t+l-1}|Y_1, Y_2, ..., Y_t) - \theta_2 E(e_{t+l-2}|Y_1, Y_2, ..., Y_t) - \cdots \\ &\quad - \theta_q E(e_{t+l-q}|Y_1, Y_2, ..., Y_t).\end{aligned}$$

For a general ARMA(p, q) process, MMSE forecasts are calculated using this equation.

- In the expression above,

$$\widehat{Y}_t(l-j) = E(Y_{t+l-j}|Y_1, Y_2, ..., Y_t),$$

  for $j = 1, 2, ..., p$. General recursive formulas can be derived to compute this conditional expectation, as we saw in the AR(1) case.

- In the expression above,

$$E(e_{t+l-k}|Y_1, Y_2, ..., Y_t) = \begin{cases} 0, & l-k > 0 \\ e_{t+l-k}, & l-k \leq 0, \end{cases}$$

  for $k = 1, 2, ..., q$. When $l - k \leq 0$, the conditional expectation

$$E(e_{t+l-k}|Y_1, Y_2, ..., Y_t) = e_{t+l-k},$$

  which can be approximated using infinite AR representations for invertible models (see pp 80, CC). This is only necessary for MMSE forecasts at early lags $l \leq q$ when q is larger than or equal to 1.

SPECIAL CASE: Consider the **ARMA(1,1)** process

$$Y_t = \theta_0 + \phi Y_{t-1} + e_t - \theta e_{t-1}.$$

For $l = 1$, we have

$$\begin{aligned}
\widehat{Y}_t(1) &= E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= E(\theta_0 + \phi Y_t + e_{t+1} - \theta e_t|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(\theta_0|Y_1, Y_2, ..., Y_t)}_{=\,\theta_0} + \underbrace{E(\phi Y_t|Y_1, Y_2, ..., Y_t)}_{=\,\phi Y_t} + \underbrace{E(e_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\,E(e_{t+1})=0} \\
&\qquad\qquad - \underbrace{E(\theta e_t|Y_1, Y_2, ..., Y_t)}_{=\,\theta e_t} \\
&= \theta_0 + \phi Y_t - \theta e_t.
\end{aligned}$$

For $l = 2$, we have

$$\begin{aligned}
\widehat{Y}_t(2) &= E(Y_{t+2}|Y_1, Y_2, ..., Y_t) \\
&= E(\theta_0 + \phi Y_{t+1} + e_{t+2} - \theta e_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(\theta_0|Y_1, Y_2, ..., Y_t)}_{=\,\theta_0} + \underbrace{E(\phi Y_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\,\phi\widehat{Y}_t(1)} + \underbrace{E(e_{t+2}|Y_1, Y_2, ..., Y_t)}_{=\,E(e_{t+2})=0} \\
&\qquad\qquad - \underbrace{E(\theta e_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\,\theta E(e_{t+1})=0} \\
&= \theta_0 + \phi\widehat{Y}_t(1).
\end{aligned}$$

It is easy to see that this pattern continues for larger lead times l; in general,

$$\widehat{Y}_t(l) = \theta_0 + \phi\widehat{Y}_t(l-1),$$

for all lead times $l > 1$. It is important to make the following observations in this special ARMA$(p = 1, q = 1)$ case:

- The MMSE forecast $\widehat{Y}_t(l)$ depends on the MA components only when $l \leq q = 1$.
- When $l > q = 1$, the MMSE forecast $\widehat{Y}_t(l)$ depends only on the AR components.
- This is also true of MMSE forecasts in higher order ARMA(p, q) models.

SUMMARY: The following notes summarize MMSE forecast calculations in general ARMA(p, q) models:

- When $l \leq q$, MMSE forecasts depend on both the AR and MA parts of the model.
- When $l > q$, the MA contributions vanish and forecasts will depend solely on the recursion identified in the AR part. That is, when $l > q$,

$$\widehat{Y}_t(l) = \theta_0 + \phi_1\widehat{Y}_t(l-1) + \phi_2\widehat{Y}_t(l-2) + \cdots + \phi_p\widehat{Y}_t(l-p).$$

- It is insightful to note that the last expression, for $l > q$, can be written as

$$\widehat{Y}_t(l) - \mu = \phi_1[\widehat{Y}_t(l-1) - \mu] + \phi_2[\widehat{Y}_t(l-2) - \mu] + \cdots + \phi_p[\widehat{Y}_t(l-p) - \mu].$$

  Therefore,

  - as a function of l, $\widehat{Y}_t(l) - \mu$ follows the same **Yule-Walker** recursion as the autocorrelation function ρ_k.
  - the roots of $\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$ determine the behavior of $\widehat{Y}_t(l) - \mu$, when $l > q$; e.g., exponential decay, damped sine waves, etc.

- For any **stationary** ARMA(p, q) process,

$$\lim_{l\to\infty} \widehat{Y}_t(l) = \mu,$$

  where $\mu = E(Y_t)$. Therefore, for large lead times l, MMSE forecasts will be approximately equal to the process mean.

- For any **stationary** ARMA(p, q) process, the variance of the **l-step ahead forecast error** satisfies

$$\lim_{l\to\infty} \text{var}[e_t(l)] = \gamma_0,$$

  where $\gamma_0 = \text{var}(Y_t)$. That is, for large lead times l, the variance of the forecast error will be close to the process variance.

- The `predict` function in R automates the entire forecasting process, providing (estimated) MMSE forecasts and standard errors of the forecast error.

Example 9.5. In Example 7.5 (pp 195, notes), we examined the bovine blood sugar data (176 observations) and we used an ARMA(1,1) process to model them. The fitted ARMA(1,1) model (using ML) is

$$Y_t - 59.0071 = 0.6623(Y_{t-1} - 59.0071) + e_t + 0.6107e_{t-1},$$

so that $\widehat{\mu} = 59.0071$, $\widehat{\phi} = 0.6623$, $\widehat{\theta} = -0.6107$, and the white noise variance estimate $\widehat{\sigma}_e^2 = 20.43$. Here are the ARMA(1,1) forecasts for lead times $l = 1, 2, ..., 10$, computed using the `predict` function in R:

```
> cows.arma11.predict <- predict(cows.arma11.fit,n.ahead=10)
> round(cows.arma11.predict$pred,3)
Start = 177
End = 186
[1] 58.643 58.766 58.847 58.901 58.937 58.961 58.976 58.987 58.994 58.998

> round(cows.arma11.predict$se,3)
Start = 177
End = 186
[1] 4.520 7.316 8.249 8.627 8.787 8.856 8.887 8.900 8.906 8.908
```

- In Figure 9.5, we display the bovine data. The full data set is from day 1-176 (one blood sugar reading per day). However, to emphasize the MMSE forecasts in the plot, we start the series at day 81.
- With $l = 1, 2, ..., 10$, the MMSE forecasts in the `predict` output and in Figure 9.5 start at day 177 and end at day 186.
- From the `predict` output and Figure 9.5, note that the predictions are all close to $\widehat{\mu} = 59.0071$, the estimated process mean. This happens because the last observed data value was $Y_{176} = 55.91133$, which is already somewhat close to $\widehat{\mu} = 59.0071$.
- Close inspection reveals that the forecasts decay (quickly) towards $\widehat{\mu} = 59.0071$ as expected.

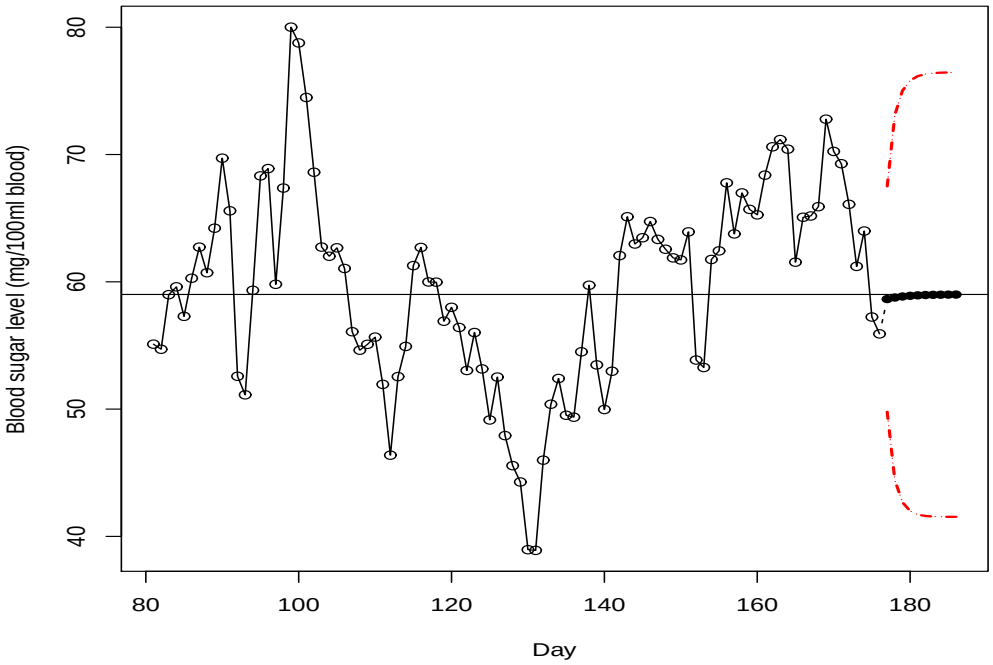


Figure 9.5: Bovine blood sugar data. The full data set is from day 1-176. This figure starts the series at day 81. ARMA(1,1) estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 10$. These lead times correspond to days 177-186.

- The variance of the l-step ahead prediction error $e_t(l)$ should satisfy

$$\lim_{l\to\infty} \text{var}[e_t(l)] = \gamma_0 = \left(\frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2}\right)\sigma_e^2,$$

  which, with $\widehat{\phi} = 0.6623$, $\widehat{\theta} = -0.6107$, and $\widehat{\sigma}_e^2 = 20.43$, is estimated to be

$$\begin{aligned}\widehat{\gamma}_0 &= \left(\frac{1 - 2\widehat{\phi}\widehat{\theta} + \widehat{\theta}^2}{1 - \widehat{\phi}^2}\right)\widehat{\sigma}_e^2 \\ &= \left[\frac{1 - 2(0.6623)(-0.6107) + (-0.6107)^2}{1 - (0.6623)^2}\right](20.43) \approx 79.407.\end{aligned}$$

- As l increases, note that the estimated standard errors $\widehat{\text{se}}[e_t(l)]$ from the `predict` output, as expected, get very close to $\sqrt{\widehat{\gamma}_0} \approx \sqrt{79.407} \approx 8.911$.

9.3.4 Nonstationary models

NOTE: For invertible ARIMA(p, d, q) models with $d \geq 1$, MMSE forecasts are computed using the same approach as in the stationary case. To see why, suppose that $d = 1$, so that the model is

$$\phi(B)(1-B)Y_t = \theta(B)e_t,$$

where $(1-B)Y_t = \nabla Y_t$ is the series of first differences. Note that

$$\begin{aligned}
\phi(B)(1-B) &= (1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1-B) \\
&= (1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) - (B - \phi_1 B^2 - \phi_2 B^3 - \cdots - \phi_p B^{p+1}) \\
&= \underbrace{1 - (1+\phi_1)B - (\phi_2 - \phi_1)B^2 - \cdots - (\phi_p - \phi_{p-1})B^p + \phi_p B^{p+1}}_{= \phi^*(B), \text{ say}}.
\end{aligned}$$

We can therefore rewrite the ARIMA$(p, 1, q)$ model as

$$\phi^*(B)Y_t = \theta(B)e_t,$$

a nonstationary ARMA$(p+1, q)$ model. We then calculate MMSE forecasts the same way as in the stationary case.

EXAMPLE: Suppose $p = d = q = 1$ so that we have an ARIMA(1,1,1) process

$$(1-\phi B)(1-B)Y_t = (1-\theta B)e_t.$$

This can be written as

$$Y_t = (1+\phi)Y_{t-1} - \phi Y_{t-2} + e_t - \theta e_{t-1},$$

a nonstationary ARMA(2,1) model. If $l = 1$, then

$$\begin{aligned}
\widehat{Y}_t(1) &= E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= E[(1+\phi)Y_t - \phi Y_{t-1} + e_{t+1} - \theta e_t | Y_1, Y_2, ..., Y_t] \\
&= \underbrace{E[(1+\phi)Y_t|Y_1, Y_2, ..., Y_t]}_{= (1+\phi)Y_t} - \underbrace{E(\phi Y_{t-1}|Y_1, Y_2, ..., Y_t)}_{= \phi Y_{t-1}} + \underbrace{E(e_{t+1}|Y_1, Y_2, ..., Y_t)}_{= E(e_{t+1})=0} \\
&\qquad - \underbrace{E(\theta e_t|Y_1, Y_2, ..., Y_t)}_{= \theta e_t} \\
&= (1+\phi)Y_t - \phi Y_{t-1} - \theta e_t.
\end{aligned}$$

If $l = 2$, then

$$\begin{aligned}
\widehat{Y}_t(2) &= E(Y_{t+2}|Y_1, Y_2, ..., Y_t) \\
&= E[(1+\phi)Y_{t+1} - \phi Y_t + e_{t+2} - \theta e_{t+1}|Y_1, Y_2, ..., Y_t] \\
&= \underbrace{E[(1+\phi)Y_{t+1}|Y_1, Y_2, ..., Y_t]}_{=\,(1+\phi)\widehat{Y}_t(1)} - \underbrace{E(\phi Y_t|Y_1, Y_2, ..., Y_t)}_{=\,\phi Y_t} + \underbrace{E(e_{t+2}|Y_1, Y_2, ..., Y_t)}_{=\,E(e_{t+2})=0} \\
&\qquad\qquad - \underbrace{E(\theta e_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\,\theta E(e_{t+1})=0} \\
&= (1+\phi)\widehat{Y}_t(1) - \phi Y_t.
\end{aligned}$$

For $l > 2$, it follows similarly that

$$\begin{aligned}
\widehat{Y}_t(l) &= E(Y_{t+l}|Y_1, Y_2, ..., Y_t) \\
&= E[(1+\phi)Y_{t+l-1} - \phi Y_{t+l-2} + e_{t+l} - \theta e_{t+l-1}|Y_1, Y_2, ..., Y_t] \\
&= (1+\phi)\widehat{Y}_t(l-1) - \phi\widehat{Y}_t(l-2).
\end{aligned}$$

Writing recursive expressions for MMSE forecasts in any invertible ARIMA(p, d, q) model can be done similarly.

RESULT: The **l-step ahead forecast error** $e_t(l) = Y_{t+l} - \widehat{Y}_t(l)$ for any invertible ARIMA(p, d, q) model has the following characteristics:

$$\begin{aligned}
E[e_t(l)] &= 0 \\
\text{var}[e_t(l)] &= \sigma_e^2 \sum_{j=0}^{l-1} \Psi_j^2,
\end{aligned}$$

where the Ψ weights correspond to those in the truncated linear process representation of the ARIMA(p, d, q) model; see pp 200 (CC).

- The first equation implies that MMSE ARIMA forecasts are **unbiased**.
- The salient feature in the second equation is that for **nonstationary** models, the Ψ weights do not "die out" as they do with stationary models.
- Therefore, for nonstationary models, the variance of the forecast error var$[e_t(l)]$ continues to increase as l does. This is not surprising given that the process is not stationary.

Example 9.6. In Example 8.7 (pp 225, notes), we examined monthly spot prices for crude oil (measured in U.S. dollars per barrel) from 1/86 to 1/06, and we used a log-transformed IMA(1,1) process to model them. The model fit (using ML) is

$$\log Y_t = \log Y_{t-1} + e_t + 0.2956e_{t-1},$$

so that $\widehat{\theta} = -0.2956$ and the white noise variance estimate is $\widehat{\sigma}_e^2 = 0.006689$. The estimated forecasts and standard errors (**on the log scale**) are given for lead times $l = 1, 2, ..., 12$ in the `predict` output below:

```
> ima11.log.oil.predict <- predict(ima11.log.oil.fit,n.ahead=12)
> round(ima11.log.oil.predict$pred,3)
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2006       4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208
2007 4.208

> round(ima11.log.oil.predict$se,3)
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2006       0.082 0.134 0.171 0.201 0.227 0.251 0.272 0.292 0.311 0.328 0.345
2007 0.361
```

- In Figure 9.6, we display the oil price data. The full data set is from 1/86 to 1/06 (one observation per month). However, to emphasize the MMSE forecasts in the plot, we start the series at month 1/98.
- With $l = 1, 2, ..., 12$, the estimated MMSE forecasts in the `predict` output and in Figure 9.6 start at 2/06 and end in 1/07.
- From the `predict` output, note that $\widehat{Y}_t(1) = \widehat{Y}_t(2) = \cdots = \widehat{Y}_t(12) = 4.208$. It is important to remember that these forecasts are on the **log scale**.
- On the original scale (in dollars), we will see later that MMSE forecasts are not constant.
- As expected from a nonstationary process, the estimated standard errors (also on the log scale) increase as l increases.

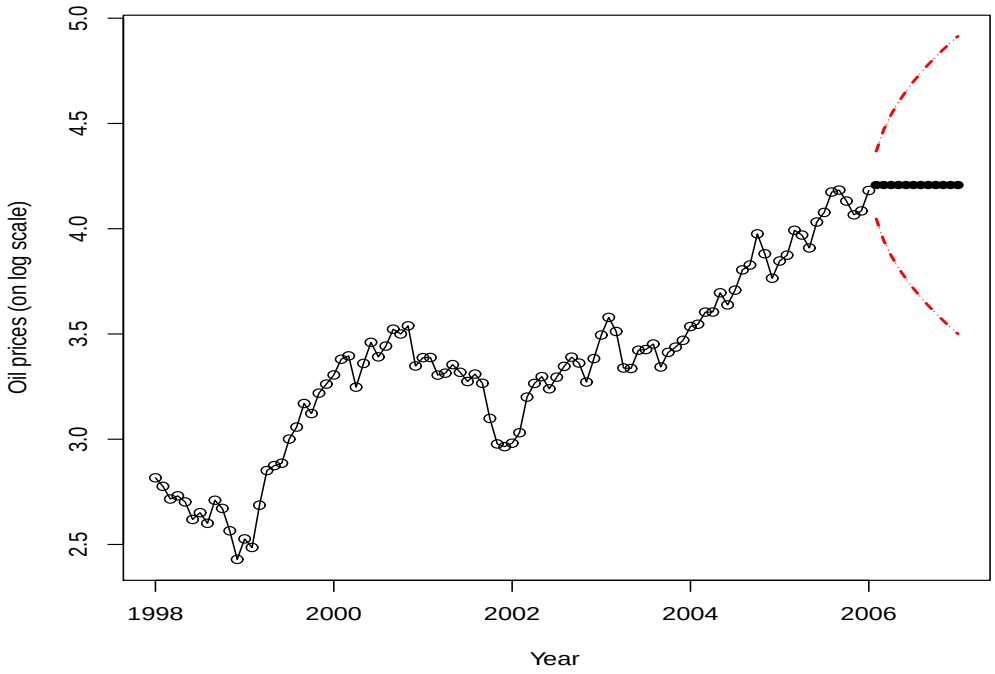


Figure 9.6: Oil price data (log-transformed). The full data set is from 1/86 to 1/06. This figure starts the series at 1/98. IMA(1,1) estimated MMSE forecasts and 95 percent prediction limits (on the log scale) are given for lead times $l = 1, 2, ..., 12$. These lead times correspond to months 2/06-1/07.

Example 9.7. In Example 1.6 (pp 7, notes), we examined the USC fall enrollment data (Columbia campus) from 1954-2010. An ARI(1,1) process provides a good fit to these data; fitting this model in R (using ML) gives the following output:

```
> enrollment.ari11.fit = arima(enrollment,order=c(1,1,0),method='ML')
> enrollment.ari11.fit
Coefficients:
         ar1
      0.3637
s.e.  0.1236
sigma^2 estimated as 1119849:  log likelihood = -469.54,  aic = 941.07
```

The fitted ARI(1,1) model is therefore

$$Y_t - Y_{t-1} = 0.3637(Y_{t-1} - Y_{t-2}) + e_t,$$

so that $\widehat{\phi} = 0.3637$ and the white noise variance estimate $\widehat{\sigma}_e^2 = 1119849$. The `predict` output from R is given below:

```
> enrollment.ari11.predict <- predict(enrollment.ari11.fit,n.ahead=10)
> round(enrollment.ari11.predict$pred,3)
Start = 2011
End = 2020
 [1] 28842.12 28973.44 29021.20 29038.56 29044.88 29047.18 29048.01 29048.32 29048.43
[10] 29048.47

> round(enrollment.ari11.predict$se,3)
Start = 2011
End = 2020
 [1] 1058.229 1789.494 2389.190 2894.460 3332.925 3723.059 4077.018 4402.947 4706.473
[10] 4991.615
```

- In Figure 9.7, we display the USC enrollment data. The full data set is from 1954-2010 (one enrollment count per year). However, to emphasize the MMSE forecasts in the plot, we start the series at year 1974.
- With $l = 1, 2, ..., 10$, the estimated MMSE forecasts in the `predict` output and in Figure 9.7 start at 2011 and end at 2020.
- From the `predict` output, note that the estimated MMSE forecasts for the next 10 years, based on the ARI(1,1) fit, fluctuate slightly.
- As expected from a nonstationary process, the estimated standard errors increase as l increases.

REMARK: As we have seen in the forecasting examples up to now, **prediction limits** are used to assess uncertainty in the calculated MMSE forecasts. We now discuss how these limits are obtained.

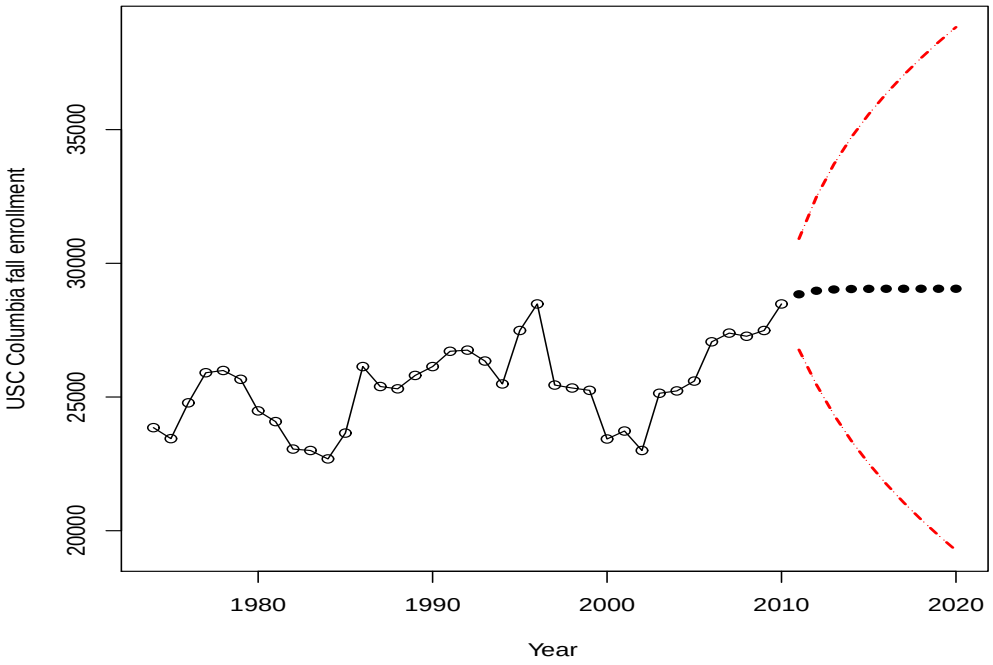


Figure 9.7: University of South Carolina fall enrollment data. The full data set is from 1954-2010. This figure starts the series at 1974. ARI(1,1) estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 10$. These lead times correspond to years 2011-2020.

9.4 Prediction intervals

TERMINOLOGY: A $100(1-\alpha)$ **percent prediction interval** for the Y_{t+l} is an interval $(\widehat{Y}_{t+l}^{(L)}, \widehat{Y}_{t+l}^{(U)})$ which satisfies

$$\text{pr}(\widehat{Y}_{t+l}^{(L)} < Y_{t+l} < \widehat{Y}_{t+l}^{(U)}) = 1 - \alpha.$$

We now derive prediction intervals for future responses with deterministic trend and ARIMA models.

NOTE: Prediction intervals and confidence intervals, while similar in spirit, have very different interpretations. A confidence interval is for a population (model) parameter, which is fixed. A prediction interval is constructed for a random variable.

9.4.1 Deterministic trend models

RECALL: Recall our deterministic trend model of the form

$$Y_t = \mu_t + X_t,$$

where μ_t is a non-random trend function and where we assume (for purposes of the current discussion) that $\{X_t\}$ is a **normally distributed** stochastic process with $E(X_t) = 0$ and $\text{var}(X_t) = \gamma_0$ (constant). We have already shown the following:

$$\begin{aligned} \widehat{Y}_t(l) &= \mu_{t+l} \\ E[e_t(l)] &= 0 \\ \text{var}[e_t(l)] &= \gamma_0, \end{aligned}$$

where $e_t(l) = Y_{t+l} - \widehat{Y}_t(l)$ is the l-step ahead prediction error. Under the assumption of normality, the random variable

$$Z = \frac{e_t(l)}{\sqrt{\text{var}[e_t(l)]}} = \frac{Y_{t+l} - \widehat{Y}_t(l)}{\sqrt{\text{var}[e_t(l)]}} = \frac{Y_{t+l} - \widehat{Y}_t(l)}{\text{se}[e_t(l)]} \sim \mathcal{N}(0,1).$$

Therefore, Z is a pivotal quantity and

$$\text{pr}\left(-z_{\alpha/2} < \frac{Y_{t+l} - \widehat{Y}_t(l)}{\text{se}[e_t(l)]} < z_{\alpha/2}\right) = 1 - \alpha.$$

Using algebra to rearrange the event inside the probability symbol, we have

$$\text{pr}\left(\widehat{Y}_t(l) - z_{\alpha/2}\text{se}[e_t(l)] < Y_{t+l} < \widehat{Y}_t(l) + z_{\alpha/2}\text{se}[e_t(l)]\right) = 1 - \alpha.$$

This shows that

$$\left(\widehat{Y}_t(l) - z_{\alpha/2}\text{se}[e_t(l)],\ \widehat{Y}_t(l) + z_{\alpha/2}\text{se}[e_t(l)]\right)$$

is a $100(1-\alpha)$ **percent prediction interval** for Y_{t+l}.

REMARK: The form the prediction interval includes the quantities $\widehat{Y}_t(l) = \mu_{t+l}$ and $\text{se}[e_t(l)] = \sqrt{\gamma_0}$. Of course, these are population parameters that must be estimated using the data.

Example 9.8. Consider the global temperature data from Example 3.4 (pp 53, notes). Fitting a linear deterministic trend model $Y_t = \beta_0 + \beta_1 t + X_t$, for $t = 1900, 1901, ..., 1997$, produces the following output in R:

```
Coefficients:             Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.219e+01  9.032e-01  -13.49   <2e-16 ***
time(globaltemps.1900)  6.209e-03  4.635e-04   13.40   <2e-16 ***

Residual standard error: 0.1298 on 96 degrees of freedom
Multiple R-squared: 0.6515,     Adjusted R-squared: 0.6479
F-statistic: 179.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

Suppose that $\{X_t\}$ is a normal white noise process with (constant) variance γ_0. The analysis in Section 3.5.1 (notes, pp 72-73) does support the normality assumption.

- The fitted model is

$$\widehat{Y}_t = -12.19 + 0.0062t,$$

  for $t = 1900, 1991, ..., 1997$.

- An **estimate** of the (assumed constant) variance of X_t is

$$\widehat{\gamma}_0 \approx (0.1298)^2 \approx 0.0168.$$

- The 1-step ahead MMSE forecast for 1998 is **estimated** to be

$$\widehat{Y}_t(1) = -12.19 + 0.0062(1997 + 1) \approx 0.198.$$

- Therefore, with

$$\widehat{\text{se}}[e_t(1)] \approx \sqrt{\widehat{\gamma}_0} \approx 0.1298,$$

  a 95 percent prediction interval for 1998 is

$$0.198 \pm 1.96 \times 0.1298 \implies (-0.056, 0.452).$$

- If we had made this prediction in 1997, we would have been 95 percent confident that the temperature deviation for 1998, Y_{1998}, falls between -0.056 and 0.452.

IMPORTANT: The formation of prediction intervals from deterministic trend models requires that the stochastic component X_t is **normally distributed** with **constant variance**. This may or may not be true in practice, but the validity of the prediction limits requires it to be true. Note also that since the margin of error

$$z_{\alpha/2}\text{se}[e_t(l)] = z_{\alpha/2}\sqrt{\gamma_0}$$

is free of l, prediction intervals have the same width indefinitely into the future.

9.4.2 ARIMA models

RECALL: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. In general, an ARIMA$(p, d, q)$ process can be written as

$$\phi(B)(1-B)^d Y_t = \theta_0 + \theta(B)e_t.$$

We have seen that the **l-step ahead forecast error** $e_t(l) = Y_{t+l} - \widehat{Y}_t(l)$ for any invertible ARIMA(p, d, q) model has the following characteristics:

$$\begin{aligned} E[e_t(l)] &= 0 \\ \text{var}[e_t(l)] &= \sigma_e^2 \sum_{j=0}^{l-1} \Psi_j^2, \end{aligned}$$

where the Ψ weights are unique to the specific model under investigation. If we additionally assume that the white noise process $\{e_t\}$ is **normally distributed**, then

$$Z = \frac{e_t(l)}{\sqrt{\text{var}[e_t(l)]}} = \frac{Y_{t+l} - \widehat{Y}_t(l)}{\sqrt{\text{var}[e_t(l)]}} = \frac{Y_{t+l} - \widehat{Y}_t(l)}{\text{se}[e_t(l)]} \sim \mathcal{N}(0,1).$$

This implies that

$$\left(\widehat{Y}_t(l) - z_{\alpha/2}\text{se}[e_t(l)],\ \widehat{Y}_t(l) + z_{\alpha/2}\text{se}[e_t(l)]\right)$$

is a $100(1-\alpha)$ **percent prediction interval** for Y_{t+l}. As we have seen in the examples so far, R gives (estimated) MMSE forecasts and standard errors; i.e., estimates of $\widehat{Y}_t(l)$ and se$[e_t(l)]$, so we can compute prediction intervals associated with any ARIMA(p, d, q) model. It is important to emphasize that normality is assumed.

Example 9.9. In Example 9.3, we examined the Lake Huron elevation data (from 1880-2006) and calculated the (estimated) MMSE forecasts based on an AR(1) model fit with lead times $l = 1, 2, ..., 20$. These forecasts, along with 95 percent prediction intervals (limits) were depicted visually in Figure 9.3. Here are the numerical values of these prediction intervals from R (I display only out to lead time $l = 10$ for brevity):

```
> # Create lower and upper prediction interval bounds
> lower.pi<-huron.ar1.predict$pred-qnorm(0.975,0,1)*huron.ar1.predict$se
> upper.pi<-huron.ar1.predict$pred+qnorm(0.975,0,1)*huron.ar1.predict$se
> # Display prediction intervals (2007-2026)
> data.frame(Year=c(2007:2026),lower.pi,upper.pi)
   Year lower.pi upper.pi
1  2007 579.6395 582.3978
2  2008 578.9851 582.6206
3  2009 578.5347 582.7004
4  2010 578.2000 582.7168
5  2011 577.9423 582.7014
6  2012 577.7395 582.6696
7  2013 577.5776 582.6301
8  2014 577.4469 582.5878
9  2015 577.3406 582.5456
10 2016 577.2534 582.5053
```

- In the R code, `$pred` extracts the estimated MMSE forecasts and `$se` extracts the estimated standard error of the forecast error. The expression `qnorm(0.975,0,1)` gives the upper 0.025 quantile of the $\mathcal{N}(0, 1)$ distribution (approximately 1.96).
- For example, we are 95 percent confident that the Lake Huron elevation level for 2015 will be between 577.3406 and 582.5456 feet.
- Note how the prediction limits (lower and upper) start to stabilize as the lead time l increases. This is typical of a **stationary** process. Prediction limits from nonstationary model fits do not stabilize as l increases.
- **Important:** The validity of prediction intervals depends on the white noise process $\{e_t\}$ being normally distributed.

9.5 Forecasting transformed series

9.5.1 Differencing

DISCOVERY: Calculating MMSE forecasts from nonstationary ARIMA models (i.e., $d \geq 1$) poses no additional methodological challenges beyond those of stationary ARMA models. It is easy to take this fact for granted, because as we have already seen, R automates the entire forecasting process for stationary and nonstationary models. However, it is important to understand why this is true so we investigate by means of an example.

EXAMPLE: Suppose that $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$ and consider the IMA(1,1) model

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1}.$$

The 1-step ahead MMSE forecast is

$$\begin{aligned}
\widehat{Y}_t(1) &= E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \\
&= E(Y_t + e_{t+1} - \theta e_t|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(Y_t|Y_1, Y_2, ..., Y_t)}_{=\, Y_t} + \underbrace{E(e_{t+1}|Y_1, Y_2, ..., Y_t)}_{=\, E(e_{t+1})=0} - \underbrace{E(\theta e_t|Y_1, Y_2, ..., Y_t)}_{=\, \theta e_t} \\
&= Y_t - \theta e_t.
\end{aligned}$$

The l-step ahead MMSE forecast, for $l > 1$, is

$$\begin{aligned}
\widehat{Y}_t(l) &= E(Y_{t+l}|Y_1, Y_2, ..., Y_t) \\
&= E(Y_{t+l-1} + e_{t+l} - \theta e_{t+l-1}|Y_1, Y_2, ..., Y_t) \\
&= \underbrace{E(Y_{t+l-1}|Y_1, Y_2, ..., Y_t)}_{=\, \widehat{Y}_t(l-1)} + \underbrace{E(e_{t+l}|Y_1, Y_2, ..., Y_t)}_{=\, E(e_{t+l})=0} - \underbrace{E(\theta e_{t+l-1}|Y_1, Y_2, ..., Y_t)}_{=\, E(e_{t+l-1})=0} \\
&= \widehat{Y}_t(l-1).
\end{aligned}$$

Therefore, we have shown that for the IMA(1,1) model, MMSE forecasts are

$$\widehat{Y}_t(l) = \begin{cases} Y_t - \theta e_t, & l = 1 \\ \widehat{Y}_t(l-1), & l > 1. \end{cases}$$

Now, let $W_t = \nabla Y_t = Y_t - Y_{t-1}$, so that W_t follows a zero-mean MA(1) model; i.e.,

$$W_t = e_t - \theta e_{t-1},$$

We have already shown that for an MA(1) process with $\mu = 0$,

$$\widehat{W}_t(l) = \begin{cases} -\theta e_t, & l = 1 \\ 0, & l > 1. \end{cases}$$

- When $l = 1$, note that

$$\widehat{W}_t(1) = -\theta e_t = \underbrace{Y_t - \theta e_t}_{= \widehat{Y}_t(1)} - Y_t = \widehat{Y}_t(1) - Y_t.$$

- When $l > 1$, note that

$$\widehat{W}_t(l) = 0 = \widehat{Y}_t(l) - \widehat{Y}_t(l-1).$$

Therefore, we have shown that with the IMA(1,1) model,

(a) forecasting the original **nonstationary** series Y_t

(b) forecasting the **stationary** differenced series $W_t = \nabla Y_t$ and then summing to obtain the forecast in original terms

are equivalent procedures. In fact, this equivalence holds when forecasting for any ARIMA(p, d, q) model!

- That is, the analyst can calculate predictions with the nonstationary model for Y_t or with the stationary model for $W_t = \nabla^d Y_t$ (and then convert back to the original scale by adding).
- The predictions in both cases will be equal (hence, the resulting standard errors will be the same too).
- The reason this occurs is that differencing is a **linear operation** (just as conditional expectation is).

9.5.2 Log-transformed series

RECALL: In Chapter 5, we discussed the **Box-Cox** family of transformations

$$T(Y_t) = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Y_t), & \lambda = 0, \end{cases}$$

where λ is the **transformation parameter**. Many time series processes $\{Y_t\}$ exhibit nonconstant variability that can be stabilized by taking logarithms. However, the function $T(x) = \ln x$ is not a linear function, so transformations on the log scale can not simply be "undone" as easily as with differenced series (differencing is a linear transformation). MMSE forecasts are not preserved under exponentiation.

THEORY: For notational purposes, set

$$Z_t = \ln Y_t,$$

and denote the MMSE forecast for Z_{t+l} by $\widehat{Z}_t(l)$, that is, $\widehat{Z}_t(l)$ is the l-step ahead MMSE forecast on the **log scale**.

- The MMSE forecast for Y_{t+l} is **not** $\widehat{Y}_t(l) = e^{\widehat{Z}_t(l)}$!! This is sometimes called the **naive forecast** of Y_{t+l}.
- The theoretical argument on pp 210 (CC) shows that the corresponding MMSE forecast of Y_{t+l} is

$$\widehat{Y}_t(l) = \exp\left\{\widehat{Z}_t(l) + \frac{1}{2}\text{var}[e_t(l)]\right\},$$

where $\text{var}[e_t(l)]$ is the variance of the l-step ahead forecast error $e_t(l) = Z_{t+l} - \widehat{Z}_t(l)$.

Example 9.10. In Example 9.6, we examined the monthly oil price data (1/86-1/01) and we computed MMSE forecasts and predictions limits for $l = 1, 2, ..., 12$ (i.e., for 2/06 to 1/07), based on an IMA(1,1) fit for $Z_t = \ln Y_t$. The estimated MMSE forecasts (on the log scale) are depicted visually in Figure 9.6. The estimated MMSE forecasts, both on the log scale and on the original scale (back-transformed), are given below:

```
> ima11.log.oil.predict <- predict(ima11.log.oil.fit,n.ahead=12)
> round(ima11.log.oil.predict$pred,3)
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2006       4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208 4.208
2007 4.208
> round(ima11.log.oil.predict$se,3)
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2006       0.082 0.134 0.171 0.201 0.227 0.251 0.272 0.292 0.311 0.328 0.345
2007 0.361
> # MMSE forecasts back-transformed (to original scale)
> oil.price.predict <-
    round(exp(ima11.log.oil.predict$pred + (1/2)*(ima11.log.oil.predict$se)^2),3)
> oil.price.predict
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2006        67.417 67.796 68.178 68.562 68.948 69.336 69.726 70.119 70.513 70.910 71.310
2007 71.711
```

For example, the MMSE forecast (on the original scale) for June, 2006 is given by

$$\widehat{Y}_t(5) = \exp\left\{4.208 + \frac{1}{2}(0.227)^2\right\} \approx 68.948.$$

NOTE: A $100(1-\alpha)$ **percent prediction interval** for Y_{t+l} can be formed by exponentiating the endpoints of the prediction interval for $Z_{t+l} = \log Y_{t+l}$. This is true because

$$1-\alpha = \text{pr}(\widehat{Z}_{t+l}^{(L)} < Z_{t+l} < \widehat{Z}_{t+l}^{(U)}) = \text{pr}\left(e^{\widehat{Z}_{t+l}^{(L)}} < Y_{t+l} < e^{\widehat{Z}_{t+l}^{(U)}}\right);$$

that is, because the exponential function $f(x) = e^x$ is strictly increasing, the two probabilities above are the same.

- For example, a 95 percent prediction interval for June, 2005 (on the log scale) is

$$4.208 \pm 1.96(0.227) \quad \Longrightarrow \quad (3.763, 4.653).$$

- A 95 percent prediction interval for June, 2005 on the original scale (in dollars) is

$$(e^{3.763}, e^{4.653}) \quad \Longrightarrow \quad (43.08, 104.90).$$

  Therefore, we are 95 percent confident that the June, 2006 oil price (had we made this prediction in January, 2006) would fall between 43.08 and 104.90 dollars.

10 Seasonal ARIMA Models

Complementary reading: Chapter 10 (CC).

10.1 Introduction

PREVIEW: In this chapter, we introduce new ARIMA models that incorporate **seasonal** patterns occurring over time. With seasonal data, dependence with the past occurs most prominently at multiples of an underlying **seasonal lag**, denoted by s. Consider the following examples:

- With **monthly** data, there can be strong autocorrelation at lags that are multiples of $s = 12$. For example, January observations tend to be "alike" across years, February observations tend to be "alike," and so on.
- With **quarterly** data, there can be strong autocorrelation at lags that are multiples of $s = 4$. For example, first quarter sales tend to be "alike" across years, second quarter sales tend to be "alike," and so on.

UBIQUITY: Many physical, biological, epidemiological, and economic processes tend to elicit seasonal patterns over time. We therefore wish to study new time series models which can account explicitly for these types of patterns. We refer to this new class of models generally as **seasonal ARIMA models**.

Example 10.1. In Example 1.2 (pp 3, notes), we examined the monthly U.S. milk production data (in millions of pounds) from January, 1994 to December, 2005.

- In Figure 10.1, we see that there are two types of trend in the milk production data:
  - an upward **linear** trend (across the years)
  - a **seasonal** trend (within years).

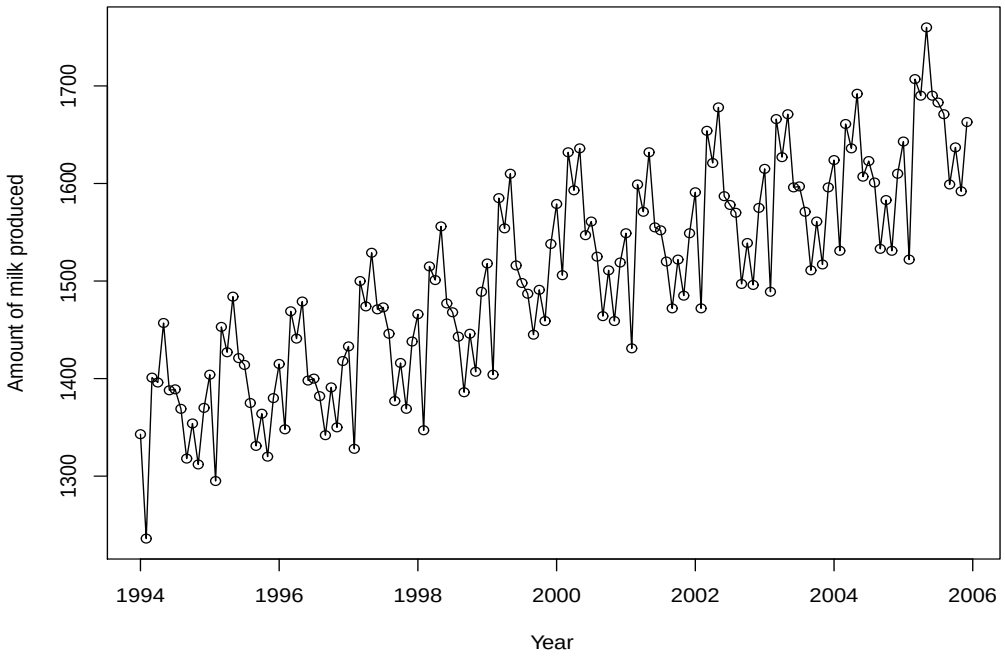


Figure 10.1: United States milk production data. Monthly production figures, measured in millions of pounds, from January, 1994 to December, 2005.

- We know the upward linear trend can be "removed" by working with first differences $\nabla Y_t = Y_t - Y_{t-1}$. This is how we removed linear trends with nonseasonal data.
- Figure 10.2 displays the series of first differences ∇Y_t. From this plot, it is clear that the upward linear trend over time has been removed. That is, the first differences ∇Y_t look stationary in the mean level.
- However, the first difference process $\{\nabla Y_t\}$ still displays a pronounced **seasonal** pattern that repeats itself every $s = 12$ months. This is easily seen from the monthly plotting symbols that I have added. How can we "handle" this type of pattern? Is it possible to "remove" it as well?

GOAL: We wish to enlarge our class of ARIMA(p, d, q) models to handle seasonal data such as these.

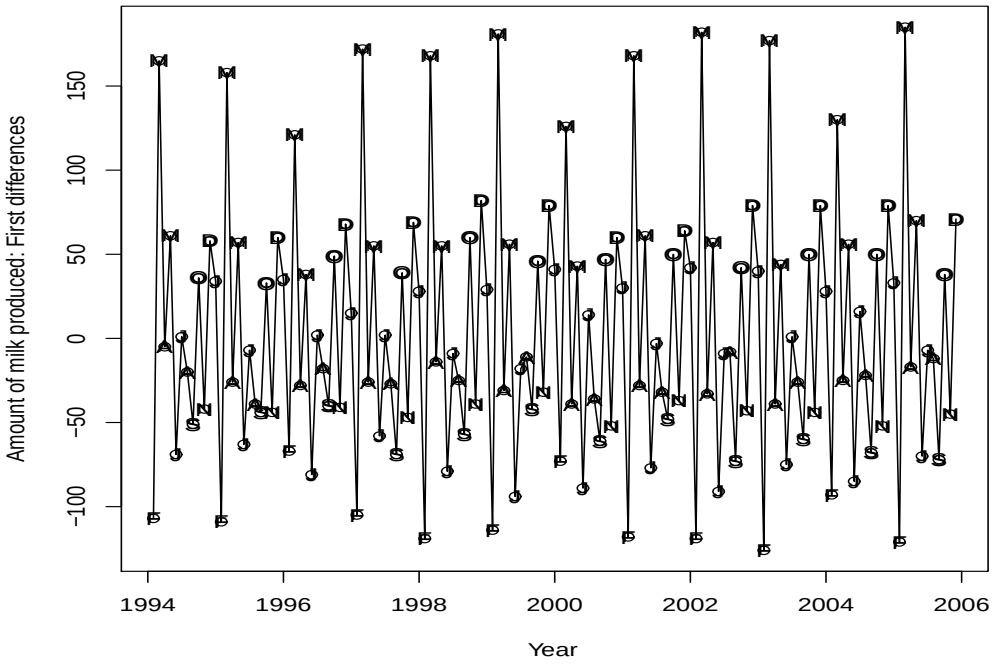


Figure 10.2: United States milk production data. First differences $\nabla Y_t = Y_t - Y_{t-1}$. Monthly plotting symbols have been added.

10.2 Purely seasonal (stationary) ARMA models

10.2.1 MA$(Q)_s$

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. A **seasonal moving average (MA) model** of **order** Q with **seasonal period** s, denoted by MA$(Q)_s$, is

$$Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs}.$$

A nonzero mean μ could be added for flexibility (as with nonseasonal models), but we take $\mu = 0$ for simplicity.

MA$(1)_{12}$: When $Q = 1$ and $s = 12$, we have

$$Y_t = e_t - \Theta e_{t-12}.$$

CALCULATIONS: For an $\text{MA}(1)_{12}$ process, note that

$$\mu = E(Y_t) = E(e_t - \Theta e_{t-12}) = E(e_t) - \Theta E(e_{t-12}) = 0.$$

The process variance is

$$\begin{aligned}\gamma_0 = \text{var}(Y_t) &= \text{var}(e_t - \Theta e_{t-12}) \\ &= \text{var}(e_t) + \Theta^2 \text{var}(e_{t-12}) - 2\Theta \underbrace{\text{cov}(e_t, e_{t-12})}_{=0} \\ &= \sigma_e^2 + \Theta^2\sigma_e^2 = \sigma_e^2(1+\Theta^2).\end{aligned}$$

The lag 1 autocorrelation is

$$\gamma_1 = \text{cov}(Y_t, Y_{t-1}) = \text{cov}(e_t - \Theta e_{t-12}, e_{t-1} - \Theta e_{t-13}) = 0,$$

because no white noise subscripts match. In fact, it is easy to see that $\gamma_k = 0$ for all k, except when $k = s = 12$. Note that

$$\begin{aligned}\gamma_{12} = \text{cov}(Y_t, Y_{t-12}) &= \text{cov}(e_t - \Theta e_{t-12}, e_{t-12} - \Theta e_{t-24}) \\ &= -\Theta \text{var}(e_{t-12}) = -\Theta\sigma_e^2.\end{aligned}$$

Therefore, the autocovariance function for an $\text{MA}(1)_{12}$ process is

$$\gamma_k = \begin{cases} \sigma_e^2(1+\Theta^2), & k = 0 \\ -\Theta\sigma_e^2, & k = 12 \\ 0, & \text{otherwise.} \end{cases}$$

Because $E(Y_t) = 0$ and γ_k are both free of t, **an $\text{MA}(1)_{12}$ process is stationary.** The autocorrelation function (ACF) for an $\text{MA}(1)_{12}$ process is

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k = 0 \\ -\dfrac{\Theta}{1+\Theta^2}, & k = 12 \\ 0, & \text{otherwise.} \end{cases}$$

NOTE: The form of the $\text{MA}(1)_{12}$ ACF is identical to the form of the nonseasonal MA(1) ACF from Chapter 4. For the $\text{MA}(1)_{12}$, the only nonzero autocorrelation occurs at the **first seasonal lag** $k = 12$, as opposed to at $k = 1$ in the nonseasonal MA(1).

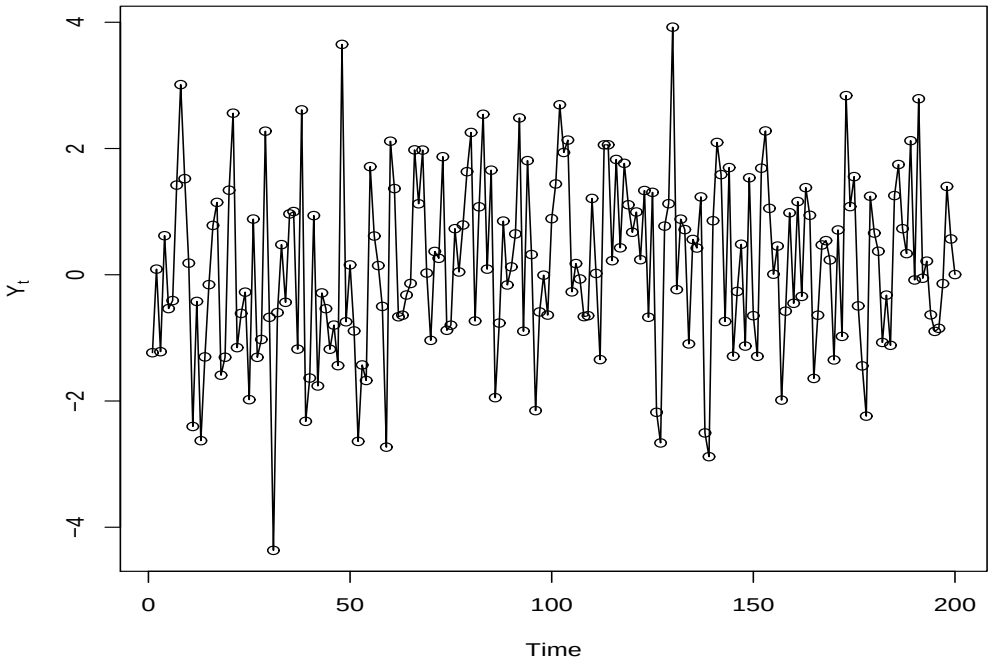


Figure 10.3: $\text{MA}(1)_{12}$ simulation with $\Theta = -0.9$, $n = 200$, and $\sigma_e^2 = 1$.

NOTE: A seasonal $\text{MA}(1)_{12}$ process is mathematically equivalent to a nonseasonal MA(12) process with

$$\theta_1 = \theta_2 = \cdots = \theta_{11} = 0$$

and $\theta_{12} = \Theta$. Because of this equivalence (which occurs here and with other seasonal models), we can use our already-established methods to specify, fit, diagnose, and forecast seasonal models.

Example 10.2. We use R to simulate one realization of an $\text{MA}(1)_{12}$ process with $\Theta = -0.9$, that is,

$$Y_t = e_t + 0.9e_{t-12},$$

where $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 200$. This realization is displayed in Figure 10.3. In Figure 10.4, we display the population (theoretical) ACF and PACF for this $\text{MA}(1)_{12}$ process and the sample versions that correspond to the simulation in Figure 10.3.

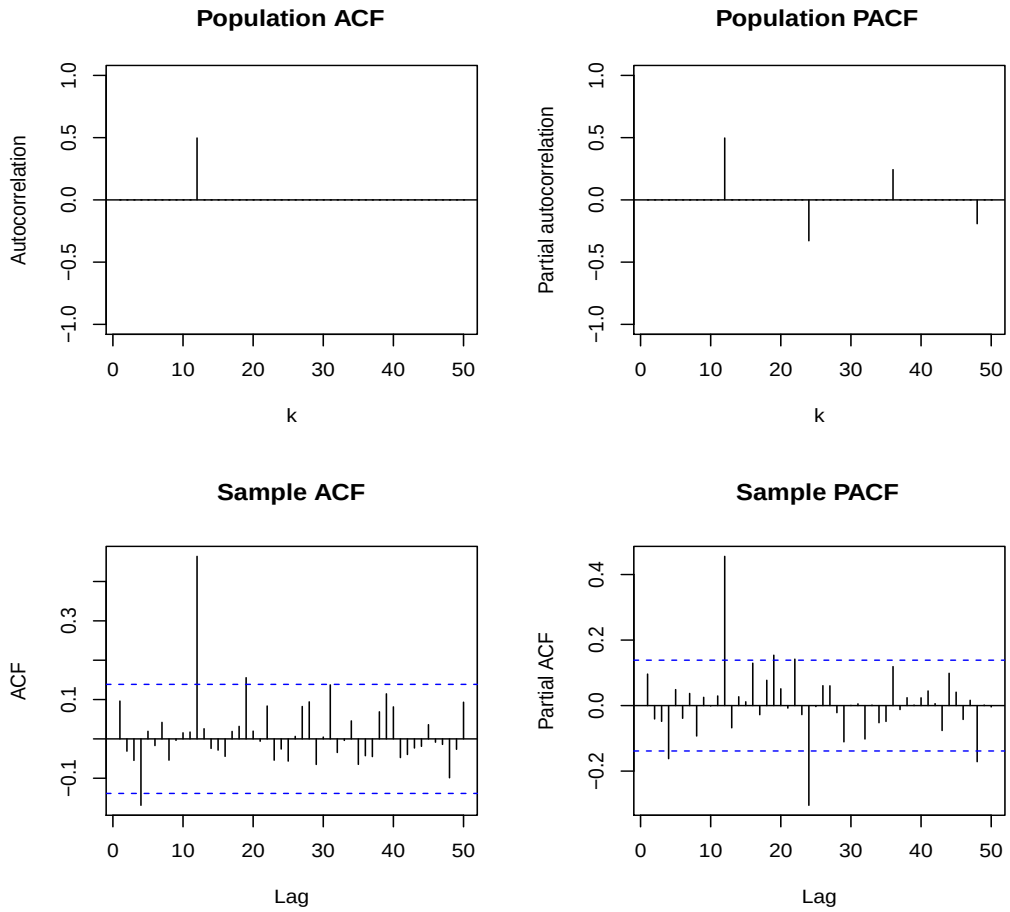


Figure 10.4: $\text{MA}(1)_{12}$ with $\Theta = -0.9$. Upper left: Population ACF. Upper right: Population PACF. Lower left (right): Sample ACF (PACF) using data in Figure 10.3.

- The population ACF and PACF (Figure 10.4; top) display the same patterns as the nonseasonal MA(1), except that now these patterns occur at **seasonal lags**.
  - The population ACF displays nonzero autocorrelation only at the first (seasonal) lag $k = 12$. In other words, observations 12 units apart in time are correlated, whereas all other observations are not.
  - The population PACF shows a decay across seasonal lags $k = 12, 24, 36, ...,$.
- The sample ACF and PACF reveal these same patterns overall. Margin of error bounds in the sample ACF/PACF are for white noise; not an $\text{MA}(1)_{12}$ process.

MA$(2)_{12}$: A seasonal MA model of order $Q = 2$ with seasonal lag $s = 12$ is

$$Y_t = e_t - \Theta_1 e_{t-12} - \Theta_2 e_{t-24}.$$

For an MA$(2)_{12}$ process, is easy to show that $E(Y_t) = 0$ and

$$\gamma_k = \begin{cases} \sigma_e^2(1 + \Theta_1^2 + \Theta_2^2), & k = 0 \\ (-\Theta_1 + \Theta_1\Theta_2)\sigma_e^2, & k = 12 \\ -\Theta_2\sigma_e^2, & k = 24 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, **an MA**$(2)_{12}$ **process is stationary**. The autocorrelation function (ACF) for an MA$(2)_{12}$ process is

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k = 0 \\ \dfrac{-\Theta_1 + \Theta_1\Theta_2}{1 + \Theta_1^2 + \Theta_2^2}, & k = 12 \\ \dfrac{-\Theta_2}{1 + \Theta_1^2 + \Theta_2^2}, & k = 24 \\ 0, & \text{otherwise.} \end{cases}$$

NOTE: The ACF for an MA$(2)_{12}$ process has the same form as the ACF for a nonseasonal MA(2). The only difference is that nonzero autocorrelations occur at the first **two seasonal lags** $k = 12$ and $k = 24$, as opposed to at $k = 1$ and $k = 2$ in the nonseasonal MA(2).

NOTE: A seasonal MA$(2)_{12}$ process is mathematically equivalent to a nonseasonal MA(24) process with

$$\theta_1 = \theta_2 = \cdots = \theta_{11} = \theta_{13} = \theta_{14} = \cdots = \theta_{23} = 0,$$

$\theta_{12} = \Theta_1$, and $\theta_{24} = \Theta_2$. This again reveals that we can use our already-established methods to specify, fit, diagnose, and forecast seasonal models.

BACKSHIFT NOTATION: In general, a seasonal MA$(Q)_s$ process

$$Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs}$$

can be expressed using backshift notation as

$$\begin{aligned} Y_t &= e_t - \Theta_1 B^s e_t - \Theta_2 B^{2s} e_t - \cdots - \Theta_Q B^{Qs} e_t \\ &= (1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}) e_t \equiv \Theta_Q(B^s) e_t, \end{aligned}$$

where $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$ is called the **seasonal MA characteristic operator**. The operator $\Theta_Q(B^s)$ can be viewed as a polynomial (in B) of degree Qs.

- As with nonseasonal processes, a seasonal MA$(Q)_s$ process is **invertible** if and only if each of the Qs roots of $\Theta_Q(B^s)$ exceed 1 in absolute value (or modulus).
- All seasonal MA$(Q)_s$ processes are **stationary**.

10.2.2 AR$(P)_s$

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. A **seasonal autoregressive (AR) model** of **order** P with **seasonal period** s, denoted by AR$(P)_s$, is

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t.$$

A nonzero mean μ could be added for flexibility (as with nonseasonal models), but we take $\mu = 0$ for simplicity.

AR$(1)_{12}$: When $P = 1$ and $s = 12$, we have

$$Y_t = \Phi Y_{t-12} + e_t.$$

- Similar to a nonseasonal AR(1) process, a seasonal AR$(1)_{12}$ process is **stationary** if and only if $-1 < \Phi < 1$. An AR$(1)_{12}$ process is automatically **invertible**.
- For an AR$(1)_{12}$ process,

$$\begin{aligned} E(Y_t) &= 0 \\ \gamma_0 = \text{var}(Y_t) &= \frac{\sigma_e^2}{1 - \Phi^2}. \end{aligned}$$

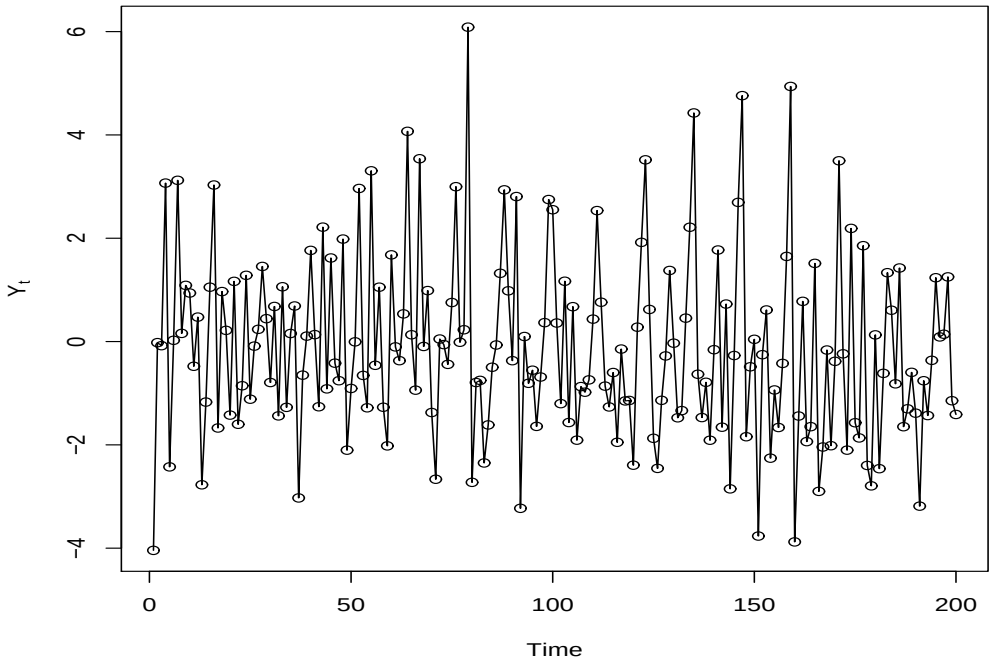


Figure 10.5: $\text{AR}(1)_{12}$ simulation with $\Phi = 0.9$, $n = 200$, and $\sigma_e^2 = 1$.

- The $\text{AR}(1)_{12}$ autocorrelation function (ACF) is given by

$$\rho_k = \begin{cases} \Phi^{k/12}, & k = 0, 12, 24, 36, ..., \\ 0, & \text{otherwise.} \end{cases}$$

- That is, $\rho_0 = 1$, $\rho_{12} = \Phi$, $\rho_{24} = \Phi^2$, $\rho_{36} = \Phi^3$, and so on, similar to the nonseasonal AR(1). The ACF $\rho_k = 0$ at all lags k that are not multiples of $s = 12$.

- A seasonal $\text{AR}(1)_{12}$ process is mathematically equivalent to a nonseasonal AR(12) process with

$$\phi_1 = \phi_2 = \cdots = \phi_{11} = 0$$

and $\phi_{12} = \Phi$.

Example 10.3. We use R to simulate one realization of an $\text{AR}(1)_{12}$ process with $\Phi = 0.9$, that is,

$$Y_t = 0.9Y_{t-12} + e_t,$$

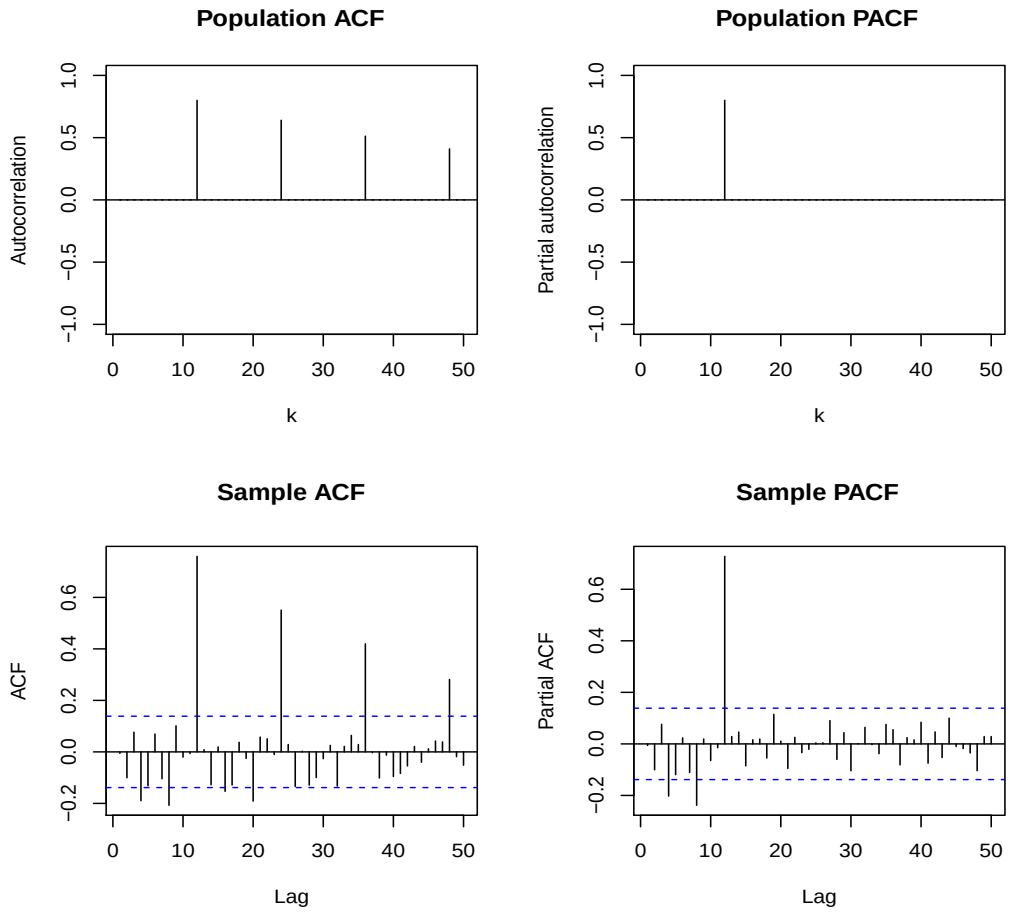


Figure 10.6: $\text{AR}(1)_{12}$ with $\Phi = -0.9$. Upper left: Population ACF. Upper right: Population PACF. Lower left (right): Sample ACF (PACF) using data in Figure 10.5.

where $e_t \sim$ iid $\mathcal{N}(0, 1)$ and $n = 200$. This realization is displayed in Figure 10.5. In Figure 10.6, we display the population (theoretical) ACF and PACF for this $\text{AR}(1)_{12}$ process and the sample versions that correspond to the simulation in Figure 10.5.

- The population ACF and PACF (Figure 10.6; top) display the same patterns as the nonseasonal AR(1), except that now these patterns occur at **seasonal lags**.
  - The population ACF displays a slow decay across the seasonal lags $k = 12, 24, 36, 48, ...,$. In other words, observations that are 12, 24, 36, 48, etc. units apart in time are correlated, whereas all other observations are not.

  - The population PACF is nonzero at the first seasonal lag $k = 12$. The PACF is zero at all other lags. This is analogous to the PACF for an AR(1) being nonzero when $k = 1$ and zero elsewhere.

- The sample ACF and PACF reveal these same patterns overall. Margin of error bounds in the sample ACF/PACF are for white noise; not an $\text{AR}(1)_{12}$ process.

AR$(2)_{12}$: A seasonal AR model of order $P = 2$ with seasonal lag $s = 12$; i.e., $\text{AR}(2)_{12}$, is

$$Y_t = \Phi Y_{t-12} + \Phi_2 Y_{t-24} + e_t.$$

- A seasonal $\text{AR}(2)_{12}$ behaves like the nonseasonal AR(2) at the seasonal lags.
  - In particular, the ACF ρ_k displays exponential decay or damped sinusoidal patterns across the seasonal lags $k = 12, 24, 36, 48, ...,$.
  - The PACF ϕ_{kk} is nonzero at lags $k = 12$ and $k = 24$; it is zero at all other lags.

- A seasonal $\text{AR}(2)_{12}$ process is mathematically equivalent to a nonseasonal AR(24) process with

$$\phi_1 = \phi_2 = \cdots = \phi_{11} = \phi_{13} = \phi_{14} = \cdots = \phi_{23} = 0,$$

  $\phi_{12} = \Phi_1$, and $\phi_{24} = \Phi_2$.

BACKSHIFT NOTATION: In general, a seasonal $\text{AR}(P)_s$ process

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t$$

can be expressed as

$$\begin{aligned} & Y_t - \Phi_1 Y_{t-s} - \Phi_2 Y_{t-2s} - \cdots - \Phi_P Y_{t-Ps} = e_t \\ \iff & (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}) Y_t = e_t \iff \Phi_P(B^s) Y_t = e_t, \end{aligned}$$

where $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ is the **seasonal AR characteristic operator**. The operator $\Phi_P(B^s)$ can be viewed as a polynomial (in B) of degree Ps.

- As with nonseasonal processes, a seasonal AR$(P)_s$ process is **stationary** if and only if each of the Ps roots of $\Phi_P(B^s)$ exceed 1 in absolute value (or modulus).
- All seasonal AR$(P)_s$ processes are **invertible**.

10.2.3 ARMA$(P, Q)_s$

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. A **seasonal autoregressive moving average (ARMA) model** of **orders** P and Q with **seasonal period** s, denoted by ARMA$(P, Q)_s$, is

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs}.$$

A nonzero mean μ could be added for flexibility (as with nonseasonal models), but we take $\mu = 0$ for simplicity.

- An ARMA$(P, Q)_s$ process is the seasonal analogue of the nonseasonal ARMA(p, q) process with nonzero autocorrelations at lags $k = s, 2s, 3s, ...,$.
- Using backshift notation, this model can be expressed as

$$\Phi_P(B^s)Y_t = \Theta_Q(B^s)e_t,$$

where the seasonal AR and MA characteristic operators are

$$\begin{aligned} \Phi_P(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \\ \Theta_Q(B^s) &= 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}. \end{aligned}$$

- Analogous to a nonseasonal ARMA(p, q) process,
  - the ARMA$(P, Q)_s$ process is **stationary** if and only if the roots of $\Phi_P(B^s)$ each exceed 1 in absolute value (or modulus)
  - the ARMA$(P, Q)_s$ process is **invertible** if and only if the roots of $\Theta_Q(B^s)$ each exceed 1 in absolute value (or modulus).

- A seasonal ARMA$(P,Q)_s$ process is mathematically equivalent to a nonseasonal ARMA(Ps,Qs) process with

$$\phi_s = \Phi_1,\ \phi_{2s} = \Phi_2, ...,\ \phi_{Ps} = \Phi_P, \quad \theta_s = \Theta_1,\ \theta_{2s} = \Theta_2, ...,\ \theta_{Qs} = \Theta_Q,$$

  and all other ϕ and θ parameters set equal to 0.

- The following table succinctly summarizes the behavior of the population ACF and PACF for seasonal ARMA$(P,Q)_s$ processes:

	AR$(P)_s$	MA$(Q)_s$	ARMA$(P,Q)_s$
ACF	Tails off at lags ks	Cuts off after	Tails off at lags ks
	$k = 1, 2, ...,$	lag Qs	$k = 1, 2, ..., s$
PACF	Cuts off after	Tails off at lags ks	Tails off at lags ks
	lag Ps	$k = 1, 2, ...,$	$k = 1, 2, ...,$

SUMMARY: We have broadened the class of stationary ARMA(p,q) models to incorporate the same type of ARMA(p,q) behavior at seasonal lags, the so-called the **seasonal ARMA$(P,Q)_s$ class** of models.

- In many ways, this "extension" is not that much of an extension, because the seasonal ARMA$(P,Q)_s$ model is essentially an ARMA(p,q) model restricted the seasonal lags $k = s, 2s, 3s, ...,$.

- That is, an ARMA$(P,Q)_s$ model, which incorporates autocorrelation at seasonal lags and nowhere else, is likely limited in application for stationary processes.

- However, if we combine these new seasonal ARMA$(P,Q)_s$ models with our traditional nonseasonal ARMA(p,q) models, we create a larger class of models applicable for use with **stationary** processes that exhibit seasonality.

- We now examine this new class of models, the so-called **multiplicative seasonal ARMA class**.

10.3 Multiplicative seasonal (stationary) ARMA models

MA$(1) \times$ **MA**$(1)_{12}$: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Consider the **nonseasonal** MA(1) model

$$Y_t = e_t - \theta e_{t-1} \quad \Longleftrightarrow \quad Y_t = (1 - \theta B)e_t$$

and the **seasonal** MA$(1)_{12}$ model

$$Y_t = e_t - \Theta e_{t-12} \quad \Longleftrightarrow \quad Y_t = (1 - \Theta B^{12})e_t.$$

- The defining characteristic of the nonseasonal MA(1) process is that the only nonzero autocorrelation occurs at lag $k = 1$.
- The defining characteristic of the seasonal MA$(1)_{12}$ process is that the only nonzero autocorrelation occurs at lag $k = 12$.

COMBINING THE MODELS: Consider taking the nonseasonal MA characteristic operator $\theta(B) = 1 - \theta B$ and the nonseasonal one $\Theta(B) = 1 - \Theta B^{12}$ and multiplying them together to get the new model

$$\begin{aligned} Y_t &= (1 - \theta B)(1 - \Theta B^{12})e_t \\ &= (1 - \theta B - \Theta B^{12} + \theta\Theta B^{13})e_t, \end{aligned}$$

or, equivalently,

$$Y_t = e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13}.$$

We call this a **multiplicative seasonal MA**$(1) \times$ **MA**$(1)_{12}$ **model**. The term "multiplicative" arises because the MA characteristic operator $1 - \theta B - \Theta B^{12} + \theta\Theta B^{13}$ is the product of $(1 - \theta B)$ and $(1 - \Theta B^{12})$. An MA$(1) \times$ MA$(1)_{12}$ process has $E(Y_t) = 0$ and

$$\rho_1 = -\frac{\theta}{1 + \theta^2} \qquad \rho_{11} = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)}$$

$$\rho_{12} = -\frac{\Theta}{1 + \Theta^2} \qquad \rho_{13} = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)}.$$

- The MA(1)$\times$MA(1)$_{12}$ process has nonzero autocorrelation at lags $k = 1$ and $k = 12$ from the nonseasonal and seasonal MA models individually.
- It has additional nonzero autocorrelation at lags $k = 11$ and $k = 13$ which arises from the multiplicative effect of the two models.
- The MA(1) $\times$ MA(1)$_{12}$ process

$$Y_t = e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13}$$

  is mathematically equivalent to a nonseasonal MA(13) process with parameters $\theta_1 = \theta$, $\theta_2 = \theta_3 = \cdots = \theta_{11} = 0$, $\theta_{12} = \Theta$, and $\theta_{13} = -\theta\Theta$.

MA(1) $\times$ **AR**(1)$_{12}$: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Consider the two models

$$Y_t = e_t - \theta e_{t-1} \quad \Longleftrightarrow \quad Y_t = (1 - \theta B)e_t$$

and

$$Y_t = \Phi Y_{t-12} + e_t \quad \Longleftrightarrow \quad (1 - \Phi B^{12})Y_t = e_t,$$

a nonseasonal MA(1) and a seasonal AR(1)$_{12}$, respectively.

- The defining characteristic of the nonseasonal MA(1) is that the only nonzero autocorrelation occurs at lag $k = 1$.
- The defining characteristic of the seasonal AR(1)$_{12}$ is that the autocorrelation decays across seasonal lags $k = 12, 24, 36, ...,$.

COMBINING THE MODELS: Consider combining these two models to form

$$(1 - \Phi B^{12})Y_t = (1 - \theta B)e_t,$$

or, equivalently,

$$Y_t = \Phi Y_{t-12} + e_t - \theta e_{t-1}.$$

We call this a **multiplicative seasonal MA**(1) $\times$ **AR**(1)$_{12}$ **process**. By combining a nonseasonal MA(1) with a seasonal AR(1)$_{12}$, we create a new process which possesses the following:

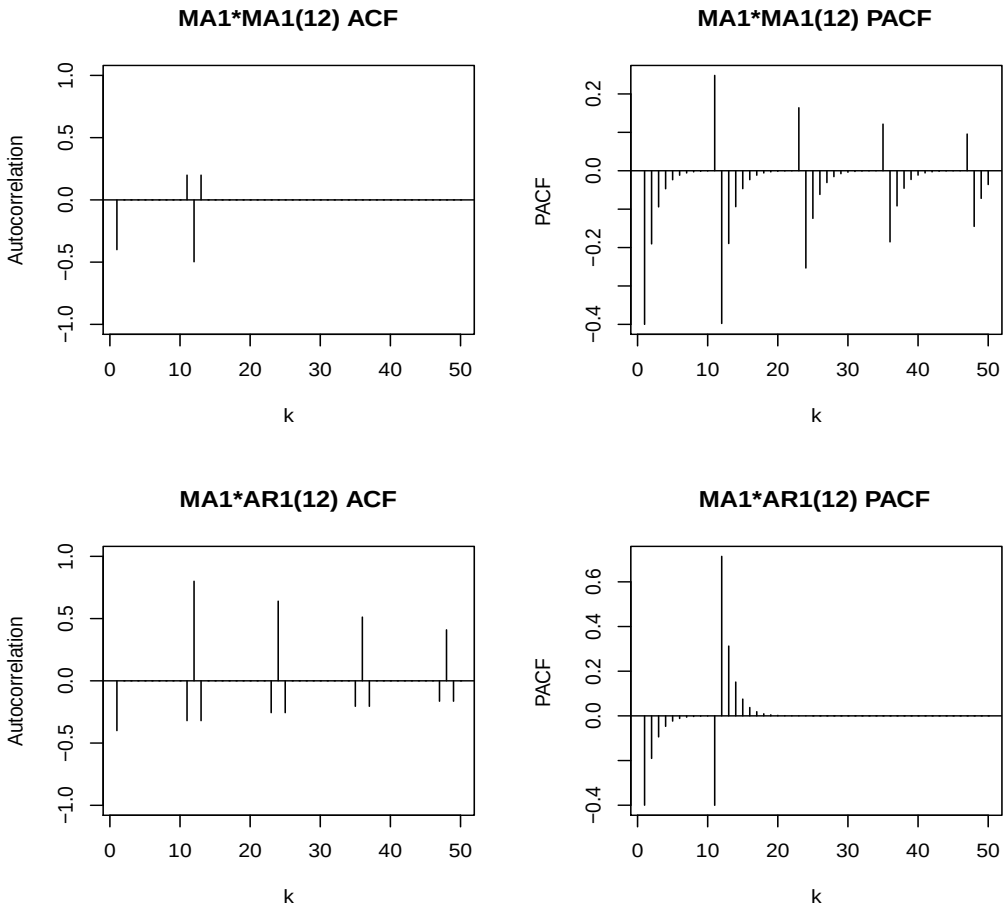


Figure 10.7: Top: Population ACF/PACF for MA(1)$\times$MA(1)$_{12}$ process with $\theta = 0.5$ and $\Theta = 0.9$. Bottom: Population ACF/PACF for MA(1) $\times$ AR(1)$_{12}$ process with $\theta = 0.5$ and $\Phi = 0.9$.

- AR-type autocorrelation at seasonal lags $k = 12, 24, 36, ...,$
- additional MA-type autocorrelation at lag $k = 1$ and at lags one unit in time from the seasonal lags, that is, at $k = 11$ and $k = 13$, $k = 23$ and $k = 25$, and so on.
- The MA(1) $\times$ AR(1)$_{12}$ process

$$Y_t = \Phi Y_{t-12} + e_t - \theta e_{t-1},$$

is mathematically equivalent to a nonseasonal ARMA(12,1) process with parameters θ, $\phi_1 = \phi_2 = \cdots = \phi_{11} = 0$, and $\phi_{12} = \Phi$.

TERMINOLOGY: Suppose $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. In general, we can combine a nonseasonal ARMA(p, q) process

$$\phi(B)Y_t = \theta(B)e_t$$

with a seasonal ARMA$(P, Q)_s$ process

$$\Phi_P(B^s)Y_t = \Theta_Q(B^s)e_t$$

to create the model

$$\phi(B)\Phi_P(B^s)Y_t = \theta(B)\Theta_Q(B^s)e_t.$$

We call this a **multiplicative seasonal (stationary) ARMA**$(p, q) \times$ **ARMA**$(P, Q)_s$ **model** with **seasonal period** s.

- This is a very flexible family of models for **stationary** seasonal processes.
  - The MA(1)$\times$MA(1)$_{12}$ and MA(1)$\times$AR(1)$_{12}$ processes (that we have discussed explicitly) are special cases.
- An ARMA$(p, q) \times$ ARMA$(P, Q)_s$ process is mathematically equivalent to a nonseasonal ARMA process with AR characteristic operator $\phi^*(B) = \phi(B)\Phi_P(B^s)$ and MA characteristic operator $\theta^*(B) = \theta(B)\Theta_Q(B^s)$.
  - Stationarity and invertibility conditions can be characterized in terms of the roots of $\phi^*(B)$ and $\theta^*(B)$, respectively.
- Because of this equivalence, we can use our already-established methods to specify, fit, diagnose, and forecast seasonal stationary models.

Example 10.4. Data file: `boardings` (TSA). Figure 10.8 displays the number of public transit boardings (mostly for bus and light rail) in Denver, Colorado from 8/2000 to 3/2006. The data have been log-transformed.

- From the plot, the boarding process appears to be relatively stationary in the mean level; that is, there are no pronounced shifts in mean level over time.

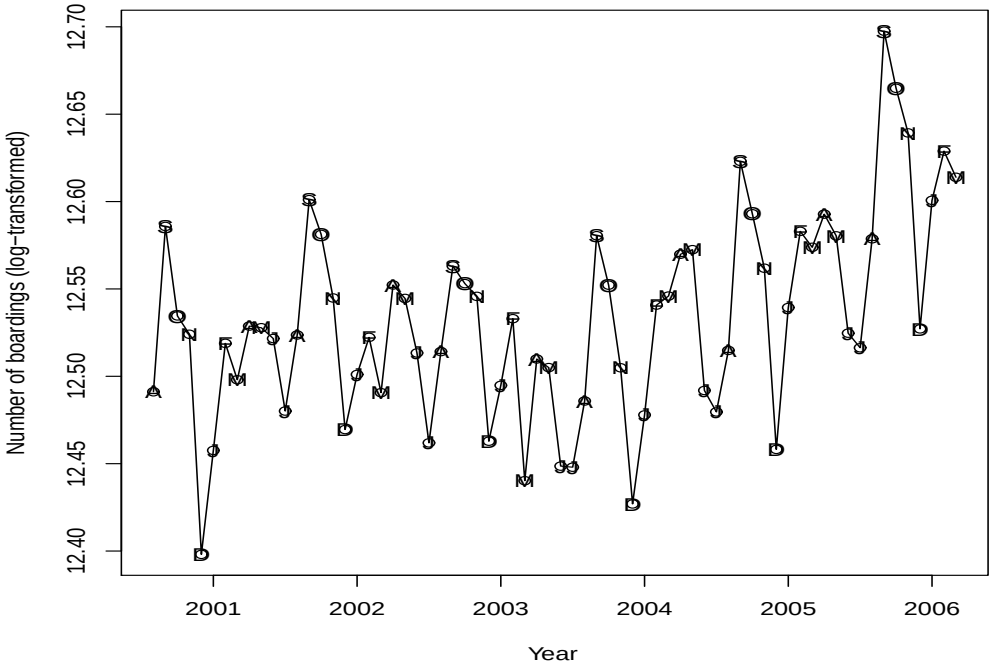


Figure 10.8: Denver public transit data. Monthly number of public transit boardings (log-transformed) in Denver from 8/2000 to 3/2006. Monthly plotting symbols have been added.

- Therefore, a member of the stationary ARMA$(p, q) \times$ ARMA$(P, Q)_s$ family of models may be reasonable for these data. The seasonal lag is $s = 12$ (the data are **monthly**).
- In Figure 10.9, we display the sample ACF and PACF for the boardings data. Note that the margin of error bounds in the plot are for a white noise process.
  - The sample ACF shows a pronounced sample autocorrelation at lag $k = 12$ and a decay afterward at seasonal lags $k = 24$ and $k = 36$.
  - The sample PACF shows a pronounced sample partial autocorrelation at lag $k = 12$ and none at higher seasonal lags.
  - These two observations together suggest a **seasonal** AR$(1)_{12}$ component.

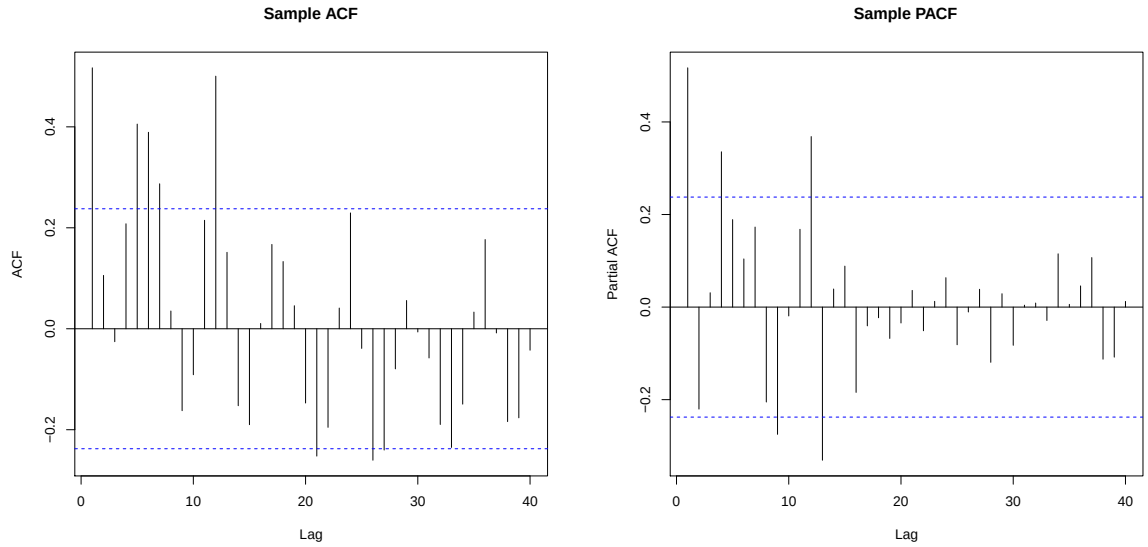


Figure 10.9: Denver public transit data. Left: Sample ACF. Right: Sample PACF.

  - Around the seasonal lags $k = 12$, $k = 24$, and $k = 36$ (in the ACF), there are noticeable autocorrelations 3 time units in both directions. This suggests a **nonseasonal** MA(3) component.

- We therefore specify an ARMA$(0, 3) \times$ ARMA$(1, 0)_{12}$ model for these data. Of course, this model at this point is **tentative** and is subject to further investigation and scrutiny.

MODEL FITTING: We use R to fit an ARMA$(0, 3) \times$ ARMA$(1, 0)_{12}$ model using maximum likelihood. Here is the output:

```
> boardings.arma03.arma10 = arima(boardings,order=c(0,0,3),method='ML',
    seasonal=list(order=c(1,0,0),period=12))
> boardings.arma03.arma10
Coefficients:
         ma1     ma2     ma3    sar1  intercept
      0.7288  0.6115  0.2951  0.8777    12.5455
s.e.  0.1186  0.1172  0.1118  0.0507     0.0354
sigma^2 estimated as 0.0006542:  log likelihood = 143.54,  aic = -277.09
```

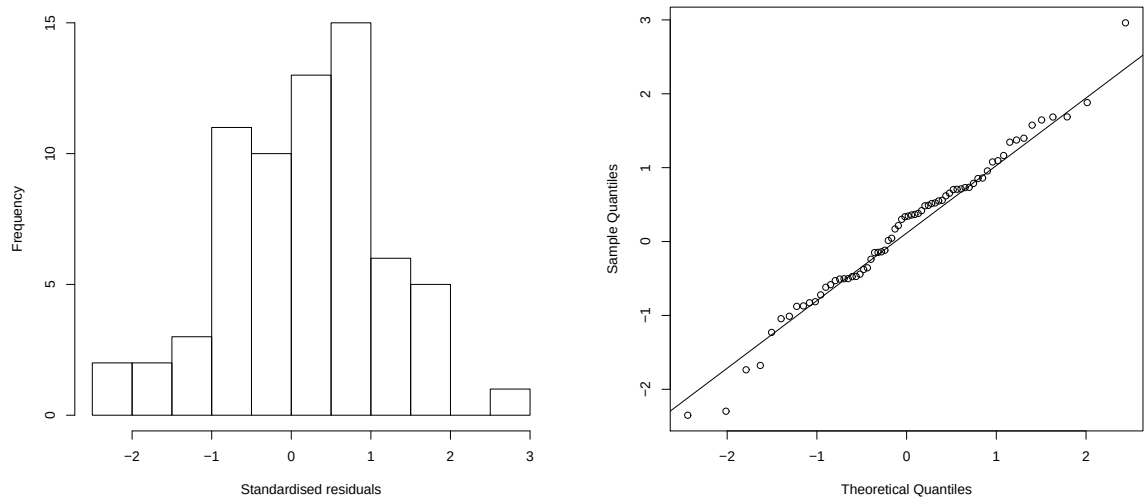


Figure 10.10: Denver public transit data. Standardized residuals from ARMA$(0,3) \times$ ARMA$(1,0)_{12}$ model fit.

Note that each of the parameter estimates is statistically different from zero. The fitted ARMA$(0,3) \times$ ARMA$(1,0)_{12}$ model (on the log scale) is

$$(1 - 0.8777B^{12})(Y_t - 12.5455) = (1 + 0.7288B + 0.6115B^2 + 0.2951B^3)e_t,$$

or equivalently,

$$Y_t = 1.5343 + 0.8777Y_{t-12} + e_t + 0.7288e_{t-1} + 0.6115e_{t-2} + 0.2951e_{t-3}.$$

The white noise variance estimate is $\widehat{\sigma}_e^2 \approx 0.0006542$.

MODEL DIAGNOSTICS: The histogram and qq plot of the standardized residuals in Figure 10.10 generally supports the normality assumption. In addition, when further examining the standardized residuals,

- the Shapiro-Wilk test does not reject normality (p-value = 0.6187)
- the runs test does not reject independence (p-value = 0.385).

Finally, the `tsdiag` output in Figure 10.11 shows no notable problems with the ARMA$(0,3)\times$ ARMA$(1,0)_{12}$ model.

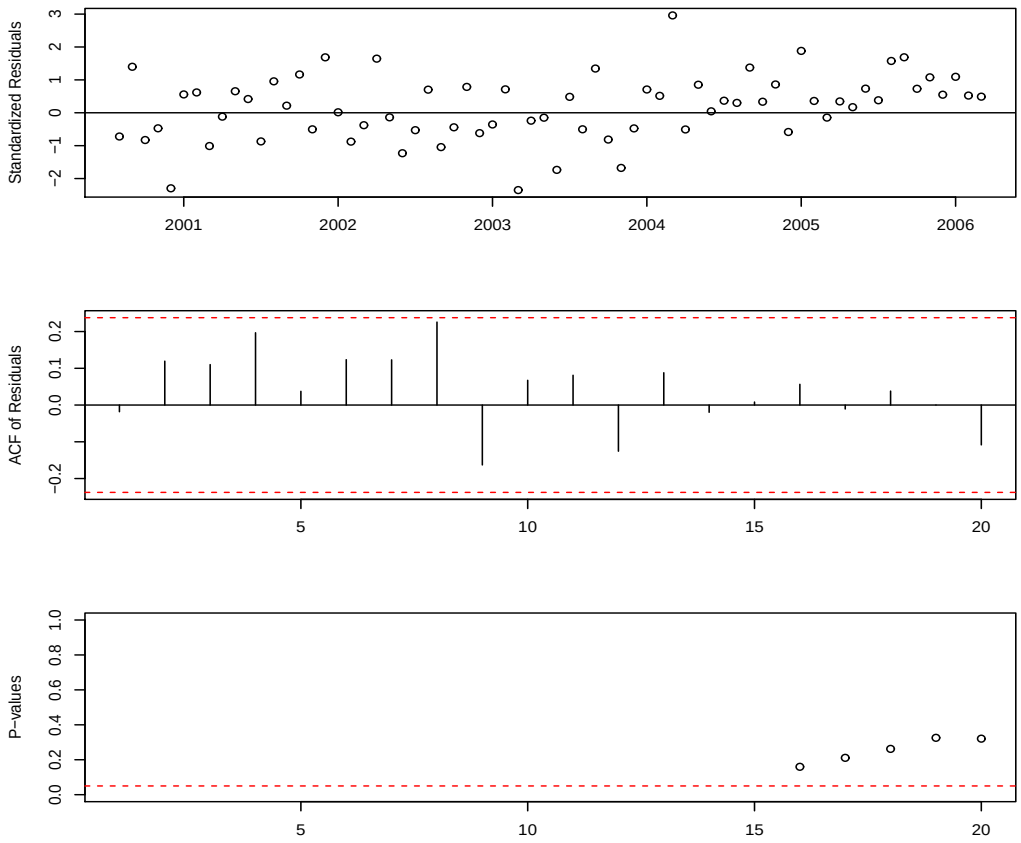


Figure 10.11: Denver public transit data. ARMA$(0,3) \times$ ARMA$(1,0)_{12}$ `tsdiag` output.

OVERFITTING: For an ARMA$(0,3) \times$ARMA$(1,0)_{12}$ model, there are 4 overfitted models. Here are the models and the results from overfitting the boarding data:

$$\begin{aligned}
\text{ARMA}(1,3) \times \text{ARMA}(1,0)_{12} &\implies \widehat{\phi} \textbf{ significant} \\
\text{ARMA}(0,4) \times \text{ARMA}(1,0)_{12} &\implies \widehat{\theta}_4 \text{ not significant} \\
\text{ARMA}(0,3) \times \text{ARMA}(2,0)_{12} &\implies \widehat{\Phi}_2 \text{ not significant} \\
\text{ARMA}(0,3) \times \text{ARMA}(1,1)_{12} &\implies \widehat{\Theta} \text{ not significant}
\end{aligned}$$

The ARMA$(1,3) \times$ ARMA$(1,0)_{12}$ fit declares a nonseasonal AR component at lag $k = 1$ to be significant, but the MA estimates at lags $k = 1, 2$, and 3 (which were all highly significant in the original fit) become insignificant in this overfitted model! Therefore, the ARMA$(1,3) \times$ ARMA$(1,0)_{12}$ overfitted model is not considered further.

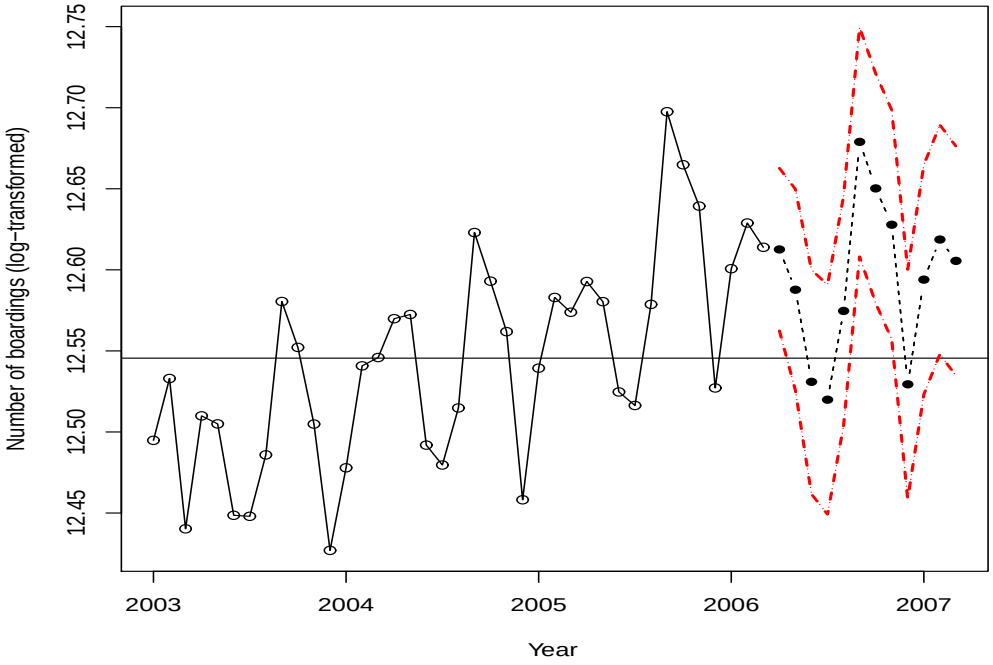


Figure 10.12: Denver public transit data. The full data set is from 8/2000-3/2006. This figure starts the series at 1/2003. ARMA$(0,3)\times$ARMA$(1,0)_{12}$ estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 12$. These lead times correspond to years 4/2006-3/2007.

FORECASTING: The estimated forecasts and standard errors (**on the log scale**) are given for lead times $l = 1, 2, ..., 12$ in the `predict` output below:

```
> boardings.arma03.arma10.predict <- predict(boardings.arma03.arma10.fit,n.ahead=12)
> round(boardings.arma03.arma10.predict$pred,3)
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2006                       12.613 12.588 12.531 12.520 12.575 12.679 12.650 12.628 12.529
2007 12.594 12.619 12.606

> round(boardings.arma03.arma10.predict$se,3)
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2006                 0.026 0.032 0.035 0.036 0.036 0.036 0.036 0.036 0.036
2007 0.036 0.036 0.036
```

- In Figure 10.12, we display the Denver boardings data. The full data set is from 8/00 to 3/06 (one observation per month). However, to emphasize the MMSE forecasts in the plot, we start the series at month 1/03.

- With $l = 1, 2, ..., 12$, the estimated MMSE forecasts in the `predict` output and in Figure 10.12 start at 4/06 and end in 3/07. It is important to remember that these forecasts are on the **log scale**. MMSE forecasts on the original scale and 95 percent prediction intervals are given below.

```
> # MMSE forecasts back-transformed (to original scale)
> denver.boardings.predict <- round(exp(boardings.arma03.arma10.predict$pred
    + (1/2)*(boardings.arma03.arma10.predict$se)^2),3)
> denver.boardings.predict
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep
2006                            300411.9 293085.1 276937.5 273911.8 289321.6 321123.3
2007 294962.7 302347.3 298397.8
          Oct      Nov      Dec
2006 312037.2 305125.6 276521.7
2007
> # Compute prediction intervals (on original scale)
> data.frame(Month=year.temp,lower.pi=exp(lower.pi),upper.pi=exp(upper.pi))
      Month lower.pi upper.pi
1  2006.250 285630.0 315752.2
2  2006.333 275318.8 311685.4
3  2006.416 258262.2 296593.3
4  2006.500 255034.1 293803.7
5  2006.583 269381.8 310332.5
6  2006.666 298991.8 344443.8
7  2006.750 290531.9 334697.8
8  2006.833 284096.7 327284.3
9  2006.916 257464.1 296603.1
10 2007.000 274634.1 316383.3
11 2007.083 281509.8 324304.2
12 2007.166 277832.5 320067.9
```

SUMMARY: The multiplicative seasonal (stationary) ARMA$(p, q) \times$ ARMA$(P, Q)_s$ family of models

$$\phi(B)\Phi_P(B^s)Y_t = \theta(B)\Theta_Q(B^s)e_t.$$

is a flexible class of time series models for **stationary** seasonal processes. The next step is to extend this class of models to handle two types of nonstationarity:

- **Nonseasonal nonstationary** over time (e.g., increasing linear trends, etc.)
- **Seasonal nonstationarity**, that is, additional changes in the seasonal mean level, even after possibly adjusting for nonseasonal stationarity over time.

10.4 Nonstationary seasonal ARIMA (SARIMA) models

REVIEW: For a stochastic process $\{Y_t\}$, the first differences are

$$\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t.$$

This definition can be generalized to any number of differences; in general, the dth differences are given by

$$\nabla^d Y_t = (1 - B)^d Y_t.$$

We know that taking $d = 1$ or (usually at most) $d = 2$ can coerce a (nonseasonal) nonstationary process into stationarity.

EXAMPLE: Suppose that we have a stochastic process defined by

$$Y_t = S_t + e_t,$$

where $\{e_t\}$ is zero mean white noise and where

$$S_t = S_{t-12} + u_t,$$

where $\{u_t\}$ is zero mean white noise that is uncorrelated with $\{e_t\}$. That is, $\{S_t\}$ is a zero mean **random walk** with period $s = 12$. For this process, taking **nonseasonal**

differences (as we have done up until now) will not have an effect on the seasonal nonstationarity. For example, with $d = 1$, we have

$$\begin{aligned}\nabla Y_t &= \nabla S_t + \nabla e_t \\ &= \nabla S_{t-12} + \nabla u_t + \nabla e_t \\ &= S_{t-12} - S_{t-13} + u_t - u_{t-1} + e_t - e_{t-1}.\end{aligned}$$

The first difference process $\{\nabla Y_t\}$ is still nonstationary because $\{S_t\}$ is a random walk across seasons; i.e, across time points $t = 12k$.

- That is, taking (nonseasonal) differences has only produced a more complicated model, one which is still nonstationary across seasons.
- We therefore need to define a new differencing operator that can remove nonstationarity across seasonal lags.

TERMINOLOGY: The **seasonal difference operator** ∇_s is defined by

$$\nabla_s Y_t = Y_t - Y_{t-s} = (1 - B^s)Y_t,$$

for a seasonal period s. For example, with $s = 12$ and monthly data, the **first seasonal differences** are

$$\nabla_{12} Y_t = Y_t - Y_{t-12} = (1 - B^{12})Y_t,$$

that is, the first differences of the January observations, the first differences of the February observations, and so on.

EXAMPLE: For the stochastic process defined earlier, that is,

$$Y_t = S_t + e_t,$$

where $S_t = S_{t-12} + u_t$, taking first seasonal differences yields

$$\begin{aligned}\nabla_{12} Y_t &= \nabla_{12} S_t + \nabla_{12} e_t \\ &= S_t - S_{t-12} + e_t - e_{t-12} = u_t + e_t - e_{t-12}.\end{aligned}$$

It can be shown that this process has the same ACF as a stationary seasonal $\text{MA}(1)_{12}$. That is, taking first seasonal differences has coerced the $\{Y_t\}$ process into stationarity.

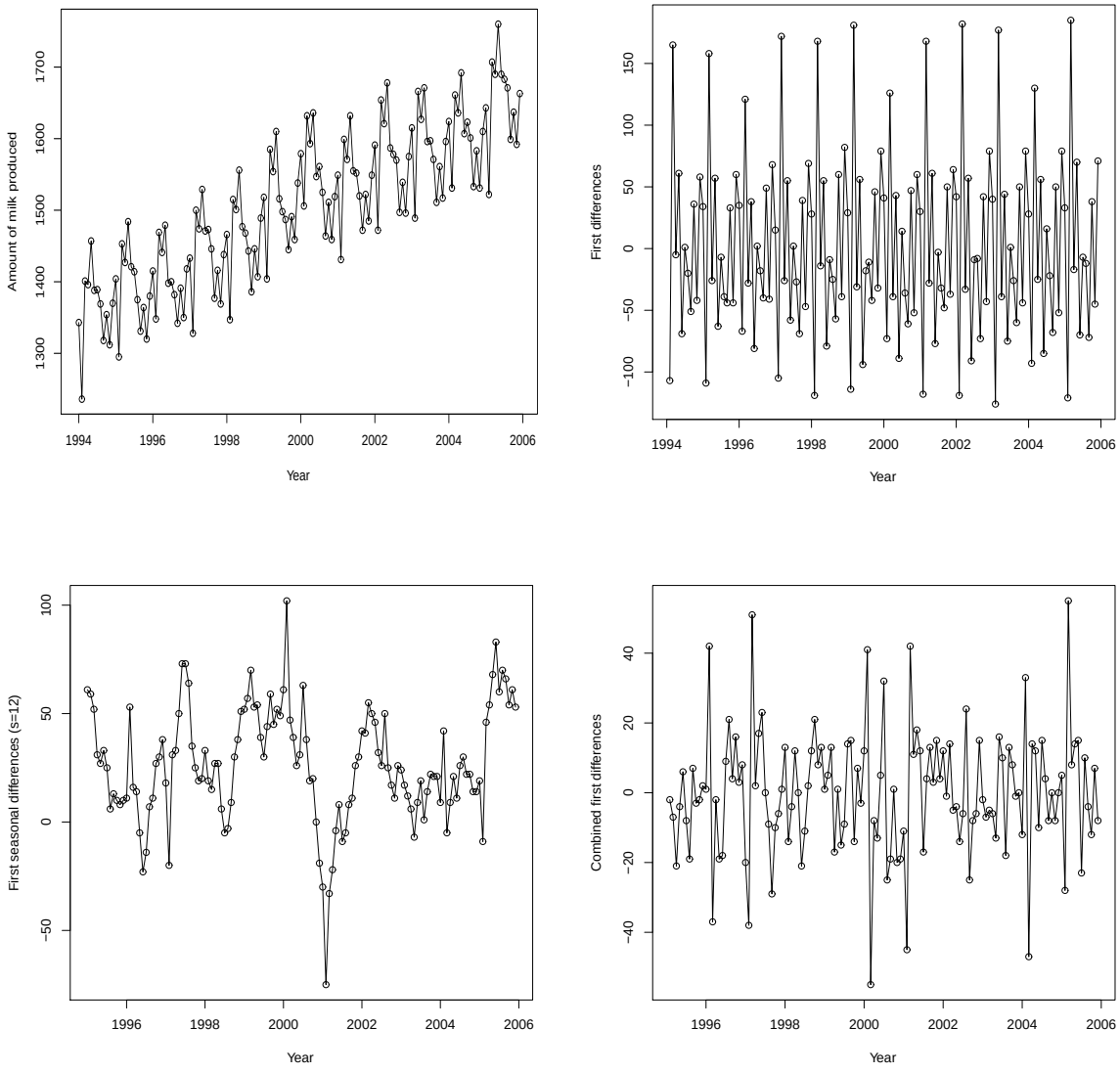


Figure 10.13: United States milk production data. Upper left: Original series $\{Y_t\}$. Upper right: First (nonseasonal) differences $\nabla Y_t = Y_t - Y_{t-1}$. Lower left: First (seasonal) differences $\nabla_{12} Y_t = Y_t - Y_{t-12}$. Lower right: Combined first (seasonal and nonseasonal) differences $\nabla \nabla_{12} Y_t$.

Example 10.5. Consider the monthly U.S. milk production data from Example 10.1. Figure 10.13 (last page) displays the time series plot of the data (upper left), the first difference process ∇Y_t (upper right), the first seasonal difference process $\nabla_{12} Y_t$ (lower left), and the combined difference process $\nabla\nabla_{12} Y_t$ (lower right). The combined difference process $\nabla\nabla_{12} Y_t$ is given by

$$\begin{aligned}\nabla\nabla_{12} Y_t &= (1-B)(1-B^{12})Y_t \\ &= (1-B-B^{12}+B^{13})Y_t \\ &= Y_t - Y_{t-1} - Y_{t-12} + Y_{t-13}.\end{aligned}$$

- The milk series (Figure 10.13; upper left) displays two trends: nonstationarity over time and a within-year seasonal pattern. A Box-Cox analysis (results not shown) suggests that no transformation is necessary for variance stabilization purposes.
- Taking first (nonseasonal) differences; i.e., computing ∇Y_t, (Figure 10.13; upper right) has removed the upward linear trend (as expected), but the process $\{\nabla Y_t\}$ still displays notable seasonality.
- Taking first (seasonal) differences; i.e., computing $\nabla_{12} Y_t$, (Figure 10.13; lower left) has seemingly removed the seasonality (as expected), but the process $\{\nabla_{12} Y_t\}$ displays still strong momentum over time.
  - The sample ACF of $\{\nabla_{12} Y_t\}$ (not shown) displays a slow decay, a sign of nonstationarity over time.
- The combined first differences $\nabla\nabla_{12} Y_t$ (Figure 10.13; lower right) look to resemble a stationary process (at least in the mean level).

REMARK: From this example, it should be clear that we can now extend the multiplicative seasonal (stationary) ARMA$(p,q) \times$ ARMA$(P,Q)_s$ model

$$\phi(B)\Phi_P(B^s)Y_t = \theta(B)\Theta_Q(B^s)e_t$$

to incorporate the two types of nonstationarity: nonseasonal and seasonal. This leads to the definition of our largest class of ARIMA models.

TERMINOLOGY: Suppose that $\{e_t\}$ is zero mean white noise with $\text{var}(e_t) = \sigma_e^2$. The **multiplicative seasonal autoregressive integrated moving average (SARIMA) model** with **seasonal period** s, denoted by $\text{ARIMA}(p,d,q) \times \text{ARIMA}(P,D,Q)_s$, is

$$\phi(B)\Phi_P(B^s)\nabla^d\nabla_s^D Y_t = \theta(B)\Theta_Q(B^s)e_t,$$

where the nonseasonal AR and MA characteristic operators are

$$\begin{aligned}\phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q),\end{aligned}$$

the seasonal AR and MA characteristic operators are

$$\begin{aligned}\Phi_P(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \\ \Theta_Q(B^s) &= 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs},\end{aligned}$$

and

$$\nabla^d\nabla_s^D Y_t = (1-B)^d(1-B^s)^D Y_t.$$

In this model,

- d denotes the number of **nonseasonal differences**. Usually $d = 1$ or (at most) $d = 2$ will provide nonseasonal stationarity (as we have seen before).
- D denotes the number of **seasonal differences**. Usually $D = 1$ will achieve seasonal stationarity.
- For many nonstationary seasonal time series data sets (at least for the ones I have seen), the most common choice for (d, D) is $(1, 1)$.

NOTE: We have the following relationship:

$$Y_t \sim \text{ARIMA}(p,d,q)\times\text{ARIMA}(P,D,Q)_s \iff \nabla^d\nabla_s^D Y_t \sim \text{ARMA}(p,q)\times\text{ARMA}(P,Q)_s.$$

The SARIMA class is very flexible. Many times series can be adequately fit by these models, usually with a small number of parameters, often less than five.

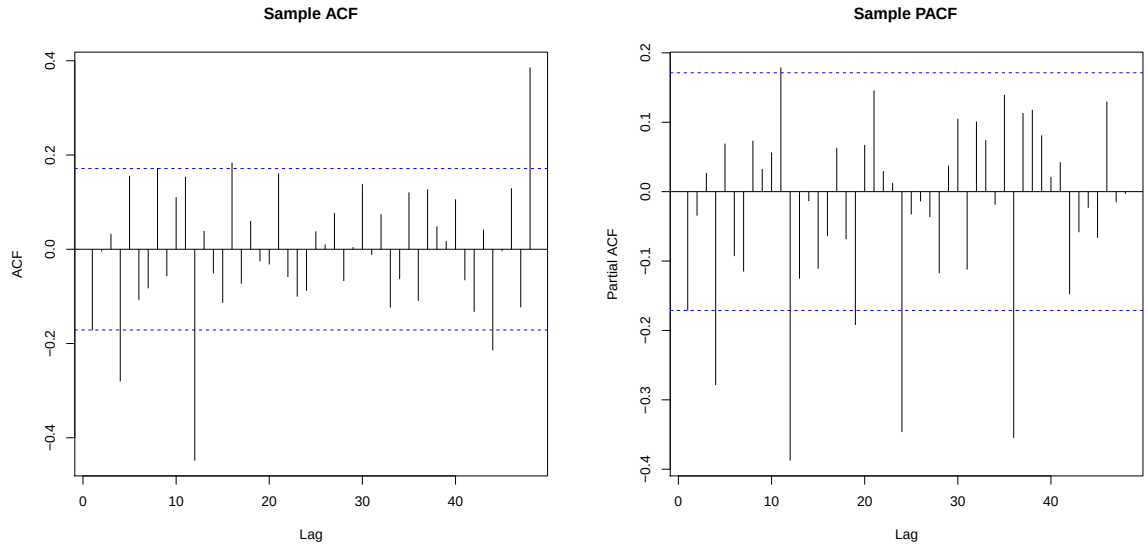


Figure 10.14: United States milk production data. Left: Sample ACF for $\{\nabla\nabla_{12}Y_t\}$. Right: Sample PACF for $\{\nabla\nabla_{12}Y_t\}$.

Example 10.5 (continued). For the milk production data in Example 10.1, we have seen that the combined difference process $\{\nabla\nabla_{12}Y_t\}$ looks to be relatively stationary. In Figure 10.14, we display the sample ACF (left) and sample PACF (right) of the $\{\nabla\nabla_{12}Y_t\}$ process. Examining these two plots will help us identify which ARMA$(p,q) \times$ ARMA$(P,Q)_{12}$ model is appropriate for $\{\nabla\nabla_{12}Y_t\}$.

- The sample ACF for $\{\nabla\nabla_{12}Y_t\}$ has a pronounced spike at seasonal lag $k = 12$ and one at $k = 48$ (but none at $k = 24$ and $k = 36$).
- The sample PACF for $\{\nabla\nabla_{12}Y_t\}$ displays pronounced spikes at seasonal lags $k = 12, 24$ and 36.
- The last two observations are consistent with the following choices:
  - $(P,Q) = (0,1)$ if one is willing to ignore the ACF at $k = 48$. Also, if $(P,Q) = (0,1)$, we would expect the the PACF to **decay** at lags $k = 12, 24$ and 36. There is actually not that much of a decay.
  - $(P,Q) = (3,0)$, if one is willing to place strong emphasis on the sample PACF.

- There does not appear to be "anything happening" around seasonal lags in the ACF, and the ACF at $k = 1$ is borderline. We therefore take $p = 0$ and $q = 0$.
- Therefore, there are two models which emerge as strong possibilities:
  - $(P, Q) = (0, 1)$: $\text{MA}(1)_{12}$ model for $\{\nabla\nabla_{12}Y_t\}$
  - $(P, Q) = (3, 0)$: $\text{AR}(3)_{12}$ model for $\{\nabla\nabla_{12}Y_t\}$.
- I have carefully examined both models. The $\text{AR}(3)_{12}$ model provides a much better fit to the $\{\nabla\nabla_{12}Y_t\}$ process than the $\text{MA}(1)_{12}$ model.
  - The $\text{AR}(3)_{12}$ model for $\{\nabla\nabla_{12}Y_t\}$ provides a smaller AIC, a smaller estimate of the white noise variance, and superior residual diagnostics; e.g., the Ljung-Box test strongly discounts the $\text{MA}(1)_{12}$ model for $\{\nabla\nabla_{12}Y_t\}$ at all lags.
- For illustrative purposes, we therefore tentatively adopt an $\text{ARIMA}(0, 1, 0) \times \text{ARIMA}(3, 1, 0)_{12}$ model for the milk production data.

MODEL FITTING: We use R to fit this $\text{ARIMA}(0, 1, 0) \times \text{ARIMA}(3, 1, 0)_{12}$ model using maximum likelihood. Here is the output:

```
> milk.arima010.arima310 =
    arima(milk,order=c(0,1,0),method='ML',seasonal=list(order=c(3,1,0),period=12))
> milk.arima010.arima310
Coefficients:
         sar1     sar2     sar3
      -0.9133  -0.8146  -0.6002
s.e.   0.0696   0.0776   0.0688
sigma^2 estimated as 121.4:  log likelihood = -512.03,  aic = 1030.05
```

The fitted model is

$$(1 + 0.9133B^{12} + 0.8146B^{24} + 0.6002B^{36}) \underbrace{(1 - B)(1 - B^{12})Y_t}_{= \nabla\nabla_{12}Y_t} = e_t.$$

The white noise variance estimate is $\widehat{\sigma}_e^2 \approx 121.4$. Note that all parameter estimates ($\widehat{\Theta}_1$, $\widehat{\Theta}_2$, and $\widehat{\Theta}_3$) are statistically different from zero (by a very large amount).

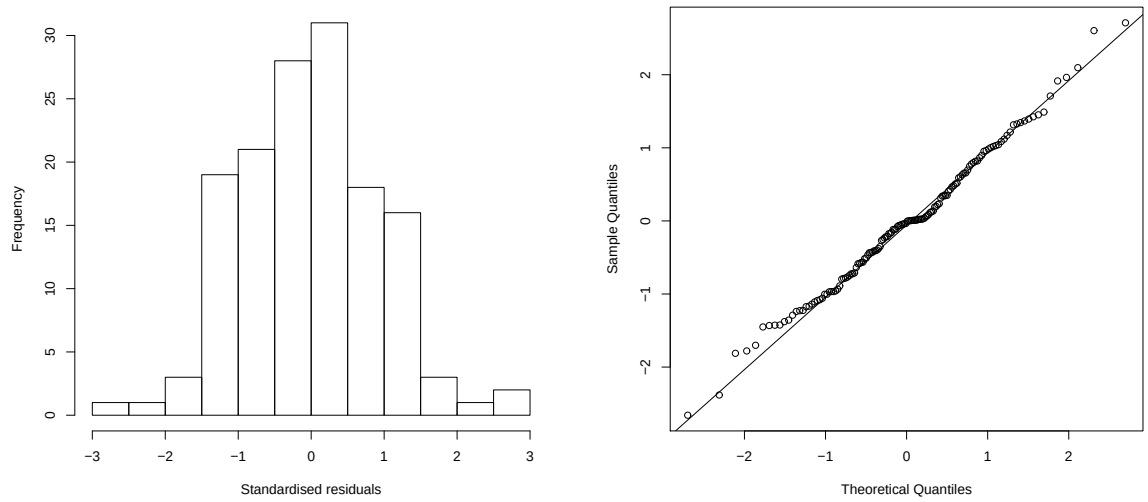


Figure 10.15: United States milk production data. Standardized residuals from ARIMA$(0,1,0) \times$ ARIMA$(3,1,0)_{12}$ model fit.

MODEL DIAGNOSTICS: The histogram and qq plot of the standardized residuals in Figure 10.15 generally supports the normality assumption. In addition, when further examining the standardized residuals from the model fit,

- the Shapiro-Wilk test does not reject normality (p-value = 0.6619)
- the runs test does not reject independence (p-value = 0.112).

Finally, the `tsdiag` output in Figure 10.16 supports the ARIMA$(0,1,0)\times$ARIMA$(3,1,0)_{12}$ model choice.

OVERFITTING: For an ARIMA$(0,1,0)\times$ARIMA$(3,1,0)_{12}$ model, there are 4 overfitted models. Here are the models and the results from overfitting:

$$\begin{aligned}
\text{ARIMA}(1,1,0) \times \text{ARIMA}(3,1,0)_{12} &\implies \widehat{\phi} \text{ not significant} \\
\text{ARIMA}(0,1,1) \times \text{ARIMA}(3,1,0)_{12} &\implies \widehat{\theta} \text{ not significant} \\
\text{ARIMA}(0,1,0) \times \text{ARIMA}(4,1,0)_{12} &\implies \widehat{\Phi}_4 \text{ not significant} \\
\text{ARIMA}(0,1,0) \times \text{ARIMA}(3,1,1)_{12} &\implies \widehat{\Theta} \text{ not significant.}
\end{aligned}$$

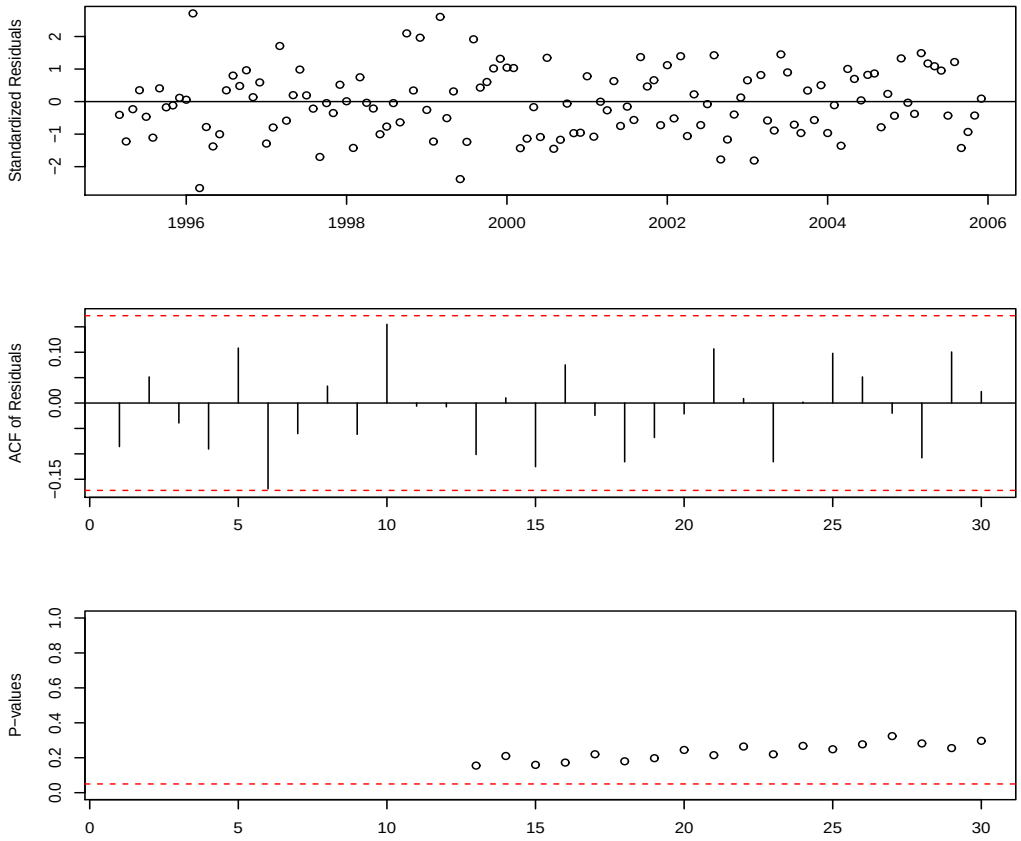


Figure 10.16: United States milk production data. ARIMA$(0,1,0) \times$ ARIMA$(3,1,0)_{12}$ `tsdiag` output.

CONCLUSION: The ARIMA$(0,1,0) \times$ ARIMA$(3,1,0)_{12}$ model does a good job at describing the U.S. milk production data. With this model, we move forward with forecasting future observations.

FORECASTING: We use R to compute forecasts and prediction limits for the lead times $l = 1, 2, ..., 24$ (two years ahead) based on the ARIMA$(0,1,0) \times$ ARIMA$(3,1,0)_{12}$ model fit. Here are the estimated MMSE forecasts and 95 percent prediction limits:

```
# MMSE forecasts
> milk.arima010.arima310.predict <- predict(milk.arima010.arima310.fit,n.ahead=24)
```

```
> round(milk.arima010.arima310.predict$pred,3)
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep
2006 1702.409 1584.302 1760.356 1728.246 1783.487 1698.330 1694.116 1680.528 1610.895
2007 1725.769 1608.022 1775.653 1742.424 1792.538 1715.007 1717.981 1695.297 1631.562
          Oct      Nov      Dec
2006 1655.054 1610.777 1689.084
2007 1679.871 1634.033 1712.183

> round(milk.arima010.arima310.predict$se,3)
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2006 11.018 15.581 19.083 22.035 24.636 26.988 29.150 31.162 33.053 34.841 36.541 38.166
2007 40.000 41.753 43.436 45.056 46.620 48.132 49.599 51.024 52.410 53.760 55.077 56.363

# Compute prediction intervals
lower.pi<-
milk.arima010.arima310.predict$pred-qnorm(0.975,0,1)*milk.arima010.arima310.predict$se
upper.pi<-
milk.arima010.arima310.predict$pred+qnorm(0.975,0,1)*milk.arima010.arima310.predict$se
## For brevity (in the notes), I display estimated MMSE forecasts only 12 months ahead.
      Month lower.pi upper.pi
1  2006.000 1680.815 1724.003
2  2006.083 1553.763 1614.840
3  2006.166 1722.954 1797.758
4  2006.250 1685.058 1771.434
5  2006.333 1735.201 1831.773
6  2006.416 1645.436 1751.225
7  2006.500 1636.983 1751.249
8  2006.583 1619.450 1741.605
9  2006.666 1546.113 1675.678
10 2006.750 1586.767 1723.340
11 2006.833 1539.158 1682.397
12 2006.916 1614.280 1763.888
```

- In Figure 10.17, we display the U.S. milk production data. The full data set is from 1/94 to 12/05 (one observation per month). However, to emphasize the MMSE forecasts in the plot, we start the series at month 1/04.

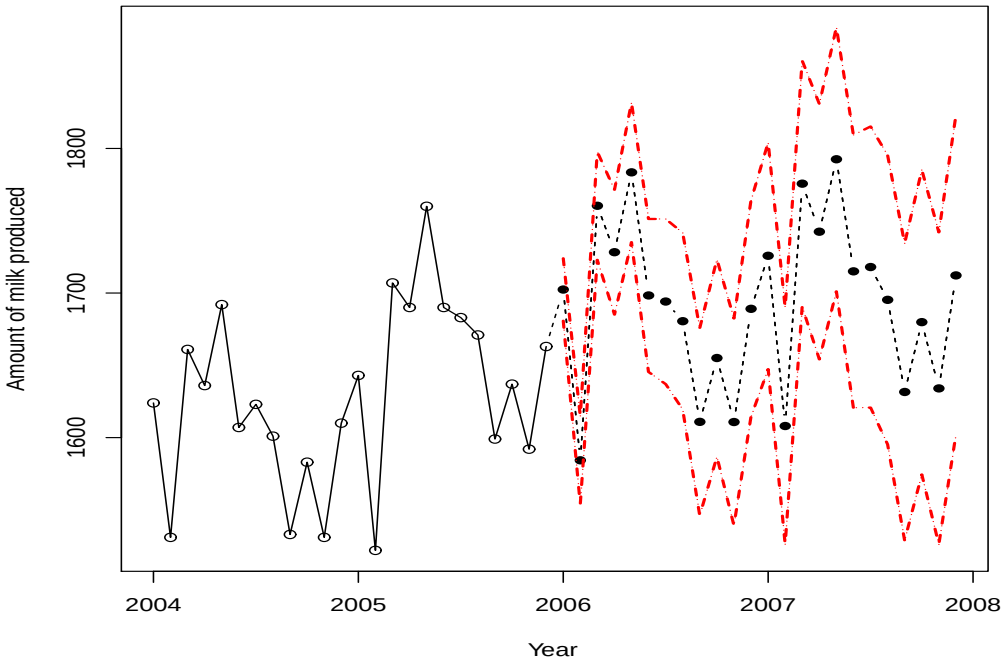


Figure 10.17: U.S. milk production data. The full data set is from 1/1994-12/2005. This figure starts the series at 1/2004. ARIMA$(0,1,0) \times$ ARIMA$(3,1,0)_{12}$ estimated MMSE forecasts and 95 percent prediction limits are given for lead times $l = 1, 2, ..., 24$. These lead times correspond to years 1/2006-12/2007.

- With $l = 1, 2, ..., 24$, the estimated MMSE forecasts in the `predict` output and in Figure 10.17 start at 1/06 and end in 12/07 (24 months).
- Numerical values of the 95 percent prediction intervals are given for 1/06-12/06 in the prediction interval output. Note how the interval lengths increase as l does. This is a byproduct of **nonstationarity**. In Figure 10.17, the impact of nonstationarity is also easily seen as l increases (prediction limits become wider).

NOTE: Although we did not state so explicitly, determining MMSE forecasts and prediction limits for seasonal models is exactly analogous to the nonseasonal cases we studied in Chapter 9. Formulae for seasonal MMSE forecasts are given in Section 10.5 (CC) for special cases.

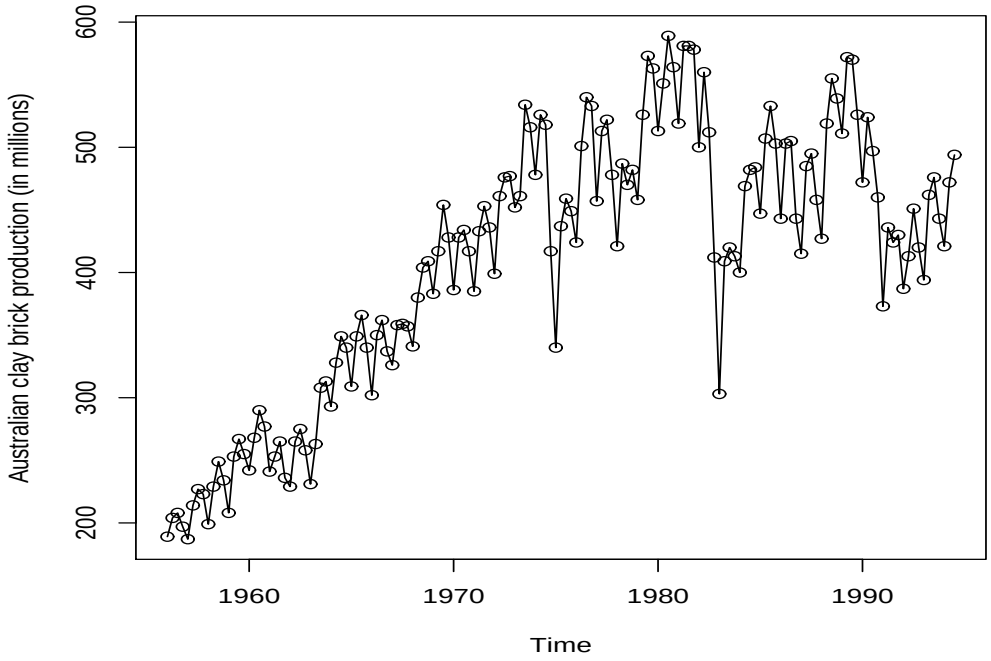


Figure 10.18: Australian clay brick production data. Number of bricks (in millions) produced from 1956-1994.

Example 10.6. In this example, we revisit the Australian brick production data in Example 1.14 (pp 15, notes). The data in Figure 10.18 represent the number of bricks produced in Australia (in millions) during 1956-1994. The data are quarterly, so the underlying seasonal lag of interest is $s = 4$.

INITIAL ANALYSIS: The first thing we should do is a Box-Cox analysis to see if a variance-stabilizing transformation is needed (there is evidence of heteroscedasticity from examining the original series in Figure 10.18).

- Using the `BoxCox.ar` function in R (output not shown) suggests that the Box-Cox transformation parameter $\lambda \approx 0.5$.
- This suggests that a square-root transformation is warranted.
- We now examine the transformed data and the relevant differenced series.

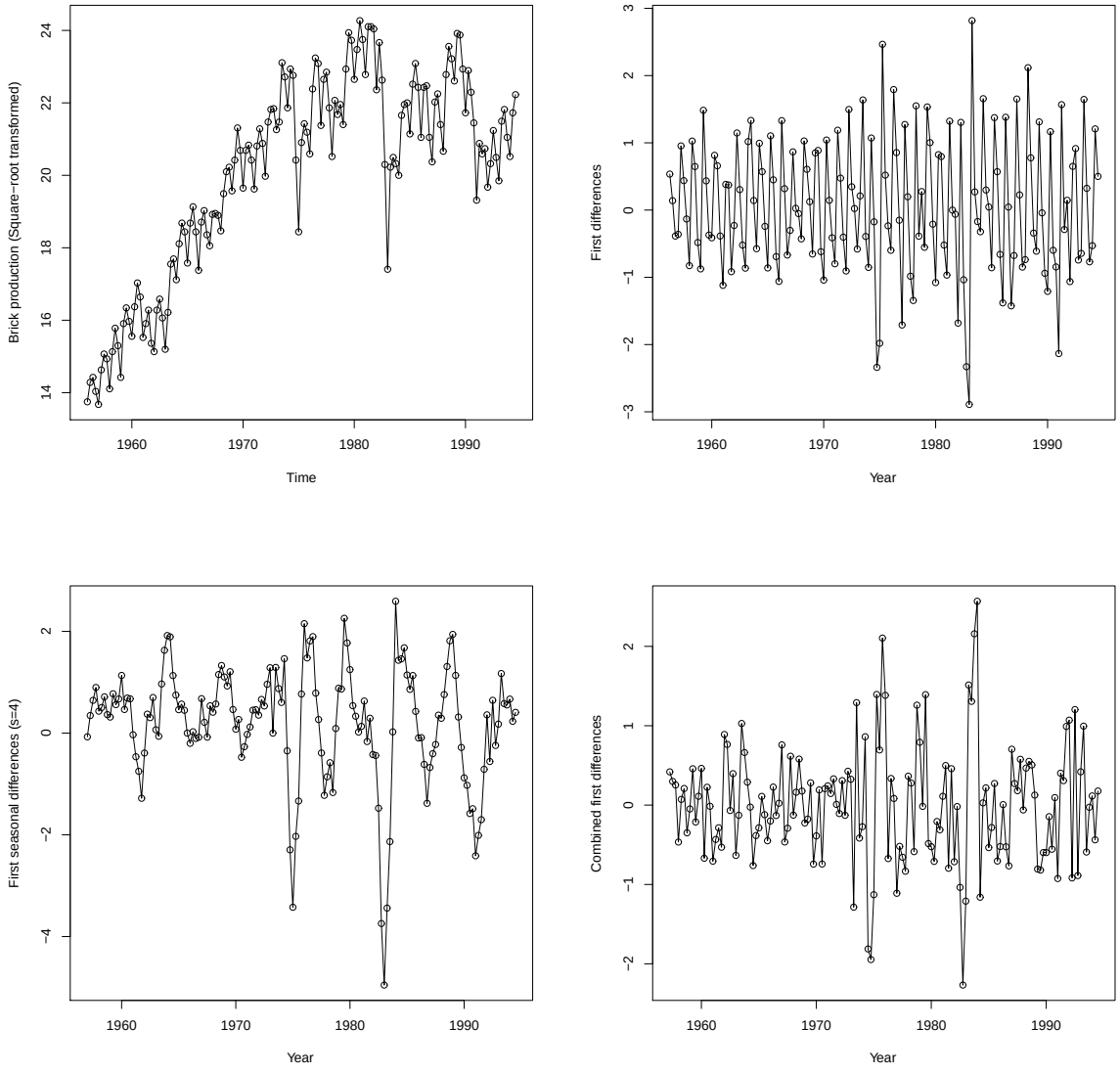


Figure 10.19: Australian clay brick production data (square-root transformed). Upper left: Original series $\{Y_t\}$. Upper right: First (nonseasonal) differences $\nabla Y_t = Y_t - Y_{t-1}$. Lower left: First (seasonal) differences $\nabla_4 Y_t = Y_t - Y_{t-4}$. Lower right: Combined first (seasonal and nonseasonal) differences $\nabla\nabla_4 Y_t$.

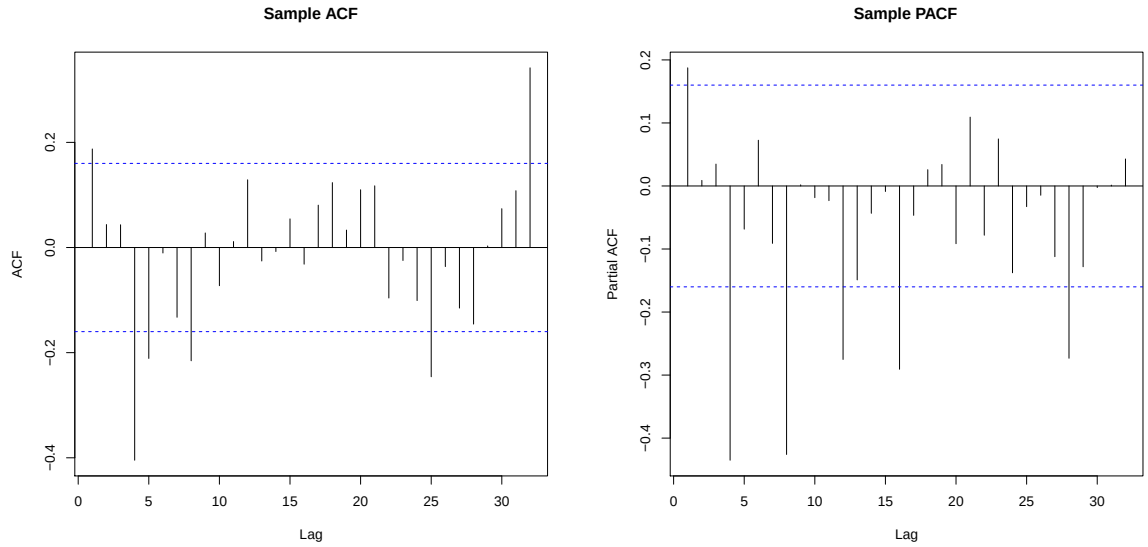


Figure 10.20: Australian clay brick production data (square-root transformed). Left: Sample ACF for $\{\nabla\nabla_4 Y_t\}$. Right: Sample PACF for $\{\nabla\nabla_4 Y_t\}$.

NOTE: The combined difference process $\nabla\nabla_4 Y_t$ in Figure 10.19 looks stationary in the mean level. The sample ACF/PACF for the $\nabla\nabla_4 Y_t$ series is given in Figure 10.20. Recall that our analysis is now on the square-root transformed scale.

ANALYSIS: Examining the sample ACF/PACF for the $\nabla\nabla_4 Y_t$ data does not lead us to one single model as a "clear favorite." In fact, there are ambiguities that emerge; e.g., a spike in the ACF at lag $k = 25$ (this is not a seasonal lag), a spike in the PACF at the seventh seasonal lag $k = 28$, etc.

- The PACF does display spikes at the first 4 seasonal lags $k = 4$, $k = 8$, $k = 12$, and $k = 16$.
- The ACF does not display consistent "action" around these seasonal lags in either direction.
- These two observations lead us to tentatively consider an AR$(4)_4$ model for the combined difference process $\{\nabla\nabla_4 Y_t\}$; i.e., an ARIMA$(0, 1, 0) \times$ ARIMA$(4, 1, 0)_4$ for the square-root transformed series.

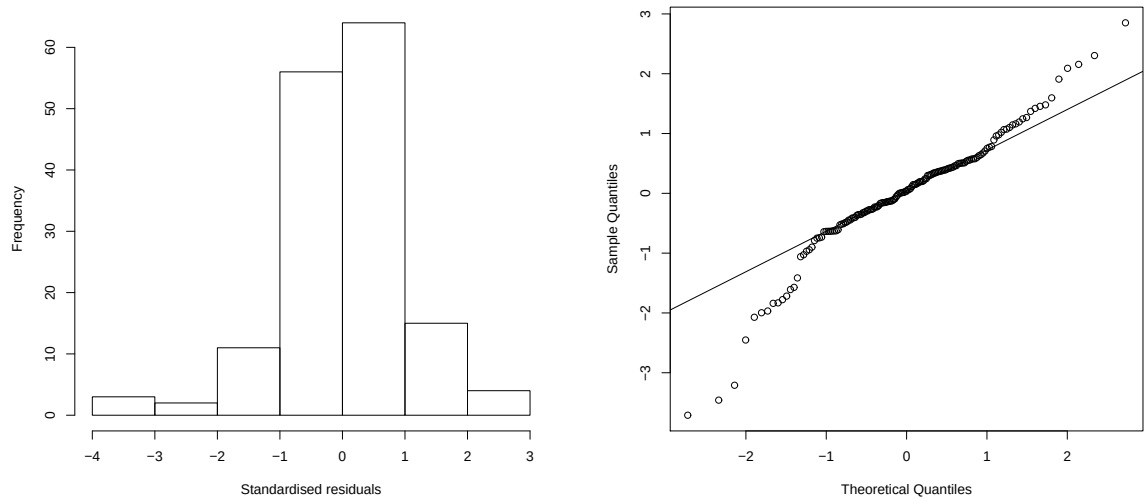


Figure 10.21: Australian clay brick production data (square-root transformed). Standardized residuals from ARIMA$(0,1,0) \times$ ARIMA$(4,1,0)_4$ model fit.

MODEL FITTING: We use R to fit this ARIMA$(0,1,0) \times$ ARIMA$(4,1,0)_4$ model using maximum likelihood. Here is the output:

```
> sqrt.brick.arima010.arima410 =
  arima(sqrt.brick,order=c(0,1,0),method='ML',seasonal=list(order=c(4,1,0),period=4))
> sqrt.brick.arima010.arima410
Coefficients:
         sar1     sar2     sar3     sar4
      -0.8249  -0.8390  -0.5330  -0.3290
s.e.   0.0780   0.0935   0.0936   0.0772
sigma^2 estimated as 0.2889:  log likelihood = -122.47,  aic = 252.94
```

The fitted model is

$$(1 + 0.8249B^4 + 0.8390B^8 + 0.5330B^{12} + 0.3290B^{16}) \underbrace{(1-B)(1-B^4)Y_t}_{= \nabla\nabla_4 Y_t} = e_t.$$

The white noise variance estimate is $\widehat{\sigma}_e^2 \approx 0.2889$. Note that all parameter estimates ($\widehat{\Theta}_1$, $\widehat{\Theta}_2$, $\widehat{\Theta}_3$, and $\widehat{\Theta}_4$) are statistically different from zero (by a very large amount).

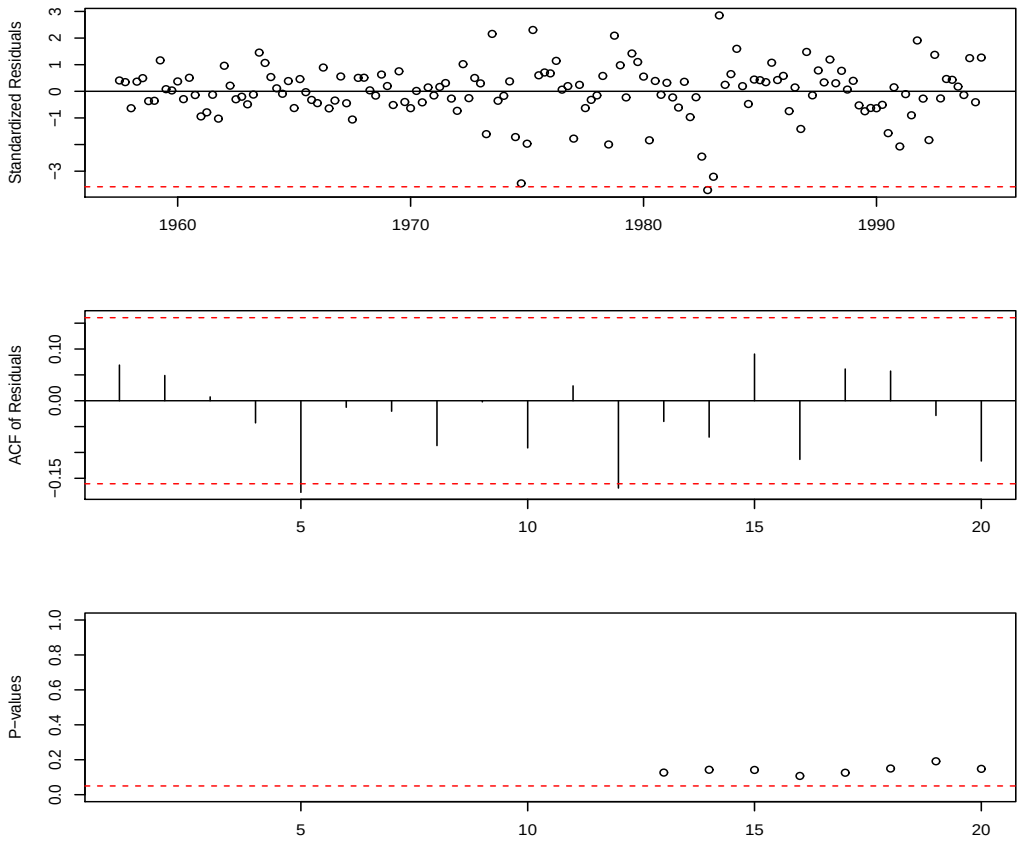


Figure 10.22: Australian clay brick production data (square-root transformed). ARIMA$(0,1,0) \times$ ARIMA$(4,1,0)_4$ `tsdiag` output.

DIAGNOSTICS: The `tsdiag` output in Figure 10.22 does not strongly refute the ARIMA$(0,1,0) \times$ ARIMA$(4,1,0)_4$ model choice, and overfitting (results not shown) does not lead us to consider a higher order model. However, the qq plot of the standardized residuals in Figure 10.21 reveals major problems with the normality assumption, and the Shapiro-Wilk test strongly rejects normality (p-value < 0.0001).

CONCLUSION: The ARIMA$(0,1,0) \times$ ARIMA$(4,1,0)_4$ model for the Australian brick production data (square-root transformed) is not completely worthless, but I would hesitate to use this model for forecasting purposes (since the normality assumption is so grossly violated). The search for a better model should continue!

10.5 Additional topics

DISCUSSION: In this course, we have covered the first 10 chapters of Cryer and Chan (2008). This material provides you with a powerful arsenal of techniques to analyze many time series data sets that are seen in practice. These chapters also lay the foundation for further study in time series analysis.

- **Chapter 11.** This chapter provides an introduction to **intervention analysis**, which deals with incorporating external events in modeling time series data (e.g., a change in production methods, natural disasters, terrorist attacks, etc.). Techniques for incorporating external covariate information and analyzing multiple time series are also presented.

- **Chapter 12.** This chapter deals explicitly with modeling **financial time series** data (e.g., stock prices, portfolio returns, etc.), mainly with the commonly used **ARCH** and **GARCH** models. The key feature of these models is that they incorporate additional heteroscedasticity that are common in financial data.

- **Chapter 13.** This chapter deals with **frequency domain** methods (spectral analysis) for periodic data which arise in physics, biomedicine, engineering, etc. The periodogram and spectral density are introduced. These methods use linear combinations of sine and cosine functions to model underlying (possibly multiple) frequencies.

- **Chapter 14.** This chapter is an extension of Chapter 13 which studies the sampling characteristics of the spectral density estimator.

- **Chapter 15.** This chapter discusses **nonlinear models** for time series data. This class of models assumes that current data are nonlinear functions of past observations, which can be a result of **nonnormality**.