Predictive Analytics: Parametric Models for Regression and Classification Using R

Ajit C. Tamhane
Northwestern University

with contributions from Edward C. Malthouse, Northwestern University

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CONTENTS

CHAPTER 1

INTRODUCTION

Statistical learning is the science of discovering patterns in the data and building models to make predictions and decisions. This term was introduced in the well-known book by Hastie, Tibshirani and Friedman (2001). In computer science the related area is known as **machine learning**, which emphasizes algorithms for model building. Both statistical learning and machine learning have become increasingly important in today's business and scientific worlds driven by the abundance of data, availability of statistical tools and advent of immense and superfast computing power.

Predictive analytics is the part of statistical learning that is concerned with making predictions based on past data. **Data mining** is the part that is concerned with discovering patterns, associations and trends in the data. In the statistical approach to predictive analytics, the focus is on building predictive models and use them to draw inferences and make predictions. In the machine learning approach, oftentimes there is no explicit model — simply a set of rules or an algorithm designed to minimize prediction errors. This book is devoted to parametric statistical models for predictive analytics including regression and classification. In this introductory chapter we provide an overview of the topic of the book in the larger landscape of statistical learning.

Although the origins of predictive analytics, such as linear regression, go more than a century back, the subject is thriving today with many modern developments in data science, including computing and optimization algorithms. In this volume we will focus on classical parametric methods while the companion volume will focus on modern, computer-intensive nonparametric methods.

1.1 Supervised Versus Unsupervised Learning

Statistical learning covers two main broad areas: **supervised learning**, which is synonymous with predictive analytics, and **unsupervised learning**. In supervised learning one of the variables is designated as a **response**, **outcome**, **dependent** or **output** variable. The goal is to model its relationship with other variables, called **predictor**, **covariate**, **independent** or **input** variables (also called **features** in machine learning). Some examples of supervised learning are: predicting customer purchase behavior as a function of past purchase history and demographic and socioeconomic data; predicting the selling price a house as a function of floor area, plot size, amenities, school district and economic indicators; classifying a patient's disease status based on a battery of laboratory tests and classifying a bank loan as in good standing or likely to default as a function of borrower's credit rating, loan amount and terms of the loan.

In unsupervised learning there is no designated response variable; all variables are of the same genre. The goal is to uncover relationships, associations and patterns that underlie the data. Some examples are: determination of latent factors in multivariate data and clustering of customers into groups with similar attributes.

Although the primary goal in predictive analytics is *prediction*, it is also generally of interest to make *inferences* and *interpretations* about the relationships between the predictor variables and the response variable. For example, it is of interest to know which attributes of a house are the key determinants of its price and how much they contribute to the price. What would raise the value of a house more relative to the costs — a finished basement or addition of a sunroom? Going beyond mere prediction, it may be of interest to get a deeper understanding of the phenomenon at hand, especially in scientific studies, by establishing causal connections between predictor variables and the response variable. In a prediction problem, on the other hand, it is of interest to get the best prediction possible regardless of the relative importance of the predictors used. Most supervised learning problems have elements of both prediction and inference.

Because there is a designated response variable in supervised learning, the accuracy of a predictive model can be tested and calibrated by comparing its predictions against the actual observed responses. Usually this is done by dividing the data randomly into a **training set** and a **test set**. The model is fitted on the training set and is tested on the test set. Thus the test set serves as an independent data set. In unsupervised learning there is no simple way to validate a model since there is no response variable.

1.2 Parametric Versus Nonparametric Models

As mentioned earlier. this volume focuses on parametric models of supervised learning. The functional form of these models is specified; only the parameters in the model are unknown and are estimated from data. In nonparametric models, on the other hand, the functional form of the model is either completely unspecified or only partially specified. Thus model fitting involves estimation of this unknown function rather than unknown parameters of a known function.

Examples of parametric models of supervised learning include the models covered in this book: multiple regression, logistic regression, discriminant analysis, generalized linear models and the Cox proportional hazards model for survival data. Examples of nonparametric models of supervised learning include classification and regression trees, random forests, support vector machines and neural nets. Examples of parametric models of un-

supervised learning include correlation analysis, principal components and factor analysis. Examples of nonparametric models of unsupervised learning include cluster analysis and association rules.

Clearly, parametric models require a lot less data for fitting than nonparametric models do. However, if the model is misspecified then we may end up fitting a wrong model. Of course, there are model diagnostics available, which help us detect model misspecification. In nonparametric models this risk is minimized since the data determine the form of the model. Parametric models are easier to fit if the model is correctly specified even if the data are not abundant. Also standard statistical inference procedures are available for them. Although, nonparametric models are more flexible and data-adaptive, they require a data-rich environment and statistical inference procedures for them generally involve bootstrap or resampling, which are highly computer intensive. So parametric models should be used when the form of the model is known or can be approximated by a simple function. Nonparametric models should be used when the functional form of the model is unknown or too complex, but sufficient data are available for its estimation.

1.3 Types of Data and Data Structures

Data are gathered on subjects (e.g., patients), objects (e.g., items or products), processes (e.g., service quality) or organizations (e.g., corporations). We refer to the entities on which the data are gathered generically as **sampling units** even though they may not be samples drawn from a larger population. If the data are gathered on sampling units but not over time, then they are referred to as **cross-sectional data**. Frequently, data are gathered over time on a given entity (e.g., quarterly sales of a company). Such data are referred to as **time-series data**. In this book, for the most part, we deal with cross-sectional data.

Data can be classified in many different ways. One way is by the measurement scale used. There are mainly two measurement scales: **numerical** or **quantitative** and **categorical** or **qualitative**. Quantitative data are of two types: data that can be treated as **continuous** for all practical purposes such as length, height, weight and time, and **discrete**, typically measured on an integer scale, such as count or frequency data. Qualitative data are also of two types: **ordinal** where data consist of ordered labels, e.g., the grade in a course or rating on a scale of 1 to 5 in a customer survey, and **nominal** where data consist of unordered labels, e.g., ethnicity or party affiliation or color of eye. Any set of distinct symbols may be assigned to distinct values of a nominal variable, including numbers, but obviously no arithmetic operations, such as averaging can be done on them. Only ordered numbers may be assigned to distinct values of an ordinal variable, e.g., 0 to 4 for grades F through A for the grade in a course. Again, arithmetic operations don't make sense for ordinal data, although they are often performed such as when computing the grade point average, when we implicitly ascribe numerical scale to ordinal data.

Another way that data can be classified is the type of study or process used to collect them. Mainly there are two types of studies used to collect data: **observational** and **experimental**. Observational data typically come from archival sources or from observational studies in which a phenomenon is observed passively and data are recorded. On the other hand, experimental data come from designed experiments in which predictor variables (called **factors** in the design of experiments terminology) are actively manipulated to assess their effects on the response variable. Observational data are more likely to be subject to confounding from selection and other biases because of possible uncontrolled and unobserved variables (called **noise factors**), whereas experimental data that come from

well-controlled randomized experiments are much less subject to such biases. It should be emphasized that if a research study is biased then no amount of data or fancy statistical analyses can rectify invalid conclusions that can result from such studies. In predictive analytics we generally work with observational data, and so it is extremely important to verify the data quality before conducting any statistical analyses.

Data structures refer to how the data are organized. In this book we only use data with a **matrix structure**, which is one of the simplest data structures. In matrix structure, the data are in the form of a matrix or an array with rows representing the cases or sampling units on which the data are collected and the columns representing the variables. Thus the same variables are measured across all cases but some cells may have missing data. More generally, one can have **relational data** having tree structure, where the variables measured on units depend on other variables measured on those units. For example, the types of complications from pregnancies can only be measured on women subjects, who have been pregnant.

Finally, in this book we only deal with **static data** as opposed to **dynamic data** that are recorded over real time, e.g., online streaming data.

1.4 Overview of Parametric Predictive Analytics

Denote the response variable by y and the predictor variables by $x_1, \ldots, x_p$. We want to build a model relating y to the x's. If y is a numerical variable such as salary, sales, etc. then we have a **regression problem**. If y is a categorical variable, such as success or failure of a treatment, brand of a detergent a customer is likely to buy or letter grade in a course, then we have a **classification problem** since the response is classified into one of several categories.

In a regression problem the predictive model is typically of the form:

$$y = E(y|x_1, \ldots, x_p) + \varepsilon = f(x_1, \ldots, x_p) + \varepsilon, \tag{1.1}$$

where $f(x_1, \ldots, x_p)$ is referred to as the **model**, which is the expected value of y, conditioned on the x's, and ε is the **random error** independent of the x's with a zero mean. The model part f is sometimes referred to as **signal** and the random part is referred to as **noise**. The goal of model fitting is to extract the signal from noisy data. Model **overfitting** occurs when noise is mistakenly fitted as part of the model, e.g., when a higher than necessary degree polynomial is fitted to a y versus x plot to account for wiggles in the data that follow essentially a linear or a quadratic trend.

More generally the response can be multivariate, i.e., there can be $q > 1$ correlated responses, $y_1, \ldots, y_q$. Fitting a regression in terms of p predictors, $x_1, \ldots, x_p$, simultaneously to $y_1, \ldots, y_q$ is referred to as **multivariate regression**, a terminology that is often erroneously used by many practitioners instead of **multiple regression** when there is a single response variable y but multiple predictors. In this book we focus exclusively on a single response variable y.

There are two types of models: **empirical models** and **mechanistic** or **theoretical models**. In an empirical model, the true form of f is unknown but is approximated by a relatively simple function such as a linear or an exponential function. Mechanistic models commonly arise in sciences and engineering and are derived from mechanistic theories underlying a given physical, chemical or biological phenomenon under study. They are often solutions to differential equations that model the phenomenon. Both types of models are parametric in nature, so fitting these models involves estimation of unknown parameters. In this book we mainly focus on building empirical models from observational data.

The simplest empirical model is the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters and ε is assumed to be a $N(0, \sigma^2)$ random variable (r.v.). This model is called a **linear model** because it is linear in the β's, not necessarily in the x's. For example, x_2 could be equal to x_1^2 or $\log x_1$ or x_3 could be equal to $x_1 x_2$. If the specified model is nonlinear in unknown parameters then it is called a **nonlinear regression model**. Mechanistic models are often nonlinear in the β's. Some models are seemingly nonlinear but can be transformed to a linear form, e.g., multiplicative models, which can be log-transformed to a linear form. Models that cannot be transformed into a linear form by such transformations are called intrinsically nonlinear. We do not discuss nonlinear models in this book. Multiple linear regression is covered in Chapters 3-6. Simple linear regression, which is a special case of multiple linear regression for a single predictor, and correlation analysis are covered in Chapter 2.

For binary responses, the logistic regression model discussed in Chapter 7 is commonly used. If y is coded as 1 for success and 0 for failure then $p = E(y)$ denotes the probability of success, which is a function of the x's. The logistic regression model postulates a linear model on the **logistic transform** $\ln[p/(1-p)]$, i.e.,

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Binary responses are assumed to follow the Bernoulli distribution. For polytomous responses, the logistic regression model needs to be extended appropriately depending on whether the responses are nominal or ordinal. These topics are covered in Chapter 7.

Another approach to predicting dichotomous or polytomous responses is discriminant analysis. A binary linear discriminant function is a linear function of the predictor variables that best discriminates between two responses. For polytomous outcomes more than one linear function is needed. Discriminant analysis is covered in Chapter 8.

A generalized linear models (GLM) extends multiple regression and logistic regression models in two different ways. First, it extends the response variable distribution to any member of the so-called **exponential family** of distributions. Both the normal and the Bernoulli distribution belong to this family but there are many others including exponential, gamma and Poisson. Second, it postulates a linear model on the so-called **link function** $g(\mu)$ where $\mu = E(y)$. In the case of multiple regression the link function is the identity function $g(\mu) = \mu$. In the case of logistic regression, the link function is the logistic transform defined above. GLM provides a powerful generalization of the linear model since it allows building predictive models for many other response distributions. Exponential and gamma distributions are used for lifetime data that arise in reliability and survival studies and in marketing (e.g., time since last order). Poisson distribution is used for count data that arise in applications such as traffic studies and marketing (e.g., number of orders). GLM's for different distributions and associated link functions are discussed in Chapter 9.

Finally, in Chapter 10 we cover analysis of survival data. A unique feature of survival data is that they are often censored, i.e., the actual survival time is not observed because the subject (e.g., patient) withdraws from the study or the study is terminated before the event of interest is observed. We cover the Cox proportional hazards regression model to analyze the effects of possible risk factors on lifetimes of patients.

CHAPTER 2

SIMPLE LINEAR REGRESSION AND CORRELATION

One of the simplest and yet commonly occurring data analytic problem is exploring the relationship between two numerical variables. In many applications one of the variables may be regarded as a **response/outcome variable** and the other as a **predictor/explanatory variable** and the goal is to find the best fitting relationship between the two. For example, it may be of interest to predict the amount of sales from advertising dollars or estimate the reduction in tumor size from the amount of radiation exposure. This is referred to as a **regression problem**. In other applications, there is no such distinction between the two variables and it is of interest to simply assess the strength of relationship between them. For example, in hereditary studies it is of interest to evaluate the degree of association between a parental trait (such as height) and the corresponding progeny trait to assess relative contributions from genetic and environmental factors. This is referred to as a **correlation problem**. We study both these problems in this chapter.

We will use the following two data sets to illustrate the methods introduced in this chapter.

EXAMPLE 2.1 (Bacteria Counts: Data)

Chatterjee and Hadi (2012) gave the data shown in Table 2.1 on the number of surviving bacteria (in hundreds) exposed to 200 kv X-rays for 15 six-minute intervals. The main question of interest is how do the bacteria decay with time, in particular, does the exponential decay law apply and if so what is the decay rate? ■

Table 2.1 Surviving bacteria count (in hundreds) at time t (in six-minute intervals)

Time (t)	Bacteria Count (N_t)	Time (t)	Bacteria Count (N_t)
1	355	9	56
2	211	10	38
3	197	11	36
4	166	12	32
5	142	13	21
6	106	14	19
7	104	15	15
8	60		

Source: Chatterjee and Hadi (2012).

EXAMPLE 2.2 (Cardiac Output Measurements: Data)

Frequently we have an accurate measurement method (often called a gold standard) that is expensive or difficult to use in practice. Hence a less accurate but cheaper and a more practical method is used. By making measurements using both methods we can correlate the two and calibrate the less accurate method against the more accurate method. A medical device company compared two methods of cardiac output measurement: an accurate but invasive method and a noninvasive but less accurate method. Cardiac outputs of 26 patients were measured using both methods. The data are shown in Table 2.2. Two questions to ask are: how well are the measurements of the two methods correlated and how accurately can we predict the actual cardiac outputs from the non-invasive method by calibrating it against the more accurate invasive method? ■

2.1 Fitting a Straight Line

2.1.1 Least Squares Method

The simplest equation one can fit to bivariate numerical data is a straight line. Denote by y the response variable and by x the predictor variable and let $\{(x_i, y_i),\ i = 1, \ldots, n\}$ denote a data set of size n. Before fitting a straight line (or more generally any other equation) to data, the first thing to do is to make a scatter plot and see whether it displays any pattern, and if so, what sort of pattern. Of course, it is foolhardy to fit a straight line if the scatter plot displays a nonlinear pattern or displays no pattern at all, i.e., if the plot is just a random scatter. Some nonlinear trends can be linearized by making suitable transformations; see Section 2.1.2 for examples.

In this section we will consider the problem of fitting a straight line $y = \beta_0 + \beta_1 x$ to the data, where β_0 and β_1 are unknown **intercept** and **slope** of the straight line. The **least squares (LS) method** is commonly used to estimate them. The **LS estimates** of β_0 and β_1, denoted by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, minimize the sum of the squared differences between the observed

Table 2.2 Cardiac output measurements (litres/min)

	Method			Method	
Patient	Invasive (x)	Non-invasive (y)	Patient	Invasive (x)	Non-invasive (y)
1	6.3	5.2	14	7.7	7.4
2	6.3	6.6	15	7.4	7.4
3	3.5	2.3	16	5.6	4.9
4	5.1	4.4	17	6.3	5.4
5	5.5	4.1	18	8.4	8.4
6	7.7	6.4	19	5.6	5.1
7	6.3	5.7	20	4.8	4.4
8	2.8	2.3	21	4.3	4.3
9	3.4	3.2	22	4.2	4.1
10	5.7	5.5	23	3.3	2.2
11	5.6	4.9	24	3.8	4.0
12	6.2	6.1	25	5.7	5.8
13	6.6	6.3	26	4.1	4.0

y_i's and their values predicted from the straight line:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{2.1}$$

We refer to Q as the **LS criterion**. The minimum of Q can be found by by setting the partial derivatives $\partial Q/\partial\beta_0$ and $\partial Q/\partial\beta_1$ equal to zero. The resulting equations are called the **normal equations** (see (2.29) and (2.30)). The solutions are

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{2.2}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of the x's and y's, respectively, and for compactness of notation, we have defined

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \quad S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{and} \quad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{2.3}$$

We refer to

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{2.4}$$

as the **LS line**, which is the best fitting straight line that minimizes the LS criterion. This line is used to predict the values of y for given x's. In particular, if $x = \overline{x}$ then it is easy to see from the formula for $\widehat{\beta}_0$ that $\widehat{y} = \overline{y}$. Thus the LS line passes through the midpoint $(\overline{x}, \overline{y})$ of the scatter plot.

An Alternative Useful Form for the LS Line

The correlation coefficient r between x and y is defined in (2.22) in Section 2.3. Using that formula we can write

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \widehat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \widehat{\beta}_1\left(\frac{s_x}{s_y}\right), \tag{2.5}$$

where s_x and s_y are the standard deviations of the x's and the y's given by (2.22). Now equation (2.4) for the LS line can be written as $\widehat{y} - \overline{y} = \widehat{\beta}_1(x - \overline{x})$. Then substituting for $\widehat{\beta}_1$ from above and rearranging the terms we get

$$\frac{\widehat{y} - \overline{y}}{s_y} = r\left(\frac{x - \overline{x}}{s_x}\right). \tag{2.6}$$

This equation has a nice interpretation that one standard deviation change in x from its mean results in a change of r standard deviation units in y from its mean. The sign of change in y is the same as the sign of change in x if $r > 0$ and the sign is reversed if $r < 0$. Thus the change in x is modulated by the magnitude of the correlation; if $r = 0$ then a change in x results in no change in y.

EXAMPLE 2.3 (Galton Data: Regression to the Mean)

In a classic study of correlation between fathers' heights (x) and sons' heights (y) based on a sample of 1078 father-son pairs, the British scientist Sir Francis Galton (1822-1911) found that the average height of fathers is $\overline{x} = 68$", the average height of sons is $\overline{y} = 69$"; thus sons are 1" taller than their fathers on the average. Furthermore, the correlation between x and y is 0.5 and the standard deviations of x and y are roughly equal $= 2.7$". Consider two fathers who are 4" taller and shorter than average, i.e., 72" and 64" in height. Then from Equation (2.6) we see that their sons will be 71" and 67" tall on the average, respectively. Note that they will not be 1" taller than their fathers. The tall fathers' sons will be taller than average but not as tall as their fathers, while short fathers' sons will be shorter than average but not as short as their fathers. Galton called this phenomenon as **regression to the mean**. ■

2.1.2 Linearizing Transformations

An example of a nonlinear relationship that can be linearized by **log-transformation** of the data is the so-called **power law**, $y = ax^b$. The log-transformed equation is $\ln y = \ln a + b \ln x$, which is a straight line on the log-log scale, so the model is called the **log-log model**. By making transformations $y \to \ln y$ and $x \to \ln x$, we get a straight line with intercept $\beta_0 = \ln a$ and slope $\beta_1 = b$.

There are many empirical applications of the power law. Here we consider its application to model the demand-price relationship in economics. Suppose y is the demand and x is the price of a product. Then differentiating both sides of the equation $\ln y = \ln a + b \ln x$, we get $dy/y = b(dx/x)$ or $b = (dy/y)/(dx/x)$. Thus b represents the relative change in demand due to a unit relative (e.g., one percent) change in price. This is known as the **price elasticity** of the product. Generally, $b < 0$ so that as price increases, demand decreases. If $b = 0$ then demand is said to be inelastic with respect to price.

Another such relationship is the so-called **exponential law**, $y = a\exp(bt)$ (where t is time), which is used to model exponential growth (for $b > 0$) or exponential decay (for $b < 0$). In this case, the log-transformed equation is $\ln y = \ln a + bx$. This model is called the **log model**. By making transformations $y \to \ln y, x \to x$, we get a straight line with

intercept $\beta_0 = \ln a$ and slope $\beta_1 = b$. By fitting a straight line to the transformed data, we can get the LS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ from which the estimates of a and b can be determined.

EXAMPLE 2.4 (Bacteria Counts: LS Estimates and Scatter Plots)

Consider the bacteria decay data in Table 2.1. The left panel of Figure 2.1 shows the scatter plot of bacteria count (N_t) versus time (t). The plot is negatively curved with a large outlier at $t = 1$. The theory suggests the exponential decay model

$$N_t = N_0 e^{bt} \quad \text{for } t \geq 0.$$

We make the logarithmic transformation yielding a straight line model: $\ln N_t = \ln N_0 + bt = \beta_0 + \beta_1 t$. The right panel of Figure 2.1 shows the scatter plot of $\ln N_t$ versus t, which shows a clear linear trend.

For the log-transformed data with $x = t$ and $y = \ln N_t$, we can calculate the following statistics:

$$\overline{x} = 8.0, \overline{y} = 4.226, S_{xx} = 280, S_{yy} = 13.516, S_{xy} = -61.159.$$

So,

$$\widehat{\beta}_1 = \frac{-61.159}{280} = -0.218 \quad \text{and} \quad \widehat{\beta}_0 = 4.226 + (0.218)(8.0) = 5.973.$$

Thus the fitted LS line is $\widehat{\ln N_t} = 5.973 - 0.218t$. So $\widehat{N}_0 = e^{5.973} = 392.68$ and the fitted exponential decay model is

$$\widehat{N}_t = 392.68 e^{-0.218t} \quad \text{for} \quad t \geq 0.$$

In radioactivity applications it is frequently of interest to estimate the half-life, denoted by $t_{0.5}$, which is the time at which the concentration of radioactive compound, or in the present case, the number of surviving bacteria, will be 50% of the initial amount. From the above model we see that $N_t/N_0 = 0.5$ when $\ln 0.5 = bt$, so $t_{0.5} = \ln 0.5/b$. Therefore the estimated half-life is $\widehat{t}_{0.5} = \ln 0.5/(-0.218) = (-0.693)/(-0.218) = 3.180$ time units, which is $3.180 \times 6 = 19.08$ minutes. ■

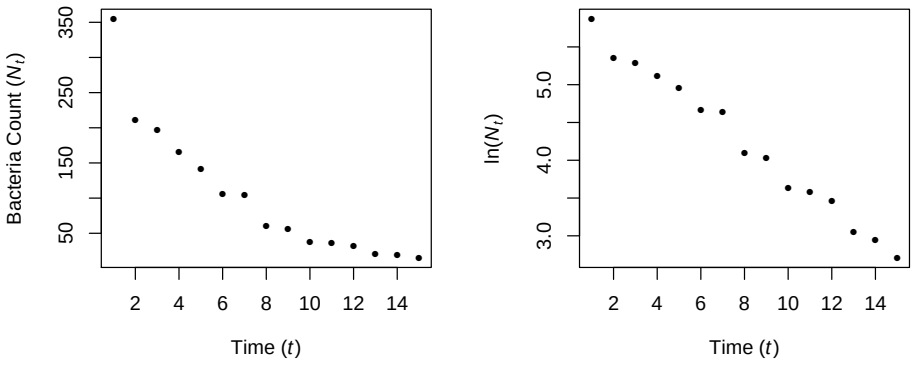


Figure 2.1 Plots of N_t versus t (left) and $\ln N_t$ versus t (right)

2.1.3 Fitted Values and Residuals

To check the model assumptions and detecting outliers, it is useful to calculate the **fitted values** and **residuals** given by

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad \text{and} \quad e_i = y_i - \widehat{y}_i \quad (i = 1, \ldots, n), \tag{2.7}$$

respectively. Figure 2.2 shows these two quantities. It can be shown that the residuals satisfy the following two linear constraints:

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i e_i = 0. \tag{2.8}$$

Thus given any $n-2$ of the n residuals, the remaining two can be determined from these two equations. Hence the **error degrees of freedom (error d.f.)** associated with the residuals is $n-2$. From the first equation above, it follows that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$; thus the average of the fitted values equals $\overline{y}$.

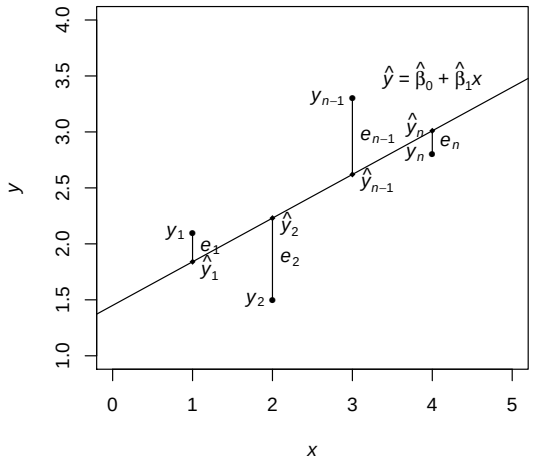


Figure 2.2 LS straight line fit

Residuals are useful for checking the goodness of fit and diagnose model violations. We will study model diagnostic uses of residuals in Chapter 4.

2.1.4 Assessing Goodness of Fit

If the straight line model is correct then the residuals obtained after filtering out the fitted straight line should be randomly distributed around zero. This can be assessed by plotting the residuals versus x_i's as shown in the following example.

EXAMPLE 2.5 (Bacteria Counts: Residual Plots)

We saw in Figure 2.1 that the plot of N_t versus t is curved, while the plot of $\ln N_t$ versus t is linear, both with negative slopes. The corresponding residual plots are shown in Figure 2.3. Notice a highly curved residual plot in the left panel and a

random residual plot in the right panel. A lesson to draw here is that the residual plot shows departures from the fitted straight line much more clearly than the scatter plot of y versus x since the linear part of the relationship has been filtered out.

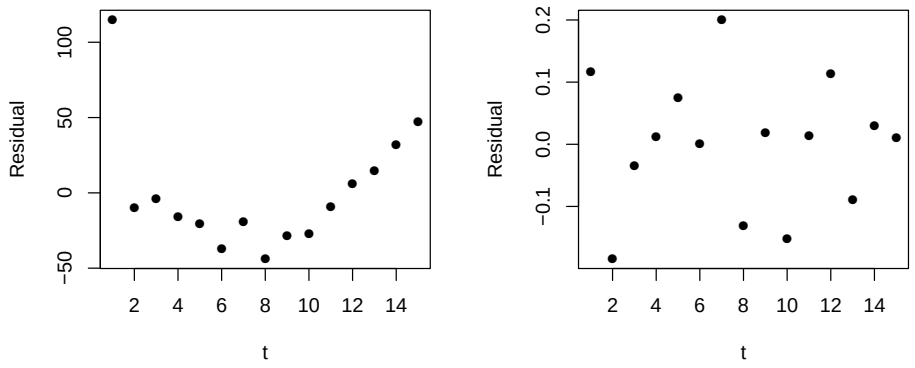


Figure 2.3 Plots of residuals from LS fits of N_t vs. t (left panel) and $\ln N_t$ vs. t (right panel)

■

A numerical measure of goodness of fit of the LS line is obtained by comparing the residual variation of the y_i's around the LS line, referred to as the **error sum of squares (SSE)**, with the total variation in the y_i's around their mean $\overline{y}$, referred to as the **total sum of squares (SST)**. The difference between SST and SSE is called the **regression sum of squares (SSR)**, as it represents the part of the total variation of the y_i's that is accounted for by regression of the y_i's on the x_i's.

To see this relationship we write the deviation of each y_i around $\overline{y}$ as the sum of two parts:

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + e_i \quad (i = 1, \ldots, n). \tag{2.9}$$

If we square both sides and sum over $i = 1, \ldots, n$, it turns out that the cross-product term,, $2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})e_i$ equals zero. So

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SSE}}. \tag{2.10}$$

Note that SSE is just the minimum value of the LS criterion Q in (2.1). We refer to (2.10) as the **analysis of variance (ANOVA) identity**.

The proportion of variation accounted for by the LS line is given by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{2.11}$$

called the **coefficient of determination**. Note that $0 \leq R^2 \leq 1$ and a higher value indicates a better fit.

In the Technical Notes section we derive the formula $\text{SSR} = \widehat{\beta}_1^2 S_{xx}$. Then SSE can be computed from $\text{SST} - \text{SSR} = S_{yy} - \widehat{\beta}_1^2 S_{xx}$. These calculations are illustrated in the following example.

EXAMPLE 2.6 (Bacteria Counts: Goodness of Fit)

Using the calculations from Example 2.4, we get

$$\text{SST} = S_{yy} = 13.516, \quad \text{SSR} = \widehat{\beta}_1^2 S_{xx} = (-0.218)^2(280) = 13.359$$

and

$$\text{SSE} = \text{SST} - \text{SSR} = 13.516 - 13.359 = 0.157.$$

Hence $R^2 = \text{SSR}/\text{SST} = 13.359/13.516 = 98.8\%$. For a straight line fit of N_t versus t, we can similarly compute $R^2 = 82.3\%$. Thus the log model provides a significantly better fit than the straight line model. ■

EXAMPLE 2.7 (Anscombe Data: Scatter Plots)

Anscombe (1973) constructed four different bivariate data sets shown in Table 2.3 such that the same LS line fits all four data sets. The scatter plots are given in Figure 2.4 along with their LS fitted lines. The scatter plot for Data Set I shows a straight line trend. The scatter plot for Data Set II is a parabola. All points in the scatter plot for Data Set III follow an almost perfect straight line except one large outlier. Finally, all observations in Data Set IV are at $x = 8$ except one at $x = 19$. Yet, the same LS line, namely $\widehat{y} = 3.0 + 0.5x$, fits all four data sets. Not only is the LS fitted line the same, but all statistics discussed in the following sections, e.g., the t-statistics for the regression coefficients, are the same for all four data sets. Thus we won't be able to tell these data sets apart if we just fit the straight lines to them without looking at the scatter plots.

What is the explanation of this seemingly bizarre result? We will see in what follows that all simple linear regression statistics depend on the raw data only through the following five quantities, whose values are the same for the four data sets:

$$\overline{x} = 9.00, \overline{y} = 7.50, S_{xx} = 110.0, S_{yy} = 41.27, S_{xy} = 55.00.$$

This result should not come as a surprise since it is similar to the univariate case where two very different data sets can have the same mean and variance. Therefore it is important to visualize the data first by making appropriate plots before computing various statistics. ■

2.2 Statistical Inferences for Simple Linear Regression

2.2.1 Simple Linear Regression Model

To test hypotheses or compute confidence intervals (CI's) on β_0 and β_1, we need to assume a probability model for the data. The standard normal theory model for simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n), \tag{2.12}$$

where the ε_i's are assumed to be independent and identically distributed (i.i.d.) $N(0, \sigma^2)$ **random errors** and σ^2 is an unknown **error variance**. This model is shown graphically in Figure 2.5. It follows that the y_i's are independent normally distributed random variables with means and variance given by

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_i \quad \text{and} \quad \text{Var}(y_i) = \sigma^2 \quad (i = 1, \ldots, n). \tag{2.13}$$

There are four assumptions implicit in this model. They are **normality**, constant variance (called **homoscedasticity**), **independence** and **linearity** of $E(y)$ with respect to x. In

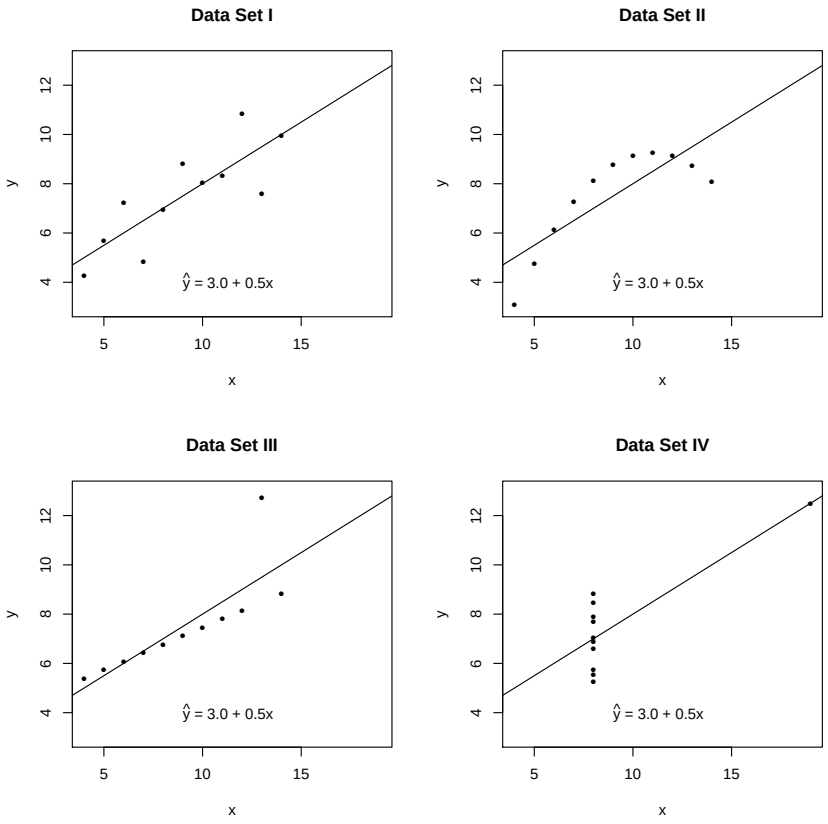


Figure 2.4 Scatter plots for Anscombe data sets with LS fitted lines

Table 2.3 Anscombe data sets

	Data Set I		Data Set II		Data Set III		Data Set IV	
No.	x	y	x	y	x	y	x	y
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
11	4	4.26	4	3.10	4	5.39	19	12.50

Source: Anscombe (1973).

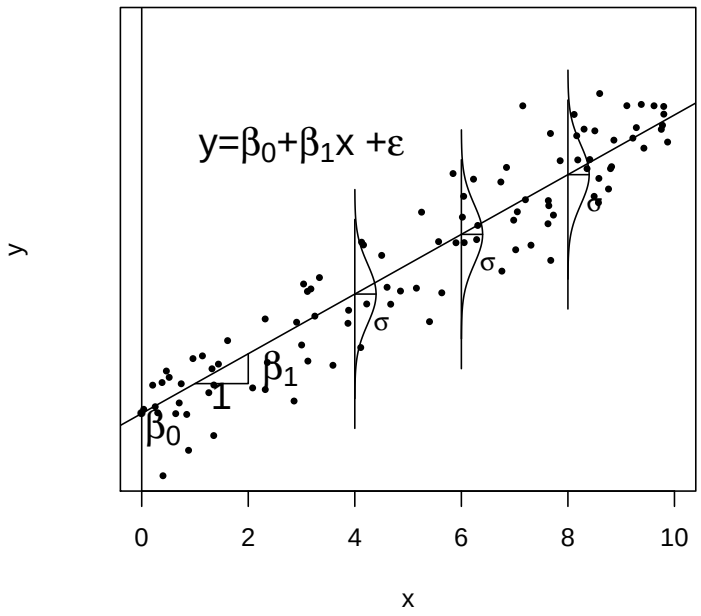


Figure 2.5 Simple linear regression model

addition, we assume that the x_i's are non-random. In Chapter 4 we will discuss residuals plots and other methods to check these assumptions and how to deal with any violations. The inferential methods given in the following sections are strictly valid only under these assumptions.

2.2.2 Inferences on β_0 and β_1

The sampling distributions of the LS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normal with $E(\widehat{\beta}_0) = \beta_0$ and $E(\widehat{\beta}_1) = \beta_1$ (i.e., they are unbiased estimators), and

$$\text{Var}(\widehat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}} \quad \text{and} \quad \text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

To estimate the error variance σ^2 we compute the sample variance of the residuals, which have $n-2$ degrees of freedom (d.f) as noted before. The mean of the residuals is zero, so the estimate of σ^2 equals

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\text{SSE}}{n-2} = \text{MSE}, \tag{2.14}$$

where MSE stands for **mean square error**. Thus $s = \sqrt{\text{MSE}}$ is the **root mean square error (RMSE)** of the residuals, which is used to estimate σ.

It can be shown that the sampling distribution of $(n-2)s^2/\sigma^2 = \text{SSE}/\sigma^2$ is χ^2 with $n-2$ d.f. independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Replacing σ^2 by its estimate s^2 in the formulae for the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and taking square roots, we get their estimated standard deviations, called the **standard errors (SE's)**:

$$\text{SE}(\widehat{\beta}_0) = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}} \quad \text{and} \quad \text{SE}(\widehat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}.$$

It then follows that

$$\frac{\widehat{\beta}_0 - \beta_0}{\text{SE}(\widehat{\beta}_0)} \sim t_{n-2} \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\text{SE}(\widehat{\beta}_1)} \sim t_{n-2},$$

where t_{n-2} denotes Student's t-distribution with $n-2$ d.f. Thus $100(1-\alpha)\%$ **confidence intervals (CI's)** on β_0 and β_1 are given by

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2}\text{SE}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm t_{n-2,\alpha/2}\text{SE}(\widehat{\beta}_1),$$

where $t_{\nu,\alpha/2}$ is the $100(1-\alpha/2)$th percentile (also called the upper $\alpha/2$ critical point) of the t-distribution with ν d.f. These critical points are tabulated in Table A.2.

The significance of the linear component of the relationship between y and x can be assessed by testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. The test statistic is

$$t = \frac{\widehat{\beta}_1}{\text{SE}(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1\sqrt{S_{xx}}}{s}. \tag{2.15}$$

The α-level test rejects H_0 if

$$|t| > t_{n-2,\alpha/2}. \tag{2.16}$$

Equivalently, we reject H_0 if the P-value of the test statistic is less than α.

EXAMPLE 2.8 (Bacteria Counts: Inferences on Intercept and Slope Coefficients)

To calculate the standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ we first calculate $s = \text{RMSE} = \sqrt{0.0121} = 0.110$. Also, $S_{xx} = 280$ as calculated before. Hence

$$\text{SE}(\widehat{\beta}_0) = 0.110\sqrt{\frac{1240}{15 \times 280}} = 0.0598 \quad \text{and} \quad \text{SE}(\widehat{\beta}_1) = \frac{0.110}{\sqrt{280}} = 0.0066.$$

Table 2.4 ANOVA table for simple linear regression

Source	SS	d.f.	MS	F
Regression	SSR	1	MSR	$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n-2$	MSE	
Total	SST	$n-1$		

So the t-statistics are

$$\beta_0:\ t=\frac{5.973}{0.0598}=99.92 \quad \text{and} \quad \beta_1:\ t=\frac{-0.218}{0.0066}=-33.22,$$

both of which are highly significant ($P<0.001$).

A 95% CI on β_1 is given by

$$\widehat{\beta}_1 \pm t_{13,0.025}\text{SE}(\widehat{\beta}_1) = -0.218 \pm (2.160)(0.0066) = [-0.232, -0.204].$$

From this we get the following 95% CI on the half-life $t_{0.5}=(\ln 0.5)/\beta_1$:

$$\left[\frac{\ln 0.5}{-0.232}, \frac{\ln 0.5}{-0.204}\right] = \left[\frac{-0.693}{-0.232}, \frac{-0.693}{-0.204}\right] = [2.987, 3.397].$$

■

2.2.3 Analysis of Variance for Simple Linear Regression

The purpose of the analysis of variance (ANOVA) is to partition the total variation in the response variable y into independent components so that each component can be attributed to a separate source of variation. In the case of simple linear regression, there are only two sources of variation, namely the variation caused by different x-values through their linear relationship with y and the residual or error variation. The ANOVA identity in (2.10) gives this decomposition.

The total d.f. is always $n-1$ since SST measures the variation of the y_i's around their mean $\overline{y}$, just as we use $n-1$ d.f. to calculate the sample variance of a sample of size n around its mean. Corresponding to the ANOVA identity (2.10), the partitioning of this total d.f. is as follows:

$$\underbrace{n-1}_{\text{Total d.f.}} = \underbrace{1}_{\text{Regression d.f.}} + \underbrace{n-2}_{\text{Error d.f.}},$$

where the regression d.f. = 1 because the regression equation has one predictor variable and the error d.f. $= n-2$ as explained before. A **sum of squares (SS)** divided by its d.f. is referred to as a **mean square (MS)**. Thus, the **mean square regression (MSR)** equals SSR/1 and the **mean square error (MSE)** equals SSE/$(n-2)$ as defined in (2.14).

It can be shown that under the null hypothesis $H_0:\beta_1=0$, the ratio $F=\text{MSR}/\text{MSE}$ has an F-distribution with 1 and $n-2$ d.f. Thus an α-level test rejects H_0 if

$$F=\frac{\text{MSR}}{\text{MSE}} > f_{1,n-2,\alpha}, \tag{2.17}$$

where $f_{1,n-2,\alpha}$ is the upper α critical point of this F-distribution. These critical points are tabulated in Table A.4. The calculations of the sums of squares and mean squares are presented in the **ANOVA table** shown in Table 2.4. The t-test given by (2.16) to test $H_0:\beta_1=0$ is equivalent to the F-test because $F=t^2$ (since $\text{MSR}=\text{SSR}=\widehat{\beta}_1^2 S_{xx}$ and $\text{MSE}=s^2$) and $f_{1,\nu,\alpha}=t^2_{\nu,\alpha/2}$. See the example below.

Table 2.5 ANOVA table for regression of ln(Bacteria Count) on Time

Source	SS	d.f.	MS	F	P
Regression	13.359	1	13.359	1103.70	0.000
Error	0.157	13	0.0121		
Total	13.516	14			

EXAMPLE 2.9 (Bacteria Counts: Analysis of Variance)

The ANOVA table for the straight line fit to the log-transformed bacteria data is shown in Table 2.5. We see that $F = 1103.70$ with 1 and 13 d.f. is highly significant ($P < 0.001$). Thus there is a significant linear component to the fit. Note that $F = 1103.70 = t^2 = (-33.22)^2$ and $f_{1,13,0.01} = 9.07 = t^2_{13,0.005} = (3.012)^2$. ■

2.2.4 Pure Error versus Model Error

We have used MSE as an estimator of σ^2 but it is an unbiased estimator only if the model is correctly specified. Otherwise the misspecified part of the model contaminates and inflates this estimator. Therefore we refer to MSE as the **model error estimator**. We are rarely certain that the specified model is correct, especially when it is an empirical model. Therefore the only way to obtain an unbiased estimator of σ^2 (called the **pure error estimator**) is by independent repeat observations at each x_i and then pooling the sample variances among these repeat observations (assuming homoscedasticity). This is generally possible only in designed experiments where repeat observations should be planned for this purpose. The pure error estimator can also be used to do a **lack of fit (LOF) test** to check if the model is misspecified. Exercise 2.13 gives the details of the LOF test.

2.2.5 Prediction of Future Observations

Having fitted the LS line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$, we often want to use it to predict a future value y^* for a given x^*. For example, suppose a straight line is fitted to data on tread wear (y) as a function of mileage (x) for a particular brand and make of a car tire. A consumer may want to use this LS line to know whether the tire that she purchases would wear out by 50,000 miles. This is a **prediction problem**. There is a related **estimation problem** that the tire manufacturer faces, namely, estimate the mean amount of tread wear at 50,000 miles for *all* tires of that brand and make. The two problems are different because in the prediction problem we want predict a future *random* outcome y^*, the actual amount of tread wear for a *random* tire, while in the estimation problem we want to estimate an unknown *fixed mean* $\mu^* = E(y^*) = \beta_0 + \beta_1 x^*$, the mean amount of tread wear for *all* tires. The **calibration problem** is the inverse of the estimation problem; see Exercise 2.6.

In both prediction and estimation problems we use the same formula for the predictor or the estimator:

$$\widehat{y}^* = \widehat{\mu}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*. \tag{2.18}$$

The difference arises when we want to calculate an interval around it. For estimating a fixed parameter we use a CI, while for predicting a random outcome we use a **prediction**

interval (PI). A $100(1-\alpha)\%$ CI for μ^* is given by

$$\widehat{\mu}^* \pm t_{n-2,\alpha/2}\text{SE}(\widehat{\mu}^*), \tag{2.19}$$

where

$$\text{SE}(\widehat{\mu}^*) = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

A $100(1-\alpha)\%$ PI for y^* is given by

$$\widehat{y}^* \pm t_{n-2,\alpha/2}\sqrt{s^2 + \text{SE}^2(\widehat{\mu}^*)} = \widehat{y}^* \pm t_{n-2,\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}. \tag{2.20}$$

The extra s^2 term added to $\text{SE}^2(\widehat{\mu}^*)$ is the estimate of the additional variance σ^2 due to the random outcome y^*. As we shall see in the example below, the PI is generally much wider than the CI. Also notice that both the PI and the CI become increasingly wider as x^* moves away from $\overline{x}$ reflecting the fact that they become increasingly less precise. Additional error may be introduced if we try to predict y^* for x^* outside the range of the observed x's since the fitted model may not hold. This is called **extrapolation**, which should be generally avoided.

EXAMPLE 2.10 (Cardiac Output Measurements: Prediction and Confidence Intervals)

The scatter plot of the cardiac output data is shown in Figure 2.6 along with the LS fitted line $\widehat{y} = -0.528 + 1.014x$. The estimated standard deviation equals $s = 0.495$ with 24 d.f. The $R^2 = 90.6\%$, which indicates a strong linear relationship.

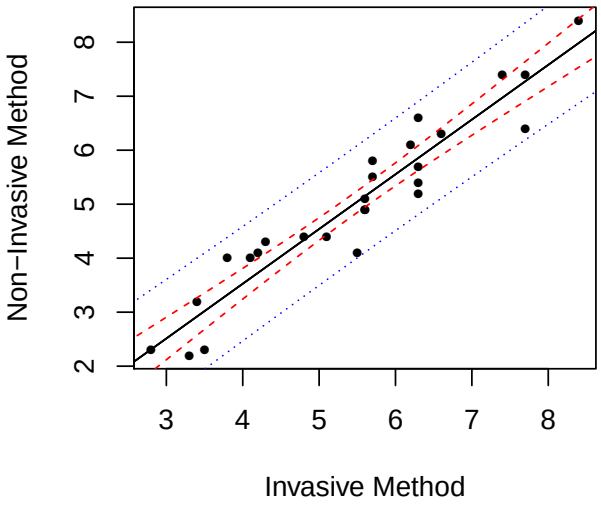


Figure 2.6 Scatter plot of cardiac outputs measured with invasive and non-invasive methods

Now suppose that we want to estimate the range of non-invasive method y^*-values that are likely to be observed for $x^* = 6$ litres/min using the invasive method. Then we would need to calculate a PI. The predicted value equals

$$\widehat{y}^* = -0.528 + 1.014 \times 6 = 5.556.$$

To compute a 95% PI for y^* we first calculate $\overline{x} = 5.469$ and $S_{xx} = 55.225$ and note that $t_{24,0.025} = 2.064$. Then a 95% PI can be calculated as

$$\begin{aligned} & 5.556 \pm 2.064 \times 0.495\sqrt{1 + \frac{1}{26} + \frac{(6-5.469)^2}{55.225}} \\ = \quad & [5.556 \pm 1.044] = [4.512, 6.600]. \end{aligned}$$

If the true cardiac output using the invasive method is 6 litres/min then the non-invasive method reading would fall in this interval with 95% confidence.

For the sake of comparison we calculate a 95% CI for μ^* if the expected cardiac output measured by the non-invasive method if $x^* = 6$ litres/min. This CI is given by

$$\begin{aligned} & 5.556 \pm 2.064 \times 0.495\sqrt{\frac{1}{26} + \frac{(6-5.469)^2}{55.225}} \\ = \quad & [5.556 \pm 0.213] = [5.343, 5.769]. \end{aligned}$$

Note that the confidence interval is much narrower than the prediction interval. The R output below gives the same results.

```
> cardiac=read.csv("c:/data/cardiac.csv")
> fit=lm(Noninvasive~Invasive,cardiac)
> summary(fit)
Call:
lm(formula = Noninvasive ~ Invasive, data = cardiac)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.52783    0.37685  -1.401    0.174
Invasive     1.01353    0.06658  15.222 7.88e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4947 on 24 degrees of freedom
Multiple R-squared:  0.9061,    Adjusted R-squared:  0.9022
F-statistic: 231.7 on 1 and 24 DF,  p-value: 7.876e-14

> predict(fit,newdata=data.frame(Invasive=6.0),interval="predict")
       fit      lwr      upr
1 5.553334 4.510229 6.596439
> predict(fit,newdata=data.frame(Invasive=6.0),interval="confidence")
       fit      lwr      upr
1 5.553334 5.340211 5.766457
```

■

2.3 Correlation Analysis

In correlation analysis we assume that both x and y are random variables (in contrast to regression analysis where x is assumed to be nonrandom) with a joint distribution. The

population correlation coefficient ρ of this joint distribution is defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where σ_{xy} is the population covariance[1] between x and y, and σ_x and σ_y are the population standard deviations of x and y.

The sample estimate of ρ, called the **Pearson correlation coefficient** and denoted by r, can be obtained by replacing σ_{xy}, σ_x and σ_y by their following sample estimates:

$$\begin{aligned} s_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{S_{xy}}{n-1}, \\ s_x &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{S_{xx}}{n-1}} \\ s_y &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}} = \sqrt{\frac{S_{yy}}{n-1}}. \end{aligned} \tag{2.21}$$

Thus the Pearson correlation coefficient is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \tag{2.22}$$

Note that $-1 \leq r \leq 1$ and the sign of r is that of S_{xy} or equivalently that of $\widehat{\beta}_1 = S_{xy}/S_{xx}$. A positive sign indicates an increasing relationship while a negative sign indicates a decreasing relationship.

The R^2 defined in (2.11) is directly related to r; in fact $R^2 = r^2$ which follows from

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{\widehat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{\text{SSR}}{\text{SST}} = R^2.$$

Keep in mind that both ρ and its sample estimate r are measures only of the linear association between x and y and not of any other type of association (e.g., general monotone association). Another point to keep in mind is that correlation does not equal causation. Thus two variables being correlated does not necessarily mean that one variable is causing the other variable. The correlation between two variables may be spurious, caused by a third variable, called a **lurking variable**. The following example illustrates this phenomenon.

EXAMPLE 2.11 (Chocolate Consumption and Nobel Laureates)

Messerli (2012) showed a highly significant correlation ($r = 0.791, P < 0.001$) between per capita chocolate consumption and per capita Nobel laureates for 23 countries. The plot is shown in Figure 2.7. Media jumped on this study quite uncritically with headlines such as "Eat Chocolate, Win the Nobel Prize," in *Reuters*, "Study Links Eating Chocolates to Winning Nobel Prizes," in *USA Today* and "Chocolate and Nobel Prizes Linked in Study," in *Forbes*. Besides many flaws in the study (see, e.g., McClintock et al. (2014)) such as the data on chocolate consumption and Nobel prizes are from different time periods and there is no evidence that the Nobel laureates themselves ate a lot of chocolates, there is a key lurking variable, namely, how affluent the countries are. Obviously, developed and affluent countries can afford to spend on a non-staple food such as chocolates and also invest in research which leads to Nobel prizes. ■

[1]The population covariance is defined as $\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$, where μ_x and μ_y are the expected values of x and y.

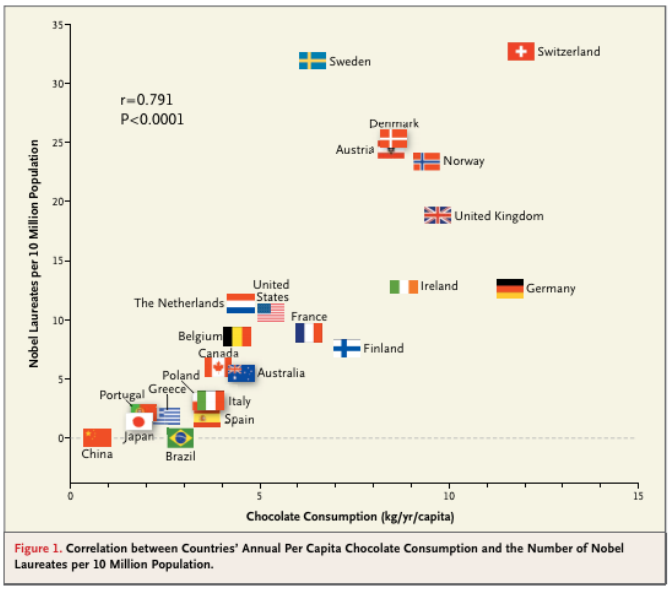


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Figure 2.7 Nobel laureates per capita versus chocolate consumption per capita for selected countries

Often, false causal effect is claimed when the observed effect is the result of regression to the mean phenomenon discussed in Example 2.3. A classic example of this was given by Tversky and Kahnemann (1973) in which the instructors in a flight school adopted a policy of positive reinforcement (e.g., praise), recommended by psychologists, after each successful flight maneuver. Unfortunately, they found that the performance typically declined at the next flight maneuver. From this they concluded that high praise has a negative effect on fliers' performance. However, the explanation lies in regression to the mean phenomenon, which results in an outcome variable regressing toward the mean, thus a high value is followed by a low value. Such changes are not caused by any policy change. Another example in the same vein is when one is trying to predict the final exam score from the midterm exam score of a student. If the correlation between the two scores is about 0.5 then as can be seen from Example 2.3, students who score higher than average on the midterm will tend to score higher than average on the final but not as high. Similarly, students who score lower than average on the midterm will tend to score lower than average on the final but not as low. To conclude from this that students who score higher than average on the midterm tend to goof off and students who score lower than average on the midterm tend to work harder would be wrong. Such an argument is called **regression fallacy**.

2.3.1 Bivariate Normal Distribution

Next we shall discuss statistical inference on ρ. First we assume a probability model for the joint distribution of (x, y). Generalizing the bell-shaped curve of the univariate normal p.d.f. to a bivariate normal p.d.f. gives a bell-shaped surface plotted over the (x, y)-plane. Whereas the univariate normal distribution has two parameters, the mean μ and variance σ^2, the bivariate normal distribution has five parameters, the mean μ_x and variance σ_x^2 for

x, the mean μ_y and variance σ_y^2 for y, and the covariance σ_{xy} or equivalently the correlation coefficient ρ between x and y.

The bivariate normal p.d.f. is given by

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}[u^2+v^2-2\rho uv]\right\} \tag{2.23}$$

for $-\infty < x, y < +\infty$, where

$$u = \frac{x-\mu_x}{\sigma_x} \quad \text{and} \quad v = \frac{y-\mu_y}{\sigma_y}.$$

Note that this p.d.f. becomes degenerate when $\rho = \pm 1$, i.e., when y is a deterministic linear function of x, say $y = a + bx$ for some constants a and $b \neq 0$. In that case the p.d.f. is concentrated on that line instead of being spread over the (x, y)-plane. Hence we restrict the range of ρ to $-1 < \rho < 1$.

The following are some useful properties of the bivariate normal distribution.

1. The marginal distributions of x and y are $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$, respectively.

2. The conditional distribution of y conditioned on x, which is obtained by dividing the joint distribution (2.23) by the $N(\mu_x, \sigma_x^2)$ p.d.f., is normal with mean and variance given by
$$E(y|x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x) \quad \text{and} \quad \text{Var}(y|x) = \sigma_y^2(1-\rho^2).$$

3. If we put
$$\beta_0 = \mu_y - \frac{\rho\sigma_y}{\sigma_x}\mu_x, \quad \beta_1 = \frac{\rho\sigma_y}{\sigma_x} \quad \text{and} \quad \sigma^2 = \sigma_y^2(1-\rho^2), \tag{2.24}$$
then we see that the conditional distribution of y is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional variance σ^2. This is exactly the simple linear regression model (2.13), which can be derived from the bivariate normal distribution of (x, y) as the conditional distribution of y conditioned on x.

4. The conditional variance σ^2 of y is smaller than the unconditional variance σ_y^2 by a factor of $(1-\rho^2)$. If $\rho = \pm 1$ then the conditional variance of y is zero because in that case, for any given x, y is fixed according to the deterministic linear relationship, $y = a + bx$.

2.3.2 Inferences on ρ

Inferences on ρ are based on the sample correlation coefficient r. When $\rho = 0$, it can be shown that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}. \tag{2.25}$$

This can be used as a t-statistic to test $H_0 : \rho = 0$. In fact, as can be readily verified, this t-statistic is algebraically the same as the t-statistic (2.15) used to test $H_0 : \beta_1 = 0$. This is not surprising since from (2.24) we know that β_1 is proportional to ρ and so $\beta_1 = 0$ if and only if $\rho = 0$.

To test a more general hypothesis such as $H_0 : \rho = \rho_0$, where $\rho_0 \neq 0$, or to obtain a CI on ρ we need the so-called noncentral distribution of r when $\rho \neq 0$. This distribution is complicated and not amenable to easy manipulation. To get around this difficulty, Sir R.A. Fisher (1890-1962) suggested the following transformation of ρ and its sample estimate,

$$\psi = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \quad \text{and} \quad \widehat{\psi} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

He showed that $\widehat{\psi}$ is asymptotically normal with

$$E(\widehat{\psi}) \approx \psi = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \quad \text{and} \quad \text{Var}(\widehat{\psi}) \approx \frac{1}{n-3}. \tag{2.26}$$

Then an approximate large sample $100(1-\alpha)\%$ CI on ψ is given by

$$\widehat{\psi} \pm z_{\alpha/2}\frac{1}{\sqrt{n-3}} = [L, U] \quad \text{(say)}. \tag{2.27}$$

From (2.26), we have $\rho = [e^{2\psi}-1]/[e^{2\psi}+1]$, which is a monotone function of ψ. Hence by substituting the lower and upper confidence limits on ψ, namely L and U, in the formula for ρ yields the following confidence limits on ρ:

$$\left[\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2U}-1}{e^{2U}+1}\right]. \tag{2.28}$$

EXAMPLE 2.12 (Cardiac Output Measurements: Inference on Correlation Coefficient)

Suppose that for the non-invasive method to be acceptable, its correlation coefficient ρ with the invasive method must be greater than 0.90. The sample correlation coefficient for these data can be computed to be $r = 0.952$, which exceeds 0.90 but we need to test whether it is significantly greater than 0.90. In other words, we need to test the one-sided hypotheses $H_0 : \rho \leq 0.90$ versus $H_1 : \rho > 0.90$ or equivalently $H_0 : \psi \leq \psi_0$ versus $H_1 : \psi > \psi_0$, where

$$\psi_0 = \frac{1}{2}\ln\left(\frac{1+0.90}{1-0.90}\right) = 1.472.$$

The test will be based on

$$\widehat{\psi} = \frac{1}{2}\ln\left(\frac{1+0.952}{1-0.952}\right) = 1.853.$$

The test statistic equals

$$z = \frac{\widehat{\psi}-\psi_0}{\sqrt{1/(n-3)}} = \frac{1.853-1.472}{\sqrt{1/(26-3)}} = 1.827.$$

The one-sided P-value of this statistic equals 0.034, which shows that H_0 can be rejected at the 0.05 level and hence we can conclude that $\rho > 0.90$.

We obtain the same result using the CI method. In this case we need to calculate a lower 95% confidence bound on ρ. First we calculate a lower 95% confidence bound on ψ:

$$L = \widehat{\psi} - z_{0.05}\frac{1}{\sqrt{n-3}} = 1.853 - 1.645\frac{1}{\sqrt{26-3}} = 1.510.$$

Hence the corresponding lower confidence bound on ρ is

$$\frac{e^{2\times 1.510}-1}{e^{2\times 1.510}+1} = 0.907.$$

Since this lower confidence bound exceeds 0.90, we can reject H_0 at the 0.05 level and conclude that $\rho > 0.90$. So the non-invasive method is acceptable. ■

2.4 Technical Notes

Many of the following derivations can be obtained as special cases of the corresponding derivations for multiple regression in Chapter 3 using matrix methods. Nevertheless, it is useful to give these special cases here without using matrix methods.

2.4.1 Derivation of the LS Estimators

To minimize the LS criterion Q with respect to β_0 and β_1, we take the first partial derivatives of Q and set them equal to 0 resulting in the following **normal equations**:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)] = 0 \implies \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.29}$$

and

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0 \implies \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{2.30}$$

An easy method to solve these equations is to rewrite the simple linear regression model (2.12) by centering the x_i's as

$$y_i = (\beta_0 + \beta_1 \overline{x}) + \beta_1(x_i - \overline{x}) + \varepsilon_i = \beta_0' + \beta_1' x_i' + \varepsilon_i \ (i = 1, \ldots, n),$$

where

$$\beta_0' = \beta_0 + \beta_1 \overline{x}, \beta_1' = \beta_1 \text{ and } x_i' = x_i - \overline{x}.$$

Then we can apply the above normal equations to estimate β_0' and β_1'. Note that the equations simplify since $\sum_{i=1}^{n} x_i' = \sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Thus from (2.29) we get

$$\beta_0' n = \sum_{i=1}^{n} y_i \implies \widehat{\beta}_0' = \overline{y} \implies \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Similarly, from (2.30) we get

$$\widehat{\beta}_1' \sum_{i=1}^{n} x_i'^2 = \sum_{i=1}^{n}(x_i - \overline{x}) y_i \implies \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

In the above we have used the fact that $\sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ since $\overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. These are the formulae (2.2) for $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

2.4.2 Sums of Squares

First note that SST $= \sum_{i=1}^{n}(y_i - \overline{y})^2 = S_{yy}$. Next note that

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}[y_i - \overline{y} - \widehat{\beta}_1(x_i - \overline{x})]^2 \quad \text{(by putting } \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}) \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= S_{yy} - 2\widehat{\beta}_1 S_{xy} + \widehat{\beta}_1^2 S_{xx} \\
&= S_{yy} - \widehat{\beta}_1^2 S_{xx},
\end{aligned}$$

where we have used the formula $\widehat{\beta}_1 = S_{xy}/S_{xx}$. Finally,

$$\text{SSR} = \text{SST} - \text{SSE} = S_{yy} - S_{yy} + \widehat{\beta}_1^2 S_{xx} = \widehat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.$$

2.4.3 Distribution of the LS Estimators

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear functions of the y_i's and hence are normally distributed. Using the fact that the x_i's are nonrandom, the expected values and variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ can be calculated as follows. First note that $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{where} \quad c_i = \frac{x_i - \overline{x}}{S_{xx}} \quad (1 \le i \le n).$$

The c_i's satisfy

$$\sum_{i=1}^{n} c_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{S_{xx}^2} = \frac{1}{S_{xx}}.$$

Hence

$$\sum_{i=1}^{n} c_i x_i = \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \overline{x})x_i = \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) = \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \overline{x})^2 = 1.$$

It follows that

$$E(\widehat{\beta}_1) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1.$$

Substituting this result in the formula for $\widehat{\beta}_0$ we get

$$E(\widehat{\beta}_0) = E(\overline{y} - \widehat{\beta}_1 \overline{x}) = E(\overline{y}) - E(\widehat{\beta}_1)\overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

Next,

$$\text{Var}(\widehat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \text{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}}.$$

Finally, using a not so difficult to prove result (see Exercise 2.5) that $\overline{y}$ and $\widehat{\beta}_1$ are independent, we can write

$$\text{Var}(\widehat{\beta}_0) = \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{x}^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}}.$$

Therefore it follows that

$$\frac{\widehat{\beta}_0 - \beta_0}{\sigma \sqrt{\sum_{i=1}^{n} x_i^2 / n S_{xx}}} \sim N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma \sqrt{1/S_{xx}}} \sim N(0,1).$$

Replacing σ by s results in t-distributions with $n-2$ d.f. for each.

2.4.4 Prediction Interval

Note that $\widehat{y}^* = \widehat{\mu}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$ is normally distributed since it is a linear combination of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, which are normally distributed. Furthermore

$$E(\widehat{y}^*) = E(\widehat{\mu}^*) = \beta_0 + \beta_1 x^* = \mu^*.$$

Next, again using the fact that $\overline{y}$ and $\widehat{\beta}_1$ are independent, we can write

$$\begin{aligned} \text{Var}(\widehat{y}^*) = \text{Var}(\widehat{\mu}^*) &= \text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \\ &= \text{Var}[\overline{y} + \widehat{\beta}_1 (x^* - \overline{x})] \\ &= \text{Var}(\overline{y}) + (x^* - \overline{x})^2 \text{Var}(\widehat{\beta}_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right]. \end{aligned}$$

Therefore,

$$\frac{\widehat{\mu}^* - \mu^*}{\sigma \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} \sim N(0,1).$$

Replacing σ by s results in a t_{n-2} random variable from which the CI (2.19) for μ^* follows.

To obtain the PI (2.20) for y^* we note that
$E(\widehat{y}^*-y^*) = \mu^*-\mu^* = 0$ and $\text{Var}(\widehat{y}^*-y^*) = \text{Var}(\widehat{y}^*)+\text{Var}(y^*) = \sigma^2\left[\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right]+\sigma^2$.
Therefore,

$$\frac{\widehat{y}^*-y^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}}} \sim N(0,1),$$

and the rest follows as before.

EXERCISES

Theoretical Exercises

2.1 (Regression through origin) In some applications the LS line is required to pass through the origin, e.g., fuel consumption as a function of the weight of the car. So we can assume the intercept to be zero. Show that the LS estimator of the slope β obtained by minimizing the LS criterion $Q = \sum_{i=1}^n (y_i - \beta x_i)^2$ equals

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

2.2 (Properties of residuals) Show that the residuals satisfy the two constraints (2.8). Further show that in the case of regression through the origin, the first constraint $\sum_{i=1}^n e_i = 0$ is not necessarily satisfied.

2.3 (Weighted least squares) Suppose that the observations y_i have different precisions and so we would like to weight them using different weights $w_i > 0$. For example, each y_i may be a sample mean of n_i i.i.d. observations so that their variances are inversely proportional to the n_i and hence we use the n_i as the weights. Show that the **weighted least squares (WLS)** estimator of the slope β for regression through the origin obtained by minimizing the LS criterion $Q = \sum_{i=1}^n w_i(y_i - \beta x_i)^2$ equals

$$\widehat{\beta} = \frac{\sum_{i=1}^n w_i x_i y_i}{\sum_{i=1}^n w_i x_i^2}.$$

2.4 (Omitted variables) Suppose that there are two predictor variables, x_1 and x_2, but we fit the straight line model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ omitting x_2. If, in fact, the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, show that

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \sum_{i=1}^n c_{i1} x_{i2} = \beta_1 + \beta_2 r_{12}\frac{s_2}{s_1},$$

where $c_{i1} = (x_{i1} - \overline{x}_1)/S_{11}$, $S_{11} = \sum_{i=1}^n (x_{i1} - \overline{x}_1)^2$, r_{12} is the sample correlation coefficient between x_1 and x_2 and s_1, s_2 are the sample SD's of x_1, x_2, respectively. Thus $\widehat{\beta}_1$ is biased with the bias given by the second term in the above expression. Under what condition is this bias zero?

2.5 (Independence between $\overline{y}$ and $\widehat{\beta}_1$) Show that $\overline{y}$ and $\widehat{\beta}_1$ are independent by carrying out the following steps.

a) Show that

$$\text{Cov}(\overline{y}, \widehat{\beta}_1) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \overline{x})\text{Cov}(\overline{y}, y_i) = 0.$$

by using the fact that $\text{Cov}(\overline{y}, y_i) = \sigma^2/n$.

b) Argue that since $\overline{y}$ and $\widehat{\beta}_1$ are jointly normally distributed, their independence follows because they are uncorrelated.

2.6 (Calibration problem) Suppose that in the tire wear problem mentioned in Section 2.2.5, the tire manufacturer wants to estimate the mileage for a given mean tread wear, e.g., the mileage which corresponds to the minimum acceptable tread depth. Thus we are given μ^* and we want to estimate the corresponding x^*.

a) Give a natural point estimate $\widehat{x}^*$ of x^*.

b) Construct a CI for x^* using the **Fieller's method** by following the steps below. Define a random variable

$$t = \frac{\widehat{\beta}_0 + \widehat{\beta}_1 x^* - \mu^*}{s\sqrt{1/n + (x^* - \overline{x})^2/S_{xx}}},$$

which is t_{n-2} distributed. Simplify the inequality $t^2 \leq t^2_{\nu,\alpha/2} = f_{1,\nu,\alpha}$, to show that a $(1-\alpha)$-level CI for x^* is given by the two roots of the quadratic equation:

$$Ax^{*2} + Bx^* + C = 0,$$

where

$$A = \widehat{\beta}_1^2 - \frac{u^2}{S_{xx}},\ B = 2\left(\frac{u^2\overline{x}}{S_{xx}} - \widehat{\beta}_1^2\widehat{x}^*\right)\ C = (\widehat{\beta}_1\widehat{x}^*)^2 - u^2\left(\frac{1}{n} + \frac{2\overline{x}^2}{S_{xx}}\right),$$

and $u = st_{n-2,\alpha/2}$.

c) For the cardiac output data in Table 2.2 calculate $\widehat{x}^*$ and a 95% CI for x^* if $\mu^* = 6$ litres/min. Use any values needed for this calculation from Example 2.10.

2.7 (t-Statistic for testing $\rho = 0$) Show that the t-statistic (2.25) for testing $\rho = 0$ is algebraically identical to the t-statistic (2.15) for testing $\beta_1 = 0$.

Applied Exercises

2.8 (Regression to the mean) Refer to Example 2.3. For the tall and short father considered in that example, calculate the expected heights of their sons for two different values of the correlation between the fathers' heights and sons' heights: $r = 0.25$ and $r = 0.75$. What do you conclude?

2.9 (Beta coefficients of stocks) The β of a stock is a coefficient that describes how the return on that stock is related to the return on a diversified stock portfolio. It is the slope coefficient in the simple linear regression model, $y = \alpha + \beta x$, where y is the return on that stock and x is the return on a benchmark stock market index representing a diversified portfolio. In this exercise we want to compare the β's of IBM and Apple versus S&P 500.

The file `IBM-Apple-SP500 RR Data.csv` contains data on percentage monthly rates of return (adjusted for dividends and stock splits) from February 2005 until September 2013 for IBM, Apple and S&P 500. These rates were calculated by downloading historical monthly prices from the Yahoo Finance website (`http://finance.yahoo.com/`). (If you prefer, you can download the more current data and use that to do the problem.)

a) Make scatter plots of rates of return of IBM versus S&P 500 and Apple versus S&P 500 and comment on them.

b) Calculate the β's for IBM and Apple versus S&P 500. Comment on the relative magnitudes of the β's. Which stock had a higher expected return relative to S&P 500?

c) Calculate the sample standard deviations (SD's) of rates of return for S&P 500, IBM and Apple. Also calculate the correlation matrix. Check that $\widehat{\beta} = rs_y/s_x$

for each stock where r is the correlation coefficient between S&P 500 and the given stock, s_x is the sample SD of S&P 500 and s_y is the sample SD of the given stock.

d) Explain based on the statistics calculated how a higher expected return is accompanied by higher volatility of the stock relative to S&P 500.

2.10 (Price elasticities of steaks) Data file `steakprices.csv` gives time series data on the prices and quantities sold of three types of beef steaks, chuck, porterhouse and rib eye,
(`http://www.aabri.com/manuscripts/08118.pdf`).

a) Estimate the price elasticities of all three steaks. Given that chuck is the least expensive cut and rib eye is the most expensive cut of beef, are the price elasticities of the three cuts in the expected order?

b) Estimate how much the demand will change if the price is increased by 10% for each cut.

2.11 (Smoking versus cancer) Data file `smoking-cancer.csv` contains data from 43 states and Washington, D.C. on the average number of cigarettes smoked (hundreds/capita) and number of deaths per 100,000 population due to four types of cancer.

a) Make scatter plots of the number of deaths due to each type of cancer versus cigarettes smoked to see what types of relationships (linear, nonlinear) exist and if there are any outliers.

b) Perform tests on the correlations to see which type of cancer deaths are most significantly correlated with cigarette smoking.

2.12 (Spearman rank correlation coefficient) The Pearson correlation coefficient measures only the extent of linear association between x and y; it does not measure the degree of monotone (increasing or decreasing) nonlinear association. Spearman's rank correlation coefficient (r_S) measures monotone association. It is simply the Pearson correlation coefficient between the ranks assigned to the original data. Let $u_i = \text{rank}(x_i)$ and $v_i = \text{rank}(y_i)$ ($i = 1, \ldots, n$), where average ranks are assigned to tied observations. Then $r_S = r_{uv}$. Compute r_S for the cardiac output data in Table 2.2 and compare it with the Pearson correlation coefficient calculated for the same data in Example 2.12.

2.13 (Lack of fit test) Suppose that the data are collected at $m \geq 2$ distinct x-values, $x_1, \ldots, x_m$ and at least at one x_i there are $n_i \geq 2$ repeat (independent) observations, $y_{i1}, \ldots, y_{in_i}$ and the total number of observations is $\sum_{i=1}^{m} n_i = n$. From each group of repeat observations we can compute the **pure error sum of squares (SSPE)** as $\text{SSPE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$. The pure error d.f equal $\sum_{i=1}^{m} (n_i - 1) = n - m$. The pure error estimator is given by $s^2 = \text{MSPE} = \text{SSPE}/(n-m)$. This is an unbiased estimator of σ^2 regardless of which particular model is fitted. The total SSE can be partitioned into $\text{SSE} = \text{SSPE} + \text{SSLOF}$. The lack of fit d.f. is $= (n-2) - (n-m) = m - 2$ and so $\text{MSLOF} = \text{SSLOF}/(m-2)$. Then $F = \text{MSLOF}/\text{MSPE}$ can be used to perform an F-test of lack of fit. If $F > f_{m-2,n-m,\alpha}$, we conclude that there is a significant at lack of fit.

Test the lack of fit of a straight line to the data given in Table 2.6 on weight loss due to corrosion as a function of iron content in 90/10 Cu-Ni alloy specimens.

Table 2.6 Corrosion weight loss data

Iron Content (%)	Weight Loss	Iron Content (%)	Weight Loss
0.01	127.6	0.95	103.9
0.01	130.1	1.19	101.5
0.01	128.0	1.44	92.3
0.48	124.0	1.44	91.4
0.48	122.0	1.96	83.7
0.71	110.8	1.96	86.2
0.71	113.1		

Source: Draper and Smith (1998), p. 98, Exercise C.

CHAPTER 3

MULTIPLE LINEAR REGRESSION: BASICS

Multiple regression is the bread and butter of predictive analytics and serves as a foundation for more advanced techniques. It takes into account the effects of multiple predictors simultaneously. Running separate simple linear regressions on each predictor can give misleading results since each regression accounts only for the direct bivariate relationship between the response variable and that predictor. Thus a simple linear regression may show strong relationship but it may be because of other intervening variables. When those variables are included in the model via multiple regression, the apparently strong relationship may be shown to be actually very weak or even with an opposite sign.

To illustrate the various concepts and techniques associated with multiple regression we will use one small example with only two predictors and 40 data points and one relatively large example with more than a dozen predictors and over 800 data points. The first example involves both prediction and inference, while the second example involves mainly prediction.

EXAMPLE 3.1 (College GPA and Entrance Test Scores: Data)

College admission committees are faced with the task of who to admit from thousands of applicants. There are many considerations including their college entrance test scores, high school GPA and rank, essays, extracurricular activities, recommendation letters and so on. In this example we will look at just one measure of academic success in college, namely graduating GPA. We will consider two predictors of that success, namely, college entrance Verbal and Math test scores. Data on 40 students

are shown in Table 3.1 and are stored in the file `GPA.csv`. Some questions of interest are: what is the relative importance of each test score in predicting the GPA and what is a good prediction model for GPA in terms of these two predictors. ■

Table 3.1 Entrance test scores and graduating GPA

Verbal	Math	GPA	Verbal	Math	GPA	Verbal	Math	GPA
81	87	3.49	83	76	3.75	97	80	3.27
68	99	2.89	64	66	2.70	77	90	3.47
57	86	2.73	83	72	3.15	49	54	1.3
100	49	1.54	93	54	2.28	39	81	1.22
54	83	2.56	74	59	2.92	87	69	3.23
82	86	3.43	51	75	2.48	70	95	3.82
75	74	3.59	79	75	3.45	57	89	2.93
58	98	2.86	81	62	2.76	74	67	2.83
55	54	1.46	50	69	1.90	87	93	3.84
49	81	2.11	72	70	3.01	90	65	3.01
64	76	2.69	54	52	1.48	81	76	3.33
66	59	2.16	65	79	2.98	84	69	3.06
80	61	2.60	56	78	2.58			
100	85	3.30	98	67	2.73			

Source: McClave, J. T. and Dietrich, F. H. (1994), p. 811.

EXAMPLE 3.2 (Used Car Prices: Data)

This example deals with building a prediction model for used car prices from a number of predictors listed in Table 3.2. Retail prices of 2005 GM cars were calculated from the data provided in the 2005 Central Edition of the *Kelly Blue Book*. All cars in this data set were less than one year old and in excellent condition. Data are in file `usedcarprices.csv`. Questions of interest are which are the best predictors of the used car price and their relative contributions to the price. ■

3.1 Multiple Linear Regression Model

3.1.1 Model in Scalar Notation

Denote the response variable by y and the predictor variables by $x_1, \ldots, x_p$. Suppose that we have n complete data vectors $(x_{i1}, \ldots, x_{ip}, y_i), \quad (i = 1, \ldots, n)$ on these variables from which we want to fit the standard multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1, \ldots, n). \tag{3.1}$$

Here $\beta_0, \beta_1, \ldots, \beta_p$ are unknown **parameters**, β_0 being the **intercept** or the **constant term**, $\beta_1, \ldots, \beta_p$ being the **regression coefficients** and the ε_i being i.i.d. $N(0, \sigma^2)$ **ran-**

Table 3.2 Variables for the used car prices example

Variable	Description
Price	Suggested retail price (response variable)
Mileage	Odometer reading in thousands of miles
Make	Division of GM (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn)
Model	Specific model of a given make
Type	Body type (convertible, coupe, hatchback, sedan, wagon)
Cylinders	Number of cylinders
Liters	Size of engine
Doors	Number of doors
Cruise	Indicator variable for cruise control (1 = cruise control)
Sound	Indicator variable for upgraded speakers (1 = upgraded)
Leather	Indicator variable for leather seats (1 = leather)

Source: Kuiper, S. (2008).

dom errors. From this model it follows that the y_i are independent $N(\mu_i, \sigma^2)$ where

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (i = 1, \ldots, n). \tag{3.2}$$

Note the following points about this model.

- This is called a **linear model** because it is linear in the β's—not necessarily in the x's. Any nonlinear functions of the x's, e.g., x^2, $\log x$ or $x_1 x_2$, may be used as predictors and the model would still be linear. For example, a pth degree polynomial model in a single predictor x:

  $$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon,$$

  is still a linear model with $x_1 = x, x_2 = x^2, \ldots, x_p = x^p$. A product term such as $x_1 x_2$ is used to model interaction between x_1 and x_2 as we shall see in Section 3.6.2.

  The reason that linearity in the β_j's is critical is that it makes the equations for finding their **least squares (LS) estimators** linear, which makes them easy to solve with closed form solutions. Furthermore, these LS estimators are then linear functions of the responses y_i's which makes their sampling distributions simple and so inferences on them straightforward.

- Some nonlinear models can be transformed into linear models. We will see some examples in Section 3.2.5. An intrinsically **nonlinear model** is nonlinear in the β's and cannot be transformed into a linear form. An example from chemical process kinetics is the model

  $$y = \frac{\beta_1}{\beta_1 - \beta_2}\left[e^{-\beta_2 t} - e^{-\beta_1 t}\right] + \varepsilon,$$

  where y is the percent reaction completed, t is the reaction time, and β_1 and β_2 are reaction rate constants. No transformation can linearize this model and nonlinear regression techniques must be used to estimate the rate constants.

3.1.2 Model in Matrix Notation

The multiple regression model and the results associated with it (e.g., the LS estimators of the β's and their sampling distributions) can be written in a compact form and these results can be derived more easily and elegantly by using the matrix notation introduced below.[1]

Let

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Here $\boldsymbol{y}$ is the **response vector**, $\boldsymbol{X}$ is the **model matrix**, $\boldsymbol{\beta}$ is the **parameter vector** and $\boldsymbol{\varepsilon}$ is the **random error vector**. From (3.2) we can write $E(y_i) = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ where $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ is the ith row vector of $\boldsymbol{X}$. Hence it follows that $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and the model (3.1) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.3}$$

Another way of saying that the ε_i are i.i.d. $N(0, \sigma^2)$ is that the random vector $\boldsymbol{\varepsilon}$ has an n-variate normal distribution with mean vector $\mathbf{0}$ (the null vector of all 0's) and covariance matrix $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. The distribution of the response vector $\boldsymbol{y}$ is then the n-variate normal distribution with mean vector $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and covariance matrix $\sigma^2\boldsymbol{I}$.

3.2 Fitting a Multiple Regression Model

3.2.1 Least Squares (LS) Method

We extend the LS method for simple linear regression to multiple regression by minimizing the **LS criterion**:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.4}$$

w.r.t. $\beta_0, \beta_1, \ldots, \beta_p$. We set the partial derivatives of Q w.r.t. the β_j equal to zero resulting in $p + 1$ simultaneous linear equations (called the **normal equations**):

$$\frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})] = 0 \quad (j = 0, \ldots, p), \tag{3.5}$$

where we have set $x_{i0} := 1$. We solve these equations for $\beta_0, \beta_1, \ldots, \beta_p$. Unique solutions to these equations exist under certain conditions on the $\boldsymbol{X}$ matrix (discussed below). The resulting LS estimators can be more easily expressed in a closed form by using the matrix notation.

The normal equations (3.5) can be written compactly as a single equation (see Section 3.7 for a derivation):

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{3.6}$$

[1] All vectors are assumed to be column vectors. A prime on a vector or a matrix denotes its transpose. Vectors and matrices are denoted by bold letters. Usually, we suppress the dimension of any vector or a matrix if it is clear from the context. When necessary, we indicate the dimension by a subscript.

If the inverse of $\boldsymbol{X}'\boldsymbol{X}$ exists then this equation has a unique solution $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{3.7}$$

which is the LS estimator of the parameter vector $\boldsymbol{\beta}$. It can be shown that the $\widehat{\beta}_j$'s satisfy the equation

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{x}_1 + \cdots + \widehat{\beta}_p\overline{x}_p. \tag{3.8}$$

In other words, the fitted regression plane passes through the centroid $(\overline{x}_1, \ldots, \overline{x}_p; \overline{y})$ of the data.

From linear algebra it is known that the inverse of $\boldsymbol{X}'\boldsymbol{X}$ exists if and only if the columns of $\boldsymbol{X}$ are linearly independent. These are the data vectors of the predictor variables. They are linearly independent if there are no linear relationships among the predictors. Otherwise one or more of them can be expressed in terms of the others, and so we can reduce them to an independent set by eliminating the extra ones. For example, suppose that a data set has three predictors: income, expenditure and saving. Since saving equals income minus expenditure, we don't need to keep all three; any two will suffice. When there are approximate or exact linear dependencies among the columns of $\boldsymbol{X}$, difficulties arise in the computation of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and hence of $\widehat{\boldsymbol{\beta}}$. This is called the **multicollinearity problem**, which we shall discuss in Chapter 4. From now on we assume that the columns of $\boldsymbol{X}$ are linearly independent, so $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists and $\widehat{\boldsymbol{\beta}}$ is unique.

We may ask "Are the LS estimators optimal in some sense?" **Gauss-Markov theorem** stated and proved in Section 3.7 provides an answer to this question.

EXAMPLE 3.3 (GPA Data: Multiple Regression)

Before running a regression of GPA on Verbal and Math scores, it is important to study relationships among them by making scatter plots and computing the correlation matrix. Scatter plots between all three of them can be combined into a single composite plot called the **matrix scatter plot** shown in Figure 3.1. This plot and the correlation matrix below are obtained by using the following R commands.

```
> gpa = read.csv("c:/data/GPA.csv")
> plot(gpa)
> cor(gpa)
```

We can see from this plot that Verbal and Math scores are fairly uncorrelated with each other while both are moderately correlated with GPA. Thus the two predictors make relatively independent and roughly equal contributions to the GPA and it makes sense to fit a linear model for GPA that includes both of them.

The above visual impressions based on the matrix scatter plot are confirmed by the correlation matrix below.

$$\begin{array}{c} \\ \text{Verbal} \\ \text{Math} \\ \text{GPA} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Verbal} & \text{Math} & \text{GPA} \end{array} \\ \begin{bmatrix} 1 & -0.107 & 0.529 \\ -0.107 & 1 & 0.573 \\ 0.529 & 0.573 & 1 \end{bmatrix} \end{array}.$$

Using the `lm` function in R we get the fitted equation given in the following output.

```
> lmfit = lm(GPA ~ Verbal + Math, data = gpa)
> summary(lmfit)

Call:
```

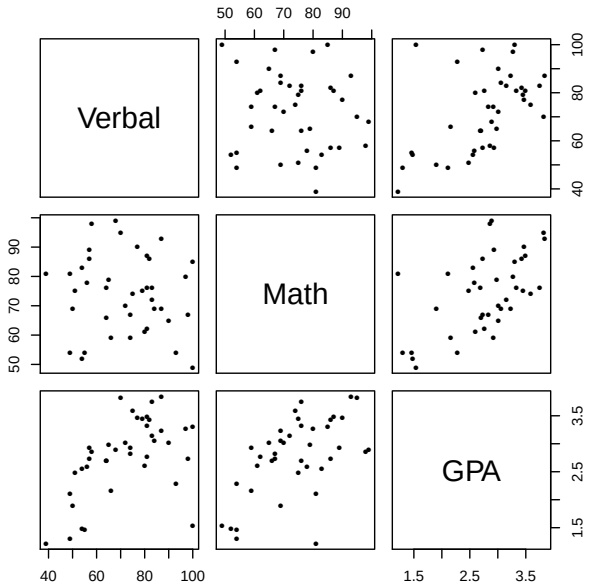


Figure 3.1 Matrix scatter plot between GPA, Verbal score and Math score

```
lm(formula = GPA ~ Verbal + Math, data = gpa)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.570537   0.493749  -3.181  0.00297 **
Verbal       0.025732   0.004024   6.395 1.83e-07 ***
Math         0.033615   0.004928   6.822 4.90e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4023 on 37 degrees of freedom
Multiple R-squared:  0.6811,    Adjusted R-squared:  0.6638
F-statistic: 39.51 on 2 and 37 DF,  p-value: 6.585e-10
```

Thus the regression equation fitted using R is

$$\widehat{\text{GPA}} = -1.5705 + 0.0257\text{Verbal} + 0.0336\text{Math}.$$

Next we illustrate fitting of this equation using (3.7). The $\boldsymbol{X}$ matrix and the $\boldsymbol{y}$ vector have 40 rows corresponding to 40 observations:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 81 & 87 \\ \vdots & \vdots & \vdots \\ 1 & 84 & 69 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} 3.49 \\ \vdots \\ 3.06 \end{bmatrix}$$

from which we can compute

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 40 & 2884 & 2960 \\ 2884 & 218048 & 212533 \\ 2960 & 212533 & 225782 \end{bmatrix}$$

and

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 1 & \cdots & 1 \\ 81 & \cdots & 84 \\ 87 & \cdots & 69 \end{bmatrix} \begin{bmatrix} 3.49 \\ \vdots \\ 3.06 \end{bmatrix} = \begin{bmatrix} 110.89 \\ 8225.68 \\ 8409.77 \end{bmatrix}.$$

Next we compute

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \begin{bmatrix} 1.506 & -8.182 \times 10^{-3} & -1.205 \times 10^{-2} \\ -8.182 \times 10^{-3} & 1.000 \times 10^{-4} & 1.310 \times 10^{-5} \\ -1.205 \times 10^{-2} & 1.310 \times 10^{-5} & 1.500 \times 10^{-4} \end{bmatrix}.$$

Finally,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{bmatrix} 1.506 & -8.182 \times 10^{-3} & -1.205 \times 10^{-2} \\ -8.182 \times 10^{-3} & 1.000 \times 10^{-4} & 1.310 \times 10^{-5} \\ -1.205 \times 10^{-2} & 1.310 \times 10^{-5} & 1.500 \times 10^{-4} \end{bmatrix} \begin{bmatrix} 110.89 \\ 8225.68 \\ 8409.77 \end{bmatrix} \\ &= \begin{bmatrix} -1.5705 \\ 0.0257 \\ 0.0336 \end{bmatrix}. \end{aligned}$$

Thus $\widehat{\beta}_0 = -1.5705, \widehat{\beta}_1 = 0.0257$ and $\widehat{\beta}_2 = 0.0336$. ■

3.2.2 Interpretation of Regression Coefficients

The following points are important to keep in mind when interpreting the regression coefficients.

1. In general, the $\widehat{\beta}_j$'s depend on which other predictors are included in the model. If the other variables in the model change then the $\widehat{\beta}_j$'s change, too. This is because the $\widehat{\beta}_j$'s measure the marginal contributions of the x_j's to y conditional on other variables in the model.

2. A variable can have an apparently large effect in presence of some variables but not in presence of other variables. Also, the signs of the $\widehat{\beta}_j$'s may change and thus the apparent direction of the effect of x_j on y may change if other variables in the model change. The only exception to this is when the columns of the $\boldsymbol{X}$ matrix are mutually orthogonal so that $\boldsymbol{X}'\boldsymbol{X}$ and its inverse are diagonal matrices. In this case, adding or deleting the variables from the model does not change the regression coefficients of the other variables; thus their contributions are independent of each other. This case typically occurs only when the so-called **orthogonal designs** are used; see Exercise 3.7.

3. The $\widehat{\beta}_j$'s have units, namely the units of y divided by the units of the x_j's. Therefore the magnitudes of the $\widehat{\beta}_j$'s depend on the units of the x_j's and cannot be directly compared to each other. For example, suppose we fit a model for the gas mileage (mpg) of a car on two variables: engine size (liters) and weight (lb). Then the coefficients have

the units of mpg/liter and mpg/lb, respectively. Comparing them is like comparing apples and oranges.

4. To address the problem of units, **standardized regression coefficients** may be used by fitting a regression model to standardized x's and y; see Section 3.6.3. Although the resulting coefficients are unitless, they inherit some of the same issues as the unstandardized regression coefficients. For instance, they also depend on other predictors included in the model.

5. Finally, neither the unstandardized nor the standardized regression coefficients take into account the costs associated with changing the x's. For example, if a marketing executive is trying to decide the numbers of TV ads and print media ads for some product, the regression coefficients quantifying their effects on the sales of the product are not sufficient to make the decision. The unit costs of different types of ads must be taken into account.

3.2.3 Fitted Values and Residuals

Fitted values and residuals are useful to assess the goodness of the LS fit of the model to the observed data as well as for model diagnostics. This latter use will be discussed in Chapter 4. Here we focus on their use for assessing the goodness of fit.

The **fitted values** are defined as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip} = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} \quad (i = 1, \ldots, n),$$

and the **fitted values vector** is defined as

$$\widehat{\boldsymbol{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' \widehat{\boldsymbol{\beta}} \\ \boldsymbol{x}_2' \widehat{\boldsymbol{\beta}} \\ \vdots \\ \boldsymbol{x}_n' \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \widehat{\boldsymbol{\beta}} = \boldsymbol{X} \widehat{\boldsymbol{\beta}}.$$

Substituting $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ from (3.7) in the above we get

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{3.9}$$

where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{3.10}$$

is called the **hat matrix**.

Next we define the **residual vector** as the vector of differences between the observed y_i's and the fitted $\widehat{y}_i$'s:

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \widehat{y}_1 \\ y_2 - \widehat{y}_2 \\ \vdots \\ y_n - \widehat{y}_n \end{bmatrix} = \boldsymbol{y} - \widehat{\boldsymbol{y}}.$$

Substituting for $\widehat{\boldsymbol{y}}$ from (3.9) in the above we get

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \tag{3.11}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Thus both $\widehat{\boldsymbol{y}}$ and $\boldsymbol{e}$ are linear transforms of the observed vector $\boldsymbol{y}$.

It can be checked that $\boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})$ are symmetric and satisfy $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H}$. Such matrices are called **projection matrices**. The $\boldsymbol{H}$ matrix projects $\boldsymbol{y}$ into $\widehat{\boldsymbol{y}}$ and the $\boldsymbol{I} - \boldsymbol{H}$ matrix projects $\boldsymbol{y}$ into $\boldsymbol{e}$. These two projections

are orthogonal to each other since $\boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H})=\boldsymbol{O}$, where $\boldsymbol{O}$ is the null matrix and so $\widehat{\boldsymbol{y}}'\boldsymbol{e}=\boldsymbol{y}'\boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}=0$. We will see applications of this in the sequel.

The n residuals are subject to $p+1$ linear constraints since it can be shown that each of the $p+1$ column vectors of $\boldsymbol{X}$ is orthogonal to $\boldsymbol{e}$ (see Section 3.7). This implies that if the constant term β_0 is included in the model then the residuals sum to zero since $\boldsymbol{e}$ is orthogonal to the first column of all 1's of $\boldsymbol{X}$ and so $\sum_{i=1}^{n} e_i=0$. Because of these linear constraints only $n-(p+1)$ of the residuals are linearly independent; the others can be deduced from them.

The residuals are used to estimate the error variance σ^2 by

$$s^2=\frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}=\frac{\text{SSE}}{n-(p+1)}=\text{MSE}, \tag{3.12}$$

where $\text{SSE}=\sum_{i=1}^{n} e_i^2$ is the **error sum of squares**, $\text{MSE}=\text{SSE}/[n-(p+1)]$ is the **mean square error** and $n-(p+1)$ is the **error degrees of freedom (d.f.)**. It can be shown that s^2 is an unbiased estimate of σ^2, i.e., $E(s^2)=\sigma^2$. So σ is estimated by $s=\sqrt{\text{MSE}}$, The **root mean square error (RMSE)**, $s=\sqrt{\text{MSE}}$, is used to estimate σ, but it is not an unbiased estimate.

3.2.4 Measures of Goodness of Fit

For measuring the goodness of the LS fit, we follow the same approach as in the case of simple linear regression. Define $\overline{\boldsymbol{y}}$ as a vector of dimension n all of whose entries are $\overline{y}$. Then we can write (2.9) in vector notation by expressing the deviation of the observation vector $\boldsymbol{y}$ from $\overline{\boldsymbol{y}}$ as the sum of two vectors:

$$\boldsymbol{y}-\overline{\boldsymbol{y}}=(\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}})+(\boldsymbol{y}-\widehat{\boldsymbol{y}})=(\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}})+\boldsymbol{e}.$$

It can be shown that $(\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}})$ and $\boldsymbol{e}$ are orthogonal vectors (see Section 3.7). Therefore from the Pythagoras theorem we get the ANOVA identity:

$$\underbrace{||\boldsymbol{y}-\overline{\boldsymbol{y}}||^2}_{\text{SST}}=\underbrace{||\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}}||^2}_{\text{SSR}}+\underbrace{||\boldsymbol{e}||^2}_{\text{SSE}}, \tag{3.13}$$

where, e.g., $||\boldsymbol{e}||^2=\boldsymbol{e}'\boldsymbol{e}=\sum_{i=1}^{n} e_i^2$ is the squared length (norm) of the vector $\boldsymbol{e}$. Here, as in Chapter 2, $\text{SST}=||\boldsymbol{y}-\overline{\boldsymbol{y}}||^2=\sum_{i=1}^{n}(y_i-\overline{y})^2$ is the **total sum of squares (SST)** and $\text{SSR}=||\widehat{\boldsymbol{y}}-\overline{\boldsymbol{y}}||^2=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$ is the **regression sum of squares (SSR)**. Note that SSE is just the minimum value of the LS criterion Q in (3.4).

The proportion of variation in y accounted for by its regression on the x's is given by

$$R^2=\frac{\text{SSR}}{\text{SST}}=1-\frac{\text{SSE}}{\text{SST}}, \tag{3.14}$$

called the **multiple coefficient of determination**. Its positive square root R $(0\le R\le 1)$ is called the **multiple correlation coefficient**. Since R measures linear association between y and multiple x's, it cannot be assigned a sign as in the case of the bivariate correlation coefficient r. By convention, we assign a positive sign to R. It can be shown that R is the Pearson correlation coefficient between the observed y_i's and the fitted $\widehat{y}_i$'s.

It should be pointed out that a high R^2 does not necessarily mean a better predictive model. To see this, observe that SSE can only decrease and hence R^2 can only increase when more predictors are added to the model, whether they are related to the response variable or not. The goal of a model is not to fit the data as closely as possible as it may result in **overfitting**. The goal is to capture the overall trend in the data with an as simple model as possible. Such a model is useful for prediction.

To avoid this drawback of R^2, sometimes **adjusted** R^2 is used as a measure of the goodness of fit of the regression model. It is defined as

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/[n-(p+1)]}{\text{SST}/(n-1)}. \tag{3.15}$$

Note that $\text{SST}/(n-1)$ does not depend on p while both SSE and $n-(p+1)$ decrease with p. So their ratio, which is the MSE, may increase or decrease as p increases. Typically R^2_{adj} initially increases with p but then decreases as additional predictors are added to the model as their marginal contributions diminish in magnitude. It can be shown (see Exercise 3.2) that $R^2_{\text{adj}} \leq R^2$ and so R^2_{adj} may become negative if R^2 is close to zero. Also, R^2_{adj} does not have a simple interpretation like that of R^2 as the proportion of variation in y explained by the fitted model.

3.2.5 Linearizing Transformations

As we saw in Chapter 2, many nonlinear models can be transformed to linear models. In addition to the power and exponential laws mentioned there, **multiplicative laws** can also be linearized and made additive by using the log-transformation.

The **Cobb-Douglas production function** in economics used to model the output of a firm (y) as a function of the capital input (x_1) and labor input (x_2) provides an example of a multiplicative law. This function has the form $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \varepsilon$ where $\beta_0, \beta_1, \beta_2$ are unknown parameters to be estimated and ε is a multiplicative random error term. By making the log-transformation, we get the linear model:

$$\ln y = \ln \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ln \varepsilon.$$

Similar to the price elasticity of demand defined in Section 2.1.2,

$$\beta_1 = \frac{(\partial y/y)}{(\partial x_1/x_1)} \quad \text{and} \quad \beta_2 = \frac{(\partial y/y)}{(\partial x_2/x_2)}$$

are called the capital and labor elasticities of output, respectively. Thus β_1 and β_2 give relative changes in the output due to unit relative changes in the corresponding inputs.

Log-transformation faces a difficulty when some responses are zero or negative. A standard practice in such cases is to add a common positive constant to all responses so that they all become positive and then take their logs. In database marketing models for predicting customer sales a large majority of the customers who are offered a sales incentive (e.g., discount coupon) do not respond and so their purchase amounts are zero. On the other hand, the distribution of the purchases is highly right-skewed with some very large purchases. For such kinds of data we can add 1 to all the purchase amounts. After log-transformation all zero purchase amounts are transformed back to zeros.

3.3 Statistical Inferences for Multiple Regression

3.3.1 Analysis of Variance for Multiple Regression

There are two sources of variation in y, namely the variation caused by the x's through their linear relationship with y and the random error or residual variation. The ANOVA identity in (3.13) gives this decomposition. Unless the column vectors of $\boldsymbol{X}$ are mutually orthogonal (i.e., the design is orthogonal), the contributions of the predictors to the variation in y are not independent of each other and cannot be partitioned into additive components, i.e., SSR cannot be expressed as the sum of the SS's due to the individual x_j's.

Table 3.3 ANOVA table for multiple linear regression

Source	SS	d.f.	MS	F
Regression	SSR	p	MSR	$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n-(p+1)$	MSE	
Total	SST	$n-1$		

As in the case of simple linear regression, we can partition the total d.f. as follows:

$$\underbrace{n-1}_{\text{Total d.f.}} = \underbrace{p}_{\text{Regression d.f.}} + \underbrace{[n-(p+1)]}_{\text{Error d.f.}}.$$

Regression d.f. equals the number of predictor variables in the linear model ($p = 1$ for simple linear regression). The error d.f. equals $n-(p+1)$ because the n residuals are subject to $p+1$ linear constraints as noted in Section 3.2.3.

The error d.f. $= n-(p+1)$ implies that n must be greater than $p+1$, i.e., the number of observations must be greater than the number of the unknown β's to be estimated, in order for any d.f. to be available for estimating the error variance σ^2. If $n = p+1$ then the error d.f. $= 0$ and SSE $= 0$ since we obtain a perfect fit, i.e., all fitted $\widehat{y}_i$'s equal to the observed y_i's, and so all residuals e_i's equal to 0. Such a model is called a **saturated model**. For example, for bivariate data $\{(x_i, y_i),\ i = 1, \ldots, n\}$ with no repeat observations on any x_i, we can fit an $(n-1)$th degree polynomial in x which will exactly pass through all the n points giving a perfect fit with SSE $= 0$ and hence $R^2 = 100\%$, but we can not estimate σ^2.

How large must n be relative to p in order to have sufficient error d.f. available for estimating σ^2 accurately? Opinions vary on this question, but a rough rule of thumb is that the error d.f. should be at least 10 and ideally 30 or more.

A sum of squares divided by its d.f. is called a **mean square (MS)**. Thus, **mean square regression (MSR)** equals SSR$/p$ and **mean square error (MSE)** equals SSE$/[n-(p+1)]$, which is used to estimate σ^2 as defined in (3.12). It can be shown that under the null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$, the ratio $F = \text{MSR}/\text{MSE}$ has an F-distribution with p and $n-(p+1)$ d.f. Thus an α-level test rejects H_0 if

$$F = \frac{\text{MSR}}{\text{MSE}} > f_{p,n-(p+1),\alpha}, \tag{3.16}$$

where $f_{p,n-(p+1),\alpha}$ is the upper α critical point of this F-distribution. These calculations are presented in the form of an **ANOVA** in Table 3.3.

Why do we test the overall null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$? Shouldn't we be testing the significance of the individual $\widehat{\beta}_j$'s directly? The problem with testing individual $\widehat{\beta}_j$'s directly without the overall F-test is that when there are many regression coefficients, some of them may turn out to be significant simply by random chance. For example, if 100 estimated regression coefficients are tested individually at 5% significance level then on the average five turn out to be significant when all the true β_j's are zero. The probability that at least one $\widehat{\beta}_j$ will turn out significant is almost 1. **Multiple testing procedures** are designed to control the probability of occurrence of false positives when testing multiple hypotheses. The book by Hochberg & Tamhane (1987) covers this area in detail.

Table 3.4 ANOVA table for linear regression of GPA on entrance test scores

Source	SS	d.f.	MS	F	P
Regression	12.7859	2	6.3930	39.51	0.000
Error	5.9876	37	0.1618		
Total	18.7735	39			

EXAMPLE 3.4 (GPA Data: ANOVA for Linear Regression)

The ANOVA table for the regression of GPA versus Verbal and Math test scores from Example 3.3 is shown in Table 3.4. We see that the F-statistic equals 39.51 with 2 and 37 d.f. which is highly significant with a P-value < 0.001. Also, $R^2 = \text{SSR}/\text{SST} = 12.7859/18.7735 = 0.681$. Thus about 2/3rd of the variation in GPA is accounted for by linear regression on Verbal and Math test scores. The estimated standard deviation is $s = \sqrt{0.1618} = 0.4023$. We will use this estimate in testing hypotheses and making confidence intervals on the β_j's.

■

3.3.2 Inferences on Regression Coefficients

The ANOVA F-test is a test of the global null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$. Having rejected this null hypothesis, we would like to know which individual β_j are different from zero or estimate their magnitudes via confidence intervals (CI's). These inferences are based on the sampling distributions of the $\widehat{\beta}_j$'s.

The LS estimator vector $\widehat{\boldsymbol{\beta}}$ has a multivariate normal distribution of dimension $p+1$ with mean vector $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and covariance matrix $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{V}$, where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Therefore the individual $\widehat{\beta}_j$ are normally distributed with means β_j and variances $\sigma^2 v_{jj}$ where v_{jj} is the jth diagonal entry ($0 \le j \le p$) of $\boldsymbol{V}$. Also it can be shown that $[n - (p + 1)]s^2/\sigma^2 = \text{SSE}/\sigma^2$ is distributed as χ^2 with $n - (p + 1)$ d.f. independent of $\widehat{\boldsymbol{\beta}}$.

From these results it follows that $\text{SE}(\widehat{\beta}_j) = s\sqrt{v_{jj}}$ and

$$\frac{\widehat{\beta}_j - \beta_j}{\text{SE}(\widehat{\beta}_j)} \sim t_{n-(p+1)} \quad (0 \le j \le p).$$

So a $100(1 - \alpha)\%$ CI for any β_j is given by

$$\widehat{\beta}_j \pm t_{n-(p+1),\alpha/2}\text{SE}(\widehat{\beta}_j) \quad (0 \le j \le p).$$

An α-level test of $H_{0j} : \beta_j = 0$ rejects if

$$|t_j| = \frac{|\widehat{\beta}_j|}{\text{SE}(\widehat{\beta}_j)} > t_{n-(p+1),\alpha/2} \quad (0 \le j \le p). \tag{3.17}$$

EXAMPLE 3.5 (GPA Data: Inferences on Regression Coefficients)

The 95% CI's on the two coefficients can be calculated as (using $t_{37,0.025} = 2.0262$ and standard errors of the regression coefficients for Verbal and Math from the R

output in Example 3.3):

$$\text{Verbal: } 0.02573 \pm 2.0262 \times 0.00402 = [0.0176, 0.0339]$$

and

$$\text{Math: } 0.03362 \pm 2.0262 \times 0.00493 = [0.0236, 0.0436].$$

Note that 0 is well outside both the intervals. This is in accord with the result from Example 3.3 that both Verbal and Math are highly significant predictors of GPA. However, this only means that there are significant linear components of Verbal and Math in their relationship with GPA, but there could be quadratic components as well. In Examples 3.8 and 3.16 we will investigate if the quadratic and interaction effects improve the model. ■

3.3.3 Confidence Ellipsoid for $\boldsymbol{\beta}$

The previous section gave separate confidence intervals for the individual β_j's. As will be seen later in some applications, it is useful to have a simultaneous (or joint) $100(1-\alpha)\%$ **confidence region** for all β's. This region turns out to be an ellipsoid centered at $\widehat{\boldsymbol{\beta}}$, and is given by

$$\left\{ \boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})' \boldsymbol{X}' \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{(p+1)s^2} \leq f_{p+1, n-(p+1), \alpha} \right\}. \tag{3.18}$$

This region can be extended to any subset of the $\boldsymbol{\beta}$ vector or more generally to any set of $r \leq p+1$ linear parametric functions $\theta_i = c_{i0}\beta_0 + \cdots + c_{ip}\beta_p$ $(1 \leq i \leq r)$. Denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)' = \boldsymbol{C}\boldsymbol{\beta}$ where $\boldsymbol{C} = \{c_{ij}\}$ is an $r \times (p+1)$ matrix with linearly independent rows (so that the θ_i's are linearly independent parameters). Then the LS estimator of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}\widehat{\boldsymbol{\beta}}$ and $\text{Cov}(\widehat{\boldsymbol{\theta}}) = \sigma^2 \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'$ (using the sandwich formula (A.7)), where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The extension of the above confidence ellipsoid for $\boldsymbol{\theta}$ is

$$\left\{ \boldsymbol{\theta} : \frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' (\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}')^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{rs^2} \leq f_{r, n-(p+1), \alpha} \right\}. \tag{3.19}$$

Using this confidence ellipsoid we can reject $H_0 : \boldsymbol{\theta} = \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ at level α if $\boldsymbol{0}$ falls outside the ellipsoid, i.e., if

$$\frac{\widehat{\boldsymbol{\theta}}' (\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}')^{-1} \widehat{\boldsymbol{\theta}}}{rs^2} > f_{r, n-(p+1), \alpha}.$$

One applications of interest of this formula is for $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p)'$. In that case $r = p$, the $\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'$ matrix is simply the $p \times p$ submatrix of $\boldsymbol{V}$ obtained by deleting the first row and the first column of $\boldsymbol{V}$ corresponding to β_0. Then rejecting H_0 if the null vector $\boldsymbol{0}$ falls outside this confidence ellipsoid is equivalent to the α-level ANOVA F-test of the overall $H_0 : \beta_1 = \cdots = \beta_p = 0$. We will see more examples of this in the next section.

EXAMPLE 3.6 (GPA Data: Confidence Ellipsoid)

For the linear model fitted to the GPA data in Example 3.3 the following covariance matrix of the regression coefficients of $(\widehat{\beta}_1, \widehat{\beta}_2)$ can be obtained by using the command `vcov(lmfit)` in R:

$$s^2 \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' = 10^{-5} \begin{bmatrix} 1.619 & 0.212 \\ 0.212 & 2.428 \end{bmatrix}.$$

The inverse of this matrix equals

$$\frac{1}{s^2}(\boldsymbol{CVC}')^1 = 10^5 \begin{bmatrix} 0.625 & -0.055 \\ -0.055 & 0.417 \end{bmatrix}.$$

Using $\widehat{\beta}_1 = 0.0257$ and $\widehat{\beta}_2 = 0.0336$, the equation for the confidence ellipse is given by

$$\frac{1}{2}[\beta_1 - 0.0257, \beta_2 - 0.0336] \left(10^5\right) \begin{bmatrix} 0.625 & -0.055 \\ -0.055 & 0.417 \end{bmatrix} \begin{bmatrix} \beta_1 - 0.0257 \\ \beta_2 - 0.0336 \end{bmatrix}$$

$$= \quad 10^5 \left[0.312(\beta_1 - 0.0257)^2 - 0.055(\beta_1 - 0.0257)(\beta_2 - 0.0336) + 0.209(\beta_2 - 0.0336)^2\right].$$

To test $H_0 : \beta_1 = \beta_2 = 0$, substitute these values in the above equation to obtain the F-statistic as

$$10^5[0.312(0.0257)^2 - 0.055(0.0257)(.0336) + 0.209(0.0336)^2] = 39.44,$$

which agrees with the F-statistic value = 39.51 in the R output in Example 3.3 except for round-off errors. Since it exceeds $f_{2,37,.05} = 3.252$, which implies that the point $\beta_1 = \beta_2 = 0$ falls outside the 95% confidence ellipse for the coefficients β_1 and β_2 as shown in Figure 3.2, we can confidently reject H_0. ■

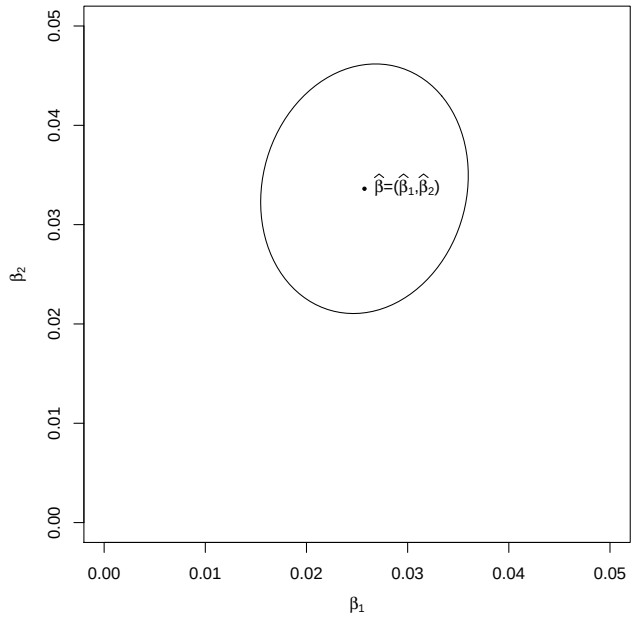


Figure 3.2 95% confidence ellipse for β_1 and β_2

3.3.4 Extra Sum of Squares Method

The **extra sum of squares (extra SS) method** is used when we want to test hypotheses on multiple β_j's simultaneously. One example of this is when we want to test whether a

partial or **reduced model** consisting of a subset of predictors fits the data not significantly worse than the **full model** with all predictors. In other words, whether adding the extra predictors improves the fit significantly. The partial model is said to be **nested** under the full model.

As an example, consider comparing a first-degree model versus a second-degree model in two variables, i.e.,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{vs.} \quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 \quad (3.20)$$

by testing $H_0 : \beta_3 = \beta_4 = 0$. The general problem of testing a partial model versus a full model can be stated as

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q \quad \text{vs.} \quad E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad (3.21)$$

where $q < p$. This problem is equivalent to testing $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$, which imposes $r = p - q$ linearly independent constraints on the β's. A special case of interest is testing the so-called **null model** (which has only the intercept term) versus the full model, i.e.,

$$E(y) = \beta_0 \quad \text{vs.} \quad E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (3.22)$$

by testing the **overall null hypothesis** $H_0 : \beta_1 = \cdots = \beta_p = 0$ against the alternative that at least one $\beta_j \neq 0$ $(1 \leq j \leq p)$.

Still another example is testing equalities among subsets of β's. For example, suppose that in a regression model for predicting annual credit card purchases, one of the variables is the age of the customer classified into three categories, young, middle-age or old. Let β_1, β_2 and β_3 be the regression coefficients for the three categories. (Actually, only two categories suffice as we shall see in Section 3.6.1 on dummy variables, but that is not germane to the discussion here.) Suppose we want to test the null hypothesis that the average credit card purchases are the same for the three age groups of the customers, i.e., test $H_0 : \beta_1 = \beta_2 = \beta_3$ by testing the following two hypotheses simultaneously:

$$H_{01} : \beta_1 - \beta_2 = 0 \quad \text{and} \quad H_{02} : \beta_1 - \beta_3 = 0.$$

There are infinitely many equivalent ways of writing these hypotheses, e.g., H_{02} can be written as $\beta_2 - \beta_3 = 0$ or $(1/2)(\beta_1 + \beta_2) - \beta_3 = 0$. It can be shown that all such representations of H_0 lead to the same test.

Both the above examples are special cases of the so-called **general linear hypothesis**, namely, that the β's are subject to a set of r linearly independent constraints $H_0 : \boldsymbol{\theta} = \boldsymbol{C\beta} = \mathbf{0}$; r is called the **rank** or the degrees of freedom (d.f.) of the hypothesis. This is the same hypothesis that was considered in the previous section and the F-test based on the simultaneous confidence ellipsoid can be shown to be the extra SS test.

The steps in the extra SS method are as follows:

1. Fit the full model and compute its SSE and SSR.
2. Fit the partial model under H_0 and compute its SSE, denoted by SSE_0, where $\text{SSE}_0 \geq$ SSE.
3. Compute the **hypothesis sum of squares**, $\text{SSH}_0 = \text{SSE}_0 - \text{SSE}$, and the **hypothesis mean square**, $\text{MSH}_0 = \text{SSH}_0/r$.
4. Compute the F-statistic
$$F = \frac{\text{MSH}_0}{\text{MSE}}.$$
5. Reject H_0 at level α if $F > f_{r,n-(p+1),\alpha}$.

This test is based on the result that under the general linear hypothesis $H_0 : \boldsymbol{C\beta} = \mathbf{0}$ of rank r, the F-statistic has the F-distribution with r and $n - (p + 1)$ d.f.

As an application of the extra SS method, consider the overall null hypothesis testing problem (3.22). We know how to compute SSE under the full model. To compute SSE_0, note that the LS estimator of β_0 under $H_0 : E(y) = \beta_0$ is just $\overline{y}$. Therefore under H_0, we have $\widehat{y}_i = \overline{y}$ for all i and $\text{SSE}_0 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \text{SST}$. Using the ANOVA identity (3.13), we get

$$\text{SSH}_0 = \text{SSE}_0 - \text{SSE} = \text{SST} - \text{SSE} = \text{SSR}.$$

The hypothesis d.f. equals $r = p$ and $\text{MSH}_0 = \text{SSR}/p = \text{MSR}$, namely the regression mean square. Thus H_0 is rejected at level α if $F = \text{MSR}/\text{MSE} > f_{p,n-(p+1),\alpha}$, which is the ANOVA F-test (3.16). The t-test (3.17) on a single β_j, which is equivalent to an F-test: reject $H_{0j} : \beta_j = 0$ if $F_j = t_j^2 > f_{1,n-(p+1),\alpha}$, can also be derived using the extra SS method.

EXAMPLE 3.7 (GPA Data: ANOVA for the Linear Model)

The ANOVA Table 3.4 for the linear model gives $\text{SSR} = 12.7859$. A question that is often of interest in such regression problems is what proportion of this SS can be attributed to each predictor, in this case to Verbal and Math scores. Using the extra SS method we can compute the SS due to each predictor by taking the difference between SSR for the full model and SSR_0 for the partial model, which omits only the predictor of interest. In the present problem the full model includes both Verbal and Math, while one partial model includes only Math (to find the SS due to Verbal) and the other partial model includes only Verbal (to find the SS due to Math). In R this calculation is performed by using the `drop1` function. The result is as follows.

```
       Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>               5.9876 -69.968
Verbal  1    6.6187 12.6063 -42.187  40.900 1.834e-07 ***
Math    1    7.5311 13.5186 -39.392  46.538 4.902e-08 ***
```

Note two things about this ANOVA table.

1. The F-test for each predictor is exactly equivalent to the t-test given in the R output from Example 3.5 in that $F = t^2$ and the P-values are identical (the former calculated from the two tails of the t_{37} distribution and the latter calculated from the upper tail of the $F_{1,37}$ distribution).

2. Further note that the SS's for Verbal and Math do not add up to SSR as can be checked from
$$\text{SS}_{\text{Verbal}} + \text{SS}_{\text{Math}} = 6.6187 + 7.5311 = 14.2098 \neq \text{SSR} = 12.7859.$$
This will be the case whenever the predictors are not orthogonal to each other and so their contributions to SSR are not mutually exclusive. Therefore it is not possible to apportion SSR into two independent components corresponding to Verbal and Math scores.

If you use the `anova` function in R then you get the following output.

```
          Df Sum Sq Mean Sq F value    Pr(>F)
Verbal     1 5.2549  5.2549  32.472 1.608e-06 ***
Math       1 7.5311  7.5311  46.538 4.902e-08 ***
Residuals 37 5.9876  0.1618
```

In this table the SS's for Verbal and Math do add up to SSR as can be checked from
$$\text{SS}_{\text{Verbal}} + \text{SS}_{\text{Math}} = 5.2549 + 7.5311 = 12.7860.$$

Table 3.5 ANOVA table for GPA vs. entrance test scores: Quadratic regression

Source	SS	d.f.	MS	F	P
Regression	17.4845	4	4.371	118.69	0.000
Residual Error	1.2890	35	0.0368		
Total	18.7735	39			

However, these SS's are calculated by adding the predictors sequentially in the order in which the predictors are specified in the `lm` function, first Verbal then Math in the present example, and attributing the increase in SSR (or equivalently decrease in SSE) to the last added predictor. These SS's are referred to as **sequential sums of squares** or **type I sums of squares** and it is easy to see that they always add up to SSR. But they depend on the order in which the predictors are entered into the regression model; if the order is changed then they change as well. Only if the predictors are orthogonal then the order of entry does not matter and the SS's of the predictors are the same as those obtained from the `drop1` function based on the extra SS method. These latter SS's are referred to as **adjusted sums of squares** or **type III sums of squares**. Only these sums of squares should be used for testing purposes and not the sequential sums of squares. ■

EXAMPLE 3.8 (GPA Data: Quadratic Model)

In Example 3.3 we fitted a model to the GPA data that was linear in both Verbal and Math scores. In this example we fit the quadratic model:

$$\text{GPA} = \beta_0 + \beta_1 \text{Verbal} + \beta_2 \text{Math} + \beta_3 \text{Verbal}^2 + \beta_4 \text{Math}^2 + \varepsilon.$$

to see if it provides a significantly better fit. The fitted quadratic model is

$$\widehat{\text{GPA}} = -11.458 + 0.189\text{Verbal} + 0.159\text{Math} - 0.0011\text{Verbal}^2 - 0.0009\text{Math}^2.$$

To check if the two quadratic terms significantly improve the fit we can do the extra SS test as follows. Taking the linear model as the partial model and the quadratic model as the full model, we get $\text{SSE}_0 = 5.9876$ and $\text{SSE} = 1.2890$ from Table 3.4 and Table 3.5, respectively. So the F-statistic for testing $\beta_3 = \beta_4 = 0$ in the quadratic model equals

$$F = \frac{(5.9876 - 1.2890)/2}{0.0368} = 63.79,$$

with 2 and 35 d.f., which is clearly highly significant ($P < 0.001$). ■

3.3.5 Prediction of Future Observations

In this section we consider the problem of predicting the response for a given set of predictors. For example, predict an entering freshman's college graduating GPA from his Verbal and Math scores. Denote the vector of the given set of predictors by $\boldsymbol{x}^* = (1, x_1^*, \ldots, x_p^*)'$, the corresponding value of y by y^* and $E(y^*)$ by $\mu^* = \boldsymbol{x}^{*'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$. As in Section 2.2.5 we will give the formulae for the confidence interval (CI) for μ^* and the prediction interval (PI) for y^*.

The unbiased estimate of both y^* and μ^* is given by

$$\widehat{y}^* = \widehat{\mu}^* = \boldsymbol{x}^{*'}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_p x_p^*.$$

By using the formula for $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and a special case (A.5) of the sandwich formula, we get

$$\text{Var}(\widehat{y}^*) = \text{Var}(\widehat{\mu}^*) = \boldsymbol{x}^{*'}\text{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}^* = \sigma^2 \boldsymbol{x}^{*'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*.$$

So

$$\text{SE}(\widehat{\mu}^*) = s\sqrt{\boldsymbol{x}^{*'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*}.$$

Then analogous to (2.19) and (2.20) we get the following formulae for the CI for μ^*:

$$\widehat{\mu}^* \pm t_{n-(p+1),\alpha/2}\, s\sqrt{\boldsymbol{x}^{*'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*} \tag{3.23}$$

and for the PI for y^*:

$$\widehat{y}^* \pm t_{n-(p+1),\alpha/2}\, s\sqrt{1 + \boldsymbol{x}^{*'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*}. \tag{3.24}$$

EXAMPLE 3.9 (GPA Data: Confidence and Prediction Intervals)

Suppose we want to estimate the graduating GPA of a typical freshman with 80 on Verbal and 90 on Math using the quadratic model fitted in Example 3.8. The estimated GPA is

$$\widehat{\text{GPA}} = -11.458 + 0.189(80) + 0.159(90) - 0.0011(80)^2 - 0.0009(90)^2 = 3.584.$$

To calculate the standard error of this estimate, note from the ANOVA Table 3.5 that $s = \sqrt{0.0368} = 0.1918$, $\boldsymbol{x}^* = (1, 80, 90, 80^2, 90^2)'$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is a 5×5 matrix not shown here. Substituting these values in the formula for $\text{SE}(\widehat{\mu}^*)$ we get $\text{SE}(\widehat{\text{GPA}}) = 0.0566$. So the 95% CI for the average graduating GPA of all freshmen with Verbal score = 80 and Math score = 90 is (using $t_{35,0.025} = 2.030$)

$$3.584 \pm (2.030)(0.0566) = [3.469, 3.699].$$

The 95% PI for the GPA of a randomly chosen freshman is

$$3.584 \pm (2.030)\sqrt{1 + (0.0566)^2} = [3.178, 3.990].$$

Note that the PI is much wider than the CI.

In R this calculation can be done by using the `predict` function as shown below.

```
> predict(qmfit,newdata=data.frame(Verbal=80,Math=90),interval="confidence")
       fit      lwr      upr
1 3.583893 3.469074 3.698712
> predict(qmfit,newdata=data.frame(Verbal=80,Math=90),interval="predict")
       fit      lwr      upr
1 3.583893 3.177738 3.990048
```

■

3.4 Weighted and Generalized Least Squares

3.4.1 Weighted Least Squares

In this section we generalize the method of weighted least squares (WLS) introduced in Exercise 2.3. In the case of multiple regression the **WLS criterion** is

$$Q = \sum_{i=1}^{n} w_i[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2, \tag{3.25}$$

where $w_i > 0 \quad (1 \leq i \leq n)$ are the weights. This criterion can be expressed in matrix notation as

$$Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3.26}$$

where $\boldsymbol{W} = \text{diag}\{w_1, \ldots, w_n\}$ is an $n \times n$ diagonal matrix. The weights may be chosen to reflect the relative importance of the observations. A common choice for w_i is $w_i \propto [\text{Var}(y_i)]^{-1}$ if the $\text{Var}(y_i)$ are not constant, i.e., if they are heteroscedastic. This case is discussed below.

Let $\boldsymbol{W}^{1/2} = \text{diag}\{w_1^{1/2}, \ldots, w_n^{1/2}\}$ so that $\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2} = \boldsymbol{W}$. Then the WLS criterion can be written as

$$\begin{aligned} Q &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{W}^{1/2}\boldsymbol{y} - \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{W}^{1/2}\boldsymbol{y} - \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{y}^* = \boldsymbol{W}^{1/2}\boldsymbol{y}$ and $\boldsymbol{X}^* = \boldsymbol{W}^{1/2}\boldsymbol{X}$. This representation shows that the WLS criterion is the same as the LS criterion (3.4) with $\boldsymbol{y}$ replaced by $\boldsymbol{y}^*$ and $\boldsymbol{X}$ replaced by $\boldsymbol{X}^*$. Therefore, using formula‘(3.7) for the LS estimator, the **WLS estimator** of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{WLS}} &= (\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}\boldsymbol{y}^* \\ &= (\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}. \end{aligned} \tag{3.27}$$

If we choose $w_i \propto [\text{Var}(y_i)]^{-1}$ then $y_i^* \propto y_i/\sqrt{\text{Var}(y_i)}$, which makes $\text{Var}(y_i^*)$ constant, thus making them homoscedastic. This is a so-called **variance stabilizing transformation**. These transformations of the response variable are discussed in Section 4.3.1. A common example of $w_i \propto [\text{Var}(y_i)]^{-1}$ occurs when each y_i is an average of n_i i.i.d. repeat observations, y_{ij} $(1 \leq j \leq n_i)$, with a common unknown variance σ^2 so that $\text{Var}(y_i) = \sigma^2/n_i$. For example, the y_{ij} $(1 \leq j \leq n_i)$ may be repeat measurements made under identical experimental conditions, but only their averages are reported. Then the w_i may be taken to be equal to n_i. In that case each term in the WLS criterion (3.25) is weighted by the sample size n_i.

Generally, the $\text{Var}(y_i)$ are completely unknown. Hence the weights w_i are also unknown. Often, it is possible to postulate a relationship between $\text{Var}(y_i)$ and $E(y_i) = \mu_i$, for example, $\text{SD}(y_i) = \sqrt{\text{Var}(y_i)} \propto \mu_i$. In that case we can use WLS regression with $w_i = 1/\mu_i^2$. However, the μ_i are themselves unknown but can be estimated by $\widehat{y}_i$. The following example illustrates WLS regression using $w_i = 1/\widehat{y}_i^2$. The $\widehat{y}_i$'s are obtained using ordinary least squares (OLS). One could iterate on the new WLS values of $\widehat{y}_i$, which results in a so-called **iteratively reweighted least squares (IRWLS)** algorithm discussed in Section 9.2.2. In the example below, we only give the first step.

EXAMPLE 3.10 (GPA Data: Weighted Least Squares)

In Example 3.3 we fitted an OLS regression model to GPA using Verbal and Math scores as predictors. Here we will fit a WLS regression model using $[\widehat{y}_i^2]^{-1}$ from the OLS regression model as weights. These weights correspond to the logarithmic transformation to stabilize error variances suggested by the plots of the residuals against fitted values; see Exercise 4.4. First we repeat the R output from Example 3.3.

```
> lmfit = lm(GPA ~ Verbal+Math , data = gpa)
> summary(lmfit)
```

```
Call:
lm(formula = GPA ~ Verbal + Math, data = gpa)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.570537   0.493749  -3.181  0.00297 **
Verbal       0.025732   0.004024   6.395 1.83e-07 ***
Math         0.033615   0.004928   6.822 4.90e-08 ***
---

Residual standard error: 0.4023 on 37 degrees of freedom
Multiple R-squared:  0.6811,    Adjusted R-squared:  0.6638
F-statistic: 39.51 on 2 and 37 DF,  p-value: 6.585e-10
```

Next we perform WLS regression as previously described.

```
> wlsfit=lm(GPA ~ Verbal+Math, weights=1/(lmfit$fitted)^2,data = gpa)
> summary(wlsfit)

Call:
lm(formula = GPA ~ Verbal + Math, data = gpa, weights
  = 1/(lmfit$fitted)^2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.958844   0.410258  -4.775 2.82e-05 ***
Verbal       0.029191   0.003780   7.723 3.16e-09 ***
Math         0.035567   0.004625   7.689 3.49e-09 ***
---
Residual standard error: 0.1495 on 37 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7531
F-statistic: 60.48 on 2 and 37 DF,  p-value: 2.179e-12
```

We see that the WLS model coefficients are qualitatively similar to those from the OLS model, but are much more significant. This is because the residual standard error s equals 0.4023 for the OLS model whereas it is only 0.1495, almost one-third, for the WLS model. Hence the standard errors of the estimated regression coefficients are correspondingly smaller. ■

3.4.2 Generalized Least Squares

The **generalized least squares (GLS)** method is a generalization of the WLS method in which $\boldsymbol{W}$ is an arbitrary symmetric positive definite $n \times n$ matrix. This generalization is useful when the y_i's are correlated besides being heteroscedastic. So the covariance structure of the y_i's is non-diagonal. Denote $\text{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma}$.

The GLS criterion is the same as the WLS criterion (3.26) except that $\boldsymbol{W}$ is a non-diagonal matrix. In the diagonal matrix case we could write $\boldsymbol{W}^{1/2} = \text{diag}\{w_1^{1/2}, \ldots, w_n^{1/2}\}$ as the "square root" of $\boldsymbol{W}$. When $\boldsymbol{W}$ is a non-diagonal matrix, it is not so simple. However, a result from linear algebra (the spectral decomposition theorem) states that there

exists a symmetric positive definite matrix $\boldsymbol{U}$ such that $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{W}$ and $\boldsymbol{U}\boldsymbol{W}^{-1}\boldsymbol{U}' = \boldsymbol{I}$. In this sense $\boldsymbol{U}$ is the "square root" of $\boldsymbol{W}$. Then the GLS criterion becomes

$$\begin{aligned} Q &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{U}'\boldsymbol{U}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{U}\boldsymbol{y} - \boldsymbol{U}\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{U}\boldsymbol{y} - \boldsymbol{U}\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{y}^* = \boldsymbol{U}\boldsymbol{y}$ and $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{X}$. It follows that the **GLS estimator** $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ of $\boldsymbol{\beta}$ is the same as that given by (3.27) where now $\boldsymbol{W}$ is a non-diagonal matrix. Note that there is no need to find the $\boldsymbol{U}$ matrix explicitly.

In analogy with $w_i \propto [\text{Var}(y_i)]^{-1}$ in the WLS case, we can choose $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$. Then using the sandwich formula (see (A.7)) we get

$$\text{Cov}(\boldsymbol{y}^*) = \boldsymbol{U}\text{Cov}(\boldsymbol{y})\boldsymbol{U}' = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' = \boldsymbol{U}\boldsymbol{W}^{-1}\boldsymbol{U}' = \boldsymbol{I}.$$

Thus not only does the transformation $\boldsymbol{y}^* = \boldsymbol{U}\boldsymbol{y}$ make the y_i^* homoscedastic, but also makes them independent. Generally, $\boldsymbol{\Sigma}$ is not known and an extension of the IRWLS algorithm must be used. However, if $\boldsymbol{\Sigma}$ is known up to a proportionality constant, e.g., $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{V}$ where σ^2 is unknown but $\boldsymbol{V}$ is known then we can choose $\boldsymbol{W} = \boldsymbol{V}^{-1}$.

3.4.3 Statistical Inference on GLS Estimator

We will focus on the GLS estimator since it generalizes the WLS estimator and has the same form. Following the same steps as in the derivation of the distribution of the LS estimator $\widehat{\boldsymbol{\beta}}$ given in Section 3.7 , it can be shown that

$$E(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = \boldsymbol{\beta} \quad \text{and} \quad \text{Cov}(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = \sigma^2(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}')^{-1}. \tag{3.28}$$

Further $\widehat{\boldsymbol{\beta}}_{\text{GLS}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}')^{-1})$ since it is a linear function of $\boldsymbol{y}$ or $\boldsymbol{y}^*$, which are multivariate normal. An unbiased estimate s^2 of σ^2 with $n - (p + 1)$ d.f. can be obtained as

$$s^2 = \frac{\text{SSE}^*}{n-(p+1)} = \frac{(\boldsymbol{y}^* - \boldsymbol{X}^*\widehat{\boldsymbol{\beta}}_{\text{GLS}})'(\boldsymbol{y}^* - \boldsymbol{X}^*\widehat{\boldsymbol{\beta}}_{\text{GLS}})}{n-(p+1)} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{GLS}})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{GLS}})}{n-(p+1)}.$$

Also, $s^2 \sim \chi^2_{n-(p+1)}/[n - (p + 1)]$ independent of $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$. Thus hypothesis tests and confidence intervals on $\widehat{\boldsymbol{\beta}}_{\text{GLS},j}$ can be derived in the same manner as in the LS case. The results in the R output in Example 3.10 are calculated using these results.

3.5 Partial Correlation Coefficient

The population partial correlation coefficient between y and any predictor x_j is the bivariate correlation coefficient between y and x_j in their conditional joint distribution conditioned on all other predictors x_k for $k \neq j$. Just as there is a direct relation (2.24) between the regression coefficient β_1 and the correlation coefficient ρ between y and x in the context of simple linear regression, there is a corresponding relation between each regression coefficient β_j and the population partial correlation coefficient between y and x_j. In fact, the population partial correlation coefficient is a normalized unitless version with range of $[-1, 1]$ of the the regression coefficient. An expression for this population partial correlation coefficient can be derived (see Section 3.7.7 in the Technical Notes section). However, here we will only consider its sample version.

We will first show how to compute the sample partial correlation coefficient in the case of two predictor variables. Let $r_{yx_2|x_1}$ denote the partial correlation coefficient between y and x_2 conditioned on x_1. The following are alternative ways to compute $r_{yx_2|x_1}$.

- Consider a set of residuals from regression of y on x_1 and another set of residuals from regression of x_2 on x_1. Then $r_{yx_2|x_1}$ is the sample correlation coefficient between these two sets of residuals. Thus the partial correlation coefficient between y and x_2 is the residual correlation between them after removing the effect of x_1 on both y and x_2 by regressing them on x_1.

- Let $\text{SSE}(x_1)$ and $\text{SSE}(x_1, x_2)$ denote the SSE's from regressions of y on x_1 and on x_1, x_2, respectively. Then
$$r^2_{yx_2|x_1} = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1)} = 1 - \frac{\text{SSE}(x_1, x_2)}{\text{SSE}(x_1)} \tag{3.29}$$
and $r_{yx_2|x_1}$ is the square root of the above, its sign being that of $\widehat{\beta}_2$ in the regression of y on x_1, x_2.

  This formula extends the definition $R^2 = 1 - \text{SSE}/\text{SST}$ by replacing SSE with $\text{SSE}(x_1, x_2)$ and SST with $\text{SSE}(x_1)$. Essentially R^2 compares the SSE of the full model consisting of p variables with SSE = SST of the null model consisting only of the intercept term, which is treated as the partial model. Here $\{x_1, x_2\}$ is the full model and $\{x_1\}$ is the partial model.

- A third method of computing $r_{yx_2|x_1}$ is from the bivariate correlations r_{yx_1}, r_{yx_2} and $r_{x_1x_2}$ using the formula:
$$r_{yx_2|x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_1})(1 - r^2_{x_1x_2})}}. \tag{3.30}$$
  This formula is the sample version of the population partial correlation coefficient given in Equation (3.36). It can also be derived from the first method given above.

The above formulae can be extended by conditioning on multiple predictors. Denote the partial correlation coefficient between y and x_p conditioned on $x_1, \ldots, x_{p-1}$ by $r_{yx_p|x_1,\ldots,x_{p-1}}$. Analogous to (3.29), we have
$$r^2_{yx_p|x_1,\ldots,x_{p-1}} = 1 - \frac{\text{SSE}(x_1, \ldots, x_p)}{\text{SSE}(x_1, \ldots, x_{p-1})}. \tag{3.31}$$
Then $r_{yx_p|x_1,\ldots,x_{p-1}}$ is the square root of the above, its sign being that of $\widehat{\beta}_p$ in the regression of y on $x_1, \ldots, x_p$.

EXAMPLE 3.11 (GPA Data: Partial Correlation Coefficients)

Denoting GPA = y, Verbal = x_1 and Math = x_2, we want to calculate $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$. By running regressions of GPA on Verbal and on Math we get $\text{SSE}(x_1) = 13.5186$ and $\text{SSE}(x_2) = 12.6063$. From the ANOVA Table 3.4 we get $\text{SSE}(x_1, x_2) = 5.9876$. Hence
$$r^2_{yx_1|x_2} = 1 - \frac{5.9876}{12.6063} = 0.5250 \quad \text{and} \quad r^2_{yx_2|x_1} = 1 - \frac{5.9876}{13.5186} = 0.5571.$$
Thus $r_{yx_1|x_2} = \sqrt{0.5250} = 0.7246$ and $r_{yx_2|x_1} = \sqrt{0.5571} = 0.7464$. The coefficients of both are positive since the corresponding regression coefficients are positive.

The same values can be obtained using the formula (3.30). The bivariate correlations are $r_{yx_1} = 0.5291, r_{yx_2} = 0.5732$ and $r_{x_1x_2} = -0.1069$. By substituting these values in (3.30) we get
$$r_{yx_1|x_2} = \frac{0.5291 + 0.1069 \times 0.5732}{\sqrt{(1 - 0.1069^2)(1 - 0.5731^2)}} = 0.7246$$

and

$$r_{yx_2|x_1} = \frac{0.5732 + 0.1069 \times 0.5291}{\sqrt{(1 - 0.1069^2)(1 - 0.5291^2)}} = 0.7464.$$

■

Test of Significance of Partial Correlation Coefficient

The test of significance on the sample partial correlation coefficient can be derived using the extra SS method of Section 3.3.4. There we saw that the test of significance for entering x_p to the model that already includes $x_1, \ldots, x_{p-1}$ is based on the following F-statistic:

$$F_p = \frac{[\text{SSE}(x_1, \ldots, x_{p-1}) - \text{SSE}(x_1, \ldots, x_p)]/1}{\text{SSE}(x_1, \ldots, x_p)/[n - (p+1)]}. \tag{3.32}$$

Using simple algebra, this F-statistic can be written as

$$F_p = \frac{r^2_{yx_p|x_1,\ldots,x_{p-1}}[n - (p+1)]}{1 - r^2_{yx_p|x_1,\ldots,x_{p-1}}}.$$

The corresponding t-statistic is the square root of F_p:

$$t_p = \sqrt{F_p} = \frac{r_{yx_p|x_1,\ldots,x_{p-1}}\sqrt{n - (p+1)}}{\sqrt{1 - r^2_{yx_p|x_1,\ldots,x_{p-1}}}}. \tag{3.33}$$

We conclude that $r_{yx_p|x_1,\ldots,x_{p-1}}$ is significant at level α if $F_p > f_{1,n-(p+1),\alpha}$ or equivalently if $|t_p| > t_{n-(p+1),\alpha/2}$. By comparing this test with (2.25), we see that it is a generalization of the test on the bivariate correlation coefficient. Furthermore this t-test is exactly the same t-test used to test the significance of $\widehat{\beta}_p$ in the multiple regression of y on $x_1, \ldots, x_p$, which follows the equivalence between the regression coefficient and the partial correlation coefficient.

■ **EXAMPLE 3.12 (GPA Data: Significance Tests on Partial Correlation Coefficients)**

The t-statistics to test the significance of $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$ are

$$t_{yx_1|x_2} = \frac{0.7246\sqrt{37}}{\sqrt{1 - 0.7246^2}} = 6.395 \quad \text{and} \quad t_{yx_2|x_1} = \frac{0.7464\sqrt{37}}{\sqrt{1 - 0.7464^2}} = 6.822,$$

which are identical to the ones for the regression coefficients given in the R output in Example 3.3. Both are highly significant. ■

3.6 Special Topics

3.6.1 Dummy Variables

A dummy variable is an indicator variable used to denote the presence or absence of a given categorical attribute. In the case of a binary variable such as treated versus untreated (control) patients in a clinical trial we need only one dummy variable, e.g., $x_1 = 0$ for control patients and $x_1 = 1$ for treated patients. Suppose the model is $E(y) = \beta_0 + \beta_1 x_1$, where y is some measure of patient response. Then the models for control and treated patients are

$$\text{Control:} \quad E(y) = \beta_0 \quad \text{and} \quad \text{Treated:} \quad E(y) = \beta_0 + \beta_1.$$

Thus β_1 is the difference between the expected response of the treated patients and of the control patients and is called the **treatment effect**.

Next consider a categorical variable with $c > 2$ categories, e.g., ethnicity with five groups: White, Black, Hispanic, Asian and Other, and suppose we code them $1, \ldots, 5$ and treat it as a numerical variable. What is wrong with this approach?

- It is wrong to treat the ethnicity variable coded as $1, \ldots, 5$ as a numerical variable since it implies a linear order. Thus if the coefficient of ethnicity is β then the coefficient for Whites will be β, that for Blacks will be 2β, etc., which is wrong since not only the coding $1, \ldots, 5$ is arbitrary for a nominal variable such as ethnicity but even if the variable were ordinal such as course grade (with categories A, B, C, D, F), the effect of grade is not necessarily linear.

- We need only $c - 1$ dummy variables to code c categories. For example, if we use $x_1, \ldots, x_5$ as dummy variables for the five ethnic groups then since exactly one $x_j = 1$ and other $x_j = 0$ for each person, so we will have $x_1 + \cdots + x_5 = 1$ for all persons. As we will see in Chapter 4, this causes multicollinearity (linear dependence among the predictor variables). Hence we need to use only four dummy variables for any four groups leaving the fifth group as the reference for comparison.

- Any category can be chosen as the reference. The β coefficients for the other categories represent the differences from that reference.

EXAMPLE 3.13 (Used Car Prices: Regression Model 1)

In this example there are six categorical variables, three with two categories (Cruise, Sound and Leather) and three with more than two categories (Make, Model and Type). In fact, Doors is also a categorical variable since it has only two values: 2 and 4. However, it is perfectly correlated with the Type variable since Coupe and Convertible always have 2 doors, while Hatchback, Sedan and Wagon always have 4 doors. So we drop Doors as a predictor variable. For now we will also ignore the Model variable.

R chooses the first listed category of any categorical variable as the reference by default. Thus it chooses Buick as the reference category for Make and Convertible as the reference category for Type. So the regression coefficients for the other categories for Make and Type represent the differences from these reference categories.

For fitting the predictive model we divided the data equally into a training set and a test set. Here we used a deterministic instead of a random split by putting all odd-numbered observations into the training set and all even-numbered observations into the test set. We did this to enable readers verify the results given here using their own codes.

We log-transformed Price so that it better satisfies the model assumptions such as normality and homoscedasticity. This point will be discussed in more detail in Chapter 4. The coefficients, their standard errors, t-statistics and P-values are listed in Table 3.6. The R^2 for the model is 95.26% and the F-statistic equals 516.8 on 15 and 386 d.f. which is highly significant.

We see that none of the dummy variables for Cruise, Sound or Leather is significant at $\alpha = 0.05$. We may drop these variables from the model. Finally, the number of cylinders is also nonsignificant with a P-value of 0.074, the reason being that the engine size (Liter) has a very high correlation of 0.958 with it. So we keep only the

Table 3.6 Statistics for used car prices regression: Model 1

Predictor	Coef	SE	t	P	Predictor	Coef	SE	t	P
Constant	4.1883	0.0222	188.352	0.000	Chevrolet	-0.0575	0.0081	-7.129	0.000
Mileage	-0.0035	0.0002	-14.227	0.000	Pontiac	-0.0402	0.0082	-4.905	0.000
Cylinder	-0.0127	0.0071	-1.790	0.074	SAAB	0.2357	0.0103	22.922	0.000
Liter	0.1095	0.0079	13.811	0.000	Saturn	-0.0450	0.0107	-4.221	0.000
Cruise	0.0097	0.0060	1.627	0.105	Coupe	-0.1404	0.0106	-13.183	0.000
Sound	0.0037	0.0046	0.801	0.424	Hatchback	-0.1536	0.0124	-12.375	0.000
Leather	0.0082	0.0048	1.694	0.091	Sedan	-0.1437	0.0092	-15.600	0.000
Cadillac	0.2004	0.0109	18.333	0.000	Wagon	-0.0730	0.0115	-6.372	0.000

engine size in the model. The resulting model is

$$\begin{aligned}\widehat{\log(\text{Price})} &= 4.169 - 0.0035\ \text{Mileage} + 0.0976\ \text{Liter} + 0.1948\ \text{Cadillac} \\ &\quad -0.0540\ \text{Chevrolet} - 0.0413\ \text{Pontiac} + 0.2463\ \text{SAAB} - 0.0463\ \text{Saturn} \\ &\quad -0.1337\ \text{Coupe} - 0.1574\ \text{Hatchback} - 0.1412\ \text{Sedan} - 0.0714\ \text{Wagon}.\end{aligned}$$

All of the variables in this model are highly significant ($P < 0.001$). The R^2 for this model is 95.10%. When this model is applied to the test set, the resulting R^2 is 95.09%, almost exactly the same. ■

EXAMPLE 3.14 (Used Car Prices: Regression Model 2)

In this example we will take the Model variable into account. Note that the Model variable is nested under the Make variable, e.g., the Century Model is available only for Buick. R uses the first Model as the reference for each make and so omits the dummy variable for that Model from the regression equation. Buick has four Models: Century, Lacrosse, Lesabre and Park Avenue. Cadillac has six Models: CST-V, CTS, Deville, STS-V6, STS-V8 and XLR-V8. Chevrolet has eight Models: AVEO, Cavalier, Classic, Cobalt, Corvette, Impala, Malibu and Monte Carlo. Pontiac has seven Models: Bonneville, G6, Grand Am, Grand Prix, GTO, Sunfire and Vibe. SAAB has five Models: 9-3, 9-3 HO, 9-5, 9-5 HO and 9-2X AWD. Finally, Saturn has only two Models: Ion and L Series.

After eliminating nonsignificant predictors, the regression coefficients for the final model along with their standard errors, t-statistics and P-values are given in Table 3.7. The R^2 for this model is 97.90%, which is a significant improvement over $R^2 = 95.10\%$ for Model 1. When this model is applied to the test set, the resulting R^2 is 97.66%, almost exactly the same. ■

3.6.2 Interactions

Often we want to model the joint effect of two or more variables on the response variable. Such a joint effect is called **interaction**. For example, sale price and discount coupon

Table 3.7 Statistics for used car prices regression: Model 2

Predictor	Coef	SE	t	P	Predictor	Coef	SE	t	P
Constant	4.1649	0.0131	318.85	0.000	Chevrolet				
Mileage	−0.0034	0.00017	−19.98	0.000	Cavalier	0.0467	0.0085	5.46	0.000
Liter	0.0556	0.0039	13.44	0.000	Classic	0.0726	0.0136	5.33	0.000
Cadillac	0.3520	0.0129	27.36	0.000	Cobalt	0.0816	0.0087	9.41	0.000
Chevrolet	−0.0727	0.0085	−8.50	0.000	Corvette	0.2711	0.0220	12.34	0.000
Pontiac	0.0870	0.0086	10.15	0.000	Impala	0.1487	0.0129	11.53	0.000
SAAB	0.3151	0.0073	42.99	0.000	Malibu	0.0969	0.0105	9.25	0.000
Coupe	−0.0828	0.0090	−9.20	0.000	Monte Carlo	0.1438	0.0139	10.34	0.000
Hatchback	−0.0786	0.0106	−7.43	0.000	Pontiac				
Sedan	−0.0908	0.0080	−11.34	0.000	Grand Am	−0.0882	0.0106	−8.32	0.000
Wagon	−0.0822	0.0102	−8.08	0.000	Grand Prix	−0.0515	0.0086	−6.00	0.000
Buick					GTO	0.0531	0.0167	3.19	0.002
Lacrosse	0.1176	0.0106	11.46	0.000	Sunfire	−0.1206	0.0147	−8.20	0.000
Lesabre	0.1573	0.0117	13.50	0.000	SAAB				
Park Avenue	−0.0806	0.0125	−6.43	0.000	9-5	0.0264	0.0097	2.74	0.006
Cadillac					9-5 HO	0.1122	0.0211	5.32	0.000
CTS	−0.0398	0.0158	−2.52	0.012	9-2X AWD	−0.0904	0.0211	−4.29	0.000
Deville	−0.0574	0.0098	−5.86	0.000	Saturn				
XLR-V8	0.0876	0.0160	5.48	0.000	L Series	0.0308	0.0134	2.30	0.022

Table 3.8 Average salaries by gender and race

		Race	
		Non-White	White
Gender	Female	\$40K	\$50K
	Male	\$45K	\$65K

have separate positive effects on the sales of a product but if both are offered together then the total effect on the sales can be greater than their sum. In this case we say that there is a positive interaction between sale price and coupon. Interaction between two or more predictor variables is typically modeled by their product, i.e., by including x_1x_2 as the interaction between x_1 and x_2. If interaction is not present then the model is said to be **additive** in the given variables.

EXAMPLE 3.15 (Interaction Between Two Categorical Variables)

Consider the fictitious average salary data in Table 3.8 for which we want to model the interaction between Gender and Race. Suppose Gender is coded as $x_1 = 0$ for females, $x_1 = 1$ for males and Race is coded as $x_2 = 0$ for non-Whites, $x_2 = 1$ for Whites. Ignoring any other predictors, the model with interaction is $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$, which can be written as four separate models:

$$E(y) = \begin{cases} \beta_0 & \text{for Female and Non-Whites} \\ \beta_0 + \beta_1 & \text{for Male and Non-Whites} \\ \beta_0 + \beta_2 & \text{for Female and Whites} \\ \beta_0 + \beta_1 + \beta_2 + \beta_3 & \text{for Male and Whites} \end{cases}$$

The estimated Male versus Female difference (called the Gender effect) is $\widehat{\beta}_1 = \$45\text{K} - \$40\text{K} = \$5\text{K}$ for Non-Whites and $\widehat{\beta}_1 + \widehat{\beta}_3 = \$65\text{K} - \$50\text{K} = \$15\text{K}$ for Whites, representing a positive $\widehat{\beta}_2 + \widehat{\beta}_3 = \$65\text{K} - \$45\text{K} = \$20\text{K}$ for Males, again representing a positive interaction $\widehat{\beta}_3 = \$10\text{K}$.

■

Next let us consider interaction between a dummy variable and a **covariate** (i.e., a continuous predictor variable). In the example from Section 3.6.1 about treated versus untreated (control) patients in a clinical trial with $x_1 = 0$ for control and $x_1 = 1$ for treated patients, suppose we include Age (x_2) as a covariate and postulate the model

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2,$$

where the x_1x_2 term represents the Treatment×Age interaction. To understand what this interaction means write this model as two separate models:

Control: $E(y) = \beta_0 + \beta_2x_2$ and Treated: $E(y) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_2$.

Thus we get two non-parallel regression lines with different intercepts and different slopes as shown in Figure 3.3. The difference between control and treated patients depends on age, so we cannot estimate the treatment effect independent of the patient's age. If $\beta_3 = 0$ then there is no interaction and we get two parallel regression lines with different intercepts, so the effect of the treatment is the same for all ages and does not depend on the age.

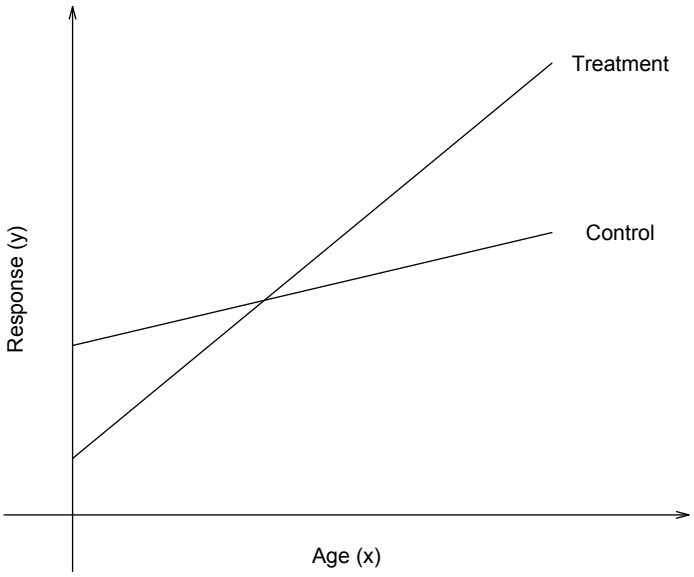


Figure 3.3 Non-parallel regression lines between response (y) and age (x) for control and treated patients

EXAMPLE 3.16 (GPA Data: Interaction Effect)

In Example 3.8 we fitted a quadratic model to the GPA data which we found to give a significantly better fit than the linear model. In this example we will examine whether adding the interaction term further improves the fit of the model significantly. Thus consider the full second degree model:

$E(\text{GPA}) = \beta_0 + \beta_1 \text{Verbal} + \beta_2 \text{Math} + \beta_3 \text{Verbal}^2 + \beta_4 \text{Math}^2 + \beta_5 \text{Verbal} \times \text{Math}.$

The regression output for this model is given below. We see that the interaction term has a P-value = 0.103, which barely fails to be significant at $\alpha = 0.10$.

```
Call:
lm(formula = GPA ~ Verbal * Math + I(Verbal^2) + I(Math^2),
 data = gpa)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.9167631  1.3544134  -7.322 1.75e-08 ***
Verbal       0.1668098  0.0212447   7.852 3.85e-09 ***
Math         0.1375972  0.0267340   5.147 1.11e-05 ***
I(Verbal^2) -0.0011082  0.0001173  -9.449 4.88e-11 ***
I(Math^2)   -0.0008433  0.0001594  -5.290 7.23e-06 ***
Verbal:Math  0.0002411  0.0001440   1.675    0.103
---
Residual standard error: 0.1871 on 34 degrees of freedom
```

Table 3.9 ANOVA Table for GPA versus entrance test scores: Quadratic regression with interaction

Source	SS	d.f.	MS	F	P
Regression	17.5827	5	3.5165	100.41	0.000
Residual Error	1.1908	34	0.0350		
Total	18.7735	39			

```
Multiple R-squared:  0.9366,    Adjusted R-squared:  0.9272
F-statistic: 100.4 on 5 and 34 DF,  p-value: < 2.2e-16
```

The ANOVA table for this second degree model with interaction is given in Table 3.9. We get the same result if we use the extra SS test by treating the quadratic model as the reduced model and this model as the full model. Then the F-statistic for $H_0 : \beta_5 = 0$ equals

$$F = \frac{(1.2890 - 1.1908)/1}{1.1908/34} = 2.806,$$

which for $\alpha = 0.10$ can be compared with $f_{1,34,0.10} = 2.859$. Since $F = 2.806 < 2.859$, we fail to reject H_0. The equivalence between the t-test on β_5 in Table **??** and the F-test calculated using the extra SS method can be verified by checking that $t = \sqrt{2.806} = 1.675$ and $t_{34,0.05} = \sqrt{f_{1,34,0.10}} = \sqrt{2.859} = 1.691$. ■

3.6.3 Standardized Regression

As mentioned earlier in Section 3.2.2, **standardized regression coefficients** are unitless and hence can be compared with each other. They are also useful for computational purposes as we will explain below.

To perform standardized regression, we first standardize all variables:

$$y_i^* = \frac{y_i - \overline{y}}{s_y} \quad \text{and} \quad x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{s_{x_j}} \quad (i = 1, \ldots, n, j = 1, \ldots, p),$$

where $\overline{y}$ and $\overline{x}_j$'s are the sample means, and s_y and s_{x_j} are the sample standard deviations of the corresponding variables. Denote the estimated standardized regression coefficients by $\widehat{\beta}_j^*$ $(j = 0, \ldots, p)$. Since all variables are standardized with 0 means, it follows from (3.8) that $\widehat{\beta}_0^* \equiv 0$. So we can omit the intercept term from the regression equation.

Let $\boldsymbol{y}^*$ denote the n-vector of standardized y_i^*'s, $\boldsymbol{X}^*$ denote the $n \times p$ matrix of standardized x_{ij}^*'s (note that there is no column corresponding to the intercept term) and $\widehat{\boldsymbol{\beta}}^* = (\widehat{\beta}_1^*, \ldots, \widehat{\beta}_p^*)'$ denote the p-vector of standardized $\widehat{\beta}_j^*$'s. Then analogous to the Equation (3.7), we have

$$\widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}\boldsymbol{y}^*.$$

The (j, k)th entry of $\boldsymbol{X}^{*'}\boldsymbol{X}^*$ equals

$$\sum_{i=1}^{n} x_{ij}^* x_{ik}^* = \frac{1}{s_{x_j} s_{x_k}} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) = \frac{(n-1)\text{Cov}(x_j, x_k)}{s_{x_j} s_{x_k}} = (n-1) r_{x_j, x_k},$$

where $\text{Cov}(x_j, x_k)$ is the sample covariance between x_j and x_k and r_{x_j,x_k} is the sample correlation coefficient between x_j and x_k. Similarly, the jth entry of $\boldsymbol{X}^{*'}\boldsymbol{y}^*$ equals $(n-1)r_{y,x_j}$, where r_{y,x_j} is the sample correlation coefficient between y and x_j.

Let $\boldsymbol{R}$ denote the $p \times p$ sample correlation matrix among $x_1, \ldots, x_p$ with entries r_{x_j,x_k} for $j \neq k$ and let $\boldsymbol{r} = (r_{y,x_1}, \ldots, r_{y,x_p})'$ denote the vector of sample correlations between y and $x_1, \ldots, x_p$. Then from the above it follows that

$$\boldsymbol{X}^{*'}\boldsymbol{X}^* = (n-1)\boldsymbol{R} \quad \text{and} \quad \boldsymbol{X}^{*'}\boldsymbol{y}^* = (n-1)\boldsymbol{r}.$$

Therefore the above formula for $\widehat{\boldsymbol{\beta}}^*$ simplifies to

$$\widehat{\boldsymbol{\beta}}^* = \boldsymbol{R}^{-1}\boldsymbol{r}. \tag{3.34}$$

Thus to calculate $\widehat{\boldsymbol{\beta}}^*$ we only need to know all the sample correlation coefficients. Then the unstandardized regression coefficients can be calculated using

$$\widehat{\beta}_j = \widehat{\beta}_j^* \left(\frac{s_y}{s_{x_j}} \right) \quad (j = 1, \ldots, p)$$

and $\widehat{\beta}_0$ can be calculated from (3.8).

This method of computation of $\widehat{\beta}_j$ is numerically more stable because all entries of $\boldsymbol{R}$ and $\boldsymbol{r}$ are between -1 and $+1$, whereas entries of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{y}$ required in the direct computation of $\widehat{\boldsymbol{\beta}}$ using (3.7) can vary over wide ranges and have diverse scales. The formula (3.34) is also useful in that it clearly pinpoints how the multicollinearity problem arises if the x_j's are highly correlated with each other in which case $\boldsymbol{R}$ is close to being singular.

EXAMPLE 3.17 (GPA Data: Standardized Regression Coefficients)

The standardized regression coefficients $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ of Verbal and Math can be computed given $\widehat{\beta}_1 = 0.0257$ and $\widehat{\beta}_2 = 0.0336$ from Example 3.3 and the standard deviations

$$s_y = 0.6938, s_{x_1} = 16.1019, s_{x_2} = 13.1481.$$

So we get

$$\widehat{\beta}_1^* = 0.0257 \left(\frac{16.1019}{0.6938} \right) = 0.5964 \quad \text{and} \quad \widehat{\beta}_2^* = 0.0336 \left(\frac{13.1481}{0.6938} \right) = 0.6367.$$

Notice that $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ are proportional to the t-statistics for β_1 and β_2, respectively, calculated in Example 3.12. ■

3.7 Technical Notes

3.7.1 Derivation of the LS Estimators

The LS estimation problem can be written as minimizing the LS criterion Q defined in (3.4), i.e.,

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} [\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}]. \tag{3.35}$$

The LS estimator $\widehat{\boldsymbol{\beta}}$ is found by taking the vector derivative,

$$\frac{dQ}{d\boldsymbol{\beta}} = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \ldots, \frac{\partial Q}{\partial \beta_p})',$$

and setting it equal to the null vector $\mathbf{0}$. The term $\boldsymbol{y}'\boldsymbol{y}$ does not contribute to the derivative since it does not involve $\boldsymbol{\beta}$. The next term $-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ is linear in $\boldsymbol{\beta}$ and its derivative can be shown to be $-2\boldsymbol{X}'\boldsymbol{y}$. Finally, the last term $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$ is quadratic in $\boldsymbol{\beta}$ and and its derivative can be shown to be $2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$. So the equation for $\widehat{\boldsymbol{\beta}}$ is

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \quad \Longrightarrow \quad (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y},$$

which is the normal equation (3.6).

3.7.2 Distribution of the LS Estimators

We use the following result: If $\boldsymbol{u} = (u_1, \ldots, u_n)'$ has an n-variate normal distribution with $E(\boldsymbol{u}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is an $m \times n$ matrix of constants with linearly independent rows with $m \leq n$ then $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{u} = (v_1, \ldots, v_m)$ has an m-variate normal distribution with mean vector $\boldsymbol{A}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$. We call this formula for the covariance matrix as the **sandwich formula** derived in (**??**) in Appendix A.

Put $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ which is a $(p+1) \times n$ matrix of constants and $\boldsymbol{u} = \boldsymbol{y}$, which is n-variate normal with $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$. Then $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{u}$ is $(p+1)$-variate normal with mean vector

$$E(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and covariance matrix

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{Cov}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{V}. \end{aligned}$$

3.7.3 Gauss-Markov Theorem:

Under the assumptions of the linear model (3.3) (the normality assumption is not needed), among all linear unbiased estimators (i.e., unbiased estimators that are linear functions of y_i's) of any linear parametric function $\theta = \boldsymbol{c}'\boldsymbol{\beta}$, the LS estimator $\widehat{\theta} = \boldsymbol{c}'\widehat{\boldsymbol{\beta}}$ has the smallest variance. Thus the LS estimator is the **best linear unbiased estimator (BLUE)**.

Proof: Let $\widetilde{\boldsymbol{\beta}}$ be any other linear unbiased estimator of $\boldsymbol{\beta}$, i.e., $E(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Write $\widetilde{\boldsymbol{\beta}} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{A}]\boldsymbol{y} = \widehat{\boldsymbol{\beta}} + \boldsymbol{A}\boldsymbol{y}$ where $\boldsymbol{A}$ is any $(p+1) \times n$ matrix of constants (so that $\widetilde{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$). For $\widetilde{\boldsymbol{\beta}}$ to be unbiased, we must have

$$E(\widetilde{\boldsymbol{\beta}}) = E(\widehat{\boldsymbol{\beta}}) + \boldsymbol{A}E(\boldsymbol{y}) = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta}.$$

Hence $\boldsymbol{A}\boldsymbol{X}$ must equal a null matrix $\boldsymbol{O}$: $(p+1) \times (p+1)$.

Next,

$$\begin{aligned} \text{Cov}(\widetilde{\boldsymbol{\beta}}) &= \text{Cov}(\widehat{\boldsymbol{\beta}} + \boldsymbol{A}\boldsymbol{y}) \\ &= \text{Cov}(\widehat{\boldsymbol{\beta}}) + \text{Cov}(\boldsymbol{A}\boldsymbol{y}) + 2\text{Cov}(\widehat{\boldsymbol{\beta}}, \boldsymbol{A}\boldsymbol{y}) \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{A}\text{Cov}(\boldsymbol{y})\boldsymbol{A}' + 2\text{Cov}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}) \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{A}\boldsymbol{A}' + 2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{Cov}(\boldsymbol{y})\boldsymbol{A}' \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{A}\boldsymbol{A}' + 2\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A}' \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{A}\boldsymbol{A}' \ \ (\text{since } \boldsymbol{X}'\boldsymbol{A}' = \boldsymbol{O}') \\ &= \text{Cov}(\widehat{\boldsymbol{\beta}}) + \sigma^2\boldsymbol{A}\boldsymbol{A}'. \end{aligned}$$

Hence $\text{Cov}(\widetilde{\boldsymbol{\beta}}) - \text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{A}\boldsymbol{A}'$, which is positive semidefinite. Therefore for any two linear unbiased estimators, $\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$, of a parametric function $\boldsymbol{c}'\boldsymbol{\beta}$, we have

$$\begin{aligned} \text{Var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}) - \text{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\beta}}) &= \boldsymbol{c}'[\text{Cov}(\widetilde{\boldsymbol{\beta}}) - \text{Cov}(\widehat{\boldsymbol{\beta}})]\boldsymbol{c} \\ &= \sigma^2\boldsymbol{c}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{c} \\ &= \sigma^2\boldsymbol{d}'\boldsymbol{d} \geq 0, \end{aligned}$$

where $\boldsymbol{d} = \boldsymbol{A}'\boldsymbol{c}$. Hence $\text{Var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}) \geq \text{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\beta}})$. ■

3.7.4 Properties of the Hat Matrix

It is straightforward to show that $\boldsymbol{HH} = \boldsymbol{H}$, hence $\boldsymbol{H}$ is a projection matrix, which projects $\boldsymbol{y}$ into $\widehat{\boldsymbol{y}}$. Next,

$$\text{tr}(\boldsymbol{H}) = \text{tr}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] = \text{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}] = \text{tr}(\boldsymbol{I}_{p+1}) = p+1,$$

where we have used the facts that $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$ if both the products are defined. This follows from the fact that the eigenvalues of a projection matrix are either 1 or 0. Furthermore, the number of nonzero eigenvalues equals the rank and the sum of the eigenvalues equals the trace. From the above it follows that $\boldsymbol{I} - \boldsymbol{H}$ is also a projection matrix (which projects $\boldsymbol{y}$ into $\boldsymbol{e}$) with $\text{tr}(\boldsymbol{I} - \boldsymbol{H}) = \text{tr}(\boldsymbol{I}) - \text{tr}(\boldsymbol{H}) = n - (p+1)$.

3.7.5 Properties of Fitted Values and Residuals

First note that since $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ and $\boldsymbol{e} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$, it follows that $\widehat{\boldsymbol{y}}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{y} = 0$, where we have used the facts that $\boldsymbol{H}$ is symmetric and $\boldsymbol{HH} = \boldsymbol{H}$. Thus $\widehat{\boldsymbol{y}}$ and $\boldsymbol{e}$ are orthogonal to each other. From this it follows that $(\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}})'\boldsymbol{e} = 0$ since $\overline{\boldsymbol{y}} = \overline{y}\mathbf{1}$, where $\mathbf{1}$ is an n-vector of all 1's and $\mathbf{1}'\boldsymbol{e} = 0$, which follows from the result that $\boldsymbol{e}$ is orthogonal to every column of $\boldsymbol{X}$ (as shown below) and $\mathbf{1}$ is the first column of $\boldsymbol{X}$ corresponding to the intercept term. This result follows from

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{X}' - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = (\boldsymbol{X}' - \boldsymbol{X}')\boldsymbol{y} = \mathbf{0},$$

where $\mathbf{0}$ is a $(p+1)$-vector of all 0's.

Finally, we derive the distributions of the fitted and residual vectors. Since $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ is a linear transform of $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, it follows that $\widehat{\boldsymbol{y}}$ has a multivariate normal distribution with mean vector, $E(\widehat{\boldsymbol{y}}) = E(\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\mu}$, and covariance matrix (using the sandwich formula (A.7)),

$$\text{Cov}(\widehat{\boldsymbol{y}}) = \text{Cov}(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{H}\text{Cov}(\boldsymbol{y})\boldsymbol{H}' = \sigma^2\boldsymbol{HIH} = \sigma^2\boldsymbol{H} \quad (\text{ since } \boldsymbol{HH} = \boldsymbol{H}).$$

Thus $\widehat{y}_i \sim N(\mu_i, \sigma^2 h_{ii})$, where h_{ii} is the ith diagonal entry of $\boldsymbol{H}$ $(1 \le i \le n)$.

Similarly, $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ has a multivariate normal distribution with the mean vector,

$$\begin{aligned} E(\boldsymbol{e}) &= (\boldsymbol{I} - \boldsymbol{H})E(\boldsymbol{y}) \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \end{aligned}$$

and covariance matrix,

$$\text{Cov}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\text{Cov}(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

Thus $e_i \sim N(0, \sigma^2(1 - h_{ii}))$ $(1 \le i \le n)$.

3.7.6 Confidence Ellipsoid for $\boldsymbol{\beta}$

As shown above, $\widehat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2\boldsymbol{V}$. Since $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ is a positive definite matrix, using the spectral decomposition theorem (see Appendix A), there exists a nonsingular $(p+1)\times(p+1)$ matrix $\boldsymbol{U}$ such that $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}^{-1}$ and $\boldsymbol{UVU}' = \boldsymbol{I}$. Make the transformation $\boldsymbol{z} = \boldsymbol{U}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Then $\boldsymbol{z}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{UVU}' = \sigma^2\boldsymbol{I}$. So the elements z_i $(1 \le i \le p+1)$ of $\boldsymbol{z}$ are i.i.d. $N(0, \sigma^2)$. Hence

$$\sum_{i=1}^{p+1} z_i^2 = \boldsymbol{z}'\boldsymbol{z} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{U}'\boldsymbol{U}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \sigma^2\chi^2_{p+1}.$$

Furthermore, from Section 3.3.2 we also know that $[n - (p+1)]s^2 \sim \sigma^2\chi^2_{n-(p+1)}$ and these two χ^2 r.v.'s are independent. Hence their ratio divided by their degrees of freedom

is distributed as $F_{p+1,n-(p+1)}$ or

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)s^2} \sim F_{p+1,n-(p+1)}.$$

The confidence ellipsoid (3.18) for $\boldsymbol{\beta}$ follows from this result.

3.7.7 Population Partial Correlation Coefficient

Consider a more general setting of a random vector $\boldsymbol{x}$ of dimension p partitioned into two subvectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of dimensions p_1 and p_2, respectively, such that $p_1 + p_2 = p$. If $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denotes the joint distribution of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $f(\boldsymbol{x}_1)$ denotes the marginal distribution of $\boldsymbol{x}_1$ then the conditional distribution of $\boldsymbol{x}_2$ conditioned on $\boldsymbol{x}_1$ is given by $f(\boldsymbol{x}_2|\boldsymbol{x}_1) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)/f(\boldsymbol{x}_1)$. In particular, suppose that $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a multivariate normal (MVN) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ partitioned corresponding to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then it can be shown that the marginal distribution $f(\boldsymbol{x}_1)$ is MVN with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{11}$ and the conditional distribution $f(\boldsymbol{x}_2|\boldsymbol{x}_1)$ is also MVN with mean vector and covariance matrix given by

$$\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \quad \text{and} \quad \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.$$

We now specialize this result to the trivariate case where two of the variables are predictor variables, x_1 and x_2, and the third variable is the response variable y. For convenience, suppose that all three variables are standardized so that they have 0 means, unit variances and their pairwise covariances are pairwise correlation coefficients $\rho_{x_1x_2}, \rho_{x_1y}$ and ρ_{x_2y}. Thus

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{x_1x_2} & \rho_{x_1y} \\ \rho_{x_1x_2} & 1 & \rho_{x_2y} \\ \rho_{x_1y} & \rho_{x_2y} & 1 \end{bmatrix}.$$

We want to derive a formula for the partial correlation coefficient $\rho_{x_2y|x_1}$. Identifying $\boldsymbol{x}_1 = x_1$ and $\boldsymbol{x}_2 = (x_2, y)'$, we have $\boldsymbol{\Sigma}_{11} = 1, \boldsymbol{\Sigma}_{12} = [\rho_{x_1x_2}, \rho_{x_1y}]$ and

$$\boldsymbol{\Sigma}_{22} = \begin{bmatrix} 1 & \rho_{x_2y} \\ \rho_{x_2y} & 1 \end{bmatrix}.$$

Hence the conditional covariance matrix of (x_2, y) conditioned on x_1 equals

$$\begin{aligned} \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} &= \begin{bmatrix} 1 & \rho_{x_2y} \\ \rho_{x_2y} & 1 \end{bmatrix} - \begin{bmatrix} \rho_{x_1x_2} \\ \rho_{x_1y} \end{bmatrix} (1)^{-1} [\rho_{x_1x_2}, \rho_{x_1y}] \\ &= \begin{bmatrix} 1 - \rho_{x_1x_2}^2 & \rho_{x_2y} - \rho_{x_1x_2}\rho_{x_1y} \\ \rho_{x_2y} - \rho_{x_1x_2}\rho_{x_1y} & 1 - \rho_{x_1y}^2 \end{bmatrix}. \end{aligned}$$

So the partial correlation coefficient is given by

$$\rho_{x_2y|x_1} = \frac{\rho_{x_2y} - \rho_{x_1x_2}\rho_{x_1y}}{\sqrt{(1-\rho_{x_1x_2}^2)(1-\rho_{x_1y}^2)}}. \tag{3.36}$$

The sample partial correlation coefficient $r_{x_2y|x_1}$ given by (3.30) is the sample version of this formula.

EXERCISES

Theoretical Exercises

3.1 (Regression through origin) Write the model $y_i = \beta x_i + \varepsilon_i \quad (i = 1, \ldots, n)$ in matrix notation. Use the formula (3.7) to derive the LS estimator of β given in Exercise 2.1.

3.2 (Relation between R^2 and R^2_{adj}) Show that

$$1 - R^2_{\text{adj}} = \left[\frac{n-1}{n-(p+1)}\right](1 - R^2),$$

and hence $R^2_{\text{adj}} \leq R^2$.

3.3 (Extra sum of squares test in terms of R^2) Let R^2_p and R^2_q denote the R^2's for the full model with p predictors and a partial model with and $q < p$ predictors. Show that the extra SS F-statistic equals

$$F = \frac{(R^2_p - R^2_q)/(p-q)}{(1 - R^2_p)/[n-(p+1)]}.$$

Suppose that $n = 26, q = 3$ and $p = 5$. Further suppose that $R^2_p = 0.90$ and $R^2_q = 0.80$. Test whether the increase in R^2 from the partial model to the full model is statistically significant at the 1% level.

3.4 (Hat matrix for simple linear regression) Show that the elements of the hat matrix $\boldsymbol{H}$ for simple linear regression are given by

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \quad \text{and} \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}} \quad (1 \leq i \neq j \leq n).$$

Further show that the $\sum_{j=1}^n h_{ij} = 1$ across every row of the hat matrix. So the fitted value $\widehat{y}_i = \sum_{j=1}^n h_{ij} y_j$ is the weighted average of all the y_j. (Hint: Use the equivalent simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $x_i \to x_i - \overline{x}$. For this model $\boldsymbol{X}'\boldsymbol{X}$ is a diagonal matrix and so $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is easy to evaluate. Note that the $\boldsymbol{H}$ matrix is the same for the two models since the fitted values obtained by the two models are the same.)

3.5 (Hat matrix for the Anscombe Data Set IV) In this problem we want to show that for the extreme observation $(x_i = 19, y_i = 12.50)$, we have $h_{ii} = 1$ and $h_{ij} = 0$ for $j \neq i$. So the fitted value of y_i at $x_i = 19$ is always equal to the observed value of y_i. Consider the problem more generally by assuming that the first $n-1$ observations all have the same $x_i = x' \ (1 \leq i \leq n-1)$ and $x_n = x''$.

a) Show that

$$\overline{x} = \frac{(n-1)x' + x''}{n} \quad \text{and} \quad S_{xx} = \left(\frac{n-1}{n}\right)(x' - x'')^2.$$

b) Next show that

$$(x_i - \overline{x})(x_n - \overline{x}) = -\frac{(n-1)(x' - x'')^2}{n^2} \quad (1 \leq i \leq n-1)$$

and

$$(x_n - \overline{x})^2 = \left(\frac{n-1}{n}\right)^2 (x' - x'')^2.$$

c) Hence conclude that $h_{in} = 0$ for $1 \leq i \leq n-1$ and $h_{nn} = 1$.

3.6 (Row and columns sums of hat matrix) Show that all rows and columns of the hat matrix for simple linear regression given in Exercise 3.4 sum to 1. This is true for

multiple regression as well as long as the constant term is included in the model but it is more challenging to prove. Try it!

3.7 (Orthogonal designs) Let the model matrix $\boldsymbol{X} = \left[\boldsymbol{X}_1 \vdots \boldsymbol{X}_2\right]$ where $\boldsymbol{X}_1$ is an $n \times p_1$ matrix and $\boldsymbol{X}_2$ is an $n \times p_2$ matrix such that $p_1 + p_2 = p$ (here we are assuming, for convenience, that there is no constant term in the model). Assume that the parameter vector $\boldsymbol{\beta}$ is similarly partitioned into two subvectors $\boldsymbol{\beta}_1$: $p_1 \times 1$ and $\boldsymbol{\beta}_2$: $p_2 \times 1$. If every column of $\boldsymbol{X}_1$ is orthogonal to every column of $\boldsymbol{X}_2$, show the following results.

a) The LS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$, respectively. Thus $\widehat{\boldsymbol{\beta}}_1$ does not depend on $\boldsymbol{X}_2$ and $\widehat{\boldsymbol{\beta}}_2$ does not depend on $\boldsymbol{X}_1$.

b) The LS estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are uncorrelated (i.e., all Corr$(\widehat{\beta}_{1j}, \widehat{\beta}_{2k}) = 0$) and so under the normality assumption they are independent.

3.8 (Omitted variables) Suppose that the true linear model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}_1 : n \times p_1, \boldsymbol{\beta}_1 : p_1 \times 1, \boldsymbol{X}_2 : n \times p_2, \boldsymbol{\beta}_2 : p_2 \times 1$. However, we mistakenly fit the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and estimate $\boldsymbol{\beta}_1$ by $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$.

a) Show that, in general, $\widehat{\boldsymbol{\beta}}_1$ is a biased estimator with

$$\text{Bias}(\widehat{\boldsymbol{\beta}}_1) = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2.$$

Note that this formula generalizes that given in Exercise 2.4.

b) Under what condition on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is $\widehat{\boldsymbol{\beta}}_1$ unbiased? How is this condition related to that in Exercise 2.4 for $\widehat{\beta}_1$ to be unbiased?

3.9 (Mean and covariance matrix of the GLS estimator) Show the results (3.28) regarding the mean and covariance matrix of the GLS estimator.

Applied Exercises

3.10 (Matrix calculation by hand) This is a hand-calculation exercise to help you get an understanding of the matrix calculations in multiple regression. Consider the following small data set. We want to fit a straight line $y = \beta_0 + \beta_1 x$ to these data.

x	1	2	3	4	5
y	2	6	7	9	10

a) Write the $\boldsymbol{X}$ matrix and the $\boldsymbol{y}$ vector.

b) Calculate $\boldsymbol{X}'\boldsymbol{X}$ and its inverse. For a 2×2 matrix, the formula for the inverse is simple:

$$\begin{bmatrix} a & c \\ d & b \end{bmatrix}^{-1} = \frac{1}{ab - cd}\begin{bmatrix} b & -c \\ -d & a \end{bmatrix}.$$

Check that the product of the original matrix and its inverse equals the identity matrix.

c) Calculate the $\boldsymbol{X}'\boldsymbol{y}$ vector.

d) Finally calculate the LS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using the formula (3.7).

3.11 (Alternate coding of categorical variables) Refer to Example 3.15 and the data in Table 3.8. Suppose that the Gender is coded as $x_1 = -1$ for females and $x_1 = +1$ for males. Similarly, Race is coded as $x_2 = -1$ for non-Whites and $x_2 = +1$ for Whites. What are the new values of $\beta_0, \beta_1, \beta_2$ and β_3? Interpret them.

3.12 (Cobb-Douglas production function) Data on 569 European companies on their capital (x_1) measured as total fixed assets (in millions of euros) at the end of 1995, labor (x_2) measured as number of workers and output (y) measured as value added (in millions of euros) are available in file `cobbdouglas.csv`. The companies in this data set are from different industry sectors in which different Cobb-Douglas production functions may apply since their capital and labor requirements are different, but we ignore this problem.

a) Fit the Cobb-Douglas production function $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$, where β_1 and β_2 are the capital and labor elasticities.

b) If $\beta_1 + \beta_2 = 1$ then it is easy to check that if capital and labor are changed by a common scaling factor then the output is changed by the same factor. In economics this is called the **constant returns to scale**. Test the null hypothesis of the constant returns to scale for these data by doing a t-test of $H_0 : \beta_1 + \beta_2 = 1$ using the estimates of $\text{Var}(\widehat{\beta}_1)$, $\text{Var}(\widehat{\beta}_2)$ and $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2)$. (These estimates can be obtained in R by using the `vcov` function.)

c) The above null hypothesis can also be tested by using the extra SS method as follows. Since the response variable must be the same for both the full model and the partial model in the extra SS method so that we can validly compare the SSE's for the two models, subtract $\ln x_2$ from both sides of the model so that the new response variable is $\ln y - \ln x_2$ and the full model is $\ln y - \ln x_2 = \ln \beta_0 + \beta_1(\ln x_1 - \ln x_2) + \beta_3 \ln x_2 + \varepsilon$ where $\beta_3 = \beta_1 + \beta_2 - 1$. Test $H_0 : \beta_3 = 0$ using the extra SS F-test and compare the result obtained in Part (b).

3.13 (Research expenditures data) Research expenditures is an important factor in the algorithm used by *US News & World* to rank graduate engineering programs. It carries 25% weight (15% for total research expenditures and 10% for research expenditures per faculty). The file `Research.csv` gives data on research expenditures in millions of $ (Research), number of faculty (Faculty) and number of PhD students (PhD) in top 30 US Universities according to *US News & World* 2017 rankings. The data are taken from ASEE profiles. We want to build a predictive model for research expenditures as a function of number of faculty and number of PhD students.

a) Make a matrix scatter plot and compute the correlation matrix of all three variables. Comment on the relationships between the variables.

b) Fit a regression model of Research versus Faculty and PhD. From this model note that PhD is a significant predictor of Research but Faculty is not. Why can't research expenditures be increased simply by increasing the number of PhD students? Given that faculty with more grants fund more PhD students (i.e., the causal arrow is Faculty $\rightarrow$ PhD) explain the apparently anomalous result obtained.

c) Calculate the partial correlation coefficients between Research and each predictor controlling for the other predictor and their t-statistics. Check that these t-statistics are the same as those given by the regression analysis.

3.14 (Sales data) Consider the following data on sales (y) of a company in 10 sales regions. The predictors are: the number of salesmen (x_1) and the amount of sales expenditures in millions of dollars (x_2).

Table 3.10 Salary Data Variables

Variable	Explanation
Salary	Annual salary in $
YrsEm	No. of years employed with the company
PriorYr	No. of years of prior experience
Educ	No. of years of education after high school
Super	No. of people supervised
Gender	M = Male, F = Female
Dept	Advertising, Engineering, Purchase, Sales

Source: McKenzie and Goldman (1999, Temco Data Set)

No.	x_1	x_2	y	No.	x_1	x_2	y
1	31	1.85	4.20	6	49	2.80	7.42
2	46	2.80	7.28	7	31	1.85	3.36
3	40	2.20	5.60	8	38	2.30	5.88
4	49	2.85	8.12	9	33	1.60	4.62
5	38	1.80	5.46	10	42	2.15	5.88

Source: Tamhane and Dunlop (2000), Example 11.7.

a) Calculate the correlation matrix $\boldsymbol{R}$ between x_1 and x_2 and the correlation vector $\boldsymbol{r}$ between y and x_1, x_2. From these bivariate correlations calculate the partial correlations $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$.

b) Calculate the standardized regression coefficients $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ from $\boldsymbol{R}$ and $\boldsymbol{r}$ for the model. How do they compare with the partial correlation coefficients $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$?

c) Check that you get the same values for $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ from the unstandardized LS estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ by scaling them appropriately. Calculate $\widehat{\beta}_0$.

d) Compare $(\widehat{\beta}_1, \widehat{\beta}_2)$ with $(\widehat{\beta}_1^*, \widehat{\beta}_2^*)$. Which variable is a better predictor of sales and why?

3.15 (Salary data) File `salaries.csv` contains data on annual salaries of 46 employees of a company and possible predictors. The variables are defined in Table 3.10. Use $\log_{10}$(Salary) as the response variable.

a) Fit a prediction model using the given data. Check that the fitted equation using Male and Purchase as reference categories for Gender and Dept categorical variables is

$$\begin{aligned}\widehat{\log_{10}(\text{Salary})} &= 4.429 + 0.0075\ \text{YrsEm} + 0.0017\ \text{PriorYr} + 0.0170\ \text{Educ} + 0.0004\ \text{Super} \\ &\quad +0.0231\ \text{Female} - 0.0388\ \text{Advert} - 0.00573\ \text{Engg} - 0.0938\ \text{Sales}.\end{aligned}$$

b) If we use Female and Sales as reference categories, what will be the new coefficients for Male and for the other three departments?

c) The coefficient of Engg is highly nonsignificant with a P-value = 0.774 in the above regression. But if Sales is used as the reference category, the coefficient

of Engg is highly significant with a P-value < 0.001. Interpret this result. If the coefficient of a dummy variable is nonsignificant, what does it tell you?

d) In the above regression, the coefficients of PriorYr and Super are nonsignificant with P-values of 0.395 and 0.631, respectively. The coefficient of Female is also nonsignificant with a P-value = 0.115 indicating nonsignificant gender difference. Drop these variables, refit the model and draw conclusions.

CHAPTER 4

MULTIPLE LINEAR REGRESSION: MODEL DIAGNOSTICS

This chapter focuses on detecting violations of standard multiple linear regression assumptions such as normality, homoscedasticity, linearity and independence. We describe methods to address these violations through data transformations and other methods. In addition to testing these model assumptions we also discuss other data problems that adversely affect regression results, namely, outliers, influential observations and multicollinearity. We discuss both graphical and formal statistical methods based on residuals.

4.1 Model Assumptions and Distribution of Residuals

The following assumptions underlie the multiple regression model, in particular about the random errors ε_i.

1. Normality: The ε_i are normally distributed.
2. Homoscedasticity: The ε_i have a constant variance σ^2.
3. Independence: The ε_i are statistically independent.
4. No outliers: No observations deviate significantly from the specified model.
5. The model is correctly specified .

These assumptions imply that the y_i are independent $N(\mu_i, \sigma^2)$ r.v.'s where $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ for $i = 1, \ldots, n$.

Under these assumptions, the distribution of the residuals has been shown (see Section 3.7) to be multivariate normal with a null mean vector and variance-covariance matrix $\sigma^2(\boldsymbol{I} - \boldsymbol{H})$, where $\boldsymbol{H} = \{h_{ij}\}$ is the hat matrix defined in (3.10). Thus each $e_i \sim N(0, \sigma^2(1 - h_{ii}))$. We will use this distributional result to test the assumptions listed above.

Note that $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ is not constant even if $\text{Var}(\varepsilon_i) = \sigma^2$ is constant. Similarly, $\text{Cov}(e_i, e_j) = \sigma^2 h_{ij}$, where the $h_{ij} \neq 0$ in general even if $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$. Thus the e_i's are not independent even if the ε_i's are. This is clear from the fact that, as noted in Section 3.2.3, the n residuals are subject to $p + 1$ linear constraints and so they are linearly dependent. Therefore, when we use residuals to test for homoscedasticity or independence of the ε_i's, we can get an anomalous result. However, if n is large relative to p then the e_i's imitate the distribution of the ε_i's closely and we can use them to test the assumptions on the ε_i's.

4.2 Checking Normality

To test the normality assumption on the ε_i's the recommended method is the normal quantile-quantile plot (called the **normal Q-Q plot** or simply the **normal plot**) of the residuals. This is a plot of the quantiles of the residuals versus theoretical standard normal distribution quantiles. Denote the ordered residuals by $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$ and define the ith ordered residual $e_{(i)}$ as the $[i/(n+1)]$th quantile of the residuals. For example, if $n = 25$ then $e_{(1)}$ is the $1/26 = 0.0385$th quantile and the corresponding standard normal distribution quantile is -1.769. The standard normal quantile axis may be labeled in terms of the cumulative standard normal probabilities going from, say, 0.001 to 0.999 (-3.090 to $+3.090$ on the quantile scale). Other definitions of a sample quantile are also used, e.g., many authors define $e_{(i)}$ as the $[(i - 0.5)/n]$th sample quantile.

The following points should be noted about normal plots.

- Standardized residuals defined in (4.5) are preferred for making a normal plot because they have a common variance equal to 1, whereas raw residuals have slightly unequal variances.
- Normality of the ε's is tested by making a normal plot of the residuals and not of the y's. The reason is that the residuals have a constant zero mean, while the means of the y's are not constant since they depend on the associated x's.
- Some typical normal plots are shown in Figure 4.1. Panel (a) shows a plot for normal data, (b) shows a plot for a right-skewed distribution such as a lognormal distribution, (c) shows a plot for a long-tailed distribution (with longer tails than the normal distribution such as a t-distribution with a small number of degrees of freedom) and (d) shows a plot for a short-tailed distribution (with shorter tails than the normal distribution such as a uniform distribution which has no tails).

Plots other than the normal plot, e.g., a histogram of the residuals, can be used but they are not as helpful in diagnosing non-normality. There are several formal normality tests, e.g., the Anderson-Darling test and the Shapiro-Wilk test. For most common applications a normal plot is usually sufficient.

It should be noted that non-normality of the data is not as serious a problem as is often perceived because the LS estimates are still approximately normally distributed for large n, thanks to the central limit theorem. Normality is a crucial assumption mainly in case of

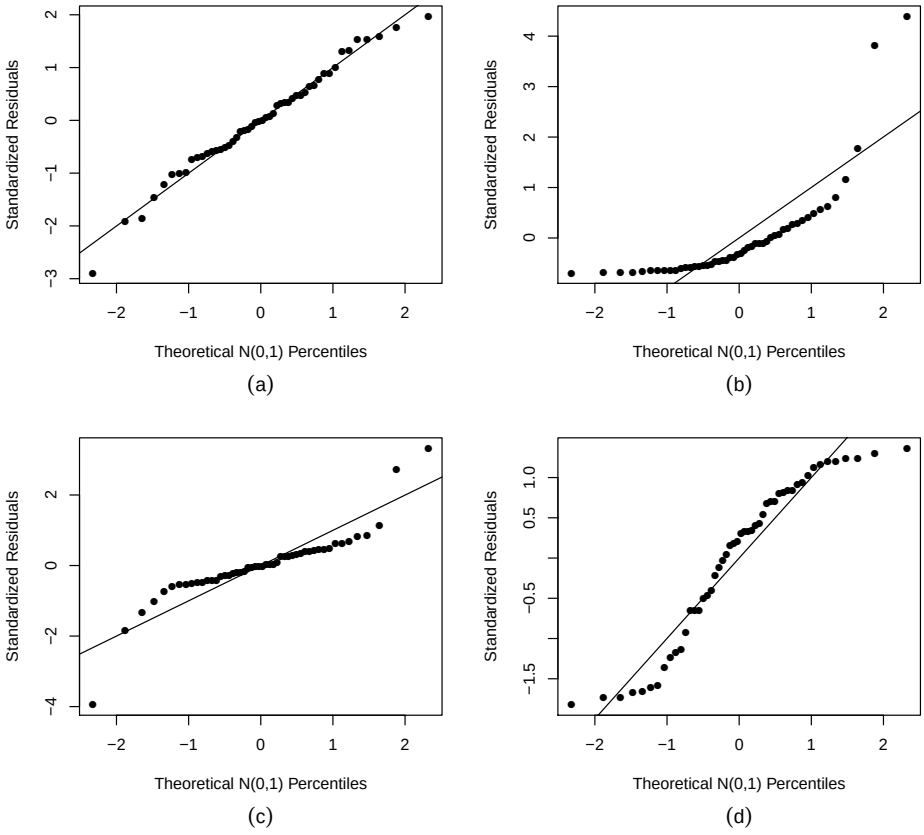


Figure 4.1 Some typical normal plots: (a) normal data, (b) right-skewed data, (c) long-tailed data, (d) short-tailed data

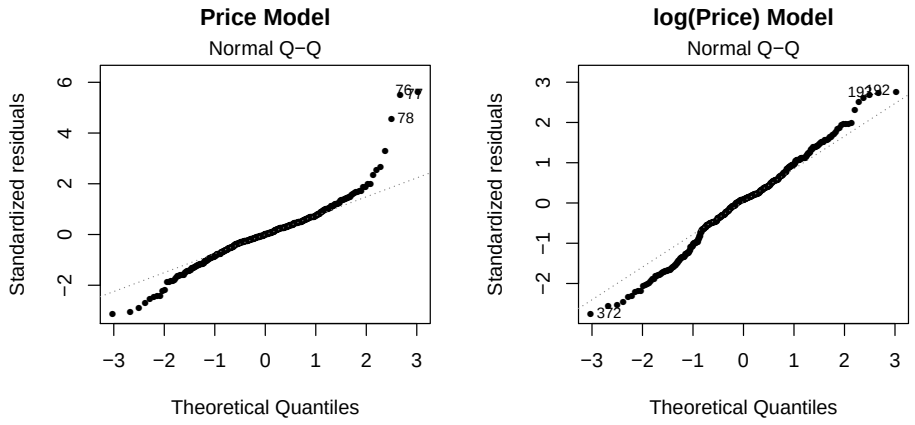


Figure 4.2 Normal plots of residuals with Price and log(Price) as the response variables

small samples in order for the normal theory inferences on the regression coefficients to be valid; the LS estimation process itself is not based on normality.

EXAMPLE 4.1 (Used Car Prices: Checking Normality)

Figure 4.2 shows the normal plots of the residuals obtained using Price and log(Price) as the response variables (with the same predictors as in Example 3.13). These plots are obtained using the following R code.

```
> carprices = read.csv("c:/data/usedcarprices.csv")
> fit1 = lm(Price ~ Mileage + Liter + factor(Make)
>   + factor(Type), carprices)
> fit2 = lm(log(Price) ~ Mileage + Liter + factor(Make)
> +factor(Type), carprices)
> plot(fit1,which=2)
> plot(fit2,which=2)
```

We see that the normal plot for Price shows more extreme departures in the upper tail than does the normal plot for log(Price). If the outliers in the upper tail are excluded then the normal plot for Price also becomes fairly linear. Thus the departures from normality are caused by outliers; the log transformation helps to mitigate those outliers. ■

4.3 Checking Homoscedasticity

Violation of the homoscedasticity assumption is a more serious problem than non-normality and can lead to invalid inferences on the regression coefficients since the test statistics incorrectly use a common pooled MSE as an estimate of the error variance. If the homoscedasticity assumption does not hold then typically $\sigma_i = \text{SD}(y_i)$ is a function of $\mu_i = E(y_i)$, which is estimated by the fitted value $\widehat{y}_i$. Therefore, to test homoscedasticity, we make the plot of the raw residuals e_i against the fitted values $\widehat{y}_i$. This is called the

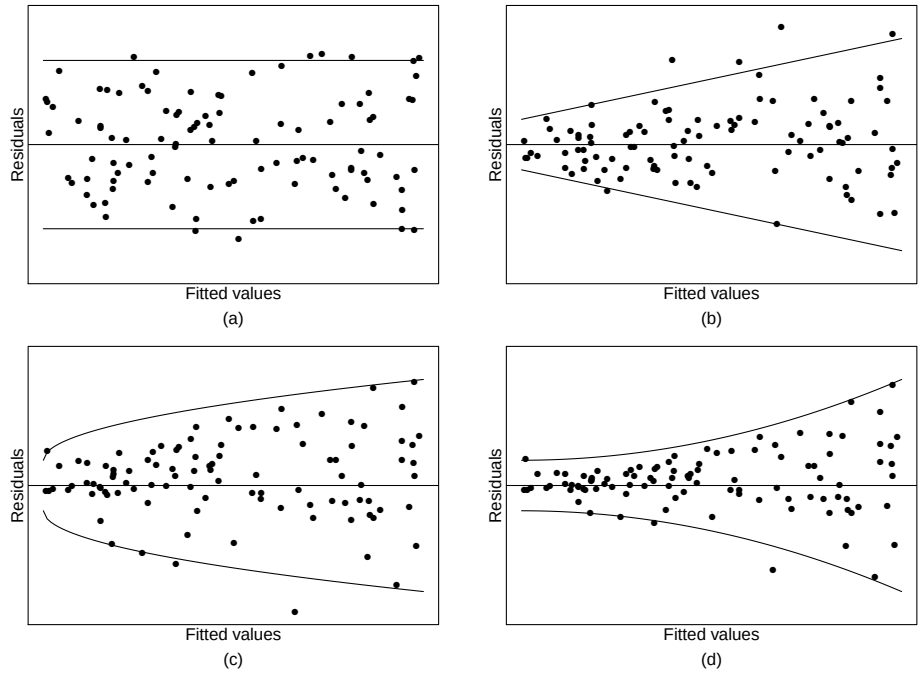


Figure 4.3 Some typical fitted values plots: (a) SD(y) is constant independent of μ, (b) SD$(y) \propto \mu$, (c) SD$(y) \propto \sqrt{\mu}$, (d) SD$(y) \propto \mu^2$

fitted values plot. Remember that the spread of the residuals is proportional to SD(e_i). If the residuals spread out evenly forming a roughly parallel band around the zero line then it indicates that the σ_i are roughly constant supporting the homoscedasticity assumption.

Note that μ_i is different for each observation and hence if σ_i is a function of μ_i then it is different for each observation thus resulting in heteroscedasticity. For example, the variability in household incomes increases with the mean income levels. If the incomes follow a lognormal distribution then σ_i is proportional to μ_i. In that case the residuals fan out approximately linearly in a funnel shape. In general, the shape of the plot tells us something about the relationship between σ_i and μ_i. Some typical fitted values plots are shown in Figure 4.3 with the corresponding relations between σ_i and μ_i. The transformations of y suggested by these relations to stabilize their variances are discussed in the next section.

4.3.1 Variance Stabilizing Transformations

Suppose that $\sigma = \text{SD}(y)$ is a known smooth function of $\mu = E(y)$, say $\sigma = g(\mu)$, and we want to find a transformation $f(y)$ such that SD$[f(y)]$ is approximately constant. Consider the first-order Taylor series expansion of $f(y)$ around μ, namely $f(y) \approx f(\mu) + (y - \mu)f'(\mu)$, where $f'(\mu)$ is the first derivative of $f(\mu)$. Then Var$(f(y))$ can be approximated as

$$\text{Var}[f(y)] \approx \text{Var}(y)[f'(\mu)]^2. \tag{4.1}$$

This is called the **delta method**. We want Var$[f(y)]$ to be constant, so we set it equal to 1 (or any other positive constant). Then putting Var$(y) = g^2(\mu)$, we get the equation $g^2(\mu)[f'(\mu)]^2 = 1$, which upon solving for $f(\mu)$ and changing the variable of integration

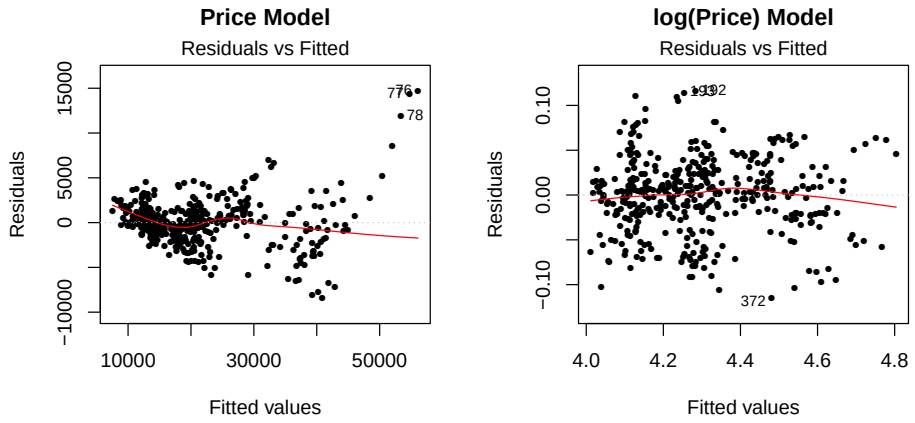


Figure 4.4 Fitted values plots of residuals with Price and log(Price) as the response variables

from μ to y gives

$$f(y) = \int \frac{dy}{g(y)}. \tag{4.2}$$

This is the desired **variance stabilizing transformation** $f(y)$.

Consider the example of lognormal data where $\text{SD}(y)$ is proportional to μ, i.e., $g(\mu) = c\mu$ where $c > 0$. The corresponding fitted values plot is shown in panel (b) of Figure 4.3. Ignoring the constant of proportionality c we get

$$f(y) = \int \frac{dy}{y} = \ln y,$$

which is the **logarithmic transformation**. If y has a lognormal distribution then, of course, $\ln y$ has a normal distribution with a constant variance.

For another example, count data (e.g., the number of calls received at a call center in one hour) are often modeled by a Poisson distribution for which we have $E(y) = \mu$ and $\text{SD}(y) = \sqrt{\mu}$. The corresponding fitted values plot is shown in panel (c) of Figure 4.3. Then

$$f(y) = \int \frac{dy}{\sqrt{y}} = 2\sqrt{y},$$

which is the **square-root transformation**.

Still another example is the **inverse transformation**, $f(y) = y^{-1}$, which arises when $\sigma = g(\mu) \propto \mu^2$. The corresponding fitted values plot is shown in panel (d) of Figure 4.3.

EXAMPLE 4.2 (Used Car Prices: Checking Homoscedasticity)

Continuing with Example 4.1, we give the fitted values plots for Price and log(Price) as the response variables in Figure 4.4. These plots are produced using the same R code as in Example 4.1 except using `which=1` in the `plot` function. Notice that the plot for Price is funnel-shaped indicating variance increasing with mean while the plot for log(Price) exhibits a random parallel band. Thus log transformation stabilizes the variance in this example. ■

4.3.2 Box-Cox Transformation

The logarithmic, square root and inverse transformations are special cases of a general family of power transformations defined by Box and Cox (1964):

$$f_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln y & \text{if } \lambda = 0. \end{cases} \tag{4.3}$$

Here λ is a parameter to be determined. Note that the log transformation is obtained as the limit of $f_\lambda(y)$ as $\lambda \to 0$ by using l'Hospital's rule.

How to choose λ? One can fit models using different transformations $f_\lambda(y)$ for a selected grid of λ-values, calculate SSE for each model and then choose λ that minimizes SSE. However, the SSE's for different λ are not comparable because they are not scale-free. To address this issue the following modification of (4.3) was suggested by Box and Cox (1964):

$$f^*_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda \widetilde{y}^{\lambda-1}} & \text{if } \lambda \neq 0 \\ \widetilde{y} \ln y & \text{if } \lambda = 0, \end{cases} \tag{4.4}$$

where $\widetilde{y} = \left(\prod_{i=1}^n y_i\right)^{1/n}$ is the geometric mean of the y_i's (assumed to be all positive). With this modification the SSE's for different λ can be compared with each other.

Usually λ-values in the range $[-1, +1]$ are of interest. Note that for given λ, $f^*_\lambda(y)$ is just a multiple of $f_\lambda(y)$ and the multiplication factor $1/(\widetilde{y})^{\lambda-1}$ is common to all observations. Therefore the two transformations are equivalent as response variables. So regression can be performed using $f_\lambda(y)$ as the response variable instead of $f^*_\lambda(y)$ after the minimizing value of λ has been determined.

EXAMPLE 4.3 (Textile Experiment Data: Finding the Best λ)

Box and Cox (1964) gave data from a 3^3 factorial experiment to study the effects of three factors on the cycles to failure (y) of worsted yarn. The three factors and their experimental levels were:

x_1: length of test specimen (250 mm, 300 mm, 350 mm)

x_2: amplitude of loading cycle (8 mm, 9 mm, 10 mm)

x_3: load (40 gm, 45 gm, 50 gm).

For convenience, we have coded the levels of all three factors as $-1, 0, +1$ in Table 4.1.

As recommended by Box and Cox, the following log-log model was fitted to the data:

$$\widehat{\log y} = 1.6784 + 4.9504 \log x_1 - 5.6537 \log x_2 - 3.5030 \log x_3;$$

here x_1, x_2, x_3 are in their original units. All the regression coefficients are highly significant. The fitted values and the normal plots are shown in Figure 4.5. These two plots are produced using `which=1:2` in the R function `plot`.

The plots look satisfactory, so we decide to stay with this model. It may be of interest to check whether the log transformation ($\lambda = 0$) of the response is in fact the best choice. For this purpose we computed the transformation (4.4) for $\lambda = -1$ to $\lambda = +1$ in steps of 0.1 and regressed them on $\log x_1, \log x_2$ and $\log x_3$. The resulting SSE's are plotted in Figure 4.6. We see that SSE is minimized at $\lambda = 0$ (with value equal to 0.1448) which is thus indeed the best choice. ■

Table 4.1 Cycles of failure of worsted yearn from a 3^3 factorial experiment

x_1	x_2	x_3	y	x_1	x_2	x_3	y	x_1	x_2	x_3	y
-1	-1	-1	674	0	-1	-1	1414	+1	-1	-1	3636
-1	-1	0	370	0	-1	0	1198	+1	-1	0	3184
-1	-1	+1	292	0	-1	+1	634	+1	-1	+1	2000
-1	0	-1	338	0	0	-1	1022	+1	0	-1	1568
-1	0	0	266	0	0	0	620	+1	0	0	1070
-1	0	+1	210	0	0	+1	438	+1	0	+1	566
-1	+1	-1	170	0	+1	-1	442	+1	+1	-1	1140
-1	+1	0	118	0	+1	0	332	+1	+1	0	884
-1	+1	+1	90	0	+1	+1	220	+1	+1	+1	360

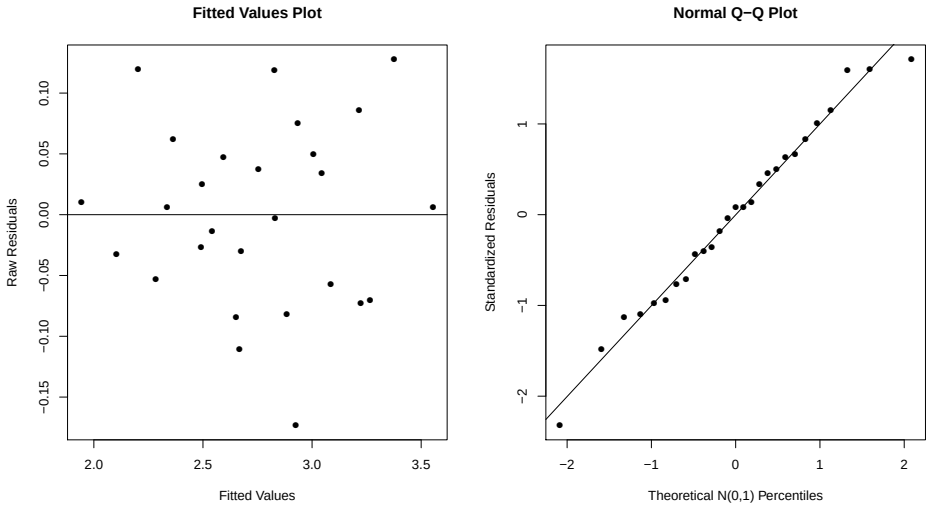


Figure 4.5 Fitted value and normal plots of residuals from the log-log model for textile data

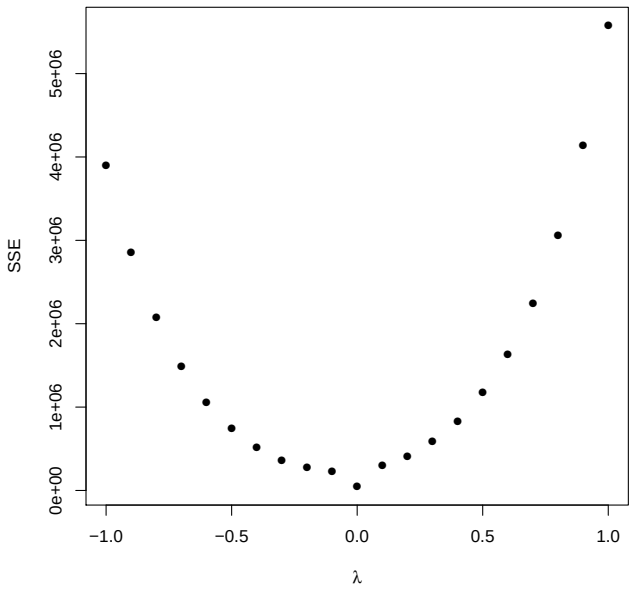


Figure 4.6 Plot of SSE versus λ

4.4 Checking Outliers

Outliers are observations that deviate significantly from the fitted model, e.g., by more than two or three standard deviations. From the properties of the residuals given in Section 4.1 it follows that $\text{SE}(e_i) = s\sqrt{1-h_{ii}}$, where $s = \sqrt{\text{MSE}}$. The standardized residual

$$e_i^* = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (1 \le i \le n) \tag{4.5}$$

is used to test if the ith observation is an outlier by checking if e_i^* exceeds a specified critical constant, e.g., 2, in absolute value.

Outliers can significantly affect the fit of the model. So it is not correct to use an observation to fit the model and also test its residual for outlierness. A better way is to fit the model by deleting that observation and then testing its residual. Denote by $\widehat{y}_{i(i)}$ the fitted value of y_i after omitting it when fitting the model. Then the **deleted residual** is defined as $e_{(i)} = y_i - \widehat{y}_{i(i)}$ $(i = 1, \ldots, n)$.

In fact, it is not necessary to fit n separate models to compute n deleted residuals by omitting one observation at a time. They can be computed from the regular residuals by using the following formula:

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \ (i = 1, \ldots, n). \tag{4.6}$$

Since $e_{(i)}$ is a scaled multiple of e_i, it follows that $\text{SE}(e_{(i)}) = \text{SE}(e_i)/(1-h_{ii})$ and so the standardized deleted residual $e_{(i)}^*$ is the same as the standardized regular residual e_i^*.

In (4.5), s is the usual RMSE computed from all residuals including e_i. Therefore, if the ith observation is an outlier then e_i will be large and will inflate $s^2 = \text{MSE}$. This will deflate the standardized residual e_i^* making it less likely to be detected as an outlier. To

overcome this problem, it has been suggested that s^2 should also be computed by deleting the ith observation, which we denote by $s^2_{(i)}$. Thus, instead of using e^*_i given by (4.5), we use

$$r^*_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}. \tag{4.7}$$

The e^*_i given by (4.5) are called **internally studentized residuals** while the r^*_i given by (4.7) above are called **externally studentized residuals**. The two are related by

$$r^*_i = e^*_i\sqrt{\frac{n-p-2}{n-p-1+e^{*2}_i}}. \tag{4.8}$$

Once again, it is not necessary to fit n separate models to compute the n separate estimates $s^2_{(i)}$. They can be computed from s^2 for the single model estimated from all observations by using the following formula:

$$s^2_{(i)} = s^2\left[\frac{(n-p-1)-e^{*2}_i}{n-p-2}\right] \quad (i=1,\ldots,n). \tag{4.9}$$

Graphical plots are also useful for identifying outliers as we have seen from Figure 4.2 where departures of a few points from a linear normal plot indicate outliers. A plot of observed y_i's versus fitted $\widehat{y}_i$'s is useful for this purpose.

Outliers (and influential observations discussed later in Section 4.7) should not be deleted without additional inspection. First they must be checked for validity and should be deleted only if they are erroneous. If they are valid observations then they may indicate model misspecification. For example, we may be fitting a straight line to data that actually follow a quadratic or an exponential model. Thus an outlier may be useful for revealing a misspecified model, which is the next topic of discussion.

EXAMPLE 4.4 (College GPA and Entrance Test Scores: Checking Outliers)

Refer to Example 3.3 where we fitted the model

$$\widehat{\text{GPA}} = -1.5705 + 0.0257\text{Verbal} + 0.0336\text{Math}.$$

Figure 4.7 gives the sequence plot of the residuals. These plots are obtained by specifying `which=4` in the `plot` function. We see that observation $\sharp 4$ appears to be an outlier. The same observation appears to be an outlier in the normal plot of the residuals. Let us check this by calculating the corresponding standardized residual. The fitted value for this observation is

$$\widehat{y}_4 = -1.5705 + 0.0257 \times 100 + 0.0336 \times 49 = 2.650,$$

so the residual is

$$e_4 = 1.54 - 2.650 = -1.110.$$

Now $s^2 = 0.1618$ from the ANOVA Table 3.4 and the leverage $h_{44} = 0.1784$ from Table 4.3. Therefore the internally studentized residual equals

$$e^*_4 = \frac{-1.110}{\sqrt{0.1618(1-0.1784)}} = -3.044,$$

which is clearly significant.

Next we will calculate the externally studentized residual. We first calculate $s^2_{(4)}$ using (4.9):

$$s^2_{(4)} = 0.1618\left[\frac{37-(-3.044)^2}{36}\right] = 0.1246.$$

Hence $r^*_4 = -1.110/\sqrt{0.1246(1-0.1784)} = -3.469$, which is even more significant than the internally studentized residual, as expected. ■

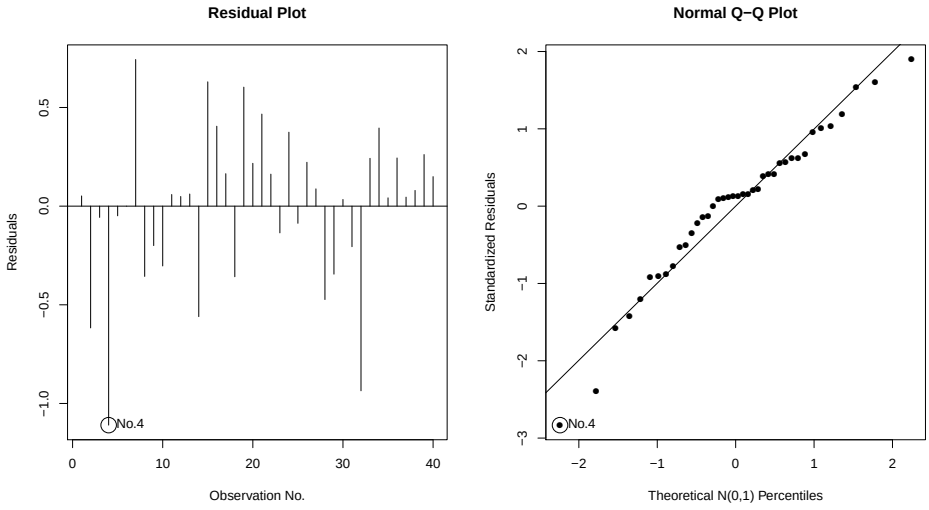


Figure 4.7 Sequence plot and normal plot of residuals for the college GPA data

4.5 Checking Model Misspecification

Models can be misspecified in any number of ways. So it is not possible to recommend a single diagnostic that will detect any type of misspecification. The most common type of misspecification is nonlinearity. One can plot residuals versus each x_j to see if they are randomly distributed or exhibit some pattern indicating departure from the assumed linearity in x_j. However, this plot could be influenced by possible nonlinearities in other predictors. Therefore better plots are needed.

The **added variables (AV) plot** (also known as the **partial regression plot**) removes the effects of other predictors by computing two sets of residuals. The first set of residuals, called the y**-residuals**, are the usual residuals, obtained by regressing y on all other predictors except x_j. The second set of residuals, called the x_j**-residuals**, are obtained by regressing x_j on all other predictors. Then the two sets of residuals are plotted against each other. If the relationship between y and x_j is linear then this plot is roughly linear. In fact, $\widehat{\beta}_j$ is the slope coefficient of the LS fit to this plot. If the relationship is nonlinear then the corresponding pattern is reflected in the plot. This plot can also be used for determining whether a predictor not currently in the model should be added to the model. If the partial regression plot for that predictor is random then that predictor should not be added. On the other hand, if the plot shows a clear trend, e.g., linear, then that variable should be added.

A variation on the added variables plot is the **component plus residuals (CR) plot** (also known as the **partial residual plot**) where the idea is to plot the **partial residuals**, defined as $e_i + \widehat{\beta}_j x_{ij}$, versus x_{ij} for $i = 1, \ldots, n$. Here the e_i are the usual residuals from the full model with all predictors and $\widehat{\beta}_j$ is the regression coefficient of x_j from the same model. Note that the partial residuals are the regular residuals with the effect of the predictor x_j

filtered out as can be seen from the following:

$$\begin{aligned} e_i + \widehat{\beta}_j x_{ij} &= y_i - \widehat{y}_i + \widehat{\beta}_j x_{ij} \\ &= y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}] + \widehat{\beta}_j x_{ij} \\ &= y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_{j-1} x_{i,j-1} + \widehat{\beta}_{j+1} x_{i,j+1} + \cdots + \widehat{\beta}_p x_{ip}]. \end{aligned}$$

If the plot is linear then no nonlinear terms in x_j should be added to the model. The interpretation of the component plus residuals plot is similar to that of the added variables plot but it is more effective in displaying the relationship of y with each x_j since the plot is made against x_j.

EXAMPLE 4.5 (College GPA and Entrance Test Scores: AV and CR Plots)

Consider the model of GPA versus Verbal and Math scores fitted in Example 3.3. We want to check whether linear terms in Verbal and Math are adequate or should quadratic terms be added to the model. We assess this graphically using the AV plots and the CR plots for the two variables shown in Figures 4.8 and 4.9. The following R code produce these plots.

```
> library("car")
> gpa = read.csv("c:/data/GPA.csv")
> fit = lm(GPA ~ Verbal + Math, data = gpa)
> crPlots(fit,"Verbal")
> crPlots(fit,"Math")
> avPlot(fit,"Verbal")
> avPlot(fit,"Math")
```

Both these plots show quadratic patterns, so quadratic terms in Math and Verbal should be added to the linear model. ■

4.6 Checking Independence

4.6.1 Tests for Independence

Lack of independence among the ε_i's typically occurs in **time-series data**. Such data are said to be autocorrelated or serially correlated. To emphasize that the observations are taken over time, we use t as the index of the observation instead of i ($t = 1, 2, \ldots, n$). For a stationary process, the first-order **autocorrelation coefficient** is defined as $\phi = \text{Corr}(y_{t-1}, y_t) = \text{Corr}(\varepsilon_{t-1}, \varepsilon_t)$. If $\phi > 0$ then the time-series of the observations is said to be positively autocorrelated while if $\phi < 0$ then it is said to be negatively autocorrelated. In time-series literature the ε_t's are called **disturbances**.

The LS estimators $\widehat{\beta}_j$ are unbiased under autocorrelation but their variances are different from those under independence. Under positive autocorrelation, the true variances of the $\widehat{\beta}_j$ are higher. Also, the MSE underestimates σ^2. Therefore the estimated variances of $\widehat{\beta}_j$ are under-biased and so the $\widehat{\beta}_j$ appear more significant than they actually are.

Runs Test

Independence can be assessed by making a **run chart** of the residuals, i.e., by plotting e_t versus t. The number of runs of positive and negative residuals can be used to test

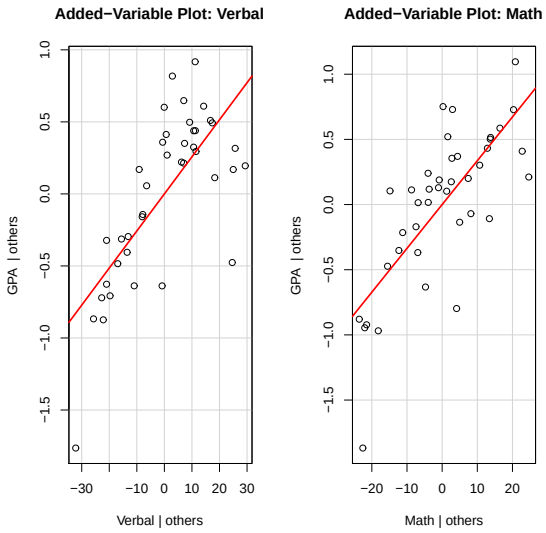


Figure 4.8 Added variables plots for Verbal and Math scores

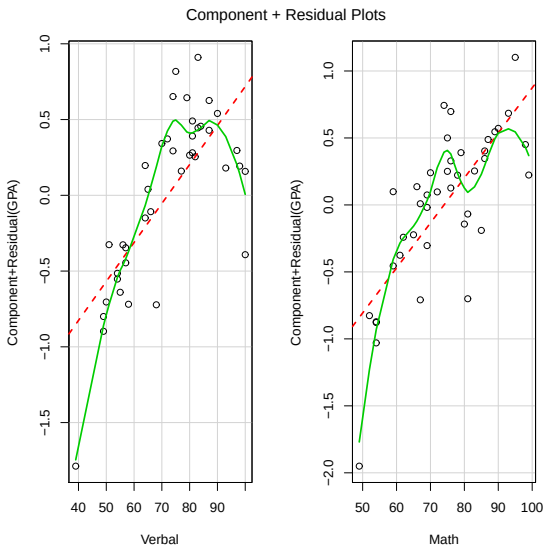


Figure 4.9 Component plus residuals plots for Verbal and Math scores

independence. A run is defined as a sequence of like-signed residual. For example, the sequence $+,+,-,-,-,+,+$ has two runs of $+$'s and one run of $-$'s for a total of three runs. If $\phi > 0$ then there will be too few runs (less than expected under independence) resulting in relatively few long cycles since each residual will tend to be followed by a like-signed residual. On the other hand, if $\phi < 0$ then there will be too many runs resulting in a zig-zag pattern since each residual will tend to be followed by an opposite signed residual.

How do we know that the number of runs is too few or too many? To answer this question we can perform the runs test. Denote the total number runs by R, the number of $+$ signs by n_1 and the number of $-$ signs by n_2, where $n_1 + n_2 = n$. Then under the null hypothesis of independence and conditioned on n_1 and n_2, it can be shown that

$$E(R) = \frac{2n_1n_2}{n} + 1, \quad \text{and} \quad \text{Var}(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}. \tag{4.10}$$

For large n_1, n_2, the statistic

$$z = \frac{R - E(R)}{\sqrt{\text{Var}(R)}}$$

can be used as a standard normal test statistic. So we can reject H_0 in favor of the alternative hypothesis $H_1 : \phi > 0$ at level α if $z < -z_\alpha$. Similarly, we can reject H_0 in favor of the alternative hypothesis $\phi < 0$ at level α if $z > z_\alpha$.

Durbin-Watson Test

This test assumes that the ε_t follow the so-called **first order autoregressive (AR(1)) model**:

$$\varepsilon_t = \phi\varepsilon_{t-1} + \eta_t \quad (t = 2, \ldots, n), \tag{4.11}$$

where the η_t are i.i.d. $N(0, \sigma_0^2)$ r.v.'s. Then it follows that $\text{Corr}(\varepsilon_{t-1}, \varepsilon_t) = \phi$. Further it is easy to show that the ε_t are $N(0, \sigma^2)$ r.v.'s with $\sigma^2 = \sigma_0^2/(1 - \phi^2)$.

For $n \gg p$, an approximate sample estimate of ϕ is given by

$$\widehat{\phi} \approx \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \tag{4.12}$$

The Durbin-Watson statistic equals

$$\begin{aligned} d &= \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \\ &= \frac{\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \\ &\approx 2(1 - \widehat{\phi}). \end{aligned} \tag{4.13}$$

Note that $0 \le d \le 4$ with $d < 2$ if $\widehat{\phi} > 0$, $d > 2$ if $\widehat{\phi} < 0$ and $d \approx 2$ if $\widehat{\phi} \approx 0$.

The Durbin-Watson test uses two critical constants, d_L and d_U, which depend on p, n and α. These constants are given in Table A.5. If the alternative hypothesis is $\phi > 0$, the test operates as follows:

1. If $d < d_L$, reject H_0 and conclude that $\phi > 0$.
2. If $d > d_U$, do not reject H_0 and conclude that there is not sufficient evidence to conclude that $\phi > 0$.
3. If $d_L \le d \le d_U$, the test is inconclusive.

If the alternative hypothesis is $\phi < 0$ then simply transform $d \to 4 - d$ and apply the same test. Note that this test has three possible decisions unlike usual hypothesis tests that involve only two decisions (reject H_0 or do not reject H_0).

4.6.2 Data Transformation to Remove First-Order Autocorrelation

First-order autocorrelations in an AR(1) model can be removed by filtering out the correlations induced by the previous observations. As an example, consider the simple linear regression model $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, where the ε_t's follow the AR(1) model (4.11). Then

$$y_t - \phi y_{t-1} = \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + \eta_t,$$

where the η_t are i.i.d. $N(0, \sigma_0^2)$. Define

$$y_t^* = y_t - \phi y_{t-1}, x_t^* = x_t - \phi x_{t-1}, \beta_0^* = \beta_0(1-\phi) \quad \text{and} \quad \beta_1^* = \beta_1.$$

Then the above model becomes

$$y_t^* = \beta_0^* + \beta_1^* x_t^* + \eta_t.$$

This model satisfies the LS assumptions, in particular, the disturbances η_t are i.i.d. So we can compute the usual LS estimates $\widehat{\beta}_0^*$ and $\widehat{\beta}_1^*$ from which we can compute $\widehat{\beta}_0 = \widehat{\beta}_0^*/(1-\phi)$ and $\widehat{\beta}_1 = \widehat{\beta}_1^*$. This procedure readily extends to multiple predictors.

To implement this procedure we need to know ϕ. We can use the estimate $\widehat{\phi}$ given by (4.12). But $\widehat{\phi}$ depends on the residuals e_t, to compute which we need to first fit a model. Therefore an iterative procedure must be used as described below.

Cochrane-Orcutt Procedure

1. Start by fitting a regression model between y_t and x_t assuming independence. Compute the residuals e_t and $\widehat{\phi}$.
2. Using the estimate $\widehat{\phi}$, compute $x_t^* = x_t - \widehat{\phi} x_{t-1}$ and $y_t^* = y_t - \widehat{\phi} y_{t-1}$.
3. Fit a regression model between y_t^* and x_t^*, and compute new residuals e_t^*. Stop if they are uncorrelated; otherwise re-estimate $\widehat{\phi}$ and return to Step 2.

Usually only one or two iterations of this procedure are necessary.

EXAMPLE 4.6 (GDP and Industrial Production:Time Series Regression)

Table 4.2 gives quarterly data on gross domestic product (GDP) in billions of dollars (adjusted for seasonality and inflation) from 2009 to 2014. The GDP data are from the Bureau of Economic Analysis website. GDP increased almost linearly during this period; so time explains nearly 99% of the variation in GDP. However, this linear trend may not continue in future. Therefore using time as a sole predictor may not give a good predictive model. In this example we use industrial production index (IPI) as a predictor (2007=100), which also increased linearly during this period. Seasonally adjusted IPI data are from the Federal Reserve Bank website.

The fitted model is $\widehat{\text{GDP}}_t = 6419.2 + 92.584 \times \text{IPI}_t$. The residual plot versus time is shown in the left panel of Figure 4.10. The plot appears to show positive autocorrelation. The sample first-order autocorrelation coefficient is $\widehat{\phi} = 0.3056$. So the Durbin-Watson statistic is $d \approx 2(1 - 0.3056) = 1.389$. We find $d_L = 1.27$ and $d_U = 1.45$ from Table A.5 for $n = 24, p = 1$ and $\alpha = 0.05$. Since d falls between d_L and d_U, the test is inconclusive. We can also do the runs test as follows. The number of runs equals $R = 8$. The number of positive signs equals $n_1 = 9$ and the number of negative signs equals $n_2 = 15$. So using (4.10) we get $E(R) = 12.25$ and $\text{Var}(R) = 5.0136$. Hence the runs test statistic is

$$z = \frac{8 - 12.25}{\sqrt{5.0136}} = -1.90,$$

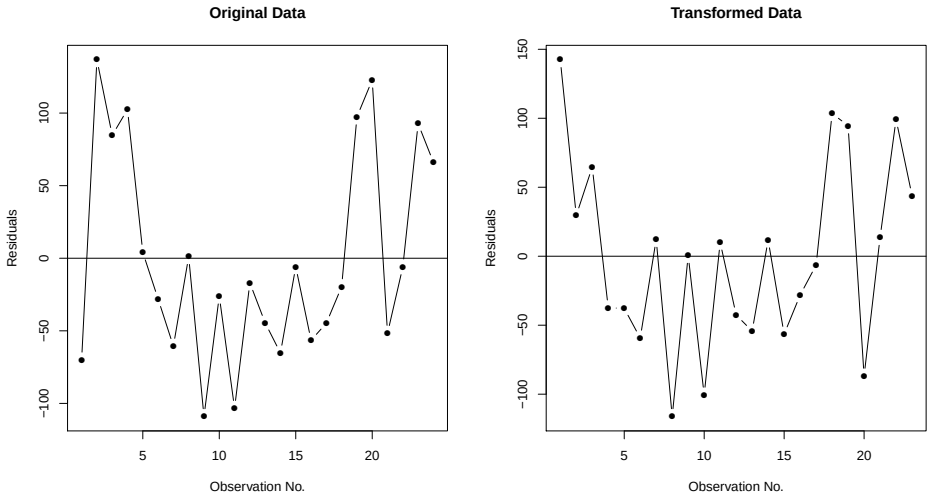


Figure 4.10 Run charts of residuals for GDP versus IPI regression for the raw data (left) and transformed data (right)

which has a two-sided P-value equal to 0.057.

We can apply the Cochrane-Orcutt procedure by using the transformation $\text{GDP}_t^* = \text{GDP}_t - 0.3056 \times \text{GDP}_{t-1}$ and $\text{IPI}_t^* = \text{IPI}_t - 0.3056 \times \text{IPI}_{t-1}$. The fitted model to the transformed data is $\widehat{\text{GDP}}_t^* = 4550.3 + 91.250 \times \text{IPI}_t^*$. The residual plot for this fit is shown in the right panel of Figure 4.10. The residuals appear to be randomly distributed with the first-order autocorrelation coefficient equal to $\widehat{\phi} = 0.1139$, a significant reduction from $\widehat{\phi} = 0.3056$. The number of runs is $R = 13$, which is close to the average and so is not statistically significant. Thus this model appears to be satisfactory.

We next calculate the estimated regression coefficients for the original model for GDP versus IPI as

$$\widehat{\beta}_0 = \frac{\widehat{\beta}_0^*}{1 - \widehat{\phi}} = \frac{4550.3}{1 - 0.1139} = 5135.20 \quad \text{and} \quad \widehat{\beta}_1 = \widehat{\beta}_1^* = 91.250.$$

■

We conclude this section with two brief remarks.

- In many applications response at time t is influenced by predictor values at previous times, e.g., $t-1$, $t-2$, etc. In such cases **lagged variables**, x_{t-1}, x_{t-2} etc. are used as predictors.

- Seasonality and cyclical variations are other common features of time series data. Such features can be modeled using a combination of sinusoidal functions. Often the period of cyclical variation is known, e.g., monthly or quarterly, which can be incorporated in the sinusoidal functions.

Table 4.2 Quarterly GDP and IPI Data from 2009 to 2014

Year	Quarter	GDP	IPI	Year	Quarter	GDP	IPI
2009	1	14375.0	86.689	2012	1	15275.0	96.132
2009	2	14355.6	84.243	2012	2	15336.7	97.024
2009	3	14402.5	85.313	2012	3	15431.3	97.402
2009	4	14541.9	86.626	2012	4	15433.7	97.976
2010	1	14604.8	88.368	2013	1	15538.4	98.980
2010	2	14745.9	90.240	2013	2	15606.6	99.445
2010	3	14845.5	91.667	2013	3	15779.9	100.053
2010	4	14939.0	92.010	2013	4	15916.2	101.250
2011	1	14881.3	92.578	2014	1	15831.7	102.223
2011	2	14989.6	92.850	2014	2	16010.4	103.659
2011	3	15021.1	94.026	2014	3	16205.6	104.701
2011	4	15190.3	94.920	2014	4	16311.6	106.134

4.7 Checking Influential Observations

The idea of fitting a model is to capture the overall pattern of variation in the response variable as a function of the predictor variables. So the fit of the model should be determined by the majority of the data and not by a few so-called **influential observations** (also called **high leverage observations**). An extreme example of an influential observation is provided by the Anscombe Data Set IV given in Table 2.3 and plotted in Figure 2.4. It is obvious that $(x = 19, y = 12.50)$ is an influential observation since it alone determines the slope of the LS line as the line must pass through it and the midpoint of the remaining 10 observations, all at $x = 8$.

4.7.1 Leverage

How can we define undue influence and how can we detect influential observations? Recall that the fitted or predicted vector is given by $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H}$ is the hat matrix. So the ith fitted value is given by

$$\widehat{y}_i = h_{i1}y_1 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n = \boldsymbol{h}_i'\boldsymbol{y},$$

where h_{ij} is the (i, j)th entry of $\boldsymbol{H}$ and $\boldsymbol{h}_i'$ is the ith row vector of $\boldsymbol{H}$. If the constant term is included in the model, it can be shown that the h_{ij} 's sum to 1 across each row. Thus $\widehat{y}_i$ is a weighted average of all observations y_j and h_{ii} is the weight on y_i in its own fitted value. If h_{ii} is too large then y_i can be said to have too much influence on its own fitted value. We refer to h_{ii} as the **leverage** of the observation. For the observation $(x = 19, y = 12.50)$ in the Anscombe data set IV, it can be shown that $h_{ii} = 1$ and all other h_{ij}'s are zero (see Exercise 3.5). So the fitted $\widehat{y}_i$ is identically equal to the observed y_i.

How large must the leverage be in order for the observation to be regarded influential? To answer this question, h_{ii} is compared with the average of the diagonal elements of $\boldsymbol{H}$. In Section 3.7 we have shown that the trace of $\boldsymbol{H}$ is $p + 1$. So the average of the diagonal

elements of $\boldsymbol{H}$ is $(p+1)/n$. A rule of thumb is to declare the ith observation as influential if h_{ii} exceeds twice this average, i.e., if

$$h_{ii} > \frac{2(p+1)}{n}. \tag{4.14}$$

Note that if $2(p+1)/n > 1$, then this rule is not applicable since it can be shown that all $h_{ii} \leq 1$.

4.7.2 Cook's Distance

In general, leverage identifies those observations as influential that are outliers in the x-space. But the observations can be influential also because they are outliers in the y-space. Cook's distance takes the effects of both these outliers into account.

Cook's distance for the ith observation measures the effect of deleting that observation on the fitted values of all observations. Let $\widehat{y}_{j(i)}$ denote the fitted value of y_j based on the regression model when the ith observation is omitted. Then Cook's distance for the ith observation is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_{j(i)})^2}{s^2(p+1)}. \tag{4.15}$$

D_i measures the total amount by which all fitted values change when the ith observation is deleted. The denominator $s^2(p+1)$ is just a scaling factor.

It can be shown that

$$D_i = \left(\frac{e_i^*}{\sqrt{p+1}}\right)^2 \left(\frac{h_{ii}}{1-h_{ii}}\right). \tag{4.16}$$

We see that D_i combines the outlierness of the ith observation in the y-space through the first term, which involves the standardized residual e_i^*, and in the x-space through the second term, which is an increasing function of the leverage h_{ii}.

Another interpretation of Cook's distance can be obtained as follows. Denote by $\widehat{\boldsymbol{\beta}}_{(i)}$ the LS estimator of $\boldsymbol{\beta}$ when the ith observation is deleted and let $\widehat{\boldsymbol{y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)}$ the corresponding fitted vector. Then the quantity $\sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_{j(i)})^2$ in the numerator of (4.15) equals

$$\begin{aligned}(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}_{(i)})'(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}_{(i)}) &= (\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)}) \\ &= (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}).\end{aligned}$$

So another formula for Cook's distance is

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(p+1)s^2}. \tag{4.17}$$

Comparing this formula with that for the $100(1-\alpha)\%$ confidence ellipsoid for the parameter vector $\boldsymbol{\beta}$ given by Equation (3.18), a possible decision rule for deciding the ith observation as influential is if $\widehat{\boldsymbol{\beta}}_{(i)}$ falls outside this ellipsoid for some chosen confidence level and hence can be regarded as a significant change in the LS estimate of $\boldsymbol{\beta}$. This is equivalent to $D_i > f_{p+1,n-(p+1),\alpha}$. Montgomery, Peck and Vining (2012, p. 216) have suggested using a confidence level of 10%-20%, which may seem very low but keep in mind that the goal here is to identify influential observations — not to estimate $\boldsymbol{\beta}$ with high level of confidence.

EXAMPLE 4.7 (College GPA and Entrance Test Scores: Checking Influential Observations)

Again refer to Example 3.3. We now check for influential observations by computing the leverage and Cook's distance values for the 40 observations which are given in Table 4.3. The threshold for the leverage is $2(p+1)/n = 6/40 = 0.15$. We see that observations 4 and 31 exceed this threshold; observation 32 falls slightly short. For Cook's distance, using $1-\alpha = 0.1$ (i.e., 10% confidence elliposid) we have $f_{3,37,0.9} = 0.1937$, which is exceeded by observations 4 and 32; observation 31 falls significantly short. Thus we may regard observations 4 and 32 as influential. ■

Table 4.3 Influence statistics for regression of GPA on entrance test scores

No.	h_{ii}	D_i	No.	h_{ii}	D_i	No.	h_{ii}	D_i
1	0.0613	0.0004	15	0.0381	0.0337	29	0.0963	0.0289
2	0.1178	0.1187	16	0.0429	0.0158	30	0.0679	0.0002
3	0.0647	0.0005	17	0.0369	0.0022	31	0.1505	0.0181
4	0.1784	0.6703	18	0.1178	0.0399	32	0.1359	0.3283
5	0.0657	0.0004	19	0.0584	0.0493	33	0.0490	0.0066
6	0.0595	0.0000	20	0.0691	0.0077	34	0.0905	0.0353
7	0.0258	0.0310	21	0.0301	0.0143	35	0.0756	0.0003
8	0.1224	0.0415	22	0.0517	0.0031	36	0.0324	0.0042
9	0.1232	0.0132	23	0.0805	0.0036	37	0.1088	0.0006
10	0.0815	0.0183	24	0.0274	0.0084	38	0.0650	0.0010
11	0.0317	0.0002	25	0.1408	0.0030	39	0.0340	0.0051
12	0.0649	0.0004	26	0.0329	0.0036	40	0.0414	0.0021
13	0.0539	0.0005	27	0.0516	0.0009			
14	0.1291	0.1099	28	0.0947	0.0533			

The next example illustrates the use of these two statistics in graphical formats.

EXAMPLE 4.8 (Used Car Prices: Checking Influential Observations)

Consider the used car prices data and the regression model fitted to Price based on the training data set in Example 3.13. In Figure 4.11 the left panel shows the sequence plot of Cook's distances in which large D_i's show up as spikes. These observations (nos. 76, 77 and 78) can be regarded as influential.

The right panel shows the plot of Cook's distance versus leverage. The same observations are clustered in the North-East corner because they have both high D_i and high h_{ii}. Hence they are identified as influential. These plots are produced using the same R code as in Example 4.1 except using `which=c(4,6)` in the `plot` function. In this example, $p = 10$ and $n = 402$, so the threshold for the leverage is $2(p+1)/n = 22/402 = 0.0547$, which is exceeded by quite a few observations as can be seen from the right panel. Using $1-\alpha = 0.1$, the threshold for Cook's distance is $f_{11,391,0.9} = 0.5049$, which is not exceeded by any of the observations including nos. 76, 77 and 78. Another threshold used in the literature is $4/[n-(p+1)]$, which equals 0.0102 in the present example, but it is exceeded by many observations as can

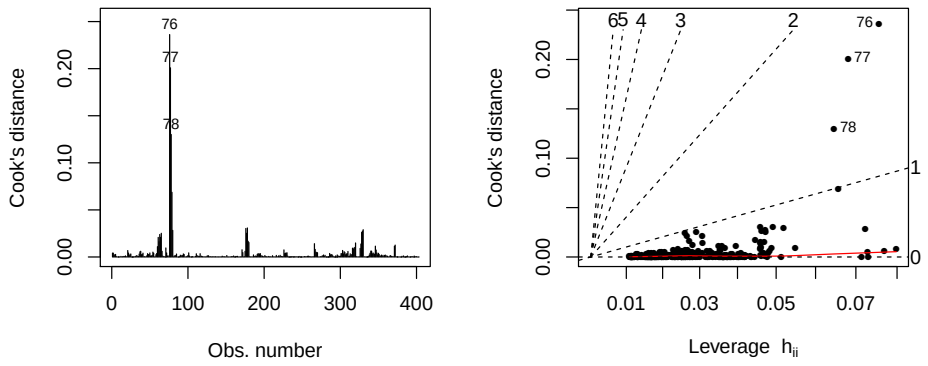


Figure 4.11 Sequence plot of Cook's distance and plot of Cook's distance versus leverage for the regression model fitted in Example 3.13

be seen from the left panel. So why only these three observations are marked as influential? A possible reason is that they have D_i values that are many times larger than any other D_i value, which is the reason they show up as especially tall spikes in the sequence plot. The dotted lines in the right panel (which are approximately straight lines) are drawn at constant values of $D_i(1-h_{ii})/h_{ii} = (e_i^*)^2/(p+1) = k^2$ (for $k = 1, \ldots, 6$). Thus large values of this ratio indicate outlier observations having large $|e_i^*|$ values. ■

There are several other deletion diagnostics similar to Cook's distance. For example, DFFITS is a standardized measure of the change in the fitted value of the ith observation due to deleting the same observation, while DFBETA is a standardized measure of the change in $\widehat{\beta}_j$ due to deleting the ith observation (thus there is a matrix of $n \times p$ DFBETA values). We do not discuss them here. Interested reader is referred to the books by Montgomery, Peck and Vining (2012) and Chatterjee and Hadi (2012) for more details.

4.8 Checking Multicollinearity

4.8.1 Multicollinearity: Causes and Consequences

As seen in Chapter 3, mathematically multicollinearity is the result of the columns of $\boldsymbol{X}$ being approximately linearly dependent. This causes $\boldsymbol{X}'\boldsymbol{X}$ to be nearly singular making it difficult (or impossible) to invert, which in turn makes the computation of $\widehat{\boldsymbol{\beta}}$ difficult and subject to numerical errors. Furthermore, $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ has large entries, which imply large variances of the $\widehat{\beta}_j$'s.

The following are some of the common causes of multicollinearity.

- The most obvious cause is structural linear dependencies among the predictors. One example is the percentages of ingradients in a chemical product that add up to 100%. A second example is dummy variables, which add up to 1 for a categorical variable if all of them are included in the model. A third example was mentioned in Chapter 3 concerning the data on income, expenditure and saving.
- In any big data set with a large number of predictors, there are often linear dependencies between subsets of the predictors which can go unnoticed.
- Predictors with spurious correlations because of omitted lurking variables.

From now on we will assume that all variables are standardized so that we are in the standardized regression setting of Section 3.6.3. To simplify the notation we will drop asterisks from $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ and assume that there is no intercept term in the model.

To see how the $\widehat{\beta}_j$'s deviate from the true β_j's in case of multicollinearity, consider the eigenvalues λ_j $(1 \leq j \leq p)$ of $\boldsymbol{R} = (n-1)^{-1}\boldsymbol{X}'\boldsymbol{X}$, namely the correlation matrix of the x_j's. Note that the $\lambda_j \geq 0$ since $\boldsymbol{R}$ is a positive semidefinite matrix (all $\lambda_j > 0$ if $\boldsymbol{R}$ is positive definite). Then

$$\sum_{j=1}^{p} \text{Var}(\widehat{\beta}_j) = \sigma^2 \text{tr}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2 \text{tr}[(n-1)\boldsymbol{R}]^{-1} = \frac{\sigma^2}{n-1}\sum_{j=1}^{p}(1/\lambda_j), \qquad (4.18)$$

since $1/\lambda_j$ are the eigenvalues of $\boldsymbol{R}^{-1}$ and the trace of a matrix equals the sum of its eigenvalues. So if some of the λ_j are close to 0 then the above quantity can blow up. Note that

$$\sum_{j=1}^{p} \text{Var}(\widehat{\beta}_j) = E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = E(\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}'\boldsymbol{\beta}. \qquad (4.19)$$

Thus if $\sum_{j=1}^{p} \text{Var}(\widehat{\beta}_j)$ blows up then the estimated parameter vector $\widehat{\boldsymbol{\beta}}$ deviates too far from the true parameter vector $\boldsymbol{\beta}$. or equivalently the squared norm (length) of the $\widehat{\boldsymbol{\beta}}$ vector tends to be much greater than that of the true $\boldsymbol{\beta}$ vector. Therefore the individual $\widehat{\beta}_j$ coefficients tend to be too large in absolute value. Ridge and lasso regressions discussed in Chapter 5 address this problem.

4.8.2 Multicollinearity Diagnostics

Pairwise Correlations

The simplest multicollinearity diagnostic is the pairwise correlations among the x's. If some or all of them are large, e.g., greater than 0.8 in absolute value then they indicate multicollinearity. Note that multicollinearity refers to linear dependencies among the x's; the correlations between the x's and y do not matter and they should be ideally large.

EXAMPLE 4.9 (Hald Cement Data: Correlation Matrix)

Hald (1952) gave the data shown in Table 4.4 on the heat evolved in calories during hardening of cement per gram (y) for 13 samples of cement and the percentages of their four ingradients: tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3) and dicalcium silicate (x_4). Obviously, the four percentages add up to 100% except for rounding errors and due to impurities. So these data are structurally multicollinear.

Table 4.4 Hald cement data

No.	x_1	x_2	x_3	x_4	y	No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5	8	1	31	22	44	72.5
2	1	29	15	52	74.3	9	2	54	18	22	93.1
3	11	56	8	20	104.3	10	21	47	4	26	115.9
4	11	31	8	47	87.6	11	1	40	23	34	83.8
5	7	52	6	33	95.9	12	11	66	9	12	113.3
6	11	55	9	22	109.2	13	10	68	8	12	109.4
7	3	71	17	6	102.7						

The correlation matrix for these data is shown below.

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 1 & 0.229 & -0.824 & -0.245 \\ 0.229 & 1 & -0.139 & -0.973 \\ -0.824 & -0.139 & 1 & 0.030 \\ -0.245 & -0.973 & 0.030 & 1 \end{bmatrix} \end{array}$$

We notice that two correlations, $\text{Corr}(x_1, x_3) = -0.824$ and $\text{Corr}(x_2, x_4) = -0.973$, are large negative. Examining the data closely we see that $x_1 + x_3$ is roughly constant equal to 20% and $x_2 + x_4$ is also roughly constant equal to 80%, so there are two approximate linear dependencies among the x's resulting in these two large negative correlations. ■

Although large pairwise correlations between the x's are indicative of multicollinearity, the converse is not necessarily true. If linear dependencies exist among multiple x's (referred to as **multivariate linear dependency**) then none of the pairwise correlations may be large; see Exercise 4.10 for an example. In the following we introduce diagnostic measures that take into account such multivariate linear dependencies.

t- and F-Statistics

Because multicollinearity results in large variances of the $\widehat{\beta}_j$'s, most of them turn out to be statistically nonsignificant. On the other hand, the ANOVA F-statistic can be highly significant. Thus one is in a dilemma that the overall fit is significant but none of the individual coefficients is significant. The Hald cement data from Example 4.9 illustrates this phenomenon.

EXAMPLE 4.10 (Hald Cement Data: Regression)

The regression output for the cement data is shown in the following output. We see that all four regression coefficients are nonsignificant, whereas the overall F-statistic is highly significant with $P = 4.756 \times 10^{-7}$ pointing to a multicollinearity problem.

```
> lsfit=lm(y~x1+x2+x3+x4,data=cement)
> summary(lsfit)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

■

Condition Number of the Correlation Matrix

Equation (3.34) from Chapter 3 gives the standardized regression coefficient vector as $\widehat{\boldsymbol{\beta}}^* = \boldsymbol{R}^{-1}\boldsymbol{r}$, where $\boldsymbol{R}$ is the correlation matrix among the x's and $\boldsymbol{r}$ is the vector of correlations between y and the x's. In that section we showed how to obtain the unstandardized regression coefficient vector $\widehat{\boldsymbol{\beta}}$ from $\widehat{\boldsymbol{\beta}}^*$. So the nub of the multicollinearity problem lies in how close $\boldsymbol{R}$ is to singularity. A standard measure used in numerical analysis for this purpose is the **condition number** of $\boldsymbol{R}$, defined as follows: Denote the eigenvalues of $\boldsymbol{R}$ by $\lambda_1, \ldots, \lambda_p$. All the eigenvalues are nonnegative because the correlation matrix is always positive semidefinite; if it is positive definite and hence nonsingular then all $\lambda_i > 0$, i.e., $\lambda_{\min} > 0$. If $\lambda_{\min} \approx 0$ then $\boldsymbol{R}$ is close to singularity indicating multicollinearity. The condition number of $\boldsymbol{R}$ is defined as the square root of the ratio of the maximum eigenvalue to the minimum eigenvalue of $\boldsymbol{R}$:

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}.$$

A rule of thumb is to decide that there is a multicollinearity problem if $\kappa > 15$.

The eigenvalues of the correlation matrix for the cement data given in Example 4.9 are:

$$\lambda_1 = 0.0016, \lambda_2 = 0.1866, \lambda_3 = 1.5761, \lambda_4 = 2.2357.$$

The condition number of the correlation matrix is then

$$\kappa = \sqrt{\frac{2.2357}{0.0016}} = 37.381,$$

which is more than twice the threshold of 15 pointing to a severe multicollinearity problem.

Variance Inflation Factors

The most widely used statistical measure of multicollinearity is the **variance inflation factor (VIF)**. To understand the idea behind the VIF, again refer to Equation (3.34) for $\widehat{\boldsymbol{\beta}}$. It follows from this equation that $\text{Cov}(\widehat{\boldsymbol{\beta}}^*) = \boldsymbol{R}^{-1}$. So $\text{Var}(\widehat{\beta}_j^*)$ is the jth diagonal entry of $\boldsymbol{R}^{-1}$. The variances of the unstandardized regression coefficients are proportional to the $\text{Var}(\widehat{\beta}_j)$. In the ideal case of uncorrelated x's, $\boldsymbol{R}$ is an identity matrix and so $\text{Var}(\widehat{\beta}_j^*) = 1$

for all j. If $\boldsymbol{R}$ is not an identity matrix then the Var$(\widehat{\beta}_j^*)$ are greater than 1 and represent the factors by which the Var$(\widehat{\beta}_j^*)$ or equivalently the Var$(\widehat{\beta}_j)$ are inflated because of the correlations among the x's. Therefore the jth diagonal entry of $\boldsymbol{R}^{-1}$ is defined as the variance inflation factor for x_j and is denoted by VIF$_j$.

As an example, consider a multiple regression problem with two predictors, x_1 and x_2. Let r denote the sample correlation coefficient between x_1 and x_2. So

$$\boldsymbol{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{R}^{-1} = \frac{1}{1-r^2}\begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}.$$

Hence VIF$_1$ = VIF$_2$ $= 1/(1-r^2)$. If $r = 0$ then there is no variance inflation but as $|r| \to 1$, VIF$_1$ = VIF$_2$ $\to \infty$.

This example is a special case of an alternative definition of the VIF's for multiple predictors. Let R_j^2 denote the R^2 from the regression of x_j on all other predictors. Then VIF$_j$ is given by

$$\text{VIF}_j = \frac{1}{1-R_j^2} \quad (j = 1, \ldots, p). \tag{4.20}$$

If x_j is approximately linearly dependent on other predictors then R_j^2 will be close to 1 and so VIF$_j$ will be large.

A common rule of thumb is to declare multicollinearity if most of the VIF$_j$ are > 10. This means that $R_j^2 > 0.90$ or more than 90% of the variation in x_j is accounted for by a linear least squares fit with respect to other predictors.

EXAMPLE 4.11 (Hald Cement Data: Variance Inflation Factors)

We calculate the VIF's for the Hald cement data using two different methods described above. In the first method we calculate the inverse of the correlation matrix between $x_1, \ldots, x_4$ given in Example 4.9:

$$\boldsymbol{R}^{-1} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 38.496 & 94.120 & 41.884 & 99.786 \\ 94.120 & 254.423 & 105.091 & 267.539 \\ 41.884 & 105.091 & 46.868 & 111.145 \\ 99.786 & 267.539 & 111.145 & 282.513 \end{bmatrix} \end{array}.$$

The diagonal entries of this matrix are the four VIF's, which can be seen to be very large.

In the second method we regress each x_j on the other three predictors. The resulting R^2's are as follows:

$$R_1^2 = 0.9740, R_2^2 = 0.9961, R_3^2 = 0.9785, R_4^2 = 0.9965.$$

So the corresponding VIF's are

$$\text{VIF}_1 = \frac{1}{1-0.9740} = 38.476, \text{VIF}_2 = \frac{1}{1-0.9961} = 254.453,$$
$$\text{VIF}_3 = \frac{1}{1-0.9785} = 46.577, \text{VIF}_4 = \frac{1}{1-0.9965} = 282.486,$$

which agree with the values obtained by the first method except for roundoff errors.

All four VIF's exceed the threshold of 10 by large margins indicating a serious multicollinearity problem. One way to address this problem is through one of the variable selection methods discussed in Chapter 6 that tells us which subset of the x's should be selected in the model. The other way is ridge or lasso regression discussed in Chapter 5. ■

EXERCISES

Theoretical Exercises

4.1 (Internally and externally studentized residuals) Prove the relationship (4.8) between internally and externally studentized residuals.

4.2 (Leverage for simple linear regression) Refer to Exercise 3.4 which gives the following expression for leverage in case of simple linear regression:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}} \quad (1 \leq i \leq n).$$

Show that the ith observation is flagged as influential using the rule of thumb (4.14) if

$$(x_i - \overline{x})^2 > \frac{3S_{xx}}{n}.$$

Interpret this result in terms of outlierness of x_i.

4.3 (Variances of the $\widehat{\beta}_j$ and VIF$_j$) Show that

$$\sum_{j=1}^{p} \text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{n-1} \sum_{j=1}^{p} \text{VIF}_j.$$

Applied Exercises

4.4 (College GPA and entrance test scores: Checking normality and homoscedasticity) Refer to Example 3.16 in which we fitted the model

$$\text{GPA} = \beta_0 + \beta_1 \text{Verbal} + \beta_2 \text{Math} + \beta_3 \text{Verbal}^2 + \beta_4 \text{Math}^2 + \beta_5 \text{Verbal} \times \text{Math} + \varepsilon.$$

The regression coefficients are given in the R output in that example.

a) Make the normal and fitted values plots of residuals. Comment on why the normality and especially the homoscedasticity assumptions seem to be violated. Does the fitted values plot suggest the log transformation of GPA?

b) Fit the same model using log(GPA) as the response variable. Make the normal and fitted values plots of residuals. Are the normality and homoscedasticity assumptions satisfied?

4.5 (Research expenditures data) Refer to Exercise 3.13 on modeling research expenditures of top 30 engineering schools using the number of faculty and the number of PhD students as predictor variables. The two scatter plots are shown in Figures 4.12 and 4.13 with each data point marked by the abbreviated name of the university. Identify the outliers and influential observations in the data using appropriate diagnostic statistics. Provide plausible explanations for why these universities are flagged.

4.6 (Employee salaries: Checking normality and homoscedasticity) Refer to Exercise 3.15. Using the same predictors that were found significant in part (a) of that exercise fit two regressions, one using Salary as the response variable and the other using log(Salary) as the response variable.

a) Make normal plots for residuals from both regressions. Has the log transformation of Salary improved normality?

b) Make fitted values plots for both sets of residuals. Has the log transformation of Salary improved homoscedasticity?

4.7 (Soft drink sales: Testing independence) Data on sales of a softdrink and advertising expenditures are available for 20 successive years. The sequence plot of residuals from

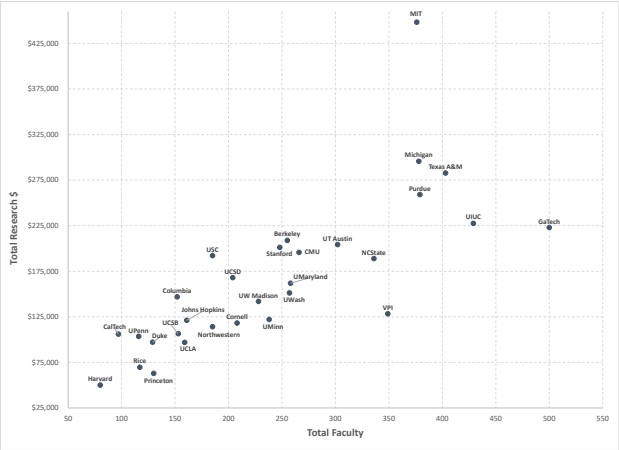


Figure 4.12 Plot of research expenditures (in millions of dollars) versus number of faculty for the top 30 graduate engineering programs

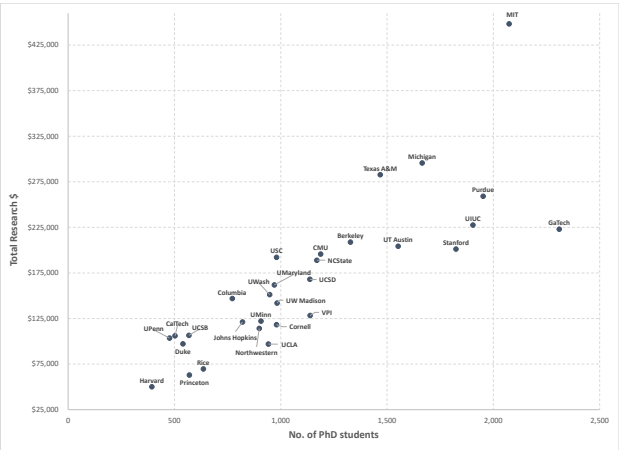
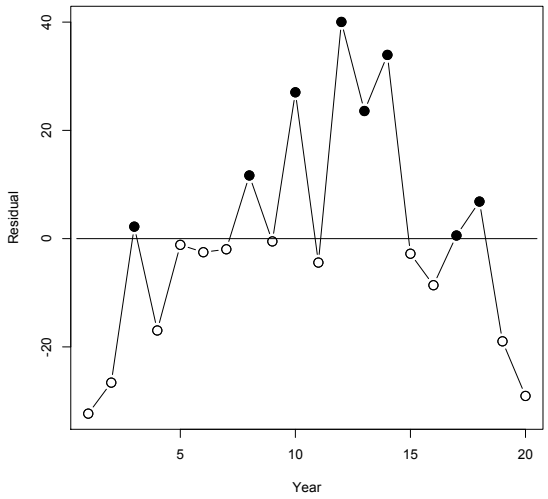


Figure 4.13 Plot of research expenditures (in millions of dollars) versus number of PhD students for the top 30 graduate engineering programs

Table 4.5 Woodbeam strength data

Observation No.	Specific Gravity	Moisture Content	Strength	Observation No.	Specific Gravity	Moisture Content	Strength
1	0.499	11.1	11.14	6	0.528	9.9	12.60
2	0.558	8.9	12.74	7	0.418	10.7	11.13
3	0.604	8.8	13.13	8	0.480	10.5	11.70
4	0.441	8.9	11.51	9	0.406	10.5	11.02
5	0.550	8.8	12.38	10	0.467	10.7	11.41

the simple linear regression of sales on advertising expenditures is shown below (open circles represent negative residuals, dark circles represent positive residuals).



a) Do the runs test to check if the autocorrelation coefficient is significantly greater than 0.
b) The sample autocorrelation coefficient for this plot is 0.46. Do the Durbin-Watson test to check if it is significantly greater than 0?
c) What model would you fit for the next iteration of the Cochrane-Orcutt procedure? How are the coefficients in the new model related to those in the original model?

4.8 (Woodbeam data: Influential observations) Table 4.5 gives data on the specific gravity (x_1), moisture content (x_2) and strength (y) of wood beams.

a) Make a scatter plot of the two predictors. Which observation appears to be influential?
b) Check if that observation is influential using the rule $h_{ii} > 2(p+1)/n$.

Table 4.6 Data illustrating multivariate linear dependence

No.	x_1	x_2	x_3	x_4	No.	x_1	x_2	x_3	x_4
1	8	1	1	1	7	2	7	0	1
2	8	1	1	0	8	2	7	0	1
3	8	1	1	0	9	2	7	0	1
4	0	0	9	1	10	0	0	0	10
5	0	0	9	1	11	0	0	0	10
6	0	0	9	1	12	0	0	0	10

c) Check if that observation is influential using Cook's distance.

d) Fit the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ from all data, and compare it with the fit obtained after omitting the influential observation. Does the fit change much?

4.9 (Anscombe data: Cook's distance) We have seen in Section 4.7.1 that the observation $(x = 19, y = 12.50)$ in the Anscombe Data Set IV is influential and that its leverage equals 1. Is the Cook's distance for this observation defined? Why or why not?

4.10 (Multivariate linear dependency) Webster, Gunst and Mason (1974) gave the data shown in Table 4.6 on four predictor variables. The data are constructed such that x_1, x_2, x_3, x_4 add up to 10 in all 12 observations except the first one where they add up to 11. Thus there is an almost exact linear dependence among the four observations.

a) Calculate the correlation matrix among the four predictors. Check that no correlation exceeds 0.5 in absolute value. Thus correlations do not provide an indication of multicollinearity.

b) Calculate the four VIF's and check that they all exceed 150 with the maximum equal to 289.23. Thus VIF's indicate serious multicollinearity.

4.11 (Acetylene data: Multicollinearity statistics) Table 4.7 gives data from Marquardt and Snee (1975) on conversion of n-heptane to acetylene (y) as a function of three reaction conditions: reactor temperature (x_1), ratio of H_2 to n-heptane (x_2) and contact time (x_3).

The following full second degree model is to be fitted to the data:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \\ &\quad + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \varepsilon. \end{aligned}$$

a) Plot the three predictor variables against each other. Also calculate the correlation coefficients between them. Do you see any indications of multicollinearity?

b) Calculate the VIF's for all the terms in the above model. Comment on your results.

c) Center x_1, x_2, x_3 by subtracting the mean of each predictor from its values. Compute the remaining terms (pairwise products and squares) from these centered values. Now calculate the VIF's for all the terms. Compare the results with those from (b). Has centering made the multicollinearity problem less severe?

Table 4.7 Acetylene data

x_1 Reactor Temperature (°C)	x_2 Ratio of H_2 to n-heptane (mole ratio)	x_3 Contact Time (sec)	y Conversion of n-heptane to Acetylene (%)	x_1 Reactor Temperature (°C)	x_2 Ratio of H_2 to n-heptane (mole ratio)	x_3 Contact Time (sec)	y Conversion of n-heptane to Acetylene (%)
1300	7.5	0.0120	49.0	1200	11.0	0.0320	34.5
1300	9.0	0.0120	50.2	1200	13.5	0.0260	35.0
1300	11.0	0.0115	50.5	1200	17.0	0.0340	38.0
1300	13.5	0.0130	48.5	1200	23.0	0.0410	38.5
1300	17.0	0.0135	47.5	1100	5.3	0.0840	15.0
1300	23.0	0.0120	44.5	1100	7.5	0.0980	17.0
1200	5.3	0.0400	28.0	1100	11.0	0.0920	20.5
1200	7.5	0.0380	31.5	1100	17.0	0.0860	29.5

CHAPTER 5

MULTIPLE LINEAR REGRESSION: SHRINKAGE AND DIMENSION REDUCTION METHODS

As we have seen in the previous chapter, multicollinearity can result in serious difficulties for LS regression. Multicollinearity is often caused by having too many predictors. In some practical examples we even have $p \gg n$, in which case regression methods that we have learned so far break down. This occurs when the samples are expensive but collecting data on many variables is relatively inexpensive. For example, spectroscopic methods are often used in chemical analysis of compounds. Spectroscopic intensity measurements are made at hundreds of frequencies. We do not consider this case here.

LS estimators have the nice property of unbiasedness. However, when multicollinearity is present, this property is not particularly useful since LS estimators have very large variances and become unstable if the data are perturbed. One can trade bias for variance, which can lead to smaller mean square error (MSE) (defined as the expected value of the squared difference between an estimator and the true parameter it is estimating) since MSE equals Bias^2 + Variance. In this chapter we discuss some alternative estimation methods that improve upon LS estimation by reducing variance substantially at the expense of introducing small bias, thus reducing overall MSE.

The point of departure of ridge and lasso regressions is the problem noted in (4.18) that under multicollinearity, the LS estimator $\widehat{\boldsymbol{\beta}}$ tends to be too long in comparison to the true $\boldsymbol{\beta}$. So these methods put a constraint on the length of the estimator of $\boldsymbol{\beta}$. In ridge regression, the length is defined by the L_2**-norm**: $\sqrt{\sum_{j=1}^{p} \beta_j^2} = \sqrt{\boldsymbol{\beta}'\boldsymbol{\beta}}$. In lasso regression the length is defined by the L_1**-norm**: $\sum_{j=1}^{p} |\beta_j|$. The specified constraint is incorporated by adding a penalty term to the LS criterion, which is then minimized. The resulting regression is

known as **penalized regression**. Putting such a constraint causes the estimators of the regression coefficients to shrink toward zero. Therefore two methods of regression are known as **shrinkage methods**.

Another class of regression methods considered in this chapter involve reducing the **high dimensionality** of the predictor space by finding a few key linear combinations of the predictors that are used to perform regression. If $p > n$ then this is necessitated by the fact that the rank of $(\boldsymbol{X}'\boldsymbol{X})$ is less than p, and so $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ needed in LS estimation cannot be computed. Even if $p < n$, $\boldsymbol{X}'\boldsymbol{X}$ may be less than full rank because of multicollinearity. **Principal components regression (PCR)** and **partial least squares (PLS)** are two methods in this class. In PCR a few key linear combinations of the predictors (called **principal components (PC's)**) are chosen to capture most of the variation among the original predictors; these PC's are then used as predictors. PLS follows a similar approach but uses the criterion of maximizing the covariance between the original predictors and the response variable for choosing the linear combinations. Thus, while PCR focuses on capturing the most variance among the x's in choosing the PC's, PLS focuses on which components are most predictive of y. These two methods of regression are known as **dimension reduction methods**.

5.1 Ridge Regression

The method of ridge regression was originally proposed by Hoerl and Kennard (1970). There have been many extensions and improvements, but here we will restrict to the basic method.

Ridge Problem

Ridge estimator of $\boldsymbol{\beta}$ is defined as the LS estimator subject to a constraint on its squared length:

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{subject to} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq d^2 \tag{5.1}$$

for some specified $d > 0$. This is a convex optimization problem, which can be reformulated using the **Lagrangian multiplier method** as

$$\min_{\boldsymbol{\beta}} \left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}\right], \tag{5.2}$$

where $\boldsymbol{\beta}'\boldsymbol{\beta}$ is the penalty term and λ is the Lagrangian multiplier. The ridge estimator of $\boldsymbol{\beta}$ is then given by

$$\widehat{\boldsymbol{\beta}}^R(\lambda) = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{5.3}$$

where $\boldsymbol{I}$ is a $p \times p$ identity matrix. Note that $\widehat{\boldsymbol{\beta}}^R(0)$ is the usual LS estimator.

As we saw in the previous chapter, multicollinearity is caused by the $\boldsymbol{X}'\boldsymbol{X}$ matrix being close to singular, which results in its minimum eigenvalue to be close to zero. The ridge estimator $\widehat{\boldsymbol{\beta}}^R(\lambda)$ can be viewed as ameliorating this singularity by adding a positive constant λ to the diagonal entries of $\boldsymbol{X}'\boldsymbol{X}$.

Note that λ is a decreasing function of d^2 with $\lambda = 0$ when $d^2 = \infty$ (the unconstrained LS minimization problem). Geometrically, the ridge estimator can be interpreted as follows. The contours of the constant values of the LS criterion $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ can be plotted in the $\boldsymbol{\beta}$ space. These contours are ellipsoids. In two dimensions they are ellipses as shown in Figure 5.3. The center of the ellipse is where the LS criterion is minimum and

corresponds to the LS estimator $\widehat{\boldsymbol{\beta}}$. As one moves away from the center, the value of the LS criterion increases. The ridge estimator is the shortest $\boldsymbol{\beta}$ vector that touches the contour with the circle of radius equal to d as shown in the figure. As d^2 gets smaller (or as λ gets larger), the ridge estimator shrinks, approaching the null vector in the limit. Thus its variance becomes zero but its bias equals $\boldsymbol{\beta}$.

5.1.1 Choice of λ

Choice of the tuning parameter λ is a difficult problem. In the **ridge trace** method the individual estimates $\widehat{\beta}_j^R(\lambda)$ $(1 \leq j \leq p)$ are plotted against λ. As λ increases, the magnitudes of the estimates shrink approaching to zero in the limit, where the bias^2 term reaches its maximum of $\boldsymbol{\beta}'\boldsymbol{\beta}$. One could choose λ where the ridge traces for most coefficients stabilize. However, such a value of λ may not exist or may be difficult to identify.

The `glmnet` library in R employs the following method to choose the optimal λ. It selects a grid of λ-values. For each value of λ in the grid, it performs m**-fold cross-validation** and calculates associated $\widehat{\boldsymbol{\beta}}^R(\lambda)$ and MSE values. Thus for each value of λ we get a range of MSE values. From these their means with one standard deviation bars around them are plotted as shown in Figure 5.2. From this plot one can find not just the optimal λ which minimizes the estimated MSE, but a range of acceptable λ-values which give MSE-values within the limits of sampling error from the estimated minimum MSE.

Regardless of how λ is chosen, ridge regression shrinks all coefficient estimates simultaneously. Some estimates are driven to zero faster than others but no estimates are generally set exactly equal to zero. Therefore no predictors are dropped from the model. As a result, ridge regression cannot be used as a variable selection method as can lasso regression discussed in the following section.

Both ridge and lasso regression can be performed using `glmnet`. This package uses standardized data but reports unstandardized regression coefficients. It is important to use the standardized data, so that all predictor variables are equally scaled so that their β coefficients are not affected by their scales and so their norms can be computed. The `glmnet` package actually minimizes a generalized criterion:

$$\frac{1}{2n}\text{SSE} + \lambda \left[\frac{(1-\alpha)}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \alpha \sum_{j=1}^{p} |\beta_j| \right],$$

where $\text{SSE} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ and the quantity in square brackets in the second term is a generalized penalty. For ridge regression $\alpha = 0$, so the penalty term is $(1/2)\boldsymbol{\beta}'\boldsymbol{\beta} = (1/2)\sum_{j=1}^{p} \beta_j^2$. For lasso regression $\alpha = 1$, so the penalty term is $\sum_{j=1}^{p} |\beta_j|$. For general α $(0 < \alpha < 1)$, the method is known as **elastic net regression**.

EXAMPLE 5.1 (Hald Cement Data: Ridge Regression)

Ridge regression was performed on the Hald cement data using `glmnet`. The following code was used to generate the results including the ridge trace plot shown in Figure 5.1. The plot of the objective function (mean-squared error) minimized using cross-validation is shown in Figure 5.2. The red dots show the averages (over all cross-validations) of the mean-squared error with one standard deviation error bars around them. Denote the optimal value of λ that minimizes MSE by $\lambda_{\min}$. The $\ln \lambda_{\min}$ is given by the first vertical dotted line on the plot; it equals -0.51, which corresponds to $\lambda_{\min} = 0.6$ as given in the R output below. To allow for sampling variation in the

cross-validation process, the $\ln \lambda_{\min}$ can be as large as about 1.15 (or $\lambda_{\min} = 3.16$) as indicated by the second vertical dotted line. It should be noted that if this R code is run again then a different interval for $\lambda_{\min}$ will generally result. The fitted ridge regression model corresponding to this value of λ is

$$\widehat{y} = 85.1108 + 1.2427x_1 + 0.2891x_2 - 0.1845x_3 - 0.3570x_4.$$

Compare this model with the LS fitted model

$$\widehat{y} = 62.4063 + 1.5511x_1 + 0.5102x_2 + 0.1019x_3 - 0.1441x_4.$$

```
> library(glmnet)
> y=cement$y
> x=model.matrix(y~.,cement)
> ridgefit=glmnet(x, y, alpha=0,lambda=seq(0,100,0.01))
> ridgecv=cv.glmnet(x, y, alpha=0,nfold=3,lambda=seq(0,100,0.1))
> plot(ridgecv)
> lambdaridge=ridgecv$lambda.min
> print(lambdaridge)
[1] 0.6
> small.lambda.index <- which(ridgecv$lambda == ridgecv$lambda.min)
> small.lambda.betas <- coef(ridgecv$glmnet.fit)[,small.lambda.index]
> plot(ridgefit,xvar="lambda", main="Coeffs of Ridge Regression",
+       type="l", xlab=expression("log_lambda"), ylab="Coeff")
> abline(h=0); abline(v=log(ridgecv$lambda.min))
> grid()
(Intercept)          x1          x2          x3          x4
 85.1107504   1.2426735   0.2891106  -0.1845128  -0.3570048
```

■

5.2 Lasso Regression

The method of lasso regression was proposed by Tibshirani (1996). This method has become extremely popular in recent years and there have been many extensions of it summarized in the book by Hastie, Tibshirani and Wainwright (2015). As in the case of ridge regression we will restrict to the basic method.

Lasso Problem

Analogous to the ridge estimation problem (5.1), the lasso estimator solves the following constrained minimization problem:

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq d, \tag{5.4}$$

where $d > 0$ is a specified constant. This is also a convex optimization problem, which can be reformulated using the Lagrangian multiplier method as

$$\min_{\boldsymbol{\beta}} \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{5.5}$$

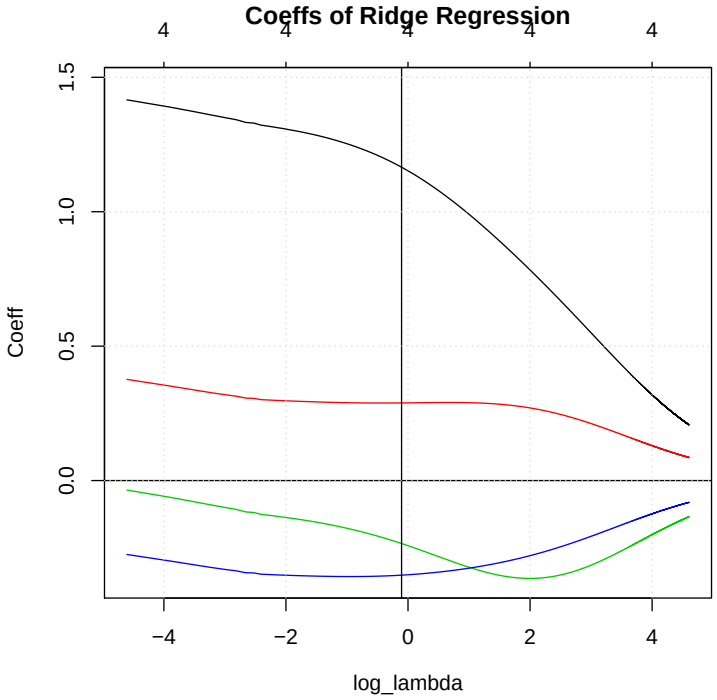


Figure 5.1 Ridge trace plot for Hald cement data

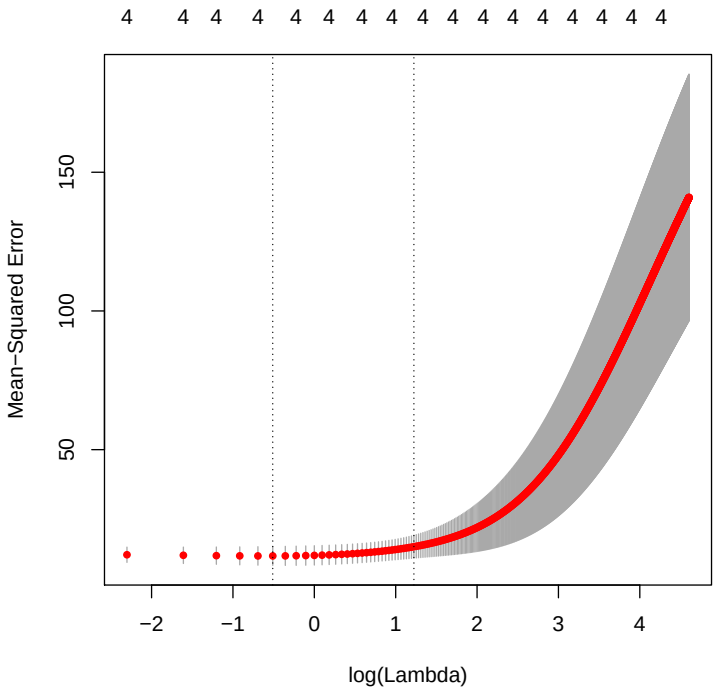


Figure 5.2 Mean-squared error plot for ridge regression of Hald cement data

This minimization problem does not have a closed form solution as in the case of ridge regression. We shall denote the lasso estimator by $\widehat{\boldsymbol{\beta}}^L(\lambda)$. The optimal λ can be chosen by the same m-fold cross-validation method given for ridge regression.

The advantage of lasso regression is that the solution to the above minimization problem generally falls at one of the corners of the hypercube $\sum_{j=1}^{p} |\beta_j| \leq d$ setting some subset of the β_j's equal to zero; see Figure 5.3. This hypercube defines the feasible region for the $\boldsymbol{\beta}$ vector with sharp corners on the β_j axes. Thus lasso regression effectively performs variable selection. On the other hand, for ridge regression, the feasible region for the $\boldsymbol{\beta}$ vector, namely $\sum_{j=1}^{p} \beta_j^2 \leq d^2$, is a smooth hypersphere with no sharp corners and so it does not set any β_j's equal to zero. For this reason, lasso regression is well-suited when there are only a few signals among many predictors, most being noise (non-signficant). This problem is referred to as **sparse regression**.

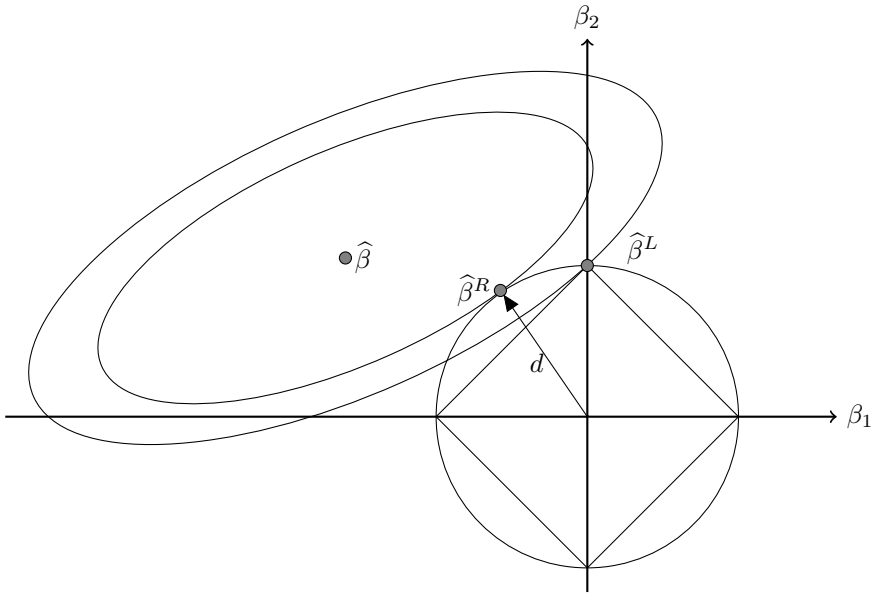


Figure 5.3 Least squares contours with ridge and lasso regression feasible regions for $p = 2$

EXAMPLE 5.2 (Hald Cement Data: Lasso Regression)

Lasso regression was performed on the Hald cement data using `glmnet`; the only change needed in the R code was to change `alpha = 0` to `alpha = 1`. The lasso trace plot is shown in Figure 5.4. The plot of the objective function (mean-squared error) minimized using cross-validation is shown in Figure 5.5. Note that $\ln \lambda_{\min}$ is about -1.43 or $\lambda_{\min} = 0.24$ as given in the R output below. But $\ln \lambda_{\min}$ can be as large as -0.07, which corresponds to $\lambda_{\min} = 0.93$. The corresponding fitted lasso regression model is

$$\widehat{y} = 72.2994 + 1.4176x_1 + 0.4061x_2 - 0.2336x_4.$$

Note that the coefficient of x_3 is zero and so x_3 is dropped from the model. The R code and the output are as follows.

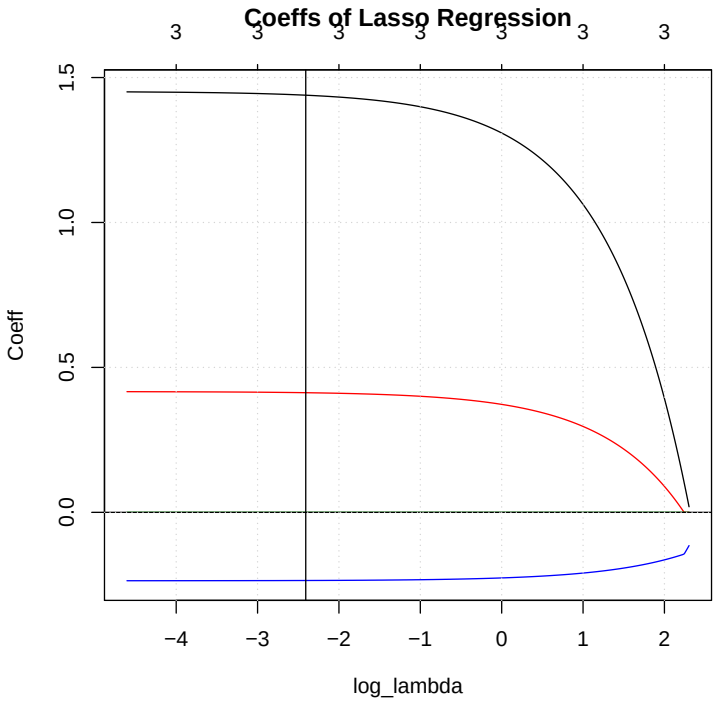


Figure 5.4 Lasso trace plot for Hald data

```
> lassofit=glmnet(x, y, alpha=1,lambda=seq(0,10,0.01))
> lassocv=cv.glmnet(x,y,alpha=1,nfold=3,lambda=seq(0,10,0.01))
> lambdalasso=lassocv$lambda.min
> print(lambdalasso)
[1] 0.24
> plot(lassofit,xvar="lambda",label=TRUE, main="Coeffs of Lasso Regression",
+     type="l", xlab=expression("log_lambda"), ylab="Coeff")
> abline(h=0); abline(v=log(lassocv$lambda.min))
> grid()
> print(small.lambda.betas)
(Intercept) (Intercept)          x1          x2          x3          x4
 72.2994407   0.0000000   1.4175553   0.4061040   0.0000000  -0.2336325
```

■

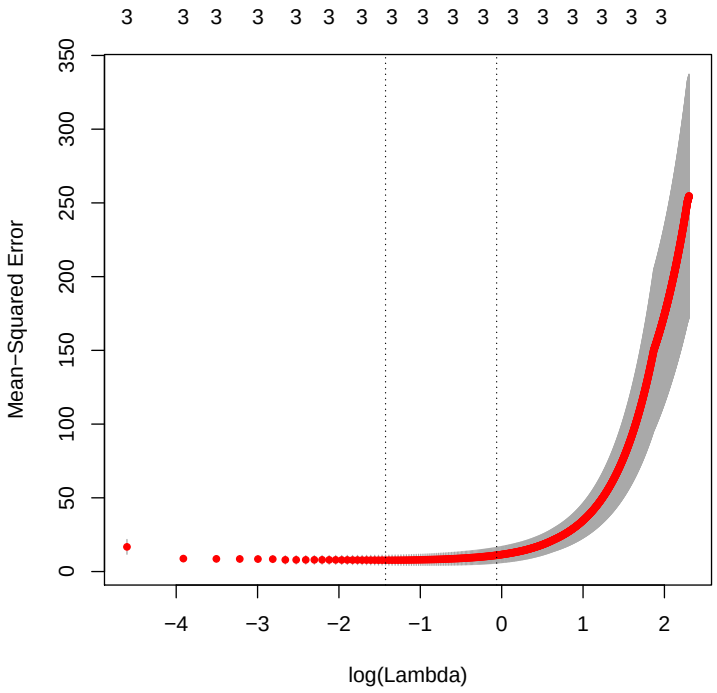


Figure 5.5 Mean-squared error plot for lasso regression of Hald data

5.3 Principal Components Analysis and Regression

5.3.1 Principal Components Analysis (PCA)

The objective of PCA is to find linear combinations (called **principal components** or **PC's**):

$$z_j = u_{1j}x_1 + \cdots + u_{pj}x_p \quad (j = 1, \ldots, p) \tag{5.6}$$

such that $\text{Var}(z_1) \geq \cdots \geq \text{Var}(z_p)$ and $\text{Corr}(z_j, z_k) = 0$ for $j \neq k$, and where the vectors $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{pj})'$ are of unit length. The idea here is that z_1 captures the most variation among the x's, z_2 captures the second most variation and is uncorrelated with z_1 and so on. We refer to z_1 as the first PC (PC1), z_2 as the second PC (PC2) and so on. Often, the first few PC's capture most of the variation among the x's and so can be used to summarize the x's without much loss of information, thus reducing the dimensionality of the data. The vectors $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{pj})' \quad (1 \leq j \leq p)$ are called the **loading vectors** of the PC's. Another term used for PC's is **factors**. This term comes from **factor analysis**, which is a closely related technique.

PCA involves finding the best orthogonal directions on which the x's should be projected in order to capture most variation among them. Along the first PC direction the x's vary the most and along the last PC direction they vary the least. Thus the most information among the x's is captured by projecting them on the first PC direction and the least information is captured by projecting them on the last PC direction. If the x's don't vary along any direction then projecting them on that direction results in a constant with no useful information.

Geometrically, the PC's correspond to the orthogonal principal axes of the ellipsoidal scatter of the data. When $p = 2$, PC1 corresponds to the major axis of the elliptical scatter and PC2 corresponds to the minor axis. Figure 5.6 shows that for a narrow elongated elliptical scatter, the major axis contains most of the information and the minor axis contains very little information.

The vector $\boldsymbol{z} = (z_1, \ldots, z_p)'$ of **PC scores** is a nonsingular orthogonal linear transformation of the data vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ of predictor variables. This transformation is defined by a $p \times p$ matrix $\boldsymbol{U} = \{u_{ij}\}$ with column vectors $\boldsymbol{u}_j \ (1 \leq j \leq p)$ so that $\boldsymbol{z} = \boldsymbol{U}'\boldsymbol{x}$. We show below that the loading vectors $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{pj})'$ are the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}$ corresponding to its eigenvalues $\lambda_j \ (1 \leq j \leq p)$. Then $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}$ and from the spectral decomposition theorem (see (A.1)) we know that $\boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$. However, from the sandwich formula (A.7) we also know that $\text{Cov}(\boldsymbol{z}) = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U}$. Therefore $\text{Cov}(\boldsymbol{z}) = \text{diag}\{\text{Var}(z_1), \ldots, \text{Var}(z_p)\} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$. Thus the z_j are uncorrelated and their variances are the ordered eigenvalues of $\boldsymbol{\Sigma}$. Next we give a computational algorithm to determine them.

First we want to find the vector $\boldsymbol{u}_1$ that maximizes $\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1 = \text{Var}(z_1)$. However, if we multiply $\boldsymbol{u}_1$ by any constant then $\text{Var}(\boldsymbol{z}_1)$ gets multiplied by that constant squared. To avoid such an artificial inflation of $\text{Var}(\boldsymbol{z}_1)$, we constrain $\boldsymbol{u}_1$ to be of unit length, i.e., $||\boldsymbol{u}_1||^2 = \boldsymbol{u}_1'\boldsymbol{u}_1 = 1$. Thus we solve the constrained maximization problem:

$$\max_{\boldsymbol{u}_1} \boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1 \quad \text{subject to} \quad \boldsymbol{u}_1'\boldsymbol{u}_1 = 1. \tag{5.7}$$

Using the Lagrangian multiplier method we get

$$\max_{\boldsymbol{u}_1} \left[\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1 - \lambda_1(\boldsymbol{u}_1'\boldsymbol{u}_1 - 1)\right]. \tag{5.8}$$

In Section 5.5.2 we show that the solution to this Lagrangian problem is $\max_{\boldsymbol{u}_1}(\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1) = \lambda_1$, which is the largest eigenvalue of $\boldsymbol{\Sigma}$, and the maximizing vector $\boldsymbol{u}_1$ is the associated

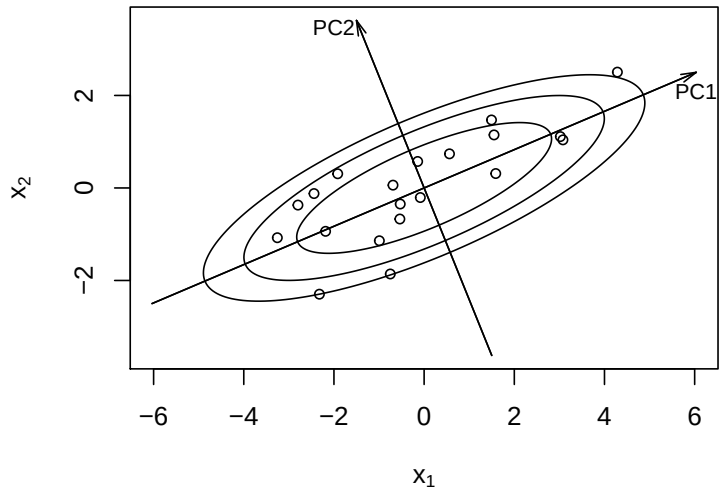


Figure 5.6 PC directions for elliptically contoured bivariate distribution

eigenvector. (To be notationally correct, (5.7) should be written as $\max_{\boldsymbol{u}} \boldsymbol{u}'\boldsymbol{\Sigma}\boldsymbol{u}$ subject to $\boldsymbol{u}'\boldsymbol{u} = 1$ and the solution to this constrained maximization problem should be denoted by $\boldsymbol{u}_1$. To simplify the notation we have denoted the variable of maximization and the solution to the constrained maximization problem both by $\boldsymbol{u}_1$. The same convention is followed below.)

The loading vector $\boldsymbol{u}_2$ of the second PC is found by the same method but with an additional constraint that $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are orthogonal, i.e., $\boldsymbol{u}_1'\boldsymbol{u}_2 = 0$. We get $\max_{\boldsymbol{u}_2} \boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_2 = \text{Var}(z_2) = \lambda_2$, the second largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_2$ is the eigenvector associated with λ_2. This process can be continued until all p PC's are computed. Usually, the process is stopped if the first $r < p$ PC's account for a high fraction (say, 90%) of the total variance among the x's defined to be $\text{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^{p} \lambda_j$.

In practice, the covariance matrix $\boldsymbol{\Sigma}$ is unknown, so the sample covariance matrix $\boldsymbol{S}$ estimated from observations $\boldsymbol{x}_i$ $(1 \leq i \leq n)$ must be used instead. We will continue to denote the eigenvalues of $\boldsymbol{S}$ by $\lambda_1, \ldots, \lambda_p$ and the associated eigenvectors by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$. Let $\boldsymbol{\Lambda}$ and $\boldsymbol{U}$ be as defined before. Then by the spectral decomposition theorem we have

$$\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}' = \boldsymbol{\Lambda} \quad \text{and} \quad \boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{U} = \boldsymbol{S}.$$

The PC scores matrix $\boldsymbol{Z}$ is obtained from the data matrix $\boldsymbol{X}$: $n \times p$ by the orthogonal rotation:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{U}. \tag{5.9}$$

The PC's in the sample space can also be found geometrically as follows. To determine the first PC direction, find the straight line that minimizes the sum of the squared projections on the line from all the data points $\boldsymbol{x}_i$. The second PC direction is found in the same way among all the lines that are perpendicular to the first PC direction. Figure 5.7 illustrates this for $p = 2$.

An important question to answer before performing PCA is whether the x's should be standardized or not. If the x's are not standardized then we will be computing linear combinations of variables with different units, e.g., age, weight and blood pressure of patients. Furthermore, inherent variations in these variables may artificially influence the

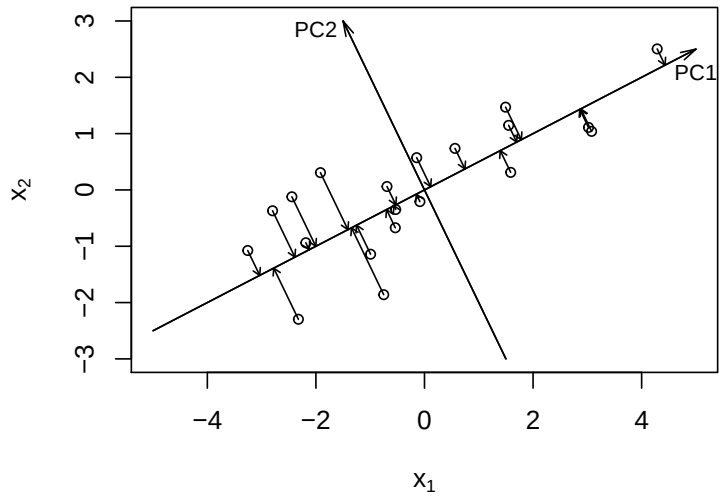


Figure 5.7 Principal component direction minimizes the sum of squared projections of all data points on it

PC's. For example, weights vary much more than ages. Because PCA aims at finding linear combinations of the original variables that maximize the variance, it puts higher loadings on variables with higher variances, in this example on the weights. Standardization makes the variables unitless and scale-free. However, if the variables are standardized then they all have a unit variance. So it is not clear what is meant by maximizing the variation among the x's. General consensus is that if the variables are commensurate, i.e., if they are measured in the same units and have roughly the same scales, then they should not be standardized; otherwise they should be standardized. Even in case of commensurate variables, standardization may be necessary if some variables have much higher inherent variability than others.

Next we give two examples of PCA using R. The "stats" module in R has two functions to perform PCA: `prcomp` and `princomp`. The variable names used are different for the two functions. For example, factor loadings are named `rotation` in `prcomp` and `loadings` in `princomp`, while factor scores are stored in a matrix labeled `x` in `prcomp` and in `scores` in `princomp`.

EXAMPLE 5.3 (GPA Data: Principal Components Analysis)

In Chapter 3 we used the GPA data in many examples. In Example 3.3 we found that the two predictors, Verbal and Math, are roughly uncorrelated and have about equal linear effects on GPA. Here we use the same data to illustrate PCA of Verbal and Math, both in unstandardized and standardized forms. The R code for performing PCA on unstandardized data and the results are shown below.

```
> gpa = read.csv("c:/data/GPA.csv")
> cov(gpa[,1:2])
         Verbal     Math
```

```
Verbal 259.2718 -22.6410
Math   -22.6410 172.8718
> fit1=prcomp(gpa[,1:2], scale=FALSE)
> summary(fit1)
Importance of components:
                          PC1     PC2
Standard deviation     16.2741 12.9344
Proportion of Variance  0.6129  0.3871
Cumulative Proportion   0.6129  1.0000
> fit1$sdev^2 # eigenvalues
[1] 264.8453 167.2983
> fit1$rotation
           PC1    PC2
Verbal -0.971 -0.239
Math    0.239 -0.971
```

The PC's are extracted from the covariance matrix between Verbal and Math shown in the above output. PC1 corresponds to $\lambda_1 = 264.8453$ and PC2 corresponds to $\lambda_2 = 167.2983$. Thus PC1 accounts for 61.29% of the total variation and PC2 accounts for the remaining 38.71%. For PC1, the loadings on Verbal and Math are -0.971 and 0.239, respectively, and for PC2, the loadings are -0.239 and -0.971, respectively. Thus PC1 puts four times as much weight on Verbal compared to Math nnd PC2 reverses these weights. PC1 and PC2 are orthogonal to each other and are normalized to have a unit length as can be verified from $0.971^2 + 0.239^2 = 1$. The PC1 and PC2 directions are shown on the scatterplot of Verbal and Math in Figure 5.8.

We can do PCA on standardized data by specifying `scale = TRUE` in the `prcomp` function. This extracts PC's from the correlation matrix. The results are as shown below.

```
> cor(gpa[,1:2])
        Verbal    Math
Verbal  1.0000 -0.1069
Math   -0.1069  1.0000

> fit2=prcomp(gpa[,1:2],scale=TRUE)
> summary(fit2)
Importance of components:
                          PC1    PC2
Standard deviation     1.0521 0.9450
Proportion of Variance 0.5535 0.4465
Cumulative Proportion  0.5535 1.0000
> fit1$sdev^2 # eigenvalues
[1] 1.1069 0.8931
> fit2$rotation
           PC1     PC2
Verbal -0.7071 -0.7071
Math    0.7071 -0.7071
```

Notice that the PC1 and PC2 loadings on both the predictors are equal in magnitude, namely $0.7071 = 1/\sqrt{2}$. This is not very useful since the loadings will be always

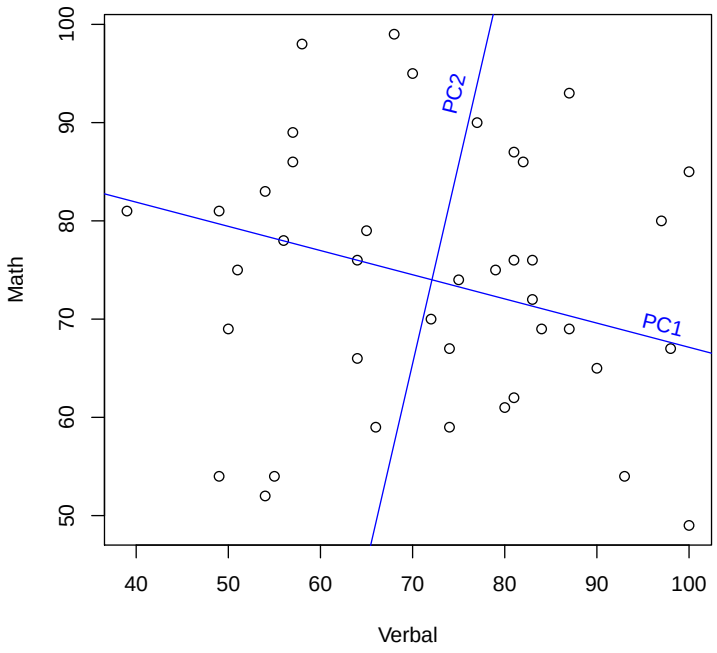


Figure 5.8 Principal component directions shown on the Verbal-Math scatterplot

equal to $\pm 1/\sqrt{2}$ for standardized data regardless of how different the variances of the two variables are, and this fact generalizes to $p > 2$ for equicorrelated data. In the present example Verbal has higher variance than Math and so should have a higher loading as found for unstandardized data. This puts into context our earlier comment that for standardized data, it is not clear what is meant by maximizing the variance among the x's. ■

Next we illustrate PCA on the Hald cement data.

EXAMPLE 5.4 (Hald Cement Data: Principal Components Analysis)

We use essentially the same R script used in the previous example; in addition we print the PC scores which are stored in a matrix labeled "x".

```
> fit1=prcomp(cement[,1:4])
> summary(fit1)
Importance of components:
                          PC1    PC2     PC3    PC4
Standard deviation     22.755 8.2156 3.52213 0.4870
Proportion of Variance  0.866 0.1129 0.02075 0.0004
Cumulative Proportion   0.866 0.9789 0.99960 1.0000
> fit1$sdev^2 # eigenvalues
[1] 517.7969  67.4964  12.4054   0.2372
> fit1$rotation
       PC1     PC2     PC3    PC4
x1  0.0678  0.6460 -0.5673 0.5062
x2  0.6785  0.0200  0.5440 0.4933
x3 -0.0290 -0.7553 -0.4036 0.5156
x4 -0.7309  0.1085  0.4684 0.4844
> fit1$x
           PC1      PC2     PC3     PC4
 [1,] -36.8218   6.8709  4.5909  0.3967
 [2,] -29.6073  -4.6109  2.2476 -0.3958
 [3,]  12.9818   4.2049 -0.9022 -1.1261
 [4,] -23.7147   6.6341 -1.8547 -0.3786
 [5,]   0.5532   4.4617  6.0874  0.1424
 [6,]  10.8125   3.6466 -0.9130 -0.1350
 [7,]  32.5882  -8.9798  1.6063  0.0818
 [8,] -22.6064 -10.7259 -3.2365  0.3243
 [9,]   9.2626  -8.9854  0.0169 -0.5437
[10,]   3.2840  14.1573 -7.0465  0.3405
[11,]  -9.2200 -12.3861 -3.4283  0.4352
[12,]  25.5849   2.7817  0.3867  0.4468
[13,]  26.9032   2.9310  2.4455  0.4116
```

The first two PC's account for nearly 98% of the total variation among x_1, x_2, x_3, x_4. PC1 is roughly proportional to the difference $x_2 - x_4$, PC2 to the difference $x_1 - x_3$, PC3 to the difference in the average of x_2 and x_3 and that of x_1 and x_4, and PC4 to the average of all four x's. The opposite signs on the factor loadings on x_2 and x_4 in

PC1 and on x_1 and x_3 in PC2 correspond to the large negative correlations between these pairs of variables as shown in Example 4.9. ■

5.3.2 Principal Components Regression (PCR)

To compare the MLR and PCR models, it will be convenient to center both the y_i's and the x_{ij}'s; thus the columns of $\boldsymbol{X}$ sum to 0. Since $\boldsymbol{Z} = \boldsymbol{XU}$, it follows that the columns of $\boldsymbol{Z}$ also sum to 0, i.e., the z_{ij} are also centered. Thus both MLR and PCR models don't have the intercept terms, β_0 and γ_0, respectively. The LS estimates of γ_0 and β_0 for the uncentered data can be readily computed from $\widehat{\gamma}_0 = \overline{y}$ and $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \widehat{\beta}_j \overline{x}_j$.

Denoting the parameter vectors for the MLR model by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and that for the PCR model by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, the two models are $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$. Since $\boldsymbol{UU}' = \boldsymbol{I}$, it follows that

$$\boldsymbol{\beta} = \boldsymbol{U\gamma} \quad \text{and} \quad \boldsymbol{\gamma} = \boldsymbol{U}'\boldsymbol{\beta}. \tag{5.10}$$

If the full model with all p predictors is the true model then the PCR model with $r < p$ PC's is a biased model. But the bias should be small if the first r PC's capture most of the variation among the x's.

Denote the submatrix of $\boldsymbol{Z}$ with only only the first $r < p$ column vectors by $\boldsymbol{Z}^{(r)}$ and the corresponding subvector of $\boldsymbol{\gamma}$ by $\boldsymbol{\gamma}^{(r)} = (\gamma_1, \ldots, \gamma_r)'$. Note that $\boldsymbol{Z}^{(r)} = \boldsymbol{XU}^{(r)}$ where $\boldsymbol{U}^{(r)}$ is the submatrix of $\boldsymbol{U}$ consisting of the first r loading vectors, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$. We have $\boldsymbol{Z}^{(r)'}\boldsymbol{Z}^{(r)} = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$. Hence

$$\widehat{\boldsymbol{\gamma}}^{(r)} = (\boldsymbol{Z}^{(r)'}\boldsymbol{Z}^{(r)})^{-1}\boldsymbol{Z}^{(r)'}\boldsymbol{y} = \text{diag}\{1/\lambda_1, \ldots, 1/\lambda_r\}\boldsymbol{Z}^{(r)'}\boldsymbol{y}. \tag{5.11}$$

Thus $\widehat{\gamma}_j^{(r)} = (1/\lambda_j)(\boldsymbol{Z}^{(r)'}\boldsymbol{y})_j$ $(1 \le j \le r)$. Furthermore, since

$$\text{Cov}(\widehat{\boldsymbol{\gamma}}^{(r)}) = \sigma^2(\boldsymbol{Z}^{(r)'}\boldsymbol{Z}^{(r)})^{-1} = \sigma^2\text{diag}\{1/\lambda_1, \ldots, 1/\lambda_r\},$$

the estimates $\widehat{\gamma}_j^{(r)}$ are uncorrelated. If some principal components z_k are added to or deleted from the model then the estimates $\widehat{\gamma}_j^{(r)}$ for $j \neq k$ are unchanged. Hence a stepwise method is not needed to build a regression model. Thus it is much easier to fit the PCR model than to fit the MLR model.

One drawback of PCR is that the PC's are not always easily interpretable, being linear combinations of the original predictors x_j. However, the final model in original variables can always be recovered using the relationship

$$\widehat{\boldsymbol{\beta}}^{(r)} = \boldsymbol{U}^{(r)}\widehat{\boldsymbol{\gamma}}^{(r)} = \boldsymbol{U}^{(r)}(\boldsymbol{Z}^{(r)'}\boldsymbol{Z}^{(r)})^{-1}\boldsymbol{Z}^{(r)'}\boldsymbol{y}. \tag{5.12}$$

Note that although $\widehat{\boldsymbol{\gamma}}^{(r)}$ has r components, in general, $\widehat{\boldsymbol{\beta}}^{(r)}$ has all p components, i.e., the model includes all p original predictors. But it is not a full model because it is not based on all PC's.

EXAMPLE 5.5 (Hald Cement Data: Principal Components Regression)

First we fit a full PCR model using all four PC's, where PC's are extracted using `prcomp` as in Example 5.4 and the scores are stored in matrix `x`.

```
> fullfit=lm(cement$y~fit1$x)
> summary(fullfit)

Call:
lm(formula = cement$y ~ fit1$x)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 95.42308    0.67840 140.659 7.30e-15 ***
fit1$xPC1    0.55365    0.03103  17.842 9.97e-08 ***
fit1$xPC2    0.91964    0.08595  10.700 5.11e-06 ***
fit1$xPC3   -0.71105    0.20048  -3.547  0.00754 **
fit1$xPC4    1.01954    1.44995   0.703  0.50190
---

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

The coefficients $\widehat{\gamma}_j$ of the first three PC's are highly significant and the intercept $\widehat{\gamma}_0 = 95.4231 = \overline{y}$.

The $\widehat{\beta}_j$ coefficients can be computed from the $\widehat{\gamma}_j$ coefficients as follows:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{U}\widehat{\boldsymbol{\gamma}} = \begin{bmatrix} 0.068 & 0.640 & -0.567 & 0.506 \\ 0.679 & 0.020 & 0.544 & 0.493 \\ -0.029 & -0.755 & -0.404 & 0.516 \\ -0.731 & 0.108 & 0.468 & 0.484 \end{bmatrix} \begin{bmatrix} 0.554 \\ 0.920 \\ -0.711 \\ 1.020 \end{bmatrix} = \begin{bmatrix} 1.551 \\ 0.510 \\ 0.102 \\ -0.144 \end{bmatrix}.$$

This matrix multiplication can be done in R using the command

```
fit1$rotation[,1:4] %*% matrix(fullfit$coefficients[2:5]).
```

Finally, using $\overline{x}_1 = 7.462, \overline{x}_2 = 48.150, \overline{x}_3 = 11.770, \overline{x}_4 = 30.000$ we can calculate

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y}_0 - \sum_{j=1}^{4} \widehat{\beta}_j \overline{x}_j \\ &= 95.423 - [(1.551)(7.462) + (0.510)(48.150) + (0.102)(11.770) + (-0.144)(30.000)] \\ &= 62.412. \end{aligned}$$

Thus the final model is

$$\widehat{y} = 62.412 + 1.551x_1 + 0.510x_2 + 0.102x_3 - 0.144x_4,$$

which is the same model (except for round-off errors) that we obtained in Example 4.10. We conclude that the full MLR and PCR models are the same.

Since only the first three PC's have significant coefficients and those three PC's account for almost 100% of the variation among the x's, we next fit a partial PCR model using the PC1, PC2 and PC3 as predictors resulting in the following output.

```
> partialfit=lm(cement$y~PC1+PC2+PC3, cbind(cement, fit1$x))
> summary(partialfit)

Call:
lm(formula = cement$y ~ PC1 + PC2 + PC3, data = cbind(cement, fit1$x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 95.42308    0.65907 144.784  < 2e-16 ***
PC1          0.55365    0.03015  18.366 1.92e-08 ***
PC2          0.91964    0.08350  11.014 1.59e-06 ***
```

```
PC3          -0.71105    0.19476  -3.651  0.00531 **
---
Residual standard error: 2.376 on 9 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.975
F-statistic: 157.3 on 3 and 9 DF,  p-value: 4.307e-08
```

Notice that the estimated regression coefficients are unchanged from the full model because of the orthogonality of the PC's. Their standard errors have changed, of course, because the MSE of the partial model has changed, both because of the increase in SSE and in error d.f. In this example, the MSE and hence the standard errors have actually decreased. Also R^2 has decreased only slightly from 98.24% to 98.13%.

The coefficients of the regression model in terms of the x's based on three PC's can be obtained using matrix multiplication as mentioned before.

```
> fit1$rotation[,1:3] %*% matrix(partialfit$coefficients[2:4])
x1  1.035031573
x2  0.007260283
x3 -0.423732982
x4 -0.637943480
```

Next we calculate

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y}_0 - \sum_{j=1}^{4} \widehat{\beta}_j \overline{x}_j \\ &= 95.423 - [(1.035)(7.462) + (0.007)(48.150) + (-0.424)(11.770) + (-0.638)(30.000)] \\ &= 111.49.\end{aligned}$$

Thus the final model is

$$\widehat{y} = 111.49 + 1.035x_1 + 0.007x_2 - 0.424x_3 - 0.638x_4,$$

which is quite different from the model obtained using all four PC's, but it gives essentially the same fit.

The functions `prcomp` and `princomp` perform PCA. To obtain the PCR model in terms of the original predictor variables, two further steps are required. First we fit a regression model with the PC's as the predictors. Next we transform the regression coefficients of the PC's to those of the x's by doing matrix multiplication as illustrated above. The function `pcr` in the `pls` library (Mevin and Wehrens, 2007) can be used to perform PCR directly. Another advantage of using `pcr` is that it allows choosing the number of PC's via cross-validation instead of fixing them *a priori*. We illustrate the use of this function below. We use the option `scale=FALSE` to fit a model with unstandardized data so that the results are comparable to those obtained above.

```
> library(pls)
> cement=read.csv("c:/data/cement.csv")
> fit3=pcr(y~., data=cement, scale=FALSE, validation="CV")
> summary(fit3)
Data:   X dimension: 13 4
        Y dimension: 13 1
Fit method: svdpc
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 10 random segments.
```

```
       (Intercept)  1 comps  2 comps  3 comps  4 comps
CV           15.66    9.286    4.466    2.785    3.006
adjCV        15.66    9.214    4.425    2.736    2.942

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps
X    86.60    97.89    99.96   100.00
y    70.13    95.36    98.13    98.24
> fit3$loadings

Loadings:
   Comp 1 Comp 2 Comp 3 Comp 4
x1          0.646 -0.567  0.506
x2  0.679         0.544  0.493
x3        -0.755 -0.404  0.516
x4 -0.731  0.108  0.468  0.484
```

The root mean square error of prediction (RMSEP) is calculated using ordinary cross-validation (CV) and bias-corrected cross-validation (called adjCV). We see that RMSEP using both CV and adjCV is minimized with three PC's, which also explain almost 100% of the variance among the x's and more than 98% of the variance of y. The loading matrix is the same as that obtained with `prcomp` except that very small loadings (< 0.1) are zeroed out. The coefficients for the regression model using three PC's are shown below.

```
> fit3$coefficients[,,3]
          x1           x2           x3           x4
 1.035031573  0.007260283 -0.423732982 -0.637943480
```

They match exactly with those obtained above using `prcomp`. ■

5.4 Partial Least Squares (PLS)

The PLS methodology was proposed by H. Wold (1966) as an econometric technique. It is now applied in a wide variety of disciplines, especially where the predictors are **high dimensional** ($p > n$) and highly correlated. PLS can also deal with multivariate responses (**multivariate regression**). Such applications arise in chemometrics, genomics, proteotomics and social sciences. For example, in chemometrics PLS is used to build a regression model to estimate or predict the percentages of various ingradients in chemical products from spectroscopic measurements at hundreds of frequencies. Thus the response is multivariate (percentages of ingradients) and predictors (spectroscopic measurements) are very large in number (p) typically exceeding the number of samples (n) used for calibration. In the following, initially we assume the multivariate response setting but then specialize it to the univariate response setting that is the focus of the present book.

Suppose that there are p predictors, $x_1, \ldots, x_p$, and q responses, $y_1, \ldots, y_q$. For convenience, assume that all variables are mean-centered. Denote the data matrix of the x's by $\boldsymbol{X} : n \times p$ and that of the y's by $\boldsymbol{Y} : n \times q$. It is desired to fit a regression model of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Because of the high dimensionality and resulting multicollinearity among the x's, this is a very difficult if not an impossible problem for LS estimation. The goal of PLS is to

find lower dimensional representations, $\boldsymbol{Z}$ and $\boldsymbol{W}$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, such that $\boldsymbol{Z}$ and $\boldsymbol{W}$ capture most of the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and then run a regression of $\boldsymbol{W}$ on $\boldsymbol{Z}$. An iterative algorithm is needed to find these matrices. The PLS1 algorithm is used for univariate response ($q = 1$) while the PLS2 algorithm is used for multivariate response ($q > 1$). Here we consider only PLS1.

PLS1 Algorithm

Since the response is univariate, the matrix $\boldsymbol{Y}$ is actually a vector $\boldsymbol{y}$; thus its dimension cannot be reduced. The idea of the algorithm is to extract loading vectors $\boldsymbol{u}_j$ for the x's such that the resulting score vectors $\boldsymbol{z}_j$ are good predictors of $\boldsymbol{y}$. This is done by sequentially extracting a single loading vector $\boldsymbol{u}_j$ at each step j by regressing the residuals of $\boldsymbol{y}$ on the residuals of $\boldsymbol{X}$ from the previous step. The final result is the reduced rank scores matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r]$. The column vectors of $\boldsymbol{Z}$, which are analogous to PC's, are called **latent variables (LV's)**. In the examples in the sequel, they are labeled as components in the R outputs but to distinguish them from PC's we refer to them as LV's.

The PLS1 algorithm operates as follows.

Step 0: Let $\boldsymbol{X}_0 = \boldsymbol{X}$ and $\boldsymbol{y}_0 = \boldsymbol{y}$. In subsequent iterations $\boldsymbol{X}_j$ and $\boldsymbol{y}_j$ consist of residuals of $\boldsymbol{X}_{j-1}$ and $\boldsymbol{y}_{j-1}$. Set $j = 1$ and go to the next step.

Step j: 1. Calculate the loading vector by

$$\boldsymbol{u}_j = \frac{\boldsymbol{X}'_{j-1}\boldsymbol{y}_{j-1}}{||\boldsymbol{X}'_{j-1}\boldsymbol{y}_{j-1}||}.$$

2. Compute the score vector

$$\boldsymbol{z}_j = \boldsymbol{X}_{j-1}\boldsymbol{u}_j.$$

Note that the columns in $\boldsymbol{X}_{j-1}$, which are highly correlated with $\boldsymbol{y}_{j-1}$ receive high loadings and vice versa.

3. Regress the predictor and response variable residuals on the scores obtained in the previous step. Denote the regression coefficient vector for $\boldsymbol{X}_{j-1}$ by $\boldsymbol{a}_j$ and the regression coefficient for $\boldsymbol{y}_{j-1}$ by b_j, where

$$\boldsymbol{a}_j = \frac{\boldsymbol{X}'_{j-1}\boldsymbol{z}_j}{\boldsymbol{z}'_j\boldsymbol{z}_j} \quad \text{and} \quad b_j = \frac{\boldsymbol{y}'_{j-1}\boldsymbol{z}_j}{\boldsymbol{z}'_j\boldsymbol{z}_j}.$$

4. Compute the residuals of $\boldsymbol{X}_{j-1}$ and $\boldsymbol{y}_{j-1}$:

$$\boldsymbol{X}_j = \boldsymbol{X}_{j-1} - \boldsymbol{z}_j\boldsymbol{a}'_j \quad \text{and} \quad \boldsymbol{y}_j = \boldsymbol{y}_{j-1} - \boldsymbol{z}_j b_j.$$

Note that $\boldsymbol{X}'_j\boldsymbol{z}_j = \boldsymbol{0}$ (where $\boldsymbol{0}$ is a p-dimensional null vector) and $\boldsymbol{y}'_j\boldsymbol{z}_j = 0$.

5. Increment $j \to j + 1$ and return to the beginning of this step until $\boldsymbol{X}_j$ and $\boldsymbol{y}_j$ become sufficiently small. Denote by r the value of j at the last step (the reduced rank of $\boldsymbol{X}$) and let $\boldsymbol{Z}^{(r)} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r]$, $\boldsymbol{U}^{(r)} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$ and $\boldsymbol{A}^{(r)} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r]$

The PLS model is obtained by regressing $\boldsymbol{y}$ on $\boldsymbol{Z}^{(r)}$. For practical use, this model needs to be expressed in terms of the original predictor matrix $\boldsymbol{X}$. This relationship is not easy to derive since $\boldsymbol{Z}^{(r)} \neq \boldsymbol{X}\boldsymbol{U}^{(r)}$ because the column vectors $\boldsymbol{u}_j$ of $\boldsymbol{U}^{(r)}$ are not obtained from $\boldsymbol{X}$ itself but from its residuals at successive steps. Also the $\boldsymbol{u}_j$ are not orthonormal vectors, so $\boldsymbol{U}^{(r)}\boldsymbol{U}^{(r)'} \neq \boldsymbol{I}$ as in the case of PCA. Using some algebra it can be shown that

$\boldsymbol{Z}^{(r)} = \boldsymbol{X}\boldsymbol{R}$ where $\boldsymbol{R} = \boldsymbol{U}^{(r)}(\boldsymbol{A}^{(r)'}\boldsymbol{U}^{(r)})^{-1}$. Thus $\widehat{\boldsymbol{\beta}}$ can be obtained by regressing $\boldsymbol{y}$ on $\boldsymbol{Z}^{(r)}$ and then premultiplying the resulting LS estimator by $\boldsymbol{R}$, which gives

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R}(\boldsymbol{Z}^{(r)'}\boldsymbol{Z}^{(r)})^{-1}\boldsymbol{Z}^{(r)'}\boldsymbol{y}.$$

Then the PLS model can be written as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

To illustrate the PLS methodology we will give two examples using the GPA data and the Hald cement data, which are both small data sets, and one example with high dimensional ($p > n$) data.

EXAMPLE 5.6 (GPA Data: Partial Least Squares)

The R script and the resulting output are as follows.

```
> library(pls)
> gpa = read.csv("c:/data/GPA.csv")
> plsfit=plsr(GPA~.,data=gpa, scale=FALSE)
> summary(plsfit)
Data:   X dimension: 40 2
        Y dimension: 40 1
Fit method: kernelpls
Number of components considered: 2
TRAINING: % variance explained
     1 comps  2 comps
X      48.44   100.00
GPA    65.05    68.11
> plsfit$loadings

Loadings:
       Comp 1 Comp 2
Verbal  0.901 -0.663
Math    0.491  0.749

               Comp 1 Comp 2
SS loadings     1.053  1.000
Proportion Var  0.526  0.500
Cumulative Var  0.526  1.026
> plsfit$coefficients
, , 1 comps

              GPA
Verbal 0.02972051
Math   0.02629076

, , 2 comps

              GPA
Verbal 0.02573212
Math   0.03361487
```

We see that the loadings on the two components are quite different from those for PCA. Here LV1 loadings on Verbal and Math are 0.901 and 0.491, respectively; both are

positive unlike those for PCA. Thus LV1 is positively sloping while PC1 is negatively sloping in the (Verbal, Math) space. This is because the PLS loadings take into account the correlations of Verbal and Math with GPA and both Verbal and Math are positively correlated with GPA. Also notice that the proportion of variance explained by the two LV's exceeds 100%. This is because the LV's are not orthogonal. Finally note that in the full model (with two LV's) coefficients of Verbal and Math are the same as those for the MLR model from Example 3.3. ■

EXAMPLE 5.7 (Hald Cement Data: Partial Least Squares)

Next we apply PLS to the Hald cement data.

```
> library(pls)
> cement=read.csv("c:/data/cement.csv")
> plsfit=plsr(y~.,data=cement, scale=FALSE)
> summary(plsfit)
Data:   X dimension: 13 4
        Y dimension: 13 1
Fit method: kernelpls
Number of components considered: 4
TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps
X    86.14    97.84    99.96   100.00
y    76.53    96.38    98.13    98.24
Loadings:
   Comp 1 Comp 2 Comp 3 Comp 4
x1          0.638  0.459  0.487
x2  0.690 -0.141 -0.538  0.512
x3        -0.726  0.518  0.501
x4 -0.741  0.250 -0.482  0.500

               Comp 1 Comp 2 Comp 3 Comp 4
SS loadings     1.036  1.016  1.001  1.000
Proportion Var  0.259  0.254  0.250  0.250
Cumulative Var  0.259  0.513  0.763  1.013
```

Note that the LV1 loadings on x_2 and x_4 are almost identical to the PC1 loadings obtained in Example 5.4 and they are roughly proportional to Corr$(x_2, y) = 0.816$ and Corr$(x_4, y) = -0.821$, respectively. We fit the model using three LV's since they capture almost all of the variance of the x's as well as of y. The regression coefficients of the x's corresponding to three LV's are as follows.

```
> plsfit$coefficients[,,3]
         x1          x2          x3          x4
 1.04728616  0.01848865 -0.41135995 -0.62687397
```

Next $\widehat{\beta}_0$ can be calculated as before:

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y}_0 - \sum_{j=1}^{4} \widehat{\beta}_j \overline{x}_j \\ &= 95.423 - [(1.047)(7.462) + (0.018)(48.150) + (-0.411)(11.770) + (-0.627)(30.000)] \\ &= 110.39.\end{aligned}$$

Thus the final model is

$$\widehat{y} = 110.39 + 1.047x_1 + 0.018x_2 - 0.411x_3 - 0.627x_4.$$

This model is almost the same as the model obtained using PCR with three PC's. ■

Next we give an example that illustrates the use of PLS for high dimensional data ($p > n$).

EXAMPLE 5.8 (Soybean Data Set: Partial Least Squares)

The data set (`soybeandata.csv`) taken from Minitab consists of spectroscopic data on $n = 54$ soybean samples at $p = 88$ frequencies. The goal is to build a predictive model for the response variable Fat (fat content of soybean sample) from the spectroscopic data. The maximum number of LV's that can be extracted from this data equals $\min(n-1, p) = 53$. Obviously, we don't want to have so many LV's. Here we give RMSEP results using 5-fold cross-validation for the first 13 LV's out of the first 20 LV's considered (results for additional LV's are suppressed to save space).

```
> library(pls)
> soybean=read.csv("c:/data/soybeandata.csv")
> plsfit=plsr(Fat~.,data=soybean, scale=FALSE, validation="CV")
> fit1=plsr(Fat~.,data=soybeantrain, ncomp=20, scale=FALSE,
  validation="CV",segments=5)
> summary(fit1)
Data:   X dimension: 54 88
        Y dimension: 54 1
Fit method: kernelpls
Number of components considered: 20

VALIDATION: RMSEP
Cross-validated using 5 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           2.391    2.527    1.721    1.407    1.243    1.181    1.129
adjCV        2.391    2.500    1.699    1.428    1.238    1.163    1.131
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       1.214    1.257    1.248     1.265     1.299     1.369     1.403
adjCV    1.194    1.201    1.191     1.207     1.236     1.289     1.308

TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X     99.547    99.73    99.85    99.94    99.98    99.99   100.00   100.00
Fat    4.427    56.02    69.28    77.27    80.62    81.70    83.87    88.26
     9 comps  10 comps  11 comps  12 comps  13 comps
X     100.00    100.00    100.00    100.00    100.00
Fat    89.02     89.59     89.95     91.12     92.05
```

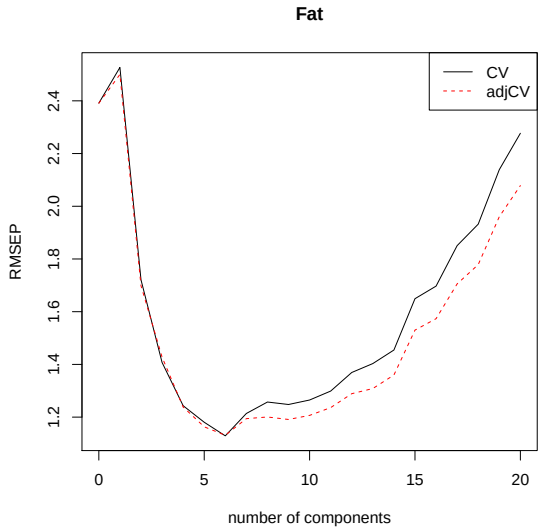


Figure 5.9 RMSEP plot for the PLS model for soybean data

From the above output we observe that RMSEP using both CV and adjCV (which are quite close) are minimized when the number of LV's is six. The percentage of the variance of the x's accounted for by six LV's is 99.99% and that of y is 81.70%.To see how RMSEP behaves as a function of the number of LV's, it is useful to plot it, which is done in Figure 5.9. We see that RMSEP drops steeply until the number of LV's equal to six and then it increases steadily.

A plot of predicted Fat values versus measured Fat values is shown in Figure 5.10. It shows that the relationship between predicted and measured values is linear with no outliers or systematic departures.

The final model consists of 88 predictors, which are the spectral measurements at 88 frequencies. It is not convenient to give the regression coefficients for so many predictors nor is it useful. Instead their plot versus the frequencies may be useful for interpretive purposes by identifying the frequencies at which the peaks in the plot occur and labeling them as most predictive of the response variable. This plot is shown in Figure 5.11. We see that the peaks in the coefficients plots occur at high frequencies. Prediction of Fat values for new data can be done by using the `predict` function in the usual manner.

Finally, Figure 5.12 gives a plot of the loadings of 88 predictors on LV1 and LV2. Notice that all the loadings on LV1 are negative with the highest negative values at frquencies numbered between 50 to 60 and 70 to 88. If we regress Fat on LV1 (i.e., the first column of the `scores` matrix) we get a positive coefficient. In other words, large values of spectrometric readings at these frequencies are predictive of low values of Fat. ■

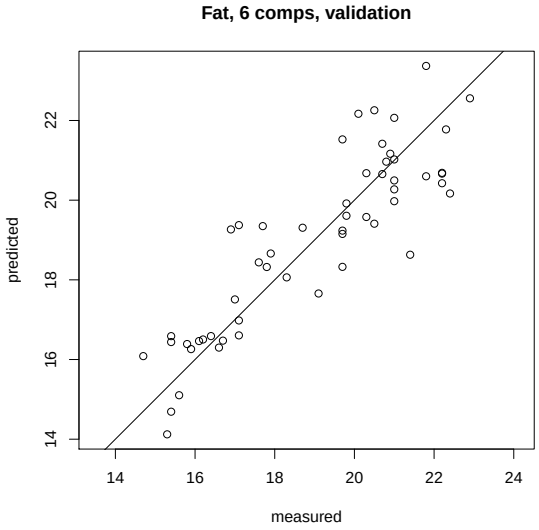


Figure 5.10 Predicted versus measured plot for soybean data

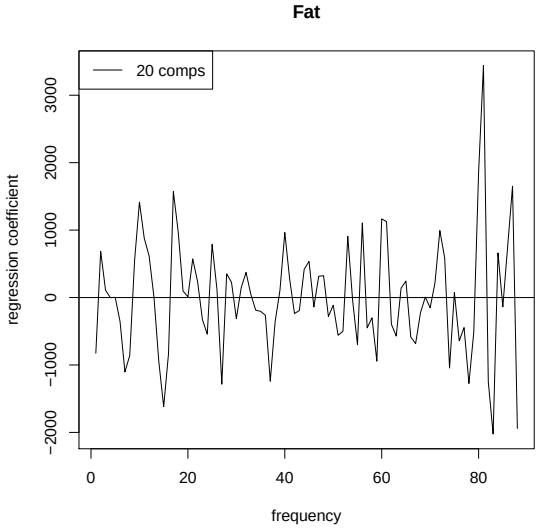


Figure 5.11 Coefficients plot for soybean data

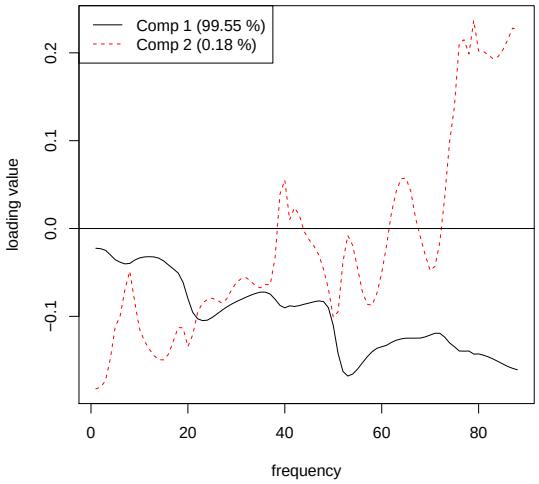


Figure 5.12 Loadings plot for soybean data

5.5 Technical Notes

5.5.1 Properties of Ridge Estimator

From (5.3) we can write $\widehat{\boldsymbol{\beta}}^R(\lambda) = [\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the LS estimator of $\boldsymbol{\beta}$. Hence

$$E[\widehat{\boldsymbol{\beta}}^R(\lambda)] = [\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\beta},$$

and so $\widehat{\boldsymbol{\beta}}^R(\lambda)$ is biased unless $\lambda = 0$ when the ridge estimator reduces to the LS estimator.

Using the sandwich formula (A.7), the covariance matrix of $\widehat{\boldsymbol{\beta}}^R(\lambda)$ for fixed λ is given by

$$\text{Cov}(\widehat{\boldsymbol{\beta}}^R(\lambda)) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} = \sigma^2\boldsymbol{V} \quad \text{(say)}. \tag{5.13}$$

The standard errors of the individual ridge estimates $\widehat{\beta}_j^R(\lambda)$ are given by $\text{SE}(\widehat{\beta}_j^R(\lambda)) = s\sqrt{v_{jj}}$ where v_{jj} is the jth diagonal entry of the $\boldsymbol{V}$ matrix defined above and s^2 is an estimate of σ^2 obtained from ridge regression.

The following method of calculating s^2 for ridge regression is given by Cule et al. (2011). First calculate the fitted vector for ridge regression:

$$\widehat{\boldsymbol{y}}^R(\lambda) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^R(\lambda) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \quad \text{(say)},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'$ is the hat matrix for ridge regression analogous to the hat matrix for multiple linear regression defined in (3.10). Hastie and Tibshirani (1990) defined the **effective error d.f.** as $\nu = n - \text{tr}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}')$. Note that for multiple linear regression $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{H}$ and hence $\nu = n - \text{tr}(\boldsymbol{H}) = n - p$ if the variables are standardized. Then the MSE estimate of σ^2 is given by

$$s^2 = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}}^R(\lambda))'(\boldsymbol{y} - \widehat{\boldsymbol{y}}^R(\lambda))}{\nu}.$$

So the significance of ridge estimates can be tested using $t_j = \widehat{\beta}_j^R(\lambda)/\text{SE}(\widehat{\beta}_j^R(\lambda))$ as approximate t-statistics with ν d.f.

The mean square error of $\widehat{\boldsymbol{\beta}}^R(\lambda)$ is given by

$$\begin{aligned}\text{MSE}(\widehat{\boldsymbol{\beta}}^R(\lambda)) &= E\left[(\widehat{\boldsymbol{\beta}}^R(\lambda)-\boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}}^R(\lambda)-\boldsymbol{\beta})\right] \\ &= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+\lambda)^2}+\lambda^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\beta},\end{aligned}$$

where the λ_j are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and λ is the tuning parameter for ridge regression. The first term represents the sum of the variances of $\widehat{\beta}_j^R(\lambda)$ and the second term represents the sum of the squares of their biases. For the ridge estimator the variance part is smaller than that of the LS estimator but the bias part is positive whereas that for the LS estimator the bias part is 0. By choosing a suitable λ we can trade off the bias against the variance and minimize $\text{MSE}(\widehat{\boldsymbol{\beta}}^R(\lambda))$.

5.5.2 Derivation of Principal Components

Setting the derivative of the objective function in (5.8) w.r.t. $\boldsymbol{u}_1$ equal to $\mathbf{0}$, we get

$$2(\boldsymbol{\Sigma}-\lambda_1\boldsymbol{I})\boldsymbol{u}_1=\mathbf{0}.$$

It follows that $\boldsymbol{u}_1$ is the eigenvector associated with the eigenvalue λ_1. To determine which of the p eigenvalues is λ_1, premultiply the above equation by $\boldsymbol{u}_1'$ resulting in

$$\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1=\lambda_1\boldsymbol{u}_1'\boldsymbol{u}_1=\lambda_1.$$

But $\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_1 = \text{Var}(z_1)$, which is what we want to maximize. Therefore λ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}$.

The constrained optimization problem to find PC2 is as follows:

$$\max_{\boldsymbol{u}_2}\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_2 \quad \text{subject to} \quad \boldsymbol{u}_2'\boldsymbol{u}_2=1 \quad \text{and} \quad \boldsymbol{u}_1'\boldsymbol{u}_2=0.$$

Using the Lagrangian multiplier λ_2 for the first constraint and μ for the second constraint, the optimization problem becomes

$$\max_{\boldsymbol{u}_2}\left[\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_2-\lambda_2(\boldsymbol{u}_2'\boldsymbol{u}_2-1)-\mu(\boldsymbol{u}_1'\boldsymbol{u}_2)\right].$$

Setting the derivative of this objective function w.r.t. $\boldsymbol{u}_2$ equal to $\mathbf{0}$, we get

$$2(\boldsymbol{\Sigma}-\lambda_2\boldsymbol{I})\boldsymbol{u}_2-\mu\boldsymbol{u}_1=\mathbf{0}.$$

As before, premultiplying the above by $\boldsymbol{u}_2'$ we get

$$2\boldsymbol{u}_2'(\boldsymbol{\Sigma}-\lambda_2\boldsymbol{I})\boldsymbol{u}_2-\mu\boldsymbol{u}_2'\boldsymbol{u}_1=2\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_2-2\lambda_2=0$$

since $\boldsymbol{u}_2'\boldsymbol{u}_1 = 0$ and $\boldsymbol{u}_2'\boldsymbol{u}_2 = 1$. Hence $\lambda_2 = \boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_2 = \text{Var}(z_2)$ is the second largest eigenvalue of $\boldsymbol{\Sigma}$. Next, premultiplying the above by $\boldsymbol{u}_1'$ we get

$$2\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_2-\lambda_2\boldsymbol{u}_1'\boldsymbol{u}_2-\mu\boldsymbol{u}_1'\boldsymbol{u}_1=2\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_2-\mu=0,$$

hence $2\boldsymbol{u}_1'\boldsymbol{\Sigma}\boldsymbol{u}_2 = \mu$. However, premultiplying (5.8) by $\boldsymbol{u}_2'$ we get

$$2\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_1-\lambda\boldsymbol{u}_2'\boldsymbol{u}_1=2\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_1=0,$$

hence $\mu = 0$ and $\boldsymbol{u}_2'\boldsymbol{\Sigma}\boldsymbol{u}_1 = \text{Cov}(z_1, z_2) = 0$. Thus PC2 is not only orthogonal to the first but is also uncorrelated with it. This process can be continued with higher principal components.

EXERCISES

Theoretical Exercises

5.1 (Derivation of ridge estimator) Derive the ridge estimator (5.3) by solving the minimization problem in (5.2).

5.2 (Ridge and lasso regression) This exercise is taken from an example in the book by James et al. (2013, p. 225). Consider a simple set up with $n = p$, no β_0 term and the $\boldsymbol{X}$ matrix is an identity matrix. In that case the LS estimates are obtained by minimizing $\sum_{i=1}^{n}(y_i - \beta_i)^2$ and so $\widehat{\beta}_i = y_i$ $(1 \leq i \leq n)$.

a) Show that the ridge estimators are given by

$$\widehat{\beta}_i^R(\lambda) = \frac{\widehat{\beta}_i}{1+\lambda}.$$

b) Show that the lasso estimators are given by

$$\widehat{\beta}_i^L(\lambda) = \begin{cases} y_i - \lambda/2 & \text{if } y_i > \lambda/2 \\ y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & \text{if } |y_i| < \lambda/2. \end{cases}$$

c) Discuss the differences between the ridge and lasso estimates in the way they shrink the LS estimators.

5.3 (Principal components of a patterned covariance matrix) In some applications, the x variables have a so-called spherical distribution, i.e., they are homoscedastic and equicorrelated, so their covariance matrix is given by

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

Assume that $\rho > 0$.

a) Show that $\boldsymbol{u}_1 = (1/\sqrt{p})(1, 1 \ldots, 1)'$ and $\lambda_1 = 1 + (p-1)\rho$.

b) For $j \geq 2$, show that

$$\boldsymbol{u}_j = \left(\frac{1}{\sqrt{(j-1)j}}, \ldots, \frac{1}{\sqrt{(j-1)j}}, \frac{-(j-1)}{\sqrt{(j-1)j}}, 0, \ldots, 0 \right)'$$

and the corresponding eigenvalues are $\lambda_2 = \cdots = \lambda_p = 1 - \rho$. Write $\boldsymbol{u}_2, \boldsymbol{u}_3$ and $\boldsymbol{u}_p$.

c) Part (a) shows that the the first principal component is proportional to the average of all x's. What is the proportion of variance accounted for by it? How does this proportion depend on ρ?

d) If ρ is negative (but $\rho > -1/(p-1)$ for $\boldsymbol{\Sigma}$ to be nonsingular), explain how the above results change.

Applied Exercises

5.4 (Acetylene data: Ridge regression) Perform ridge regression on the acetylene data. Use optimum λ. given in Exercise 4.11.

5.5 (Acetylene data: Lasso regression) Repeat the above exercise for lasso regression. Are any regression coefficients set equal to zero?

5.6 (Acetylene data: PCR) Perform principal components regression on the acetylene data.

5.7 (Soybean data: PCR) Fit a PCR model with Moisture as the response variable using the `soybean.csv` data set. Choose the first five PC's for fitting the model.

5.8 (Soybean data: PLS) Fit a PLS model with Moisture as the response variable using the `soybean.csv` data set. Choose the first five LV's for fitting the model. Compare the results with those obtained in the previous exercise using the PCR model.

5.9 (Gasoline data: PLS) The `pls` library includes `gasoline` data set, which consists of octane number (octane) and NIR spectra (NIR) of 60 gasoline samples. Each NIR spectrum consists of 401 diffuse reflectance measurements from 900 to 1700 nm. Analyze this high dimensional data using partial least squares following the steps in Example 5.8 and draw conclusions.

CHAPTER 6

MULTIPLE LINEAR REGRESSION: VARIABLE SELECTION AND MODEL BUILDING

Thus far we have assumed that the regression model is fully specified and only its parameters are to be estimated. In practice, often we have a choice of many predictors and the goal is to select a small subset of them that provides a parsimonious yet well-fitting predictive model. In this chapter we discuss methods for predictor variable selection.

We will use two examples to illustrate the methods presented in this chapter. The first example deals with the Hald cement data from Example 4.9. This is a small data set that involves extreme multicollinearity. As we have seen, there are large negative correlations between x_1 and x_3, and x_2 and x_4. However, it is is unclear which variable out of each pair should be retained in the model. Ridge regression performed in Example 5.1 is ambivalent on this issue since it shrinks all four regression coefficients. Lasso regression performed in Example 5.2 drops x_3 from the model but keeps the other three variables. We will analyze this data set by applying the methods of variable selection to see which variables should be kept in the model.

The second example deals with the used car prices data from Example 3.13. This is a large data set that involves many predictor variables. Any data set with a large number of variables is likely to have some dependencies among them. Therefore it is of interest to screen them to find a small subset that gives a good predictive model.

6.1 Best Subset Selection

Denote by m the number of predictor variables, $x_1, \ldots, x_m$, available for selection in the model. Some of them could be functions of the original variables, e.g., x_2 could be x_1^2 or x_3 could be the interaction $x_1 x_2$. Let $p \leq m$ denote the number of predictors in a given model. There are a total of $2^m - 1$ possible models excluding the null model, which has no predictors. (We assume that the intercept is included in every model.) For every $p \leq m$, there are $\binom{m}{p}$ models of size p. The best subset selection method aims at selecting the best model according to a specified criterion by evaluating it for all models. Some commonly used criteria are described below. Efficient algorithms are available to evaluate these criteria without having to actually fit the models. Nevertheless, if m is large then this method may not be feasible because the total number of possible models can be prohibitively large.

6.1.1 Model Selection Criteria

A p-variable model consists of a subset of size p from $x_1, \ldots, x_m$. For convenience, we will renumber the p variables in the model as $x_1, \ldots, x_p$, keeping in mind that these are not necessarily the first p variables from the set $x_1, \ldots, x_m$. To keep the notation simple, we will indicate the dependence of the criterion only on the size p of the model and not on the particular subset of variables included, e.g., we will denote the R^2 criterion for the model consisting of $x_1, \ldots, x_p$ by R_p^2 rather than by $R^2(x_1, \ldots, x_p)$.

R_p^2 Criterion: This criterion compares models based on their R^2 values, which are measures of the goodness of fit. Since R^2 can only increase by adding more variables to the model, it is trivially maximized by including all m variables. Therefore, if this criterion is to be used for model selection, then some condition must be put on it, e.g., choose the smallest model whose R^2 does not increase significantly by adding more variables. The test given in Exercise 3.3 can be used for this purpose. In general, the R_p^2 criterion tends to produce models that are too large. Furthermore, although these models provide good fits, they do not necessarily give good predictions.

We can plot R_p^2 as a function of p as shown in Figure 6.1 for the Hald cement data. We see that $\max R_p^2$ values for each p, where the maximum is taken over all $\binom{m}{p}$ models of size p, increases sharply initially as p increases but soon reaches a plateau with only marginal increases beyond it. This point corresponds to the desired model.

Adjusted R_p^2 Criterion: This criterion was introduced in Section 3.2.4 as a modification of R_p^2 in order to incorporate penalty for the number of variables in the model. It is defined as

$$R^2_{\text{adj},p} = 1 - \frac{\text{SSE}_p/[n-(p+1)]}{\text{SST}/(n-1)} = 1 - \frac{\text{MSE}_p}{\text{MST}}.$$

Note that MST does not depend on p. As p increases, SSE_p decreases but so do the error degrees of freedom. Therefore MSE_p does not necessarily decrease. $R^2_{\text{adj},p}$ reaches a maximum when MSE_p reaches a minimum at some p as shown in Figure 6.1. Thus maximizing $R^2_{\text{adj},p}$ (or equivalently minimizing MSE_p) is a well-defined criterion unlike maximizing R_p^2.

Mallows' C_p Criterion: Mallows' (1973) C_p statistic is an approximately unbiased estimate of the sum of the standardized mean squared errors of prediction (MSEP) of all

n observations. It is given by

$$C_p = \frac{\text{SSE}_p}{\widehat{\sigma}^2} + 2(p+1) - n,$$

where $\widehat{\sigma}^2$ is an unbiased estimate of σ^2. The goal is to minimize C_p. See Section 6.4.1 for the derivation of C_p.

Usually, the MSE for the full model is used for $\widehat{\sigma}^2$, i.e., $\widehat{\sigma}^2 = \text{MSE}_m = \text{SSE}_m/[n - (m+1)]$. This gives

$$C_m = \frac{\text{SSE}_m}{\text{MSE}_m} + 2(m+1) - n = n - (m+1) + 2(m+1) - n = m + 1.$$

It can be shown that if a p-variable model has zero bias then $E(C_p) \approx p + 1$. Since $C_m = m + 1$, we can say that the full model has zero bias and dropping variables from the full model will add bias. If we add more variables to a model, p increases and SSE_p decreases but the penalty term $2(p+1)$ increases. So the minimum C_p is generally attained at some intermediate value of p as shown in Figure 6.2 for the Hald cement data. Note that C_p charges a stiffer penalty for the number of variables than does $R^2_{\text{adj},p}$. Therefore C_p tends to select a more parsimonious model.

AIC_p and BIC_p Criteria: AIC_p stands for **Akaike's information criterion** and is given by

$$\text{AIC}_p = n \ln \text{SSE}_p + 2(p+1) - n \ln n.$$

This is the same formula derived in Equation (9.14) except for the constant term $n \ln 2\pi + n$. AIC_p measures the "information loss" (using the Kullback-Leibler information measure) because of not fitting the true model. Minimizing AIC_p is equivalent to maximizing the expected information in a model subject to a penalty term for the number of variables in the model.

There are many variants of the AIC_p criterion. BIC_p stands for Schwarz's **Bayesian information criterion** and is given by

$$\text{BIC}_p = n \ln \text{SSE}_p + (p+1) \ln n - n \ln n.$$

Note that the multiplying factor $\ln n$ used in the penalty term for the number of variables is greater than 2 used in AIC_p if $n \geq 8$. So the BIC_p criterion charges a stiffer penalty for the number of variables than does AIC_p. Therefore it tends to result in a more parsimonious model.

EXAMPLE 6.1 (Hald Cement Data: Best Subsets Regression)

The values of the above subset selection criteria for the Hald cement data are shown in Table 6.1. For each p, the two best models (in terms of the largest R^2_p) are shown. We see that the $R^2_{\text{adj},p}$, C_p and the BIC_p criteria choose $\{x_1, x_2\}$ as the best model whereas the AIC_p criterion chooses $\{x_1, x_2, x_4\}$ as the best model, which was also chosen by lasso regression; see Example 5.2. However, its AIC_p is only marginally smaller than that of the $\{x_1, x_2\}$ model.

We illustrate the calculation of the criteria for the $\{x_1, x_2\}$ model. For this model $\text{SST} = 2715.76, \text{SSE}(x_1, x_2) = 57.90$ and $\text{MSE}(x_1, x_2) = 5.79$. For the full model $\text{SSE}(x_1, x_2, x_3, x_4) = 47.86$ and $\text{MSE}(x_1, x_2, x_3, x_4) = 5.98$. Therefore

$$R^2_p = 1 - \frac{57.90}{2715.76} = 97.87\%, R^2_{\text{adj},p} = 1 - \frac{5.79}{2715.76/12} = 97.44\%,$$

$$C_p = \frac{57.90}{5.98} + 2(2+1) - 13 = 2.678,$$

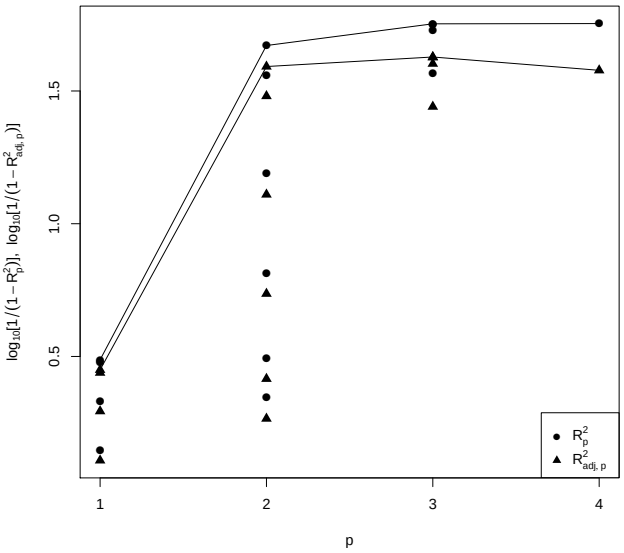


Figure 6.1 $\log_{10}\left(\frac{1}{1-R_p^2}\right)$ and $\log_{10}\left(\frac{1}{1-R_{\text{adj},p}^2}\right)$ as functions of p for the Hald cement data

Table 6.1 Values of the criteria for two best subsets of each size for the Hald cement data

p	Variables	R_p^2	$R_{\text{adj},p}^2$	C_p	AIC$_p$	BIC$_p$
1	x_4	67.45%	64.50%	138.75	58.85	59.98
	x_2	66.62%	63.60%	142.49	59.18	60.31
2	x_1, x_2	97.87%	97.44%	2.678	25.42	27.11
	x_1, x_4	97.25%	96.69%	5.503	28.75	30.44
3	x_1, x_2, x_4	98.23%	97.64%	3.018	24.97	27.23
	x_1, x_2, x_3	98.23%	97.64%	3.042	25.01	27.27
4	x_1, x_2, x_3, x_4	98.24%	97.36%	5.0	26.94	29.76

$$\text{AIC}_p = 13\ln\left(\frac{57.90}{13}\right) + 2(2+1) = 25.418,$$

and

$$\text{BIC}_p = 13\ln\left(\frac{57.90}{13}\right) + (2+1)\ln 13 = 27.113.$$

Figure 6.1 shows the plot of $\log_{10}[1/(1-R_p^2)]$ and $\log_{10}[1/(1-R_{\text{adj},p}^2)]$ versus p. The log scale is used to clearly bring out the differences between R_p^2 and $R_{\text{adj},p}^2$ near their maximum values for each p.

Figure 6.2 shows the plot of $\log_{10} C_p$ versus p. In this plot we have chosen the log scale to shrink the wide range of C_p values (from 2.678 to 202.567). ■

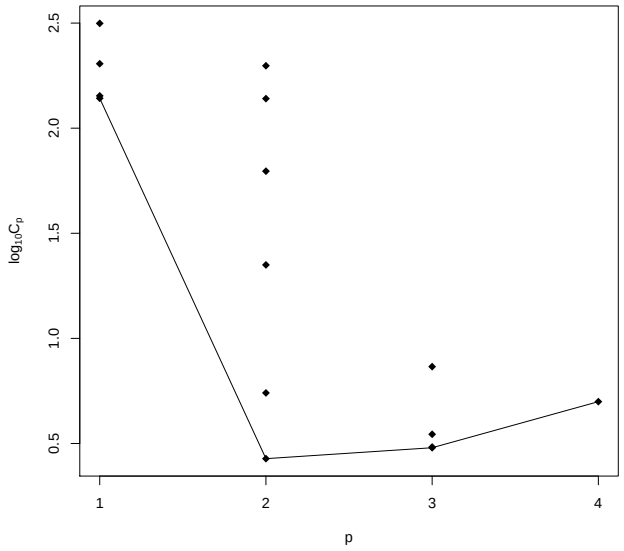


Figure 6.2 $\log_{10} C_p$ as a function of p for the Hald cement data

EXAMPLE 6.2 (Hald Cement Data: Best Subsets Regression Using R)

The following R script produces two best models of each size in terms of the C_p criterion. It uses the library `leaps`.

```
> cement = read.csv("c:/data/cement.csv")
> library(leaps)
> best=leaps(cement[,1:4], cement[,5], method="Cp", nbest=2,
> names=names(cement)[1:4])
> data.frame(size=best$size,Cp=best$Cp,best$which)
```

The output is as follows.

```
  size         Cp    x1    x2    x3    x4
1    2 138.730833 FALSE FALSE FALSE  TRUE
2    2 142.486407 FALSE  TRUE FALSE FALSE
3    3   2.678242  TRUE  TRUE FALSE FALSE
4    3   5.495851  TRUE FALSE FALSE  TRUE
5    4   3.018233  TRUE  TRUE FALSE  TRUE
6    4   3.041280  TRUE  TRUE  TRUE FALSE
7    5   5.000000  TRUE  TRUE  TRUE  TRUE
```

We see that the minimum C_p model is $\{x_1, x_2\}$ with $C_p = 2.678$.

■

EXAMPLE 6.3 (Used Car Prices: Best Subsets Selection)

We applied best subsets regression to the used car prices training data set from Example 3.13 with $n = 402$. The model with the smallest $C_p = 13.03$ (this value depends on what is used as the full model and so may vary) includes 13 variables (Mileage, Cylinders, Liter, Cruise, Buick, Cadillac, Chevrolet, SAAB, Convertible, Coupe, Hatchback, Sedan) as shown below.

$$\begin{aligned}\widehat{\log_{10}(\text{Price})} &= 4.082 - 0.0036\,\text{Mileage} - 0.0141\,\text{Cylinder} + 0.1116\,\text{Liter} + 0.0098\,\text{Cruise} \\ &\quad +0.0408\,\text{Buick} + 0.2464\,\text{Cadillac} - 0.1292\,\text{Chevrolet} + 0.2794\,\text{SAAB} \\ &\quad +0.0720\,\text{Convertible} - 0.0668\,\text{Coupe} - 0.0792\,\text{Hatchback} - 0.0708\,\text{Sedan}.\end{aligned}$$

Recall that Mileage is expressed in thousands of miles. The only nonsignficant variable in the above model is Cruise with P-value equal to 0.098. Surprisingly, the Cylinder variable is significant with P-value equal to 0.042 despite its high correlation of 0.958 with the Liter variable. As a comparison, $C_p = 16.05$ for the 10-variable model fitted in Example 3.13 which includes Liter but not the Cylinder variable. We would probably choose that model as it has less number of variables all of which are highly significant ($P < 0.001$), although its C_p is higher. ■

6.2 Stepwise Regression

As the name suggests, stepwise regression, builds a single model by entering or removing predictors in a stepwise manner (one at a time) according to a set of rules. Note that stepwise regression does not evaluate all models and does not optimize some well-defined criterion of "bestness" as does the best subset selection method.

There are two basic versions of stepwise regression: forward and backward. The forward version starts from the null model and enters variables one at a time. The backward version starts from the full model and removes variables one at a time. For each version there are two types of algorithms. **Forward stepwise algorithm** allows the option of removing the variables entered at previous steps; similarly the **backward stepwise algorithm** allows the option of entering the variables removed at previous steps. On the other hand, **forward selection algorithm** proceeds only in one direction with no option of removing the previously entered variables, while **backward elimination algorithm** proceeds also only in one direction with no option of entering the previously removed variables. We will focus on the forward stepwise algorithm in the present discussion, although in case of multicollinearity a backward algorithm is recommended. Exercise 6.2 gives a data set which illustrates this point.

Different criteria can be used to enter and remove variables in the stepwise algorithm. The classical criterion is the t or equivalently the F-statistic associated with the predictor being considered for entry or removal. We will present the algorithm in terms of this criterion because the significance tests used are the familiar ones introduced in Chapter 3. The `lm.stepwise` function in R uses the AIC criterion as illustrated in Example 6.4.

Suppose that a model includes variables $x_1, \ldots, x_{p-1}$ and we want to decide whether to enter a new variable x_p. Similarly, if the model includes $x_1, \ldots, x_p$ and we want to decide whether to remove one of the variables, say x_p. To make either of these decisions the F_p-statistic defined in (3.32) can be used; the F_p-statistic is denoted by F_{in} or F_{out} depending on whether we are making the enter or remove decision. In each case the partial F-statistic F_p is computed and compared with a critical constant f_{in} for the enter decision or with a critical constant f_{out} for the remove decision where $f_{\text{in}} \geq f_{\text{out}}$. If $F_p > f_{\text{in}}$ then x_p is entered

in the model and if $F_p \leq f_{\text{out}}$ then x_p is removed from the model. Note that these are not formal α-level F-tests since the tests are sequential and p changes at every step.

The algorithm is best represented in the form of a flowchart shown in Figure 6.3. Note that we need $f_{\text{in}} \geq f_{\text{out}}$; otherwise the algorithm can go into an infinite loop. For example, if $f_{\text{in}} = 3.0$ and $f_{\text{out}} = 4.0$ and suppose $F_p = 3.5$ then the variable x_p can enter the model since $F_p > 3.0$ but at the next step x_p will be removed from the model since $F_p < 4.0$. Thus the algorithm will keep recycling x_p in and out of the model.

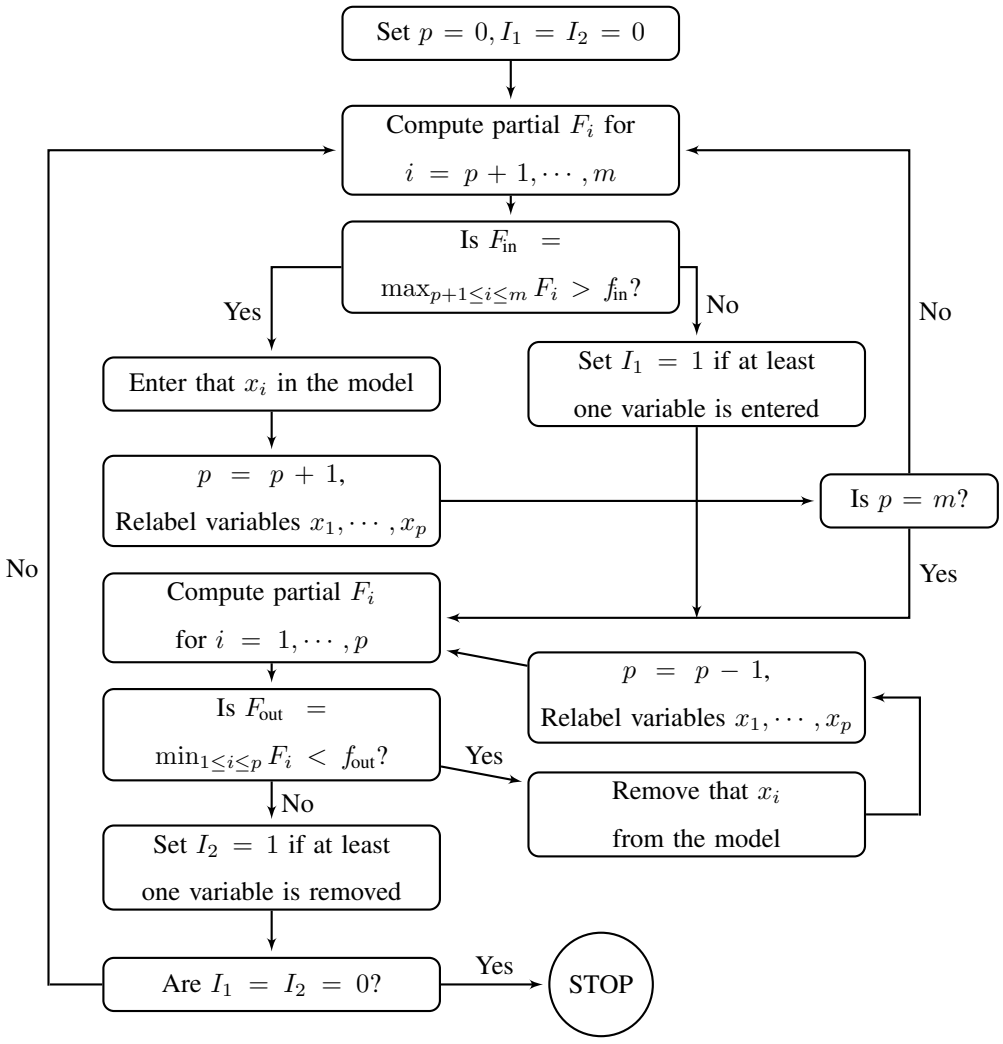


Figure 6.3 Flow chart for forward stepwise algorithm sec6:stepflowchart

EXAMPLE 6.4 (Hald Cement Data: Stepwise Regression)

In this example we will illustrate the stepwise regression algorithm for the Hald cement data from Table 4.4. We will set $f_{\text{in}} = f_{\text{out}} = 4.0$.

Step 1: At Step 1, we decide which variable to enter in the model first. Since there are no variables in the model to begin with, the maximum partial F-statistic corresponds to the maximum absolute bivariate correlation coefficient $|r_{yx_j}|$. The four correlation coefficients are

$$r_{yx_1} = 0.7307, r_{yx_2} = 0.8163, r_{yx_3} = -0.5347, r_{yx_4} = -0.8213,$$

and the corresponding SSE's are

$$\text{SSE}(x_1) = 1265.7, \text{SSE}(x_2) = 906.1, \text{SSE}(x_3) = 1939.3, \text{SSE}(x_4) = 883.9.$$

Also, $\text{SST} = \text{SSE}(\emptyset) = 2715.76$ from the ANOVA table in Example 4.10, where $\emptyset$ denotes a null model with only the intercept term. The maximum absolute bivariate correlation coefficient is $|r_{yx_4}|$, which corresponds to the minimum SSE or $\max F_{\text{in}}$. Using the extra SS method, we calculate the F_{in} for x_4:

$$F_{\text{in}} = \frac{\text{SSE}(\emptyset) - \text{SSE}(x_4)}{\text{MSE}(x_4)} = \frac{2715.76 - 883.9}{883.9/(13-2)} = 22.745 > 4.0.$$

So x_4 enters the model.

Step 2: At Step 2, we choose between x_1, x_2 and x_3 to decide which variable, if any, enters the model next. The variable that gives the smallest SSE will be the candidate. The three SSE values are as follows:

$$\text{SSE}(x_1, x_4) = 74.8, \text{SSE}(x_2, x_4) = 868.9, \text{SSE}(x_3, x_4) = 175.7.$$

Since adding x_1 gives the smallest SSE, it is a candidate for the entry into the model. To test if it meets the entry criterion, we calculate its F_{in} statistic:

$$F_{\text{in}} = \frac{\text{SSE}(x_4) - \text{SSE}(x_1, x_4)}{\text{MSE}(x_1, x_4)} = \frac{883.9 - 74.8}{74.8/(13-3)} = 108.17 > 4.0.$$

Therefore x_1 enters the model.

Before going to the next step, we need to check if x_4, which entered at Step 1, can be removed from the model. So we calculate its F_{out} statistic:

$$F_{\text{out}} = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_4)}{\text{MSE}(x_1, x_4)} = \frac{1265.7 - 74.8}{74.8/(13-3)} = 159.20 > 4.0.$$

Therefore x_4 cannot be removed from the model.

Step 3: Next we check whether x_2 or x_3 can be added to the model. Toward this end we calculate

$$\text{SSE}(x_1, x_2, x_4) = 47.97, \text{SSE}(x_1, x_3, x_4) = 50.84.$$

Since adding x_2 gives a smaller SSE, it is the candidate to enter the model next. To test if it meets the entry criterion, we calculate its F_{in} statistic:

$$F_{\text{in}} = \frac{\text{SSE}(x_1, x_4) - \text{SSE}(x_1, x_2, x_4)}{\text{MSE}(x_1, x_2, x_4)} = \frac{74.8 - 47.97}{47.97/(13-4)} = 5.034 > 4.0.$$

Therefore x_2 enters the model.

Before going to the next step, we need to check if either x_1 or x_4 can be removed from the model. So we calculate

$$\text{SSE}(x_1, x_2) = 57.9 \quad \text{and} \quad \text{SSE}(x_2, x_4) = 868.9.$$

Thus removing x_4 gives a smaller SSE. So we check if it can be removed by calculating its F_{out} statistic:

$$F_{\text{out}} = \frac{\text{SSE}(x_1, x_2) - \text{SSE}(x_1, x_2, x_4)}{\text{MSE}(x_1, x_2, x_4)} = \frac{57.9 - 47.97}{47.97/(13-4)} = 1.863 < 4.0.$$

Therefore we remove x_4 from the model. Recall that x_4 was the first variable to enter the model but it is removed at this step.

It is easy to check that neither x_1 nor x_2 can be removed from the model.

Step 3: Since x_4 was just removed from the model, it cannot re-enter (its F_{in} at this step equals its F_{out} at the previous step). Thus it only remains to check whether x_3 can be entered in the model. We calculate $\text{SSE}(x_1, x_2, x_3) = 48.11$. Hence its

$$F_{\text{in}} = \frac{\text{SSE}(x_1, x_2) - \text{SSE}(x_1, x_2, x_3)}{\text{MSE}(x_1, x_2, x_3)} = \frac{57.9 - 48.11}{48.11/(13-4)} = 1.832 < 4.0.$$

Therefore x_3 cannot enter the model and the algorithm stops with the final model consisting of two variables: x_1 and x_2, which is the same model obtained using the best subsets regression with the C_p criterion.

■

It should be noted that stepwise regression in R uses the AIC criterion and not partial F-tests. Variables are added or removed from the model in a stepwise manner to minimize AIC. The following example illustrates this using the same Hald cement data.

EXAMPLE 6.5 (Hald Cement Data: Stepwise Regression Using R)

Stepwise regression in R finds the model with the minimum AIC by deleting (in case of backward elimination) or adding (in case of forward selection) one variable at each step. Thus it is equivalent to the best subsets regression using the AIC criterion. We can run a backward elimination algorithm using the following R script.

```
> cement = read.csv("c:/data/cement.csv")
> fit1 = lm(y ~ ., cement)
> step(fit1)
```

The resulting output is as follows.

```
Start:  AIC=26.94
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS    AIC
- x3    1    0.1091 47.973 24.974
- x4    1    0.2470 48.111 25.011
- x2    1    2.9725 50.836 25.728
<none>              47.864 26.944
- x1    1   25.9509 73.815 30.576

Step:  AIC=24.97
y ~ x1 + x2 + x4

       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- x4    1      9.93  57.90 25.420
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629

Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)
```

```
Coefficients:
(Intercept)           x1           x2           x4
    71.6483       1.4519       0.4161      -0.2365
```

We see from the output that the algorithm starts with the full model, deletes one variable at a time and calculates the AIC of the resulting model. The results are arranged in the increasing order of AIC. The full model has AIC $= 26.94$. The best model of size three is $\{x_1, x_2, x_4\}$ and has AIC $= 24.97$. From the next step we see that removing any more variables from this model increases AIC. So the overall best model is $\{x_1, x_2, x_4\}$.

Obviously the same result is obtained if we run the forward selection algorithm using the following R script:

```
> fit2 = lm(y ~ 1, cement)
> step(fit2, scope=~x1+x2+x3+x4)
```

The resulting output is as follows.

```
Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4

       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
<none>               883.87 58.852
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444

Step:  AIC=28.74
y ~ x4 + x1

       Df Sum of Sq     RSS    AIC
+ x2    1     26.79   47.97 24.974
+ x3    1     23.93   50.84 25.728
<none>               74.76 28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190.92 1265.69 63.519

Step:  AIC=24.97
y ~ x4 + x1 + x2
```

Table 6.2 Stepwise addition of variables for used car prices regression model

Step	Variable	Partial r	F_{in}	Step	Variable	Partial r	F_{in}
1	Liter	0.590	213.6	6	Wagon	0.390	70.73
2	SAAB	0.785	639.58	7	Chevrolet	-0.244	24.90
3	Cadillac	0.743	491.51	8	Pontiac	-0.155	23.62
4	Mileage	0.457	104.65	9	Saturn	-0.206	17.31
5	Convertible	0.501	132.94	10	Cylinder	-0.142	8.07

```
       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- x4    1      9.93  57.90 25.420
+ x3    1      0.11  47.86 26.944
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629

Call:
lm(formula = y ~ x4 + x1 + x2, data = cement)

Coefficients:
(Intercept)           x4           x1           x2
    71.6483      -0.2365       1.4519       0.4161
```

■

EXAMPLE 6.6 (Used Car Prices: Stepwise Regression)

We applied stepwise regression with $f_{\text{in}} = f_{\text{out}} = 4.0$ on the used car prices training data set from Example 3.13 with $n = 402$. The response variable was $\log_{10}(\text{Price})$. The variables are listed in Table 6.2 in the sequence in which they entered the model along with their partial correlation coefficients and the F_{in} statistics. Note that in this example no variables were removed at any step. Thus the procedure operated as if it is a forward selection procedure. Similar to the best model according to C_p found in Example 6.3, this model also includes the Cylinder variable along with Liter. The fitted equation for this model is

$$\begin{aligned}\widehat{\log(\text{Price})} \;=\;& 4.064 - 0.0035\,\text{Mileage} + 0.1177\,\text{Liter} + 0.2075\,\text{Cadillac} \\ & -0.0572\,\text{Chevrolet} - 0.0393\,\text{Pontiac} + 0.2406\,\text{SAAB} - 0.0475\,\text{Saturn} \\ & +0.1423\,\text{Convertible} + 0.0672\,\text{Wagon} - 0.0183\,\text{Cylinder}.\end{aligned}$$

Once again, recall that Mileage is expressed in thousands of miles.

The C_p statistic for this model can be calculated as follows.The full model consists of the following 15 variables: Mileage, Cylinder, Liter, Cruise, Sound, Leather, Buick, Cadillac, Chevrolet, Pontiac, SAAB, Convertible, Coupe, Hatchback and Sedan. This full model has MSE $= 1.5875 \times 10^{-3}$ and SSE for the above fitted model is 0.6274.

Therefore

$$C_p = \frac{0.6274}{1.5875 \times 10^{-3}} + 2(11) - 402 = 15.2214.$$

■

6.3 Model Building

Regression modeling is an iterative process. Below we give suggested steps in this process.

- **Univariate exploration of the data:** Examine each variable for outliers, wrong or incosistent data entries, missing values, etc. If possible, correct outliers and wrong data entries and fill in missing values as appropriate. If outliers and missing values are a small fraction of the total sample then one may discard those observations. It is better to discard a few observations than to discard some variables as they may be important predictors.

  Examine univariate distributions of all variables and in cases of highly skewed distributions make suitable transformations to symmetrize the distributions. Transformations may also be made based on subject matter knowledge by consulting subject matter experts. Note that the predictor variables don't need to be normally or even symmetrically distributed (in fact, dummy variables are binary). However, highly skewed distributed predictors are likely to result in influential observations and so it is recommended to symmetrize them. Generally, logarithmic or square root transformation is useful for this purpose.

- **Bivariate exploration of the data:** Make scatter plots and compute correlations between all variables (feasible only when there is a modest number of variables, say, less than 10) to identify any bivariate outliers, influential observations and highly correlated x's causing multicollinearity. Bivariate scatter plots between y and each x help suggest linearizing transformations for x's. It is recommended not to transform y at this step unless a common linearizing transformation of y works for all x's or if a subject matter based transformation exists for y. For example, if one wants to predict the gas mileage (miles per gallon or mpg) of a car from its weight, engine size, etc. then an inverse transformation (gallons per mile or gpm) is preferable since the latter is proportional to the amount of energy spent to move the car one mile which in turn is proportional to its weight, engine size, etc.

- **Interactions:** There are automatic interaction detection softwares available which screen all possible two-factor interactions. However, we recommend including only those interactions that are supported by subject matter knowledge.

- **Training set and test set:** Randomly divide the data into a training and a test set. The training set should be large enough to build a reliable model and have enough error d.f. If the data set is sufficiently large, use a 50:50 split; otherwise make the training set bigger.

- **Fit several candidate models:** Use best subsets regression or stepwise regression to come up with three or four good models. If there are many variables (e.g., $p \geq 20$) then the best subsets regression approach may not be feasible in which case it may be

advisable to first select a subset of variables by using stepwise regression followed by the best subsets regression on that subset.

- **Compare the candidate models:** Evaluate the candidate models based on C_p and other criteria such as residual plots and other diagnostics. Choose two or three good models. Check if further transformations of variables are necessary.

- **Selection of the final model:** Choose the final model by comparing the contending models in terms of the total prediction error by applying them to the test set. More generally, the data may be divided into $m \geq 2$ random subsets. One subset is used as a test set while the remaining $m-1$ subsets are used as a training set. This process is repeated m times by leaving out one subset as a test set each time and fitting the model on the remaining $m-1$ subsets. Prediction errors are averaged over all m splits of the data into training and test sets. This is known as m**-fold cross-validation**. If m equals the total sample size n then in each iteration of the algorithm, a single observation is set aside as a test set. The model is fitted by deleting one observation at a time and the prediction error for the deleted observation is calculated. This is known as **leave-one-out (LOO) cross-validation**.

- **Practical checks on the final model:** Check that the model includes the key variables based on subject matter knowledge and their coefficients have the correct signs. The model should pass the standard statistical tests as well as the the subject matter criteria.

6.4 Technical Notes

6.4.1 Derivation of C_p Statistic

C_p is an approximately unbiased sample estimate of the **mean squared error of prediction (MSEP)** using a p-variable model. Let $\widehat{y}_{i,p}$ denote the fitted or predicted value of y_i. The MSEP equals

$$\begin{aligned}
\text{MSEP}_p &= \sum_{i=1}^{n} E\left[(\widehat{y}_{i,p} - y_i)^2\right] \\
&= \sum_{i=1}^{n} E\left[\{(\widehat{y}_{i,p} - E(y_i)) - (y_i - E(y_i))\}^2\right] \\
&= \sum_{i=1}^{n} E\left[(\widehat{y}_{i,p} - E(y_i))^2\right] + \sum_{i=1}^{n} E\left[(y_i - E(y_i))^2\right] \\
&= \sum_{i=1}^{n} E\left[(\widehat{y}_{i,p} - E(y_i))^2\right] + n\sigma^2.
\end{aligned}$$

In the above the cross-product term cancels out since $y_i - E(y_i) = 0$. We can ignore $n\sigma^2$ since it is a constant. Therefore minimizing MSEP_p is equivalent to minimizing Γ_p, where

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E\left[(\widehat{y}_{i,p} - E(y_i))^2\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E\left[\{(\widehat{y}_{i,p} - E(\widehat{y}_{i,p})) + (E(\widehat{y}_{i,p}) - E(y_i))\}^2\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n} \left\{E\left[(\widehat{y}_{i,p} - E(\widehat{y}_{i,p}))^2\right] + \left[(E(\widehat{y}_{i,p}) - E(y_i))^2\right]\right\}.\end{aligned}$$

In the above the cross-product term cancels out since $E[(\widehat{y}_{i,p} - E(\widehat{y}_{i,p}))] = 0$. Note that the first term inside the summation is $\text{Var}(\widehat{y}_{i,p})$ and the second term is $[\text{Bias}(\widehat{y}_{i,p})]^2$. In Section 3.7, we have shown that $\text{Cov}(\widehat{\boldsymbol{y}}) = \text{Cov}(\boldsymbol{H}\boldsymbol{y}) = \sigma^2\boldsymbol{H}$ and $\text{tr}(\boldsymbol{H}) = p + 1$. Hence $\sum_{i=1}^{n}\text{Var}(\widehat{y}_{i,p}) = \sigma^2\text{tr}(\boldsymbol{H}) = \sigma^2(p+1)$. Denote the total bias squared term by $B_p = \sum_{i=1}^{n}\left[(E(\widehat{y}_{i,p}) - E(y_i))^2\right]$. Then

$$\Gamma_p = \frac{1}{\sigma^2}B_p + (p+1).$$

It can be shown that $E(\text{SSE}_p) = B_p + [n - (p+1)]\sigma^2$. (This equation shows that $\text{MSE}_p = \text{SSE}_p/[n-(p+1)]$ is an unbiased estimate of σ^2 when $B_p = 0$.) Therefore an approximately unbiased estimator of Γ_p is given by

$$C_p = \frac{\text{SSE}_p}{\widehat{\sigma}^2} - [n - (p+1)] + (p+1) = \frac{\text{SSE}_p}{\widehat{\sigma}^2} + 2(p+1) - n.$$

Here $\widehat{\sigma}^2$ is some unbiased estimate of σ^2, which is usually taken to be the MSE from the full model based on the assumption that the full model has zero bias, i.e., $B_m = 0$.

EXERCISES

Theoretical Exercises

6.1 (Flow chart for backward stepwise algorithm) Modify the flow chart in Figure 6.3 for backward stepwise algorithm.

Applied Exercises

6.2 (Hamilton data) Hamilton (1987) gave the data shown in Table 6.3.

a) Make a matrix scatter plot of all three variables and calculate pairwise correlations between them. Comment.

b) Regress y on x_1, x_2. What is the R^2? How can you explain R^2 being close to 1 when both x_1 and x_2 have low correlations with y?

c) Why might the forward stepwise algorithm give a wrong result for this data set while the backward stepwise algorithm might give the right result?

6.3 (Best subset and stepwise regression) Table 6.4 lists the SSE's obtained by fitting different models involving three predictor variables, x_1, x_2, x_3, based on a total of $n = 20$ observations. (The constant term is included in all models.)

a) Complete the table by filling in the values of p (the number of predictor variables in the model), error d.f., MSE_p, $R^2_{\text{adj},p}$, C_p and AIC_p statistics.

Table 6.3 Hamilton data

x_1	x_2	y	x_1	x_2	y
2.23	9.66	12.37	3.04	7.71	12.86
2.57	8.94	12.66	3.26	5.11	10.84
3.87	4.40	12.00	3.39	5.05	11.20
3.10	6.64	11.93	2.35	8.51	11.56
3.39	4.91	11.06	2.76	6.59	10.83
2.83	8.52	13.03	3.90	4.90	12.63
3.02	8.04	13.13	3.16	6.96	12.46
2.14	9.05	11.44			

Source: Hamilton (1987).

Table 6.4 SSE's for all possible models with three predictors

Variables in Model	SSE_p	p	Error d.f.	MSE_p	$R^2_{\text{adj},p}$	C_p	AIC_p
None	950						
x_1	720						
x_2	630						
x_3	540						
x_1, x_2	595						
x_1, x_3	425						
x_2, x_3	510						
x_1, x_2, x_3	400						

b) Which models will be selected as the best using the $R^2_{\text{adj},p}$, C_p and AIC$_p$ criteria? Which model will you choose and why?
c) Suppose that stepwise regression is to be carried out with $f_{\text{in}} = f_{\text{out}} = 4.0$. Which variable would be the first to enter the model? What is its F_{in} value?
d) Which will be the second variable to enter the equation? What is its F_{in} value? What is its partial correlation coefficient with respect to y controlling for the first variable that entered the model?
e) Will the first variable that entered the model be removed upon the entry of the second variable? Check by doing the partial F-test.
f) Will stepwise regression enter the third variable in the model, i.e., will it choose the full model? Check by doing the partial F-test.

6.4 (Used car prices data: Comparing two models on the test set) In Examples 6.3 and 6.6 we obtained two different models for predicting the price using the training data. Evaluate these two models on the test data by computing the SSE's for both models. Which model will you select based on the SSE for the test set as well as whether all variables are included in the model are statistically significant at the 5% level.

CHAPTER 7

LOGISTIC REGRESSION AND CLASSIFICATION

In previous chapters we studied multiple regression where the response variable was numerical, ideally interval-scaled. In this chapter we study a regression methodology for categorical responses. The simplest categorical response is **binary** or **dichotomous**, e.g., a treatment outcome is a success or failure; a customer buys or does not buy. More generally, the response may have multiple categories (referred to as **multinomial** or **polytomous response**); furthermore, these categories may be **nominal** or **ordinal**. For example, a customer makes a choice among several nominal alternatives such as make of a car or mode of transportation. On the other hand, some outcomes are intrinsically ordinal such as stock recommendation (buy, hold or sell) or grade in a course (A, B, C, D, F). Polytomous responses typically have a small, finite range and do not have an underlying interval scale even if integers $1, 2, 3, \ldots$ may be used to code the ordinal categories. Multiple regression is inappropriate for such data. So we need a new methodology. Logistic regression is such a methodology.

In the case of logistic regression, prediction corresponds to **classification** of observations into one of several categorical outcomes. For example, a physician uses a battery of lab tests to determine whether a patient has a certain disease or not; a fraud detection algorithm classifies an online credit card transaction as fraudulent or legitimate depending on the past transaction history of the customer, origination of transaction etc.; a spam filter in an email system flags emails suspected as spam or not depending on the occurrence of some key words, proportion of capitalized words, etc. Such classifiers are known as **binary classifiers**.

Table 7.1 Variables for art museum visits data set

Variable	Description
Visit	At least one visit during the year: 1 (Yes), 0 (No)
Age	1 (18-19 yrs.), 2 (20-24 yrs.), . . . , 11 (65-69 yrs.), 12 (70+ yrs.)
Gender	0 (Male), 1 (Female)
Children	1 (Yes), 0 (No)
Married	1 (Yes), 0 (No)
Education	1 (8th grade or less), 2 (Some high school), . . . , 8 (Post graduate degree)
Income	1 (< $20K), 2 ($20K-$30K), . . . , 9 ($90K-$100K), 10 ($100K-$125K), 11 ($125K-$150K), 12 (> $150K)
County of Residence	1 (Chicago), 2 (Suburban Cook), 3 (Lake), 4 (DuPage), 5 (Other)

The following two examples will be analyzed in this chapter.

EXAMPLE 7.1 (Art Museum Visits: Data)

A survey of 2626 Chicago adults was conducted in 1995 by the Kellogg Center for Cultural Marketing at Northwestern University to determine which individual attributes are most predictive of a person visiting an art museum in the Chicago area. Data were collected on the variables listed in Table 7.1 and are stored in file `art.csv`.

We will use this data set to illustrate both simple logistic regression and multiple logistic regression. In the former case we will use Education as the only predictor. There are 19 missing observations on Education, so only 2607 complete observations will be used for simple logistic regression. Overall, there are 28 observations with missing data on at least one variable. So for multiple logistic regression 2598 complete observations will be used. Of these 2598 observations, 1676 are no visits (64.5%) and 922 are visits (35.5%). ■

EXAMPLE 7.2 (MBA Admissions: Data)

Johnson and Wichern (2002) have given data on 85 MBA admission decisions (1 = admit, 2 = wait list, 3= don't admit) by a business school as a function of two predictors: applicant's GPA score and GMAT score. The data are stored in file `MBA.csv`. The plot of the data is shown in Figure 7.1. It can be seen that the clusters corresponding to the three groups of admission decisions are fairly well-separated. The means of GPA and GMAT of the three groups are calculated as shown below. We notice that the GPA scores are well separated; the mean GMAT score of the "admit" group is much higher than that for the "wait list" and "don't admit" groups, which are almost identical.

```
> meanGPA <- sapply(1:3, function(x) mean(MBA$GPA[MBA$admit==x]))
> meanGMAT <- sapply(1:3, function(x) mean(MBA$GMAT[MBA$admit==x]))
```

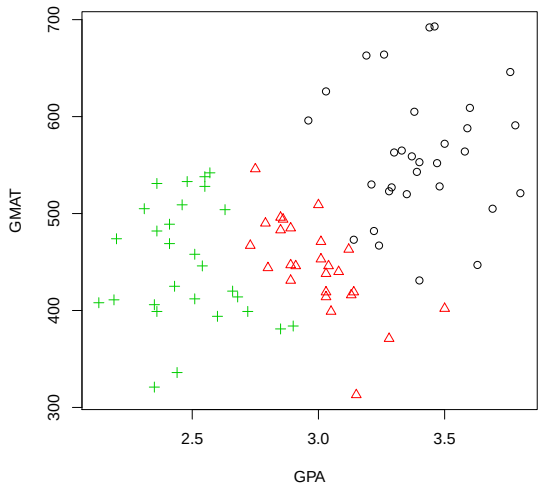


Figure 7.1 Plot of GMAT vs. GPA for MBA admissions data (circle=admit, triangle=wait list, cross=don't admit)

```
> meanGPA
[1] 3.403871 2.992692 2.482500
> meanGMAT
[1] 561.2258 446.2308 447.0714
```

We will use this data set to illustrate multinomial logistic regression, both nominal by disregarding the order among the three decisions and ordinal by taking the order into account. ■

7.1 Simple Logistic Regression

7.1.1 Model

We will begin with the simplest case of a dichotomous response variable y and a single predictor variable x. Following the standard convention, assume that y is coded as 1 or 0 depending on whether the outcome, generically labeled as a "success" or "failure," respectively. Denote the probability of success by $p(x)$ to indicate its dependence on the predictor x. Just as in multiple regression, we want to model $E(y|x) = P(y = 1|x) = p(x)$ as a function of x. Note that $p(x)$ is not to be confused with the notation p used to denote the number of predictors.

Why can't we use the simple linear regression model $p(x) = \beta_0 + \beta_1 x$? First of all, $p(x)$, being a probability, must lie between 0 and 1 but $\beta_0 + \beta_1 x$ is unconstrained, so predictions made from this model may fall outside the interval $[0, 1]$. Second, since y is a

Bernoulli r.v., its variance, $p(x)[1-p(x)]$, is not constant, being a function of x. So the homoscedasticity assumption is violated. Logistic regression addresses these issues.

Denote the **odds** of success by

$$\psi(x) = \frac{p(x)}{1-p(x)},$$

which is the ratio of the probability of success to the probability of failure. This ratio ranges between 0 and ∞, so we take its log thus mapping it to $[-\infty, \infty]$. Logistic regression defines a linear model on the **log-odds** or **logit**:

$$\ln \psi(x) = \ln \left[\frac{p(x)}{1-p(x)} \right] = \beta_0 + \beta_1 x \tag{7.1}$$

or equivalently

$$p(x) = \frac{\psi(x)}{1+\psi(x)} = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}. \tag{7.2}$$

This is known as the **logistic response function** or **logistic transform**.

The logistic function is an S-shaped curve as shown in Figure 7.2. It has a positive slope if $\beta_1 > 0$ and a negative slope if $\beta_1 < 0$. Just as in linear regression, β_1 has the interpretation of being the change in the log-odds of success for a unit change in x as can be seen from the following:

$$\ln \left[\frac{\psi(x+1)}{\psi(x)} \right] = \ln \psi(x+1) - \ln \psi(x) = (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) = \beta_1.$$

Thus

$$\frac{\psi(x+1)}{\psi(x)} = \exp(\beta_1).$$

This is called the **odds ratio**, being the ratio of odds of success for $x+1$ versus that for x. As an example, suppose that the probability of success for given x is $p(x) = 0.1$ or the odds of success are $\psi(x) = 0.1/0.9 = 0.111$. Furthermore, suppose that $\beta_1 = 0.3$. Then if x is increased by one unit, the odds of success will increase by $\exp(0.3) = 1.350$, i.e., the odds will increase to $(1.350)(0.111) = 0.150$. So the probability of success will increase to $0.150/1.150 = 0.130$.

A special case of interest is when the predictor variable is binary, for example, in a clinical trial suppose the control group is coded as $x = 0$ and the treatment group is coded as $x = 1$. Then $\exp(\beta_1)$ gives the odds ratio of success for the treatment compared to the control. If $\beta_1 = 0$ then the odds ratio is 1, which means that the odds of success are the same for the control and the treatment, so the treatment has no effect. On the other hand, if $\beta_1 > 0$ then the treatment is more effective than the control and if $\beta_1 < 0$ then the treatment is less effective than the control.

Logistic transform is just one of many possible functions that can be used to model $p(x)$. Essentially, any cumulative distribution function (c.d.f.) can be used for $p(x)$ since it is a nondecreasing function taking values between 0 and 1 as its argument goes from $-\infty$ to $+\infty$ (logistic function is the c.d.f. of the logistic distribution; see the Technical Notes section). Two examples are:

- **Probit function**: $p(x) = \Phi(\beta_0 + \beta_1 x)$ where $\Phi(x)$ is the standard normal c.d.f. This function is similar in shape to the logistic function shown in Figure 7.7.
- **Complementary log-log function**: $p(x) = 1 - \exp(-\exp(\beta_0 + \beta_1 x))$ or $\ln[-\ln(1-p(x))] = \beta_0 + \beta_1 x$, which is the c.d.f. of the extreme value distribution. This function is used to model very low probabilities associated with rare events.

These functions are analytically not as tractable as the logistic function and hence are not as commonly used.

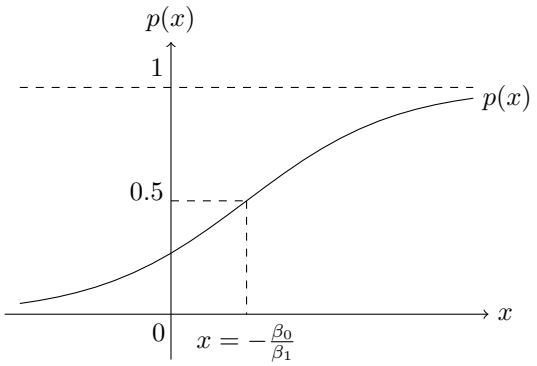


Figure 7.2 Logistic Response Function

7.1.2 Parameter Estimation

Suppose that we have n independent observations $y_1, \ldots, y_n$ corresponding to the predictor values $x_1, \ldots, x_n$, where y_i is the response outcome (0 or 1) for the ith observation. We want to estimate the parameters β_0 and β_1 of the model (7.1) from these data. The least squares (LS) method cannot be used in this case since the observed value of the logistic transform $\ln[y_i/(1-y_i)]$ is either $-\infty$ if $y_i = 0$ or $+\infty$ if $y_i = 1$ (unless the data are grouped in which case each y_i is a proportion; see Example 7.3). Therefore we will use the **maximum likelihood estimation (MLE)** method, which is a more general method with certain optimality properties. See Appendix B for a primer on the MLE method.

In the present case, the random data are the y_i which have Bernoulli distributions with success probabilities

$$p_i = p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]}.$$

This Bernoulli distribution can be written as

$$f(y_i|\beta_0, \beta_1) = (p_i)^{y_i}(1-p_i)^{1-y_i} = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0. \end{cases}$$

Hence the likelihood function equals

$$L = L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left[(p_i)^{y_i}(1-p_i)^{1-y_i}\right] = \prod_{i=1}^{n} \left(\frac{p_i}{1-p_i}\right)^{y_i} \times \prod_{i=1}^{n}(1-p_i)$$

and the log-likelihood function equals

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \ln(1-p_i) \\ &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln\left[1 + \exp(\beta_0 + \beta_1 x_i)\right]. \end{aligned} \tag{7.3}$$

The MLE's $\widehat{\beta}_0$ and $\widehat{\beta}_1$ maximize $\ln L$ w.r.t. β_0 and β_1. These maximizing values can be found by setting the partial derivatives of $\ln L$ w.r.t. β_0 and β_1 equal to zero and solving

the resulting equations. Now,

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]} \\ &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p_i\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_1} &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]} \\ &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i p_i.\end{aligned}$$

Since $p_i = E(y_i)$, the equations obtained by setting the above partial derivatives equal to zero can be written as

$$E\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i \quad \text{and} \quad E\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i y_i. \tag{7.4}$$

Thus they equate certain expected quantities to the corresponding observed quantities. As we shall see in Chapter 8, this is a common feature of all generalized linear models. We will also see there that an iteratively reweighted least squares algorithm can be used to solve these nonlinear simultaneous equations.

EXAMPLE 7.3 (Art museum visits: Simple logistic regression estimation)

The grouped data for eight increasing levels of Education, coded 1-8, are shown in Table 7.2.

Table 7.2 Art museum visit data

	Visit		
Education	Yes	No	Total
1	7	24	31
2	24	92	116
3	92	408	500
4	53	196	249
5	271	439	710
6	172	277	449
7	107	96	203
8	199	150	349
Total	925	1682	2607

Since the data are grouped, we can calculate the sample proportion f_i of "yes" responses at each Education level $x_i = i$ and the corresponding sample logistic transforms as shown in Table 7.3. Figure 7.3 shows the plot of the logistic transform versus Education level. The plot appears roughly linear with a characteristic "hockey stick" shape resulting from the fact that the Education level must exceed some threshold

(≥ 5: some college or university studies) before its effect on the probability of visiting an art museum becomes apparent. The equation of the LS fitted straight line is $\widehat{y} = -1.9067 + 0.2587x$, where $\widehat{y}$ is the predicted log-odds of a "yes" response and x is the Education level.

This preliminary analysis justifies fitting a simple logistic regression model. The following R output shows that the fitted logistic regression model is $\widehat{y} = -2.3083 + 0.3281x$, which is qualitatively similar to the previous LS fitted equation. The odds of visiting an arts museum go up by a factor $\exp(0.3281) = 1.388$ if the Education level goes up by one level.

```
> fit = glm(Visit ~ Education, family=binomial, data = art)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3083     0.1407  -16.41   <2e-16 ***
Education     0.3281     0.0251   13.07   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3391.1  on 2606  degrees of freedom
Residual deviance: 3205.4  on 2605  degrees of freedom
  (19 observations deleted due to missingness)
AIC: 3209.4
```

The deviances are defined in Section 7.3.1.

The estimated probabilities:

$$\widehat{p}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)} \quad (i = 1 \ldots, n) \tag{7.5}$$

for this model are given in Table 7.3. The following R code and the resulting output show the calculation of these probabilities as well as the classification of observations (using the classification rule $\widehat{p}_i > 0.3$ implying "Yes").

```
> vector1=data.frame(Education=c(1,2,3,4,5,6,7,8))
> probs1=predict(fit1,newdata=vector1,type="response")
> probs1
        1         2         3         4         5         6
0.1213012 0.1608361 0.2101743 0.2697819 0.3390380 0.4159458
        7         8
0.4971766 0.5785568
> pred1=rep("No", 8)
> pred1[probs1>0.3]="Yes"
> pred1
[1] "No"  "No"  "No"  "No"  "Yes" "Yes" "Yes" "Yes"
```

■

Table 7.3 Sample proportions and logistic transforms for the art museum visit data

Education	Proportion (f_i)	$\ln[f_i/(1-f_i)]$	$\ln[\widehat{p}_i/(1-\widehat{p}_i)]$	$\widehat{p}_i$
1	0.2258	-1.2321	-1.9802	0.1213
2	0.2069	-1.3437	-1.6521	0.1608
3	0.1840	-1.4895	-1.3240	0.2102
4	0.2129	-1.3078	-0.9959	0.2697
5	0.3817	-0.4824	-0.6678	0.3390
6	0.3831	-0.4765	-0.3397	0.4159
7	0.5271	0.1085	-0.0116	0.4971
8	0.5702	0.2827	0.3165	0.5985

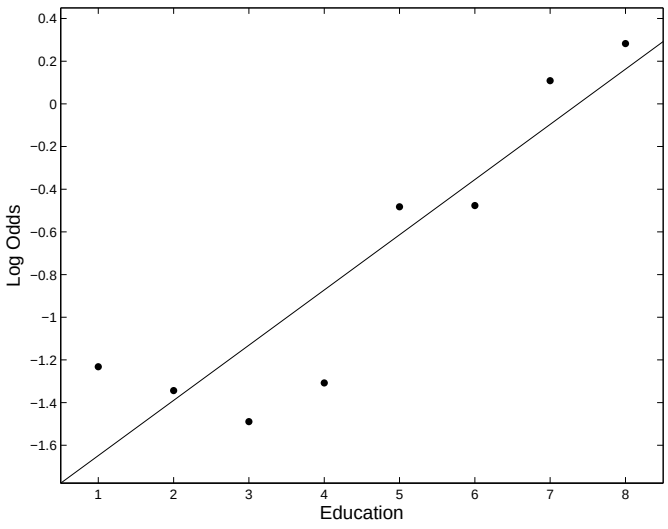


Figure 7.3 Plot of $\ln\left(\frac{f_i}{1-f_i}\right)$ versus Education

7.1.3 Inferences on Parameters

To test hypotheses and make confidence intervals on β_0 and β_1 we use the following result: For large n, $(\widehat{\beta}_0, \widehat{\beta}_1)$ is approximately bivariate normal with mean vector (β_0, β_1) and **asymptotic covariance matrix** $\boldsymbol{V}$, which is the inverse of the **information matrix**:

$$\begin{aligned} \boldsymbol{\mathcal{I}} &= E\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta_0^2} & \frac{\partial^2 \ln L}{\partial \beta_0 \beta_1} \\ \frac{\partial^2 \ln L}{\partial \beta_0 \beta_1} & \frac{\partial^2 \ln L}{\partial \beta_1^2} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n p_i(1-p_i) & \sum_{i=1}^n x_i p_i(1-p_i) \\ \sum_{i=1}^n x_i p_i(1-p_i) & \sum_{i=1}^n x_i^2 p_i(1-p_i) \end{bmatrix}. \end{aligned} \tag{7.6}$$

The derivation of the entries of $\boldsymbol{\mathcal{I}}$ is given in Example B.3. The diagonal entries v_{00} and v_{11} of $\boldsymbol{V} = \boldsymbol{\mathcal{I}}^{-1}$ are the asymptotic variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. In practice, we need to replace the p_i in the information matrix by their sample estimates given by (7.5) and then invert the matrix to obtain the estimated asymptotic covariance matrix $\widehat{\boldsymbol{V}}$ with diagonal entries $\widehat{v}_{00}$ and $\widehat{v}_{11}$. To test $\beta_1 = 0$ we can use the **Wald statistic** statistic $z = \widehat{\beta}_1/\sqrt{\widehat{v}_{11}}$ as a standard normal test statistic. A large sample $100(1-\alpha)\%$ CI for β_1 is given by

$$\widehat{\beta}_1 - z_{\alpha/2}\sqrt{\widehat{v}_{11}} \le \beta_1 \le \widehat{\beta}_1 + z_{\alpha/2}\sqrt{\widehat{v}_{11}}.$$

An analogous formula can be used for the CI on β_0.

EXAMPLE 7.4 (Art museum visits: Simple logistic regression inference)

In this example we will test a hypothesis and compute a CI on β_1 and on the associated odds ratio. As seen from the R output from Example 7.3, the asymptotic standard error of $\widehat{\beta}_1$ is $\text{SE}(\widehat{\beta}_1) = 0.0251$. We will verify this by hand calculation. Using estimated probabilities $\widehat{p}_i$ given in Table 7.3, we calculate the following entries of the estimated information matrix (7.6):

$$\sum_{i=1}^8 n_i \widehat{p}_i(1-\widehat{p}_i) = 555, \sum_{i=1}^8 n_i x_i \widehat{p}_i(1-\widehat{p}_i) = 2966, \sum_{i=1}^8 n_i x_i^2 \widehat{p}_i(1-\widehat{p}_i) = 17435,$$

where $x_i = 1, \ldots, 8$. The estimated asymptotic covariance matrix then equals

$$\begin{bmatrix} 555 & 2966 \\ 2966 & 17435 \end{bmatrix}^{-1} = \begin{bmatrix} 0.01983 & -0.00337 \\ -0.00337 & 0.00063 \end{bmatrix}.$$

Thus $\text{SE}(\widehat{\beta}_1) = \sqrt{0.00063} = 0.0251$. So a large sample 95% CI on β_1 is

$$0.3281 \pm 1.96 \times 0.0251 = [0.2789, 0.3773].$$

Since this interval excludes 0, the null hypothesis $H_0 : \beta_1 = 0$ can be rejected. The corresponding z-statistic equals $z = 0.3281/0.0251 = 13.07$, which is highly significant. A 95% CI on the odds ratio equals $[\exp(0.2789), \exp(0.3773)] = [1.3217, 1.4583]$, which shows that the odds ratio is significantly greater than 1. ■

7.2 Multiple Logistic Regression

7.2.1 Model and Inference

Extension of simple logistic regression to multiple logistic regression is straightforward. Assume that we have p predictors, $x_1, \ldots, x_p$. Let $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ denote the predic-

tor vector and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ denote the regression coefficient vector. Further let $p(\boldsymbol{x})$ denote the probability of success and $\psi(\boldsymbol{x}) = p(\boldsymbol{x})/[1 - p(\boldsymbol{x})]$ denote the odds of success. A straightforward extension of the simple logistic regression model (7.1) gives the multiple logistic regression model:

$$\ln \psi(\boldsymbol{x}) = \ln \left[\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} \right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}'\boldsymbol{\beta} \tag{7.7}$$

or equivalently

$$p(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})}. \tag{7.8}$$

The regression coefficients β_j have the same interpretation as before, namely $\exp(\beta_j)$ is the odds ratio of success when x_j is increased by one unit, keeping all other x's fixed.

The MLE's of the β_j's can be computed by setting the partial derivatives of the log-likelihood function equal to zero and solving the resulting simultaneous equations. These equations have a relatively simple form, which is a generalization of Equation (7.4). Let $\boldsymbol{X}$ be the $n \times (p + 1)$ model matrix as defined in Chapter 3, Section 3.1.2 (including the first column of all 1's) and let $\boldsymbol{p} = (p_1, \ldots, p_n)'$ be the vector of the success probabilities $p_i = E(y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/[1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})]$. Then the MLE equations are given by

$$\boldsymbol{X}'\boldsymbol{p} = \boldsymbol{X}'\boldsymbol{y} \quad \text{or equivalently} \quad \sum_{i=1}^{n} \boldsymbol{p}_i \boldsymbol{x}_i = \sum_{i=1}^{n} y_i \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i'$ is the ith row of $\boldsymbol{X}$. This is a system of $p + 1$ simultaneous nonlinear equations in $p + 1$ unknowns $\beta_0, \ldots, \beta_p$ (since the p_i are nonlinear functions of $\beta_0, \ldots, \beta_p$).

Similarly the Hessian matrix of mixed second partial derivatives of the log-likelihood function evaluated at the MLE's of the β_j's gives the observed information matrix which upon inversion yields the estimated asymptotic covariance matrix of the $\widehat{\beta}_j$'s. The following example illustrates these calculations.

EXAMPLE 7.5 (Art museum visits: Multiple logistic regression)

Before fitting a multiple logistic regression model, it is useful to perform some exploratory analyses to assess the relationships between the proportions of people visiting art museums and the various predictors. Frequency tables for the three binary predictors, Gender, Children and Married, are shown in Table 7.4. From these tables we see that females, people with no children and unmarried people are more likely to visit an art museum than their corresponding counterparts.

The frequency table for the County variable is shown in Table 7.5. We see that there is a significant county effect with Chicago Cook county having the largest proportion of art museum visitors compared to collar counties.

Finally, Figure 7.4 shows the plots of logistic transforms of proportions visiting art museums versus Income and Age. (Plot against Education is shown in Figure 7.3.) We see that the relationship with Income is roughly linear but the relationship with Age, if any, is more complex. When included in the model, Age turns out to be a nonsignificant predictor, so we omit it in the following.

The R commands for fitting the multiple logistic regression model are as follows.

```
> fit = glm(Visit ~ Education+Income+Gender+Children+Married,
> factor(County), family=binomial, data = art)
> summary(fit)
```

The R output is shown below. We see that all the predictors are highly significant. Their regression coefficients represent the change in the log-odds (or the log odds

Table 7.4 Frequency tables for Gender, Children and Married variables for art museum visit data

		Visit?			Yes
Predictor	Category	Yes	No	Total	Proportion
Gender	Female	472	749	1221	0.387
	Male	455	938	1393	0.327
Children	No	668	1075	1743	0.383
	Yes	262	621	883	0.297
Married	No	434	617	1051	0.413
	Yes	496	1079	1575	0.315

Table 7.5 Frequency table for County variable for art museum visit data

	Visit?			Yes
County	Yes	No	Total	Proportion
Chicago Cook	106	84	190	0.558
Suburban Cook	434	747	1181	0.368
Lake	121	246	367	0.330
DuPage	90	111	201	0.448
Other	179	508	687	0.261

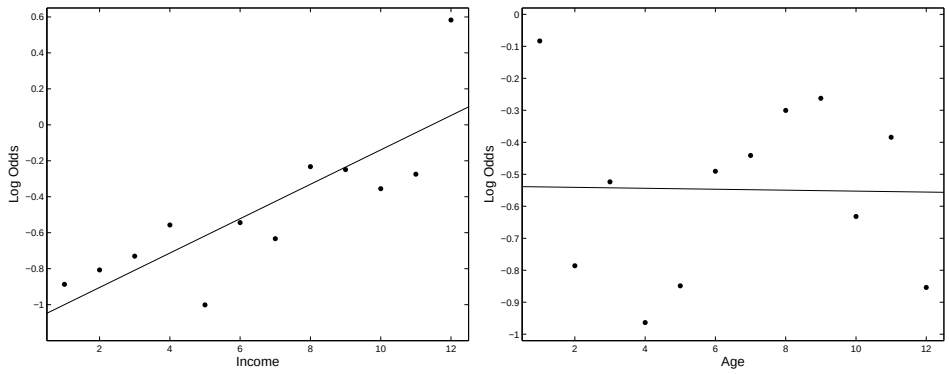


Figure 7.4 Plots of logistic transforms versus Income and Age

ratio) of visiting an art museum if the value of the predictor is increased by one unit for a numerical variable and with respect to the reference category for a categorical variable. The effects of Education and Income are positive, as suggested by the plots, indicating that the odds of visiting an art museum increase as these variables increase in magnitude. The effects of categorical predictors are in agreement with what we saw in the frequency tables given above.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.26621    0.21079  -6.007 1.89e-09 ***
Education         0.30223    0.02768  10.918  < 2e-16 ***
Income            0.07168    0.01627   4.406 1.05e-05 ***
Gender           -0.42181    0.08958  -4.709 2.49e-06 ***
Children         -0.34137    0.09867  -3.460 0.000541 ***
Married          -0.49891    0.09990  -4.994 5.91e-07 ***
factor(County)2 -0.59726    0.17009  -3.511 0.000446 ***
factor(County)3 -0.74408    0.19613  -3.794 0.000148 ***
factor(County)4 -0.43072    0.21822  -1.974 0.048409 *
factor(County)5 -0.96831    0.18330  -5.283 1.27e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3379.6  on 2597  degrees of freedom
Residual deviance: 3068.2  on 2588  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 3088.2
```

■

7.3 Likelihood Ratio (LR) Test

The LR test extends the extra sum of squares test discussed in Section 3.3.4 for multiple regression to more general parametric models. Consider a multiple logistic regression model with p predictors, which we refer to as the **full model (FM)**, and suppose we want to test if a **partial model (PM)** with a subset of $q < p$ predictors provides an equally good fit. Labeling the q predictors in the partial model as $x_1, \ldots, x_q$, we want to test the null hypothesis $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$, i.e., whether the extra predictors, $x_{q+1}, \ldots, x_p$, can be dropped from the model.

Denote the maximums of the likelihood functions for the two models by $L_{\max}(\text{PM})$ and $L_{\max}(\text{FM})$, where the maximums are obtained by substituting the MLE's of the β_j's under the respective models in the likelihood function. Then the **LR test statistic** for testing H_0 is given by

$$G^2 = -2 \ln \left[\frac{L_{\max}(\text{PM})}{L_{\max}(\text{FM})} \right] = -2 \left[\ln L_{\max}(\text{PM}) - \ln L_{\max}(\text{FM}) \right]. \tag{7.9}$$

Under H_0, the LR statistic can be shown to be asymptotically (as $n \to \infty$) chi-square distributed with $p - q$ d.f. So we can reject H_0 at level α if $G^2 > \chi^2_{p-q,\alpha}$ and conclude that the full model provides a significantly better fit than the partial model.

Usually the LR test for comparing PM with FM is conducted by comparing their deviances, defined in the next section.

7.3.1 Deviance

The **deviance** (referred to as **residual deviance** in the R output) of a given model plays the same role in logistic regression (or more generally in generalized linear models; see Chapter 9) as does the SSE in multiple regression. Similar to SSE, the larger the deviance, the poorer the fit of the given model. It is defined as the LR test statistic for comparing the given model (M) with the so-called **saturated model (SM)**, which has as many parameters as the number of distinct observations and thus provides an "exact" fit to the data. In the LR test parlance, the saturated model with n parameters is the full model and the given model with $p+1 < n$ parameters (p predictor variables) is the partial model. The deviance of the model M equals

$$D^2 = -2[\ln L_{\max}(\text{M}) - \ln L_{\max}(\text{SM})]. \tag{7.10}$$

For the binary logistic regression model,

$$\ln L_{\max}(\text{M}) = \sum_{i=1}^{n} \left[y_i \ln \widehat{p}_i + (1 - y_i) \ln(1 - \widehat{p}_i) \right],$$

where

$$\widehat{p}_i = \frac{\exp(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})}$$

and $\widehat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. Under the saturated model SM we set $\widehat{p}_i = y_i = 0$ or 1, thus yielding

$$\ln L_{\max}(\text{SM}) = \sum_{i=1}^{n} \left[y_i \ln y_i + (1 - y_i) \ln(1 - y_i) \right].$$

Using l'Hospital's rule, it is easy to show that $\ln L_{\max}(\text{SM}) = 0$. This is also explained by the fact that since SM fits the data exactly, $L_{\max}(\text{SM}) = 1$ and hence $\ln L_{\max}(\text{SM}) = 0$. However, this is not true in general as we shall see for some generalized linear models in Chapter 8. Even for binary logistic regression models, if the data are grouped then $L_{\max}(\text{SM}) < 1$ and so $\ln L_{\max}(\text{SM}) < 0$ as Exercise 7.3 asks you to show.

Under the null hypothesis that model M provides as good a fit as does the saturated model SM, asymptotically D^2 is chi-square distributed with $n - (p + 1)$ d.f. Thus the null hypothesis can be rejected at level α if $D^2 > \chi^2_{n-(p+1),\alpha}$. This is referred to as the **goodness of fit test** of model M. However, it is not a very useful test in practice since it usually gives a significant result as the comparison is with the saturated model which provides an "exact" fit. For example, from the R output in Example 7.5 we see that the deviance $D^2 = 3068.2$ on $n - (p + 1) = 2598 - (9 + 1) = 2588$ d.f., which is highly significant implying that the multiple logistic regression model in that example is not a good fit.

Another goodness of fit test that is sometimes used is the **Pearson chi-square test**. It requires the data to be grouped into homogeneous groups so that the estimated response probability is approximately the same for all observations in each group. This requirement is of course satisfied if there is only one categorical predictor variable as in Example 7.3 where Education with eight categories is the only categorical predictor variable.

Let $g \geq 2$ be the number of groups, n_i and s_i be the sample size and the number of successes and $\widehat{p}_i$ be the estimated success probability for the ith group ($i = 1, \ldots, g$) using

a given model. Then the chi-square statistic is given by

$$X^2 = \sum_{i=1}^{g} \frac{(s_i - n_i\widehat{p}_i)^2}{n_i\widehat{p}_i(1-\widehat{p}_i)}. \tag{7.11}$$

Since g is the effective number of distinct observations for grouped data, this statistic has $g-(p+1)$ d.f.

The **null deviance** (denoted by D_0^2) is the deviance of the **null model (NM)** that has no predictor variables — only the intercept term. Null deviance is analogous to SST in multiple regression. The null model assumes a common success probability for all cases given by

$$p_0 = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}.$$

It is easy to show that the MLE of p_0 is $\widehat{p}_0 = s/n$, the proportion of successes. Therefore

$$L_{\max}(\text{NM}) = (\widehat{p}_0)^s\,(1-\widehat{p}_0)^{n-s} = \left(\frac{s}{n}\right)^s \left(\frac{n-s}{n}\right)^{n-s}.$$

Since $\ln L_{\max}(\text{SM}) = 0$, the **null deviance** equals

$$D_0^2 = -2[s\ln s + (n-s)\ln(n-s) - n\ln n].$$

We can compare the deviance D^2 of the given model with the null deviance D_0^2 to test $H_0 : \beta_1 = \cdots = \beta_p = 0$. This is referred to as the **overall significance test** of the model and is similar to the ANOVA F-test for multiple regression. From the LR test it follows that this H_0 can be rejected at level α if $D_0^2 - D^2 > \chi^2_{p,\alpha}$. Applying this test to the multiple logistic regression model in Example 7.5, we get $D_0^2 - D^2 = 3379.6 - 3068.2 = 311.4$ on 9 d.f. (corresponding to the 9 coefficients under test) which is highly significant. So H_0 can be rejected and we conclude that at least one of the $\beta_j \neq 0$.

EXAMPLE 7.6 (Art museum visits: Comparison of simple versus multiple logistic regression models)

Let us compare the simple logistic regression model that uses only Education as the predictor with the multiple logistic regression model that uses six predictors (including Education). A subset of the data consisting of 2598 observations is used to fit the multiple logistic regression model in Example 7.5 as compared to the larger data set consisting of 2607 observations used to fit the simple logistic regression model in Example 7.3 because more observations are missing on additional predictors used in the former. In order for their deviances and d.f. to be comparable, we refitted the simple logistic regression model on the subset data set resulting in the following output.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.30444    0.14088  -16.36   <2e-16 ***
Education    0.32765    0.02515   13.03   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3379.6  on 2597  degrees of freedom
Residual deviance: 3195.3  on 2596  degrees of freedom
AIC: 3199.3
```

For this model $D^2 = 3195.3$ with 2596 d.f. whereas for the multiple logistic regression model $D^2 = 3068.2$ with 2588 d.f. as seen before. So the test statistic is $3195.3 - 3068.2 = 127.1$ with $2596 - 2588 = 8$ d.f. It is readily seen to be highly significant. So we reject the null hypothesis and conclude that the multiple logistic regression model provides a significantly better fit than the simple logistic regression model. ■

7.3.2 Akaike information criterion (AIC)

The Akaike information criterion (AIC) was introduced in Section 6.1.1 for multiple regression. More generally it is given by

$$\text{AIC} = -2 \ln L_{\max}(\text{M}) + 2(p+1), \tag{7.12}$$

where one can regard $2(p+1)$ as a penalty for the number of predictors in the model. Since in the case of binary logistic regression, $\ln L_{\max}(\text{SM}) = 0$, we get

$$\text{AIC} = D^2 + 2(p+1).$$

Thus AIC is a combination of a measure of the goodness of fit (D^2) and complexity of the model ($2(p+1)$). As a check of this formula, note from the R output in Example 7.5 that $D^2 = 3068.2$ and AIC $= 3068.2 + 2(9+1) = 3088.2$.

7.4 Logistic Regression Model Selection and Model Diagnostics

Best subsets and stepwise regression methods discussed in Chapter 6 can be applied to select the best logistic regression model. Minimizing AIC is a well-defined model selection criterion. One cannot simply aim at minimizing the deviance as it would lead to the full model just as in multiple regression minimizing SSE (or equivalently maximizing R^2) leads to the full model. Therefore a predictor variable should be added to a model only if it reduces the deviance significantly at some preassigned α level using a chi-square test. The following example illustrates use of `stepAIC` and `bestglm` functions in R for this purpose.

EXAMPLE 7.7 (Art museum visits: Stepwise logistic regression)

In Example 7.5 we fitted a multiple logistic regression model to the art museum visits data. In this example we apply stepwise regression to the same data. Forward stepwise and backward stepwise both result in the same full model. For the forward stepwise method the R commands are given below.

```
> fit = glm(Visit ~ 1, binomial, art2)
> stepAIC(fit, scope=~ Income+ Gender + Children + Married
  + Education + County)
```

Table 7.6 gives the best model (in terms of the smallest AIC) after adding each new variable in a stepwise manner. We see that the decrease in deviance at each step exceeds $\chi^2_{1,.05} = 3.843$ (for the addition of the County variable it exceeds $\chi^2_{4,.05} = 9.488$), so the addition of each variable is justified based on the deviance criterion as well. ■

Table 7.6 Stepwise Regression Results for Art Museum Visits Data

p	Added Variable	d.f.	Deviance	AIC
0	None	0	3363.96	3365.96
1	Education	1	3184.09	3188.09
2	County	4	3132.86	3144.86
3	Married	1	3102.81	3116.81
4	Gender	1	3086.18	3102.18
5	Income	1	3068.69	3086.69
6	Children	1	3056.22	3076.22

Next we consider some diagnostics for logistic regression. Two types of residuals are used for this purpose. **Deviance residuals** are defined as

$$d_i = \text{sign}(y_i - \widehat{p}_i)\sqrt{-2\left[y_i \ln \widehat{p}_i + (1-y_i)\ln(1-\widehat{p}_i)\right]}.$$

Note that the deviance $D^2 = \sum_{i=1}^{n} d_i^2$. These residuals can be used to check for outliers as in the multiple regression case, e.g., if $|d_i| > 3$ then declare the observation to be an outlier. Note that we don't test normality or homoscedasticity in the case of logistic regression since the response variables are Bernoulli distributed which are heteroscedastic.

Influential observations are generally detected by calculating the changes in D^2 by deleting individual observations and identifying those observations as influential which cause the biggest changes defined below:

$$\Delta D_i^2 = D^2 - D_{(i)}^2,\ i = 1, 2, \ldots, n,$$

where $D_{(i)}^2$ is the D^2 statistic calculated by deleting the ith observation. There are no formal statistical tests for how large ΔD_i^2 should be to be regarded as large. One can make a sequence plot of these changes against the observation index i and identify any spikes in the plot as influential observations. One can use the leverage measures used to detect influential observations and variance inflation factors (VIF's) used to detect multicollinearity in the context of logistic regression as they depend solely on the $\boldsymbol{X}$ matrix.

7.5 Binary Classification Using Logistic Regression

7.5.1 Measures of Correct Classification

Let $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ denote the vector of predictors for a future observation and let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ denote the estimated parameter vector of the logistic regression model. Then we can calculate the estimated probability of "success" as

$$\widehat{p}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})}.$$

We compare $\widehat{p}(\boldsymbol{x})$ with a specified **cutoff probability** p^* such that if $\widehat{p}(\boldsymbol{x}) > p^*$ then the observation is classified as a "success," otherwise as a "failure." For example, we may use $p^* = 1/2$ thinking that a better than 50:50 chance is sufficient to predict a "success." However, this may not be the right thing to do for various reasons, one reason being the

costs of misclassification of a true "success" and a true "failure" may be unequal. For instance, the costs of misdiagnosing a patient may be very different depending on the severity of the disease and the cost and side-effects of the treatment. Another reason is that the true proportion of "successes" in the population may be much lower or higher than 50% and this should be factored into the decision rule. For example, for a rare disease with less than 0.1% prevalence rate, p^* should be correspondingly lower. Therefore it is worthwhile to study different choices of p^*.

The results of classifying n observations can be represented in the form of a 2×2 table, called the **confusion matrix**, shown below.

		Classified		
		Negative	Positive	Row Total
True	Negative	n_{00}	n_{01}	n_0
	Positive	n_{10}	n_{11}	n_1

The proportion of true positives that are classified as negatives, namely n_{01}/n_0, is called the **false positive rate** and the proportion of true negatives that are classified as positives, namely n_{10}/n_1, is called the **false negative rate**. In hypothesis testing terminology, they correspond to the **type I error rate** and the **type II error rate**, respectively. Alternative but equivalent measures can be defined in terms of correct classifications as follows:

$$\text{Specificity } = 1 - \text{False positive rate} = \frac{n_{00}}{n_0}$$

and

$$\text{Sensitivity } = 1 - \text{False negative rate} = \frac{n_{11}}{n_1}.$$

These two measures are complementary to each other in that if one increases the other decreases. As p^* is increased, the false positive rate decreases and hence specificity increases. On the other hand, the false negative rate increases and hence sensitivity decreases. We can choose p^* to minimize the total number of misclassifications, $n_{01} + n_{10}$. If the costs of misclassifications are unequal, say c_0 is the cost of a false positive and c_1 is the cost of a false negative then we can minimize the total cost of misclassification, $C = c_0 n_{01} + c_1 n_{10}$. Note that only the cost ratio c_0/c_1 matters, not the actual costs. If the cost of a false positive is twice that of the cost of a false negative then it suffices to minimize $C = 2n_{01} + n_{10}$.

In machine learning and information retrieval the following variants of the above measures are used:

$$\text{Precision} = P = \frac{n_{11}}{n_{01} + n_{11}}$$

and

$$\text{Recall} = R = \frac{n_{11}}{n_{10} + n_{11}} = \frac{n_{11}}{n_1}.$$

In the context of information retrieval the true positives and negatives correspond to the relevant and irrelevant items (e.g., documents), respectively. The predicted positives and negatives correspond to the retrieved and non-retrieved items, respectively. Thus precision is the proportion of the retrieved items that are relevant and recall is the proportion of the relevant items that are retrieved. The same kind of complementarity holds between precision and recall as between specificity and sensitivity. Therefore both cannot be increased simultaneously. A common practice is to maximize the harmonic mean of the two, called the F_1**-score**:

$$F_1 = \left[\frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)\right]^{-1} = \frac{2PR}{P+R}.$$

EXAMPLE 7.8 (Art museum visits: Determination of optimum cutoff probability)

Suppose that we want to determine the optimum cutoff probability p^* and the associated predictive power of the multiple logistic regression model fitted in Example 7.5. To get unbiased estimates of these quantities, we split the data into a training set and a test set. We did this by taking the training set to be all odd-numbered observations and the test set to be all even-numbered observations. Thus, of the total 2598 complete observations, exactly half, i.e., 1299 observations are in each of the two sets.

We used the following R script to form the training and test sets and compute the confusion matrix using a cutoff probability $p^* = 0.5$ for illustration.

```
> art = read.csv("c:/data/art.csv")
> evenrow = seq(2,nrow(art),2)
> oddrow = seq(1,nrow(art),2)
> train = art[oddrow,]
> test = art[evenrow,]
> #Fit a multiple logistic regression model
> fit = glm(Visit ~ Education+Income+Gender+Children+Married,
> factor(County),family=binomial, data = train)
> summary(fit)
> testpredict=predict(fit,newdata=test)
> tab=table(test$Visit, testpredict>.50)
> tab
```

From the confusion matrix, we can calculate the **correct classification rate (CCR)**, given by $(n_{00} + n_{11})/n$ using the following commands

```
> CCR=sum(diag(tab))/sum(tab)
> CCR
```

The CCR's for $p^* = 0$ and $p^* = 1$ serve as useful limiting values. If $p^* = 0$ then all true positives are correctly classified but all true negatives are misclassified, so CCR $= 459/1299 = 0.3534$. Similarly, if $p^* = 1$ then all true negatives are correctly classified but all true positives are misclassified, so CCR $= 840/1299 = 0.6467$.

By a numerical search we find the optimum $p^* = 0.44$ corresponding to the confusion matrix shown below which gives maximum CCR $= (797 + 75)/1299 = 0.6713$.

```
  FALSE TRUE
0   797   43
1   384   75
```

```
[1] 0.6712856
```

In this example CCR is a fairly flat function of p^* in the search interval $[0.40, 0.60]$, so almost any value in this interval is equally good.

The values of the other measures defined above for $p^* = 0.44$ are as follows:

$$\text{Sensitivity} = \frac{75}{384+75} = 0.1634, \text{Specificity} = \frac{797}{797+43} = 0.9488,$$

$$\text{Precision} = \frac{75}{75+43} = 0.6356, \text{Recall} = \frac{75}{384+75} = 0.1634.$$

The F_1-score equals

$$F_1 = \frac{2(0.6356)(0.1634)}{0.6356 + 0.1634} = 0.2600.$$

We see that this model is good for predicting non-visitors (its specificity is high) but not so good for predicting visitors (its sensitivity is low). Recalling that in the entire sample, there were 35.5% visitors (see Example 7.1), sensitivity is less than half the true proportion of visitors. If we think of visitors as relevant items and non-visitors as irrelevant items then 63.56% of the retrieved items are relevant but only 16.34% of the relevant items are retrieved by this binary classifier. The F_1-score is correspondingly low. ■

7.5.2 Receiver Operating Characteristic (ROC) Curve

The ROC curve is used to assess the performance of a binary classifier for different choices of the cutoff probability p^*. It is obtained by plotting sensitivity versus $1-$ specificity by varying p^*.

Denote by $\widehat{p}_i = \widehat{p}(\boldsymbol{x}_i)$ the estimated probability of a "positive" (or a "success") for the ith observation with the predictor vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$. For simplicity, assume that all $\widehat{p}_i$ are distinct. Suppose that we start with $p^* = 1$. Then all $\widehat{p}_i$'s are $< p^*$, so all observations are classified as "negatives" (or "failures"). Therefore sensitivity $= 0$ and specificity $= 1$ or $1-$ specificity $= 0$. This results in the lower left hand corner point or the origin. The reverse happens when $p^* = 0$ and we get the upper right hand corner point. As p^* is decreased from 1 to 0, every time it crosses one of the calculated $\widehat{p}_i$ values we get one more observation classified as a "positive." If it is a true "positive" then sensitivity goes up by $1/n_1$; if it is a true "negative" then sensitivity remains unchanged but specificity goes down by $1/n_0$ or $1-$ specificity goes up by $1/n_0$. Thus the ROC curve is a nondecreasing step function as shown in Figure 7.5 along with a smooth approximation to it.

The ideal binary classifier perfectly discriminates between the true "positives" and the true "negatives." This will happen if the estimated probabilities $\widehat{p}_i$ for all true negatives are strictly less than those for all true positives, so that there is a perfect separation. In that case it is easy to check that the ROC curve traverses the path along the left vertical edge and then along the top horizontal edge of the square in Figure 7.5 as p^* decreases from 1 to 0. On the other hand, if the $\widehat{p}_i$'s for the true negatives and the true positives are uniformly mixed then the ROC curve traverses the path along the 45° line. A typical ROC curve falls between these two extremes.

The discriminating power of a binary classification model can be quantified by the **area under the curve (AUC)** of the ROC. Note that for the perfectly separated data, AUC $= 1$ and for perfectly mixed data, the ROC curve is the 45° line with AUC $= 1/2$.

The AUC can be shown to be the so-called **concordance index** γ, which is defined as follows. Let n_0 be the number of observations with $y_i = 0$ and n_1 be the number of observations with $y_i = 1$. The number of pairs of observations (i, j) with $y_i = 0, y_j = 1$ is $N = n_0 n_1$. Let $(\widehat{p}_i, \widehat{p}_j)$ denote the estimated success probabilities using a given model for the pair (i, j). If $\widehat{p}_i < \widehat{p}_j$ then the pair is said to be concordant, if $\widehat{p}_i > \widehat{p}_j$ then the pair is said to be discordant and if $\widehat{p}_i = \widehat{p}_j$ then the pair is said to be tied. Let N_c, N_d, N_t be the numbers of concordant, discordant and tied pairs with $N_c + N_d + N_t = N$. Then the concordance index is defined as $\gamma = (N_c + 0.5N_t)/N$.

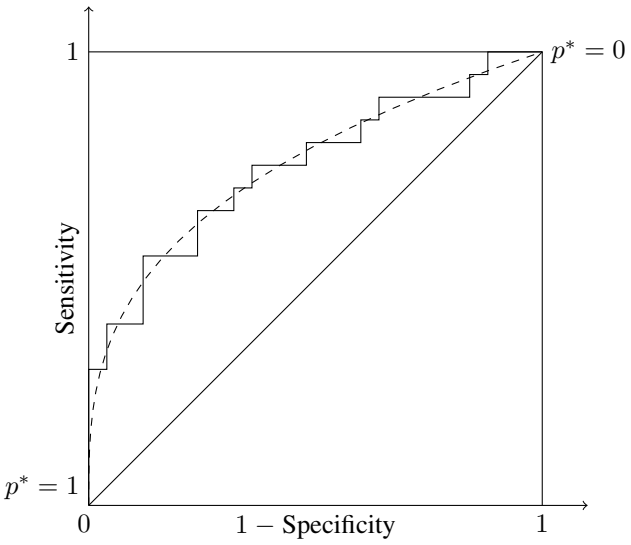


Figure 7.5 Typical ROC curve

EXAMPLE 7.9 (Art museum visits: AUC calculation for simple logistic regression)

In general, the AUC calculation is difficult and requires a computer program. However, it is relatively straightforward when there is a single ordered categorical predictor, as in the case of Education in Example 7.3 so that the $\widehat{p}_i$ are naturally ordered as seen in Table 7.3. Let n_{i0} and n_{i1} denote the number of "no" and "yes" responses for the ith Education level ($i = 1, \ldots, 8$). Here $\sum_{i=1}^{8} n_{i0} = n_0 = 1682$ and $\sum_{i=1}^{8} n_{i1} = n_1 = 925$. The n_{i0} and n_{i1} for each i are given in Table 7.2. Then since $\widehat{p}_i < \widehat{p}_j$ for $i < j$, the sum of the products $n_{i0}n_{j1}$ gives the total number of concurrences and the sum of the products $n_{i1}n_{j0}$ gives the total number of discordances. Thus

$$\begin{aligned} N_c &= \sum_{i=1}^{7} \sum_{j=i+1}^{8} n_{i0}n_{j1} = \sum_{i=1}^{7} n_{i0} \left(\sum_{j=i+1}^{8} n_{j1} \right) \\ &= 24(24 + \cdots + 199) + 92(92 + \cdots + 199) + \cdots + 96(199) = 892,008, \\ N_d &= \sum_{i=1}^{7} \sum_{j=i+1}^{8} n_{i1}n_{j0} = \sum_{i=1}^{7} n_{i1} \left(\sum_{j=i+1}^{8} n_{j0} \right) \\ &= 7(92 + \cdots + 150) + 24(408 + \cdots + 150) + \cdots + 107(150) = 406,807 \end{aligned}$$

and

$$N_t = N - N_c - N_d = 1682 \times 925 - 892,008 - 406,807 = 257,035.$$

So AUC equals

$$\text{AUC} = \gamma = \frac{N_c + 0.5N_t}{N} = \frac{892,008 + 0.5 \times 257,035}{1682 \times 925} = 0.6559.$$

■

EXAMPLE 7.10 (Art museum visits: ROC curves for simple and multiple logistic regression models)

The ROC curves can be plotted and their AUC's can be computed using the R function `plot.roc`, which requires the libraries `pROC` and `Rcpp`. Here is the R code:

```
> library(pROC)
> art = read.csv("c:/data/art.csv")
> summary(art)
> artnomiss1=na.omit(art[,c(1,6)])
> artnomiss2=na.omit(art[,-2])
> fit1 = glm(Visit ~ Education, family=binomial, data = artnomiss1)
> plot.roc(artnomiss1$Visit, fit1$fitted.values,xlab="Specificity")
> fit2 = glm(Visit ~ Education+Income+Gender+Children+Married,
>  factor(County)+ family=binomial, data = artnomiss2)
> plot.roc(artnomiss2$Visit, fit2$fitted.values,xlab="Specificity")
```

The output of this code is shown below.

```
Call:
plot.roc.default(x = artnomiss1$Visit, predictor = fit1$fitted.values,
    xlab = "Specificity")

Data: fit1$fitted.values in 1682 controls (artnomiss1$Visit 0) < 925 cases
(artnomiss1$Visit 1).
Area under the curve: 0.6559

Call:
plot.roc.default(x = artnomiss2$Visit, predictor = fit2$fitted.values,
    xlab = "Specificity")

Data: fit2$fitted.values in 1676 controls (artnomiss2$Visit 0) < 922 cases
    (artnomiss2$Visit 1).
Area under the curve: 0.6996
```

The ROC curves produced by this code are shown in Figure 7.6. Note that the AUC for the simple logistic regression model is 0.6559 as calculated in Example 7.9 and that for the multiple logistic regression model is 0.6996. Thus there is an increase in AUC due to inclusion of additional predictors, which helps to better discriminate between visitors and non-visitors. Also note that the horizontal axis in these plots is marked "Specificity" instead of "1 − Specificity" but the labels on the axis are reversed and go from 1 to 0 instead of 0 to 1. Thus what is plotted is actually 1 − Specificity.

The ROC curve for the simple logistic regression model does not appear to be a step function. This is due to the fact that since there are only 8 groups, changes in sensitivity and specificity occur simultaneously only at 8 points. Exercise 7.10 asks you to verify the plot of this ROC curve.

■

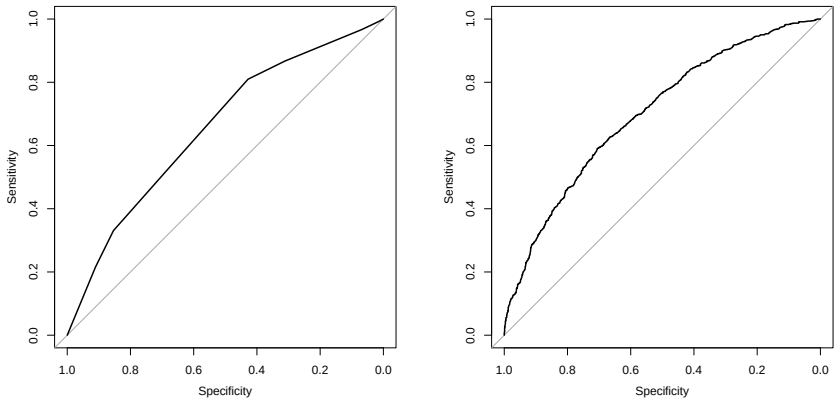


Figure 7.6 ROC curves for the simple and multiple logistic regression models from Examples 7.3 and 7.5

7.6 Polytomous Logistic Regression

As mentioned in the introduction of this chapter, polytomous responses may be nominal or ordinal. Different models are necessary for these two types of responses. When the response is binary, it makes no difference whether it is nominal or ordinal, so we use the regular binary logistic regression. From here on, assume that the the number of possible categorical responses, m, is at least 3, and the responses are coded numerically as $y = 1, \ldots, m$. Let $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ denote the vector of predictor variables. The probability of response $y = k$ is denoted by $p_k = P(y = k|\boldsymbol{x}) = p_k(\boldsymbol{x})$.

7.6.1 Logistic Regression for Nominal Response

In binary logistic regression we use a single logit, which compares the probability of success with the probability of failure. For general m, we use $m - 1$ logits, which use any one response as the reference and compare the probabilities of the remaining $m - 1$ responses to it. For convenience, we choose the last response as reference and consider the logits $\ln(p_k/p_m),\ k = 1, \ldots, m - 1$. This results in $m - 1$ logistic models:

$$\ln\left(\frac{p_k}{p_m}\right) = \beta_{0k} + \beta_{1k}x_1 + \cdots + \beta_{pk}x_p = \boldsymbol{x}'\boldsymbol{\beta}_k,\ (k = 1, \ldots, m - 1), \qquad (7.13)$$

where $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})'$ is an unknown parameter vector to be estimated, Note that we have a different $\boldsymbol{\beta}_k$ vector for each response $k = 1, \ldots, m - 1$.

The interpretation of the coefficient β_{jk} is similar to that of the β_1 coefficient for binary logistic regression. It is the change in the log-odds of response k relative to that of the reference category m when the predictor variable x_j is increased by one unit keeping all other variables fixed. As an example, suppose $\beta_{jk} = 0.5$ then $\exp(0.5) = 1.649$. Hence the odds of outcome k versus outcome m increase by a factor of 1.649 if x_j is increased by one unit.

From the logistic model (7.13) it follows that

$$p_k = p_m \exp(\boldsymbol{x}'\boldsymbol{\beta}_k),\ k = 1, \ldots, m - 1.$$

Hence

$$\sum_{k=1}^{m} p_k = p_m \sum_{k=1}^{m-1} \exp(\boldsymbol{x}'\boldsymbol{\beta}_k) + p_m = 1.$$

Solving for p_m we get

$$p_m = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(\boldsymbol{x}'\boldsymbol{\beta}_j)}$$

and substituting back in the formula for p_k we get

$$p_k = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta}_k)}{1 + \sum_{j=1}^{m-1} \exp(\boldsymbol{x}'\boldsymbol{\beta}_j)}.$$

Estimation

The maximum likelihood method is used to estimate the parameters of the model (7.13). Consider the data $(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n$ where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ is the vector of x variables for the ith observation and $y_i = k$ is the corresponding response for $k = 1, \ldots, m$. Let z_{ik} be the indicator variable of the outcome y_i, i.e., $z_{ik} = 1$ if $y_i = k$ and $z_{ij} = 0$ for $j \neq k$. Let $p_{ik} = p_k(\boldsymbol{x}_i) = P(z_{ik} = 1|\boldsymbol{x}_i) = P(y_i = k|\boldsymbol{x}_i)$. Note $\sum_{k=1}^{m} p_{ik} = 1$ for all observations $i = 1, \ldots, n$.

The likelihood function is given by

$$L = \prod_{i=1}^{n} \prod_{k=1}^{m} (p_{ik})^{z_{ik}},$$

where the inside product includes only those p_{ik} for which $z_{ik} = 1$. The log-likelihood function equals

$$\ln L = \sum_{i=1}^{n} \sum_{k=1}^{m} z_{ik} \ln p_{ik}.$$

The MLE $\widehat{\boldsymbol{\beta}}_k = (\widehat{\beta}_{0k}, \widehat{\beta}_{1k}, \ldots, \widehat{\beta}_{pk})'$ maximizes $\ln L$. The asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by the inverse of the information matrix, which can be used to make inferences on the parameters of the model.

Classification

To classify the outcome for any observation with given $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ vector, we can calculate the estimated probabilities of all m outcomes:

$$\widehat{p}_k(\boldsymbol{x}) = \exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}_k)\widehat{p}_m(\boldsymbol{x})\ (k = 1, \ldots, m-1) \quad \text{where} \quad \widehat{p}_m(\boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}_j)}.$$

Then we can predict the outcome k^* as that value of k, which maximizes $\widehat{p}_k(\boldsymbol{x})$. We refer to this as the **maximum probability classifier**.

Suppose we have available **prior probabilities**, $\pi_1, \ldots, \pi_m$, of the m possible responses, where $\sum_{k=1}^{m} \pi_k = 1$. Then their **posterior probabilities** are given by

$$\widehat{p}_k^*(\boldsymbol{x}) = \frac{\pi_k \widehat{p}_k(\boldsymbol{x})}{\sum_{j=1}^{m} \pi_j \widehat{p}_j(\boldsymbol{x})} \quad (1 \leq k \leq m).$$

The **Bayes classifier** (see Section 8.3) classifies $\boldsymbol{x}$ to that group k^*, which maximizes $\widehat{p}_k^*(\boldsymbol{x})$. If the prior probabilities are equal, $\pi_k = 1/m$ for all k then the Bayes classifier reduces to the maximum probability classifier. We will discuss the rationale behind the Bayes classifier in the next chapter.

EXAMPLE 7.11 (MBA Admissions: Nominal Logistic Regression Model)

The following R script is used to fit the nominal regression model. Note that the reference level 3 corresponds to the "don't admit" decision.

```
> library(mlogit)
> MBA=read.csv("c:/data/MBA.csv")
> MBA1 = mlogit.data(data = MBA, choice="admit", shape="wide",
> varying=NULL)
> fit1=mlogit(admit ~ 0 | GPA + GMAT, data = MBA1, reflevel = "3")
> summary(fit1)
```

The fitted model using the above script is shown below.

```
Coefficients :
                  Estimate  Std. Error t-value Pr(>|t|)
1:(intercept) -1941.67562  2905.80083 -0.6682   0.5040
2:(intercept) -1718.06457  2901.24843 -0.5922   0.5537
1:GPA           504.80288   783.89213  0.6440   0.5196
2:GPA           463.52086   783.33688  0.5917   0.5540
1:GMAT            1.16625     1.64324  0.7097   0.4779
2:GMAT            0.96982     1.63663  0.5926   0.5535

Log-Likelihood: -4.3078
McFadden R^2:  0.95376
Likelihood ratio test : chisq = 177.7 (p.value = < 2.22e-16)
```

Thus the estimated parameters of the nominal logistic regression model with $x_1 =$ GPA and $x_2 =$ GMAT are
$\widehat{\beta}_{01} = -1941.68, \widehat{\beta}_{02} = -1718.06, \widehat{\beta}_{11} = 504.80, \widehat{\beta}_{12} = 463.52, \widehat{\beta}_{21} = 1.17, \widehat{\beta}_{22} = 0.97.$

Surprisingly none of the coefficients is significant in this model. This is probably the result of a near complete separation between the three groups as seen in Figure 7.1 in which case fitting of the model is most difficult. This may seem counterintuitive but can be seen from the following example. Consider fitting a simple logistic regression model to the data in which all x-values corresponding to $y = 0$ are less than those corresponding to $y = 1$, so the two clusters are completely separated. Then the MLE of the slope β_1 of the logistic response curve must approach ∞ and so its MLE does not converge. Exercise 7.2 gives a small data set to illustrate this phenomenon. ■

7.6.2 Logistic Regression for Ordinal Response

Now suppose the responses are ordered: $1 < 2 < \cdots < m$. Then it makes sense to define cumulative probabilities $P(y \le k),\ k = 1, \ldots, m$. Note that $P(y \le m) = 1$. Next define **cumulative logits**:

$$\ln\left[\frac{P(y \le k)}{P(y > k)}\right], \quad k = 1, \ldots, m-1.$$

A linear model is postulated on these cumulative logits. The part of the model that depends on the predictor variables will be assumed to be common to all cumulative logits. Thus let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ denote a common parameter vector and $\boldsymbol{x} = (x_1, \ldots, x_p)'$ denote the predictor variable vector. Note that these vectors don't include the intercept term β_0 as is

the case with other models in this chapter. The final model is given by

$$\ln\left[\frac{P(y \le k)}{P(y > k)}\right] = \beta_{0k} + \boldsymbol{x}'\boldsymbol{\beta},\ k = 1, \ldots, m-1, \tag{7.14}$$

where $\beta_{01} < \beta_{02} < \cdots < \beta_{0,m-1}$. Note that this model is equivalent to

$$P(y \le k) = \frac{\exp(\beta_{0k} + \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\beta_{0k} + \boldsymbol{x}'\boldsymbol{\beta})},\ k = 1, \ldots, m-1. \tag{7.15}$$

The β_{0k} terms are constrained to be non-decreasing to ensure that the cumulative logits (and hence the cumulative probabilities) are non-decreasing.

Note from (7.14) that the difference in log-odds of two individuals i and j with covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for any category k is

$$\ln\left[\frac{P(y \le k|\boldsymbol{x}_i)}{P(y > k|\boldsymbol{x}_i)}\right] - \ln\left[\frac{P(y \le k|\boldsymbol{x}_j)}{P(y > k|\boldsymbol{x}_j)}\right] = (\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta},$$

which is independent of $k = 1, \ldots, m-1$. Because of this property, this model is referred to as the **proportional odds model**.

The derivation of model (7.15) is given in Section 7.7. The derivation assumes that there is a latent (unobservable) continuous response z and m intervals defined by unknown cutpoints $\beta_{01} < \cdots < \beta_{0,m-1}$ such that if z falls in the kth interval, $\beta_{0,k-1} \le z < \beta_{0k}$, then we observe the outcome $y = k$ as shown in Figure 7.8. The latent response z is assumed to have a logistic distribution with with mean $\mu = \boldsymbol{x}'\boldsymbol{\beta} = \beta_1 x_1 + \cdots + \beta_p x_p$.

Instead of the logistic distribution if we assume that $z \sim N(\mu = \boldsymbol{x}'\boldsymbol{\beta}, 1)$ then

$$P(y \le k) = P(z \le \beta_{0k}) = \Phi(\beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta})$$

or equivalently $\Phi^{-1}[P(y \le k)] = \beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta}$, where $\Phi(\cdot)$ is the standard normal c.d.f. and $\Phi^{-1}(\cdot)$ is its inverse. This is called the **probit model**. However, this model is not analytically tractable since $\Phi^{-1}(\cdot)$ cannot be expressed in a closed algebraic form. On the other hand, the standard logistic c.d.f. $F(x) = \exp(x)/[1 + \exp(x)] = p$ (say) and its inverse $F^{-1}(p) = \ln[p/(1-p)] = x$ have simple closed algebraic forms. As shown in Figure 7.7, the standard normal and standard logistic distributions are quite close. Now put $p = P(y \le k)$ and $x = \beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta}$ in these expressions, which yield the model (7.14), except for a "$-$" sign on $\boldsymbol{x}'\boldsymbol{\beta}$ instead of the "+" sign. This sign is a matter of convention. R uses the $-$ sign on the $\boldsymbol{x}'\boldsymbol{\beta}$ term.

Estimation

Once again, we use the maximum likelihood method to estimate the parameters of the model (7.14). Assume that the data are in the same format as for nominal logistic regression. We have

$$\begin{aligned} p_{ik} &= P(y_i = k) \\ &= P(y_i \le k) - P(y_i \le k-1) \\ &= \frac{\exp(\beta_{0k} + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\beta_{0k} + \boldsymbol{x}_i'\boldsymbol{\beta})} - \frac{\exp(\beta_{0,k-1} + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\beta_{0,k-1} + \boldsymbol{x}_i'\boldsymbol{\beta})}. \end{aligned}$$

The likelihood function is given by

$$L = \prod_{i=1}^{n}\prod_{k=1}^{m}(p_{ik})^{z_{ik}},$$

where the inside product includes only those p_{ik} for which $z_{ik} = 1$. The log-likelihood function equals

$$\ln L = \sum_{i=1}^{n}\sum_{k=1}^{m} z_{ik} \ln p_{ik}.$$

The MLE's $\widehat{\beta}_{01}, \ldots, \widehat{\beta}_{0,m-1}$ and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ maximize $\ln L$. The asymptotic variance-covariance matrix of the MLE's of the β's is given by the inverse of the information matrix, which can be used to make inferences on the parameters of the model.

Classification

To classify the outcome for any observation with given $\boldsymbol{x} = (x_1, \ldots, x_p)'$ vector, we can calculate the estimated probabilities of all m outcomes:

$$\widehat{p}_k(\boldsymbol{x}) = \frac{\exp(\widehat{\beta}_{0k} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})}{1 + \exp(\widehat{\beta}_{0k} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})} - \frac{\exp(\widehat{\beta}_{0,k-1} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})}{1 + \exp(\widehat{\beta}_{0,k-1} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})} \ (k = 1 \ldots, m-1)$$

and

$$\widehat{p}_m(\boldsymbol{x}) = 1 - \frac{\exp(\widehat{\beta}_{0,m-1} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})}{1 + \exp(\widehat{\beta}_{0,m-1} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}})}.$$

Then we can predict that outcome k^*, which maximizes $\widehat{p}_k(\boldsymbol{x})$. The prior probabilities can be incorporated in this rule in the same way as was done for the nominal logistic regression model.

EXAMPLE 7.12 (MBA Admissions: Ordinal Logistic Regression Model)

We will use the MBA admissions data as in Example 7.12 but we will take the order among the three groups into account to fit an ordinal logistic regression model using the following R script.

```
> library(ordinal)
> MBA$admit.ordered= as.ordered(MBA$admit)
> fit2=clm(admit.ordered~GPA+GMAT,data=MBA)
> summary(fit2)
```

The output is shown below.

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
GPA  -26.20487    7.46633  -3.510 0.000449 ***
GMAT  -0.05928    0.01932  -3.069 0.002150 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -111.24      32.10  -3.465
2|3   -99.61      28.81  -3.458
```

The negative coefficients of GPA and GMAT seem counterintuitive since increasing either one or both should increase the probability of an applicant to be admitted or wait-listed, i.e., increase $P(y \leq 1)$ and $P(y \leq 2)$. The explanation is that R fits the model

$$\ln\left[\frac{P(y \leq k)}{P(y > k)}\right] = \beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta}, \ k = 1, \ldots, m-1,$$

so the signs of the coefficients are reversed. Thus the actual fitted models are:

$$\ln\left[\frac{P(\text{admit})}{P(\text{wait-list or don't admit})}\right] = -111.24 + 26.205 \times \text{GPA} + 0.059 \times \text{GMAT}$$

and

$$\ln\left[\frac{P(\text{admit or wait-list})}{P(\text{don't admit})}\right] = -99.61 + 26.205 \times \text{GPA} + 0.059 \times \text{GMAT}.$$

Suppose we want to predict the admission decision for an applicant whose GPA $=$ 3.20 and GMAT $= 450$. Then using the function

```
> predict(fit2,newdata=data.frame(GPA=3.20,GMAT=450))
```

we get the predicted probabilities of the three outcomes as

```
          1         2             3
1 0.3307963 0.6691856 1.810068e-05
```

Thus there is a 33% chance that this student will be admitted and 67% chance that the student will be wait-listed; there is virtually no chance that the student will be denied admission.

For illustration purposes, we check the above probabilities by hand calculation. Toward this end, we first calculate two scores for this student:

$$X_1 = -111.24 + 26.2049 \times 3.20 + 0.0593 \times 450 = -0.6993$$

and

$$X_2 = -99.61 + 26.205 \times 3.20 + 0.059 \times 450 = 10.9307.$$

Then

$$P(y \le 1) = \frac{\exp(-0.6993)}{1+\exp(-0.6993)} = 0.3320 \text{ and } P(y \le 2) = \frac{\exp(10.9307)}{1+\exp(10.9307)} \approx 1.$$

Thus

$$p_1 = 0.3320, p_2 \approx 1 - 0.3320 = 0.6680 \quad \text{and} \quad p_3 \approx 0.$$

Suppose that historically 40% of the applicants are admitted, 20% are wait-listed and 40% are denied admission. We may use these percentages as prior probabilities and calculate posterior probabilities by combining them with the predicted probabilities calculated above as follows:

$$p_1^*(\boldsymbol{x}) = \frac{(0.40)(0.33)}{(0.40)(0.33)+(0.20)(0.67)} = 0.496, p_2^*(\boldsymbol{x}) = \frac{(0.20)(0.67)}{(0.40)(0.33)+(0.20)(0.67)} = 0.504$$

and $p_3^*(\boldsymbol{x}) \approx 0$. Thus, according to the Bayes classification rule, there is almost 50:50 chance that this applicant will be admitted or wait-listed. ■

7.7 Technical Notes

To derive the ordinal logistic regression model, we first introduce the **logistic distribution**. If x has a logistic distribution with mean (location parameter) μ and scale parameter τ (the standard deviation of x equals $\sigma = \pi\tau/\sqrt{3}$) then the p.d.f. and the c.d.f. of the standardized variable $z = (x-\mu)/\tau$ are given by

$$f(z) = \frac{\exp(z)}{[1+\exp(z)]^2} \quad \text{and} \quad F(z) = \frac{\exp(z)}{[1+\exp(z)]},$$

respectively. In Figure 7.7 we have plotted the $N(0,1)$ p.d.f. (shown by a dashed curve) and the logistic p.d.f. (shown by a solid curve) also with $\mu = 0$ and $\sigma = 1$. As can be seen the two p.d.f. curves match quite closely except in the center.

We can think of the observable ordinal outcome, $y = 1, 2, \ldots, m$, as resulting from an underlying unobservable continuous **latent variable** z as follows. Suppose that the real axis is divided into m intervals by unknown cutpoints $-\infty < \beta_{01} < \beta_{02} < \cdots <$

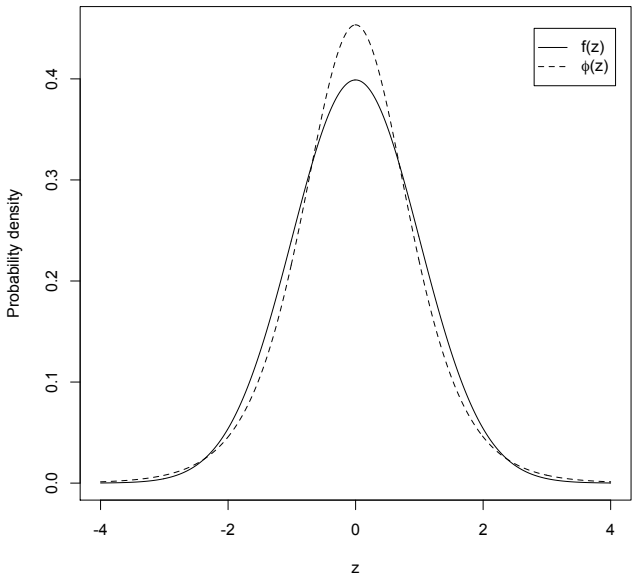


Figure 7.7 Plots of logistic density $f(z)$ (solid curve) and the standard normal density $\phi(z)$ (dashed curve) both with $\mu = 0$ and $\sigma = 1$

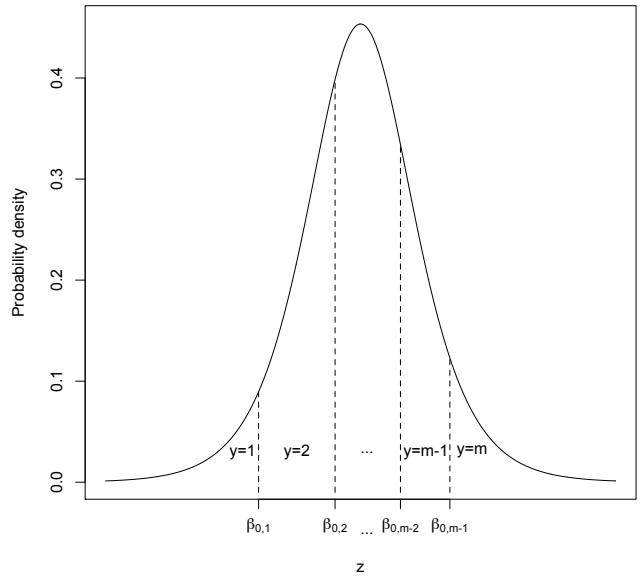


Figure 7.8 Latent variable distribution with cutpoints resulting in ordinal outcomes

$\beta_{0,m-1} < \infty$ such that if z falls in the kth interval, $[\beta_{0,k-1}, \beta_{0k}]$, then we observe the discrete outcome $y = k$, as shown in Figure 7.8. Next we assume that z has a logistic distribution with parameters μ and $\tau = 1$ and μ depends on the predictor variable vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ through a linear predictor, $\mu = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector. Then we have

$$P(y \le k) = P\left(z \le \beta_{0k}\right) = \frac{\exp(\beta_{0k} - \mu)}{1 + \exp(\beta_{0k} - \mu)} = \frac{\exp(\beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta})},$$

and hence

$$\ln\left[\frac{P(y \le k)}{P(y > k)}\right] = \beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta},\ k = 1, \ldots, m-1.$$

This is the same model as given in (7.15) except for a change of sign of the $\boldsymbol{x}'\boldsymbol{\beta}$ term. This reduces to the simple logistic regression model (7.1) if $m = 2$ in which case there is only one cutpoint $\beta_{01} = \beta_0$.

EXERCISES

Theoretical Exercises

7.1 **(2×2 contingency table)** Consider a clinical trial to compare a new treatment (coded as $x = 1$) with a control (coded as $x = 0$) for some disease. Denote their success probabilities by p_1 and p_0. Suppose that there are n_0 patients in the control group of whom

s_0 are successes and n_1 patients in the treatment group of whom s_1 are successes. The MLE's of p_0 and p_1 can be shown to be the sample proportions of successes $\widehat{p}_0 = s_0/n_0$ and $\widehat{p}_1 = s_1/n_1$.

a) Show that
$$\widehat{\beta}_1 = \ln \widehat{\psi} = \ln[\{\widehat{p}_1/(1-\widehat{p}_1)\}/\{\widehat{p}_0/(1-\widehat{p}_0)\}],$$
where $\widehat{\beta}_1$ is the MLE of β_1 in the logistic response model and $\ln \widehat{\psi}$ is the sample log-odds ratio.

b) Using the information matrix (7.6) derive the formula:
$$\widehat{\text{Var}}(\ln \widehat{\psi}) \approx \frac{1}{n_0 \widehat{p}_0(1-\widehat{p}_0)} + \frac{1}{n_1 \widehat{p}_1(1-\widehat{p}_1)}.$$

7.2 (Nonconvergence of MLE's in logistic regression) This exercise is based on Allison (2008). First consider completely separated data in the following table.

x	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
y	0	0	0	0	0	1	1	1	1	1

Because these data are symmetric, it can be shown that β_0 in the simple logistic regression model can be taken to be zero. So the likelihood function can be treated as a function only of the slope parameter β_1.

a) Plot the log-likelihood function with respect to β_1 for these data and check that it approaches the maximum value of 0 (i.e., the likelihood function approaches the maximum value of 1) as $\beta_1 \to \infty$. So the MLE of β_1 does not exist and the algorithm to find it does not converge.

b) Next consider quasi-separated data obtained by adding two observations $(x, y) = (0, 0)$ and $(x, y) = (0, 1)$ to the above data set and repeat the above exercise. Check that the log-likelihood function approaches a number less than 0 as $\beta_1 \to \infty$. So again the MLE of β_1 does not exist and the algorithm to find it does not converge.

7.3 (Deviance for grouped data) Consider the goodness of fit test given in Section 7.3.1 based on D^2, which treats the data as ungrouped. The following alternative test takes into account the grouping of the data into $g \geq 2$ groups.

a) It is easily shown that the MLE's of the success probabilities p_i for the grouped data under the saturated model are the sample proportions of successes $f_i = s_i/n_i$. Hence show that the deviance statistic for testing the goodness of fit is given by

$$\begin{aligned} D^2 &= -2\sum_{i=1}^{g} [\{s_i \ln \widehat{p}_i + (n_i - s_i)\ln(1-\widehat{p}_i)\} - \{s_i \ln f_i + (n_i - s_i)\ln(1-f_i)\}] \\ &= -2\sum_{i=1}^{g} \left[s_i \ln\left(\frac{\widehat{p}_i}{f_i}\right) + (n_i - s_i)\ln\left(\frac{1-\widehat{p}_i}{1-f_i}\right)\right]. \end{aligned}$$

How many d.f. does D^2 have?

b) Apply this deviance statistic and the Pearson chi-square statistic to the art museum visit data in Table 7.3, where the grouping variable is Education.

7.4 (Derivation of the logistic regression model assuming the covariate x is Poisson distributed) The simple logistic regression model can be derived as a conditional probability model, $P(y = 1|x)$, assuming that the covariate x has a Poisson distribution. As an example, consider automated screening of emails for phishing attempts. Suppose there

is a set of key words that are identified with phishing attempts and we want to use the frequency x of these key words as a covariate to classify incoming emails as phishing or non-phishing. Assume that the distributions of x conditioned on y ($y = 0$: non-phishing emails, $y = 1$: phishing emails) are

$$f_0(x) = f(x|y=0) = \frac{e^{-\mu_0}\mu_0^x}{x!} \quad \text{and} \quad f_1(x) = f(x|y=1) = \frac{e^{-\mu_1}\mu_1^x}{x!} \quad \text{for} \quad x = 0, 1, \ldots.$$

Also assume prior probabilities $1-\pi$ and π for non-phishing and phishing emails, respectively. Apply the Bayes formula to show that the posterior probability of $y = 1$ conditioned on x is given by

$$\begin{aligned} P(y=1|x) &= \frac{1}{1+\exp\left\{-\left[(\mu_0-\mu_1)) + \ln\left(\frac{\pi}{1-\pi}\right) + x\ln\left(\frac{\mu_1}{\mu_0}\right)\right]\right\}} \\ &= \frac{1}{1+\exp\{-(\beta_0+\beta_1 x)\}}, \end{aligned}$$

which is the logistic regression model. Here

$$\beta_0 = (\mu_0-\mu_1) + \ln\left(\frac{\pi}{1-\pi}\right) \quad \text{and} \quad \beta_1 = \ln\left(\frac{\mu_1}{\mu_0}\right),$$

Note that if $\mu_1 > \mu_0$ then $\beta_1 > 0$ and so $P(y=1|x)$ is an increasing function of x.

7.5 (Derivation of the logistic regression model assuming the covariate x is exponentially distributed) The diagnosis of some diseases are predicated on high or low values of the count of some blood chemical. For example, high values of the white blood cell count (WBC) is indicative of leukemia. Suppose we assume that the WBC count is exponentially distributed and derive a model for the conditional probability of leukemia given the WBC count (x). Assume that the distributions of x conditioned on y ($y = 0$: non-leukemia patients, $y = 1$: leukemia patients) are

$$f_0(x) = f(x|y=0) = \lambda_0 e^{-\lambda_0 x} \quad \text{and} \quad f_1(x) = f(x|y=1) = \lambda_1 e^{-\lambda_1 x} \quad \text{for} \quad x \geq 0.$$

The means of x for the two groups of patients are $\mu_0 = 1/\lambda_0$ and $\mu_1 = 1/\lambda_1$. Assume prior probabilities $1-\pi$ and π for non-leukemia and leukemia patients, respectively. Apply the Bayes formula to show that the posterior probability of $y = 1$ conditioned on x is given by

$$\begin{aligned} P(y=1|x) &= \frac{1}{1+\exp\left\{-\left[\ln\left(\frac{\lambda_1}{\lambda_0}\right) + \ln\left(\frac{\pi}{1-\pi}\right) + (\lambda_0-\lambda_1)x\right]\right\}} \\ &= \frac{1}{1+\exp\{-(\beta_0+\beta_1 x)\}}, \end{aligned}$$

which is the logistic regression model. Here

$$\beta_0 = \ln\left(\frac{\lambda_1}{\lambda_0}\right) + \ln\left(\frac{\pi}{1-\pi}\right) \quad \text{and} \quad \beta_1 = \lambda_0 - \lambda_1,$$

Note that if $\lambda_1 < \lambda_0$, i.e., $\mu_1 > \mu_0$, then $\beta_1 > 0$ and so $P(y=1|x)$ is an increasing function of x.

Applied Exercises

7.6 (2×2 contingency table) In a clinical trial to compare prednisone therapy (active control) with prednisone + VCR therapy (treatment) for leukemia the following data were obtained.

	Outcome		
Therapy	Success	Failure	Row Total
Prednisone	14	7	21
Prednisone + VCR	38	4	42

a) Calculate the sample log-odds ratio and show that it is significantly different from zero. Use the formula for $\widehat{\text{Var}}(\ln \widehat{\psi})$ from Exercise 7.1 to perform the test.

b) Do a two-sample z-test of $H_0 : p_0 = p_1$ using the test statistic

$$z = \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}(1-\widehat{p})[1/n_0 + 1/n_1]}},$$

where $\widehat{p}_0$ and $\widehat{p}_1$ are the sample proportions of successes for the Prednisone (Control) and Prednisone + VCR (Treatment) groups, respectively and $\widehat{p} = (n_0\widehat{p}_0 + n_1\widehat{p}_1)/(n_0+n_1)$ is the pooled (overall) sample proportion of successes. Compare the result with that of the test on the odds ratio from Part (a).

7.7 (Simpson's Paradox) Data are available in the data file `UCBAdmissions.csv` (derived from the data file `UCBAdmission` in the R library) and are summarized in Table 7.7 on 4526 applicants (2691 men and 1835 women) who applied for admission to six departments in a university. The admission rate for men was 44.5% (1198/2691) and that for women was 30.4% (557/1835). This naturally raised the question of sex discrimination. Although the overall admission rate was 14.1% lower for women than that for men, the admission rate for women was actually higher than that for men in 4 out of 6 departments, as can be seen from Table 7.7. This is called the Simpson's paradox.

a) Explain why Simpson's paradox occurs for these data.

b) Calculate the sample log odds ratio for Men vs. Women given their admission rates at the bottom of Table 7.7. Fit a logistic regression model to the data using only Gender (1 = Men, 0 = Women) as the predictor. Check that this sample log odds ratio is equal to the estimated β coefficient for Gender. Show this in general when the only predictor is a binary variable.

c) Next fit a second logistic regression model by including the Department as an additional predictor. Explain why the Gender coefficient changes sign from positive to negative, and how this illustrates Simpson's paradox.

7.8 (Radiation therapy) Twenty four cancer patients were treated with radiation therapy for different number of days (x) and the presence ($y = 0$) or absence ($y = 1$) of tumor was observed.

Table 7.7 Admissions Data for Men and Women

Dept.	Men			Women			Total		
	♯ Apply	♯ Admit	% Admit	♯ Apply	♯ Admit	% Admit	♯ Apply	♯ Admit	% Admit
A	825	512	62.1%	108	89	82.4%	993	620	64.4%
B	560	353	63.0%	25	17	68.0%	585	577	63.2%
C	325	120	36.9%	593	202	34.1%	918	322	35.1%
D	417	138	33.1%	375	131	34.9%	792	269	34.0%
E	191	53	27.7%	393	94	23.9%	584	147	25.2%
F	373	22	5.9%	341	24	7.0%	714	46	6.4%
Total	2691	1198	44.5%	1835	557	30.4%	4526	1755	38.8%

Days (x)	Response (y)	Days (x)	Response (y)
21	1	51	1
24	1	55	1
25	1	25	0
26	1	29	0
28	1	43	0
31	1	44	0
33	1	46	0
34	1	46	0
35	1	51	0
37	1	55	0
43	1	56	0
49	1	58	0

Source: Tanner (1996), p. 28.

a) Fit a binary logistic regression model to the data.
b) Calculate a 95% confidence interval for the odds of absence of tumor vs. presence of tumor if the number of days of therapy is increased by 5 days.
c) Calculate the estimated success probabilities $\widehat{p}_i$ for the 24 patients in the sample. Find the optimum threshold p^* that maximizes the correct classification rate (CCR). Calculate sensitivity, specificity and the F_1-score for this p^*.
d) Plot the ROC curve for this data set using the binary logistic regression model fitted in Part (a). What is the AUC for this ROC curve?

7.9 (Odds ratios for coronary disease data) Logistic regression analysis was done on data from a random sample of patients in a hospital about half of whom had coronary disease. The following coefficients were estimated along with their standard errors for three predictors: Age ($\widehat{\beta} = 0.0906, \text{SE} = 0.0184$), Cholesterol ($\widehat{\beta} = 0.0755, \text{SE} = 0.0136$) and Sex: Female = 0, Male = 1 ($\widehat{\beta} = 0.035, \text{SE} = 0.0148$).

a) What is the odds ratio of coronary disease for males vs. females. Calculate a 95% confidence interval for it.
b) If the odds of coronary disease for a female with Age = 50 and Cholesterol = 180 are 1 in 10, what are the odds for a male with Age = 60 and Cholesterol = 200? What is the corresponding probability of coronary disease for that male?

7.10 (ROC curve for simple logistic regression model for the art museum visits data)

a) Make a table of Sensitivity versus $1 -$ Specificity using the art museum visit data in Table 7.2.
b) Plot Sensitivity versus $1 -$ Specificity to obtain the ROC curve and check that it matches with the one in the left panel of Figure 7.6.

7.11 (Pregnancy duration) Kutner et al. (2005) gave data on 102 women whose pregnancy durations were classified as 1 = Preterm (less than 36 weeks), 2 = Intermediate term (36-37 weeks) and 3 = Full term (more than 37 weeks). The predictor variables are Nutrition (higher scores mean better nutrition), Alcohol use (1 = yes, 0 = no), Smoking history (1 = yes, 0 = no) and Age (1 = less than 20 years, 2 = 21 to 30 years, 3 = greater than 30 years). Treat Age as a categorical variable and use Age = 2 as the reference category since

mothers in this age group are known to have the lowest risk of pre-term delivery. Divide the data by putting all odd-numbered observations into the training set and all even-numbered observations into the test set.

a) Fit a nominal logistic regression model to the training set and make predictions for the test set using the maximum probability rule. What is the correct classification rate and how does it break down among the three categories?

b) Repeat the above exercise by fitting an ordinal logistic regression model. Do you get better predictions?

7.12 (Mammography testing history) Hosmer and Lemeshow (1989) gave data on 412 women who were asked about their mammography testing history with possible responses (y): Never = 0 (234 responses), Within the past year = 1 (104 responses) and More than one year ago = 2 (74 responses). There are two predictors: family history (HIST) of breast cancer (mother or sister) with values No = 0 and Yes = 1, and perception of benefit (PB) of mammography on a scale of 5-20 with low values representing high perception of benefit. Divide the data by putting all odd-numbered observations into the training set and all even-numbered observations into the test set.

a) Fit a nominal logistic regression model to the training set and make predictions for the test set using the maximum probability rule. What is the correct classification rate and how does it break down among the three categories?

b) Repeat the above exercise by fitting an ordinal logistic regression model. Make sure that you order the responses so that $0 < 2 < 1$. Do you get better predictions?

7.13 (Program choices by high school students) Entering high school students make a choice from three programs: academic, general and vocational. The file `program.csv` contains data on the program choices of 200 students along with the following possible predictors: reading, writing, math and science test scores, gender, type of school (public or private) and socio-economic status (seslow, sesmiddle and seshigh). The data are taken from `https://stats.idre.ucla.edu/sas/dae/multinomiallogistic-regression/`.

a) Fit a nominal logistic regression model using the best subset of predictors that minimizes AIC. Use the academic program as a reference response category. Which predictors are retained in the model? Interpret their effects on the program choice: vocational vs. academic and general vs. academic.

b) Calculate the probabilities of three program choices for a male student from high ses and a private school with median scores on four tests: reading = 50, writing = 54, math = 52 and science = 53 (note that some of these predictors may not be in the final model). Which choice is this student likely to make?

c) Repeat the above for the ordinal logistic regression model. Compare the results for the two models, in particular, with respect to the predictors in the final model and their interpretations.

d) Compute the classification matrices for the two models and the correct classification rates (CCR's). Which model gives a higher CCR?

CHAPTER 8

DISCRIMINANT ANALYSIS

Discriminant analysis is sometimes used as an alternative to logistic regression for classification. However, there is a fundamental conceptual difference between the two. In logistic regression we condition on predictors $x_1, \ldots, x_p$ and regard the binary outcome y as random with a Bernoulli distribution. We then model the logit of the expected value of this distribution (which is the success probability) as a function of the x's. On the other hand, in discriminant analysis we condition on the outcome $y = 0$ or 1 and x's as random with different distributions depending on the value of y. Because of this difference, in discriminant analysis we often refer to y as a group variable with values of y as the group or population labels. As an example, different groups of customers (e.g., high volume buyers, low volume buyers and non-buyers) may be characterized by certain attributes (predictors) and these attributes have different distributions depending on different groups that the customers belong to.

The predictors are ideally required to be numerical. Generally, we assume that the distribution of predictors is multivariate normal (the normality assumption) with a different mean vector for each group and a common covariance matrix across all groups (the homoscedasticity assumption). Under these assumptions, if there are two groups of observations, then we can derive a linear function of the predictors which best discriminates between the two groups. If there are $m > 2$ groups then we need more linear functions. This methodology is referred to as **linear discriminant analysis (LDA)**. If the homoscedasticity assumption is dropped then the discriminating functions are quadratic in predictors and the methodology is referred to as **quadratic discriminant analysis (QDA)**. We will use

the MBA admissions data from Chapter 7 to illustrate LDA; we will not cover QDA in this book.

8.1 Two-Group Discriminant Analysis

The linear discriminant function can be derived in several different ways. The simplest way is to use the minimum distance rule. Suppose we have p numerical predictors and we have a sample of n_i observations $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ $(j = 1, \ldots, n_i, i = 1, 2)$. Let $\overline{x}_{ik}$ denote the sample mean of the kth predictor from the ith group given by

$$\overline{x}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ijk} \ (1 \leq k \leq p, i = 1, 2).$$

Then we can form the sample mean vectors $\overline{\boldsymbol{x}}_i = (\overline{x}_{i1}, \ldots, \overline{x}_{ip})'$ $(i = 1, 2)$.

Consider a new observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$. The squared distance d_i^2 between $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}_i$ using the **Euclidean distance** can be computed as

$$d_i^2 = (x_1 - \overline{x}_{i1})^2 + \cdots + (x_p - \overline{x}_{ip})^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}}_i)'(\boldsymbol{x} - \overline{\boldsymbol{x}}_i) \quad (i = 1, 2). \tag{8.1}$$

The minimum distance rule assigns $\boldsymbol{x}$ to that group i for which d_i^2 is smallest, i.e., $\boldsymbol{x}$ is closest to the mean $\overline{\boldsymbol{x}}_i$.

There are two problems with the Euclidean distance function.

1. It combines variables with different units and scales of measurement, e.g., blood pressure and cholesterol level.
2. It does not take into account the different variances of the variables as well as correlations among them. For example, the variables with large variances should be weighted less and vice versa. Similarly, if two variables are highly correlated then both should not be highly weighted.

One way to address these problems is to use standardized variables for calculating the Euclidean distances, in other words,

$$d_i^2 = \left(\frac{x_1 - \overline{x}_{i1}}{s_1}\right)^2 + \cdots + \left(\frac{x_p - \overline{x}_{ip}}{s_p}\right)^2 \ (i = 1, 2),$$

where s_k^2 is the pooled (from both groups) sample variance of the kth variable given by

$$s_k^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)s_{1k}^2 + (n_2 - 1)s_{2k}^2\right] \ \ (1 \leq k \leq p),$$

where

$$s_{ik}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ijk} - \overline{x}_{ik})^2 \ (i = 1, 2).$$

Although the above definition of d_i^2 addresses the problems of different scales of measurement and different variances, it does not address the problem of correlations among the variables. For this purpose, we need to take into account also the sample covariances. Let $\boldsymbol{S} = \{s_{k\ell}\}$ denote the pooled sample covariance matrix. The diagonal entries of $\boldsymbol{S}$ are the sample variances $s_{kk} = s_k^2$ and the off-diagonal entries are the sample covariances $s_{k\ell}$ for $k \neq \ell$. The pooled sample covariances are given by

$$s_{k\ell} = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)s_{1k\ell} + (n_2 - 1)s_{2k\ell}\right] \ (1 \leq k < \ell \leq p),$$

where

$$s_{ik\ell} = \frac{1}{n_i - 1}\left[\sum_{j=1}^{n_i} (x_{ijk} - \overline{x}_{ik})(x_{ij\ell} - \overline{x}_{i\ell})\right] \ (1 \leq k < \ell \leq p, i = 1, 2).$$

We can then generalize (8.1) to

$$d_i^2 = (\boldsymbol{x} - \overline{\boldsymbol{x}}_i)' \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}_i) \ (i = 1, 2). \tag{8.2}$$

This is called the squared **Mahalnobis distance** between the new observation $\boldsymbol{x}$ and the sample mean vector $\overline{\boldsymbol{x}}_i$ of the ith group.

The minimum distance rule classifies the observation $\boldsymbol{x}$ to the group that has the smallest d_i^2. Simplifying the decision rule for classifying $\boldsymbol{x}$ to group 1, we get

$$\begin{aligned} d_1^2 < d_2^2 &\iff (\boldsymbol{x} - \overline{\boldsymbol{x}}_1)' \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}_1) < (\boldsymbol{x} - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}_2) \\ &\iff \boldsymbol{x}' \boldsymbol{S}^{-1} \boldsymbol{x} - 2 \overline{\boldsymbol{x}}_1' \boldsymbol{S}^{-1} \boldsymbol{x} + \overline{\boldsymbol{x}}_1' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_1 < \boldsymbol{x}' \boldsymbol{S}^{-1} \boldsymbol{x} - 2 \overline{\boldsymbol{x}}_2' \boldsymbol{S}^{-1} \boldsymbol{x} + \overline{\boldsymbol{x}}_2' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_2 \\ &\iff \overline{\boldsymbol{x}}_1' \boldsymbol{S}^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_1' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_1 > \overline{\boldsymbol{x}}_2' \boldsymbol{S}^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_2' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_2 \\ &\iff L_1 > L_2. \end{aligned} \tag{8.3}$$

where

$$L_i = \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_i' \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_i \quad (i = 1, 2). \tag{8.4}$$

L_i is called the **linear discriminant score** for the ith group ($i = 1, 2$). The quadratic term, $\boldsymbol{x}' \boldsymbol{S}^{-1} \boldsymbol{x}$, which is common to both d_1^2 and d_2^2, gets canceled from both sides of the above inequality. If the homoscedasticity assumption is dropped then we cannot pool the sample covariance matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ for the two groups. In that case, the inequality $d_1^2 < d_2^2$ becomes $Q_1 > Q_2$, where

$$Q_i = -\frac{1}{2} \boldsymbol{x}' \boldsymbol{S}_i^{-1} \boldsymbol{x} + \overline{\boldsymbol{x}}_i' \boldsymbol{S}_i^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_i' \boldsymbol{S}_i^{-1} \overline{\boldsymbol{x}}_i \quad (i = 1, 2), \tag{8.5}$$

which is called the **quadratic discriminant function (QDF)**.

8.1.1 Fisher's Linear Discriminant Function (LDF)

Since the classification rule (8.3) for the two groups case depends only on a single difference $L_1 - L_2$, it can be reformulated in terms of a single linear function of $\boldsymbol{x}$ given by:

$$\text{LD} = \text{LD}(\boldsymbol{x}) = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1} \boldsymbol{x}, \tag{8.6}$$

called Fisher's linear discriminant function (LDF). As shown in Figure 8.1, $\text{LD} = \text{LD}(\boldsymbol{x})$ is the projection of the point $\boldsymbol{x}$ on the straight line defined by this function. Let $\text{LD}_1 = \text{LD}(\overline{\boldsymbol{x}}_1)$ and $\text{LD}_2 = \text{LD}(\overline{\boldsymbol{x}}_2)$ denote the projections of the group means $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ on the same line. Then (8.3) is equivalent to classifying $\boldsymbol{x}$ to group 1 if LD is closer to LD_1, otherwise to group 2.

Let $d_1 = |\text{LD} - \text{LD}_1|$ and $d_2 = |\text{LD} - \text{LD}_2|$ be the distances of LD from LD_1 and LD_2, respectively. Then this rule is equivalent to classifying observation $\boldsymbol{x}$ to Group 1 if $d_1 < d_2$, which simplifies to

$$\text{LD} > \overline{\text{LD}} \ \text{ if } \ \text{LD}_1 > \text{LD}_2 \ \text{ and } \ \text{LD} < \overline{\text{LD}} \ \text{ if } \ \text{LD}_1 < \text{LD}_2, \tag{8.7}$$

where $\overline{\text{LD}} = (1/2)(\text{LD}_1 + \text{LD}_2)$. The normal to the line $(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1} \boldsymbol{x}$ at $\overline{\text{LD}}$ is the maximally separating line between the clusters of data from the two groups.

Fisher derived this discriminant function using a geometric approach. He aimed to find a linear transformation of the observation vectors $\boldsymbol{x}_{ij}$ to univariate values $y_{ij} = \boldsymbol{c}' \boldsymbol{x}_{ij} = c_1 x_{ij1} + \cdots + c_p x_{ijp}$ ($i = 1, 2; 1 \le j \le n_i$) such that the transformed values y_{1j} from group 1 and y_{2j} from group 2 are maximally separated in terms of the squared t-statistic (or equivalently the F-statistic) for comparing their means:

$$t^2(\boldsymbol{c}) = \frac{(\overline{y}_1 - \overline{y}_2)^2}{s_y^2 \, (1/n_1 + 1/n_2)}.$$

We can ignore the term $(1/n_1 + 1/n_2)$ in maximization. Here $\overline{y}_1$ and $\overline{y}_2$ are the sample means of y_{1j} and y_{2j}, respectively, and s_y^2 is the pooled sample variance of the y_{ij} given

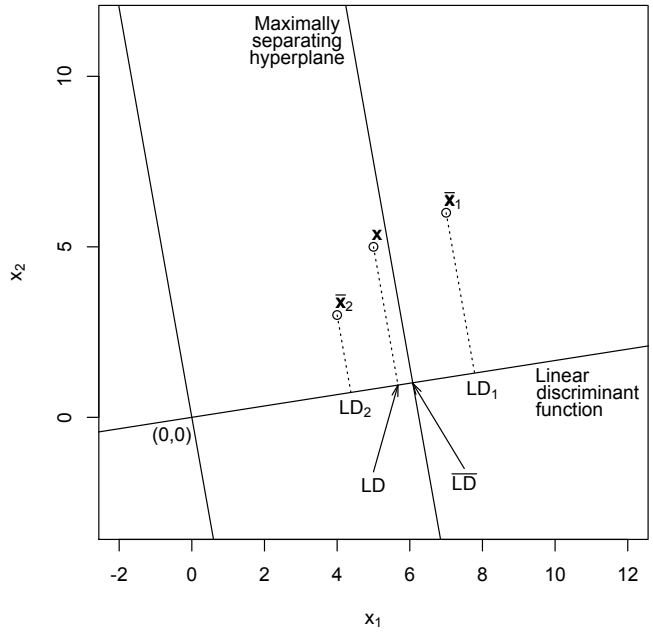


Figure 8.1 Linear discriminant function for two groups

by

$$s_y^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2\right].$$

Now note that $\overline{y}_1 = \boldsymbol{c}'\overline{\boldsymbol{x}}_1, \overline{y}_2 = \boldsymbol{c}'\overline{\boldsymbol{x}}_2$ and $s_y^2 = \boldsymbol{c}'\boldsymbol{S}\boldsymbol{c}$. Hence

$$t^2(\boldsymbol{c}) \propto \frac{[\boldsymbol{c}'(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)]^2}{\boldsymbol{c}'\boldsymbol{S}\boldsymbol{c}}.$$

The Technical Notes section extends this to multiple groups.

Next we give two small numerical examples to illustrate the calculation of linear discriminating functions and classification.

EXAMPLE 8.1 (Calculation of LDF's for Uncorrelated x_1 and x_2)

Consider two-group discriminant analysis with two predictors, x_1 and x_2. Suppose the mean vectors of the two groups are $\overline{\boldsymbol{x}}_1 = (7, 6)'$ and $\overline{\boldsymbol{x}}_2 = (4, 3)'$, and the sample covariance matrix of x_1 and x_2 is a diagonal matrix (i.e., x_1 and x_2 are uncorrelated):

$$\boldsymbol{S} = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}.$$

Its inverse is

$$\boldsymbol{S}^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}.$$

So the LDF's are

$$L_1 = (7,6)\begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(7,6)\begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}\begin{bmatrix} 7 \\ 6 \end{bmatrix} = \frac{7}{4}x_1 + \frac{2}{3}x_2 - \frac{65}{8}$$

and

$$L_2 = (4,3)\begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(4,3)\begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} = x_1 + \frac{1}{3}x_2 - \frac{5}{2}.$$

Consider a new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$. Then we have $L_1 = (7/4)5 + (2/3)5 - 65/8 = 3.958$ and $L_2 = 5 + (1/3)5 - 5/2 = 4.167$. Since $L_1 < L_2$, this observation is classified to group 2.

Using Fisher's LDF we have

$$\text{LD} = (7-4, 6-3)\begin{bmatrix} 1/4 & 0 \\ 0 & 1/9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{3}{4}x_1 + \frac{3}{9}x_2.$$

For the new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$, we have $\text{LD} = (3/4)5 + (3/9)5 = 5.147$. Furthermore, $\text{LD}_1 = (3/4)7 + (3/9)6 = 7.250$ and $\text{LD}_2 = (3/4)4 + (3/9)3 = 4.000$, so $\overline{\text{LD}} = 5.625$. Since $\text{LD} < \overline{\text{LD}}$ the observation is classified to group 2. Note that $|\text{LD} - \text{LD}_1| = |5.147 - 7.250| = 2.103 > |\text{LD} - \text{LD}_2| = |5.147 - 4.000| = 1.147$. Thus LD is closer to LD_2 than to LD_1. ■

EXAMPLE 8.2 (Calculation of LDF's for Correlated x_1 and x_2)

Assume the same variances as in the above example but suppose that the sample $\text{Corr}(x_1, x_2) = 0.5$ so that $\text{Cov}(x_1, x_2) = 0.5\sqrt{4 \times 9} = 3$. Thus the sample covariance matrix is

$$\boldsymbol{S} = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}.$$

Its inverse equals

$$\boldsymbol{S}^{-1} = \begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}.$$

So the LDF's are

$$L_1 = (7,6)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(7,6)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} 7 \\ 6 \end{bmatrix} = \frac{5}{3}x_1 + \frac{1}{9}x_2 - \frac{37}{6}$$

and

$$L_2 = (4,3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(4,3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} = x_1 - 2.$$

For a new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$, we have $L_1 = (5/3)5 + (1/9)5 - (37/6) = 2.722$ and $L_2 = 5 - 2 = 3.000$. Since $L_1 < L_2$, this observation is classified to group 2.

Using Fisher's LDF we get

$$\text{LD} = (7-4, 6-3)\begin{bmatrix} 1/3 & -1/9 \\ -1/9 & 4/27 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{2}{3}x_1 + \frac{1}{9}x_2.$$

For the new observation $\boldsymbol{x} = (x_1, x_2)' = (5,5)'$, we have $\text{LD} = (2/3)5 + (1/9)5 = 3.889$. Furthermore, $\text{LD}_1 = (2/3)7 + (1/9)6 = 5.333$ and $\text{LD}_2 = (2/3)4 + (1/9)3 = 3.000$, so $\overline{\text{LD}} = 4.167$. Since $\text{LD} < \overline{\text{LD}}$ the observation is classified to group 2. Note that $|\text{LD} - \text{LD}_1| = |3.889 - 5.333| = 1.444 > |\text{LD} - \text{LD}_2| = |3.889 - 3.000| = 0.889$. Thus LD is closer to LD_2 than to LD_1. So the classification decision is not affected by

the introduction of correlation in this example. Figure 8.1 is based on this calculation. ■

8.2 Multiple Group Discriminant Analysis

Extension from two-groups to multiple groups is fairly straightforward. The sample data available from the ith group consists of n_i observations, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})$, $j = 1, \ldots, n_i$ $(1 \leq i \leq m)$ and the total sample size is $n = \sum_{i=1}^{m} n_i$. From these data we can calculate the sample mean vectors $\overline{\boldsymbol{x}}_i$ and the sample covariance matrix $\boldsymbol{S}$ pooled from the group sample covariance matrices $\boldsymbol{S}_i$ as follows:

$$\boldsymbol{S} = \frac{1}{n-m}\left[(n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_m - 1)\boldsymbol{S}_m\right].$$

Following the same development as in the two-group case, the minimum Mahalnobis distance function rule leads to m linear discriminant functions, $L_1, \ldots, L_m$, as defined in (8.4). A new observation $\boldsymbol{x}$ is classified to that group for which L_i is the largest.

Fisher's approach to LDA is a bit more complicated and its derivation is given in the Technical Notes section. Whereas in the case of two groups we have a single discriminant function LD, in the case of $m > 2$ groups we have $q = \min(m-1, p)$ discriminant functions, denoted by $\text{LD}_1, \ldots, \text{LD}_q$. These discriminant functions are ordered in terms of the extent to which they separate the m groups; LD_1 separates the groups maximally, LD_2 is next, and so on. They are uncorrelated with each other. The fraction of separation captured by any discriminant function is measured by the proportion of the trace of a certain matrix (explained in Section 8.4.1). We may decide to use only the first $r < q$ discriminant functions if they capture most of the separation.

To classify a new observation $\boldsymbol{x}$, we project it on to the first $r \leq q$ discriminant function axes giving $\text{LD}_1(\boldsymbol{x}), \ldots, \text{LD}_r(\boldsymbol{x})$, referred to as **discriminant scores**, where $\text{LD}_k(\boldsymbol{x}) = \boldsymbol{c}_k'\boldsymbol{x}$ and $\boldsymbol{c}_k = (c_{k1}, \ldots, c_{kp})'$ is the coefficient vector of LD_k $(1 \leq k \leq r)$. We can similarly compute the r discriminant scores for the means $\overline{\boldsymbol{x}}_1, \ldots, \overline{\boldsymbol{x}}_m$; denote them by $\text{LD}_1(\overline{\boldsymbol{x}}_i), \ldots, \text{LD}_r(\overline{\boldsymbol{x}}_i)$ $(1 \leq i \leq m)$. Then we compute the distances between the vector of discriminant scores of $\boldsymbol{x}$ from the vector of discriminant scores of each group mean $\overline{\boldsymbol{x}}_i$ and assign $\boldsymbol{x}$ to that group for which the distance is the shortest. Since the discriminant scores are uncorrelated, we can use the Euclidean distance. Thus define the squared distance between the vector of discriminant scores of $\boldsymbol{x}$ from the vector of discriminant scores of $\overline{\boldsymbol{x}}_i$ by

$$d_i^2 = [\text{LD}_1(\boldsymbol{x}) - \text{LD}_1(\overline{\boldsymbol{x}}_i)]^2 + \cdots + [\text{LD}_r(\boldsymbol{x}) - \text{LD}_r(\overline{\boldsymbol{x}}_i)]^2, \quad i = 1, \ldots, m.$$

We classify $\boldsymbol{x}$ to that group for which d_i^2 is smallest.

EXAMPLE 8.3 (MBA Admissions: Discriminant Analysis)

We use the MBA admissions data from Example 7.2 to illustrate multiple group discriminant analysis. For computation of linear discriminant functions we used `Minitab` which gives the following output.

```
Linear Discriminant Function for Groups

                1        2        3
Constant  -240.37  -177.32  -133.90
GPA        106.25    92.67    78.09
GMAT         0.21     0.17     0.17
```

As in Example 7.12, suppose we want to classify a new observation GPA $= 3.20$ and GMAT $= 450$. So we calculate

$$\begin{aligned} L_1 &= -240.37 + 106.25(3.20) + 0.21(450) &= 194.13, \\ L_2 &= -177.32 + 92.67(3.20) + 0.17(450) &= 195.72, \\ L_3 &= -133.90 + 78.09(3.20) + 0.17(450) &= 192.49. \end{aligned}$$

Since L_2 is the largest, the highest probability is assigned to the "wait-list" decision.

If we apply these discriminant functions to the MBA admissions data from which they are estimated then we get the following confusion matrix:

```
                    True Group
Put into Group      1      2      3
1                  27      1      0
2                   4     25      2
3                   0      0     26
Total N            31     26     28
N correct          27     25     26
```

So the CCR is $(27 + 25 + 26)/85 = 91.8\%$.

The `MASS` package in R calculates Fisher's LDF's. We run the following script to calculate these functions and to predict the probabilities of the three decisions.

```
> library(MASS)
> fit=lda(admit~GPA+GMAT,data=MBA,prior=c(1,1,1)/3)
> fit
> predict(fit,newdata=data.frame(GPA=3.20,GMAT=450))
> prob = fitted(fit, outcome= FALSE)
```

The R output is as follows.

```
Prior probabilities of groups:
        1         2         3
0.3333333 0.3333333 0.3333333

Group means:
       GPA     GMAT
1 3.403871 561.2258
2 2.992692 446.2308
3 2.482500 447.0714

Coefficients of linear discriminants:
             LD1         LD2
GPA  5.017202736  1.85401003
GMAT 0.008503148 -0.01448967

Proportion of trace:
   LD1    LD2
0.9644 0.0356
```

We see that LD_1 captures 96.44% of the separation. However, we shall use both LD_1 and LD_2 for illustration purposes in this example. The discriminant scores for

the new observation (GPA=3.20, GMAT=450) are obtained by the following matrix multiplication:

$$\begin{bmatrix} 5.0172 & 0.0085 \\ 1.8540 & -0.0145 \end{bmatrix} \begin{bmatrix} 3.20 \\ 450 \end{bmatrix} = \begin{bmatrix} 19.8800 \\ -0.5922 \end{bmatrix}.$$

A similar calculation gives the discriminant scores for the three group means:

$$\begin{bmatrix} 5.0172 & 0.0085 \\ 1.8540 & -0.0145 \end{bmatrix} \begin{bmatrix} 3.4039 & 2.9927 & 2.4825 \\ 561.23 & 446.23 & 447.07 \end{bmatrix} = \begin{bmatrix} 21.8485 & 18.8079 & 16.2553 \\ -1.8270 & -0.9219 & -1.8800 \end{bmatrix}.$$

Then we calculate the squared Euclidean distances between the discriminant scores for the new observation and the three group means as follows:

$$\begin{aligned} d_1^2 &= (19.8800 - 21.8485)^2 + (-0.5922 + 1.8270)^2 &&= 5.400, \\ d_2^2 &= (19.8800 - 18.8079)^2 + (-0.5922 + 0.9219)^2 &&= 1.258, \\ d_3^2 &= (19.8800 - 16.2553)^2 + (-0.5922 + 1.8800)^2 &&= 14.797. \end{aligned}$$

Since d_2^2 is the smallest of the three, we assign this student to the "wait-list" category. This result is in agreement with the previous result obtained using the Mahalnobis distance method, as it should be, since the two methods are equivalent. ■

8.3 Bayesian Classification

Suppose that prior information is available about the prevalence of the different groups and this information can be quantified in terms of **prior probabilities** $\pi_i = P(y = i)$ $(1 \leq i \leq m)$ where $\sum_{i=1}^m \pi_i = 1$. The **posterior probability** that a new observation $\boldsymbol{x}$ belongs to group i, $P(y = i|\boldsymbol{x})$, can be computed by combining the prior probability π_i with the probability density function (p.d.f.) of $\boldsymbol{x}$ assuming that it comes from group i, namely $f(\boldsymbol{x}|y = i) = f_i(\boldsymbol{x})$, by applying the **Bayes formula**:

$$\begin{aligned} \widehat{p}_i^*(\boldsymbol{x}) &= P(y = i|\boldsymbol{x}) \\ &= \frac{P(y = i) f(\boldsymbol{x}|y = i)}{f(\boldsymbol{x})} \\ &= \frac{\pi_i f_i(\boldsymbol{x})}{\sum_{j=1}^m \pi_j f_j(\boldsymbol{x})} \quad (1 \leq i \leq m). \end{aligned} \tag{8.8}$$

The **Bayes classifier** assigns the observation $\boldsymbol{x}$ to that group i, which maximizes the posterior probability $\widehat{p}_i^*(\boldsymbol{x})$.

We assume that $f_i(\boldsymbol{x})$ is a **multivariate normal (MVN) distribution** with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$ and a common covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{k\ell}\}$ where the diagonal elements are $\sigma_{kk} = \sigma_k^2 = \text{Var}(x_k)$ and the off-diagonal elements are $\sigma_{k\ell} = \text{Cov}(x_k, x_\ell)$ $(1 \leq k < \ell \leq p)$. Thus the m MVN distributions differ in their mean vectors but have a common covariance matrix (the homoscedasticity assumption). This MVN p.d.f. has the formula:

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right\}, \tag{8.9}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. This formula generalizes the well-known formula for the univariate normal p.d.f. and that for the bivariate normal distribution (see (2.23)). It may be noted that if we replace the population mean vector $\boldsymbol{\mu}_i$ by the sample mean vector $\overline{\boldsymbol{x}}_i$ and the population covariance matrix $\boldsymbol{\Sigma}$ by the sample covariance matrix $\boldsymbol{S}$ then $f_i(\boldsymbol{x}) \propto \exp\{-(1/2)d_i^2\}$. Thus a small d_i^2 implies a large value of the p.d.f. $f_i(\boldsymbol{x})$. In other words, the minimum Mahalnobis distance rule for classification of an observation $\boldsymbol{x}$

is equivalent to the maximum p.d.f. rule: classify the observation $\boldsymbol{x}$ to that group i which maximizes the p.d.f. $f_i(\boldsymbol{x})$.

Substituting the MVN p.d.f. in (8.8) and canceling the common terms from the numerator and denominator, we get

$$\widehat{p}_i^*(\boldsymbol{x}) = \frac{\pi_i \exp\left\{\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - (1/2)\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right\}}{\sum_{j=1}^m \pi_j \exp\left\{\boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - (1/2)\boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right\}} \quad (1 \le i \le m).$$

Replacing $\boldsymbol{\Sigma}$ by the sample covariance matrix $\boldsymbol{S}$ and $\boldsymbol{\mu}_i$ by the sample mean vector $\overline{\boldsymbol{x}}_i$, we see that the estimates of the above posterior probabilities simplify to

$$\widehat{p}_i^*(\boldsymbol{x}) = \frac{\pi_i \exp(L_i)}{\sum_{j=1}^m \pi_j \exp(L_j)}, \tag{8.10}$$

where the L_i are the linear discriminant functions defined in (8.4). We assign the new observation $\boldsymbol{x}$ to that group, which has the highest posterior probability. If the prior probabilities are equal, $\pi_i = 1/m$, then the Bayes classification rule reduces to the minimum Mahalnobis distance rule or the maximum probability rule (8.3).

EXAMPLE 8.4 (MBA Admissions: Posterior Probability Calculations)

In Example 8.3 we calculated $L_1 = 194.13, L_2 = 195.72$ and $L_3 = 192.49$. To simplify the calculation of the posterior probabilities we can subtract a common number, say 190, from all three L_i's since the common factor $\exp(190)$ would cancel from both the numerator and denominator. If the prior probabilities are equal then we get

$$\widehat{p}_1^* = \frac{\exp(4.13)}{\exp(4.13) + \exp(5.72) + \exp(2.49)} = 0.164,$$
$$\widehat{p}_2^* = \frac{\exp(5.72)}{\exp(4.13) + \exp(5.72) + \exp(2.49)} = 0.804,$$
$$\widehat{p}_3^* = \frac{\exp(2.49)}{\exp(4.13) + \exp(5.72) + \exp(2.49)} = 0.032.$$

So this student will be wait-listed. In Example 7.12 using ordinal logistic regression we obtained the probability of admit to be 33% and the probability of wait-list to be 67%, so the probabilities predicted by the two methods are different but are similarly ordered.

If the prior probabilities are unequal, say, $\pi_1 = 0.3, \pi_2 = 0.2$ and $\pi_3 = 0.5$ then the posterior probabilities are calculated as follows.

$$\widehat{p}_1^* = \frac{0.3\exp(4.13)}{0.3\exp(4.13) + 0.2\exp(5.72) + 0.5\exp(2.49)} = 0.218,$$
$$\widehat{p}_2^* = \frac{0.2\exp(5.72)}{0.3\exp(4.13) + 0.2\exp(5.72) + 0.5\exp(2.49)} = 0.712,$$
$$\widehat{p}_3^* = \frac{0.5\exp(2.49)}{0.3\exp(4.13) + 0.2\exp(5.72) + 0.5\exp(2.49)} = 0.070.$$

These probabilities are somewhat different from the previous ones but the classification decision is unchanged. ■

8.4 Technical Notes

8.4.1 Derivation of Fisher's Linear Discriminant Functions

First consider one-way univariate analysis of variance (ANOVA) with data x_{ij} $(1 \leq i \leq m, 1 \leq j \leq n_i)$ from $m \geq 2$ groups. Let $\overline{x}_{i\cdot}$ denote the mean of the ith group and $\overline{x}_{\cdot\cdot}$ denote the grand mean of all x_{ij}. Then the total sum of squares is defined as $T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{\cdot\cdot})^2$, which can be partitioned into the between sum of squares, $B = \sum_{i=1}^{m} n_i(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$ and the within sum of squares, $W = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i\cdot})^2$, so that $T = B + W$. We aim to maximally separate the groups by maximizing the between groups variation relative to the within groups variation. This is done by maximizing the ratio B/W, which is proportional to the ANOVA F-statistic.

In one-way multivariate analysis of variance (MANOVA) we have multivariate data $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ $(1 \leq i \leq m, 1 \leq j \leq n_i)$ from $m \geq 2$ groups. Analogous to B, W and T, we have matrices $\boldsymbol{B}, \boldsymbol{W}$ and $\boldsymbol{T}$, referred to as between, within and total sums of squares and cross-product (SSCP) matrices defined as follows:

$$\boldsymbol{B} = \sum_{i=1}^{m} n_i(\overline{\boldsymbol{x}}_{i\cdot} - \overline{\boldsymbol{x}}_{\cdot\cdot})(\overline{\boldsymbol{x}}_{i\cdot} - \overline{\boldsymbol{x}}_{\cdot\cdot})',$$

$$\boldsymbol{W} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{i\cdot})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{i\cdot})',$$

and

$$\boldsymbol{T} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot\cdot})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot\cdot})',$$

where $\overline{\boldsymbol{x}}_{i\cdot}$ and $\overline{\boldsymbol{x}}_{\cdot\cdot}$ are vector analogs $\overline{x}_{i\cdot}$ and $\overline{x}_{\cdot\cdot}$, respectively. They satisfy the MANOVA identity $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W}$. As in Section 8.1.1, this multivariate problem can be reduced to the univariate ANOVA problem by taking linear combinations $y_{ij} = \boldsymbol{c}'\boldsymbol{x}_{ij} = c_1 x_{ij1} + \cdots + c_p x_{ijp}$. Then it is not difficult to show that the between sum of squares for the y_{ij}'s equals $B = \boldsymbol{c}'\boldsymbol{B}\boldsymbol{c}$ and the within sum of squares equals $W = \boldsymbol{c}'\boldsymbol{W}\boldsymbol{c}$. Thus we want to find $\boldsymbol{c}$ to maximize

$$\frac{\boldsymbol{c}'\boldsymbol{B}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{W}\boldsymbol{c}}.$$

Since this ratio is invariant to any scalar multiple of $\boldsymbol{c}$, we put the constraint $\boldsymbol{c}'\boldsymbol{W}\boldsymbol{c} = 1$. Then by the Lagrangian multiplier method we need to maximize

$$f(\boldsymbol{c}, \lambda) = \boldsymbol{c}'\boldsymbol{B}\boldsymbol{c} - \lambda(\boldsymbol{c}'\boldsymbol{W}\boldsymbol{c} - 1),$$

where λ is the Lagrangian multiplier. Write

$$\frac{\partial f(\boldsymbol{c}, \lambda)}{\partial \boldsymbol{c}} = 2\boldsymbol{B}\boldsymbol{c} - 2\lambda\boldsymbol{W}\boldsymbol{c} = \boldsymbol{0} \;\Rightarrow (\boldsymbol{B} - \lambda\boldsymbol{W})\boldsymbol{c} = \boldsymbol{0} \;\Rightarrow (\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda\boldsymbol{I})\boldsymbol{c} = \boldsymbol{0},$$

where $\boldsymbol{0}$ is a null vector of dimension p. This is an eigenvalue problem. The solution $\boldsymbol{c} = \boldsymbol{c}_1$ to this equation is the eigenvector of $\boldsymbol{W}^{-1}\boldsymbol{B}$ corresponding to the largest eigenvalue $\lambda = \lambda_1$, which is the solution to the determinantal equation $|\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda\boldsymbol{I}| = 0$. It can be shown that λ_1 is the maximum of the ratio $(\boldsymbol{c}'\boldsymbol{B}\boldsymbol{c})/(\boldsymbol{c}'\boldsymbol{W}\boldsymbol{c})$. The first linear discriminant function LD_1 equals $\boldsymbol{c}_1'\boldsymbol{x}$. Note that $\boldsymbol{W}^{-1}\boldsymbol{B}$ is a matrix analog of B/W.

Now $\boldsymbol{W}^{-1}\boldsymbol{B}$ is of rank $q = \min(m-1, p)$ and so has q nonzero eigenvalues $\lambda_1 > \cdots > \lambda_q$ and associated eigenvectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$. These eigenvectors are mutually orthogonal w.r.t. $\boldsymbol{W}$, i.e., $\boldsymbol{c}_i'\boldsymbol{W}\boldsymbol{c}_j = 0$ for all $i \neq j$. Thus we have q linear discriminant functions, $\mathrm{LD}_1, \ldots, \mathrm{LD}_q$. Their associated eigenvalues quantify the proportion of trace of $\boldsymbol{W}^{-1}\boldsymbol{B}$

(which is the sum of the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}$) that they explain with LD_1 explaining the largest fraction of the trace.

EXERCISES

Theoretical Exercises

8.1 (Fisher's LDF for two groups) For two groups show that the coefficient vector for Fisher's linear discriminant function is given by $\boldsymbol{c} = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$.

8.2 (Equivalence between the LDF and Fisher's LDF) Show that the decision rule (8.3) in terms of LDF's is equivalent to the decision rule (8.7) in terms of Fisher's LDF.

Applied Exercises

8.3 (Coronary heart disease data) Baseline measurements were made on 832 white males free of coronary heart disease (CHD) on three risk factors: age, diastolic blood pressure (DBP) and cholesterol level (CHL). By the end of the study period, 71 subjects had developed CHD while 761 did not (NCHD). The following linear discriminant functions (L_{NCHD} and L_{CHD}) were computed.

Predictor	NCHD	CHD
Const.	-23.561	-28.726
Age	0.027	0.072
DBP	0.338	0.360
CHL	0.075	0.079

a) Based on the coefficients of these linear discriminant functions explain why the probability of CHD increases with increases in all three risk factors?

b) For a person of 50 years of age, 95 mm of Hg diastolic blood pressure and cholesterol level of 210 mg/dL, calculate the posterior probabilities of the CHD and NCHD assuming equal prior probabilities.

c) The mean vector for the NCHD group is $(44.81, 86.99, 201.27)'$ and that for the CHD group is $(56.86, 95.62, 221.51)'$. Calculate the Euclidean distances d_1 and d_2 of a new observation vector $(50, 95, 210)'$ from the NCHD and CHD group means. To which group does the Euclidean distance rule classify this person to? Why does this classification differ from that obtained in (b)?

8.4 (Fisher's iris data) Fisher (1936) used this data set to introduce discriminant analysis. The data consist of measurements on the sepal and petal dimensions of 50 samples of three species of iris flowers (iris setosa, iris virginica and iris versicolor). The data are in file `Iris.csv`.

a) Do a discriminant analysis of the data. Give Fisher's linear discriminant functions.

b) For an iris flower with the following dimensions: sepal length = 5.5 mm, sepal width = 3.0 mm, petal length = 4.0 mm, petal width = 1.5 mm, calculate the posterior probabilities of the three species assuming equal prior probabilities.

CHAPTER 9

GENERALIZED LINEAR MODELS

In previous chapters we studied mainly two classes of models: multiple regression and logistic regression. In multiple regression the response variable was assumed to have the normal distribution, while in logistic regression the response variable was assumed to have the Bernoulli distribution. In polytomous logistic regression the response variable was assumed to have the multinomial distribution. In practice many other types of response variable distributions are encountered, e.g., in survival and reliability studies, lifetimes may be assumed to be exponentially or gamma distributed. In other applications count type of response variables are encountered, which may be assumed to be Poisson distributed. Some examples of count data are number of defects in quality control studies, number of traffic accidents in transportation studies and number of disease cases in epidemiological studies. **Generalized linear models (GLM's)** introduced by Nelder and Wedderburn (1972) deal with these and many other response distributions which belong to a class called the **exponential family**.

GLM's use another key concept, called the **link function**. In multiple regression, a linear model is postulated on $E(y) = \mu$. In logistic regression, a linear model is postulated on the logistic transform of $E(y) = P(y = 1) = p$. A link function generalizes this idea. It is a function of $E(y)$ on which a linear model is postulated. GLM's employ these two generalizations to provide a class of useful predictive models.

9.1 Exponential Family and Link Function

9.1.1 Exponential Family

A distribution belonging to the **exponential family** has the following general form for its p.d.f. or p.m.f.:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right\} \tag{9.1}$$

for some functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. This form of the distribution is known as the **canonical form** and θ is referred to as the **natural parameter**, which is the main parameter of interest. On the other hand, ϕ is a nuisance parameter (not a main parameter of interest), and is referred to as the **dispersion parameter**. In many examples $a(\phi) = \phi = 1$ and so it may be dropped from (9.1) and the expression can be simplified.

For any exponential family distribution it can be shown that $E(y)$ and $\text{Var}(y)$ are given by (see the Technical Notes section for the derivation)

$$E(y) = \mu = b'(\theta) \quad \text{and} \quad \text{Var}(y) = a(\phi)b''(\theta), \tag{9.2}$$

where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$. $\text{Var}(y)$ is generally a function of μ, so it is often denoted by $V(\mu)$, called the **variance function**.

Here are four common examples of distributions belonging to the exponential family.

Normal Distribution

The normal distribution with mean μ and variance σ^2 can be expressed in the form (9.1) as follows:

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]\right\}. \end{aligned}$$

for $-\infty < y < \infty$. Hence we have

$$a(\phi) = \sigma^2, \theta = \mu, b(\theta) = \frac{\mu^2}{2} \text{ and } c(y;\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right].$$

Note that μ is the natural parameter, which is the mean of the distribution as follows from $b'(\theta) = (d/d\mu)(\mu^2/2) = \mu$ and σ^2 is the dispersion parameter, which is the variance of the distribution as follows from $a(\phi)b''(\theta) = \sigma^2(d^2/d\mu^2)(\mu^2/2) = \sigma^2$.

Binomial Distribution

The binomial distribution is the distribution of the number of successes in n i.i.d. Bernoulli trials each with success probability p. It can be expressed in the form (9.1) as follows:

$$\begin{aligned} f(y;n,p) &= \binom{n}{y}p^y(1-p)^{n-y} \\ &= \binom{n}{y}\left(\frac{p}{1-p}\right)^y(1-p)^n \\ &= \exp\left[y\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\binom{n}{y}\right]. \end{aligned}$$

for $y = 0, \ldots, n$. Hence we have

$$a(\phi) = 1, \theta = \ln\left(\frac{p}{1-p}\right), b(\theta) = -n\ln(1-p), c(y;\phi) = \ln\binom{n}{y}.$$

Thus the logistic transform $\ln[p/(1-p)]$ is the natural parameter.

It can be checked that the mean of the binomial distribution is

$$E(y) = b'(\theta) = \frac{db(\theta)}{dp}\frac{dp}{d\theta} = \frac{db(\theta)}{dp}\left(\frac{d\theta}{dp}\right)^{-1} = \frac{n}{1-p}\left(\frac{1}{p(1-p)}\right)^{-1} = np$$

and the variance is

$$\text{Var}(y) = a(\phi)b''(\theta) = \frac{d(np)}{dp}\frac{dp}{d\theta} = n\left(\frac{d\theta}{dp}\right)^{-1} = n\left(\frac{1}{p(1-p)}\right)^{-1} = np(1-p).$$

Poisson Distribution

The Poisson distribution can be expressed in the form (9.1) as follows:

$$\begin{aligned} f(y;\mu) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp(y\ln\mu - \mu - \ln y!). \end{aligned}$$

for $y = 0, 1, 2, \ldots$. Hence we have

$$a(\phi) = 1, \theta = \ln\mu, b(\theta) = \mu \text{ and } c(y;\phi) = -\ln y!.$$

Thus $\ln\mu$ is the natural parameter.

It can be checked that the mean of the Poisson distribution is

$$E(y) = b'(\theta) = \frac{d\mu}{d(\ln\mu)} = \left(\frac{d\ln\mu}{d\mu}\right)^{-1} = \mu$$

and the variance is

$$\text{Var}(y) = a(\phi)b''(\theta) = \frac{d\mu}{d(\ln\mu)} = \left(\frac{d\ln\mu}{d\mu}\right)^{-1} = \mu.$$

Gamma Distribution

The gamma distribution can be expressed in the form (9.1) as follows:

$$\begin{aligned} f(y;\lambda,\alpha) &= \frac{1}{\Gamma(\alpha)}\lambda^\alpha e^{-\lambda y}y^{\alpha-1} \\ &= \exp\{-\lambda y + \alpha\ln\lambda + (\alpha-1)\ln y - \ln\Gamma(\alpha)\} \end{aligned} \tag{9.3}$$

for $y \geq 0$, where $\Gamma(\alpha)$ is the **gamma function**.[1] Hence we have

$$\phi = \alpha, a(\phi) = 1, \theta = -\lambda, b(\theta) = -\alpha\ln\lambda, c(y,\phi) = (\alpha-1)\ln y - \ln\Gamma(\alpha).$$

Here α is the dispersion (shape) parameter and λ is known as the scale parameter. For $\alpha = 1$, we get the exponential distribution.

The mean and variance of y are given by

$$E(y) = b'(\theta) = \frac{d}{d(-\lambda)}(-\alpha\ln\lambda) = \frac{\alpha}{\lambda} \quad \text{and} \quad \text{Var}(y) = a(\phi)b''(\theta) = \frac{d}{d(-\lambda)}\left(\frac{\alpha}{\lambda}\right) = \frac{\alpha}{\lambda^2}.$$

9.1.2 Link Function

A link function is a monotone differentiable function $g(\mu)$ of $\mu = E(y)$. If $g(\mu)$ is chosen to be equal to the natural parameter θ then $g(\mu)$ is known as the **canonical link**

[1]For all α except for negative integers, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ and $\Gamma(0) = \Gamma(1) = 1$. $\Gamma(\alpha)$ satisfies the recursive relation $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. If α is a positive integer then we get $\Gamma(\alpha) = (\alpha-1)!$.

function. In general, $g(\mu)$ can be arbitrary and we denote it by η. A GLM posits a linear model on η:

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}'\boldsymbol{\beta}, \tag{9.4}$$

where η is called the **linear predictor**. If $\eta = \theta$, i.e., if we use the canonical link function as the linear predictor then the estimation of the regression parameters is simplified. For the gamma distribution, the natural parameter $\lambda = \alpha/\mu$, but the inverse function $g(\mu) = 1/\mu$ is generally used as the canonical link function. In some cases it is more convenient to use a link function different from the canonical link function. For instance, for the gamma distribution, $\ln \mu$ may be used as the link function instead of $1/\mu$ to ensure non-negativity constraint on μ.

It is important to remember that the fitting algorithm applies the link function to the mean of y. That transformation is not applied to y itself. Thus if we use the log-link function, y is not log-transformed.

9.2 Estimation of Parameters of GLM

9.2.1 Maximum Likelihood Estimation

Let $y_1, \ldots, y_n$ denote independent observations, each y_i having the same exponential family distribution (9.1) but with a different θ_i, which depends on the predictors $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ $(i = 1, \ldots, n)$. Assuming the canonical link function, we have

$$g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i'\boldsymbol{\beta} \quad (i = 1, \ldots, n). \tag{9.5}$$

We use the maximum likelihood method to estimate the unknown regression parameters β_j's. The equation for finding the MLE of $\boldsymbol{\beta}$ has the general form (see the Technical Notes section for a derivation):

$$\boldsymbol{X}'\boldsymbol{\mu} = \boldsymbol{X}'\boldsymbol{y}. \tag{9.6}$$

Here $\boldsymbol{X}$ is the $n \times (p+1)$ model matrix with row vectors $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ $(i = 1, 2, \ldots, n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, where $\mu_i = E(y_i)$. The above matrix equation can be written as a column vector of equations

$$\sum_{i=1}^{n} \mu_i \boldsymbol{x}_i = \sum_{i=1}^{n} y_i \boldsymbol{x}_i \quad \text{or} \quad E\sum_{i=1}^{n} y_i \boldsymbol{x}_i = \sum_{i=1}^{n} y_i \boldsymbol{x}_i. \tag{9.7}$$

This is a system of $p+1$ equations in $p+1$ unknowns, $\beta_0, \beta_1, \ldots, \beta_p$. Notice that this system of equations equates certain linear combinations (defined by the predictor vectors $\boldsymbol{x}_i$'s) of the $E(y_i)$'s to the same linear combinations of the observed values of the y_i's.

Once the MLE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained, the fitted or predicted values of the y_i are calculated from $\widehat{y}_i = \widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})$. For example, in Poisson regression $g(\mu_i) = \ln \mu_i$ and so $\widehat{y}_i = \exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})$. In gamma regression $g(\mu_i) = 1/\mu_i$ and so $\widehat{y}_i = 1/\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$.

As an illustration of (9.6), consider the LS estimator of $\boldsymbol{\beta}$ for multiple regression. Putting $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ in (9.6), we see that this reduces to $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$, which is the same equation (3.6).

For another example, consider the simple logistic regression model (7.1). In this case

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ and } \boldsymbol{\mu} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}.$$

Simplifying the matrix equation we get

$$p_1\begin{pmatrix}1\\x_1\end{pmatrix}+\cdots+p_n\begin{pmatrix}1\\x_n\end{pmatrix}=y_1\begin{pmatrix}1\\x_1\end{pmatrix}+\cdots+y_n\begin{pmatrix}1\\x_n\end{pmatrix}.$$

This is equivalent to

$$\sum_{i=1}^{n}p_i=\sum_{i=1}^{n}y_i \quad \text{and} \quad \sum_{i=1}^{n}x_ip_i=\sum_{i=1}^{n}x_iy_i,$$

which is Equation (7.4).

Although (9.6) looks simple, it is not easy to solve since in general since $\mu_i = E(y_i)$ is a nonlinear function of $\boldsymbol{\beta}$, e.g., in logistic regression $p_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/[1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]$. Therefore an iterative algorithm described below is used to solve it. Only in the case of multiple regression, (9.6) is a linear equation in $\boldsymbol{\beta}$, which it has a closed form solution (3.7).

9.2.2 Iteratively Reweighted Least Squares (IRWLS) Algorithm

In the above we discussed the MLE approach to fit the general linear model $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ $(1 \le i \le n)$. In this section we give a weighted least squares (WLS) approach to fit this model. In the Technical Notes section we show that the WLS approach is asymptotically equivalent to the MLE approach .

For the WLS approach we define a new response variable z_i by expanding $g(y_i)$ around $g(\mu_i)$ using the first order Taylor series:

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = z_i \quad (1 \le i \le n). \tag{9.8}$$

It follows that

$$\text{Var}(z_i) = \text{Var}(y_i - \mu_i)[g'(\mu_i))]^2 = V(\mu_i)[g'(\mu_i)]^2 \quad (1 \le i \le n). \tag{9.9}$$

Let $\boldsymbol{z} = (z_1, \ldots, z_n)'$ denote the response vector and consider fitting the linear model $E(\boldsymbol{z}) = \boldsymbol{X}\boldsymbol{\beta}$. Since, in general, the variances of the z_i's are unequal we use the WLS estimate of $\boldsymbol{\beta}$: $\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}$, where $\boldsymbol{W} = \text{diag}\{w_1, \ldots, w_n\}$ and the weights $w_i = 1/\text{Var}(z_i)$ (see Exercise **??**). As can be seen from (9.8) and (9.9), these weights as well as the z_i's are functions of the μ_i's and hence of the unknown $\boldsymbol{\beta}$ since $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. So they need to be iteratively estimated. At the rth iteration, the estimate $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}}^{(r)} = \left(\boldsymbol{X}'\widehat{\boldsymbol{W}}^{(r)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{(r)}\widehat{\boldsymbol{z}}^{(r)}, \tag{9.10}$$

where

$$\widehat{z}_i^{(r)} = g(\widehat{\mu}_i^{(r)}) + (y_i - \widehat{\mu}_i^{(r)})g'(\widehat{\mu}_i^{(r)}). \tag{9.11}$$

The estimated weight matrix $\widehat{\boldsymbol{W}}^{(r)}$ is a diagonal matrix with diagonal elements

$$\widehat{w}_i^{(r)} = \left[V(\widehat{\mu}_i^{(r)})\{g'(\widehat{\mu}_i^{(r)})\}^2\right]^{-1},$$

where

$$\widehat{\mu}_i^{(r)} = g^{-1}\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}^{(r)}\right).$$

The starting value $\widehat{\boldsymbol{\beta}}^{(0)}$ may be taken to be the LS estimate $\widehat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ from which $\widehat{\mu}_i^{(0)}$, $\widehat{\boldsymbol{W}}^{(0)}$ and $\widehat{\boldsymbol{z}}^{(0)}$ can be computed.

EXAMPLE 9.1 (IRWLS Algorithm for Logistic Regression)

Suppose that we have binary observations y_i, where $y_i = 1$ with probability p_i and $y_i = 0$ with probability $1 - p_i$. We want to fit the logistic regression model:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \quad (i = 1, \ldots, n)$$

to these data. In this case $E(y_i) = \mu_i = p_i$ and $\text{Var}(y_i) = V(p_i) = p_i(1 - p_i)$. Further,

$$g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) \quad \text{and} \quad g'(p_i) = \frac{1}{p_i(1-p_i)}.$$

Hence

$$z_i = \ln\left(\frac{p_i}{1-p_i}\right) + \frac{y_i - p_i}{p_i(1-p_i)}$$

and

$$w_i = \left[V(p_i)\{g'(p_i)\}^2\right]^{-1} = \left[p_i(1-p_i)\left(\frac{1}{p_i(1-p_i)}\right)^2\right]^{-1} = p_i(1-p_i).$$

At the rth iteration the estimate of p_i is given by

$$\widehat{p}_i^{(r)} = \frac{\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}^{(r)})}{1 + \exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}^{(r)})}.$$

This estimate is substituted in the above expressions for z_i and w_i to yield $\widehat{z}_i^{(r)}$ and $\widehat{w}_i^{(r)}$. The initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}$ may be obtained by running least squares regression of the $\boldsymbol{x}_i$'s on the 0-1 responses y_i's. ■

9.3 Deviance and AIC

Deviance as a measure of goodness of fit for a given model M in comparison to the saturated model SM was defined in Equation (7.10) as

$$D^2 = -2\left[\ln L_{\max}(\text{M}) - \ln L_{\max}(\text{SM})\right] \tag{9.12}$$

for the logistic regression model. This definition applies generally to any exponential family distribution. Here $L_{\max}(\text{M})$ is the maximum of the likelihood function under the model M and $L_{\max}(\text{SM})$ is obtained by substituting $\widehat{\mu}_i = y_i$, i.e., by estimating the unknown means of the y_i's by the observed y_i's (thus having as many parameters in the model as the number of observations).

The Akaike information criterion (AIC) of a given model M with $p + 1$ parameters is defined as

$$\text{AIC} = -2\ln L_{\max}(\text{M}) + 2(p+1) = D^2 - \ln L_{\max}(\text{SM}) + 2(p+1). \tag{9.13}$$

For the logistic regression model, $\ln L_{\max}(\text{SM}) = 0$ and so AIC simplifies to $D^2 + 2(p+1)$. However, this is not the case in general for other regression models as illustrated in the following examples.

EXAMPLE 9.2 (Deviance for normal distribution)

For the normal distribution, the likelihood function equals

$$L = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2}\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2}.$$

The log-likelihood function equals

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2.$$

If σ^2 is regarded as a known parameter then

$$\ln L_{\max}(\mathrm{M}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2,$$

where $\widehat{\mu}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and

$$\ln L_{\max}(\mathrm{SM}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - y_i)^2 = -\frac{n}{2}\ln(2\pi\sigma^2).$$

Therefore

$$D^2 = -2[\ln L_{\max}(\mathrm{M}) - \ln L_{\max}(\mathrm{SM})] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2 = \frac{\mathrm{SSE}}{\sigma^2}.$$

Note that $D^2 \sim \chi^2_{n-(p+1)}$. The `glm` function in R reports SSE as the residual deviance and SST as the null deviance ignoring the scaling factor $1/\sigma^2$.

For calculation of AIC, `glm` regards σ^2 as an unknown parameter and uses its MLE $\widehat{\sigma}^2 = \mathrm{SSE}/n$. Also, it counts σ^2 among the number of estimated parameters. Then

$$\begin{aligned} -2\ln L_{\max}(\mathrm{M}) &= n\ln 2\pi + n\ln\left(\frac{\mathrm{SSE}}{n}\right) + \left(\frac{n}{\mathrm{SSE}}\right)\mathrm{SSE} \\ &= n\ln 2\pi + n\ln(\mathrm{SSE}/n) + n. \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{AIC} &= -2\ln L_{\max}(\mathrm{M}) + 2(p+1) \\ &= n\ln 2\pi + n\ln\left(\frac{\mathrm{SSE}}{n}\right) + n + 2(p+1). \end{aligned} \tag{9.14}$$

Frequently, the constant term $n\ln 2\pi + n$ is omitted as is done in the AIC formula given in Chapter 6 since only the differences in AIC's matter when comparing different models.

These formulae are illustrated by refitting the multiple regression model of Example 3.3 to the GPA versus college entrance test scores data using the `glm` function in the following example. ■

EXAMPLE 9.3 (College GPA and Entrance Test Scores: Deviance and AIC)

To check these results we fitted a multiple regression model GPA $= \beta_0 + \beta_1$Verbal$+$ β_2Math $+ \varepsilon$ using the `glm` function resulting in the following output.

```
glm(formula = GPA ~ Verbal + Math, family = gaussian, data = gpa)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.570537   0.493749  -3.181  0.00297 **
Verbal       0.025732   0.004024   6.395 1.83e-07 ***
Math         0.033615   0.004928   6.822 4.90e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 0.1618257)

    Null deviance: 18.7735  on 39  degrees of freedom
Residual deviance:  5.9876  on 37  degrees of freedom
AIC: 45.547
```

```
Number of Fisher Scoring iterations: 2
```

Note that the null deviance = SST = 18.7735 and the residual deviance = SSE = 5.9876 by comparing them with the ANOVA Table 3.4. Next the AIC can be calculated using (9.14) as follows:

$$\text{AIC} = 40\ln 5.9876 + 40 - 40\ln 40 + 40\ln 2\pi + 2(3+1) = 45.547.$$

Although $p = 2$ in this example, `glm` uses $p = 3$ by counting σ^2 as an additional unknown parameter. ■

Note that for the normal distribution, $D^2 = \text{SSE} = \sum_{i=1}^n e_i^2$. More generally, for any exponential family, by writing $D^2 = \sum_{i=1}^n d_i^2$ we may define d_i as the **deviance residual** with its sign being that of $y_i - \widehat{\mu}_i$. We will see examples of this definition below. These residuals can be used to detect outliers in a similar manner as in the normal distribution case.

EXAMPLE 9.4 (Deviance for Poisson distribution)

The likelihood function for the Poisson distribution is

$$L = \prod_{i=1}^n \left[\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right]$$

and the log-likelihood function is

$$\ln L = \sum_{i=1}^n \left[-\mu_i + y_i \ln \mu_i - \ln y_i!\right].$$

Hence the maximum of the log-likelihood function under the given model is

$$\ln L_{\max}(\text{M}) = \sum_{i=1}^n \left[-\widehat{\mu}_i + y_i \ln \widehat{\mu}_i - \ln y_i!\right],$$

where $\widehat{\mu}_i = \exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})$. The log-likelihood function under the saturated model is obtained by setting $\widehat{\mu}_i = y_i$ yielding

$$\ln L_{\max}(\text{SM}) = \sum_{i=1}^n \left[-y_i + y_i \ln y_i - \ln y_i!\right].$$

Hence we get

$$D^2 = -2[\ln L_{\max}(\text{M}) - \ln L_{\max}(\text{SM})] = 2\sum_{i=1}^n \left[y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right]. \quad (9.15)$$

The deviance residuals are given by

$$d_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{2\left[y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right]} \ (i = 1, \ldots, n). \quad (9.16)$$

■

EXAMPLE 9.5 (Deviance for gamma distribution)

Assume that the response variables y_i follow the gamma distributions:

$$f(y_i) = \frac{1}{\Gamma(\alpha)}\lambda_i^\alpha y_i^{\alpha-1} e^{-\lambda_i y_i},$$

with a common dispersion parameter α (the homoscedasticity assumption). Note $E(y_i) = \mu_i = \alpha/\lambda_i$ or $\lambda_i = \alpha/\mu_i$. Although λ_i is the natural parameter, `glm`

simply uses the inverse link function $g(\mu_i) = 1/\mu_i$. The likelihood function is

$$L = \prod_{i=1}^{n} \left[\frac{1}{\Gamma(\alpha)} \lambda_i^{\alpha} y_i^{\alpha-1} e^{-\lambda_i y_i} \right]$$

and the log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \left[-\ln \Gamma(\alpha) + \alpha \ln \lambda_i + (\alpha - 1) \ln y_i - \lambda_i y_i \right].$$

Hence the maximum of the log-likelihood function under the given model is

$$\ln L_{\max}((\mathbf{M}) = \sum_{i=1}^{n} \left[-\ln \Gamma(\alpha) + \alpha \ln \widehat{\lambda}_i + (\alpha - 1) \ln y_i - \widehat{\lambda}_i y_i \right].$$

Substituting $\widehat{\lambda}_i = \alpha/\widehat{\mu}_i$ where $1/\widehat{\mu}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$, we get

$$\ln L_{\max}(\mathbf{M}) = \sum_{i=1}^{n} \left[-\ln \Gamma(\alpha) + \alpha \ln \left(\frac{\alpha}{\widehat{\mu}_i} \right) + (\alpha - 1) \ln y_i - \left(\frac{\alpha}{\widehat{\mu}_i} \right) y_i \right].$$

The log-likelihood function under SM is obtained by setting $\widehat{\mu}_i = y_i$ or equivalently $\widehat{\lambda}_i = \alpha/y_i$, thus yielding

$$\ln L_{\max}(\mathbf{SM}) = \sum_{i=1}^{n} \left[-\ln \Gamma(\alpha) + \alpha \ln \left(\frac{\alpha}{y_i} \right) + (\alpha - 1) \ln y_i - \left(\frac{\alpha}{y_i} \right) y_i \right].$$

Hence the deviance equals

$$\begin{aligned} D^2 &= -2[\ln L_{\max}\mathbf{M}) - \ln L_{\max}(\mathbf{SM})] \\ &= 2\alpha \sum_{i=1}^{n} \left[\left(\frac{y_i}{\widehat{\mu}_i} - 1 \right) - \ln \left(\frac{y_i}{\widehat{\mu}_i} \right) \right]. \end{aligned} \tag{9.17}$$

Notice that this D^2 is simply the D^2 for the exponential regression multiplied by the scale parameter α. However, AIC $= -2 \ln L_{\max}(M) + 2(p+1)$ is a more complex function of α.

The deviance residuals are given by

$$d_i = \text{sign}(y_i - \widehat{\mu}_i) \sqrt{2\alpha \left[\left(\frac{y_i}{\widehat{\mu}_i} - 1 \right) - \ln \left(\frac{y_i}{\widehat{\mu}_i} \right) \right]} \; (i = 1, \ldots, n). \tag{9.18}$$

■

9.4 Poisson Regression

Poisson regression of count data is a common application of GLM. **Log-linear models** used to analyze contingency tables (tables of count data cross-classified by multiple categorical variables) is an example of Poisson regression where all covariates are categorical. More generally, the x's can be any type of covariates. The canonical link is $g(\mu) = \ln \mu$. The model is fitted using the `glm` function by specifying `family=Poisson`.

A limitation of the Poisson distribution is that its variance equals its mean. It does not allow the flexibility of having variance greater or less than mean (referred to as **overdispersion** or **underdispersion**, respectively). Overdispersion is more common primarily due to omitted predictors. There are several options available to model count data having over or underdispersion. One option is to include a scale parameter ϕ in the model so that the variance function equals $V(\mu) = \phi\mu$. We can estimate ϕ by $\widehat{\phi} = D^2/[n - (p+1)]$ or by $\widehat{\phi} = X^2/[n - (p+1)]$ after the model is fitted, where X^2 is the Pearson chi-square statistic.

Table 9.1 Melanoma Cases and Rates (per 100,000) Data

No.	Area	Age	Population	Cases	Rate
1	1	1	2,880,262	61	2.118
2	1	2	564,535	76	13.462
3	1	3	592,983	98	16.527
4	1	4	450,740	104	23.073
5	1	5	270,908	63	23.255
6	1	6	161,850	80	49.428
7	2	1	1,074,246	64	5.958
8	2	2	220,407	75	34.028
9	2	3	198,119	68	34.323
10	2	4	134,084	63	46.985
11	2	5	70,708	45	63.642
12	2	6	34,233	27	78.871

Source:

`https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/`

The estimated regression coefficients $\widehat{\beta}$'s are unaffected by the introduction of the dispersion parameter ϕ but their standard errors are scaled by $\sqrt{\widehat{\phi}}$. Thus the $\widehat{\beta}$'s are less significant if $\widehat{\phi} > 1$ for overdispersed data and more significant if $\widehat{\phi} < 1$ for underdispersed data. The reason for the $\widehat{\beta}$'s to be unaffected by the introduction of the dispersion parameter is that Equation (9.6) for computing $\widehat{\beta}$'s is independent of the dispersion parameter and depends only on the means μ_i. For the same reason, as we shall see in Section 9.5, exponential regression and gamma regression give the same fitted model since the exponential distribution is a special case of the gamma distribution for the dispersion parameter $\alpha = 1$.

Another option is to use the **negative binomial distribution** to model the data; see Exercise 9.1. This distribution also belongs to the exponential family. For more options see the books by Myers et al. (2010) and Hardin and Hilbe (2012).

EXAMPLE 9.6 (Melanoma Cancer Data: Poisson Regression of Cases)

The `melanoma.csv` file contains the data reported by Koch et al. (1986) from the Third National Cancer Survey. This data set contains the number of new melanoma cases (Cases) from 1969 to 1971 among white males in two areas (Area 1, Area 2) of the country for the following age groups: < 35, 35-44, 45-54, 55-64, 65-74, > 74, which are coded as $1, \ldots, 6$ under the variable Age. The variable Population gives the size of the population at risk. The variable Rate is the incidence rate of new melanoma cases per 100,000 population. The data are shown in Table 9.1.

We are mainly interested in the effect of Age on Cases. However, using Age as a numerical predictor would imply that its effect varies linearly, so we will use it as a categorical variable (factor). We will also adjust for Area. In Figure 9.1 we have plotted Cases versus six age groups for the two areas. We see that the two areas differ significantly with Area 1 having higher numbers of case than Area 2 for all age groups except one. The R output for Poisson regression is shown below.

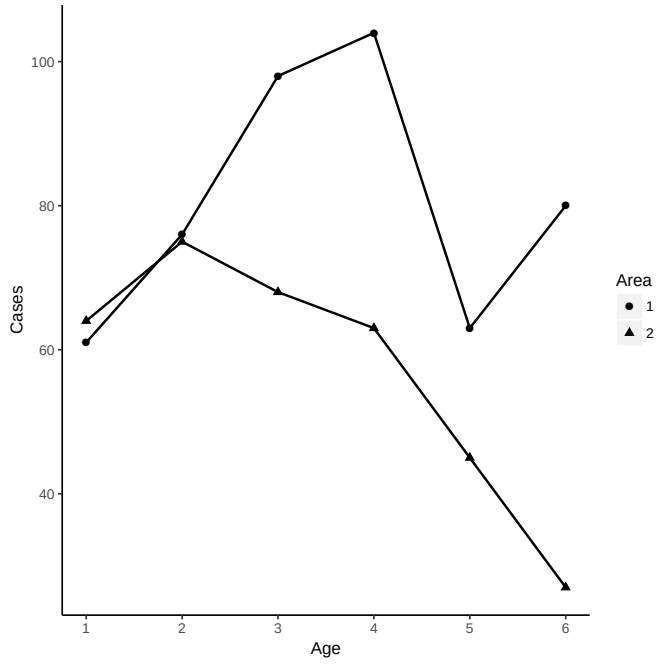


Figure 9.1 Melanoma cases for two areas versus age groups

```
> fit1=glm(Cases~Area+factor(Age), data=melanoma, family=poisson(log))
> summary(fit1)

Call:
glm(formula = Cases ~ Area + factor(Age), family = poisson,
    data = melanoma)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.6352     0.1342  34.540  < 2e-16 ***
Area         -0.3431     0.0707  -4.853 1.21e-06 ***
factor(Age)2  0.1890     0.1209   1.563   0.1181
factor(Age)3  0.2837     0.1184   2.395   0.0166 *
factor(Age)4  0.2897     0.1183   2.449   0.0143 *
factor(Age)5 -0.1462     0.1314  -1.113   0.2658
factor(Age)6 -0.1555     0.1317  -1.181   0.2378
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.240  on 11  degrees of freedom
Residual deviance: 22.259  on  5  degrees of freedom
AIC: 108.48

Number of Fisher Scoring iterations: 4

> SSE1 = sum((melanoma$Cases-fit1$fitted)^2) # SSE
> SSE1
[1] 1285.294
```

As seen in Figure 9.1, Area is highly significant with the coefficient of Area being negative, i.e., Area 1 has higher incidence of melanoma cases than Area 2. We also see a clear nonlinear effect of age with incidence of melanoma cases increasing compared to the youngest group until the age of 64 and then slightly decreasing (although the decline is not statistically significant). The SSE for this model is 1285.294.

To illustrate the calculation of a fitted value, consider Area = 2 and Age = 3. Then $\ln \widehat{\mu} = 4.6352 - 0.3431 + 0.2837 = 4.5758$ and so $\widehat{y} = \widehat{\mu} = e^{4.5758} = 97.106$. The observed number of cases at this combination is 68. So the residual is $68 - 97.106 = -29.106$. The deviance residual is

$$\text{sign}(68 - 97.106)\sqrt{2\left[68 \ln\left(\frac{68}{97.106}\right) - (68 - 97.106)\right]} = -3.123.$$

Notice that this residual is on log scale.

For comparison sake, we performed multiple regression using the square-root transformation of Cases as the response variable. The results are shown below.

```
> fit2=lm(sqrt(Cases)~Area+factor(Age), data=melanoma)
> summary(fit2)

Call:
lm(formula = sqrt(Cases) ~ Area + factor(Age), data = melanoma)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.0949     1.1541   8.747 0.000324 ***
Area          -1.4598     0.5960  -2.449 0.057976 .
factor(Age)2   0.7839     1.0323   0.759 0.481867
factor(Age)3   1.1677     1.0323   1.131 0.309286
factor(Age)4   1.1625     1.0323   1.126 0.311224
factor(Age)5  -0.5824     1.0323  -0.564 0.596996
factor(Age)6  -0.8349     1.0323  -0.809 0.455374
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.032 on 5 degrees of freedom
Multiple R-squared:  0.7267,    Adjusted R-squared:  0.3987
```

```
F-statistic: 2.216 on 6 and 5 DF,  p-value: 0.2002

> SSE2 = sum((melanoma$Cases-(fit2$fitted)^2)^2) # SSE
> SSE2
[1] 1297.955
```

We see that the effects of all predictors are similar to those obtained from Poisson regression, but they are all nonsignificant except the effect of Area, which is close to being significant. Surprisingly, the SSE for this model is only slightly higher than that for the Poisson regression model. ■

9.4.1 Poisson Regression for Rates

Poisson regression is used to model count data but often the counts of events depend on the amount of exposure. In Example 9.6, the number of cases depends on the population size. In epidemiological studies, the number of cases of a disease depends on the amount or time of exposure of some pollutant. In traffic studies the number of car accidents depends on the number of cars on the road or the traffic density. In such applications it is more appropriate to model the event rates rather than the counts of events while maintaining the Poisson nature of the counts.

For the ith observation, denote the number of cases by y_i, the covariate vector by $\boldsymbol{x}_i$ and the size of the exposure by N_i. Assume that y_i is Poisson distributed with mean μ_i. The mean event rate for the ith observation is then μ_i/N_i. We fit a linear model $\ln(\mu_i/N_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. This model can be written as $\ln \mu_i = \ln N_i + \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\ln N_i$ can be regarded as an additional covariate except that its β coefficient is fixed at 1. Such a term is called an **offset**.

This model can be fitted using `glm` as follows. The observed rates $r_i = y_i/N_i$ are treated as responses. The r_i are not in general integers as required by the Poisson distribution. This problem can be handled by specifying `family=quasipoisson`, which is an option available in the `MASS` library. Furthermore, we use the N_i as weights. These steps are illustrated in the example below.

EXAMPLE 9.7 (Melanoma Cancer Data: Poisson Regression of Rates)

In Figure 9.2 we have plotted Rate versus Age for the two areas. Notice the systematic increasing trends in Rate with Age for both areas. Cases showed decreasing trends with Age in Figure 9.1, but that was an artifact of decreasing population sizes with Age. Also observe that the number of melanoma cases are significantly higher for Area 1, but the incidence rate is significantly higher for Area 2. Again, this is the result of substantially smaller population sizes in Area 2 for each Age.

In the following we fit the Poisson regression model to Rate using the log link. If we use Cases/Population as the response variable then only the intercept term will be different; all other regression coefficients will be the same. Since Rate is defined as the number of Cases per 100,000 of Population, the intercept term for Cases/Population will be the intercept term for Rate $- \ln 10^4 = 0.03519 - 11.5129 = -11.4777$.

The R output is shown below.

```
Call:
```

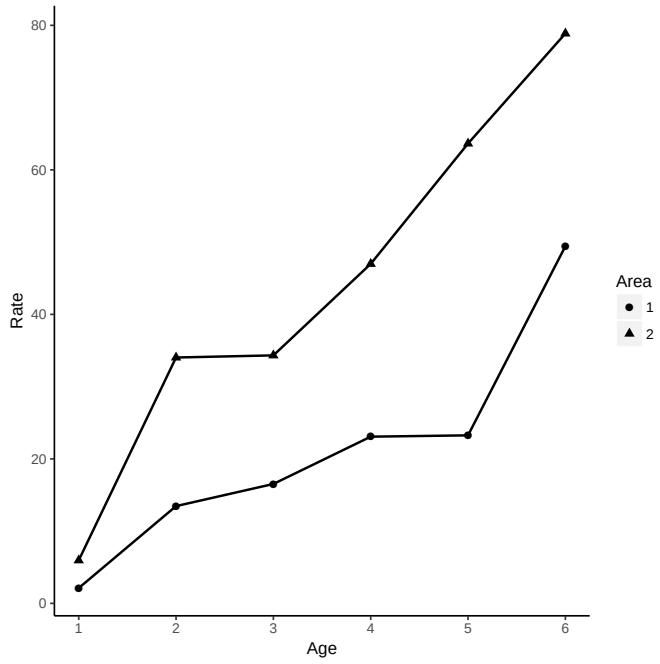


Figure 9.2 Melanoma Rates for Two Areas Versus Age Group

```
glm(formula = Rate ~ Area + factor(Age), family = quasipoisson,
    data = melanoma, weights = Population)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.03519    0.15135   0.232 0.825383
Area          0.81949    0.07855  10.432 0.000139 ***
factor(Age)2  1.79730    0.13373  13.439 4.08e-05 ***
factor(Age)3  1.91304    0.13098  14.605 2.72e-05 ***
factor(Age)4  2.24173    0.13087  17.130 1.24e-05 ***
factor(Age)5  2.36566    0.14545  16.265 1.60e-05 ***
factor(Age)6  2.94461    0.14603  20.164 5.55e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family taken to be 122308)

    Null deviance: 89579678  on 11  degrees of freedom
Residual deviance:   621520  on  5  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

> SSE3 = sum((melanoma$Cases-(fit4$fitted)*(melanoma$Population)
```

```
> /100000)^2) # SSE
> SSE3
[1] 350.7735
```

Notice that the SSE for this model is nearly one-fourth of the SSE's for the two models fitted for Cases. (Note that SSE in both examples is the sum of squared errors between the observed number of Cases and the predicted number of Cases.) The reason is that the Rate is a much more well-behaved variable and hence can be more accurately modeled.

An alternative approach to modeling the Rate is to treat Rate/10^5 = Cases/Population as a binomial proportion and use the logistic regression model with Population sizes as weights. The R output for this analysis is shown below.

```
Call:
glm(formula = Cases/Population ~ Area + factor(Age), family = binomial,
    data = melanoma, weights = Population)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -11.47811    0.13688  -83.86   <2e-16 ***
Area           0.81973    0.07104   11.54   <2e-16 ***
factor(Age)2   1.79756    0.12093   14.86   <2e-16 ***
factor(Age)3   1.91330    0.11845   16.15   <2e-16 ***
factor(Age)4   2.24211    0.11835   18.95   <2e-16 ***
factor(Age)5   2.36608    0.13153   17.99   <2e-16 ***
factor(Age)6   2.94531    0.13207   22.30   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 895.9558  on 11  degrees of freedom
Residual deviance:   6.2102  on  5  degrees of freedom
AIC: 92.431

Number of Fisher Scoring iterations: 4

> SSE4 = sum((melanoma$Cases-(fit4$fitted)*(melanoma$Population))^2) # SSE
> SSE4
[1] 350.347
```

Notice that the fitted model is almost identical to the Poisson regression model; also the SSE is also almost exactly the same. The reason is that the proportion p_i of people having melanoma in each (Area, Age) category is very small and so $\ln\left(\frac{p_i}{1-p_i}\right) \approx \ln p_i$. Thus the link function is approximately $\ln p_i$, the same as for the Poisson regression model for Rate. ■

Table 9.2 Leukemia Survival (in Days) Data

No.	$\log_{10}$(WBC)	Survival	No.	$\log_{10}$(WBC)	Survival
1	3.36	65	10	3.85	143
2	2.88	156	11	3.97	56
3	3.63	100	12	4.51	26
4	3.41	134	13	4.54	22
5	3.78	16	14	5.00	1
6	4.02	108	15	5.00	1
7	4.00	121	16	4.72	5
8	4.23	4	17	5.00	65
9	3.73	39			

9.5 Gamma Regression

As noted before, in gamma regression the canonical link is the inverse function $g(\mu) = 1/\mu$. To fit the model we use the `glm` function and specify `family=Gamma`. As in the case of Poisson regression, the dispersion parameter α is estimated after the specified linear model is fitted. The estimated value of α is printed by leaving out the value of the dispersion parameter in the `summary(fit)` statement. A particular value of α can be specified in the `summary(fit)` statement. For example, if we want to fit the exponential regression model then we use the statement `summary(fit,dispersion=1)`. The estimates of the regression coefficients are the same in both cases but the standard errors are multiplied by a factor $\sqrt{\widehat{\alpha}}$ in the former case. This factor follows from the fact that $\text{Var}(y) \propto \alpha$. Thus if $\widehat{\alpha} < 1$ then the estimated regression coefficients are more significant and vice versa.

EXAMPLE 9.8 (Leukemia Survival and WBC Count)

Table 9.2 gives data on the survival times of 17 leukemia patients and their $\log_{10}$(WBC) counts (WBC = white blood cell count). It is known that WBC count and survival are negatively correlated, This is demonstrated in the plot of Survival versus log(WBC) in Figure 9.3.

We first fit the exponential regression model to these data as shown in the following R output.

```
Call:
glm(formula = Survival ~ logWBC, family = Gamma, data = leukemia)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.034657   0.018626  -1.861   0.0628 .
logWBC       0.013528   0.005519   2.451   0.0142 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

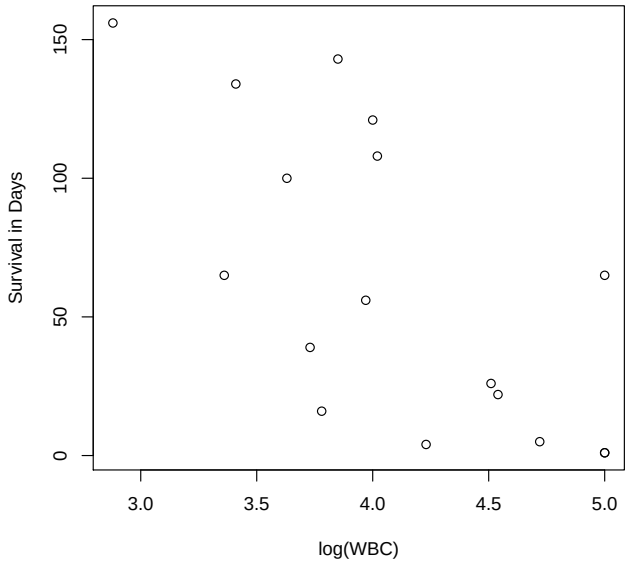


Figure 9.3 Survival in Days versus $\log_{10}$(WBC) for Leukemia Patients

```
(Dispersion parameter for Gamma family taken to be 1)

    Null deviance: 26.282  on 16  degrees of freedom
Residual deviance: 20.956  on 15  degrees of freedom
AIC: 175.46

Number of Fisher Scoring iterations: 6
```

The gamma regression model is obtained by simply printing the summary of the fit without specifying the dispersion parameter. The estimated dispersion parameter is $\widehat{\alpha} = 0.78134$. We get the same regression coefficients as for exponential regression but their standard errors are multiplied by $\sqrt{0.78134} = 0.884$ as shown in the R output below. So the t-values are scaled up by $1/0.884 = 1.131$ and hence are more significant. However, the deviance residuals are not scaled by $\sqrt{\widehat{\alpha}}$ as they should be according to the formula (9.18). Similarly, the deviance is not scaled by $\widehat{\alpha}$ as it should be. Thus the residual deviance for the gamma regression should be $0.78134 \times 20.956 = 16.374$ instead of 20.956, which is for the exponential regression.

```
Call:
glm(formula = Survival ~ logWBC, family = Gamma,
data = leukemia)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

log(WBC)	Survival	$\widehat{y}_i$	e_i	d_i
3.36	65	92.618	−27.618	−0.3344
2.88	156	232.361	−76.361	−0.3736
3.63	100	69.206	30.794	0.3921
3.41	134	87.158	46.842	0.4633
3.78	16	60.684	−44.684	−1.0925
4.02	108	50.696	57.304	0.8650
4	121	51.401	69.599	0.9979
4.23	4	44.314	−40.314	−1.7293
3.73	39	63.281	−24.281	−0.4479
3.85	143	57.386	85.614	1.0760
3.97	56	52.496	3.504	0.0653
4.51	26	37.945	−11.945	−0.3556
4.54	22	37.369	−15.369	−0.4869
5	1	30.319	−29.319	−2.2112
5	1	30.319	−29.319	−2.2112
4.72	5	34.252	−29.252	−1.4631
5	65	30.319	34.681	0.8732

```
(Intercept) -0.034657   0.016465  -2.105   0.0526 .
logWBC       0.013528   0.004879   2.773   0.0142 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.7813441)

    Null deviance: 26.282  on 16  degrees of freedom
Residual deviance: 20.956  on 15  degrees of freedom
AIC: 175.46

Number of Fisher Scoring iterations: 6
```

The fitted values for this model are given by $\widehat{y}_i = 1/(-0.0347 + 0.0135x_i)$ and the residuals are given by $e_i = y_i - \widehat{y}_i$. The SSE $= \sum_{i=1}^{n} e_i^2 = 33574.12$. The deviance residuals are calculated using (9.18). These calculations are readily done in a spreadsheet and are shown in Table 9.8.

We see that many of the fitted values deviate significantly from the observed Survival values, which means that the predictions made by this model are not very accurate. The main reason for the prediction inaccuracies is survival times fluctuate widely. As an alternative, we fit a simple linear regression model to log(Survival) resulting in the following R output.

```
Call:
glm(formula = log(Survival) ~ logWBC, family = gaussian, data = leukemia)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0738     2.0393   5.430 6.96e-05 ***
logWBC       -1.8829     0.4925  -3.823  0.00166 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 1.518751)

    Null deviance: 44.978  on 16  degrees of freedom
Residual deviance: 22.781  on 15  degrees of freedom
AIC: 59.22

Number of Fisher Scoring iterations: 2
```

The fitted values for this model are given by $\widehat{y}_i = \exp(11.0738 - 1.8829x_i)$ and the residuals are given by $e_i = y_i - \widehat{y}_i$. The SSE $= \sum_{i=1}^{n} e_i^2 = 49600.27$, which is almost 50% higher than that for the Poisson regression model. So this model does even a poorer job of prediction.

An alternative approach to fitting a model with log-transformation is to use the gamma family with the log link, i.e., instead of log-transforming the response variable Survival as done above, fit a linear model to $\ln \mu$. The R output is shown below.

```
Call:
glm(formula = Survival ~ logWBC, family = Gamma(log),
    data = leukemia)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.4775     1.6034   5.287 9.13e-05 ***
logWBC       -1.1093     0.3872  -2.865   0.0118 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.9388638)

    Null deviance: 26.282  on 16  degrees of freedom
Residual deviance: 19.457  on 15  degrees of freedom
AIC: 173.97

Number of Fisher Scoring iterations: 8
```

The fitted values for this model are given by $\widehat{y}_i = \exp(8.4775 - 1.1093x_i)$ and the residuals are given by $e_i = y_i - \widehat{y}_i$. The SSE $= \sum_{i=1}^{n} e_i^2 = 27211.34$, which is the least of all three models. ■

9.6 Technical Notes

9.6.1 Mean and Variance of the Exponential Family of Distributions

In this section we derive the formulae (9.2). We use the following two relationships from (B.7):

$$E\left[\frac{d\ln L}{d\theta}\right]=0 \quad \text{and} \quad E\left[\left(\frac{d\ln L}{d\theta}\right)^2\right]=-E\left[\frac{d^2\ln L}{d\theta^2}\right].$$

Substituting (9.1) for L we get

$$E\left[\frac{d\ln L}{d\theta}\right]=\frac{1}{a(\phi)}E[y-b'(\theta)]=0 \quad \Longrightarrow E(y)=\mu=b'(\theta).$$

Next

$$-E\left[\frac{d^2\ln L}{d\theta^2}\right]=\frac{1}{a(\phi)}b''(\theta)$$

and

$$E\left[\left(\frac{d\ln L}{d\theta}\right)^2\right]=\frac{1}{a^2(\phi)}E(y-b'(\theta))^2=\frac{1}{a^2(\phi)}\text{Var}(y).$$

Equating these two expressions we get $\text{Var}(y)=V(\mu)=a(\phi)b''(\theta)$.

9.6.2 MLE of $\boldsymbol{\beta}$ and Its Evaluation Using the IRWLS Algorithm

The log-likelihood function for GLM equals

$$\ln L(\boldsymbol{\beta})=\sum_{i=1}^{n}\left[\frac{1}{a(\phi)}[y_i\theta_i-b(\theta_i)]+c(y_i,\phi)\right].$$

To find the MLE of $\boldsymbol{\beta}$, we treat ϕ as a known constant; it is estimated separately from the deviance residuals after estimating $\boldsymbol{\beta}$. This is similar to how σ^2 is estimated in multiple regression from residuals by $\text{MSE}=\text{SSE}/[n-(p+1)]$ after estimating $\boldsymbol{\beta}$.

We assume that $a(\phi)$ is known and constant. If we use the canonical link function $\theta_i=\boldsymbol{x}_i'\boldsymbol{\beta}$ then we get

$$\begin{aligned}\frac{d\ln L(\boldsymbol{\beta})}{d\boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{d\ln L(\boldsymbol{\beta})}{d\theta_i}\frac{d\theta_i}{d\boldsymbol{\beta}}\\ &= \frac{1}{a(\phi)}\sum_{i=1}^{n}[y_i-b'(\theta_i)]\boldsymbol{x}_i \quad (\text{since } \theta_i=\boldsymbol{x}_i'\boldsymbol{\beta},\ \text{so } d\theta_i/d\boldsymbol{\beta}=\boldsymbol{x}_i)\\ &= \frac{1}{a(\phi)}\sum_{i=1}^{n}[y_i-\mu_i]\boldsymbol{x}_i.\end{aligned}$$

Setting this derivative equal to $\mathbf{0}$ (the null vector) and canceling the constant factor $1/a(\phi)$ results in Equation (9.6) by expressing it in matrix notation.

Next we show that solving this equation is asymptotically equivalent to solving Equation (9.10) with $z_i=g(\mu_i)+(y_i-\mu_i)g'(\mu_i)$ as the response variables and $w_i=1/\text{Var}(z_i)$ as the weights. The jth element of the vector $d\ln L(\boldsymbol{\beta})/d\boldsymbol{\beta}$ can be written as

$$\frac{\partial\ln L(\boldsymbol{\beta})}{\partial\beta_j}=\sum_{i=1}^{n}\frac{\partial\ell_i}{\partial\beta_j},$$

where

$$\ell_i=\ln L_i=\frac{1}{a(\phi)}[y_i-b(\theta_i)]+c(y_i,\phi) \quad (i=1,\ldots,n).$$

By the chain rule, we have

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{9.19}$$

These derivatives can be written as follows:

$$\begin{aligned}
\frac{d\ell_i}{d\theta_i} &= \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}, \\
\frac{d\theta_i}{d\mu_i} &= \left[\frac{d\mu_i}{d\theta_i}\right]^{-1} = [b''(\theta_i)]^{-1} = \frac{a(\phi)}{V(\mu_i)} \quad \text{(using (9.2))}, \\
\frac{\partial \eta_i}{\partial \beta_j} &= \frac{\partial(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \beta_j} = x_{ij}.
\end{aligned}$$

The second equation above can be reexpressed in terms of the weights w_i using (9.9) as follows:

$$\frac{d\theta_i}{d\mu_i} = \frac{a(\phi)}{V(\mu_i)(d\eta_i/d\mu_i)^2}\left(\frac{d\eta_i}{d\mu_i}\right)^2 = \frac{a(\phi)}{\text{Var}(z_i)}\left(\frac{d\eta_i}{d\mu_i}\right)^2 = w_i a(\phi)\left(\frac{d\eta_i}{d\mu_i}\right)^2.$$

Substituting in (9.19) we get

$$\frac{d\ell_i}{d\beta_j} = \frac{y_i - \mu_i}{a(\phi)} w_i a(\phi)\left(\frac{d\eta_i}{d\mu_i}\right)^2 \frac{d\mu_i}{d\eta_i} x_{ij} = w_i x_{ij}(y_i - \mu_i)\frac{d\eta_i}{d\mu_i}.$$

Thus we have to solve

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} w_i x_{ij}(y_i - \mu_i)\frac{d\eta_i}{d\mu_i} = 0 \quad (j = 0, 1, \ldots, p).$$

We use the Fisher scoring algorithm from Section B.3 to solve this equation. To simplify the notation we denote the current estimate $\widehat{\boldsymbol{\beta}}^{(r)}$ by $\widehat{\boldsymbol{\beta}}$ and the new estimate $\widehat{\boldsymbol{\beta}}^{(r+1)}$ by $\widehat{\boldsymbol{\beta}}^*$. Then the recursion equation of the Fisher scoring algorithm (see (B.8) becomes

$$\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} + \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\beta}})\frac{d \ln L(\widehat{\boldsymbol{\beta}})}{d\boldsymbol{\beta}},$$

where $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})$ is the expected information matrix and $d \ln L(\widehat{\boldsymbol{\beta}})/d\boldsymbol{\beta}$ is the derivative $d \ln L(\boldsymbol{\beta})/d\boldsymbol{\beta}$ both evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Multiplying both sides of the above equation by $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})$ we get

$$\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}^* = \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} + \frac{d \ln L(\widehat{\boldsymbol{\beta}})}{d\boldsymbol{\beta}}. \tag{9.20}$$

To derive an expression for the (j,k)th element of $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})$ we compute

$$E\left[-\left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right)\right] = E\sum_{i=1}^{n} -\left[(y_i - \mu_i)x_{ij}\frac{\partial\{w_i(d\eta_i/d\mu_i)\}}{\partial \beta_k} - w_i x_{ij}\frac{\partial(y_i - \mu_i)}{\partial \beta_k}\frac{d\eta_i}{d\mu_i}\right].$$

The first term has expectation zero since $E(y_i - \mu_i) = 0$. The second term equals

$$\sum_{i=1}^{n} w_i x_{ij}\frac{d\eta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n} w_i x_{ij}\frac{\partial \eta_i}{\partial \beta_k} = \sum_{i=1}^{n} w_i x_{ij} x_{ik}.$$

Thus

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \text{ and } \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}.$$

where $\widehat{\boldsymbol{W}}$ is the estimated diagonal weight matrix evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

The jth element of the R.H.S. of (9.20) equals

$$\begin{aligned}\left(\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}\right)_j + \left(\frac{d\ln L(\widehat{\boldsymbol{\beta}})}{d\boldsymbol{\beta}}\right)_j &= \sum_{k=0}^{p}\boldsymbol{\mathcal{I}}_{jk}(\widehat{\boldsymbol{\beta}})\widehat{\beta}_k + \sum_{i=1}^{n}\widehat{w}_i x_{ij}(y_i-\widehat{\mu}_i)\frac{d\widehat{\eta}_i}{d\mu_i} \\ &= \sum_{k=0}^{p}\sum_{i=1}^{n}\widehat{w}_i x_{ij}x_{ik}\widehat{\beta}_k + \sum_{i=1}^{n}\widehat{w}_i x_{ij}x_{ij}(y_i-\widehat{\mu}_i)\frac{d\widehat{\eta}_i}{d\mu_i} \\ &= \sum_{i=1}^{n}\widehat{w}_i x_{ij}\sum_{k=0}^{p}x_{ik}\widehat{\beta}_k + \sum_{i=1}^{n}\widehat{w}_i x_{ij}x_{ij}(y_i-\widehat{\mu}_i)\frac{d\widehat{\eta}_i}{d\mu_i}\end{aligned}$$

Now note that

$$\sum_{k=0}^{p}x_{ik}\widehat{\beta}_k = \widehat{\eta}_i.$$

Substituting in the above and combining the first and second terms we get the above equal to

$$\sum_{i=1}^{n}\widehat{w}_i x_{ij}\left\{\widehat{\eta}_i + (y_i-\widehat{\mu}_i)\frac{d\widehat{\eta}_i}{d\mu_i}\right\} = \sum_{i=1}^{n}\widehat{w}_i x_{ij}\widehat{z}_i,$$

where $\widehat{z}_i$ is given by (9.11). In vector notation the R.H.S. of (9.20) equals $\boldsymbol{X}\widehat{\boldsymbol{W}}\boldsymbol{z}$. Thus we get the final equation (9.10) for the IRWLS algorithm.

EXERCISES

Theoretical Exercises

9.1 (Negative binomial distribution) The negative binomial distribution is the probability distribution of the number of trials y needed to get a fixed number $r \geq 1$ of successes in a sequence of i.i.d. Bernoulli trials, each with success probability p. It is given by

$$f(y;r,p) = \binom{y-1}{r-1}p^r(1-p)^{y-r},$$

where p is the unknown parameter of interest and r is specified. For $r = 1$, it reduces to the geometric distribution.

a) Express the negative binomial distribution in the exponential family form. Give the functions $a(\phi), b(\theta)$ and $c(y,\phi)$.
b) Find the mean μ and variance σ^2 of this distribution.
c) Show that the distribution is overdispersed ($\sigma^2 > \mu$) if $p < 1/2$ and underdispersed ($\sigma^2 < \mu$) if $p > 1/2$.

9.2 (MLE score equations for exponential and Poisson distributions) Derive the MLE score equations for estimating the regression parameter vector $\boldsymbol{\beta}$ if the response variable distribution is exponential or Poisson, and show that they have the form (9.6).

Applied Exercises

9.3 (Airline injury incidents) The file `Airline-Injury.csv` contains data from Chatterjee and Hadi (2012) on the number of injury incidents and the proportion of the total number of flights out of New York for 9 airlines.

Fit three different models to predict the number of injury incidents (y) as a function of the proportion of the total number of flights (x): (1) simple linear regression without any

transformation of y, (2) simple linear regression with square-root transformation of y and (3) Poisson regression. Calculate the SSE for each model. Which model do you prefer and why?

9.4 (Automobile traffic accidents) The file `crashdata2014.csv` contains data (provided by Professor Hani Mahmassani, Director of the Transportation Center, Northwestern University) on 77 variables for 82,744 automobile crashes in Illinois towns in 2014. The variables include location of the accident (county and township), day and time of the accident, driving conditions, presence of traffic control device, weather conditions, etc. The descriptions of these variables are given in the Word document `Illinois Crash Data Variable Description.docx`. Most of these variables are not relevant as predictor variables for predicting the number of accidents. Furthermore, each variable has many categories, for example, there are 102 counties, time of the day has 24 categories (hours) and so on. An R code was written using the library `dplyr` to create new categorical variables from selected variables as defined below. In addition, a new variable called Weight was created which takes a value of 5 for Weekday and 2 for Weekend, as a measure of the exposure time, to take into account that the number of accidents are likely to be proportional to the number of days in each category, keeping all other conditions fixed. It can be used to perform weighted Poisson regression. Finally these variables were grouped to create a Count variable, which is the number of cases (accidents) in each group. The resulting data are saved in a data file `crashdata2014-summary.csv`. The marginal counts for each new variable are as follows (the numbers are the counts of crashes in each category).

Variable	Counts
Day	Weekday: 59,949; Weekend: 22,795
Time	Night: 16,865; Morning: 18,892; Midday: 23,399; Evening: 23,588
TrafficControl	No Control: 41,651; Control: 38,072; Unknown: 3021
Road	Dry: 58,089; Wet: 18,676; Other: 5979
Light	Daylight: 52,178; Dawn/Dusk: 3076, Dark: 23,538; Unknown: 3952
Weather	Clear: 64,058; Rain/Snow: 12,117; Poor Visibility: 1753, Other: 4816

Note that there are a total of $2 \times 4 \times 3 \times 3 \times 4 \times 4 = 1152$ grouping combinations.

a) Do unweighted Poisson regression with Count as the response variable. Comment on the significance of each variable and whether the sign of each coefficient is as expected.

b) Do weighted Poisson regression with Count as the response variable. Check if the results change much.

c) (Optional) Write an R code that creates the `crashdata2014-summary.csv` data file from the raw data file `crashdata2014.csv`.

9.5 (Auto insurance claims) The data file `claims.csv` contains data from a Canadian insurance company for policy years 1956 and 1957. The variables defined in Table 9.4 describe the different categories of insured drivers, the premiums paid by them, the number of claims they filed and the cost to the company of paying the claims. The response variable is Cost.

Table 9.3 New Categorical Variables Created from Existing Variables from Raw Data for Exercise 9.4

Current Variable (Values)	New Variable	Recoded Values
DayofWeekCode (1-7)	Day	Weekday (1-5), Weekend (6-7)
CrashHour (1-24)	Time	Morning (6-10), Midday (11-15), Evening (16-20), Night 21-5)
TrafficControlDeviceCode (1-14,99)	Traffic-Control	No control (1), Control (2-14), Unknown (99)
RoadSurfaceConditionCode (1-6, 9)	Road	Dry (1), Wet (2,3,4), Other (5,6,9)
LightConditionCode (1-5, 9)	Light	Daylight (1), Dawn/Dusk (2,3), Dark (4,5), Unknown (9)
WeatherCode (1-9)	Weather	Clear (1), Rain/Snow (2,3,5), Poor Visibility (4,8), Other (7,9)

Table 9.4 Description of the Variables in the Automobile Insurance Claims Data

Variable	Description
Merit	3 licensed and accident free ≥ 3 years
	2 licensed and accident free 2 years
	1 licensed and accident free 1 year
	0 all others
Class	1 pleasure, no male operator < 25
	2 pleasure, non-principal male operator < 25
	3 business use
	4 unmarried owner or principal operator < 25
	5 married owner or principal operator < 25
Insured	Earned car years
Premium	Earned premium in 1000s (adjusted to what the premium would have been had all cars been written at 01 rates)
Claims	Number of claims
Cost	Total cost of the claim in 1000s of dollars

Source: `https://www.statistics.ma.tum.de/fileadmin/w00bdb/www/czado/lec8.pdf`

a) Fit a gamma regression model relating Cost of insurance to the attributes of the drivers, the premiums paid out and the claims filed by them. Which variables are significant? What is the estimate of the dispersion parameter?
b) Perform the best subsets regression to find the model with the minimum AIC.

CHAPTER 10

SURVIVAL ANALYSIS

Origins of survival analysis lie in actuarial science of lifetimes and death rates of people. As a discipline, survival analysis has grown from many applications in biomedical fields. However, its methods are applicable in any area where the variable of interest is time to an event. In actuarial and biomedical applications the event of interest is typically death or recurrence of an adverse outcome such as tumor or hospitalization. In engineering reliability studies the event of interest is failure of an item. In marketing applications the event of interest is the next purchase by a customer. Generally, the goal is to model the lifetime as a function of some predictor variables, e.g., prognostic variables of a patient or load conditions in a reliability study or demographic and socioeconomic attributes and past purchase history of a customer in a marketing study. For convenience, we will use the biomedical terminology of death of a patient as the event of interest.

A unique feature of the time to an event data is that often they are censored; in other words, the event is not observed within the time-frame of the study. Thus all we know is that the lifetime is greater than the censoring time. For example, the death or failure may not occur before the study is terminated or a patient may withdraw or drop out of the study before the event of interest occurs or a customer may not place a new order before data collection stops. This is called **right censoring**.

In some applications the data are **left censored** as happens when the start of the lifetime is not observed. For example, in the study of incubation time from HIV infection to onset of AIDS, some patients may enter the study already infected by HIV virus. So the start of the incubation time is unknown. We will not consider left censoring in this chapter.

The challenge in analyzing censored data is that the information is incomplete. Methods discussed in this chapter such as lifetables, Kaplan-Meier survival curves and Cox's proportional hazards regression model take into account censoring of the lifetimes. These methods are different from other methods discussed in this book in that the first two methods are nonparametric while the third method is semiparametric, i.e., one part of the model is unspecified and so is nonparametric while the other part, which is similar to multiple regression, is parametric.

10.1 Hazard Rate and Survival Distribution

Suppose that the lifetime is a continuous random variable (r.v.) denoted by T (here we have deviated from the notational convention followed elsewhere in the book by denoting a r.v. by an upper case letter and its observed value by the corresponding lower case letter). Let $f(t)$ denote the p.d.f. and $F(t) = P(T \leq t)$ denote the c.d.f. of T. Then $S(t) = 1 - F(t) = P(T > t)$ is called the **survival distribution** of T. The **hazard rate**, denoted by $\lambda(t)$, is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \left[\frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \right].$$

In words, $\lambda(t)$ is the instantaneous **failure rate** at time t given that the patient survives until time t. An explicit expression for $\lambda(t)$ can be obtained as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \left[\frac{1}{\Delta t} \frac{P(t < T \leq t + \Delta t)}{P(T > t)} \right] = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\frac{f(t)\Delta t}{1 - F(t)} \right] = \frac{f(t)}{S(t)}. \quad (10.1)$$

Note that

$$\lambda(t) = -\frac{d}{dt}(\ln(S(t))).$$

Hence

$$S(t) = \exp\left(-\int_0^t \lambda(u) du \right) = \exp(-\Lambda(t)), \quad (10.2)$$

where $\Lambda(t)$ is called the **cumulative hazard**.

EXAMPLE 10.1 (Hazard Rate for the Exponential Distribution)

The exponential distribution is the simplest continuous lifetime distribution. Its p.d.f. equals $f(t) = \lambda \exp(-\lambda t)$ and its c.d.f. equals $F(t) = 1 - \exp(-\lambda t)$; thus its survival distribution equals $S(t) = \exp(-\lambda t)$. Hence its hazard rate equals

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda,$$

which is constant with respect to time.

The **memoryless property** of the exponential distribution derives from its constant hazard rate. This property says that the conditional probability that a patient having survived until time t, will survive for another u time units is independent of t. This can be checked as follows:

$$P(T > t + u | T > t) = \frac{P(T > t + u)}{P(T > t)} = \frac{e^{-\lambda(t+u)}}{e^{-\lambda t}} = e^{-\lambda u},$$

which is independent of t. ■

The constant hazard rate for the exponential distribution limits its use in practice. Many real life phenomena exhibit increasing hazard rates, e.g., machine parts become more failure-prone as they age due to wear and tear; the same is true of living beings. The

Weibull distribution (see Exercise 10.2) generalizes the exponential distribution and allows modeling of increasing or decreasing hazard rates. The gamma distribution introduced in Chapter 8 is another generalization of the exponential distribution.

10.2 Kaplan-Meier Estimator

The **Kaplan-Meier (KM) estimator** (also known as the **product-limit (PL) estimator**) gives a nonparametric estimate of the survival distribution. The idea of the estimator is very simple and yet powerful. Suppose we begin with n patients some of whom die or are censored as the study progresses. Let $m \leq n$ be such event times. Suppose that the event times are ordered so that $t_1 < t_2 < \cdots < t_m$ some of which may be censored. If the data are binned then the events (whether death or censored) are assumed to occur at the end of each time interval. Let n_i be the number of patients at risk just before time t_i (i.e., those who are still alive). Suppose c_i of them are censored and d_i of them die at time t_i. Then the number of patients at risk just before time t_{i+1} equal $n_{i+1} = n_i - c_i - d_i$ $(1 \leq i \leq m-1)$ where $n_1 = n$. The estimated hazard rate at time t_i is

$$\widehat{\lambda}(t_i) = \frac{d_i}{n_i}.$$

To find the KM estimator of the survival function, note that $P(T > t_j) = P(T > t_{j-1})P(T \neq t_j)$ or equivalently $S(t_j) = S(t_{j-1})[1 - \lambda(t_j)]$ for $j \geq 1$. Applying this formula recursively we get

$$S(t_j) = \prod_{i=1}^{j}[1 - \lambda(t_i)].$$

Now $\lambda(t_i)$ can be estimated by $\widehat{\lambda}(t_i)$ given above. Further noting that $\widehat{S}(t)$ is constant for $t_i \leq t < t_{i+1}$, leads to the following KM estimator of the survival function:

$$\widehat{S}(t) = \prod_{t_i \leq t}[1 - \widehat{\lambda}(t_i)] = \prod_{t_i \leq t}\left[1 - \frac{d_i}{n_i}\right]. \tag{10.3}$$

These calculations are presented in the form of a **lifetable**. KM survival curves can be plotted from lifetable calculations as illustrated in the example below.

If the death times are not tied, i.e., if there is at most one death at any given time, then it is easy to see that the KM estimator for $t_i \leq t < t_{i+1}$ reduces to

$$\widehat{S}(t) = \prod_{j \leq i}\left(1 - \frac{1}{n-j+1}\right)^{\delta_j} = \prod_{j \leq i}\left(\frac{n-j}{n-j+1}\right)^{\delta_j},$$

where δ_j is an indicator variable for censoring with $\delta_j = 0$ if t_j is censored and $\delta_j = 1$ if t_j is not censored.

EXAMPLE 10.2 (AML Data: Kaplan-Meier Survival Curves)

Tableman and Kim (2003) have given data on the times to relapse for acute myelogenous leukemia (AML) patients in a clinical trial under two treatment arms: extended treatment ("Maintained") and non-extended treatment ("Nonmaintained"). The data for the 11 Maintained group of patients are 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, and 161+ weeks, where the + sign indicates a censored observation. The data for the 12 Nonmaintained group of patients are 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45.

Table 10.1 Calculation of Kaplan-Meier survival curves for AML data (+ indicates a censored observation)

Maintained							Nonmaintained						
t_i	n_i	c_i	d_i	$\widehat{\lambda}(t_i)$	$1-\widehat{\lambda}(t_i)$	$\widehat{S}(t_i)$	t_i	n_i	c_i	d_i	$\widehat{\lambda}(t_i)$	$1-\widehat{\lambda}(t_i)$	$\widehat{S}(t_i)$
0	11	0	0	0	0.000	1.000	0	12	0	0	0.000	1.000	1.000
9	11	0	1	0.091	0.909	0.909	5	12	0	2	0.167	0.833	0.833
13+	10	1	1	0.100	0.900	0.818	8	10	0	2	0.200	0.800	0.667
18	8	0	1	0.125	0.875	0.716	12	8	0	1	0.125	0.875	0.583
23	7	0	1	0.143	0.857	0.614	16+	7	1	0	0.000	1.000	0.583
28+	6	1	0	0.000	1.000	0.614	23	6	0	1	0.167	0.833	0.486
31	5	0	1	0.200	0.800	0.491	27	5	0	1	0.200	0.800	0.389
34	4	0	1	0.250	0.750	0.368	30	4	0	1	0.250	0.750	0.292
45+	3	1	0	0.000	1.000	0.368	33	3	0	1	0.333	0.667	0.194
48	2	0	1	0.500	0.500	0.184	43	2	0	1	0.500	0.500	0.097
161+	1	1	0	0.000	1.000	0.184	45	1	0	1	1.000	0.000	0.000

Note that the estimated survival function remains constant from one uncensored event to the next regardless of any censored events in between (e.g., $\widehat{S}(t) = 0.614$ from $t = 23$ to $t = 31^-$ in the Maintained treatment group regardless of the censored time $t = 28$ in between). This is clear from the fact that at a censored time, $\widehat{\lambda}(t_i) = 0$ and so $1 - \widehat{\lambda}(t_i) = 1$.

The survival curves for the two treatment arms are shown in Figure 10.1. We see that the survival curve for the Maintained treatment lies uniformly above that for the Nonmaintained treatment thus showing that the Maintained treatment improves survival, i.e., lengthens time to relapse. These survival curves are obtained using the following R code:

```
> library(survival)
> fit<- survfit(Surv(time,status) ~ x,data=aml)
> plot(fit,col=1:2)
> legend("topright", paste(" ",c("Maintained","Nonmaintained")),
  col=1:2, lty=c(1 ,2))
```

■

Since $\widehat{S}(t)$ is an estimate of the true survival function $S(t)$ at some fixed time t, we can assess its accuracy through its variance. The **Greenwood formula** gives an estimate of the variance of $\widehat{S}(t)$:

$$\widehat{\text{Var}}(\widehat{S}(t)) = [\widehat{S}(t)]^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{10.4}$$

and $\text{SE}[\widehat{S}(t)] = [\widehat{\text{Var}}(\widehat{S}(t))]^{1/2}$. The following example illustrates this calculation.

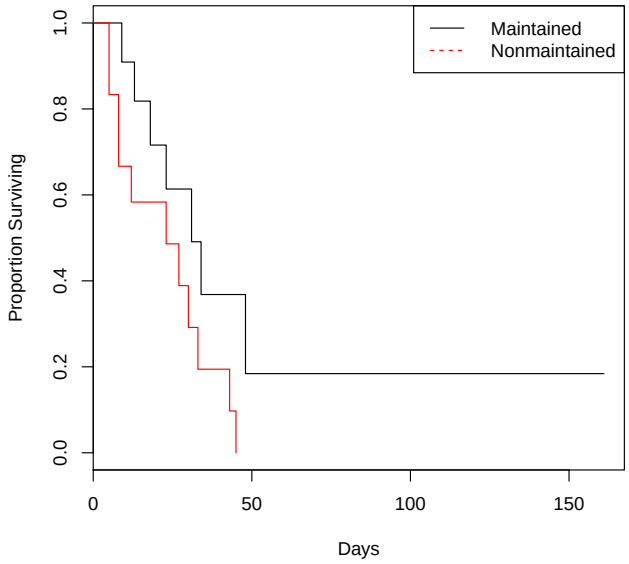


Figure 10.1 Kaplan-Meier survival curves for Maintained and Nonmaintained treatment arms

EXAMPLE 10.3 (AML Data: Standard Error Calculation for Kaplan-Meier Survival Curves)

For illustration purposes, consider $\widehat{S}(t) = 0.716$ for $18 \leq t < 23$ for the Maintained group calculated in Table 10.1. Then $t_1 = 9, t_2 = 13, t_3 = 18$ are $\leq t$ (any censored time $t_i \leq t$ with $d_i = 0$ can be omitted from the variance calculation). Thus for $18 \leq t < 23$ we have

$$\widehat{\text{Var}}(\widehat{S}(t)) = (0.716)^2 \left[\frac{1}{11 \times 10} + \frac{1}{10 \times 9} + \frac{1}{8 \times 7} \right] = 0.0195.$$

Hence $\text{SE}(\widehat{S}(t)) = \sqrt{0.0195} = 0.1397$. The standard errors for all values of $\widehat{S}(t)$ for both treatment arms calculated using R are given in Table 10.2. These standard errors can be used to draw pointwise confidence intervals around the estimated survival function. For example, for $18 \leq t < 23$ for the Maintained treatment, the 95% confidence interval is

$$0.716 \pm 1.96 \times 0.1397 = [0.442, 0.990].$$

■

10.3 Log Rank Test

We saw in Figure 10.1 that the survival distribution for the Maintained group lies uniformly above that of the Nonmaintained group, but is the difference statistically significant? Denoting the two survival distributions by $S_1(t)$ and $S_2(t)$, we would like to test the null hypothesis $H_0 : S_1(t) = S_2(t)$ for all t.

Table 10.2 Calculation of standard errors for Kaplan-Meier survival curves for AML data

Maintained			Nonmaintained		
t_i	$\widehat{S}_1(t_i)$	SE$(\widehat{S}_1(t_i))$	t_i	$\widehat{S}_2(t_i)$	SE$(\widehat{S}_2(t_i))$
0	1.000	0.0000	0	1.000	0.000
9	0.909	0.0867	5	0.833	0.1076
13+	0.818	0.1163	8	0.667	0.1361
18	0.716	0.1397	12	0.583	0.1423
23	0.614	0.1526	16+	0.583	0.1423
28+	0.614	0.1526	23	0.486	0.1481
31	0.491	0.1642	27	0.389	0.1470
34	0.368	0.1627	30	0.292	0.1387
45+	0.368	0.1627	33	0.194	0.1219
48	0.184	0.1535	43	0.097	0.0919
161+	0.184	0.1535	45	0.000	N/A

Table 10.3 Number of deaths, survivals and at risk patients at time t_i

		Death		
		Yes	No	
Treatment	Maintained	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
	Nonmaintained	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
		d_i	$n_i - d_i$	n_i

We can test the difference between $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ for some fixed t as follows. Say $t = 32$. Then $\widehat{S}_1(t) = 0.491$ and $\widehat{S}_2(t) = 0.292$; the respective standard errors are 0.1642 and 0.1387. Since $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ are independent, the standardized z-statistic is

$$z = \frac{\widehat{S}_1(t) - \widehat{S}_2(t)}{\sqrt{\widehat{\text{Var}}(\widehat{S}_1(t) - \widehat{S}_2(t))}} = \frac{0.491 - 0.292}{\sqrt{0.1642^2 + 0.1387^2}} = 0.926.$$

So the difference is not statistically significant.

Log rank test provides a test of the *overall* difference between the two survival distributions by cumulating the differences between $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ at the observed death times. Let $t_1 < t_2 < \cdots < t_m$ denote the observed death times in the combined data set of the two treatments. The deaths may occur in one treatment arm or both. For each time point t_i let n_{1i} and n_{2i} denote the numbers at risk and d_{1i} and d_{2i} denote the numbers of deaths from the two treatment arms with $n_i = n_{1i} + n_{2i}$ being the total number at risk at time t_i. This data can be summarized in a 2×2 table shown in Table 10.3.

It is well-known that if we fix the row and column margins (i.e., n_{1i}, n_{2i}, d_i and $n_i - d_i$) then d_{1i} determines the remaining entries in the table and the distribution of d_{1i} under H_0 is hypergeometric given by

$$f(d_{1i}) = \frac{\binom{n_{1i}}{d_{1i}}\binom{n_{2i}}{d_{2i}}}{\binom{n_i}{d_i}}.$$

This distribution can be used in **Fisher's exact test** to compare the two groups at some chosen time t_i as illustrated in the example below. The mean and variance of d_{1i} are given by

$$E(d_{1i}) = \frac{d_i n_{1i}}{n_i} \quad \text{and} \quad \text{Var}(d_{1i}) = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}.$$

EXAMPLE 10.4 (AML Data: Fisher's Exact Test)

Consider the AML data shown in Table 10.1 at $t_1 = 5$ days. Suppose we want to compare the Maintained group with the Non-Maintained group at this time. We have $d_{11} = 0, n_{11} = 11$ for the Maintained group and $d_{21} = 2, n_{21} = 12$ for the Non-Maintained group. The hypergeometric probability of this outcome, conditioned on the number of deaths $d_1 = d_{11} + d_{21} = 2$ is

$$\frac{\binom{11}{0}\binom{12}{2}}{\binom{23}{2}} = \frac{6}{23} = 0.261.$$

The P-value of Fisher's exact test for comparing the Maintained group versus the Non-Maintained group is the cumulative tail probability of all the outcomes at least as extreme as the observed outcome. In this case, the extreme outcomes are those with $d_{11} \leq 0$, so the observed outcome with $d_{11} = 0$ is the only outcome in this set and hence $P = 0.261$, which is not significant.

Note that

$$E(d_{11}) = \frac{2 \times 11}{23} = 0.957, E(d_{21}) = \frac{2 \times 12}{23} = 1.043$$

and

$$\text{Var}(d_{11}) = \text{Var}(d_{21}) = \frac{11 \times 12 \times 2 \times 21}{23^2 \times 22} = 0.522.$$

The sample sizes are too small to apply the normal approximation to d_{11} and perform the z-test, but just for illustration purposes we calculate

$$z = \frac{d_{11} - -E(d_{11})}{\sqrt{\text{Var}(d_{11})}} = \frac{0 - 0.957}{\sqrt{0.522}} = -1.325,$$

which has $P = 0.093$. Thus the normal approximation is not close to the exact $P = 0.261$. ■

The logrank statistic is obtained as the standardized sum of the deviations of the d_{1i} from $E(d_{1i})$ summed over all time points:

$$z = \frac{\sum_{i=1}^m d_{1i} - \sum_{i=1}^m E(d_{1i})}{\left(\sum_{i=1}^m \text{Var}(d_{1i})\right)^{1/2}}. \tag{10.5}$$

Asymptotically (for large n_{1i}, n_{2i}), this statistic has a standard normal distribution under H_0.

EXAMPLE 10.5 (AML Data: Log Rank Test)

For the AML data shown in Table 10.1, the number of distinct observed event times are $m = 15$. Table 10.5 shows the calculation of the logrank statistic.

Time	d_{1i}	d_{2i}	n_{1i}	n_{2i}	d_i	n_i	$E(d_{1i})$	Var(d_{1i})
5	0	2	11	12	2	23	0.9565	0.4764
8	0	2	11	10	2	21	1.0476	0.4739
9	1	0	11	8	1	19	0.5789	0.2438
12	0	1	10	8	1	18	0.5556	0.2469
13	1	0	10	7	1	17	0.5882	0.2422
18	1	0	8	6	1	14	0.5714	0.2449
23	1	1	7	6	2	13	1.0769	0.4556
27	0	1	6	5	1	11	0.5454	0.2479
30	0	1	5	4	1	9	0.5556	0.2469
31	1	0	5	3	1	8	0.6250	0.2344
33	0	1	4	3	1	7	0.5714	0.2449
34	1	0	4	2	1	6	0.6667	0.2222
43	0	1	3	2	1	5	0.6000	0.2400
45	0	1	3	1	1	4	0.7500	0.1875
48	1	0	2	0	1	2	1	0

We calculate $\sum_{i=1}^{15} d_{1i} = 7, \sum_{i=1}^{15} E(d_{1i}) = 10.6893$ and $\sum_{i=1}^{15}$ Var$(d_{1i}) = 4.0076$. Hence the logrank statistic equals $z = (7 - 10.6893)/\sqrt{4.0076} = -1.8429$ with a one-sided P-value $= 0.0326$. So we can reject $H_0 : S_1(t) = S_2(t)$ for all t at $\alpha = 0.05$ and conclude that the Maintained group experiences significantly longer survival times overall than the Non-Maintained group. The R code for performing the logrank test and its output are shown below. Note that R reports a two-sided test with $\chi^2 = (-1.8429)^2 = 3.40$ with a two-sided P-value $= 0.0653$, which is twice that of the one-sided P-value of 0.0326.

```
> library(survival)
> survdiff(Surv(time,status) ~ x,data=aml)
Call:
survdiff(formula = Surv(time, status) ~ x, data = aml)

                N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27       3.4
x=Nonmaintained 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

■

10.4 Cox's Proportional Hazards Model

Cox (1972) proposed a novel regression model for censored lifetime as a function of a set of covariates. This model postulates how the hazard rate depends on the covariates. Let $\lambda(t) = \lambda(t|\boldsymbol{x})$ denote the hazard rate at time t given the covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ denote the corresponding unknown regression coefficient vector. Then Cox's model is

$$\lambda(t) = \lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p) = \lambda_0(t) \exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{10.6}$$

where $\lambda_0(t)$, called the **base hazard rate**, is independent of $\boldsymbol{x}$ and is assumed to be completely unspecified. Thus the model has a nonparametric component $\lambda_0(t)$ and a parametric component $\exp(\boldsymbol{x}'\boldsymbol{\beta})$. Hence it is called a **semiparametric model**. Note that this model does not have an intercept term β_0 since $\exp(\beta_0)$ can be absorbed in $\lambda_0(t)$.

This model has an interesting property that if there are two individuals, i and j, with covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ then the ratio of their hazard rates (called the **hazard ratio**) at any time t:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\lambda_0(t) \exp(\boldsymbol{x}_j'\boldsymbol{\beta})} = \exp((\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta})$$

is independent of t. Hence this is called the **proportional hazards (PH) model**. This property is similar to the proportional odds property of the ordinal logistic regression model mentioned in Section 7.6.2.

A special case of interest is when the only covariate is a single treatment factor with two levels, say treatment ($x_1 = 1$) and control ($x_1 = 0$). Then the hazard ratio is

$$\frac{\lambda(t|x_1 = 1)}{\lambda(t|x_1 = 0)} = \frac{\lambda_0(t) \exp(\beta_1)}{\lambda_0(t)} = \exp(\beta_1),$$

which is constant. If $\beta_1 < 0$ then the hazard rate for the treatment is less than that for the control. Thus negative β_1 represents an effective treatment since it reduces the hazard rate. More generally, $\exp(\beta_j)$ represents the ratio of hazard rates if x_j is increased by one unit, keeping all other covariates fixed.

10.4.1 Estimation

Suppose we have data on the lifetimes t_i, covariate vectors $\boldsymbol{x}_i$ and censoring indicators δ_i ($\delta_i = 0$ if t_i is censored and $\delta_i = 1$ if t_i is not censored) for $i = 1, \ldots, n$. Denote by $C = \{i : \delta_i = 0\}$ the set of censored individuals and by $D = \{i : \delta_i = 1\}$ the set of died individuals. Assume that $t_1 < t_2 < \ldots < t_n$ and thus there are no ties among the lifetimes. For convenience, suppose that the individuals are labeled so that the ith individual is associated with the observation at time t_i. Since $\lambda_0(t)$ is arbitrary and unspecified, it is not possible to find the MLE of $\boldsymbol{\beta}$ by maximizing the full likelihood. Cox (1972) suggested to use the likelihood conditioned on the observed lifetimes of individuals.

Consider the event time t_i at which individual $i \in D$ dies (we assume that at most one individual may die at each time t_i). Let $R(t_i)$ denote the **risk set** of all individuals who are alive and still under observation at time t_i^- and hence are at risk of dying at time t_i. Note that $R(t_i)$ includes individuals who either die or are censored at time greater than or equal to t_i. Then the conditional probability that out of all the individuals in this risk set, the particular individual i dies at time t_i is given by

$$\frac{\lambda_0(t_i) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \lambda_0(t_i) \exp(\boldsymbol{x}_j'\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j'\boldsymbol{\beta})} = \frac{\psi_i}{\sum_{j \in R(t_i)} \psi_j},$$

where $\psi_j = \exp(\boldsymbol{x}_j'\boldsymbol{\beta})$ is the jth individual's **risk score**. One can view this as an urn model in which one individual is drawn to die out of all the individuals in the risk set $R(t_i)$ and the probability of drawing individual $i \in R(t_i)$ is proportional to ψ_i. In the above, the unknown base hazard function $\lambda_0(t_i)$ cancels from the numerator and denominator in each term and so its unknown form does not affect the MLE of $\boldsymbol{\beta}$.

The so-called **partial likelihood** is obtained by regarding the observed death times as independent resulting in

$$L = \prod_{i \in D} \left[\frac{\psi_i}{\sum_{j \in R(t_i)} \psi_j} \right]. \tag{10.7}$$

The MLE of $\boldsymbol{\beta}$ maximizes L or equivalently $\ln L$.

Note that the survival times t_i enter the partial likelihood L only through the risk sets $R(t_i)$ and hence L is a function only of the ranks of the t_i's. Therefore any monotone transformation, such as the log transformation, of the t_i's does not affect the partial likelihood and hence the MLE of $\boldsymbol{\beta}$.

EXAMPLE 10.6 (Toy Example)

Consider a toy example to illustrate the partial likelihood calculation. Suppose there are three patients, aged 60, 55 and 50, and age (x_1) is the only covariate. The patients are observed at three time points: $t_1 < t_2 < t_3$. Patient 1 dies at t_1, Patient 2 dies at t_2, while Patient 3 survives past t_3 and is censored since the study is terminated at t_3. Then the risk sets are

$$R(t_1) = \{1, 2, 3\}, R(t_2) = \{2, 3\}, R(t_3) = \{3\}.$$

Suppose we want to fit the model

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 x_1).$$

At time t_1, Patient 1 dies out of the risk set $R(t_1)$, at time t_2, Patient 2 dies out of the risk set $R(t_2)$ and at time t_3, no death is observed and so there is no contribution to the partial likelihood, which therefore equals

$$L = \left[\frac{\exp(60\beta_1)}{\exp(60\beta_1) + \exp(55\beta_1) + \exp(50\beta_1)} \right] \left[\frac{\exp(55\beta_1)}{\exp(55\beta_1) + \exp(50\beta_1)} \right].$$

Note again that only the order $t_1 < t_2 < t_3$ matters, their numerical values do not enter into the partial likelihood. ■

The MLE of $\boldsymbol{\beta}$ can be found by maximizing the log-likelihood function:

$$\ln L = \sum_{i \in D} \ln \psi_i - \sum_{i \in D} \ln \sum_{j \in R(t_i)} \psi_j = \sum_{i \in D} \boldsymbol{x}_i'\boldsymbol{\beta} - \sum_{i \in D} \ln \sum_{j \in R(t_i)} \psi_j.$$

Now,

$$\frac{d(\boldsymbol{x}_i'\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \boldsymbol{x}_i,$$

and

$$\frac{d\psi_j}{d\boldsymbol{\beta}} = \frac{d(\exp(\boldsymbol{x}_j'\boldsymbol{\beta}))}{d\boldsymbol{\beta}} = \exp(\boldsymbol{x}_j'\boldsymbol{\beta}) \frac{d(\boldsymbol{x}_j'\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \psi_j \boldsymbol{x}_j.$$

Hence

$$\frac{d}{d\boldsymbol{\beta}} \sum_{i \in D} \ln \sum_{j \in R(t_i)} \psi_j = \sum_{i \in D} \frac{\frac{d}{d\boldsymbol{\beta}} (\sum_{j \in R(t_i)} \psi_j)}{\sum_{j \in R(t_i)} \psi_j} = \sum_{i \in D} \frac{\sum_{j \in R(t_i)} \psi_j \boldsymbol{x}_j}{\sum_{j \in R(t_i)} \psi_j}.$$

Therefore the MLE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the solution to the equation

$$\frac{d \ln L}{d\boldsymbol{\beta}} = \sum_{i \in D} \boldsymbol{x}_j - \sum_{i \in D} \frac{\sum_{j \in R(t_i)} \psi_j \boldsymbol{x}_j}{\sum_{j \in R(t_i)} \psi_j} = 0.$$

Asymptotically, $\widehat{\boldsymbol{\beta}}$ can be shown to be fully efficient and normally distributed, and its asymptotic covariance matrix can be obtained by inverting the Hessian matrix of second partial derivatives of $\ln L$ in the usual manner.

10.4.2 Examples

In the first example we apply the PH model to the AML data and perform a two sample test of Maintained versus Nonmaintained groups. The only covariate here is the group membership variable indicating whether the individual belongs to the Maintained group ($x = 0$) or to the Nonmaintained group ($x = 1$).

EXAMPLE 10.7 (AML Data: Proportional Hazards Model)

The R code used to perform this analysis and the resulting output are shown below.

```
> library(survival)
> fit1 <- coxph(Surv(time, status) ~ x, data=aml)
> summary(fit1)
Call:
coxph(formula = Surv(time, status) ~ x, data = aml)

  n= 23, number of events= 18

                 coef exp(coef) se(coef)     z Pr(>|z|)
xNonmaintained 0.9155    2.4981   0.5119 1.788   0.0737 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

               exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained     2.498     0.4003    0.9159     6.813

Concordance= 0.619  (se = 0.073 )
Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.06581
Wald test            = 3.2  on 1 df,   p=0.07371
Score (logrank) test = 3.42  on 1 df,   p=0.06454
```

We see that the fitted model is $\lambda(t) = \lambda_0(t)\exp(0.9155x)$. So the hazard for the Nonmaintained group is $\exp(0.9155) = 2.498$ or almost 2.5 times that of the Maintained group. The results of the likelihood ratio (LR) test, Wald test and logrank test, agree closely. The Wald statistic equals $[\widehat{\beta}/\text{SE}(\widehat{\beta})]^2 = [0.9155/0.5119]^2 = 1.788^2 = 3.20$ which is a chi-square statistic with 1 d.f. The P-value is approximately 0.07 and so $\widehat{\beta}_1$ is not significant at $\alpha = 0.05$.

A large sample 95% C.I. on the hazard ratio between the nonmaintained group and the maintained group is obtained by first obtaining a 95% C.I. on β_1 as

$$\widehat{\beta}_1 \pm z_{.025}\text{SE}(\widehat{\beta}_1) = 0.9155 \pm 1.960 \times 0.5119 = [-0.0878, 1.9188].$$

Hence the corresponding 95% C.I. on the hazard ratio equals

$$[\exp(-0.08780), \exp(1.9188)] = [0.9159, 6.8128],$$

which is given in the R output above. Note that these CI's are in agreement with the test result above in that the CI for β_1 includes 0 and correspondingly the CI for the hazard ratio includes 1. ■

The second example involves multiple covariates including a treatment factor with two levels.

EXAMPLE 10.8 (Recidivism Study: Proportional Hazards Model)

Rossi et al. (1980) performed an experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from prison. The goal of the study was to evaluate whether financial aid (the treatment factor) helps these prisoners become productive citizens and not get arrested again. Half the prisoners were randomly assigned to receive financial aid. The following variables were recorded for each prisoner.

1. Week: Week of first arrest after release or censoring time.
2. Arrest: The event indicator (1 if arrested, 0 if not arrested).
3. Aid: Financial aid (yes or no)
4. Age: In years at the time of release.
5. Race: Black or other.
6. Work: Yes if the individual had full-time work experience prior to incarceration, no if he did not.
7. Married: Married or not married.
8. Parole: Released on parole or not.
9. Prior: Number of prior convictions.
10. Education: An ordinal variable with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary).

The data are in file `recid.csv`. The file also contains data on the employment status for each of 52 weeks. This is a time-dependent covariate, which may vary from week to week. In this example we ignore this variable. It is included in the analysis in Exercise 10.11.

First we compare the Aid versus No Aid group by plotting their Kaplan-Meier curves and performing the logrank test. This plot ignores other covariates. The Kaplan-Meier curves are shown in Figure 10.2. We see that the survival curves for the two groups overlap until about 20 weeks but later the survival curve for the Aid group lies above that for the No Aid group. Thus in the first 20 weeks prisoners get arrested at roughly the same rate whether they received financial aid or not. After 20 weeks, prisoners who received financial aid stay out of the jail longer.

The logrank test results are shown below. We see that the difference between the two survival distributions is nearly signficant at the 0.05 level.

```
> recid=read.csv("c:/data/recid.csv")
> fit <- coxph(Surv(Week, Arrest) ~ Aid, data=recid)
```

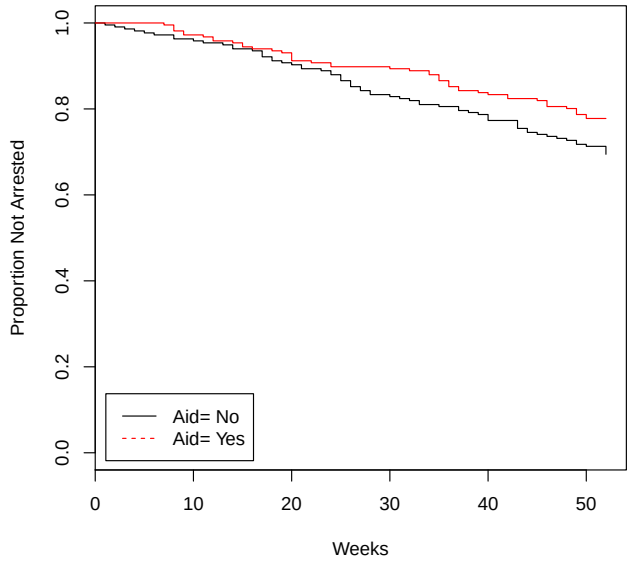


Figure 10.2 Kaplan-Meier survival curves for Aid and No Aid groups for Recidivism data

```
> summary(fit)
Call:
coxph(formula = Surv(Week, Arrest) ~ Aid, data = recid)

  n= 432, number of events= 114

       coef exp(coef) se(coef)      z Pr(>|z|)
Aid -0.3691    0.6914   0.1897 -1.945   0.0517 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    exp(coef) exp(-coef) lower .95 upper .95
Aid    0.6914      1.446    0.4767     1.003

Concordance= 0.546  (se = 0.024 )
Rsquare= 0.009   (max possible= 0.956 )
Likelihood ratio test= 3.84  on 1 df,   p=0.05013
Wald test            = 3.78  on 1 df,   p=0.05174
Score (logrank) test = 3.83  on 1 df,   p=0.05042
```

Next we fit a full PH model with all covariates using the following R code.

```
> library(survival)
> recid=read.csv("c:/data/recid.csv")
> fit1 <- coxph(Surv(Week, Arrest) ~ ., data=recid)
```

```
> summary(fit1)
```

The output is as follows.

```
Call:
coxph(formula = Surv(Week, Arrest) ~ ., data = recid)

  n= 432, number of events= 114

              coef exp(coef) se(coef)      z Pr(>|z|)
Aid       -0.35963   0.69794  0.19180 -1.875  0.06079 .
Age       -0.05768   0.94395  0.02187 -2.638  0.00835 **
Race       0.34554   1.41276  0.30907  1.118  0.26356
Work      -0.11439   0.89191  0.21311 -0.537  0.59145
Married   -0.42496   0.65380  0.38209 -1.112  0.26605
Parole    -0.08991   0.91401  0.19568 -0.459  0.64589
Prior      0.08469   1.08838  0.02919  2.902  0.00371 **
Education -0.18578   0.83046  0.13153 -1.412  0.15782
---

Concordance= 0.656  (se = 0.027 )
Rsquare= 0.079   (max possible= 0.956 )
Likelihood ratio test= 35.35  on 8 df,   p=2.31e-05
Wald test            = 33.74  on 8 df,   p=4.529e-05
Score (logrank) test = 35.1  on 8 df,   p=2.568e-05
```

Note that the effect of Aid is now less significant ($P = 0.0608$). Of the remaining covariates only Age and Prior are significant. Including only these three variables another PH model is fitted whose results are shown below.

```
Call:
coxph(formula = Surv(Week, Arrest) ~ Aid + Age + Prior, data = recid)

  n= 432, number of events= 114

          coef exp(coef) se(coef)      z Pr(>|z|)
Aid   -0.34695   0.70684  0.19025 -1.824 0.068197 .
Age   -0.06711   0.93510  0.02085 -3.218 0.001289 **
Prior  0.09689   1.10174  0.02725  3.555 0.000378 ***
---

Concordance= 0.63  (se = 0.027 )
Rsquare= 0.065   (max possible= 0.956 )
Likelihood ratio test= 29.05  on 3 df,   p=2.189e-06
Wald test            = 27.94  on 3 df,   p=3.741e-06
Score (logrank) test = 29.03  on 3 df,   p=2.203e-06
```

Now Age and Prior are much more significant but Aid is even less significant ($P = 0.0682$). Our overall conclusion is that Aid has a positive effect on the recidivism of the prisoners but not very significant. ■

10.4.3 Time-Dependent Covariates

In the basic Cox model (10.6) we have assumed that the covariates are fixed. But in practice many covariates vary with time. For example, many lab measurements such as cholesterol level or hemoglobin A1c vary with time. We don't consider the age of a patient as a time-dependent covariate since it increases at the same rate for every patient, so its value at the baseline is used as a fixed covariate. If the time-dependent nature of covariates is ignored in the analysis then misleading conclusions may result as shown in Example 10.9. More generally, the β coefficients in the Cox model can be functions of time, but we do not discuss this extension of the model.

Denote a time-dependent covariate x_j as $x_j(t)$. By letting $x_j'(t) = x_j(t-s)$ we can model lagged effect of a variable where $s \geq 1$ is a specified lag. More generally, denote the covariate vector as $\boldsymbol{x}(t)$, some of whose components may be functions of time whereas other components may be fixed in time. Then the Cox model may be written as

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{x}'(t)\boldsymbol{\beta}).$$

Note that this model does not have the proportional hazards property since the ratio of hazards for any pair of individuals is no longer fixed but is a function of time. The proportional hazards property of the Cox model only holds if the covariates are not time-dependent. The partial likelihood is still the product of the conditional probabilities at different death times t_i; each conditional probability being the probability that the ith patient dies at time t_i conditioned on all the patients in the risk set $R(t_i)$.

To analyze time-dependent covariate data using the `survival` library, the data need to be set up in what is called the **long format**. Essentially, this format consists of multiple rows of entries for each subject for successive time intervals, $\texttt{tstart} \leq t < \texttt{tstop}$, such that the values of all covariates remain fixed over these time intervals. Note that each time interval is open on the right except the last interval when either death or censoring occurs at time `tstop`. We set `tstart` for any interval (except the first one) equal to `tstop` for the previous interval. Any changes in the data are assumed to take place at `tstart` of the following interval and are included in that interval.

As an example, consider the recidivism data which has a time-varying covariate, employment status (0 if unemployed, 1 if employed) of the prisoner, for each of 52 weeks. The first prisoner was arrested in Week 20 and was not employed throughout those 20 weeks. The second prisoner was employed from Week 9 until Week 14 and was arrested in Week 17. So the data for these two prisoners (ignoring other covariates) can be represented in long format as shown below.

Prisoner	`tstart`	`tstop`	Employment Status	Arrest
1	0	20	0	1
2	0	9	0	0
2	9	14	1	0
2	14	17	0	1

EXAMPLE 10.9 (Stanford Heart Transplant Study)

Crowley and Hu (1977) reported a heart transplant study conducted at Stanford University Medical School between October 1, 1967 and April 1, 1974. There were 103 cardiac patients enrolled. Patients had to wait until a suitable donor heart was available. Of the 103 patients, 30 died before receiving a transplant while 4 patients

had still not received a transplant when the study ended. Only 24 of the 69 patients who received a transplant were still alive at termination.

The raw data file `jasa`, which is part of the `survival` library, consists of a number of variables for each patient. Several data manipulations have to be done and certain anomalies in the data have to be fixed in order to obtain the variables in the form ready for analysis. The R code for these data manipulations is taken from a preprint by Therneu, Crowson and Atkinson (2018) and is shown below. The code creates two data files; `tdata.csv` in which the trt variable is regarded as fixed and so there are 103 rows of data (one for each patient) and `sdata.csv` in which the trt variable is regarded as time-dependent and there are 170 rows of data (one or more for each patient) in the long format.

```
> jasa$subject <- 1:nrow(jasa) #we need an identifier variable
> tdata <- with(jasa, data.frame(subject = subject,
    futime= pmax(.5, fu.date - accept.dt),
    txtime= ifelse(tx.date== fu.date,
    (tx.date -accept.dt) -.5,
    (tx.date - accept.dt)),
    fustat = fustat
))
> sdata <- tmerge(jasa, tdata, id=subject,
    death = event(futime, fustat),
    trt = tdc(txtime),
    options= list(idname="subject"))
> sdata$age <- sdata$age -48
> sdata$year <- as.numeric(sdata$accept.dt - as.Date("1967-10-01"))
  /365.25
> write.csv(sdata, "c:/data/sdata.csv")
> tdata$year <- as.numeric(jasa$accept.dt - as.Date("1967-10-01"))
  /365.25
> tdata$trt = as.numeric(!is.na(tdata$txtime))
> tdata$survtime=tdata$futime
> tdata$age=jasa$age-48
> tdata$death=jasa$fustat
> tdata$surgey=jasa$surgery
> write.csv(tdata, "c:/data/tdata.csv")
```

The variables used in the analysis are listed below.

trt	= 0 if transplant was not done, 1 if transplant was done
age	= baseline age of the patient
surgery	= 0 if there was no prior heart surgery, 1 if there was prior heart surgery
year	= time since the start of the study until the enrollment of the patient

The data on the first 10 subjects from the `tdata.csv` are shown in Table 10.4. The data on the same 10 subjects from the `sdata.csv` are shown in Table 10.5. Notice that there are two rows of data each for subjects 4, 7 and 10. Subject 3 received the transplant on Day 15 but died on the same day; thus there is only a single row of data for that subject.

Table 10.4 Stanford Heart Transplant data regarding the trt variable as not time-dependent (first 10 subjects)

subject	trt	year	age	death	surgery	survtime
1	0	0.123	30.845	1	0	49
2	0	0.255	51.836	1	0	5
3	1	0.266	54.297	1	0	15
4	1	0.490	40.263	1	0	38
5	0	0.608	20.786	1	0	17
6	0	0.701	54.595	1	0	2
7	1	0.780	50.869	1	0	674
8	0	0.835	45.350	1	0	39
9	0	0.857	47.162	1	0	84
10	1	0.862	42.502	1	0	57

Table 10.5 Stanford Heart Transplant data in long format regarding the trt variable as time-dependent (first 10 subjects)

subject	trt	year	age	death	surgery	`tstart`	`tstop`
1	0	0.123	30.845	1	0	0	49
2	0	0.255	51.836	1	0	0	5
3	1	0.266	54.297	1	0	0	15
4	0	0.490	40.263	0	0	0	35
4	1	0.490	40.263	1	0	35	38
5	0	0.608	20.786	1	0	0	17
6	0	0.701	54.595	1	0	0	2
7	0	0.780	50.869	0	0	0	50
7	1	0.780	50.869	1	0	50	674
8	0	0.835	45.350	1	0	0	39
9	0	0.857	47.162	1	0	0	84
10	0	0.862	42.502	0	0	0	11
10	1	0.862	42.502	1	0	11	57

First we give the results of fitting the Cox model to the `tdata.csv`. We see that the trt variable is highly significant.

```
> library(survival)
> tdata=read.csv("c:/data/tdata.csv")
> tfit<-coxph(Surv(survtime, death) ~ trt+age+surgery+year, data= tdata)
> tfit

Call:
coxph(formula = Surv(survtime, death) ~ trt + age + surgery +
    year, data = tdata)

           coef exp(coef) se(coef)     z       p
trt     -1.7045    0.1819   0.2826 -6.03 1.6e-09
age      0.0575    1.0592   0.0147  3.92 9.0e-05
surgery -0.3178    0.7278   0.3767 -0.84   0.399
year    -0.1177    0.8890   0.0692 -1.70   0.089

Likelihood ratio test=48.8  on 4 df, p=6.47e-10
n= 103, number of events= 75
```

However, regarding the trt variable as fixed is not correct since the patients who died early did not have a chance to get a matching heart donor and so their deaths may be wrongly attributed to not getting a transplant which would make the trt variable more significant than it actually is. Therefore we must take into account the waiting time to get a transplant by treating trt as a time-dependent covariate. Below we give the results of fitting the Cox model to the `sdata`. We see that the trt variable is now highly nonsignificant.

```
> library(survival)
> sdata=read.csv("c:/data/sdata.csv")
> sfit<-coxph(Surv(tstart, tstop, death) ~ trt+age + surgery + year,
    data= sdata)
> sfit

Call:
coxph(formula = Surv(tstart, tstop, death) ~ trt + age + surgery +
    year, data = sdata)

           coef exp(coef) se(coef)     z     p
trt     -0.0129    0.9872   0.3133 -0.04 0.967
age      0.0272    1.0276   0.0137  1.98 0.047
surgery -0.6371    0.5288   0.3672 -1.73 0.083
year    -0.1464    0.8638   0.0705 -2.08 0.038

Likelihood ratio test=15.1  on 4 df, p=0.00447
n= 170, number of events= 75
```

■

EXERCISES

Theoretical Exercises

10.1 (Geometric distribution: Hazard rate) The geometric distribution gives the probability of the first success on the tth trial when making successive independent Bernoulli trials, each with success probability θ. It is given by

$$f(t) = P(T = t) = \theta(1-\theta)^{t-1}, \quad t = 1, 2, \ldots.$$

and is a discrete analog of the exponential distribution. Using the hazard rate formula for discrete lifetime:

$$\lambda(t) = \frac{P(T=t)}{P(T \geq t)}, \quad t = 1, 2, \ldots,$$

show that the hazard rate for the geometric distribution is constant and equal to θ.

10.2 (Weibull distribution: Hazard rate) The Weibull distribution is used widely to model failure times in engineering reliability applications. Its p.d.f. is given by

$$f(t; \lambda, \gamma) = \lambda\gamma\,(\lambda t)^{\gamma-1} \exp\left(-(\lambda t)^{\gamma}\right) \quad \text{for} \quad t \geq 0,$$

where λ is the scale parameter and γ is the shape parameter. For $\gamma = 1$, we get the exponential distribution. Another way to think of the Weibull distribution is that if $U = (\lambda T)^{\gamma}$ has the unit exponential distribution then T has the Weibull distribution.

a) Show that the survival function of the Weibull distribution is given by

$$S(t) = \exp\left(-(\lambda t)^{\gamma}\right).$$

b) Show that the hazard rate of the Weibull distribution is given by

$$\lambda(t) = \lambda\gamma\,(\lambda t)^{\gamma-1}.$$

c) Show that the hazard rate of the Weibull distribution is increasing in t if $\gamma > 1$, decreasing in t if $\gamma < 1$ and constant if $\gamma = 1$ (the exponential distribution).

10.3 (Uncensored data: Survival function estimation) Show that if there are no censored observations then $\widehat{S}(t)$ is simply 1 minus the empirical c.d.f. of T, i.e., $\widehat{S}(t) = 1 - \sum_{j=1}^{i} d_j/n$, which is just the binomial proportion of the patients still surviving at time t for $t_i \leq t < t_{i+1}$.

10.4 (Greenwood formula) Derive the Greenwood formula (10.4). (Hint: First find

$$\text{Var}[\ln(\widehat{S}(t))] = \sum_{t_i \leq t} \text{Var}[\ln(1 - \widehat{\lambda}(t_i))]$$

using the delta method (4.1). Since $\widehat{\lambda}(t_i) = d_i/n_i$ is a binomial proportion, use $\text{Var}(\widehat{\lambda}(t_i)) = \lambda(t_i)(1-\lambda(t_i))/n_i$. Finally obtain $\text{Var}(\widehat{S}(t))$ from $\text{Var}[\ln(\widehat{S}(t))]$ by re-applying the delta method to the transformation $\widehat{S}(t) = \exp(\ln(\widehat{S}(t)))$.

10.5 (Exponential data: One sample problem) Consider n independent observations $(\boldsymbol{x}_i, t_i, \delta_i)$, where t_i is the survival time, $\boldsymbol{x}_i$ is the covariate vector and δ_i is the censoring indicator ($\delta_i = 1$ if t_i is observed and $\delta_i = 0$ if t_i is censored) of the ith patient ($i = 1, \ldots, n$). Let m denote the number of uncensored observations (deaths) and $n - m$ denote the number of censored observations. Assume that the t_i are exponentially distributed and follow the Cox proportional hazards model with a constant (with respect to time) hazard rate $\lambda_i = \lambda_0 \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ for the ith patient.

a) Write the full likelihood function for this model and derive the equations for finding the MLE's of λ_0 and $\boldsymbol{\beta}$.

b) Why is the partial likelihood approach not necessary nor advisable in this case?

c) What is the MLE of λ_0 if there are no covariates, i.e., the t_i are identically distributed?

10.6 (Exponential data: Two sample problem) Consider the same set up as in the previous exercise but now assume that the only covariate is the indicator variable for the group, $x_i = 0$ for the placebo group (denoted by P) and $x_i = 1$ for the treatment group (denoted by T). Thus $\lambda_i = \lambda_0$ for $i \in P$ and $\lambda_i = \lambda_0 \exp(\beta)$ for $i \in T$. Further denote the set of censored observations (from both the placebo group and the treatment group) by C and the set of uncensored observations (deaths) by D. Let n_0 and n_1 be the number of patients in the two groups of whom $m_0 = |P \cap D|$ and $m_1 = |T \cap D|$ are uncensored (deaths), respectively.

a) Show that the MLE's of λ_0 and β are
$$\widehat{\lambda}_0 = \frac{1}{\bar{t}_0} \quad \text{and} \quad \widehat{\beta} = \ln\left(\frac{\bar{t}_0}{\bar{t}_1}\right).$$
b) Interpret $\widehat{\beta}$. Explain how $\widehat{\beta} < 0$ implies that the treatment is effective.

10.7 (Survival function for the Cox regression model) Use (10.2) to show that for the Cox regression model, for the ith individual with the covariate vector $\boldsymbol{x}_i$ and the risk score $\psi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, the survival function is given by
$$S_i(t) = [S_0(t)]^{\psi_i},$$
where $S_0(t)$ is the survival function corresponding to the baseline hazard rate. Thus the higher the risk score of an individual the lower is his/her survival distribution.

Applied Exercises

10.8 (Cellphone data: Kaplan-Meier curves and logrank test) A cellular service provider keeps data on how many months their customers maintained service with the company before they switched to another company. If a customer did not switch to another service provider then the service time on the customer is censored. The file `cellphone data.csv` contains 4912 records of customers (88 have missing data on at least one variable) with data on the following variables: Months = the number of months of service, Account-Type = Business (B) or Individual (I), Churn = censor indicator (0 if no, 1 if yes), Line-Count = number of phone lines served).

a) Make Kaplan-Meier curves for business and individual customers. Which customers maintain their service longer? How long do most customers maintain the service? Does this seem to be related to the standard service contract period of two years that the cellphone providers used to have before this restriction was removed?
b) Do the logrank test to check if there is a significant difference between the two survival curves.

10.9 (Cellphone data: Cox model) Fit a Cox proportional hazards model with Months as the response variable and the Account-Type as the predictor variable. Fit another Cox proportional hazards model with both the Account-Type and Line-Count as predictor variables. Note that the Account-Type goes from being a highly significant variable in the first model to a highly nonsignificant variable in the second model. What could explain this change? Note that the business customers generally have more lines.

10.10 (Air Miles Reward Program: Cox model) Air Miles Reward Program (AMRP) is a loyalty program of an airline for redeeming miles for a reward. Data on 5330 program members are in the file `AMRP.csv`. All members in the data set have redeemed at least

once. (The data was missing on many variables for the members who had never redeemed and so they are not included in this data set.) The goal is to model the time until the next (i.e., the second) redemption, so customers who have redeemed only once are censored. The variables included in the data set are as follows.

1. t: time in days until the second redemption or censoring
2. censored: 1 for censored, 0 for event (redemption)
3. totredeem: total number of miles redeemed in the past
4. prevcat: category of the previous redemption (travel, merchandise, gift certificate, entertainment)
5. prefood: previous miles earned/day in food (grocery)
6. pregas: previous miles earned/day gas
7. prebonus: previous miles earned/day that were bonus miles, e.g., double miles under a certain promotion. This measures if they chase promotions.
8. baselen: how long they have been a member.

Find the strongest predictors of the redemption time.

10.11 (Recidivism study: Time-dependent employment status) In Example 10.8 we analyzed the recidivism data using Aid, Age, Race, Work, Married, Parole, Prior and Education as covariates with Week as the response variable and Arrest as the censoring indicator. We did not use the time-dependent covariate employment status (0 if not employed, 1 if employed), which changes every week and is denoted by Emplt for the tth week ($t = 1, \ldots, 52$) in the data file `recid.csv`. Note that if the person is arrested in the sth week then Emplt is marked NA for $t > s$.

Fit the Cox PH model to the recidivism data with the employment status as a time-dependent covariate in addition to those fixed covariates used in Example 10.8. Comment on any changes in significance of the variables.

APPENDIX A

SOME RESULTS FROM MATRIX ALGEBRA AND MULTIVARIATE DISTRIBUTIONS

A.1 Results from Matrix Algebra

We assume the basic knowledge of matrix algebra including arithmetic operations with vectors and matrices. We will review a few advanced concepts that are useful in linear models.

We use bold letters to denote vectors and matrices with lower case letters denoting vectors and upper case letters denoting matrices. Their dimensions are not generally indicated notationally and their elements are denoted by the respective unbolded letters with appropriate subscripts, e.g., a_i for an element of vector $\boldsymbol{a}$ and a_{ij} for an element of matrix $\boldsymbol{A}$. All vectors are assumed to be column vectors and transpose of a vector or a matrix is indicated by putting a prime on its symbol.

A symmetric $m \times m$ matrix $\boldsymbol{A}$ is said to be **positive definite** if for all non-null vectors $\boldsymbol{a} = (a_1, \ldots, a_m)'$, the **quadratic form** $\boldsymbol{a}'\boldsymbol{A}\boldsymbol{a} > 0$. If $\boldsymbol{a}'\boldsymbol{A}\boldsymbol{a} = 0$ for some non-null vector $\boldsymbol{a}$ then $\boldsymbol{A}$ is said to be **positive semidefinite** or **non-negative definite**. An inverse of a positive definite matrix (denoted by $\boldsymbol{A}^{-1}$) exists and $\boldsymbol{A}$ is said to be **non-singular**. In that case $\boldsymbol{A}^{-1}$ is also positive definite. If $\boldsymbol{A}$ positive semidefinite then $\boldsymbol{A}^{-1}$ does not exist and $\boldsymbol{A}$ is said to be **singular**.

As we saw in Chapter 3, the $(p+1) \times (p+1)$ symmetric matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ arises in multiple regression in the computation of the LS estimates and their covariance matrix. For

this inverse to exist $\boldsymbol{X}'\boldsymbol{X}$ must be positive definite, which can be checked as follows. Let $\boldsymbol{a} = (a_1, \ldots, a_{p+1})'$ be a non-null vector. Denoting $\boldsymbol{b} = \boldsymbol{X}\boldsymbol{a} = (b_1, \ldots, b_{p+1})'$, we have

$$\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a} = \boldsymbol{b}'\boldsymbol{b} = \sum_{i=1}^{p+1} b_i^2 \geq 0$$

for all non-null vectors $\boldsymbol{a}$. Note that $\boldsymbol{b}$ is a linear combination of the columns of $\boldsymbol{X}$ with the coefficients of the linear combination being $a_1, \ldots, a_{p+1}$. The above inequality will be an equality iff $\boldsymbol{b} = \boldsymbol{X}\boldsymbol{a}$ is a null vector for some vector $\boldsymbol{a}$, which means that the columns of $\boldsymbol{X}$ are linearly dependent. Therefore for $\boldsymbol{X}'\boldsymbol{X}$ to be positive definite and hence invertible, the columns of $\boldsymbol{X}$ must be linearly independent.

Another useful concept from linear algebra is that of **eigenvalues** (also called **singular values**) and **eigenvectors**. The eigenvalue λ and its associated eigenvector $\boldsymbol{u}$ of an $m \times m$ matrix $\boldsymbol{A}$ are defined by the equation

$$[\boldsymbol{A} - \lambda\boldsymbol{I}]\boldsymbol{u} = \boldsymbol{0},$$

where $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{0}$ is a null vector. This is a linear system of equations in unknowns $\boldsymbol{u} = (u_1, \ldots, u_m)'$. By Crámer's rule this system has a nontrivial solution iff the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$ is singular, i.e., iff its determinant vanishes:

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \det\begin{bmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} - \lambda \end{bmatrix} = 0.$$

This is a polynomial equation of degree m and thus has m solutions, $\lambda_1, \ldots, \lambda_m$, not necessarily distinct. If they are distinct then there exist associated m eigenvectors, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, which are mutually orthogonal, i.e., $\boldsymbol{u}_i'\boldsymbol{u}_j = 0$ for all $i \neq j$. Furthermore if all eigenvectors are scaled to be of unit length then the matrix $\boldsymbol{U}$ whose columns are the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ is an orthogonal matrix, i.e., $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$.

Let $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$ be the diagonal matrix with eigenvalues of $\boldsymbol{A}$ as its entries. . Then the **spectral decomposition theorem** (also known as the **Jordan decomposition theorem**) states that

$$\boldsymbol{U}\boldsymbol{A}\boldsymbol{U}' = \boldsymbol{\Lambda} \quad \text{or} \quad \boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{U} = \boldsymbol{A}. \tag{A.1}$$

From this composition it follows that

$$\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{m} \lambda_i \quad \text{and} \quad \det(\boldsymbol{A}) = \det(\boldsymbol{\Lambda}) = \prod_{i=1}^{m} \lambda_i.$$

In this sense the essential information about $\boldsymbol{A}$ is contained in its eigenvalues and eigenvectors.

Singular value decomposition (SVD) is a generalization of spectral decomposition. Let $\boldsymbol{A}$ be an $m \times p$ matrix of rank $r \leq p$. Then $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}', \tag{A.2}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$, $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$. The column vectors of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}'$ and the column vectors of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}'\boldsymbol{A}$. The eigenvalues of both $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}'$ are $\lambda_1^2, \ldots, \lambda_r^2$, the remaining eigenvalues being 0. SVD has many applications; in the regression context it is used in the computation of the LS estimates and in the computation of principal component scores.

A.2 Results from Multivariate Distributions

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ be a random vector where $x_1, x_2, \ldots, x_m$ are jointly distributed random variables (r.v.'s) with means $E(x_i) = \mu_i$, variances

$$\text{Var}(x_i) = E[(x_i - \mu_i)^2] = \sigma_{ii} = \sigma_i^2$$

and covariances

$$\text{Cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] = \sigma_{ij}.$$

The **mean vector** of $\boldsymbol{x}$ equals

$$\boldsymbol{\mu} = E(\boldsymbol{x}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_m) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}.$$

The **covariance matrix** of $\boldsymbol{x}$ (denoted by $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x})$)[1] is an $m \times m$ matrix with diagonal elements, $\text{Var}(x_i) = \sigma_{ii} = \sigma_i^2$, and off-diagonal elements, $\text{Cov}(x_i, x_j) = \sigma_{ij}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}.$$

It is easy to see that $\boldsymbol{\Sigma}$ can be expressed as

$$\begin{aligned} \boldsymbol{\Sigma} &= E\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_m - \mu_m \end{bmatrix} \begin{bmatrix} x_1 - \mu_1, & x_2 - \mu_2, & \cdots, & x_m - \mu_m \end{bmatrix} \right) \\ &= E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})']. \end{aligned} \tag{A.3}$$

Since $\text{Cov}(x_i, x_j) = \text{Cov}(x_j, x_i)$, we have $\sigma_{ij} = \sigma_{ji}$ for all $i \neq j$; therefore $\boldsymbol{\Sigma}$ is a symmetric matrix. Furthermore, $\boldsymbol{\Sigma}$ is a **positive semidefinite matrix**, i.e., for all non-null vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$, we have $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$. In fact, if the r.v.'s x_i's are linearly independent, i.e., if there is no vector $\boldsymbol{a} \neq \boldsymbol{0}$ such that $\boldsymbol{a}'\boldsymbol{x} = \sum a_i x_i$ equals a constant, then $\boldsymbol{\Sigma}$ is a **positive definite matrix**. These results follow from (A.5) below.

We know from basic probability that if a, b, c and d are constants ($a, b \neq 0$), x and y are r.v.'s, and $u = ax + c$ and $v = by + d$ then

$$\text{Cov}(u, v) = ab\text{Cov}(x, y) \text{ and } \text{Corr}(u, v) = \pm\text{Corr}(x, y),$$

where the sign is + if $ab > 0$ and the sign is $-$ if $ab < 0$. Thus, if the r.v.'s x and y are linearly transformed then the additive constants c and d have no effect on the covariance; furthermore, the multiplicative constants a and b have no effect on the correlation except through their signs.

More generally, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ be two random vectors and $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)'$ be two vectors of constants.

[1] We use the same notation $\text{Cov}(\cdot)$ to denote different types of covariances, e.g., $\text{Cov}(\boldsymbol{x})$ denotes the covariance matrix of a random vector $\boldsymbol{x}$, $\text{Cov}(x_i, x_j)$ denotes the covariance of two scalar r.v.'s x_i and x_j, and $\text{Cov}(\boldsymbol{x}, \boldsymbol{y})$ denotes the covariance matrix between two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., the matrix of covariances between the r.v.'s x_i and the r.v.'s y_j.

Let

$$u = \sum_{i=1}^{m} a_i x_i = \boldsymbol{a}'\boldsymbol{x} \text{ and } v = \sum_{j=1}^{n} b_j y_j = \boldsymbol{b}'\boldsymbol{y}.$$

Further let $\boldsymbol{\Omega} = \text{Cov}(\boldsymbol{x}, \boldsymbol{y})$ denote an $m \times n$ covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$ whose elements are $\omega_{ij} = \text{Cov}(x_i, y_j)$ $(1 \le i \le m, 1 \le j \le n)$. Note that $\boldsymbol{\Omega}$ is not a symmetric matrix if $\boldsymbol{x} \ne \boldsymbol{y}$ even if $m = n$. Then

$$\text{Cov}(u, v) = \text{Cov}(\boldsymbol{a}'\boldsymbol{x}, \boldsymbol{b}'\boldsymbol{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \text{Cov}(x_i, y_j) = \boldsymbol{a}'\boldsymbol{\Omega}\boldsymbol{b}. \tag{A.4}$$

A special case of the above formula is obtained by putting $u = v$:

$$\text{Var}(\boldsymbol{a}'\boldsymbol{x}) = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \text{Cov}(x_i, x_j) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}. \tag{A.5}$$

Note that $\text{Var}(\boldsymbol{a}'\boldsymbol{x}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \ge 0$ for $\boldsymbol{a} \ne \boldsymbol{0}$ and equals 0 iff $\boldsymbol{a}'\boldsymbol{x}$ equals a constant. This shows that $\boldsymbol{\Sigma}$ is positive semidefinite and is in fact positive definite if the r.v.'s x_i's are linearly independent.

Suppose that $\boldsymbol{x}$ is an $m \times 1$ random vector and $\boldsymbol{A}$ is a $p \times m$ matrix of constants. Let $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x}$. It is readily shown that

$$E(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}E(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\mu}. \tag{A.6}$$

Then using (A.3), it follows that

$$\begin{aligned} \text{Cov}(\boldsymbol{u}) &= \text{Cov}(\boldsymbol{A}\boldsymbol{x}) \\ &= E[(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{\mu})(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{\mu})'] \\ &= E[\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{A}'] \\ &= \boldsymbol{A}E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})']\boldsymbol{A}' \\ &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'. \end{aligned} \tag{A.7}$$

This is known as the **sandwich formula**, which generalizes (A.5).

A.3 Multivariate Normal Distribution

The random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ has a multivariate normal (MVN) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (denoted by $\boldsymbol{x} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) if the joint probability density function (p.d.f.) of $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \tag{A.8}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. In the above it is assumed that $\boldsymbol{\Sigma}$ is invertible or equivalently positive definite. We will only consider this nonsingular case of the MVN distribution.

The marginal p.d.f. of each component r.v. x_i is $N(\mu_i, \sigma_i^2)$. If $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$, i.e., if $\text{Cov}(x_i, x_j) = \sigma_{ij} = 0$ for all $i \ne j$, and thus the x_i's are uncorrelated, then the joint p.d.f. (A.8) of $\boldsymbol{x}$ factors into the product of the marginal p.d.f.'s of x_i's:

$$f(\boldsymbol{x}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\right\}.$$

Therefore the x_i's are independent and are distributed as $N(\mu_i, \sigma_i^2)$. The converse of this result is immediate. Therefore if $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ is MVN distributed then the x_i's are independent if and only if they are uncorrelated.

The following is a useful property of the MVN distribution: If $\boldsymbol{x} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then any fixed (nonrandom) nonsingular linear transformation of $\boldsymbol{x}$ also has an MVN distribution. Specifically, let $\boldsymbol{A}$ be a $p \times m$ non-random matrix with linearly independent rows. Then $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x} = (u_1, u_2, \ldots, u_p)'$ has an MVN distribution of dimension p with the mean vector and covariance matrix given by

$$E(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{\mu} \text{ and } \text{Cov}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'.$$

Two cases of this result are of particular interest.

1. Since $\boldsymbol{\Sigma}$ is invertible it can be shown that there exists an $m \times m$ symmetric matrix, say $\boldsymbol{P}$, such that $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}' = \boldsymbol{I}$ and $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}$. Then
$$\boldsymbol{z} = \boldsymbol{P}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{A.9}$$
is MVN with
$$E(\boldsymbol{z}) = \boldsymbol{0} \text{ and } \text{Cov}(\boldsymbol{z}) = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}' = \boldsymbol{I},$$
i.e., $z_1, z_2, \ldots, z_m$ are i.i.d. $N(0, 1)$ r.v.'s. Thus (A.9) is a standardizing transformation.

2. Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$ be a vector of constants. Then $u = \boldsymbol{a}'\boldsymbol{x} = \sum a_i x_i$ is univariate normal with
$$E(u) = \boldsymbol{a}'\boldsymbol{\mu} = \sum_{i=1}^{m} a_i \mu_i \text{ and } \text{Var}(u) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \sigma_{ij}.$$

APPENDIX B

PRIMER ON MAXIMUM LIKELIHOOD ESTIMATION

B.1 Maximum Likelihood Estimation

The method of maximum likelihood estimation was proposed by Sir R. A. Fisher. Let y be a r.v., either discrete or continuous, with probability mass or density function (both abbreviated as p.d.f.) $f(y|\theta)$. Here θ is an unknown parameter, which we want to estimate from an i.i.d. random sample $y_1, \ldots, y_n$ drawn from this distribution. We can view the joint p.d.f. of $y_1, \ldots, y_n$, ,

$$f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta),$$

as the probability of their occurrence if the true parameter is θ. (If the y_i are discrete then this is the probability; if the y_i are continuous then $f(y_1, \ldots, y_n|\theta)dy_1 \ldots dy_n$ is the probability element.) This is called the **likelihood function**, viewed as a function of θ for given $y_1, \ldots, y_n$ (whereas the joint p.d.f. is a function of $y_1, \ldots, y_n$ for given θ and has the probability interpretation given above). We denote the likelihood function by

$$L(\theta) = L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta).$$

We ask the question: what value of θ makes the observed data $y_1, \ldots, y_n$ most likely? This value of θ, which maximizes the likelihood function, is called the **maximum likelihood estimator (MLE)** of θ and is denoted by $\widehat{\theta}$.

Since the log is a monotone increasing function, maximizing $L(\theta)$ is equivalent to maximizing its log. So we define the **log-likelihood function** as

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i|\theta). \tag{B.1}$$

To keep things simple we will restrict to the so-called regular case where the log-likelihood function is differentiable and concave, so that its maximum can be found by setting the derivative of $lnL(\theta)$ equal to zero and solving the resulting equation:

$$\frac{d \ln L(\widehat{\theta})}{d\theta} = \left[\frac{d \ln L(\theta)}{d\theta}\right]_{\theta=\widehat{\theta}} = 0. \tag{B.2}$$

We assume that the solution to this equation, which is the MLE $\widehat{\theta}$, exists and is unique.

EXAMPLE B.1 (MLE of Bernoulli Parameter)

Suppose $f(y|\theta)$ is a Bernoulli distribution with

$$f(y|\theta) = \theta \quad \text{if} \quad y = 1 \quad \text{and} \quad f(y|\theta) = 1 - \theta \quad \text{if} \quad y = 0,$$

where θ is the probability of success on a single Bernoulli trial. This distribution can be written compactly as

$$f(y|\theta) = \theta^y (1-\theta)^{1-y} \text{ for } y = 0, 1.$$

The likelihood function is then given by

$$L(\theta) = \prod_{i=1}^{n} \left[\theta^{y_i}(1-\theta)^{1-y_i}\right] = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{\sum_{i=1}^n (1-y_i)} = \theta^s (1-\theta)^{n-s},$$

where $s = \sum_{i=1}^n y_i$ is the number of successes in n Bernoulli trials. The log-likelihood function is

$$\ln L(\theta) = s \ln \theta + (n-s)\ln(1-\theta).$$

Taking the derivative of $\ln L(\theta)$ w.r.t. θ and setting it equal to zero, we get

$$\frac{d \ln L(\theta)}{d\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0.$$

The solution to this equation is $\widehat{\theta} = s/n$, which is the MLE. Note that this is simply the proportion of successes. It is easy to see that the second derivative of $\ln L(\theta)$ at $\theta = \widehat{\theta}$ is negative and hence $\widehat{\theta}$ indeed gives the maximum.

Instead of working with n i.i.d. Bernoulli outcomes, we can work directly with the distribution of their sum $s = \sum_{i=1}^n y_i$. The distribution of s is binomial:

$$f(s|\theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s},$$

which is the likelihood function of θ. Note that this likelihood function differs from the previous one obtained using the Bernoulli distribution only in the multiplication factor $\binom{n}{s}$. But this factor is irrelevant to maximizing the likelihood function w.r.t. θ since it does not involve θ. Therefore we get the same MLE $\widehat{\theta} = s/n$. ■

B.2 Large Sample Inference on MLE's

Next we discuss inference on the MLE. Fisher showed that, under certain regularity conditions, asymptotically (as $n \to \infty$) the MLE $\widehat{\theta}$ is approximately normally distributed

with mean θ and asymptotic variance $[\mathcal{I}(\theta)]^{-1}$, where

$$\mathcal{I}(\theta) = E\left[\left(\frac{d\ln L(\theta)}{d\theta}\right)^2\right] = -E\left[\frac{d^2\ln L(\theta)}{d\theta^2}\right] \tag{B.3}$$

is the so-called **Fisher information**. The above identity is derived in Section B.4. A plug-in estimate of $\mathcal{I}(\theta)$, denoted by $\mathcal{I}(\widehat{\theta})$, can be obtained by substituting $\widehat{\theta}$ for θ in the expected value. Another approach is to substitute $\widehat{\theta}$ in the second derivative expression of the log-likelihood function without taking the expected value. The former quantity is called the **expected information** while the latter quantity is called the **observed information**. They equal under certain regularity conditions. In that case

$$\mathcal{I}(\widehat{\theta}) = -\left[\frac{d^2\ln L(\theta)}{d\theta^2}\right]_{\theta=\widehat{\theta}} = -\sum_{i=1}^{n}\left[\frac{d^2\ln f(y_i|\theta)}{d\theta^2}\right]_{\theta=\widehat{\theta}}.$$

A large sample $100(1-\alpha)\%$ CI on θ is given by

$$\widehat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{\mathcal{I}(\widehat{\theta})}} \le \theta \le \widehat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{\mathcal{I}(\widehat{\theta})}}.$$

EXAMPLE B.2 (Inference on the MLE of Bernoulli Parameter)

We first evaluate $\mathcal{I}(\theta)$. Using the result from Example B.1 we get

$$\frac{d^2\ln L(\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}.$$

Hence

$$\mathcal{I}(\theta) = \frac{E(s)}{\theta^2} + \frac{E(n-s)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}.$$

Thus $[\mathcal{I}(\theta)]^{-1} = \theta(1-\theta)/n$, which is in fact the *exact* variance of the MLE $\widehat{\theta} = s/n$. It is estimated by $\widehat{\theta}(1-\widehat{\theta})/n$.

We can check that we get the same result if we calculate the observed information. Thus

$$\mathcal{I}(\widehat{\theta}) = \frac{s}{\widehat{\theta}^2} + \frac{n-s}{(1-\widehat{\theta})^2} = \frac{n\widehat{\theta}}{\widehat{\theta}^2} + \frac{n(1-\widehat{\theta})}{(1-\widehat{\theta})^2} = \frac{n}{\widehat{\theta}(1-\widehat{\theta})}.$$

Using this result, we get the following formula for the large sample $100(1-\alpha)\%$ C.I. on θ:

$$\widehat{\theta} - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} \le \theta \le \widehat{\theta} + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}},$$

a formula that we learn in an elementary statistics course. ■

Next we extend these results to multiple parameters. Suppose the distribution of y depends on $p \ge 2$ unknown parameters, $\theta_1, \ldots, \theta_p$, represented as a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and denote the distribution of y by $f(y|\boldsymbol{\theta})$. Assume that we have an i.i.d. sample $y_1, \ldots, y_n$ from this distribution. Then the log-likelihood function equals

$$\ln L(\boldsymbol{\theta}) = \ln\left[\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})\right] = \sum_{i=1}^{n}\ln f(y_i|\boldsymbol{\theta}).$$

The MLE's of the θ_j's are obtained by solving the p simultaneous equations obtained by setting the partial derivatives of the log-likelihood function w.r.t. the θ_j's equal to zero:

$$\frac{\partial\ln L(\boldsymbol{\theta})}{\partial\theta_j} = \sum_{i=1}^{n}\frac{\partial\ln f(y_i|\boldsymbol{\theta})}{\partial\theta_j} = 0 \quad (1 \le j \le p).$$

To make inferences on the $\widehat{\theta}_j$'s we need to calculate the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$. The diagonal elements of this matrix are the asymptotic variances of the $\widehat{\theta}_j$'s. This covariance matrix is the inverse of the **information matrix** $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \{I_{jk}(\boldsymbol{\theta})\}$, where

$$\boldsymbol{\mathcal{I}}_{jk}(\boldsymbol{\theta}) = -E \sum_{i=1}^{n} \left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right] \quad (1 \le j < k \le p).$$

The estimated information matrix can be obtained in two ways as before. One way is to compute the above expected information matrix and substitute $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ in it. The other way is to compute the observed information matrix by substituting $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ in the second derivative expressions of the log-likelihood function.

EXAMPLE B.3 (Information matrix for simple logistic regression)

In this example we derive the formula (7.6) for the information matrix for simple logistic regression. From the first derivatives of the log-likelihood function we can calculate the following second derivatives:

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial \beta_0^2} &= -\sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} = -\sum_{i=1}^{n} p_i(1 - p_i), \\
\frac{\partial^2 \ln L}{\partial \beta_0 \partial \beta_1} &= -\sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} = -\sum_{i=1}^{n} x_i p_i(1 - p_i), \\
\frac{\partial^2 \ln L}{\partial \beta_1^2} &= -\sum_{i=1}^{n} \frac{x_i^2 \exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} = -\sum_{i=1}^{n} x_i^2 p_i(1 - p_i).
\end{aligned}$$

Note that these second derivatives do not involve the y_i's, which are the random quantities. Therefore there are no expected values to be taken. Hence both methods of calculating the estimated information matrix give the same result. ■

B.3 Newton-Raphson and Fisher Scoring Algorithms

We will present these algorithms for the case of a single unknown parameter θ and then generalize them to the multiparameter case. They provide an iterative way to find the MLE $\widehat{\theta}$, which satisfies

$$\frac{d \ln L(\widehat{\theta})}{d\theta} \equiv \frac{d \ln L(\theta)}{d\theta}\Big|_{\theta=\widehat{\theta}} = 0, \tag{B.4}$$

which is referred to as the **MLE score equation**. We begin with an initial guess $\widehat{\theta}_0$ for its solution. By the first-order Taylor series expansion of $d \ln L(\widehat{\theta})/d\theta$ around $\widehat{\theta}_0$ we get

$$\frac{d \ln L(\widehat{\theta})}{d\theta} \approx \frac{d \ln L(\widehat{\theta}_0)}{d\theta} + (\widehat{\theta} - \widehat{\theta}_0) \frac{d^2 \ln L(\widehat{\theta}_0)}{d\theta^2}. \tag{B.5}$$

We set the RHS of the above equation equal to 0 and solve for $\widehat{\theta}$ to obtain the first approximation $\widehat{\theta}_1$ to the solution of the MLE equation:

$$\widehat{\theta}_1 = \widehat{\theta}_0 - \left[\frac{d^2 \ln L(\widehat{\theta}_0)}{d\theta^2} \right]^{-1} \left(\frac{d \ln L(\widehat{\theta}_0)}{d\theta} \right).$$

Repeating this procedure we get the following recursion at the rth iteration:

$$\widehat{\theta}_{r+1} = \widehat{\theta}_r - \left[\frac{d^2 \ln L(\widehat{\theta}_r)}{d\theta^2}\right]^{-1} \left(\frac{d \ln L(\widehat{\theta}_r)}{d\theta}\right).$$

Recall that $-d^2 \ln L(\widehat{\theta}_r)/d\theta^2 = \mathcal{I}(\widehat{\theta}_r)$, namely the observed information evaluated at the current estimate $\widehat{\theta}_r$. Therefore the above iterative equation can be equivalently written as

$$\widehat{\theta}_{r+1} = \widehat{\theta}_r + [\mathcal{I}(\widehat{\theta}_r)]^{-1} \frac{d \ln L(\widehat{\theta}_r)}{d\theta}. \tag{B.6}$$

The algorithm converges when a specified convergence criterion is met, e.g., when $|\widehat{\theta}_{r+1} - \widehat{\theta}_r| < \varepsilon$ for some specified tolerance $\varepsilon > 0$. Note that not only does this algorithm give the MLE $\widehat{\theta}$ but it also gives the asymptotic variance $\text{Var}(\widehat{\theta}) = [\mathcal{I}(\widehat{\theta})]^{-1}$ at the final step.

Fisher's **score statistic** is defined as

$$U(\theta) = \frac{d \ln L(\theta)}{d\theta}.$$

In the next section we show under certain regularity conditions that its mean and variance are given by

$$E[U(\theta)] = E\left[\frac{d \ln L(\theta)}{d\theta}\right] = 0 \text{ and } \text{Var}[U(\theta)] = -E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right]. \tag{B.7}$$

The equation (B.4) is equivalent to solving $U(\widehat{\theta}) = 0$. The Fisher scoring algorithm is essentially the same as the Newton-Raphson algorithm except that the expected information evaluated at $\widehat{\theta}_r$, i.e., $E[\mathcal{I}(\theta)]_{\theta=\widehat{\theta}_r}$ is used instead of the observed information $I(\widehat{\theta}_r)$ in the recursion (B.5).

In the multiparameter case, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ denote the unknown parameter vector of interest and let $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ denote the associated information matrix. Then the above recursion generalizes to the following:

$$\widehat{\boldsymbol{\theta}}_{r+1} = \widehat{\boldsymbol{\theta}}_r + [\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_r)]^{-1} \frac{d \ln L(\widehat{\boldsymbol{\theta}}_r)}{d\boldsymbol{\theta}}. \tag{B.8}$$

Once again Fisher's scoring algorithm uses the expected value of the information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

B.4 Technical Notes

In this section we derive the identity (B.3) and provide an outline of the proof of the asymptotic normality of the MLE. The information from a single observation y with p.d.f. $f(y, \theta)$ is defined as

$$\mathcal{I}_1(\theta) = E\left[\left(\frac{d \ln L(\theta)}{d\theta}\right)^2\right] = \int \left(\frac{d \ln L(\theta)}{d\theta}\right)^2 f(y, \theta) dy.$$

Under appropriate regularity conditions we can move the derivative inside the integral sign and write

$$\begin{aligned} E\left(\frac{d \ln L(\theta)}{d\theta}\right) &= \int \frac{d \ln f(y, \theta)}{d\theta} f(y, \theta) dy \\ \int \frac{d \ln f(y, \theta)}{d\theta} f(y, \theta) dy & \\ &= \frac{d}{d\theta} \int f(y, \theta) dy \\ &= 0 \quad (\text{since } \textstyle\int f(y, \theta) dy = 1). \end{aligned}$$

Differentiating under the integral sign in the above equation we get

$$\begin{aligned}
& \int \left[\frac{d^2 \ln f(y,\theta)}{d\theta^2} f(y,\theta) + \frac{d \ln f(y,\theta)}{d\theta} \frac{df(y,\theta)}{d\theta} \right] dy \\
= & \int \left[\frac{d^2 \ln f(y,\theta)}{d\theta^2} + \frac{d \ln f(y,\theta)}{d\theta} \frac{df(y,\theta)}{d\theta} \frac{1}{f(y,\theta)} \right] f(y,\theta) dy \\
= & \int \left[\frac{d^2 \ln f(y,\theta)}{d\theta^2} + \left(\frac{d \ln f(y,\theta)}{d\theta} \right)^2 \right] f(y,\theta) dy \\
= & \ 0.
\end{aligned}$$

Hence

$$\int \left(\frac{d \ln f(y,\theta)}{d\theta} \right)^2 f(y,\theta) dy = - \int \frac{d^2 \ln f(y,\theta)}{d\theta^2} f(y,\theta) dy.$$

By putting $L(\theta) = f(y,\theta)$, namely the likelihood function from a single observation, we can write the above equation as

$$\mathcal{I}_1(\theta) = E \left[\left(\frac{d \ln L(\theta)}{d\theta} \right)^2 \right] = -E \left[\frac{d^2 \ln L(\theta)}{d\theta^2} \right],$$

which gives an alternative expression for $\mathcal{I}(\theta)$.

The information from n i.i.d. observations $y_1, \ldots, y_n$ with a common p.d.f. $f(y,\theta)$ equals $\mathcal{I}(\theta) = \mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$. This follows since from (B.1) we can write

$$\mathcal{I}(\theta) = -E \left[\left(\frac{d^2 \ln L}{d\theta^2} \right) \right] = -E \left[\sum_{i=1}^{n} \frac{d^2 f(y_i,\theta)}{d\theta^2} \right] = \sum_{i=1}^{n} -E \left[\frac{d^2 f(y_i,\theta)}{d\theta^2} \right] = nI_1(\theta),$$

where we have used the fact that $\mathcal{I}_1(\theta)$ is the same for all n i.i.d. observations and so $\mathcal{I}(\theta) = n\mathcal{I}_1(\theta)$.

Next consider the proof of the asymptotic normality of the MLE $\widehat{\theta}$. Using the same first order Taylor series expansion as in (B.5) we get

$$\begin{aligned}
\widehat{\theta} - \theta & \approx \frac{d \ln L(\theta)/d\theta}{-d^2 \ln L(\theta)/d\theta^2} \\
& = \frac{(1/n) \sum_{i=1}^{n} (d \ln f(y_i,\theta)/d\theta)}{-(1/n) \sum_{i=1}^{n} \left(d^2 \ln f(y_i,\theta)/d\theta^2 \right)}.
\end{aligned}$$

The numerator is the mean of n i.i.d. Fisher score statistics $U_i = d \ln f(y_i,\theta)/d\theta$ and so by the central limit theorem approaches a normal distribution with mean equal to $E(U_i) = 0$ and variance equal to $\text{Var}(U_i)/n = \mathcal{I}(\theta)/n$. The denominator is also the mean of n i.i.d. random variables $d^2 \ln f(y_i,\theta)/d\theta^2$, each of which has expectation $\mathcal{I}(\theta)$. Hence by the law of large numbers the denominator approaches $\mathcal{I}(\theta)$ in probability. Hence $\widehat{\theta} - \theta$ approaches $N(0, \mathcal{I}^{-1}(\theta)/n)$ in distribution.

APPENDIX C

PRIMER ON R

You must first install R from the CRAN website on your computer. We will use the bacteria data from Table 2.1 to illustrate some basic R commands. The first step is to read the data into R. This data set is small and can be typed directly into R with the following command:

```
> bacteria = data.frame(t=1:15,
  Nt=c(355,211,197,166,142,106,104,60,56,38,36,32,21,19,15))
```

It is more common to have large data sets that would be impractical to type into R directly. Data are often available in the form of a comma delimited file. The bacteria data are in `bacteria.csv` posted at the book's website. This file can be read into R as follows:

```
> bacteria = read.csv("bacteria.csv")
```

Either command creates an object of class `data.frame` composed of two variables, `t` and `Nt`. The variables in a data frame can be displayed with the `names` command, the dimensions with the `dim` command, and descriptive statistics with the `summary` command:

```
> class(bacteria)
```

```
[1] "data.frame"
> names(bacteria)
[1] "t"  "Nt"
> dim(bacteria)
[1] 15  2
> summary(bacteria)
       t              Nt
 Min.   : 1.0   Min.   : 15.0
 1st Qu.: 4.5   1st Qu.: 34.0
 Median : 8.0   Median : 60.0
 Mean   : 8.0   Mean   :103.9
 3rd Qu.:11.5   3rd Qu.:154.0
 Max.   :15.0   Max.   :355.0
```

The plots in Figure 2.1 can be produced with the `plot` command:

```
> plot(bacteria)
> plot(bacteria$t, bacteria$Nt, xlab="Time (t)",
  ylab="Bacteria Count (Nt)")
> plot(bacteria$t, log(bacteria$Nt),xlab="Time(t)",ylab="ln(Nt)")
```

The first `plot` command plots the first variable in the data frame on the x-axis and the second on the y-axis. Alternatively, we could reference individual variables within a data frame by appending a `$` and the variable name to the data frame name. We can apply functions such as `log` within the `plot` command. Axis labels are optional, and can be specified with the `xlab` and `ylab` arguments.

Pearson correlations are computed by default with the `cor` command, and Spearman correlations with the `method="spearman"` argument. As with the `plot` command, we can apply functions to variables within a call to the `cor` function.

```
> cor(bacteria)
            t         Nt
t   1.0000000 -0.9074223
Nt -0.9074223  1.0000000
> cor(bacteria, method="spearman")
    t Nt
t   1 -1
Nt -1  1
> cor(bacteria$t, log(bacteria$Nt))
[1] -0.9941623
```

Regression models can be estimated with the `lm` function. The first argument should be a formula that specifies the model. Formulas use the $\sim$ symbol instead of the "=" sign:

```
fit = lm(log(Nt) ~ t, bacteria)
```

As with the `plot` command, we can apply functions such as the `log` within a formula. The second argument gives the data frame. The fitted model is stored in an object called `fit` and is a `list` of class `lm`. Lists in R can contain many different objects, which are displayed with the `names` function:

```
> class(fit)
[1] "lm"
> names(fit)
 [1] "coefficients"  "residuals"     "effects"       "rank"
```

```
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> fit$coefficients
(Intercept)           t
  5.9731603  -0.2184253
> fit$call
lm(formula = log(Nt) ~ t, data = bacteria)
```

Different components can be accessed with a $ sign, as we did with data frames. The coefficients match those computed in Example 2.4. Residuals are stored in the vector `fit$residuals`, fitted values in `fit$fitted.values`, etc. A more common way to display the results of a regression model is to apply different functions to a fitted object. For example, `summary` displays the parameter estimates, standard errors, t-statistics, P-values, R^2 and the ANOVA F-test:

```
> summary(fit)

Call:
lm(formula = log(Nt) ~ t, data = bacteria)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.973160   0.059778   99.92  < 2e-16 ***
t           -0.218425   0.006575  -33.22 5.86e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.11 on 13 degrees of freedom
Multiple R-squared:  0.9884,Adjusted R-squared:  0.9875
F-statistic:  1104 on 1 and 13 DF,  p-value: 5.86e-14
```

This illustrates how R is object oriented. The summary function detects the class of an object and does something appropriate. In the case of a data frame `summary` computes the five number summary and mean for each variable. For the objects of class `lm` it computes the above output. Likewise, the `plot` function detects the class, produces a scatter plot for data frames, and different diagnostic plots for the objects of class `lm`, including the residual plot in Figure 2.3. We can superimpose the fitted regression line on a scatter plot with the `abline` function:

```
> plot(fit)        # produces diagnostic plots
> plot(bacteria$t, log(bacteria$Nt), xlab="Time (t)", ylab="ln(Nt)")
> abline(fit)      # superimpose fitted regression line
```

The `summary` function displays the residual standard error and the results of the F-test. A more detailed ANOVA table can be displayed with the `anova` function:

```
> anova(fit)
Analysis of Variance Table

Response: log(Nt)
          Df  Sum Sq Mean Sq F value   Pr(>F)
t          1 13.3587 13.3587  1103.7 5.86e-14 ***
Residuals 13  0.1573  0.0121
```

Confidence intervals for the β's are computed with the `confint` function:

```
> confint(fit)
                 2.5 %     97.5 %
(Intercept)  5.8440175  6.1023030
t           -0.2326291 -0.2042214
```

Predicted values for the data used to estimate the model are stored in `fit$fitted.values`. The model can be applied to other data with the `predict` function, e.g., to predict $\ln(N_t)$ for $t = 6.5$ we use

```
> predict(fit, data.frame(t=6.5))
       1
4.553396
```

Prediction intervals are computed by adding the argument `interval="pred"` and confidence intervals are computed by adding the argument `interval="conf"`.

```
> predict(fit, data.frame(t=6.5), interval="pred")
       fit      lwr      upr
1 4.553396 4.307003 4.799789
```

We show how to use R to fit the multiple regression model from Example 3.13. The `lm` function in R fits this model as shown in the R code below.

```
> fit = lm(log(price, base=10) ~ I(mileage/1000) + cylinder + liter + cruise
    + sound + leather + make + type, bluetrain)
```

The additional `base=10` argument to the `log` function changes the base of the logs. The expression `I(mileage/1000)` converts mileage to thousands; the `I()` indicates that R should interpret the `/` symbol as division.

```
> summary(fit)

Call:
lm(formula = log(price, base = 10) ~ I(mileage/1000) + cylinder +
    liter + cruise + sound + leather + make + type, data = bluetrain)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.1883262  0.0222367 188.352  < 2e-16 ***
I(mileage/1000) -0.0035341  0.0002484 -14.227  < 2e-16 ***
cylinder        -0.0127217  0.0071063  -1.790   0.0742 .
liter            0.1094932  0.0079280  13.811  < 2e-16 ***
cruise           0.0096903  0.0059557   1.627   0.1045
sound            0.0037310  0.0046581   0.801   0.4236
leather          0.0081627  0.0048177   1.694   0.0910 .
makeCadillac     0.2004296  0.0109327  18.333  < 2e-16 ***
makeChevrolet   -0.0575070  0.0080661  -7.129 5.00e-12 ***
makePontiac     -0.0402394  0.0082032  -4.905 1.38e-06 ***
makeSAAB         0.2357121  0.0102830  22.922  < 2e-16 ***
makeSaturn      -0.0450062  0.0106617  -4.221 3.03e-05 ***
typeCoupe       -0.1403903  0.0106495 -13.183  < 2e-16 ***
typeHatchback   -0.1536999  0.0124206 -12.375  < 2e-16 ***
typeSedan       -0.1436521  0.0092084 -15.600  < 2e-16 ***
typeWagon       -0.0730266  0.0114614  -6.372 5.33e-10 ***
---
```

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.03984 on 386 degrees of freedom
Multiple R-squared:  0.9526,Adjusted R-squared:  0.9507
F-statistic: 516.8 on 15 and 386 DF,  p-value: < 2.2e-16
```

APPENDIX D

PROJECTS

A data analysis project involving a large data set should be an integral part of any course based on this book. I generally assign a team project for teams of 3-4 students (the same project to all teams) at just past the midpoint of the course when multiple regression and logistic regression have been covered in the class. The project involves applying these two methodologies to the data set. Two sample projects are described in the following.

D.1 Catalog Sales Project 1

- **Business Situation**: A retail company sells upscale clothing on its website and via catalogs, which help drive customers to the website. All customers in the data file were sent a catalog mailing in early fall 2012 and purchases made by them during fall 2012 were recorded. There is one row for each customer. The `targdol` is the response variable, which is the purchase amount during fall 2012; `targdol` $= 0$ indicates that the customer did not make a purchase, i.e., the customer was a non-respondent. The remainder of variables are potential predictor variables which give information about the customer as of the time of the mailing. The data are stored in the file `catalog sales data.csv`.
- **Data**: There are a total 101,532 customers, of whom 9571 (9.43%) are respondents, i.e., they have `targdol` > 0. The data are randomly divided into a training set with 50418 observations and the remaining 51,114 into a test set. (`train` is the indicator

variable for training or test set: `train` =1 for the training set, `train` =0 for the test set). The definitions of the variables are as follows.

Variable	Description
targdol	dollar purchase resulting from catalog mailing
datead6	date added to file
datelp6	date of last purchase
lpuryear	latest purchase year
slstyr	sales (\$) this year
slslyr	sales (\$) last year
sls2ago	sales (\$) 2 years ago
sls3ago	sales (\$) 3 years ago
slshist	LTD dollars
ordtyr	number of orders this year
ordlyr	number of orders last year
ord2ago	number of orders 2 years ago
ord3ago	number of orders 3 years ago
ordhist	LTD orders
falord	LTD fall orders
sprord	LTD spring orders
train	training/test set indicator (1 = training, 0 = test)

LTD means "life-to-date," i.e., since the customer purchased for the first time.

- **Goal**: Build a predictive model for `targdol` based on the training set and then test it on the test set.
- **Strategy for Building Predictive Models**: Straightforward multiple regression will not work since more than 90% are non-respondents with `targdol` $= 0$. A two-step model fitting approach (logistic regression followed by multiple regression) is recommended using the training set.
  1. Based on preliminary analyses, transform the data and include any interactions as appropriate.
  2. First develop a binary logistic regression model for $y > 0$, where `targdol` is denoted by y.
  3. Next develop a multiple regression model using data with $y > 0$ only. The resulting model will enable us to estimate the conditional expectation $E(y|y > 0)$ for given values of predictors. To address the right skew in `targdol`, log-transform it.
- **Evaluating the Fitted Models**: Fit two oe three candidate models using the above approach. These models should meet the usual criteria such as significant coefficients, satisfactory residual plots, good fit as measured by R^2 or R^2_{adj}, parsimony and interpretability of the model etc. The final choice of the model depends on how well it is able to predict `targdol` values for the test set. To calculate the predicted values of `targdol` do the following steps.

1. Use the binary logistic regression model fitted to the training set to estimate the probability of being a responder for each customer in the test set, which is the probability $P(y > 0)$.

2. Next, for each observation in the test set calculate the predicted values of y using the multiple regression model fitted to the training set with $y > 0$. So these predicted values actually give estimates of $E(y|y > 0)$.

3. For each observation in the test set multiply the predicted y by the estimated probability $P(y > 0)$ to obtain the estimate of the unconditional expectation of y using the formula
$$E(y) = E(y|y > 0)P(y > 0).$$
This formula gives the predicted values $\widehat{y}_i$ for observations in the test set. These predicted values can be used to evaluate the two numerical criteria (and possibly others) to evaluate the fitted models on the test set.

   Statistical Criterion : Mean square error of prediction (MSEP) $= \left[\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right]/[n-(p+1)]$; $p+1$ is the number of β coefficients in the multiple regression model derived from the training set.

   Financial Criterion : Select the top 1000 customers (prospects) from the test set who have the highest E(`targdol`). Then find their total actual purchases. This is the payoff and should be as high as possible.

- **Hints**:

1. These data are dirty and you will have much cleaning up to do. Some errors in the data are as follows. If you run a histogram or frequency distribution of the `datelp6` variable among only those with `targdol` > 0 you will see that, for the most part, `datelp6` equals one of two distinct dates in the calendar year. It is as if the person who prepared the data binned them into six-month bins. There are also other inconsistencies in the data, e.g., `falord` + `sprord` is not equal to `ordhist` in about 9% of the cases. Similarly, the year of the latest purchase obtained from `lpuryear` variable and from `datelp6` variable do not always agree. Some of these errors result because when two variables measure the same thing, both are not updated.

2. There are other inconsistencies in the data as well. For example, in a few cases the number of orders are not recorded but there are sales amounts, date added to file does not match with date of last purchase, etc. You will need to make reasonable decisions to resolve the inconsistencies. Just clearly state in your report how you resolved the inconsistencies.

3. It is known in data base marketing that generally the best predictors for deciding whether a customer will respond to a catalog are (1) recency of last purchase and (2) consistency of past purchases. Recency can be readily deduced from the date of last purchase. Consistency can be coded as an interaction of the last 1, 2 or 3 years of sales or orders. You would need to create such interaction variables.

4. The significant predictors for the logistic regression model will be generally different from those for the multiple regression model.

D.2 Catalog Sales Project 2

Here is another catalog sales data set which can be modeled following the same strategy as described above for Project 1. The same criteria may be used for model validation and evaluation. The response variable for this data set is `targamnt`. The total sample size for this data set is 106,284 of which 5698 are respondents (5.36%), i.e., have `targamnt` > 0. The data set is divided randomly into a training set (`training.csv`) consisting of 52,844 observations and test set (`test.csv`) consisting of 53,440 observations. The descriptions of the variables are as follows.

```
-----------------------------------------
Variable  Description
-----------------------------------------
recmon    Months since last order
ordcls1   5 Year Product Class 1 Orders
ordcls2   5 Year Product Class 2 Orders
ordcls3   5 Year Product Class 3 Orders
ordcls4   5 Year Product Class 4 Orders
ordcls5   5 Year Product Class 5 Orders
ordcls6   5 Year Product Class 6 Orders
ordcls7   5 Year Product Class 7 Orders
salcls1   5 Year Sales Product Class 1
salcls2   5 Year Sales Product Class 2
salcls3   5 Year Sales Product Class 3
salcls4   5 Year Sales Product Class 4
salcls5   5 Year Sales Product Class 5
salcls6   5 Year Sales Product Class 6
salcls7   5 Year Sales Product Class 7
ord185    Order Yr 1, Prom 85 (Y/N)
ord285    Order Yr 2, Prom 85 (Y/N)
ord385    Order Yr 3, Prom 85 (Y/N)
ord485    Order Yr 4, Prom 85 (Y/N)
tof       Time on File
totord    Lifetime Orders
totsale   Lifetime Sales
targamnt  Prom 85 Sales in Targ Wndw
-----------------------------------------
```

You should consider creating new predictor variables from the given set. For example, `aoa` = average order amount = `totsale/totord` or `pr` = purchase rate = `totord/tof` may be good predictors of `targamnt` in the multiple regression model. (Note that `tof` measures how long someone has been a customer.) There may be other "interaction" variables that are good predictors.

APPENDIX E

REFERENCES

1. Allison, P. D. (2008), "Convergence failures in logistic regression," *SAS Global Forum*, Paper No. 360-2008.

2. Anscombe, F. J. (1973),"Graphs in statistical analysis," *The American Statistician*, **27**, 17-21.

3. Box, G. E. P. and Cox, D. R. (1964), "An analysis of transformation (with discussion)," *Journal of the Royal Statistical Society, Ser, B*, **26**, 211-252.

4. Chatterjee S. and Hadi, A. S. (2012), *Regression Analysis by Example*, 5th edition, New York: Wiley.

5. Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187-220.

6. Crowley, J. and Hu, M. (1977), "Covariance analysis of heart transplant survival data," *Journal of the American Statistical Association*, **72**, 27-36.

7. Cule, E., Vineis, P. and De lorio, M. (2011), "Significance testing in ridge regression for genetic data," *BMC Bioinformatics*, doi 10.1186/147-2105-12-372.

8. Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd edition, New York: Wiley.

9. Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*,**7**, 179188.

10. Hamilton, D. J. (1987),"Sometimes $R^2 > r^2_{y \cdot x_1} + r^2_{y \cdot x_2}$, correlated variables are not always redundant," *The American Statistician*, **41**, 129-132.

11. Hardin, J. W. and Hilbe, J. M. (2012), *Generalized Linear Models and Extension*, 3rd edition, College Station, TX: Stata Press.

12. Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, 2nd edition, New York: Springer.

13. Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press.

14. Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: Wiley.

15. Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Biased estimation for non-orthogonal problems," *Technometrics*, **12**, 69-82.

16. Hosmer, D. W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley.

17. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, New York: Springer.

18. Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, 5th edition, Prentice Hall: Upper Saddle River, NJ.

19. Koch, G.G., Atkinson, S.S. and Stokes, M.E. (1986), "Poisson regression," *Encyclopedia of Statistical Sciences*, **7**, (edited by N.L. Johnson and S. Kotz), New York: Wiley, 32-42.

20. Kuiper, S. (2008), "Introduction to multiple regression: How much is your car worth?," *Journal of Statistics Education*, **16 (3)**
(http://www.amstat.org/publications/jse/v16n3/datasets.kuiper.html).

21. Kutner, M. H., Nachtsheim, C. J., Neter, J., Wasserman, W. and Li, W. (2005), *Applied Linear Statistical Models*, 5th edition, New York: McGraw-Hill/Irwin

22. Mallows, C. L. (1973), "Some comments on C_p," *Technometrics*, **15**, 661-673.

23. Marquardt, D. W. and Snee, R. D. (1975), "Ridge regression in practice," *The American Statistician*, **29**, 3-20.

24. McClave, J. T. and Dietrich, F. H. (1994), *Statistics*, 6th edition, New York: Dellen-MacMillan.

25. McClintock, S., Stangl, D. and Cetinkya-Rundel, M. (2014), "The real secret to genius? Reading between lines," *Chance*, **27**, 41-44.

26. McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edition, Chapman & Hall: London.

27. Messerli, F.H. (2012), "Chocolate consumption, cognitive function and Nobel laureates," *New England Journal of Medicine*, **367**, 1562-1564.

28. Mevik, B-H and Wehrens, R. (2007), "The pls package: Principal component and partial least squares regression in R," *Journal of Statistical Software*, **18 (2)**, 1-23.

29. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012), *Introduction to Linear Regression Analysis*, 5th edition, New York: Wiley.

30. Myers, R. H., Montgomery, D. C. and Vining, G. G. (2010), *Generalized Linear Models: With Applications in Engineering and the Sciences*, 1st edition, New York: Wiley.

31. Rossi, P. H., Berk, R. A. and Lenihan, K. J. (1980). *Money, Work and Crime: Some Experimental Results*, New York: Academic Press.

32. Tableman, M.; and Kim, J. S. (2003), *Survival Analysis Using S*, Chapman and Hall/CRC

33. Tamhane, A. C. and Dunlop, D. D. (2000), *Statistics and Data Analysis: From Elementary to Intermediate*, Prentice Hall: Upper Saddle River, NJ.

34. Tanner, M. (1996), *Tools for Statistical Inference*, 3rd edition, New York: Springer.

35. Therneu, T., Crowson, C., and Atkinson, E. (2018): "Using time dependent covariates and time dependent coefficients in the Cox model," (`https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf`).

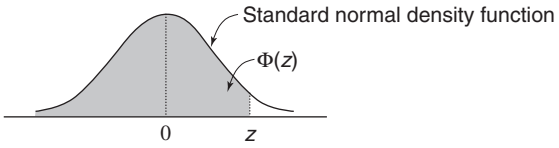
36. Tibshirani, R. (1996), "Regression shrinking and selection via the lasso," *Journal of the Royal Statistical Society, Ser. B*, **58**, 267-288.

37. Tversky, A. and Kahnemann, D. (1973), "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, **5**, 207-232.

38. Webster, J. T., Gunst, R. F. and Mason, R. L. (1974), "Latent root regression analysis," *Technometrics*, **16**, 513-522.

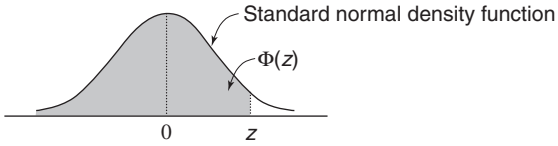
39. Wold, H. (1966), "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis* (ed. P. R. Krishniah), pp. 391-420, New York: Academic Press.

Table C.1 Standard Normal c.d.f. $\Phi(z) = P(Z \leq z)$



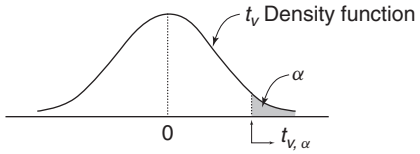
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0394	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0,0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table C.1 Standard Normal c.d.f. $\Phi(z) = P(Z \leq z)$ (*Continued*)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0,4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

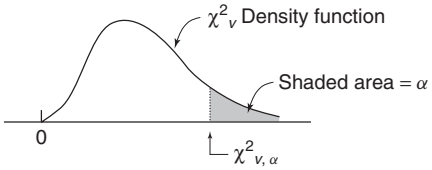
Table C.2 Critical Values $t_{\nu,\alpha}$ for the t-Distribution



	α						
ν	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Source: Reprinted with permission of Pearson Education, Inc.

Table C.3 Critical Values $\chi^2_{\nu,\alpha}$ for Chi-Square Distribution

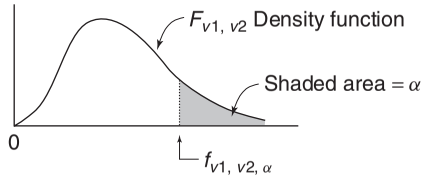


	α									
ν	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.420	65.473
40[a]	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

[a]For $\nu > 40$, $x^2_{\nu,\alpha} \simeq \nu\left(1 - \frac{2}{9\nu} + z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3$.

Source: Reprinted with permission of Pearson Education, Inc.

Table C.4 Critical Values $f_{v_1,v_2,\alpha}$ for F-Distribution

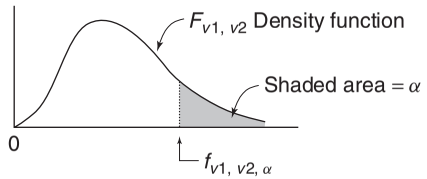


$\alpha = 0.01$

Degrees of freedom for denominator (v_2)	Degrees of freedom for numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052.0	4999.5	5403.0	5625.0	5764.0	5859.0	5928.0	5982.0	6022.0	6056.0	6106.0	6157.0	6209.0	6235.0	6261.0	6287.0	6311.0	6339.0	6366.0
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	9.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	9.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.96	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	1.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49

20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	
22	7.95	5.72	4.81	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38	
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.51	1.47	1.32	1.00	

Table C.4 Critical Values $f_{\nu_1,\nu_2,\alpha}$ for F-Distribution (*Continued*)



		Degrees of freedom for numerator (ν_1)																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
$\alpha = 0.05$																				
Degrees of freedom for denominator (ν_2)	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.49	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.69	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88

20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.81	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.91	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.59	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.81	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00