MATH 484-564 Regression

Version Sept 22nd 2024

Contents

Chapter 1

Introduction

1.1 What is Statistical Learning?

Suppose we are the (highly paid) consultants hired by a client to investigate the association between advertising and sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with the advertising budgets for the product in each markets for the three different media: TV, radio, and newspaper. The data are displayed in Fig 1.1 below.

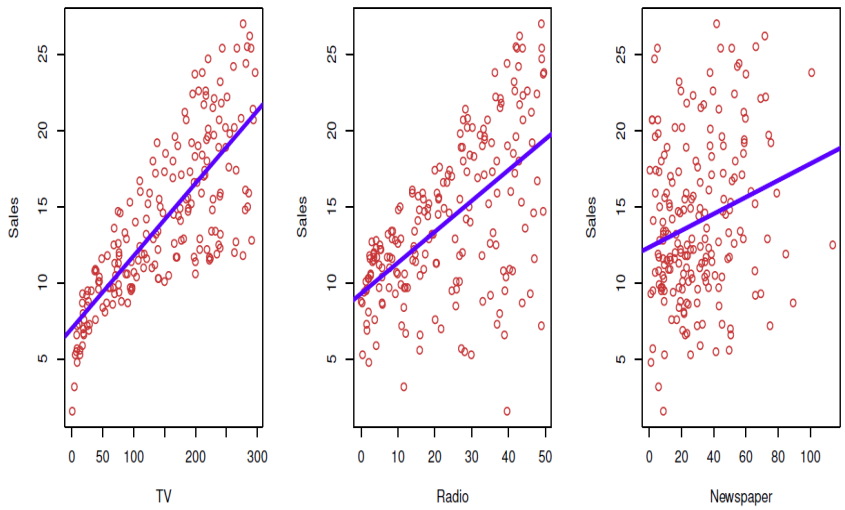


Figure 1.1: The figure above is the Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable, as describe later in the course. In other words, each blue line represents a simple model that can be used to predict sales, using TV, radio, and newspaper, respectively

.

We have been told that it is not possible for our client to directly increase sales of the product. However, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increase sales.

In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

In this setup, the advertising budgets are *input variables* while sales is an *output variable*. We will denote the input variables using the symbol X, with a subscript to distinguish them. So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget.

The inputs go by many different names, such as *predictors, independent variables, features* or sometimes just *variables*.

The output variable, in this case, sales, is often called the *response* or *dependent variable*, and typically denoted by Y. Throughout this note, we will be using all of these terms interchangeably.

More generally, assume that we observe a quantitative response Y and p different predictors, $X_1, X_2, \cdots, X_p$. Suppose that there is some relationship between Y and $X = (X_1, X_2, \cdots, X_p)$, which can be written in the form

$$Y = f(X) + \epsilon \tag{1.1}$$

Here f is some fixed but unknown function of $X_1, X_2, \cdots, X_p$, and ϵ is a random *error term*, which is independent of X and has mean zero.

In this formulation, f represent the *sytematic* information that X provides about Y.

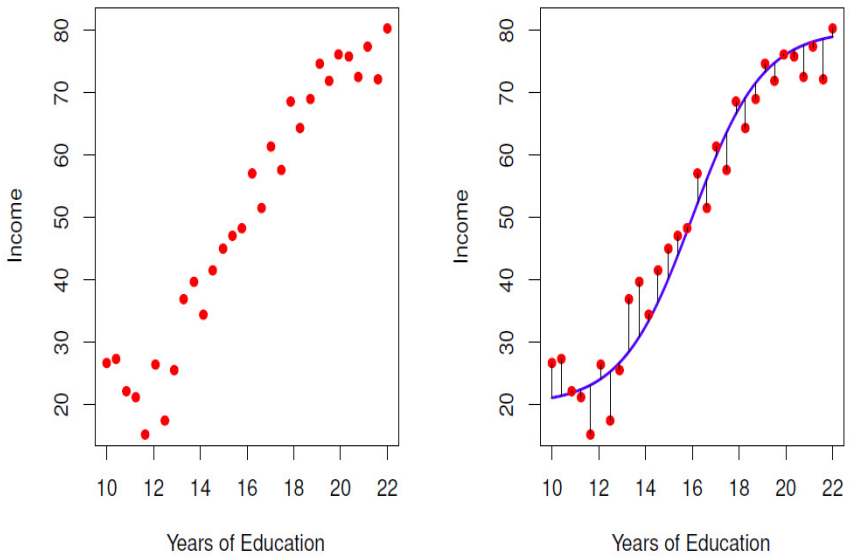


Figure 1.2: The figure above relates to the Income data set. Left: The red dots are the observed values of income (in thousands of dollars) and years of education for 30 individuals. Right: The blue curve represents the true underlying relationship between income and years of education. Which is generally unknown (but is known in this case because the data is simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.

As an example, consider the left-hand panel of Fig 1.2, a plot of income vs years of education for 30 individuals in the Income data set. The plot suggests that one might be able to predict income using years of education. However, the function f that connects the input variable to the output variable is in general unknown.

In this situation one must estimate f based on the observed points. Since Income is a simulated data set, f is known and is shown by the blue curve in the right-hand panel of Fig 1.2. We note that some of the 30 observations lie above the blue curve and some lie below it; overall, the errors have approximately mean zero.

In general, the function f may involve more than one input variable. In Fig 1.3 we plot income as a function of years of education and seniority. Here f is two-dimensional surface that must be estimated based on the observed data.

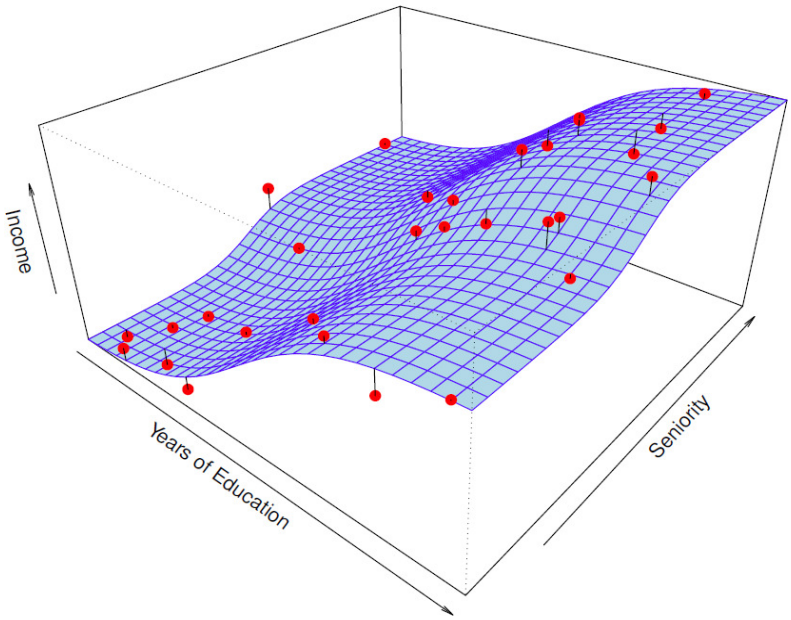


Figure 1.3: The figure above displays income as a function of years of education and seniority in the Income data set. The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

In this class, we will be studying how to estimate the function f based on the data given.

1.2 Why Estimate f?

There are two main reasons that we may wish to estimate f, namely, *prediction* and *inference*. We will discuss each in turn.

1.2.1 Prediction

In many situation, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\widehat{Y} = \widehat{f}(X) \tag{1.2}$$

where $\widehat{f}$ represents our estimation for f, and $\widehat{Y}$ represents the resulting prediction for Y. In this setting, $\widehat{f}$ is often treated as a black-box, in the sense that one is not typically concerned with the exact form of $\widehat{f}$, provided that it yields accurate predictions for Y.

As an example, suppose that $X_1, \cdots, X_p$ are characteristics of a patient's blood sample that can be easily measured in a lab, and Y is a variable encoding the patient's risk for a severe adverse reaction to a particular drug. It is natural to seek to predict Y using X, since we can then avoid giving the drug in question to patients who are at high risk of an adverse reaction, that is, patients for who the estimate of Y is high.

The accuracy of $\widehat{Y}$ as a prediction for Y depends on two quantities which we will called the *reducible error* and *irreducible error*. In general, $\widehat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is *reducible* because we can potentially improve the accuracy of $\widehat{f}$ by using the most appropriate statistical learning technique to estimate f. However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\widehat{Y} = f(X)$, our prediction would still have some error in it. This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X. Therefore, variability associated with ϵ also affects the accuracy of our prediction. This is known as *irreducible* error, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ.

Let us see why the irreducible error is larger than zero. The quantity ϵ may contain unmeasured variables that are useful in predicting Y: since we don't measure them, f cannot use them for prediction. For example, the risk of an adverse reaction, Y, might vary for a given patient on a given day, depending on manufacturing variation in the drug itself or the patient's general feeling of well-being on that day. Which we are not able to set as one of our input variables X.

Consider a given estimate $\widehat{f}$ and a set of predictors X, which yields the prediction $\widehat{Y} = \widehat{f}(X)$. Assume for a moment that both $\widehat{f}$ and X are fixed, so that the only variability comes from ϵ. Then we have the following

$$\begin{aligned} E(Y - \widehat{Y})^2 &= E[f(X) + \epsilon - \widehat{f}(X)]^2 \\ &= E\left[(f(X) - \widehat{f}(X))^2 + 2(f(X) - \widehat{f}(X))\epsilon + \epsilon^2\right] \\ &= E\left[(f(X) - \widehat{f}(X))^2\right] + 2(f(X) - \widehat{f}(X))E\left[\epsilon\right] + E\left[\epsilon^2\right] \\ &= \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned} \tag{1.3}$$

as $E(\epsilon) = 0$, hence $E(\epsilon^2) = \text{Var}(\epsilon)$, where $\text{Var}(\epsilon)$ represent the variance associated with the error term ϵ.

From above we see that the expected value (or average) of the squared difference between the predicted and actual value of Y (i.e. the term $E(Y - \widehat{Y})^2$) can be written as the sum of reducible and irriducible terms.

The focus of this course is on techniques for estimating f with the aim of minimizing the reducible errors.

It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. In the sense that even if we can make the term $\left[f(X) - \widehat{f}(X)\right]^2$ in Eq 1.3 vanishes, the term $E(Y - \widehat{Y})^2$ still not able to be made small, thanks to the irreducible term $\text{Var}(\epsilon)$.

1.2.2 Inference

We are often interested in understanding the association between Y and $X_1, \cdots, X_p$. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. Suppose now, we do not want to treat $\widehat{f}$ as a black box and wants to know its exact form. In this setting, one may be interested in answering the following questions:

- *Which predictors are associated with the response?* It is often the case that only a small fraction of the available predictors are substantially associated with Y. Identifying the few important predictors among a larger set of possible variables can be extremely useful, depending on the application.
- *What is the relationship between the response and each predictor?* Some predictors may have a positive relationship with Y, in the sense that larger values of the predictor are associated with larger values of Y. Other predictors may have the opposite relationship. Depending on the complexity of f, the relationship between the response and a given predictor may also depend on the values of the other predictors.
- *Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?* Historically, most methods for estimating f have taken a linear form. In some situation, such an assumption is reasonable or even desirable. But often the true relationship is more complicated, in which case a linear model may not provide an accurate representation of the relationship between the input and output variables.

In our class, we will see a number of examples that fall into the prediction setting, the inference setting, or a combination of the two.

As an example, consider a company that is interested in conducting a direct-marketing campaign. The goal is to identify individuals who are likely to respond positively to a mailing, based on observations of demographic variables measured on each individual. In this case, the demographic variables serve as predictors, and response to marketing campaign (either positive or negative) serves as the outcome. The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants to accurately predict the response using the predictors. This is an example of modeling for prediction.

In contrast, consider the Advertising data illustrated in Fig 1.1. One may be interested in answering questions such as:

- Which media are associated with sales?
- Which media generate the biggest boost in sales? or
- How large of an increase in sales is associated with a given increase in TV advertising?

This situation falls into the inference paradigm.

Another example involves modeling the brand of a product that a customer might purchase based on variables such as price, store location, discount levels, competition price, and so forth. In this situation one might really be most interested in the association between each variable and the probability of purchase. For instance, *to what extent is the product's price associated with sales?* This is an example of modeling for inference.

Finally, some modeling could be conducted both for prediction and inference. For example, in the real estate setting, one may seek to relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, and so forth. In this case one might be interested in the association between each individual input variable and housing price. For instance, *how much extra will a house be worth if it has a view of the river?* This is an inference problem. Alternatively, one may simply be interested in predicting the value of a home given its characteristics: *is this house under- or over-valued?* This is a prediction problem.

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f maybe appropriate. For example, linear models allow for relatively simple and interpretable inference, but may not yield as accurate prediction as some other approaches. In contrast, some of the highly non-linear approaches that we discuss in the later part of the class can potentially provide quite accurate predictions for Y, but this comes at the expense of a less interpretable model for which the inference is more challenging.

1.3 How Do We Estimate f?

Throughout this class, we explore many linear and non-linear approaches for estimating f. However, these methods generally share certain characteristics. We provide an overview of these shared characteristics in this section.

We assume that we have observed a set of n different data points. For example, in Fig 1.2 we observed $n = 30$ data points. These observations are called *training data* because we will use these observations to train, or teach, our method how to find f.

Let x_{ij} represent the value of the jth predictor, or input, for observation i, where $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

Correspondingly, let y_i represent the response variable for the ith observation. Then our training data consists of

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

where $x_i = (x_{i1}, x_{12}, \cdots, x_{ip})^T$.

That is, in the long form, the training data look like:

$$\underbrace{x_{11}}_{\text{1st observrvation for the variable } x_1}, \cdots, \underbrace{x_{1p}}_{\text{1st observrvation for the variable } x_p} \quad \underbrace{y_1}_{\text{1st respond variable for the 1st observation}}$$

$$\underbrace{x_{21}}_{\text{2nd observrvation for the variable } x_1}, \cdots, \underbrace{x_{2p}}_{\text{2nd observrvation for the variable } x_p} \quad \underbrace{y_2}_{\text{2nd respond variable for the 2nd observation}}$$

all the way to

$$\underbrace{x_{n1}}_{\text{n-th observrvation for the variable } x_1}, \cdots, \underbrace{x_{np}}_{\text{n-th observrvation for the variable } x_p} \quad \underbrace{y_n}_{\text{n-th respond variable for the n-th observation}}$$

Our goal is to apply statistical learning method to the training data in order to estimate the unknown function f. In other words, we want to find a function $\widehat{f}$ such that $Y \approx \widehat{f}(X)$ for any observation (X, Y). Broadly speaking, most of the statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*. We now briefly discuss these two types of approaches.

1.3.1 Parametric Methods

Parametric methods involve a two-step model-based approach.

1. First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1.4}$$

   This is a *linear model*, which will be discussed extensively later in the course.

   Once we have assumed that f is linear, the problem of estimating f is greatly simplified. Instead of having to estimate an entirely p-dimensional function $f(X)$, one only needs to estimate the $p + 1$ coefficients, $\beta_0, \beta_1, \cdots, \beta_p$.

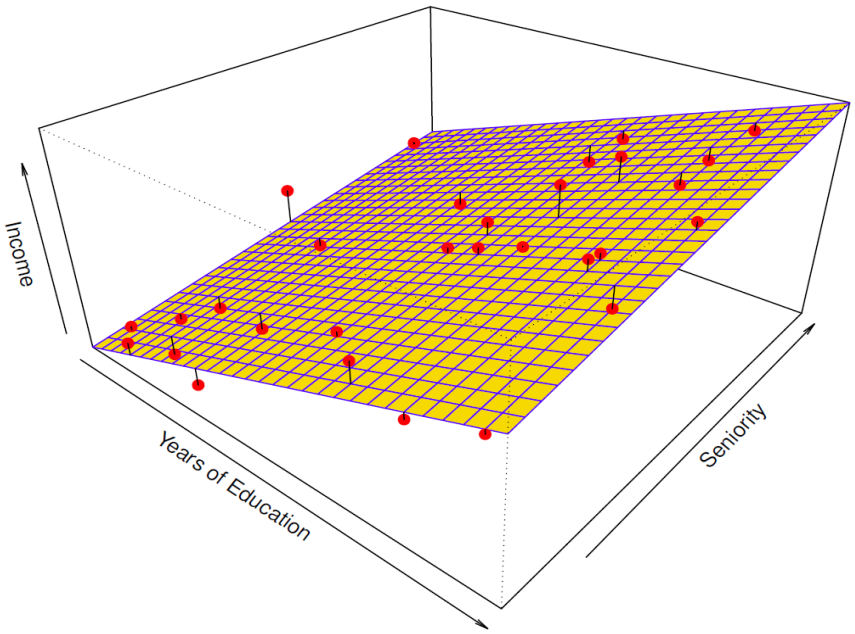


Figure 1.4: A linear model fit by least squares to the Income data from Fig 1.3. The observations are shown in red, and the yellow plane indicated the least squares fit to the data.

2. After a model has been selected, we need a procedure that uses the training data to *fit* or *train* the model. In the case of the linear model (1.4), we need to estimate the parameters $\beta_0, \beta_1, \cdots, \beta_p$. That is, we want to find values of these parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

   The most common approach to fitting the model (1.4) is referred to as *(ordinary) least squares*, which we will discuss later.

The model-based approach just described is referred to as *parametric*; it reduced the problem of estimating f down to one of estimating a set of parameters. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters, such as $\beta_0, \beta_1, \cdots, \beta_p$ in the linear model (1.4), than it is to fit an entirely arbitrary function f.

The potential disadvantage of a parametric approach is that the model we choose will usually not match the true known form of f. If the chosen model is too far from the true f, then our estimate will be poor. We can try to address this problem by choosing *flexible* models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as *overfitting* the data, which essentially means they follow the errors, or *noise* too closely. These issues are discussed throughout this course.

Figure 1.4 shows an example of the parametric approach applied to the Income data set from Fig 1.3. We have fit a liner model of the form

$$income \approx \beta_0 + \beta_1 \times education + \beta_2 \times seniority$$

Since we have assumed a linear relationship between the response and the two predictors, the entire fitting problem reduced to estimating β_0, β_1 and β_2, which we do using least squares linear regression.

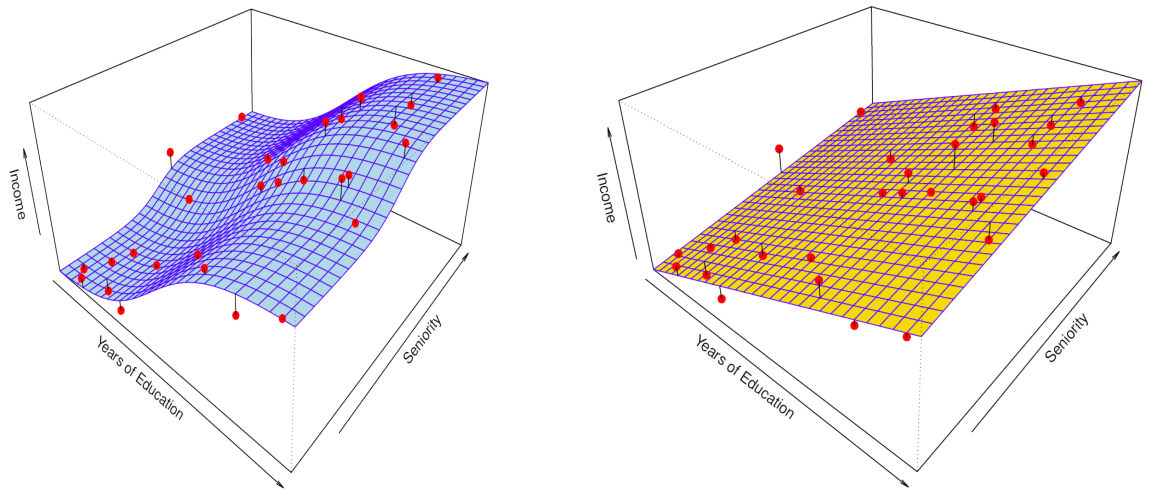


Figure 1.5: Note the Figure 1.3 on the left and Fig 1.4 on the right.

Comparing Figure 1.3 to Fig 1.4, we can see that the linear fit given in Fig 1.4 is not quite right: the true f has some curvature that is not captured in the linear fit. However, the linear fit still appears to do a reasonable job of capturing the positive relationship between years of education and income, as well as the slightly less positive relationship between seniority and income. It may be that with such a small number of observations, this is the best we can do.

1.3.2 Non-Parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.

But non-parametric approaches do suffer from a major disadvantages: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

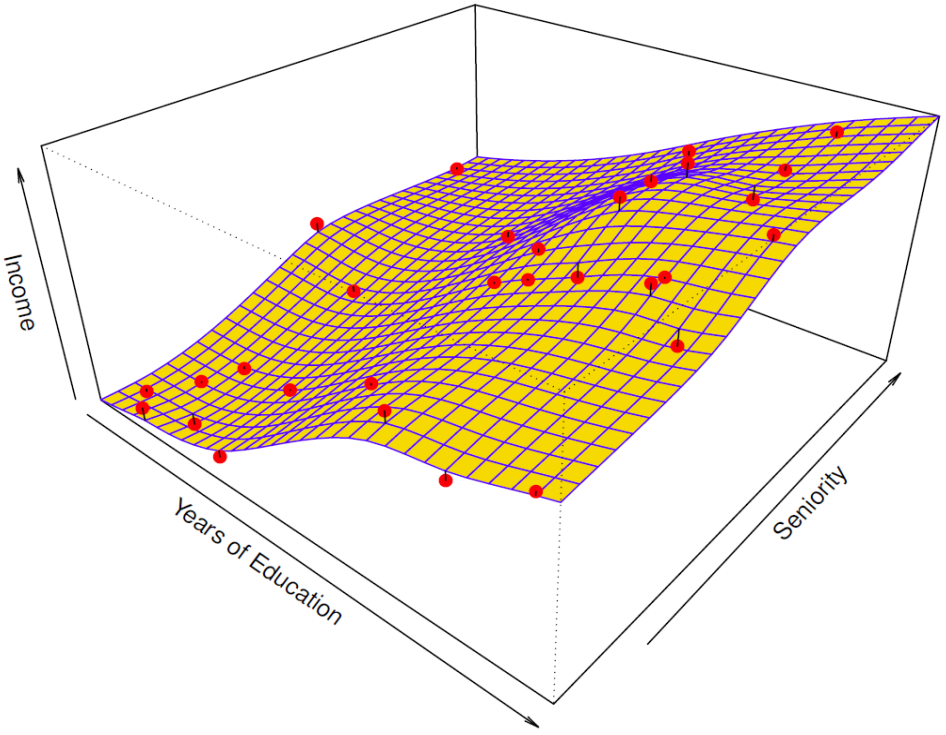


Figure 1.6: A rough thin-plate spline fit to the Income data from 1.3. This fit makes zero errors on the training data.

An example of a non-parametric approach to fitting the Income data is shown in Fig 1.6. A *thin-plate spline* is used to estimate f. This approach does not impose any pre-specified model on f. It instead attempts to produce an estimate for f that is as close as possible to the observed data, subject to the fit, that is, the yellow surface in Fig 1.6 being *smooth*. In this case, the non-parametric fit has produced a remarkably accurate estimate of the true f shown in Fig 1.3. In fact, the resulting estimate fits the observed data perfectly.

However, the spline fit shown in Fig 1.6 is far more variable than the true function f, from Fig 1.3. See the comparison of the figures below. This is an example of overfitting the data, which we discussed earlier.

Overfitting is an undesirable situation because the fit obtained will not yield accurate estimates of the response on new observations that were not part of the original training data set. Basically, overfitting occurs when a model learns the training data too well, capturing noise and random fluctuations in the data rather than the underlying relationships. As a result, the model performs poorly on new, unseen data because it has essentially memorized the training data rather than learning the general pattern.

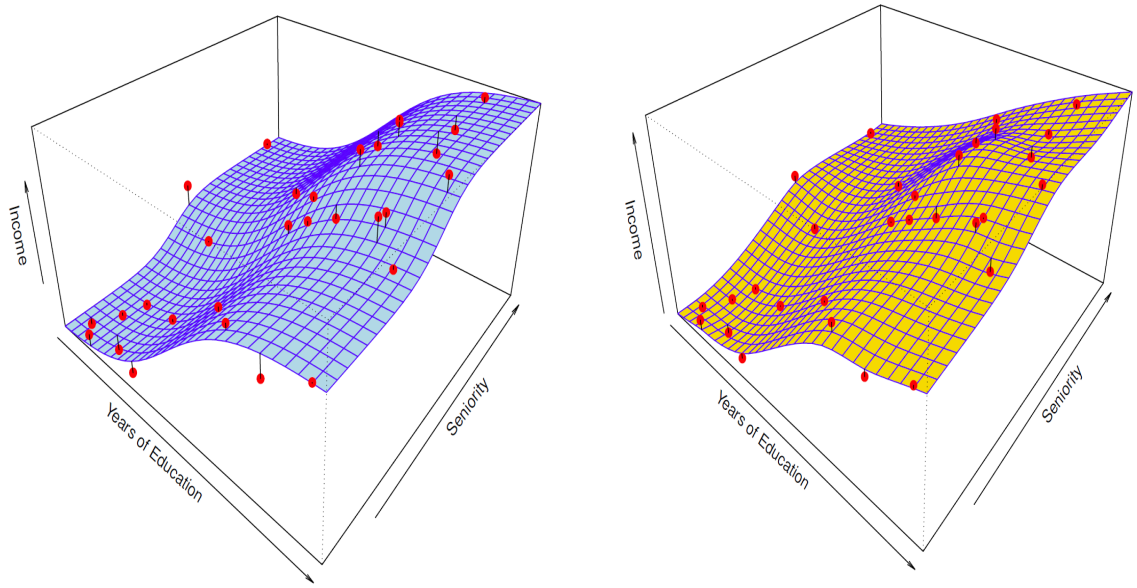


Figure 1.7: This is a comparison figure between the true function f, see Fig 1.3 on the left and the spline fit shown in Fig 1.6 on the right. Note the spline fit shown on the right is far more variable than the true function f shown on the left.

1.3.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

Of the many methods that we will examine in this course, some are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes to estimate f. For example, linear regression is a relatively inflexible approach, because it can only generate linear functions such as lines, Fig 1.1 or plane, Fig 1.4. Other methods, such as the thin plate splines shown in Fig 1.6 are considerably more flexible because they can generate a much wider range of possible shapes to estimate f.

One might reasonably ask the following question: *why would we ever choose to use a more restrictive method instead of a very flexible approach?* There are several reasons that we might prefer a more restrictive model.

If we are mainly interested in inference, then restrictive models are much more interpretable. For example, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the replationship between Y and $X_1, X_2, \cdots, X_p$.

In contrast, very flexible approaches, such as the splines method can lead to more complicated estimates of f that is difficult to understand how any individual predictor is associated with the response.

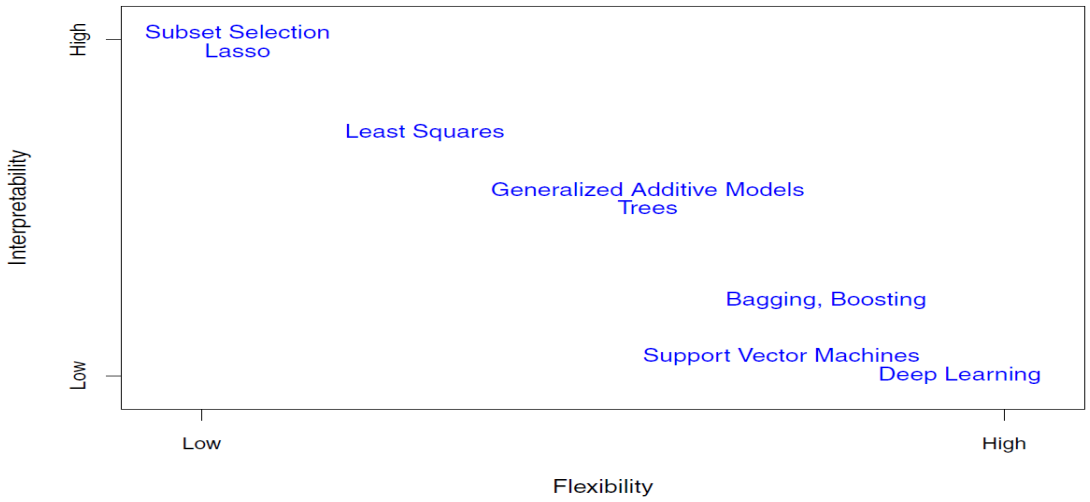


Figure 1.8: A representative of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of method increases, its interpretability decreases.

Fig 1.8 above provides an illustration of the trade-off between flexibility and interpretability for some of the methods that we will cover in this course.

1.3.4 Supervised Versus Unsupervised Learning

Most statistical learning problems fall into one of the two categories: *supervised* or *unsupervised*. The examples that we have discussed so far in this chapter all fall into the supervised learning domain. For each observation of the predictor measurement(s), $x_i, i = 1, \cdots, n$ there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference.) Many classical statistical learning methods such as linear regression and logistic regression, as well as more modern approaches such as GAM, boosting, and support vector machines, operate in the supervised learning domain. The vast majority of this course is devoted to this setting.

By contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation $i = 1, \cdots, n$, we observe a vector of measurements x_i but no associated response y_i. It is not possible to fit a linear regression model, since there is no response variable to predict. In this setting, we are in some sense working blind; the situation is referred to as *unsupervised* because we lack a response variable that can supervise our analysis.

In unsupervised learning, we can seek to understand the relationship between the variables or between observations. One statistical learning tool that we may use in this setting is *cluster analysis*, or clustering. The goal of cluster analysis is to ascertain, on the basis of $x_1, \cdots, x_n$, whether the observations fall into relatively distinct groups. For example, in a market segmentation study we might observe multiple characteristics (variables) for potential customers, such as zip code, family income, and shopping habits. We might believe that the customers fall into different groups, such as big spenders vs low spenders. If the information about each customer's spending patterns were available, then a supervised analysis would be possible. However, this information is not available, that is, we do not know whether each potential customer is a big spender or not. In this setting, we can try to cluster the customers on the basis of the variables

measured, in order to identify distinct groups of potential customers.

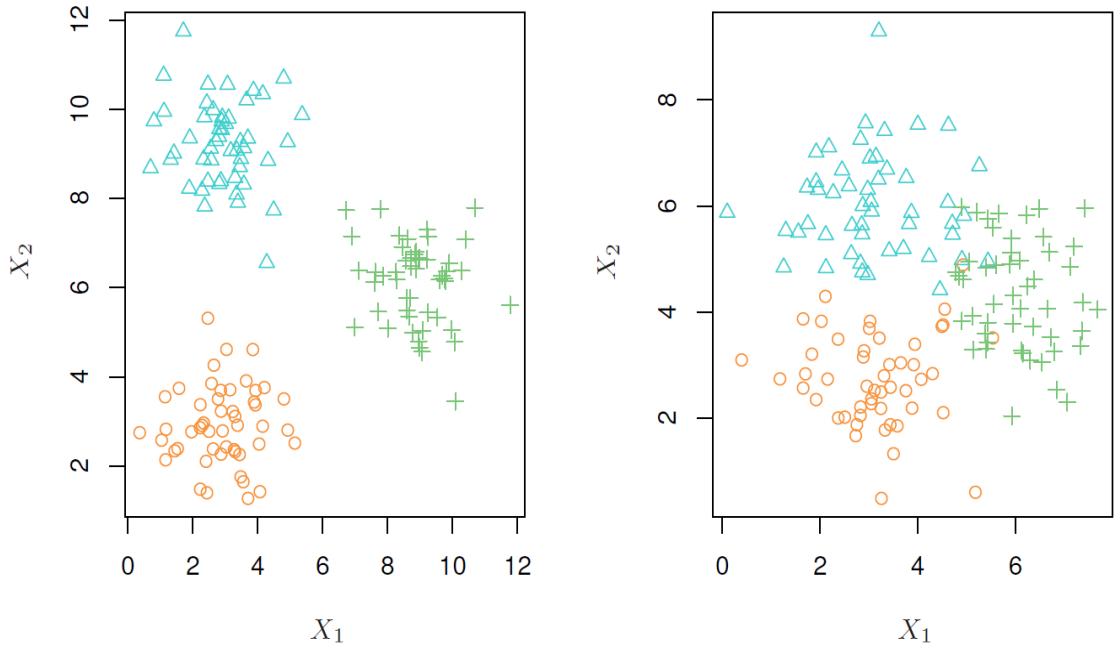


Figure 1.9: The figure shows a clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated in this setting, a clustering approach should succesfully identify the three groups. Right: There is some overlapping among the groups. Now the clustering is more challenging.

Fig 1.9 provides a simple illustration of the clustering problem. We have plotted 150 observations with measurements on two variables, X_1 and X_2. Each observation corresponds to one of three distinct groups. For illustrative purposes, we have plotted the members of each group using different colors and symbols. However, in practice the group memberships are unknown, and the goal is to determine the group to which each observation belongs. In the left-hand panel of Fig 1.9, this is a relatively easy task because the groups are well-separated. By contrast, the right-hand panel illustrates a more challenging setting in which there is some overlap between the groups. A clustering method could not be expected to assign all of the overlapping points to their correct group (blue, green, or orange).

In the example shown in Fig 1.9, there are only two variables, and so one can simply visually inspect the scatter plots of the observations in order to identify cluster. However, in practice, we often encounter data sets that contain many more than two variables. In this case, we cannot easily plot the observations.

1.3.5 Regression Versus Classification Problems

Variables can be characterized as either *quantitative* or *qualitative* (also known as *categorical*). Quantitative variables take on numerical values. Examples include a person's age, height, or income, the value of a house, and the price of a stock. In contrast, qualitative variables take on values in one of K different *classes*, or categories. Examples of qualitative variables include a person's marital status (married or not), the brand of a product purchased (Brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

We tend to refer to problems with a quantitative response as *regression* problem, while those involving a qualitative response are often referred to as *classification* problems. However, the distinction is not always that crisp. Least squares linear regression is used with a quantitative response, wheres logistic regression is typically used with a qualitative (two-class, or *binary*) response. Thus, despite its name, logistic regression is a classification method. But since it estimates class probability, it can be thought of as a regression method as well. Some statistical methods, such as K-nearest neighbors and boosting, can be used in the case of either quantitative or qualitative responses.

We tend to select statistical learning methods on the basis of whether the response is quantitative or qualitative; i.e. we might use linear regression when quantitative and logistic regression when qualitative. However, whether the *predictors* are qualitative or quantitative is generally considered less important. Most of the statistical learning methods can be applied regardless of the predictor variable type, provided that any qualitative predictors are properly *coded* before the analysis is performed.

1.4 Assessing Model Accuracy

In this course you will get a chance to see other statistical learning methods that extend beyond the standard regression approach. Why is it necessary to introduce so many different statistical learning approaches, rather than just a single *best* method? *There is no free lunch in statistics*: no one method dominated all others over all possible data sets. On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set. Hence it is an important task to decide for any given set of data which method produces the best results. Selecting the best approach can be one of the most challenging part of performing statistical leaning in practice.

In this section, we discuss some of the most important concepts that arise in selecting a statistical learning procedure for a specific data set. As the course progresses, we will explain how the concept presented here can be applied in practice.

1.4.1 Measuring the Quality of Fit

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data. That is, we need to quantify the extend to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly-used measure is the *mean squared error*(MSE), given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2, \tag{1.5}$$

where $\widehat{f}(x_i)$ is the prediction that $\widehat{f}$ gives for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if some of the observations, the predicted responses differ substantially.

The MSE in (1.5) is computed using the training data that was used to fit the model, and so should more accurately be referred to as the *training MSE*. But in general, we do not really care how well the method works on the training data. Rather, *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data*. Why

this is what we care about? Suppose that we are interested in developing an algorithm to predict a stock's price based on previous stock returns. We can train the method using stock returns from the past 6 months. But we don't really care how well our method predicts last week's stock price. We instead care about how well it will predict tomorrow's price or next month's price. On a similar note, suppose that we have clinical measurements (e.g. weight, blood pressure, height, age, family history of disease) for a number of patients, as well as information about whether each patient has diabetes. We can use these patients to train a statistical learning method to predict risk of diabetes based on clinical measurements. In practice, we want this method to accurately predict diabetes risk for *future patients* based on their clinical measurements. We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the method, since we already know which of those patients have diabetes.

To state it more mathematically, suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, and we obtain the estimate $\widehat{f}$. We can then compute $\widehat{f}(x_1), \widehat{f}(x_2), \cdots, \widehat{f}(x_n)$. If these are approximately equal to $y_1, y_2, \cdots, y_n$, then the training MSE given by (1.5) is small. However, we are really not interested in whether $\widehat{f}(x_i) \approx y_i$; instead, we want to know whether $\widehat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a *previously unseen test observation not used to train the statistical learning method*. We want to choose the method that gives the lowest *test* MSE, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

$$\text{Ave}\left(y_0 - \widehat{f}(x_0)\right)^2 \tag{1.6}$$

the average squared prediction error for these test observations (x_0, y_0). We'd like to select the model for which this quantity is as small as possible.

How can we go about trying to select a method that minimizes the test MSE? In some settings, we may have a test data set available, that is, we may have access to a set of observations that were not used to train the statistical learning method. We can then simply evaluate (1.6) on the test observations, and select the learning method for which the test MSE is smallest. But what if no test observations are available? In that case, one might imagine simply selecting a statistical learning method that minimizes the training MSE (1.5). This seems like it might be a sensible approach, since the training MSE and the test MSE appear to be closely related. Unfortunately, there is a fundamental problem with this strategy: there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. Roughly speaking, the problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE. For those methods, the training set MSE can be quite small, but the test MSE is often much larger.

Fig 1.10 illustrates this phenomenon on a simple example. In the left-hand panel of Fig 1.10, we have generated observations from $Y = f(X) + \epsilon$ with the true f given by the black curve. The orange, blue and green curves illustrate three possible estimates for f obtained using methods with increasing levels of flexibility. The orange line is the linear regression fit, which is relatively inflexible. The blue and green curves were produced using *smoothing splines*, with different level of smoothness. It is clear that as the level of flexibility increases, the curves fit the observed data more closely. The green curve is the most flexible and matches the data very well; however, we observe that it fits the true f (shown in black) poorly because it is too wiggly. By adjusting the level of flexibility of the smooth spline fit, we can produce many different fits to the data.

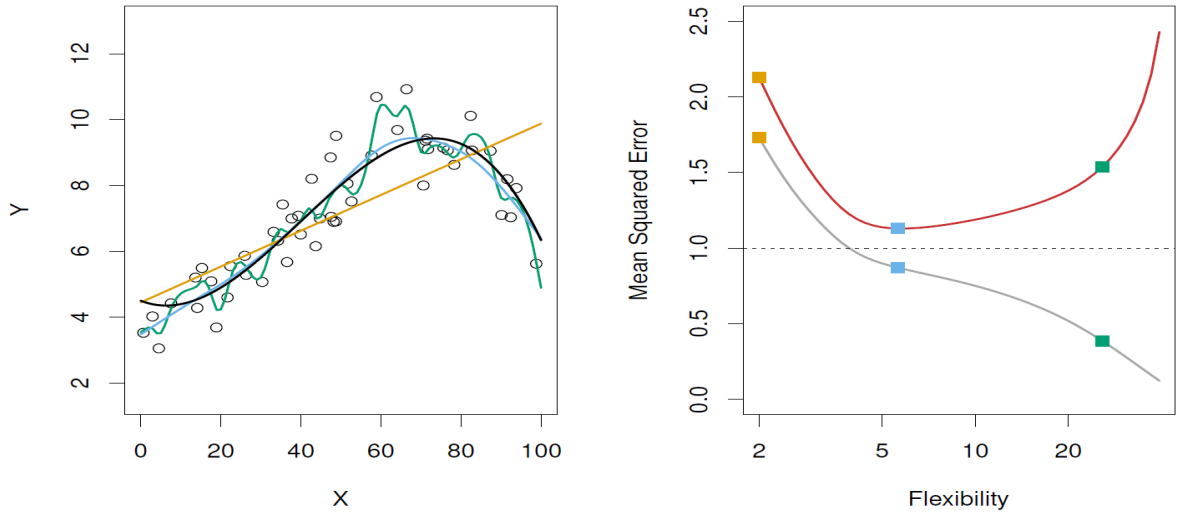


Figure 1.10: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve) and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

We now move to the right-hand panel of Fig 1.10. The grey curve displays the average training training MSE as a function of flexibility, or more formally the *degree of freedom*, for a number of smoothing splines. The degree of freedom is a quantity that summarizes the flexibility of a curve. The orange, blue, and green squares indicate the MSEs associated with the corresponding curves in the left-hand panel. A more restricted and hence smoother curve has fewer degrees of freedom than a wiggly curve, note that in Fig 1.10, linear regression is at the most restrictive end, with two degree of freedom. The training MSE declines monotonically as flexibility increases. In this example the true f is non-linear, and so the orange linear fit is not flexible enough to estimate f well. The green curve has the lowest training MSE of all three methods, since it corresponds to the most flexible of the three curves fit in the left-hand panel.

In this example, we know the true function f, and so we can also compute the test MSE over a very large test set, as a function of flexibility. (Of course, in general f is unknown, so this will not be possible.) The test MSE is displayed using the red curve in the right-hand panel of Fig 1.10. As with the training MSE, the test MSE initially declines as the level of flexibility increases. However at some point the test MSE levels off and then starts to increase again. Consequently, the orange and green curves both have high test MSE. The blue curve minimizes the test MSE, which should not be surprising given that visually it appears to estimate f the best in the left-hand panel of Fig 1.10. The horizontal dashed line indicates $\text{Var}(\epsilon)$, the irreducible error in

$$E(Y - \widehat{Y})^2 = \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \tag{1.7}$$

which corresponds to the lowest achievable test MSE among all possible methods. Hence, the smoothing spline represented by the blue curve is close to optimal.

In the right-hand panel of Fig 1.10, as the flexibility of the statistical learning method increases, we observed a monotone decrease in the training MSE and a *U-shape* in the test MSE. This is a fundamental property of statistical leaning that holds regardless of the particular data

set at hand and regardless of the statistical method being used. As model flexibility increases, training MSE will decrease, but the test MSE may not. When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and maybe picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f. When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don't exist in the test data. Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE. Overfitting refers specifically to the case in which a less fexible model would have yielded a smaller test MSE.

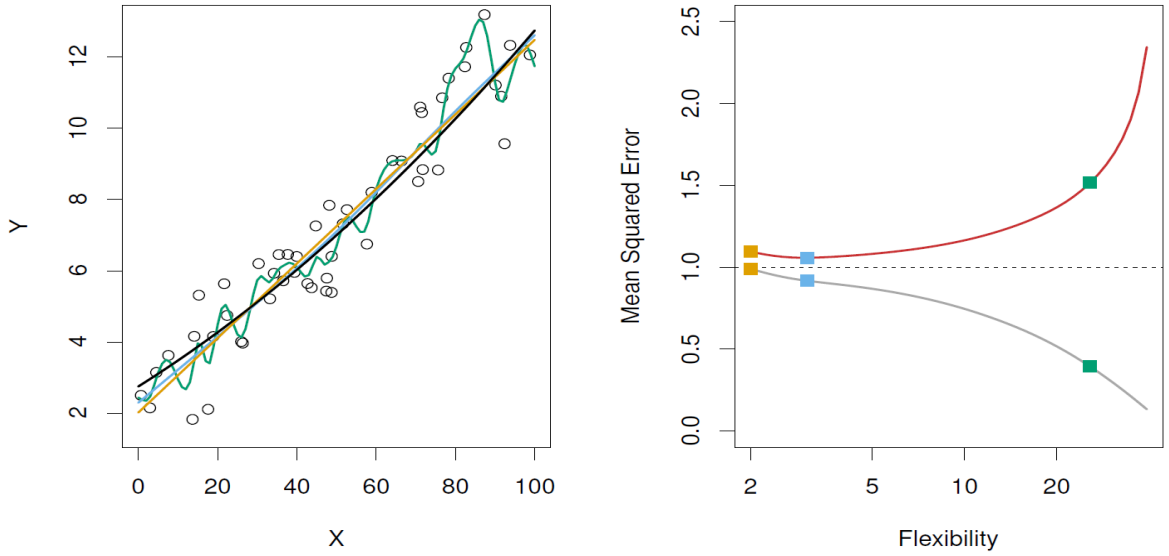


Figure 1.11: Details are as in Fig 1.10, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

Fig 1.11 provides another example in which the true f is approximately linear. Again we observe that the training MSE decreases monotonically as the model flexibility increases, and that there is a U-shape in the test MSE. However, because the truth is close to linear, the test MSE only decreases slightly before increasing again, so that the orange least squares fit is substantially better than the highly flexible green curve.

Finally, Fig 1.12 displays an example in which f is highly non-linear. The training and test MSE curves still exhibit the same general patterns, but now there is a rapid decrease in both curves before the test MSE starts to increase slowly.

In practice, one can usually compute the training MSE with relative easy, but estimating test MSE is considerably more difficult because usually no test data are available. As the previous three examples illustrate, the flexibility level corresponding to model with the minimal test MSE can vary considerably among data sets. Throughout this course, we discuss a variety of approaches that can be used in practice to estimate this minimum point. One important method is *cross-validation*, which is a method for estimating test MSE using the training data.

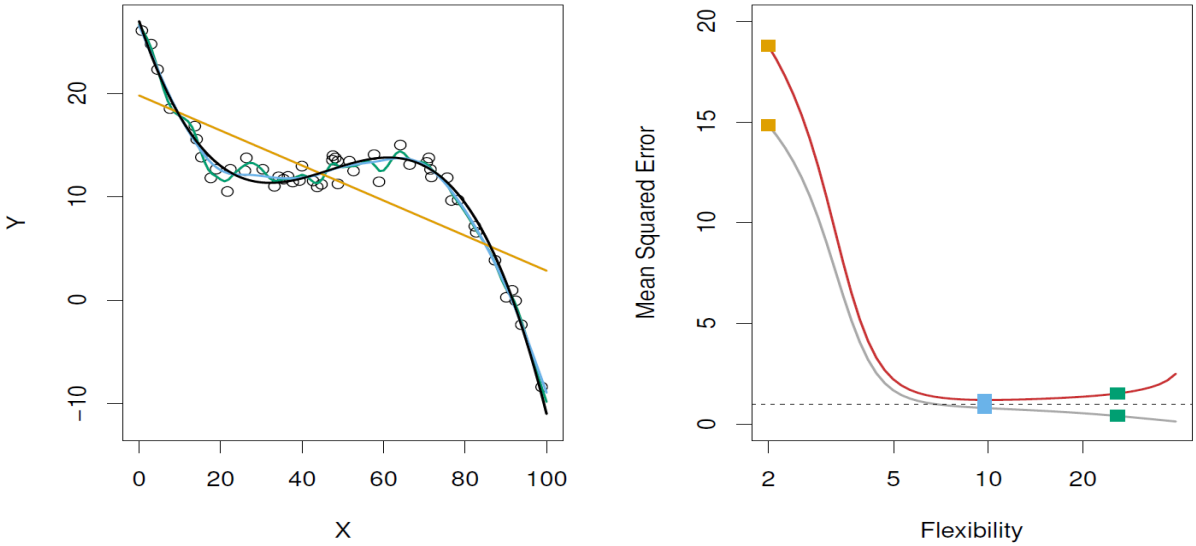


Figure 1.12: Details are as in Fig 1.10, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.

1.4.2 The Bias-Variance Trade-Off

The U-shape observed in the test MSE curves (Figures 1.10-1.12) turns out to be the result of two competing properties of statistical learning methods. It is possible to show that the expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the *variance* of $\widehat{f}(x_0)$, the squared *bias* of $\widehat{f}(x_0)$ and the variance of the error term ϵ. That is

$$E\left(y_0 - \widehat{f}(x_0)\right)^2 = \text{Var}\left(\widehat{f}(x_0)\right) + \left[\text{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \text{Var}(\epsilon). \tag{1.8}$$

Here the notation $E\left(y_0 - \widehat{f}(x_0)\right)^2$ defines the *expected test* MSE at x_0, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0. The overall expected test MSE can be computed by averaging $E\left(y_0 - \widehat{f}(x_0)\right)^2$ over all possible values of x_0 in the test set.

Equation 1.8 tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves *low variance* and *low bias*. Note that variance is inherently a non-negative quantity, and squared bias is also non-negative quantity. Hence, we see that the expected test MSE can never lie below $\text{Var}(\epsilon)$, the irreducible error from (1.7).

What do we mean by the *variance* and *bias* of a statistical learning method? Variance refers to the amount by which $\widehat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\widehat{f}$. But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in $\widehat{f}$. In general, more flexible statistical methods have higher variance. Consider the green and orange curves in Fig 1.10. The flexible green curve is following the observations very closely. It has high variance because changing any one of these data points may cause the estimate $\widehat{f}$ to change considerably. In contrast, the orange least

squares line is relatively inflexible and has low variance, because moving any single observation will likely cause only a small shift in the position of the line.

On the other hand, bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. For example, linear regression assumes that there is a linear relationship between Y and $X_1, X_2, \cdots, X_p$. It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f. In Fig 1.12, the true f is substantially non-linear, so no matter how many training observations we are given, it will not be possible to produce an accurate estimate using linear regression. In other words, linear regression results in high bias in this example. However, in Fig 1.11 the true f is very close to linear, and so given enough data, it should be possible for linear regression to produce an accurate estimate. Generally, more flexible methods result in less bias.

As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases. As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines. However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens the test MSE increases. Note that we observed this pattern of decreasing test MSE followed by increasing test MSE in the right-hand panels of Figures 1.10-1.12.

The three plots in Fig 1.13 illustrate Equation 1.8 for the examples in Figures 1.10-1.12. In each case the blue solid curve represents the squared bias, for different level of flexibility, while the orange curve corresponds to the variance. The horizontal dashed line represents $\text{Var}(\epsilon)$, the irreducible error. Finally, the red curve, corresponding to the test set MSE, is the sum of these three quantities. In all three cases, the variance increases and the bias decreases as the method's flexibility increases. However, the flexibility level corresponding to the optimal test MSE differs considerably among the three data sets, because the squared bias and variance change at different rates in each of the data sets. In the left-hand panel of Figure 1.13, the bias initially decreases rapidly, resulting in an initial sharp decrease in the expected test MSE. On the other hand, in the center panel of Figure 1.13 the true f is close to linear, so there is only small decrease in bias as flexibility increases, and the test MSE only declines slightly before increasing rapidly as the variance increases. Finally, in the right-hand panel of Fig 1.13, as flexibility increases, there is a dramatic decline in bias because the true f is very non-linear. There is also very little increase in variance as flexibility increases. Consequently, the test MSE declines substantially before experiencing a small increase as model flexibility increases.

The relationship between bias, variance, and test set MSE given in Equation 1.8 and displayed in Figure 1.13 is refereed to as the *bias-variance trade-off*. Good test set performance of a statistical learning method requires low variance as well as low squared bias. This is referred to as a trade-off because it is easy to obtain a method with extremely low bias but high variance (for instance, by drawing a curve that passes through every single training observation) or a method with very low variance but high bias (by fitting a horizontal line to the data). The challenge lies in finding a method for which both variance and the squared bias are low. This trade-off is one of the most important recurring themes in this course.

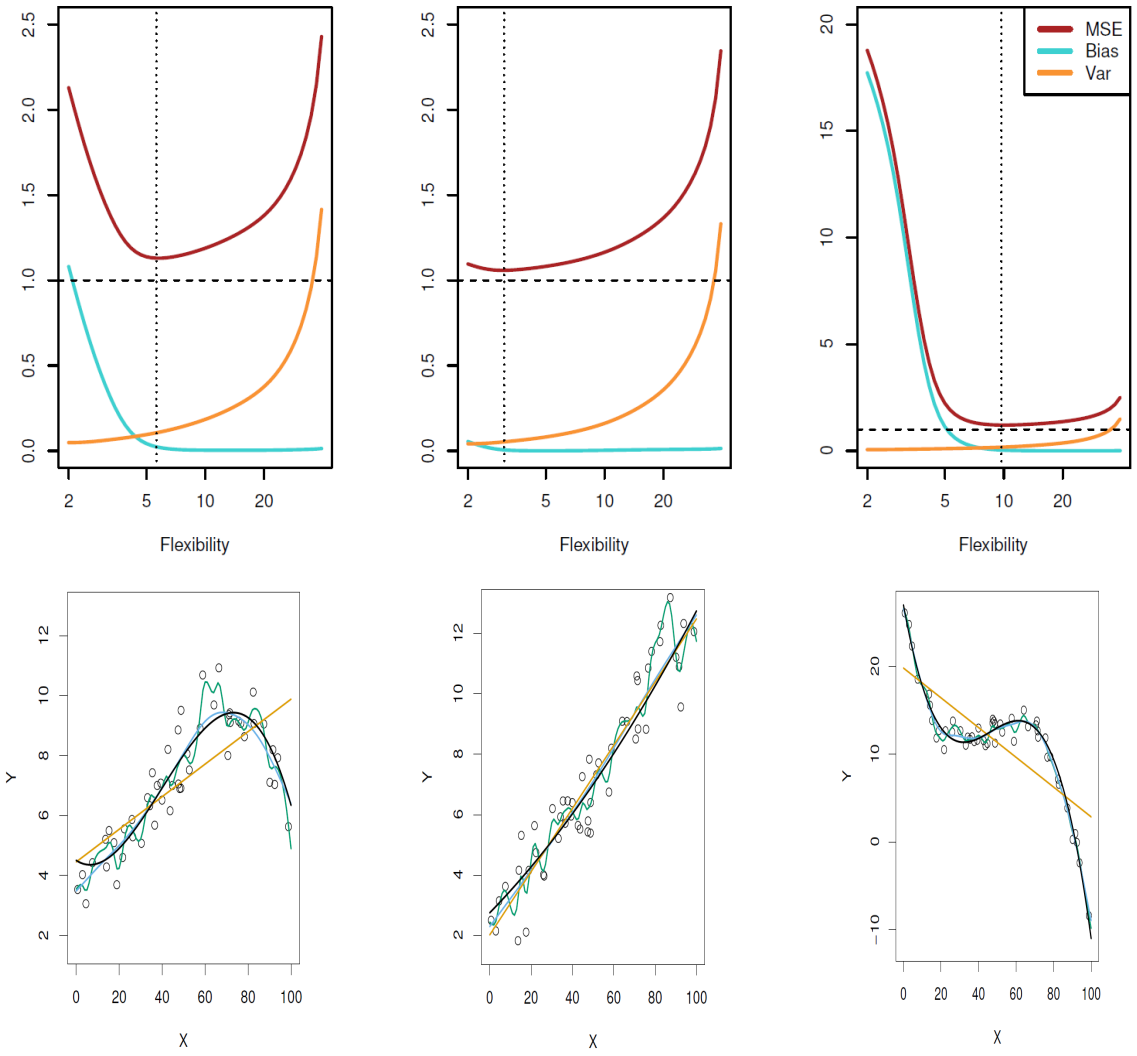


Figure 1.13: Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and the test MSE (red curve) for the three data sets in Figures 1.10-1.12. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

In a real-life situation on which f is unobserved, it is generally not possible to explicitly compute the test MSE, bias, or variance for a statistical learning method. Nevertheless, one should always keep the bias-variance trade-off in mind. In this course and future courses, we explore methods that are extremely flexible and hence can essentially eliminate bias. However, this does not guarantee that they will outperform a much simpler method such as linear regression. To take an extreme example, suppose that the true f is linear. In this situation linear regression will have no bias, making it very hard for a more flexible method to compete. In contrast, if the true f is highly non-linear and we have an ample number of training observations, then we may do better using a highly flexible approach, as in Figure 1.12. In a later chapter we discuss cross-validation, which is a way yo estimate the test MSE using the training data.

We now look at the mathematical derivation of the bias-variance decomposition of the test mean squared error,

$$E\left(y_0 - \widehat{f}(x_0)\right)^2 = \text{Var}\left(\widehat{f}(x_0)\right) + \left[\text{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \text{Var}(\epsilon).$$

given in (1.8).

Let us recall the following definitions and setup:

- $y_0 = f(x_0)$ is the true response for the input x_0.
- $\widehat{f}(x_0)$ is the predicted response for the input x_0.
- ϵ is the noise in the response, which is assumed to be a random variable with zero mean and variance σ^2.
- The true response can be modeled as:

  $$y_0 = f(x_0) + \epsilon,$$

  where f is the true underlying function. Here, $f(x_0)$ is the true, deterministic value of the function f at x_0, while y_0 is the observed value at x_0 which includes both the true value $f(x_0)$ and the random noise ϵ.
- The test mean squared error, at x_0 is defined as: $E\left(y_0 - \widehat{f}(x_0)\right)^2$, and refers to the average test MSE that we would obtain if we repeatedly estimated $\widehat{f}$ using a large number of training sets, and tested each at x_0.

We can decompose the test MSE as follows:

Subtracting both sides of $y_0 = f(x_0) + \epsilon$ by $\widehat{f}(x_0)$ we obtain

$$y_0 - \widehat{f}(x_0) = f(x_0) + \epsilon - \widehat{f}(x_0) \tag{1.9}$$

Expand the square of (1.9) we write

$$\left(f(x_0) + \epsilon - \widehat{f}(x_0)\right)^2 = \left(f(x_0) - \widehat{f}(x_0)\right)^2 + 2\left(f(x_0) - \widehat{f}(x_0)\right)\epsilon + \epsilon^2$$

Hence, with the properties of expectation (the expectation ranges over different choices of the training set), we get

$$E\left(\left(y_0 - \widehat{f}(x_0)\right)^2\right) = E\left(\left(f(x_0) - \widehat{f}(x_0)\right)^2\right) + E\left(2\left(f(x_0) - \widehat{f}(x_0)\right)\epsilon\right) + E\left(\epsilon^2\right) \tag{1.10}$$

Notice that the term $f(x_0) - \widehat{f}(x_0)$ is independent of ϵ. To understand why, take note of the fact that $\widehat{f}(x_0)$ is the predicted value based on a learning process using a training dataset. The training data consists of pairs (x_i, y_i) where $y_i = f(x_i) + \epsilon_i$, and importantly, ϵ_i are noise terms in the training data. When evaluating the model, the term $y_0 = f(x_0) + \epsilon$ is from a test dataset (new observation). The noise term ϵ in y_0 is a new realization, separate from the noise in the training data. This explains why ϵ is independent of $\widehat{f}(x_0)$ and the rest is clear. With this we have the cross-term

$$E\left(2\left(f(x_0) - \widehat{f}(x_0)\right)\epsilon\right) = 2E\left(f(x_0) - \widehat{f}(x_0)\right)E(\epsilon) = 0$$

since $E(\epsilon) = 0$. Also with the error having mean zero, the term $E(\epsilon^2)$ is the variance of the noise, σ^2. Therefore Equation (1.10) becomes

$$E\left(\left(y_0 - \widehat{f}(x_0)\right)^2\right) = E\left(\left(f(x_0) - \widehat{f}(x_0)\right)^2\right) + \sigma^2 \tag{1.11}$$

Next, we need to decompose $E\left(\left(f(x_0)-\widehat{f}(x_0)\right)^2\right)$. From basic statistics we know that bias at x_0 is defined as:

$$\text{Bias}\left(\widehat{f}(x_0)\right) = E\left(\widehat{f}(x_0)\right) - f(x_0)$$

while the variance at x_0 is defined as:

$$\text{Var}\left(\widehat{f}(x_0)\right) = E\left[\left(\widehat{f}(x_0) - E\left(\widehat{f}(x_0)\right)\right)^2\right]$$

Hence we obtain the decomposition of $E\left(\left(f(x_0)-\widehat{f}(x_0)\right)^2\right)$ as:

$$\begin{aligned}
E\left[\left(f(x_0)-\widehat{f}(x_0)\right)^2\right] &= E\left[\left(f(x_0)+\left(-E(\widehat{f}(x_0))+E(\widehat{f}(x_0))\right)-\widehat{f}(x_0)\right)^2\right] \\
&= E\left[\left(\left(f(x_0)-E(\widehat{f}(x_0)\right)+\left(E(\widehat{f}(x_0))-\widehat{f}(x_0)\right)\right)^2\right] \\
&= E\left[\left(f(x_0)-E(\widehat{f}(x_0)\right)^2\right] + 2E\left\{\left[f(x_0)-E(\widehat{f}(x_0))\right]\left[E(\widehat{f}(x_0))-\widehat{f}(x_0)\right]\right\} \\
&\quad + E\left[\left(E(\widehat{f}(x_0))-\widehat{f}(x_0)\right)^2\right]
\end{aligned}$$

The cross-product term in the above equation is equal to zero as $\left[f(x_0)-E(\widehat{f}(x_0))\right]$ is a constant term, hence

$$\begin{aligned}
E\left\{\left[f(x_0)-E(\widehat{f}(x_0))\right]\left[E(\widehat{f}(x_0))-\widehat{f}(x_0)\right]\right\} &= \left[f(x_0)-E(\widehat{f}(x_0))\right] E\left\{\left[E(\widehat{f}(x_0))-\widehat{f}(x_0)\right]\right\} \\
&= \left[f(x_0)-E(\widehat{f}(x_0))\right]\left\{\left[E(\widehat{f}(x_0))-E(\widehat{f}(x_0))\right]\right\} \\
&= 0
\end{aligned}$$

Therefore we get

$$\begin{aligned}
E\left[\left(f(x_0)-\widehat{f}(x_0)\right)^2\right] &= E\left[\left(f(x_0)-E(\widehat{f}(x_0)\right)^2\right] + 2E\left\{\left[f(x_0)-E(\widehat{f}(x_0))\right]\left[E(\widehat{f}(x_0))-\widehat{f}(x_0)\right]\right\} \\
&\quad + E\left[\left(E(\widehat{f}(x_0))-\widehat{f}(x_0)\right)^2\right] \\
&= E\left[\underbrace{\left(f(x_0)-E(\widehat{f}(x_0)\right)^2}_{\text{this is a constant term}}\right] + E\left[\left(E(\widehat{f}(x_0))-\widehat{f}(x_0)\right)^2\right] \\
&= \left(f(x_0)-E(\widehat{f}(x_0)\right)^2 + E\left[\left(E(\widehat{f}(x_0))-\widehat{f}(x_0)\right)^2\right] \\
&= \left[\text{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \text{Var}\left(\widehat{f}(x_0)\right) \qquad (1.12)
\end{aligned}$$

Together with (1.11), we conclude that

$$E\left(y_0-\widehat{f}(x_0)\right)^2 = \text{Var}\left(\widehat{f}(x_0)\right) + \left[\text{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \text{Var}(\epsilon).$$

1.4.3 The Classification Setting

Thus far, our discussion of model accuracy has been focused on the regression setting. But many of the concepts that we have encountered, such as the bias-variance trade-off, transfer over to the classification setting with only some modifications due to the fact that y_i is no longer quantitative. Suppose that we seek to estimate f on the basis of training observations $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, where now $y_1, \cdots, y_n$ are qualitative. The most common approach for quantifying the accuracy of our estimate $\widehat{f}$ is the training *error rate*, the proportion of mistakes that are made if we apply our estimate $\widehat{f}$ to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i) \tag{1.13}$$

Here $\widehat{y}_i$ is the predicted class label for the ith observation using $\widehat{f}$. And $I(y_i \neq \widehat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \widehat{y}_i$ and zero if $y_i = \widehat{y}_i$. If $I(y_i \neq \widehat{y}_i) = 0$ then the ith observation was classified correctly by our classification method; otherwise it was misclassified. Hence Equation 1.13 computes the fraction of incorrect classification.

Equation 1.13 is referred to as the *training error* rate because it is computed based on the data that was used to train our classifier. As in the regression setting, we are most interested in the error rates that result from applying our classifier to test observations that were not used in training. The *test error* rate associated with a set of test observations of the form (x_0, y_0) is given by

$$\text{Ave}\left(I(y_0 \neq \widehat{y}_0)\right), \tag{1.14}$$

where $\widehat{y}_0$ is the predicted class label that results from applying the classifier to the test observation with predictor x_0. A good classifier is the one for which the test error (1.14) is smallest.

The Bayes Classifier

It is possible to show (the proof of this is left as an exercise) that the test error rate given in (1.14) is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor x_0 to class j for which

$$\Pr(Y = j | X = x_0) \tag{1.15}$$

is largest. Note that (1.15) is a *conditional probability*: it is the probability that $Y = j$, given the observed predictor vector x_0. This very simple classifier is called the *Bayes classifier*. In a two-class problem where there are only two possible response, say class 1 or class 2, the Bayes classifier corresponds to predicting class one if $\Pr(Y = 1|X = x_0) > 0.5$, and class two otherwise.

Figure 1.14 provides an example using a simulated data set in a two dimensional space consisting of predictors X_1 and X_2. The orange and blue circles correspond to training observations that belong to two different classes. For each value of X_1 and X_2, there is a different probability of the response being orange or blue. Since this is simulated data, we know how the data were generated and we can calculate the conditional probabilities for each value of X_1 and X_2. The orange shaded region reflects the set of points for which $\Pr(Y = \text{orange}|X)$ is greater than 50%, while the blue shaded region indicates the set of points for which the probability is below 50%. The purple dashed line represents the points where the probability is exactly 50%. This is called the Bayes decision boundary. The Bayes classifier's prediction is determined by the Bayes decision boundary; an observation that falls on the orange side of the boundary will be

assigned the orange class, and similarly an observation on the blue side of the boundary will be assigned to the blue class.

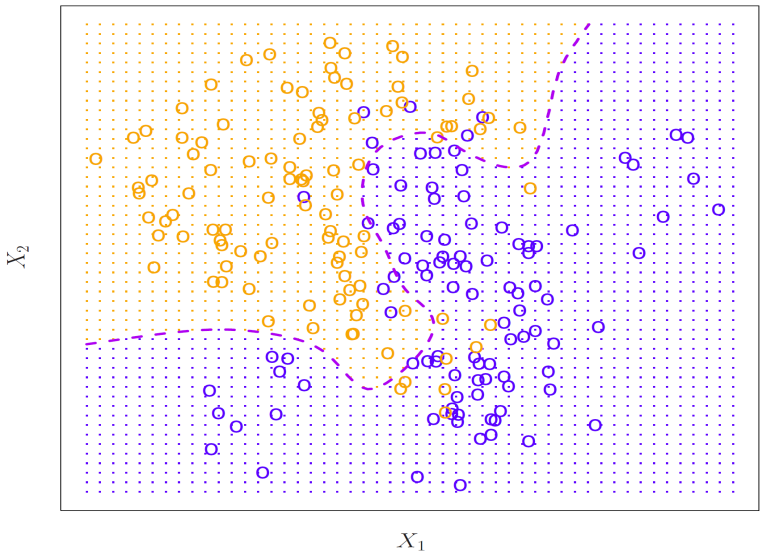


Figure 1.14: A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

The Bayes classifier produces the lowest possible test error rate, called the *Bayes error rate*. Since the Bayes classifier will always choose the class for which (1.15) is largest, the error rate will be

$$1 - \max_j \Pr(Y = j|X = x_0)$$

at $X = x_0$. In general, the overall Bayes error rate is given by

$$1 - E\left(\max_j \Pr(Y = j|X)\right) \tag{1.16}$$

where the expectation averages the probability over all possible values of X. For our simulated data, the Bayes error rate is 0.133. It is greater than zero, because the classes overlap in the true population so $\max_j \Pr(Y = j|X = x_0) < 1$ for some values of x_0. The Bayes error rate is analogous to the irreducible error, discussed earlier.

K-Nearest Neighbors

In theory we would always like to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible. Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods. Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest *estimated* probability. One such method is the K-nearest neighbors (KNN) classifier. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of the points in $\mathcal{N}_0$ whose response values equal j:

$$\Pr\left(Y = j|X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j). \tag{1.17}$$

Finally, KNN classifies the test observation x_0 to the class with the largest probability from (1.17).

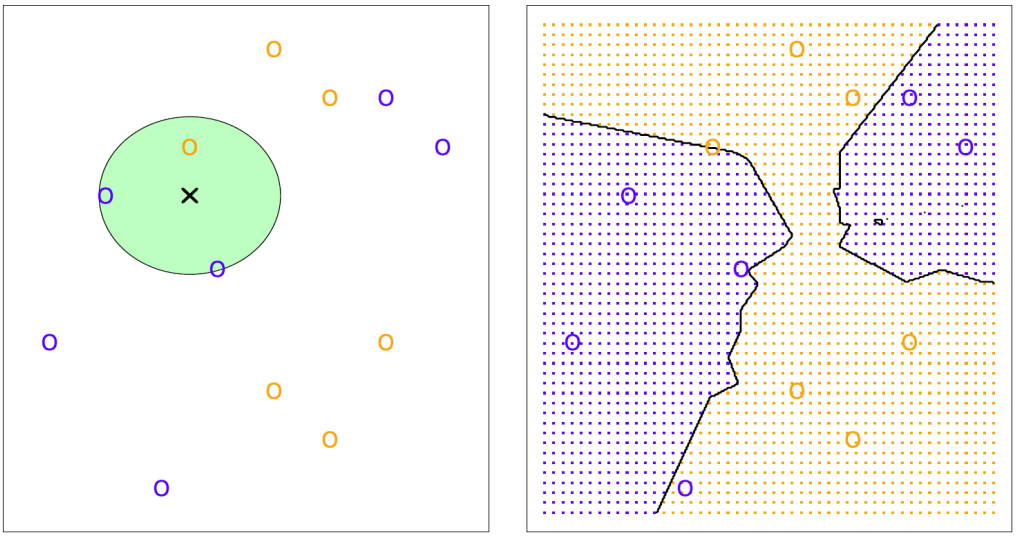


Figure 1.15: The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

Figure 1.15 provides an illustrative example of the KNN approach. In the left-hand panel, we have plotted a small training data set consisting of six blue and six orrange observations. Our goal is to make a prediction for the point labeled by the black cross. Suppose that we choose $K = 3$. Then KNN will first identify the three observations that are closet to the cross. This neighborhood is shown as a circle. It consists of two blue points and one orange point, resulting in estimated probabilities of 2/3 for the blue and 1/3 for the orange class. Hence KNN will predict that the black cross belongs to the blue class. In the right-hand panel of Figure 1.15 we have applied the KNN approach with $K = 3$ at all of the possible values for X_1 and X_2, and have drawn in the corresponding KNN decision boundary.

Despite the fact that it is a very simple approach, KNN can often produce classifiers that are surprisingly close to the optimal Bayes classifier. Figure 1.16 displays the KNN decision boundary, using $K = 10$, when applied to the larger simulated data set from Figure 1.14. Notice that even though the true distribution is not know by the KNN classifier, the KNN decision boundary is very close to that of the Bayes classifier. The test error rate using KNN is 0.1363, which close to the Bayes error rate of 0.1304.

The choice of K has a drastic effect on the KNN classifier obtained. Figure 1.17 displays two KNN fits to the simulated data from Figure 1.14, using $K = 1$ and $K = 100$. When $K = 1$, the decision boundary is overly flexible and finds patterns in the data that don't correspond to the Bayes decision boundary. This corresponds to a classifier that has low bias but very high variance. As K grows, the method becomes less flexible and produced a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier. On this simulated data set, neither $K = 1$ nor $K = 100$ give good predictions: they have test error rates of 0.1695 and 0.1925 respectively.

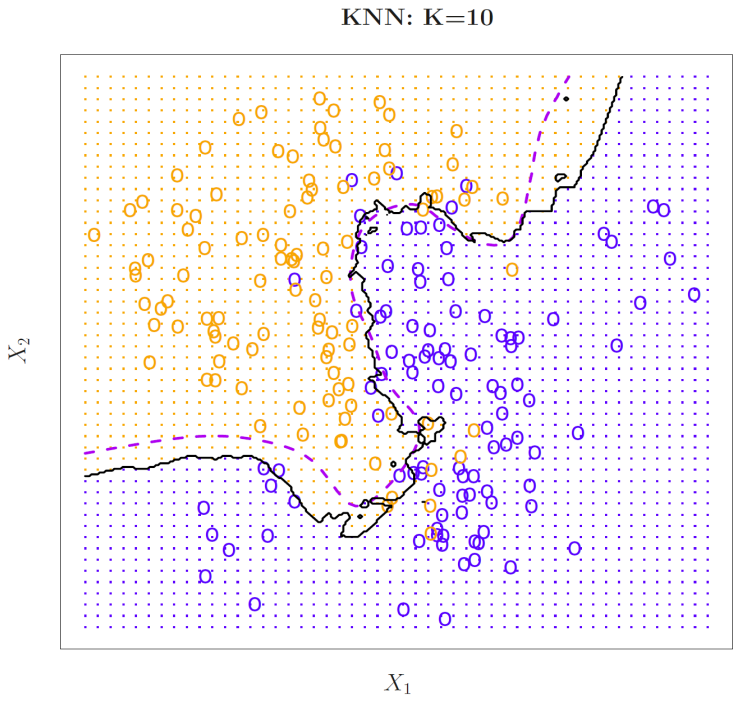


Figure 1.16: The black curve indicates the KNN decision boundary on the data from Figure 1.14, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.

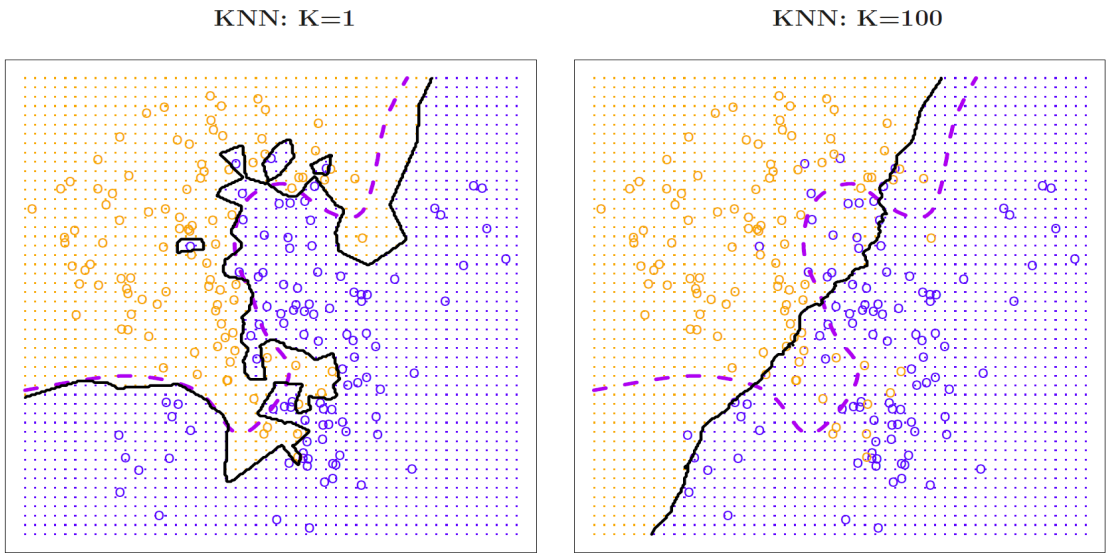


Figure 1.17: A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 1.14. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.

Just as in the regression setting, there is not a strong relationship between the training error rate and the test error rate. With $K = 1$, the KNN training error rate is 0, but the test error rate may be quite high. In general, as we use more flexible classification methods, the training error rate will decline but the test error rate may not. In Figure 1.18, we have plotted the KNN test and training errors as a function of $1/K$. As $1/K$ increases, the method becomes more flexible. As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits a characteristic U-shape, declining at first (with a minimum at approximately $K = 10$) before increasing again when the method becomes excessively flexible and overfits.

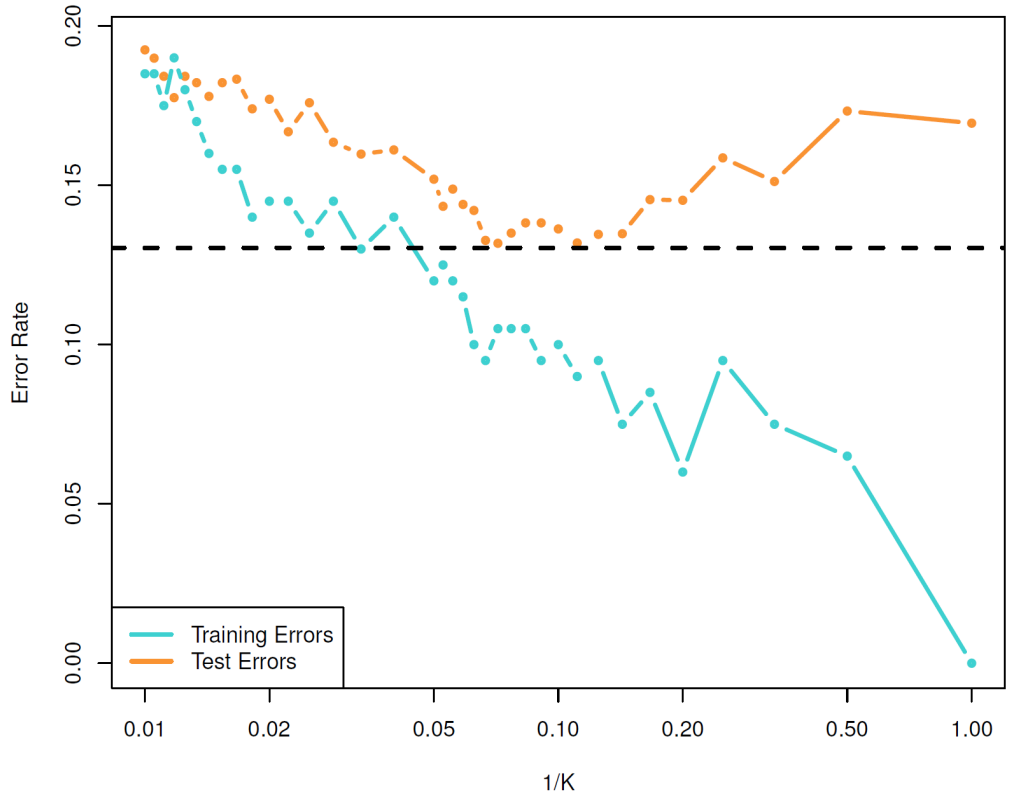


Figure 1.18: The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 1.14, as the level of flexibility (assessed using $1/K$ on the log scale) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.

In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method. The bias-variance tradeoff, and the resulting U-shape in the test error, can make this a difficult task. In a later section, we return to this topic and discuss various methods for estimating test error rates and thereby choosing the optimal level of flexibility for a given statistical learning method.

1.5 Reference

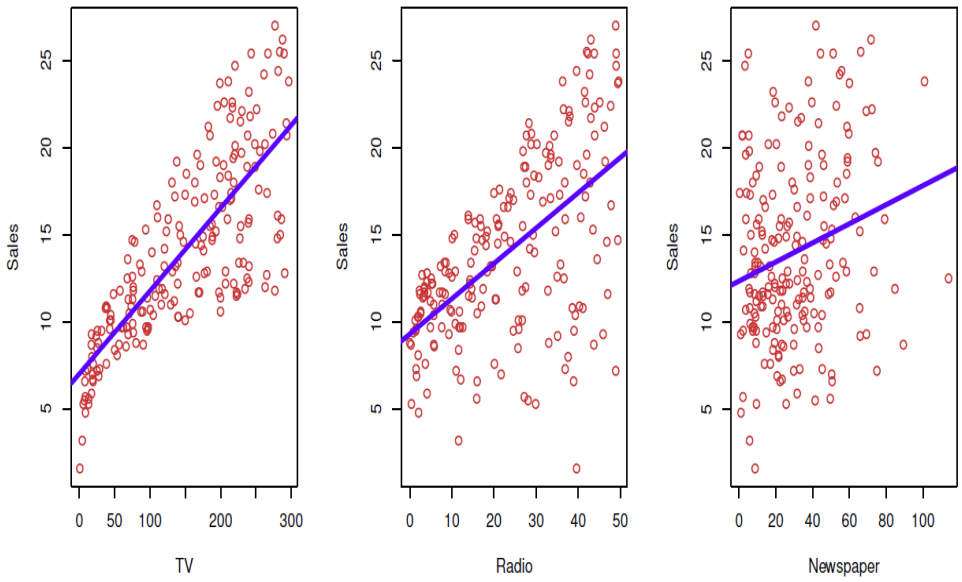
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. An Introduction to Statistical Learning. 2nd ed. Springer Texts in Statistics. New York, NY: Springer.

Chapter 2

Linear Regression

The method of *linear regression* is a very simple approach for supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response, Y. It has been around for a long time and is a topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern statistical learning approaches, linear regression is still a useful and widely used statistical learning method.

Recall the Advertising data from Chapter 1. Figure 1.1, (reproduced below) shows sales (in thousands of unit) for a particular product as a function of advertising budgets (in thousands of dollars) for TV, radio, and newspaper media. Suppose that in our role as statistical consultants we are asked to suggest, on the basis of this data, a marketing plan for next year that will result in high product sales. What information would be useful in order to provide such a recommendation? Here are a few important questions that we might seek to address:



1. *Is there a relationship between advertising budget and sales?*
   Our first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales. If the evidence is weak, then one might argue that no money should be spent on advertising!

2. *How strong is the relationship between advertising budget and sales?*
   Assuming that there is a relationship between advertising and sales, we would like to know the strength of this relationship. Does knowledge of the advertising budget provide a lot of information about product sales?

3. *Which media are associated with sales?*
   Are all three media, TV, radio, and newspaper, associated with sales, or are just one or two of the media associated? To answer this question, we must find a way to separate out the individual contribution of each medium to sales when we have spent money on all three media.

4. *How large is the association between each medium and sales?*
   For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurate can we predict this amount of increase?

5. *How accurate can we predict future sales?*
   For any given level of TV, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of the prediction?

6. *Is the relationship linear?*
   If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictor or the response so that linear regression can be used.

7. *Is there synergy among the advertising media?*
   Perhaps spending \$50,000 on TV advertising and \$50,000 on radio advertising is associated with higher sales than allocating \$100,000 to either TV or radio individually. In marketing, this is known as a *synergy* effect, while in statistics it is called an *interaction* effect.

It turns out that linear regression can be used to answer each of these questions. We will first discuss all these questions in a general context, and then return to them in this specific context later.

2.1 Simple Linear Regression

Simple linear regression is a very straightforward approach for predicting a quantitative response Y on the basis of the single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X \tag{2.1}$$

You might read "$\approx$" as "*is approximately modeled as*". We will also sometimes describe (2.1) by saying that we are regressing Y on X (or Y onto X).

For example, X may represent TV advertising, and Y may represent sales. Then we can regress sales onto TV by fitting the model

$$sales \approx \beta_0 + \beta_1 \times TV.$$

In Equation (2.1), β_0 and β_1 are two unknown constants that represnt the *intercept* and *slope* terms in the linear model. Together, β_0 and β_1 are known as the model *coefficients* or *parameters*. Once we have used our training data to produce estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x, \tag{2.2}$$

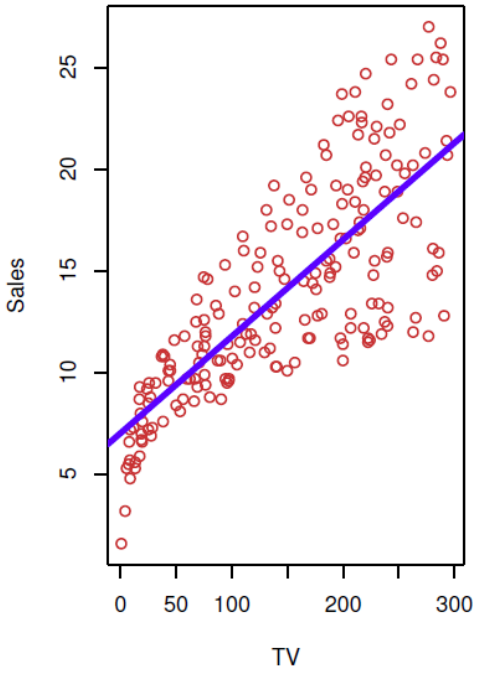
where $\widehat{y}$ indicates a prediction of Y on the basis of $X = x$. Here we use a *hat* symbol, $\widehat{\ }$, to denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response.

2.1.1 Estimating the Coefficients

In practice, β_0 and β_1 are unknown. So before we can use (2.1) to make predictions, we must use data to estimate the coefficients. Let

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

represent n observation pairs, each of which consists of a measurement of X and a measurement of Y. In the Advertising example, this data set consists of the TV advertising budget and product sales in $n = 200$ different markets.

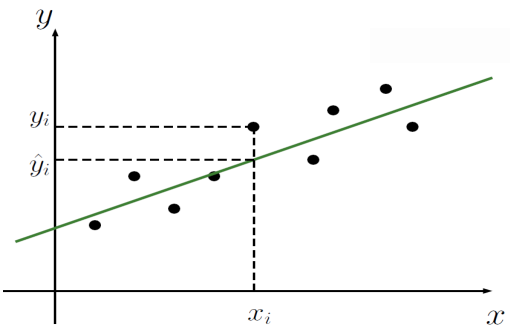


Our goal is to obtain coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that the linear model (2.1) fits the available data well, that is, so that

$$y_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad \text{for} \quad i = 1, \cdots, n$$

In other words, we want to find an intercept $\widehat{\beta}_0$ and a slope $\widehat{\beta}_1$ such that the resulting line is as close as possible to the $n = 200$ data points. There are a number of ways of measuring *closeness*. However, by far the most common approach involves minimizing the *least squares* criterion, and we take that approach in this chapter. Alternative approaches will be considered later.

Let $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \widehat{y}_i$ represents the ith *residual*, that is the difference between the ith observed response value and the ith response value that is predicted by our linear model.



We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

or equivalently as

$$\text{RSS} = (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_0 - \widehat{\beta}_1 x_2)^2 + \cdots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2 \tag{2.3}$$

The least squares approach chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the RSS. Using some basic calculus, one can show that the minimizers are

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}, \end{aligned} \tag{2.4}$$

where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i \quad \text{and} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i,$$

Another equivalent way of writing (2.4) above is

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \\ \widehat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}, \end{aligned} \tag{2.5}$$

where

$$S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \tag{2.6}$$

and

$$S_{xy} = \sum_{i=1}^n y_i (x_i - \overline{x}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n} \tag{2.7}$$

Here is the derivation of (2.4). Let

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2.$$

Notice that x_i and y_i are known quantities, hence S is a function of variables β_0 and β_1. We now estimate β_0 and β_1 by minimizing $S(\beta_0, \beta_1)$ above. Taking the derivatives of S with respect to β_0 and β_1, we see that the *least squares estimators* of β_0 and β_1, written as $\hat{\beta}_0$ and $\hat{\beta}_1$ (which minimize S), must satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{2.8}$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \tag{2.9}$$

From (2.8)

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

we obtain

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

which gives the following

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{2.10}$$

Next, from (2.9)

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

we obtain

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

$$\sum_{i=1}^{n}\left(x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2\right) = 0$$

therefore, we have the following

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{2.11}$$

Therefore from (2.10) and (2.11) we have the following system of equations, which sometimes referred to as the *normal equations*.

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.12}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2.13}$$

Dividing (2.12) by n and with the following sample means

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Equation (2.12) simplifies to

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{2.14}$$

Substituting (2.14) into (2.13) we get

$$\left(\overline{y} - \hat{\beta}_1 \overline{x}\right)\sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Finally, solving for $\hat{\beta}_1$, we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}},$$

which are what we set up to show.

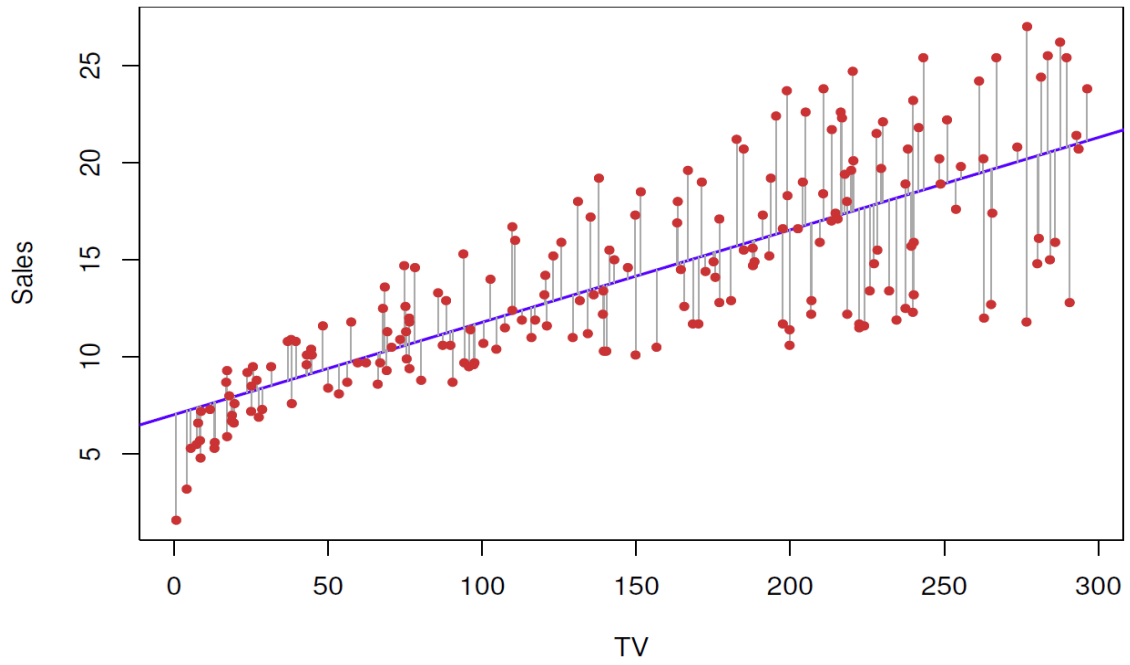


Figure 2.1: For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend on the left of the plot.

Fig 2.1 above displays the simple linear regression fit to Advertising data, where $\widehat{\beta}_0 = 7.03$ and $\widehat{\beta}_1 = 0.0475$. In other words, according to this approximation, an additional $1,000 spent on TV advertising is associated with selling approximately 47.5 additional unit of product. (Recall that Sales is in thousands of unit, and TV advertisement is in thousand of dollars.)

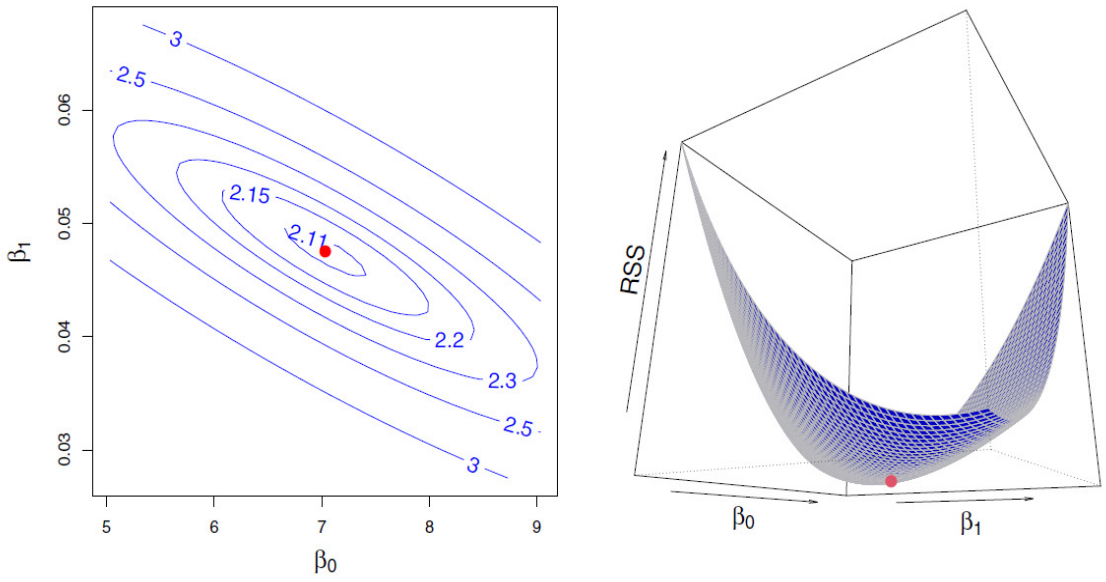


Figure 2.2: Contour and three-dimensional plots of the RSS on the Advertising data, using sales as response and TV as the predictor. The red dot correspond to the least squared estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, given by (2.4).

For Fig 2.2, we have computed RSS for a number of values β_0 and β_1, using the advertising data with sales as the response and TV as the predictor. In each plot, the red dot represents the pair of least squared estimates $(\widehat{\beta}_0, \widehat{\beta}_1)$, given by (2.4). These values clearly minimize the RSS.

2.1.2 Assessing the Accuracy of the Coefficient Estimates

As seen from (1.1), we assume that the true relationship between X and Y takes the form

$$Y = f(X) + \epsilon$$

for some unknown function f, where ϵ is a mean-zero random error term. If f is to be approximated by a linear function, then we can write the relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.15}$$

Here β_0 is the intercept term, that is, the expected value of Y when $X = 0$, and β_1 is the slope, the average increase in Y associated with a one-unit increase in X.

The error term ϵ is a catch-all for what we miss with this simple model. This is because the true relationship is probably not linear, there may be other variables that cause variation in Y, and there may be measurement error. We typically assume that the error term is independent of X.

The model given by (2.15) defines the *population regression line*, which is the best linear approximation to the true relationship between X and Y. The least squares regression coefficient estimate (2.4), namely

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}.\end{aligned}$$

characterize the the least squares line (2.2),

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

Let us now look at the following example. The left-hand panel of Fig 2.3 displays these two line in a simple simulated example. We create 100 random Xs, and generated 100 corresponding Ys from the model

$$Y = 2 + 3X + \epsilon \tag{2.16}$$

where ϵ was generated from a normal distribution with mean zero. The red line in the left-hand panel of Fig 2.3 displays the *true* relationship, $f(X) = 2 + 3X$, while the blue line is the least squares estimate based on the observed data. The true relationship is generally not known for real data, but the least squares line can always be computed using the coefficient estimates in (2.4), In other words, in real applications, we have access to a set of observations from which we can compute the least squares line; however, the population regression line is unobserved.

In the right-hand panel of Fig 2.3 we have generated ten different data sets from the model given by (2.16) and plotted the corresponding ten least squares lines. Notice that different data sets generated from the same true model result in slightly different least squares lines, but the unobserved population regression line does not change.

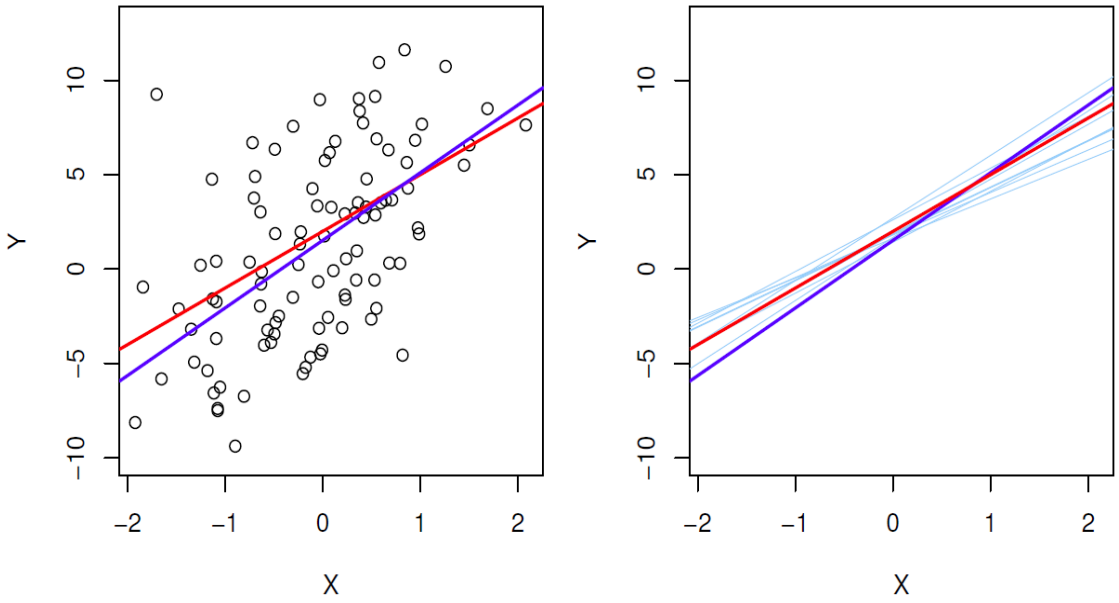


Figure 2.3: A simulated data set. Left: The red line represents the true relationship $f(X) = 2+3X$, which is known as the population regression line. The blue line is the least squares line, it is the least squares estimate for $f(X)$ based on the observed data, shown in black. Right: The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares line are quite close to the population regression line.

At first glance, the difference between the population regression line and the least squares line may seem subtle and confusing. We only have one data set, and so what does it mean that two different line describe the relationship between the predictor and the response?

Fundamentally, the concept of these two lines is a natural extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population. For example, suppose that we are interested in knowing the population mean μ of some random variable Y. Unfortunately, μ is unknown but we have access to n observations from Y, namely $y_1, \cdots, y_n$,which we can use to estimate μ. A reasonable estimate is $\widehat{\mu} = \overline{y}$, where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

is the sample mean. The sample mean and the population mean are different, but in general the sample mean will provide a good estimate of the population mean. (In fact, the sample mean is an *unbiased estimator* for the population mean).

In the same way, the unknown coefficients β_0 and β_1 in linear regression define the population regression line. We seek to estimate these unknown coefficients using $\widehat{\beta}_0$ and $\widehat{\beta}_1$ given in (2.4). These coefficients define the least squares line.

The analogy between linear regression and estimation of the mean of a random variable is an apt one based on the concept of *bias*. If we use the sample mean $\widehat{\mu}$ to estimate μ, this estimate is *unbiased*, in the sense that on average, we expect $\widehat{\mu}$ to equal to μ, and write $E(\widehat{\mu}) = \mu$.

What exactly does this mean? It means that on the basis of one particular set of observations $y_1, \cdots, y_n$, $\widehat{\mu}$ might overestimate μ, and on the basis of another set of observations, $\widehat{\mu}$ might underestimate μ. But if we could average a huge number of estimates of μ obtained from a huge number of sets of observations, then the average would exactly equal μ. Hence, an unbiased estimator does not *systematically* over- or under-estimate the true parameter.

The property of unbiasedness holds for the least squares coefficient estimate given by (2.4) as well, namely, if we estimate β_0 and β_1 on the basis of a particular data set, then our estimates won't be exactly equal to β_0 and β_1. But if we could average the estimates obtained over a huge number of data sets, then the average of these estimates would be spot on! In fact, we can see from the right-hand panel of Fig 2.3 that the average of many least squares, each estimated from a separate data set, is pretty close to the true population regression line.

We now provide the mathematical details on the above discussion.

From (2.5) we see that by minimizing the sum of squares S

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

we obtained

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}} = \sum_{i=1}^{n} c_i y_i \tag{2.17}$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}, \quad i = 1, \cdots, n,$$

as well as

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Using the properties of expectation and (2.17) we get

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) \\ &= \sum_{i=1}^{n} c_i E(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= \sum_{i=1}^{n} c_i \left(E(\beta_0 + \beta_1 x_i) + E(\epsilon_i)\right) \\ &= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \end{aligned} \tag{2.18}$$

since $E(\epsilon_i) = 0$ by assumption. We leave it as an exercise for you to show that

$$\sum_{i=1}^{n} c_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} c_i x_i = 1$$

Therefore, from (2.18) we conclude that

$$E(\hat{\beta}_1) = \beta_1$$

We have shown that $\hat{\beta}_1$ is an **unbiased estimator** of β_1.

Next we show that $\widehat{\beta}_0$ is also an unbiased estimator of β_0. From

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

and using the linear property of expectation as well as $E(\epsilon_i) = 0$, we obtain

$$\begin{aligned} E(\hat{\beta}_0) &= E(\overline{y} - \hat{\beta}_1 \overline{x}) \\ &= E\left(\frac{\sum_{i=1}^n y_i}{n}\right) - \overline{x} E(\hat{\beta}_1) \\ &= \frac{1}{n}\sum_{i=1}^n E(y_i) - \overline{x}\beta_1 \\ &= \frac{1}{n}\left(\sum_{i=1}^n (\beta_0 + \beta_1 x_i)\right) - \overline{x}\beta_1 \\ &= \frac{1}{n}\left(n\beta_0 + \beta_1 n\overline{x}\right) - \overline{x}\beta_1 \\ &= \beta_0 + \beta_1\overline{x} - \overline{x}\beta_1 = \beta_0 \end{aligned}$$

Hence, we say that $\hat{\beta}_0$ is an **unbiased estimator** of β_0.

We continue the analogy with the estimation of the population mean μ of a random variable Y. A natural question is as follow: How accurate is the sample mean $\widehat{\mu}$ as an estimate of μ? We have established that the average of $\widehat{\mu}$'s over many data sets will be very close to μ, but that a single estimate $\widehat{\mu}$ may be a substantial underestimate or overestimate of μ. How far off will that single estimate of $\widehat{\mu}$ be? In general, we answer this question by computing the *standard error* of $\widehat{\mu}$, written as $\mathrm{SE}(\widehat{\mu})$. Recall from elementary statistics that we have the following well-known formula

$$\mathrm{Var}(\widehat{\mu}) = \mathrm{SE}(\widehat{\mu}) = \frac{\sigma^2}{n} \tag{2.19}$$

where σ is the standard deviation of each of the realizations y_i of Y. (Note: This formula holds provided that the n observations are uncorrelated.)

Roughly speaking, the standard error tells us the average amount that this estimate $\widehat{\mu}$ differs from the actual value μ. Equation 2.19 also tells us how this deviation shrinks with n, that is, the more observations we have, the smaller the standard error of $\widehat{\mu}$. In a similar vein, we can wonder how close $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are to the true values β_0 β_1. To compute the standard errors associated with $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we use the following formulas:

$$\mathrm{SE}(\widehat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right], \tag{2.20}$$

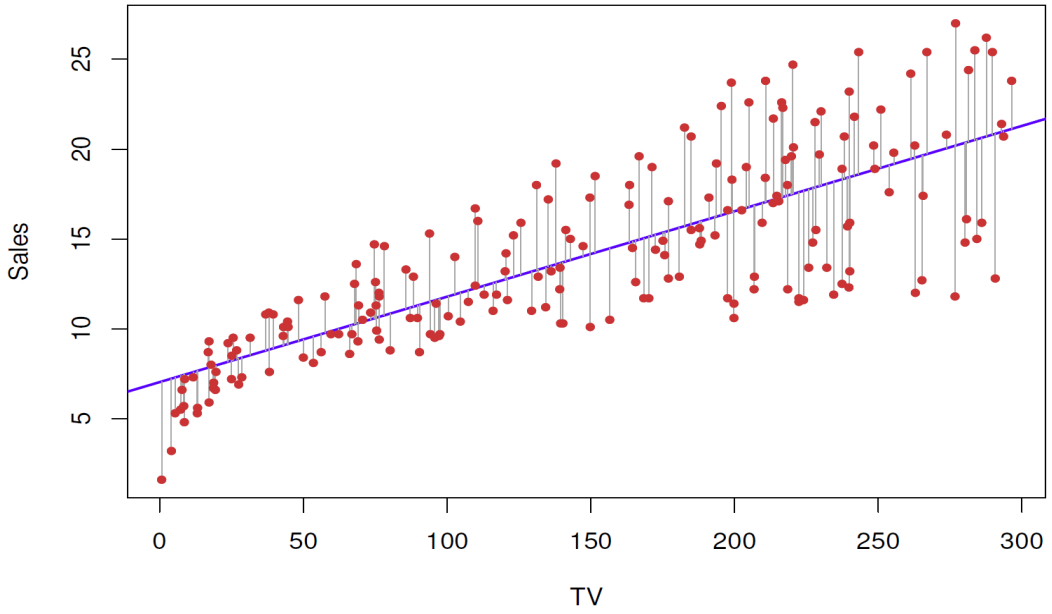
$$\mathrm{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}, \tag{2.21}$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$. Using the notation we saw in (2.6), namely, $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$, and that $\mathrm{SE}^2 = \mathrm{Variance}$, we can also write the above equation as

$$\mathrm{Var}(\widehat{\beta}_0) = \mathrm{SE}(\widehat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}}{S_{xx}}\right], \tag{2.22}$$

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{S_{xx}}, \tag{2.23}$$

For these formulas to be strictly valid, we need to assume that the errors ϵ_i, for each observation have a **common variance** σ^2 (also known as homoscedasticity condition) and are **uncorrelated**. Notice from Fig 2.1, reproduced below, that the errors do not have constant variance (heteroscedasticity condition), this can have several issues which we will discuss later in the course. However, our least square estimates remain unbiased.



Notice in the formula

$$\mathrm{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

that $\mathrm{SE}(\widehat{\beta}_1)$ is smaller when the x_i are more spread out, intuitively, we have more *leverage* to estimate a slope when this is the case.

We also see from

$$\mathrm{SE}(\widehat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right]$$

that $\mathrm{SE}(\widehat{\beta}_0)$ would be the same as $\mathrm{SE}(\widehat{\mu})$ if $\overline{x}$ were zero (in which case $\widehat{\beta}_0$ would be equal to $\overline{y}$, see (2.4)).

Let us now look at the derivation of (2.20) and (2.21). Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{S_{xx}} = \sum_{i=1}^n c_i y_i$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}, \quad i = 1, \cdots, n$$

hence

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^n c_i y_i \right) = \sum_{i=1}^n \sum_{j=1}^n c_i c_j \mathrm{Cov}(y_i, y_j)$$

Because the error terms ϵ_i are uncorrelated, so are the terms y_i, namely $\mathrm{Cov}(y_i, y_j) = 0$, for $i \neq j$. Hence the above equation gives

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^n c_i y_i \right) = \sum_{i=1}^n c_i^2 \mathrm{Var}(y_i) \tag{2.24}$$

With

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

we have, since $\text{Var}(y_i) = \sigma^2$ that

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \text{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2, \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right)^2 \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \frac{\sigma^2}{S_{xx}^2} S_{xx} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Which is what we want to show. Next using

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

and

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

the variance of $\hat{\beta}_0$ is computed as

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \text{Var}(\overline{y} - \hat{\beta}_1 \overline{x}) \\
&= \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\hat{\beta}_1) - 2\overline{x}\text{Cov}(\overline{y}, \hat{\beta}_1)
\end{aligned}$$

Now, using the property we observed in (2.24), we get

$$\begin{aligned}
\text{Var}(\overline{y}) &= \text{Var}\left(\frac{\sum_{i=1}^{n} y_i}{n} \right) \\
&= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(y_i) \\
&= \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}
\end{aligned}$$

since $\text{Var}(y_i) = \sigma^2$. We will leave it as an exercise for you to show that

$$\text{Cov}(\overline{y}, \hat{\beta}_1) = 0$$

Using

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \text{Var}(\overline{y}) = \frac{\sigma^2}{n} \quad \text{and} \quad \text{Cov}(\overline{y}, \hat{\beta}_1) = 0$$

we conclude that

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \text{Var}(\overline{y} - \hat{\beta}_1 \overline{x}) \\
&= \text{Var}(\overline{y}) + \overline{x}^2 \text{Var}(\hat{\beta}_1) - 2\overline{x}\text{Cov}(\overline{y}, \hat{\beta}_1) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)
\end{aligned}$$

These are the facts that we set up to show.

In general σ^2 is not known, but can be estimated from the data. This estimation of σ is known as the *residual standard error*, and is given by the formula

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}.$$

Remark: Strictly speaking, when σ^2 is estimated from the data we should write $\widehat{\text{SE}}(\widehat{\beta})_1$ to indicate that an estimate has been made, but for simplicity of notation we will drop this extra "hat".

RSE^2 turns out to be an unbiased estimator of σ^2. To understand this, we need to understand the various "sum of squares" defined below:

- Total Sum of Squares (Also denoted as TSS)

$$\text{SS}_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{2.25}$$

- Regression Sum of Squares

$$\text{SS}_R = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \tag{2.26}$$

- Residual Sum of Squares

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.27}$$

Here is a relationship between the three quantities above:

Lemma 2.1.

$$SS_T = SS_R + RSS \tag{2.28}$$

Proof. Notice that

$$\begin{aligned}
\text{SS}_T &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\
&= \text{RSS} + \text{SS}_R + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y})
\end{aligned}$$

To complete the proof, we need to show that the sum of the cross-product term is zero:

Consider the sum of the cross-product term:

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y})$$

Multiply the expression out to obtain

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)\widehat{y}_i - \overline{y}\sum_{i=1}^{n}(y_i - \widehat{y}_i). \tag{2.29}$$

Since $\sum_{i=1}^n (y_i - \widehat{y}_i) = 0$ by the property of residuals in the least squares estimate, see (2.8), the second term of (2.29) vanishes, hence

$$\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \sum_{i=1}^n (y_i - \widehat{y}_i)\widehat{y}_i. \tag{2.30}$$

The sum $\sum_{i=1}^n (y_i - \widehat{y}_i)\widehat{y}_i$ in (2.30) is also zero. To see this, notice that $\widehat{y}$ is a linear combination of x_i, as $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Now from (2.8) and (2.9) we have

$$\sum_{i=1}^n (y_i - \widehat{y}_i)\, \widehat{y}_i = \widehat{\beta}_0 \underbrace{\sum_{i=1}^n (y_i - \widehat{y}_i)}_{\text{zero by (2.8)}} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^n (y_i - \widehat{y}_i)\, x_i}_{\text{zero by (2.9)}} = 0.$$

Hence we conclude that the sum $\sum_{i=1}^n (y_i - \widehat{y}_i)\widehat{y}_i$ in (2.30) is zero, therefore the sum of the cross-product in (2.29) is zero and that

$$\mathrm{SS}_T = \mathrm{RSS} + \mathrm{SS}_R + 2\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \mathrm{RSS} + \mathrm{SS}_R$$

which is what we would like to show. □

Another useful result we need is

Lemma 2.2.

$$SS_R = \widehat{\beta}_1^2 S_{xx} \tag{2.31}$$

Proof. This is because

$$\begin{aligned}
\mathrm{SS}_R &= \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^n [\underbrace{(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}_{\widehat{y}_i} - \underbrace{(\widehat{\beta}_0 + \widehat{\beta}_1 \overline{x})}_{\overline{y},\ \text{see (2.4)}}]^2 \\
&= \sum_{i=1}^n \widehat{\beta}_1^2 (x_i - \overline{x})^2 \\
&= \widehat{\beta}_1^2 S_{xx}
\end{aligned}$$

□

Lastly, to understand the estimation of σ^2, we need to know how to compute the expected value of $\overline{y}^2$.

Lemma 2.3. *In a simple linear regression model with $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are independent and identically distributed with mean 0 and variance σ^2, we have*

$$E(\bar{y}^2) = (\beta_0 + \beta_1 \bar{x})^2 + \frac{\sigma^2}{n}$$

Proof. Let us compute $E(\bar{y}^2)$. First, express $\bar{y}$ in terms of the model components:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum_{i=1}^n \epsilon_i$$

Now, square $\bar{y}$ and expand out the expression to obtain

$$\bar{y}^2 = (\beta_0 + \beta_1\bar{x})^2 + 2(\beta_0 + \beta_1\bar{x})\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2$$

Take the expectation of both sides we see that

$$E[(\bar{y})^2] = \underbrace{E\left[(\beta_0 + \beta_1\bar{x})^2\right]}_{\text{Part I}} + \underbrace{2E\left[(\beta_0 + \beta_1\bar{x})\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)\right]}_{\text{Part II}} + \underbrace{E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right]}_{\text{Part III}}$$

Evaluate each part separately.

For Part I, since $\beta_0 + \beta_1\bar{x}$ is a constant, we get

$$E\left[(\beta_0 + \beta_1\bar{x})^2\right] = (\beta_0 + \beta_1\bar{x})^2$$

For Part II, since $E[\epsilon_i] = 0$, we obtain

$$E\left[2(\beta_0 + \beta_1\bar{x})\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)\right] = 2(\beta_0 + \beta_1\bar{x}) \cdot E\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right] = 0$$

For Part III, We have

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}E[\epsilon_i^2] + \frac{1}{n^2}\sum_{i\neq j}E[\epsilon_i\epsilon_j]$$

Since the errors are independent (a weaker assumption such as uncorrelated will do) we can write

$$E[\epsilon_i^2] = \sigma^2 \quad \text{as} \quad E[\epsilon] = 0$$

$$E[\epsilon_i\epsilon_j] = 0 \text{ for } i \neq j$$

Therefore

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Combining Part I, II and III, we conclude that

$$E[\bar{y}^2] = (\beta_0 + \beta_1\bar{x})^2 + \frac{\sigma^2}{n}$$

□

We now see how to use the residual sum of squares to estimate σ^2, the errors' variance, when it is unknown. To do this, we need to establish the fact that

$$E(\text{RSS}) = (n-2)\sigma^2. \tag{2.32}$$

The Equation 2.32 then immediately implies that

$$\frac{\text{RSS}}{n-2}$$

is an unbiased estimator for σ^2 as

$$E\left(\frac{\text{RSS}}{n-2}\right) = \frac{1}{n-2}E(\text{RSS}) = \frac{1}{n-2}(n-2)\sigma^2 = \sigma^2$$

Theorem 2.1. *Using our regression set up, we have* $E(RSS) = (n-2)\sigma^2$.

Proof. From (2.28) we write $\text{RSS} = \text{SS}_T - \text{SS}_R$. Using the alternative form of SS_T and (2.31) we write

$$\text{RSS} = \text{SS}_T - \text{SS}_R = \underbrace{\left(\left[\sum_{i=1}^{n} y_i^2\right] - n\overline{y}^2\right)}_{\text{SS}_T} - \widehat{\beta}_1^2 S_{xx}.$$

Using the formula

$$E(X^2) = E(X)^2 + \text{Var}(X) \tag{2.33}$$

for any random variable X, we have

$$\begin{aligned}
E(\text{RSS}) &= \left(\sum_{i=1}^{n} E(y_i^2)\right) - nE(\overline{y}^2) - E(\widehat{\beta}_1^2)S_{xx} \\
&= \sum_{i=1}^{n} \underbrace{\left[E(y_i)^2 + \text{Var}(y_i)\right]}_{\text{using (2.33) for } E(y_i^2)} - nE(\overline{y}^2) - E(\widehat{\beta}_1^2)S_{xx} \\
&= \sum_{i=1}^{n} \left[(\beta_0 + \beta_1 x_i)^2 + \sigma^2\right] - n\underbrace{\left[(\beta_0 + \beta_1\overline{x})^2 + \frac{\sigma^2}{n}\right]}_{\text{Lemma 2.3}} - \underbrace{\left(\beta_1^2 + \frac{\sigma^2}{S_{xx}}\right)S_{xx}}_{\text{using (2.23) and (2.33) for } E(\widehat{\beta}_1^2)} \\
&= \left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)^2\right] + n\sigma^2 - n(\beta_0 + \beta_1\overline{x})^2 - \sigma^2 - \beta_1^2 S_{xx} + \sigma^2 \\
&= (n-2)\sigma^2.
\end{aligned}$$

Remark: The last line in the equation above is obtained by expanding everything out as well as using $S_{xx} = (\sum_{i=1}^{n} x_i^2) - n\overline{x}^2$.

□

Remark: Take note of other terminologies. In

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-2} = \text{MS}_{\text{Res}}$$

the quantity MS_{Res} is called the *residual mean square*, while the square root of $\widehat{\sigma}^2$, that is,

$$\sqrt{\frac{\text{RSS}}{n-2}}$$

is the residual standard error, RSE, which is also known as the *standard error of regression*. Also, because $\widehat{\sigma^2}$ depends on the residual sum of squares, any violation of the assumptions on the model errors or any missspecification of the model form may seriously damage the usefulness of $\widehat{\sigma^2}$ as an estimate of σ^2.

Standard errors can be used to compute *confidence intervals*. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. The range is defined in terms of lower and upper limits computed from the sample data. A 95% confidence interval has the following property: if we take repeated samples and construct the confidence interval for each sample, 95% of the intervals will contain the true unknown value of the parameter.

For linear regression, the 95% confidence interval for β_1 approximately takes the form

$$\widehat{\beta}_1 \pm 2 \cdot \text{SE}(\widehat{\beta}_1) \tag{2.34}$$

That is, there is approximately a 95% chance that the interval

$$\Big(\widehat{\beta}_1 - 2 \cdot \text{SE}(\widehat{\beta}_1),\ \widehat{\beta}_1 + 2 \cdot \text{SE}(\widehat{\beta}_1)\Big) \tag{2.35}$$

will contain the true value of β_1. Similarly, a confidence interval for β_0 approximately takes the form

$$\widehat{\beta}_0 \pm 2 \cdot \text{SE}(\widehat{\beta}_0) \tag{2.36}$$

Remark: Equations (2.34) and (2.36) are only approximation, we will give the precise form later in this section.

In the case of the TV advertising data, the 95% confidence interval for β_0 is (6.130, 7.935) and the 95% confidence interval for β_1 is (0.042, 0.053). Therefore, we can conclude that in the absence of any TV advertising, sales will, on average, fall somewhere between 6,130 and 7,935 units. Furthermore, for each \$1,000 increase in TV advertising, there will be an average increase in sales of between 42 and 53 units.

Standard errors can also be used to perform *hypothesis test* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

$$H_0 : \text{There is no relationship between } X \text{ and } Y \tag{2.37}$$

versus the *alternative hypothesis*

$$H_a : \text{There is some relationship between } X \text{ and } Y \tag{2.38}$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

since if $\beta_1 = 0$ then the model $Y = \beta_0 + \beta_1 X + \epsilon$ reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y. To test the null hypothesis, we need to determine whether $\widehat{\beta}_1$, our estimate for β_1, is sufficiently far from zero that we can be confident that β_1 is non-zero. We now provide the details on how this is done. This in turns will also give a precise version of the confidence interval described above.

Notice that we have not prescribe a distribution for the error terms ϵ_i in all of our previous discussion. What we have assumed were that the errors ϵ_i in $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ have mean zero and constant variance σ^2. However, to perform the hypothesis testing and interval estimation in regression, we need to assume additionally that the errors ϵ_i are *independent and identically distributed normal random variables* with mean zero and variance σ^2, and write

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Hence the observations y_i are now normal and independently distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2, and we write

$$y_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Notice that the mean depend on the value of x_i, hence the distribution will not be identical for all i, but these are still independent.

Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{S_{xx}} = \sum_{i=1}^n c_i y_i$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}, \quad i = 1, \cdots, n$$

hence from

$$y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

we know that $\hat{\beta}_1$ is also normally distributed, as it is the sum of independent and normally distributed random variables, y_i. Also, since $E(\widehat{\beta}_1) = \beta_1$ and $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2/S_{xx}$, see (2.23), we get

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Since $\hat{\beta}_0$ can also be written as a linear combination of y_i, the distribution of $\hat{\beta}_0$ will also be normal with mean and variance as shown below, see also (2.22)

$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$$

Back to $\widehat{\beta}_1$. Since

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathrm{N}(0,1)$$

hence

$$\sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathrm{N}(0,1)$$

and it is actually independent of

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-2}.$$

Now, recall from your previous statistics classes for the following fact:

If Z and X_n^2 are independent r.v. with $Z \sim \mathrm{N}(0,1)$ and X_n^2 has a chi-square distribution with n degree of freedom, then the r.v. T_n defined below has a t-distribution with n degree of freedom, that is

$$T_n := \frac{Z}{\sqrt{X_n^2/n}} \sim t_n. \tag{2.39}$$

Applying (2.39) to

$$\sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathrm{N}(0,1) \quad \text{and} \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

we obtain

$$\begin{aligned} T_{n-2} &:= \frac{\sqrt{S_{xx}}(\hat{\beta}_1 - \beta_1)/\sigma}{\sqrt{\frac{RSS/\sigma^2}{n-2}}} \\ &= \sqrt{\frac{(n-2)S_{xx}}{RSS}}(\hat{\beta}_1 - \beta_1) \sim t_{n-2} \end{aligned} \tag{2.40}$$

With the conclusion established (2.40), we can get back to the question on whether $\beta_1 = 0$ for

$$y = \beta_0 + \beta_1 x + \epsilon$$

Let us perform hypothesis testing:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0 \tag{2.41}$$

With the significance level set at the level α, we

$$\begin{aligned} \text{Reject} \quad H_0, \quad \text{if} \quad \sqrt{\frac{(n-2)S_{xx}}{RSS}}|\hat{\beta}_1| > t_{\alpha/2,n-2} \\ \text{Do not reject} \quad H_0, \quad \text{otherwise} \end{aligned} \tag{2.42}$$

Here, $t_{\alpha/2,n-2}$ is the value such that $P(t_{n-2} > t_{\alpha/2,n-2}) = \alpha/2$.

Another way of saying this, is that, the test can be performed by computing the value of the test statistic, ν, (a number computed from the data) below

$$\nu = \sqrt{\frac{(n-2)S_{xx}}{RSS}}\hat{\beta}_1$$

then rejecting H_0 if $|\nu| > t_{\frac{\alpha}{2},n-2}$.

Also, it follows from

$$\sqrt{\frac{(n-2)S_{xx}}{RSS}}(\hat{\beta}_1 - \beta_1) \sim t_{n-2}$$

that, for any $\alpha, 0 < \alpha < 1$, a $100(1-\alpha)\%$ confidence interval for the true β_1 is

$$\left(\hat{\beta}_1 - \sqrt{\frac{RSS}{(n-2)S_{xx}}}\ t_{\alpha/2,n-2}, \quad \hat{\beta}_1 + \sqrt{\frac{RSS}{(n-2)S_{xx}}}\ t_{\alpha/2,n-2}\right)$$

In order to have a better understanding of the above mentioned procedure on hypothesis testing, it is best that we review some of the related concepts here.

After we stated the hypotheses (2.41) and selected the significance level α. We have to choose a decision rule. That is, if $\widehat{\beta}_1$ (computed from data) lays outside of the interval $(-c, c)$, then we will reject H_0. Hence $|\widehat{\beta}_1 - 0| > c$ is referred to as a rejection region of the test, see Fig 2.4.

We next need to determine the size of c. To do that, we will have to refer to the level of significance α, which is defined to be the probability of committing a type-I error based on the rule we just established. Here, the type-I error is defined as the probability of rejecting H_0 given that H_0 is true. See the table below

Table 2.1: Type I and Type II Errors in Hypothesis Testing

Actual State	**Decision: Reject** H_0	**Decision: Do not reject** H_0
H_0 is true	Type I Error, α	Correct Decision
H_0 is false	Correct Decision	Type II Error, β

In our case, this is written as

$$\begin{aligned}\alpha &= P\,(\text{ Reject } H_0 \mid \text{H}_0 \text{ is true}) \\ &= P\left(\left|\widehat{\beta}_1 - \beta_1\right| > c \mid \text{H}_0 \text{ is true}\right) \\ &= P\left(\left|\widehat{\beta}_1 - 0\right| > c\right) \end{aligned} \tag{2.43}$$

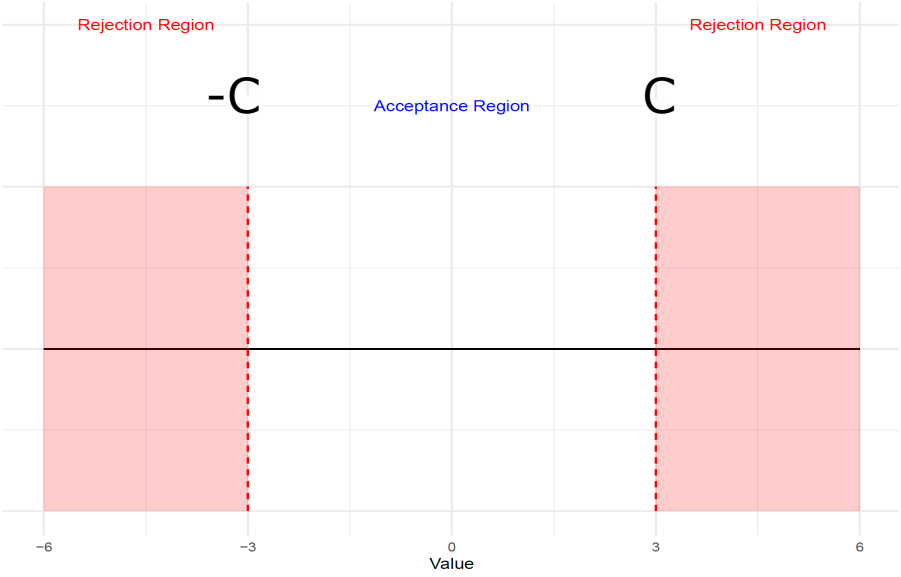


Figure 2.4: The figure above show the decision rule of the hypothesis test in (2.41), namely if $\widehat{\beta}_1$ (computed from the data) lays outside of the interval $(-c, c)$, then we will conclude that the true population parameter $\beta_1 \neq 0$. In other words, if $\widehat{\beta}_1$ lays in the region $|\widehat{\beta}_1 - 0| > c$, then we will reject H_0 and conclude that $\beta_1 \neq 0$.

Therefore, continue from (2.43) we get

$$\begin{aligned}\alpha &= P\left(\left|\widehat{\beta}_1 - 0\right| > c\right) \\ &= P\left(\widehat{\beta}_1 > c\right) + P\left(\widehat{\beta}_1 < -c\right) \\ &= P\left(\sqrt{\frac{(n-2)S_{xx}}{RSS}}\widehat{\beta}_1 > c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) + P\left(\sqrt{\frac{(n-2)S_{xx}}{RSS}}\widehat{\beta}_1 < -c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) \\ &= P\left(t_{n-2} > c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) + P\left(t_{n-2} < -c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right), \end{aligned} \tag{2.44}$$

this is because

$$\sqrt{\frac{(n-2)S_{xx}}{RSS}}(\hat{\beta}_1 - \beta_1) \sim t_{n-2},$$

with $\beta_1 = 0$, see (2.40). Since the t-distribution is symmetrical, we can write (2.44) as

$$\alpha = 2P\left(t_{n-2} > c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right)$$

or

$$P\left(t_{n-2} > c\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) = \frac{\alpha}{2}.$$

Let us denote $t_{\alpha/2,n-2}$ to be the value (on x-axis, as indicated by the right red arrow, in the Fig 2.5 below) such that

$$P(t_{n-2} > t_{\alpha/2,n-2}) = \frac{\alpha}{2}$$

then

$$c\sqrt{\frac{(n-2)S_{xx}}{RSS}} = t_{\alpha/2,n-2} \tag{2.45}$$

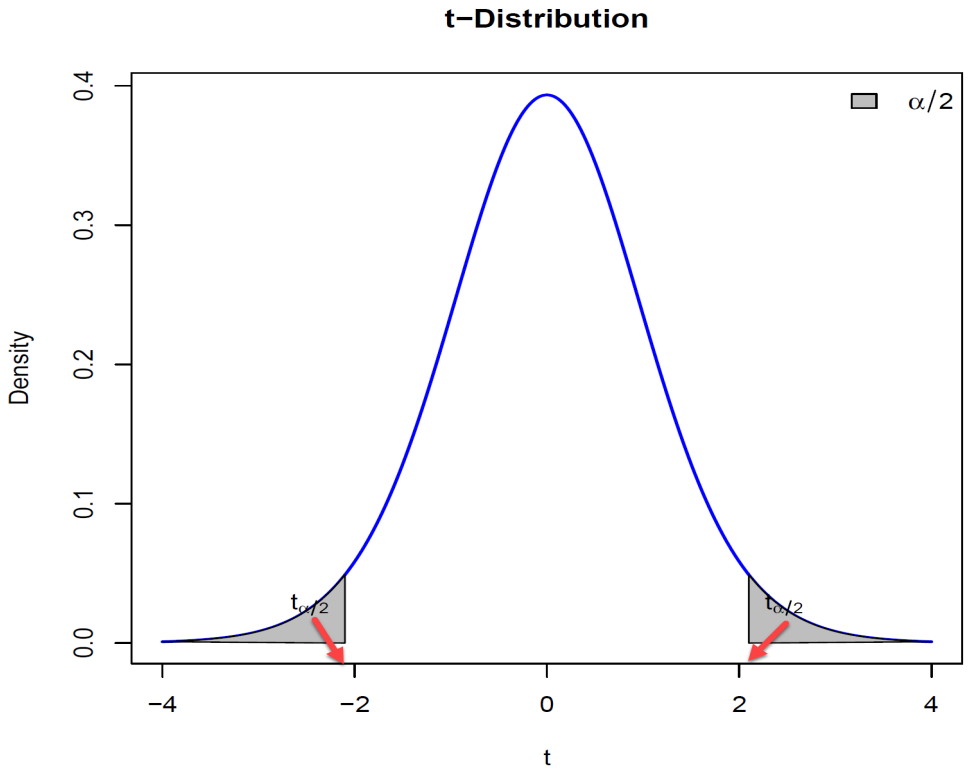


Figure 2.5: This is a general plot for a t-distribution (with $df = n-2$) where the shaded region has a total area of α. Hence, by symmetry each part of the shaded regions has an area of $\alpha/2$.

Therefore, from (2.45) we see that

$$c = t_{\alpha/2,n-2}\left(\frac{(n-2)S_{xx}}{RSS}\right)^{-\frac{1}{2}}$$

Hence for the hypothesis test (2.41) with α predetermined, our decision rule says that:

Reject H_0 if

$$|\widehat{\beta}_1 - 0| > c = t_{\alpha/2,n-2}\left(\frac{(n-2)S_{xx}}{RSS}\right)^{-\frac{1}{2}} \tag{2.46}$$

or when

$$\sqrt{\frac{(n-2)S_{xx}}{RSS}}|\widehat{\beta}_1| > t_{\alpha/2,n-2}. \tag{2.47}$$

which is what we saw in (2.42).

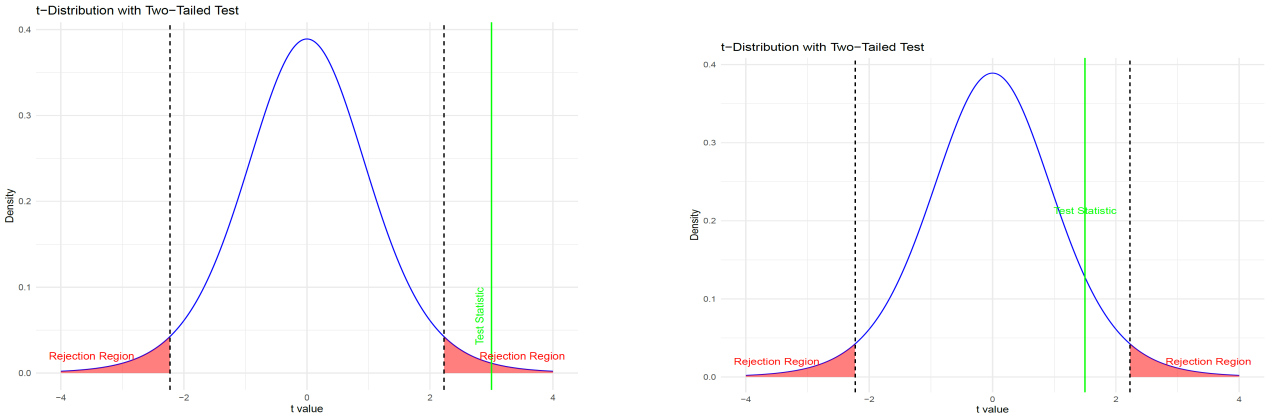
For simplicity of notation let us denote ν to be the *test statistic* shown below

$$\nu = \sqrt{\frac{(n-2)S_{xx}}{RSS}}\ \widehat{\beta}_1 \tag{2.48}$$

With that we will reject H_0 if

$|\nu| > t_{\frac{\alpha}{2},n-2}$. This is how we obtain the result we presented earlier.

Note that $t_{\frac{\alpha}{2},n-2}$ is the value on the x-axis, as indicated by the right red arrow, in the Fig 2.5 while $-t_{\frac{\alpha}{2},n-2}$ is the value indicated by the left red arrow in the same figure mentioned earlier.



(a) The test statistic $|\nu| > t_{\frac{\alpha}{2},n-2}$ and H_0 is rejected

(b) H_0 failed to be rejected as $|\nu| \ngtr t_{\frac{\alpha}{2},n-2}$

Figure 2.6: The figure shows when to reject H_0 based on the location of the test statistic ν.

Remark: Since we also write

$$\frac{RSS}{n-2} = MS_{Res}$$

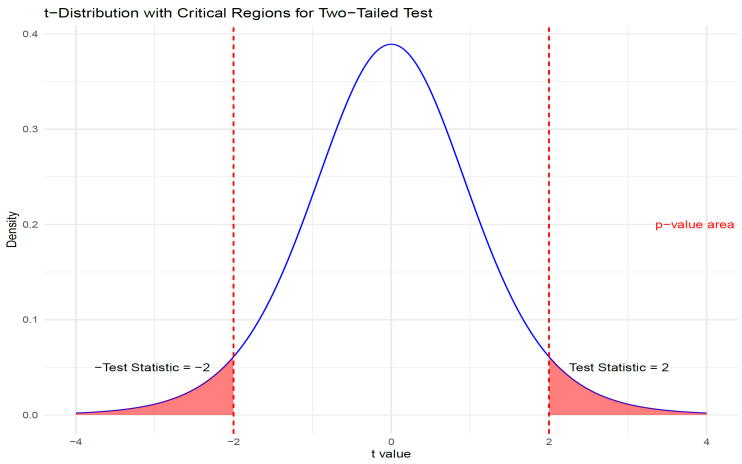
the test statistic ν can also be written as

$$\nu = \sqrt{\frac{(n-2)S_{xx}}{RSS}}\ \widehat{\beta}_1 = \frac{\widehat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}.$$

Another very common tool for hypothesis testing is using the concept of *p-value*. The p-value is the probability of observing a t-statistic as extreme as, or more extreme than, the observed value under the null-hypothesis.

Roughly speaking we interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.

Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response. We *reject the null hypothesis* - that is, we declare a relationship to exist between X and Y - if the p-value is small enough. Typical p-value cutoffs for rejecting the null hypothesis are 5% or 1%.



As an example, suppose the test statistic computed from the data is ν, then the p-value of the test is given by

$$p - \text{value} = 2 \times P(T > |\nu|)$$

The figure on the left assume that the computed test statistic is 2, hence the area of the shaded region is the p-value of the test.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Figure 2.7: For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of \$1,000 in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the sales) variable is in thousands of units, and the TV variable is in thousands of dollars.

Figure 2.7 provides details of the least squares model for the regression of the number of units sold on TV advertising budget for the Advertising data. Notice that the coefficients for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are very large relative to their standard errors, so the t-statistics are also large. The probability of seeing such values if H_0 is true are virtually zero. Hence we can conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

As for constructing a $(1-\alpha) \times 100\%$ confidence interval of a population parameter, say, θ, what we need to do is to build an interval with length $2m$ such that

$$P(\theta \in (\widehat{\theta} - m, \widehat{\theta} + m)) = 1 - \alpha$$

after α has been predetermined. After that, we need to determine how big is the value of m. Hence the $(1-\alpha) \times 100\%$ confidence interval for β_1 is constructed by computing the value of m such that

$$P(\beta_1 \in (\widehat{\beta}_1 - m, \widehat{\beta}_1 + m)) = 1 - \alpha$$

Equivalently we have

$$\begin{aligned} P\left(-m < \widehat{\beta}_1 - \beta_1 < m\right) &= P\left(-m\sqrt{\frac{(n-2)S_{xx}}{RSS}} < \sqrt{\frac{(n-2)S_{xx}}{RSS}}\left(\widehat{\beta}_1 - \beta_1\right) < m\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) \\ &= P\left(-m\sqrt{\frac{(n-2)S_{xx}}{RSS}} < t_{n-2} < m\sqrt{\frac{(n-2)S_{xx}}{RSS}}\right) = 1 - \alpha \quad (2.49) \end{aligned}$$

From here, we see that (also refer to Fig 2.8))

$$m\sqrt{\frac{(n-2)S_{xx}}{RSS}} = t_{\alpha/2,n-2}$$

therefore

$$m = \sqrt{\frac{RSS}{(n-2)S_{xx}}} t_{\alpha/2,n-2}$$

and from (2.49) we conclude that the $(1-\alpha) \times 100\%$ confidence interval of β_1 is given by

$$\left(\widehat{\beta}_1 - \sqrt{\frac{RSS}{(n-2)S_{xx}}} t_{\alpha/2,n-2},\ \widehat{\beta}_1 + \sqrt{\frac{RSS}{(n-2)S_{xx}}} t_{\alpha/2,n-2} \right)$$

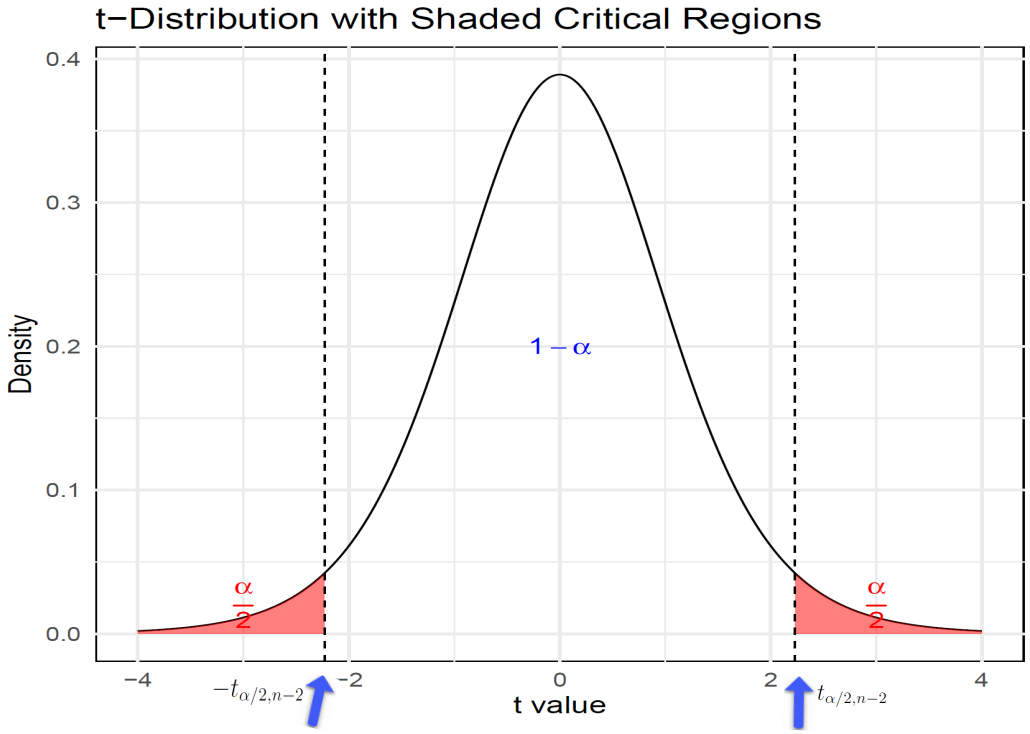


Figure 2.8: This is a general plot for a t-distribution (with $df = n-2$) where the shaded region has a total area of α . Notice that the value $t_{\alpha/2,n-2}$ on the x-axis indicated by the blue right arrow is a number such that $P(t_{n-2} > t_{\alpha/2,n-2}) = \alpha/2$.

2.1.3 Assessing the Accuracy of the Model

Once we have rejected the null hypothesis (2.37)

$$H_0 : \text{There is no relationshiop between } X \text{ and } Y$$

in favor of the alternative hypothesis (2.38)

$$H_a : \text{There is some relationshiop between } X \text{ and } Y$$

it is natural to want to quantify *the extent to which the model fits the data*. The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

Figure (2.9) displays the RSE, the R^2 statistic, and the F-statistic (to be discussed later) for the linear regression of number of units sold on TV advertising budget.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Figure 2.9: For the Advertising data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

Residual Standard Error

Recall from the model (2.15)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

that associated with each observation is an error term ϵ. Due to the presence of these error terms, even if we knew the true regression line (i.e. even if β_0 and β_1 were known), we would not be able to perfectly predict Y from X. The RSE is an estimate of the standard deviation of ϵ. (See Theorem 2.1).

$$\underbrace{Y - (\beta_0 + \beta_1 X)}_{\text{observation deviate from the true true regression line}} = \underbrace{\epsilon}_{\text{RSE ia an estimate of the standard deviation of } \epsilon}$$

Roughly speaking, RSE is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}. \tag{2.50}$$

In the case of the advertising data, we see from the linear regression output in Fig 2.9 that the RSE is 3.26. In other words, actual sales in each market deviate from the true regression line by approximately 3,260 units, on average.

Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average. Of course, whether or not 3,620 units is an acceptable prediction error depends on the problem context. In the advertising data set, the mean values of sales over all markets is approximately 14,000 units, and so the percentage error is $3,260/14,000 = 23\%$.

The RSE is considered a measure of the *lack of fit* of the model (2.15) to the data. If the predictions obtained using the model are very close to the true outcome values, that is, $\widehat{y}_i \approx y_i$, for $i = 1, \cdots, n$, then (2.50) will be small, and we can conclude that the model fits the data very well. (Remark: The prediction $\widehat{y}_i \approx y_i$, for $i = 1, \cdots, n$ will make RSE small (see (2.50)), hence the deviation of the response, y, and the true regression line, $\beta_0 + \beta_1 x$, will be small, on average. In this sense, we say that the model fit the data well). On the other hand, if $\widehat{y}_i$ is very far from y_i for one or more observations, then RSE could be quite large, indicating that the model doesn't fit the data well.

2.1.4 R^2 Statistic

The RSE provides an absolute measure of lack of fit of the model (2.15) to the data. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit. It takes the form of a *proportion*–the proportion of variance explained–and so it always takes on a value between 0 and 1.

To calculate R^2, we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2.51}$$

where from (2.25) and (2.27) we see that

$$\text{Total Sum of Squares} = \text{TSS} = SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

while

$$\text{Residual Sum of Squares} = \text{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 .$$

Recall, from Lemma 1, namely, $\text{TSS} = \text{SS}_R + \text{RSS}$, hence the term $\text{TSS} - \text{RSS}$ is actually the regression sum of squares, $\text{SS}_R = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$, (see (2.26)).

In order to understand R^2 better, let us look at the following quantities (also, see Fig 2.10):

$$\underbrace{y_i}_{\text{observed}} = \underbrace{\widehat{y}_i}_{\text{fit}} + \underbrace{(y_i - \widehat{y}_i)}_{\text{deviation from fit}} \tag{2.52}$$

Subtracting $\overline{y}$ from both sides of (2.52), we obtain

$$\underbrace{y_i - \overline{y}}_{\text{deviation from mean}} = \underbrace{\widehat{y}_i - \overline{y}}_{\substack{\text{deviation due to fit} \\ \text{(explained variation)}}} + \underbrace{(y_i - \widehat{y}_i)}_{\substack{\text{residual (unexplained} \\ \text{variation)}}}$$

First, note that the term *variability* (variation and so on) refers to the degree to which data points in a set differ from each other and from the mean of the data set. It indicates how spread out or dispersed the values are within a data set. High variability means the data points are more spread out from the mean, while low variability indicates they are closer to the mean.

Second, the term $\widehat{y}_i - \overline{y}$ represents the deviation of the predicted value $\widehat{y}_i$ from the mean of the observed value, namely, $\overline{y}$. Essentially, it indicates how much the predicted value (based on the regression model) deviates from the mean of y. This term has been labeled above as "explained variation" because the term captures the extend to which the regression model accounts for the variability in the observed data.

Third, the terminology *unexplained variation* means variation that cannot be explained by the relationship due to chance or other variables. This is the residual or leftover variation in the dependent variable y that is not explained by the independent variable(s) in the regression model. It represents the discrepancies between the actual observed values of y and the values predicted by the regression model. In summary, the unexplained variation, $y_i - \widehat{y}_i$ captures the portion of variability in the dependent variable that the regression model fails to explain, highlighting the presence of factors or sources of variation not accounted for by the model.

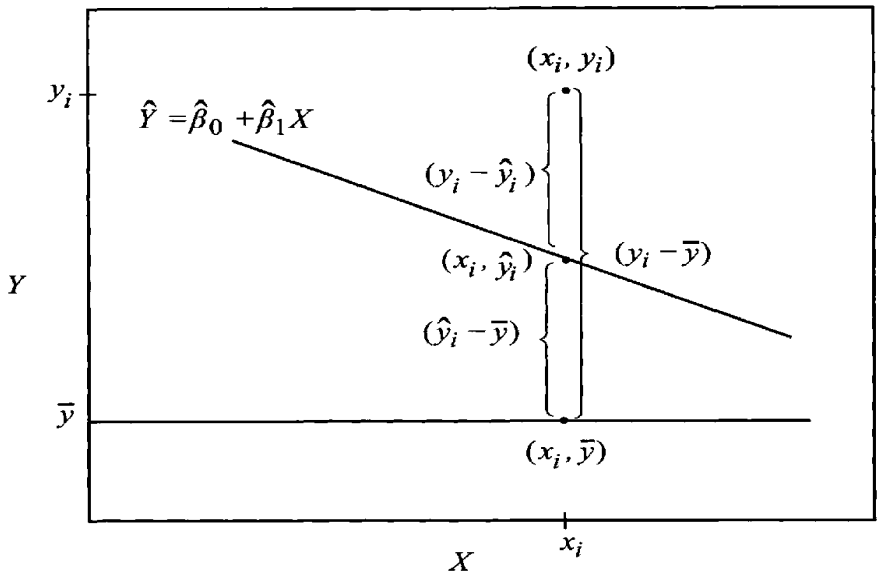


Figure 2.10: The figure above illustrates various quantities computed after fitting a regression line to data.

Hence we see that TSS measures the total variance in the response Y, and can be thought of as the amount of variability in the response before the regression is performed. (Note: The computation of TSS doesn't involve any information related to regression). In contrast, RSS measures the amount of variability that is left unexplained after performing the regression line.

Therefore $\text{TSS} - \text{RSS} = \text{SS}_R$ measures the amount of variability in the response that is explained by performing the regression, and R^2 measures the *proportion of variability in* Y *that can be explained using* X.

From

$$R^2 = \frac{\overbrace{\text{TSS} - \text{RSS}}^{\text{variability in } Y \text{ that is explained using } X}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

we see that an R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response is explained by the regression, since the term RSS is much smaller than TSS (or since $\text{TSS} - \text{RSS} \approx \text{TSS}$). Intuitively, you can think of the term RSS as the "bad term", so when RSS is small, the "good part", SS_R will be bigger.

A number near 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance σ^2 is high, or both. Intuitively, R^2 near 0, means RSS is almost equal to TSS, and since $\text{TSS} = \text{SS}_R + \text{RSS}$, in this situation, the "good term" SS_R will be small. Therefore, we have the conclusion stated above.

In the Fig 2.9, the R^2 was 0.61, and so just under two-thirds of the variability in sales is explained by a linear regression on TV.

The R^2 statistic in (2.51) has an interpretational advantage over the RSE (2.50), since unlike RSE, it always lie between 0 and 1. However, it can still be challenging to determine what is a *good* R^2 value, and in general, this will depend on the application. For instance, in certain problems in physics, we may know that the data truely comes from a linear model with a small

residual error. In this case, we would expect to see an R^2 value that is extremely close to 1, and a substantially smaller R^2 value might indicate a serious problem with the experiment in which the data were generated.

On the other hand, in typical applications in biology, psychology, marketing, and other domains, the linear model (2.15) is at best an extremely rough approximation to the data, and the residual errors due to other unmeasured factors are often very large. In this setting, we would expect only a very small proportion of the variance in the response to be explained by the predictor, and an R^2 value well below 0.1 might be more realistic.

The R^2 statistic is a measure of the linear relationship between X and Y. Recall that *correlation*, defined as

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{2.53}$$

is also a measure of the linear relationship between X and Y. This suggests that we might be able to use $r = \text{Cor}(X, Y)$ instead of R^2 in order to access the fit of the linear model. In fact, it can be shown that in the simple linear regression setting, $R^2 = r^2$. In other words, the squared correlation and the R^2 statistic are identical. However, in the next section we will discuss the multiple linear regression problem, in which we use several predictors simultaneously to predict the response. The concept of correlation between the predictors and the response does not extend automatically to this setting, since correlation quantifies the association between a single pair of variables rather than between a large number of variables. We will see that R^2 fills this role.

2.1.5 Summary

At the beginning of this Chapter, we have posted a series of questions which some of them we can now answer.

- Is there a relationship between advertising budget and sales?

In the case of simple regression where we regress sales onto TV advertising budget, based on table below we see that, with high certainty, there is a relationship between the TV advertis-

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

ing budget and sales. This can be seen from the extremely low p-value associated with the hypothesis testing on whether or not $\beta_1 = 0$.

- How strong is the relationship between advertising budget and sales?

The strength of the relationship between the TV advertising budget and sales can be obtained by looking at the correlation r. This measures the strength and direction of the linear relationship between two variables. The value of r ranges from -1 to 1. A value close to 1 indicates a strong positive relationship, a value close to -1 indicates a strong negative relationship, and a value close to 0 indicates no linear relationship.

We can also obtain the strength of the relationship by looking at the R^2 statistic. In simple linear regression, this is the square of the correlation r mentioned earlier. The test statistic

R^2 measures the proportion of variability in the dependent variable (sales) that can be explained using the independent variable (TV advertising budget). An R^2 close to 1 indicates a strong relation. Based on the table below, we see that R^2 is 0.612. In our context, this could be seen as a reasonably strong relationship between the TV advertising budget and sales.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

- How accurately can we predict future sales?

To measure how accurately we can predict future sales using a regression model, several metrics can be utilized. One such metric is the R^2 statistic. While R^2 is primarily used to measure how well the model fits the data, it can also give an indication of predictive power. A higher R^2 suggests that the model explains a large portion of the variation and thus likely to predict future outcomes more accurately. See also the discussion in Section 2.2.3.

Other questions raised at the beginning of this Chapter can be answered after the discussion of multiple linear regression. See Section 2.2.3.

2.2 Multiple Linear Regression

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. For example, in the Advertising data, we have examined the relationship between sales and TV advertising. We also have data for the amount of money spent advertising on the radio and in newspapers, and we may want to know whether either of these two media is associated with sales. How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

One option is to run three separate simple linear regressions, each of which uses a different advertising medium as a predictor. For instance, we can fit a simple linear regression to predict sales on the basis of the amount spent on radio advertisements. Results are shown in Fig 2.11 (top figure). We find that a \$1,000 increase in spending on radio advertising is associated with an increase of around 203 units. Figure 2.11 (bottom figure) contains the least square coefficients for a simple linear regression of sales onto newspaper advertising budget. A \$1,000 increase in newspaper advertising budget is associated with an increase in sales of approximately 55 units.

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Figure 2.11: More simple linear regression models for the Advertising data. Coefficients of the simple linear regression model for number of units sold on Top: radio advertising budget and Bottom: newspaper advertising budget. A \$1,000 increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units. (Note that the sales variable is in thousands of units, and the radio and newspaper variables are in thousand of dollars.)

However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory. First of all, it is unclear how to make a single prediction of sales given the three advertising media budgets, since each of the budgets is associated with a separate regression equations. Second, each of the three regression equations ignores the other two media in forming estimates for the regression coefficients. We will see shorty that if the media budgets are correlated with each other in the 200 markets in our data set, then this can lead to very misleading estimates of the association between each media budget and sales.

Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model (2.15) so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.54}$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*. In the advertising example, (2.54) becomes

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon. \tag{2.55}$$

2.2.1 Estimating the Regression Coefficients

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \cdots, \beta_p$ in (2.54) are unknown, and must be estimated. Given estimates $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$, we can make predictions using the formula

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p \tag{2.56}$$

The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \cdots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_p x_{ip} \right)^2 . \end{aligned} \tag{2.57}$$

The values $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$ that minimize (2.57) are the multiple least squares regression coefficient estimates.

Unlike the simple linear regression estimates given in (2.4), the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra, as demonstrated below.

To see what are these estimators, let us define the least squares function S below

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

We then minimize S with respect to $\beta_0, \beta_1, \cdots, \beta_p$. (Again, note that y_i and x_{ij} are known quantities, the variables in S are the βs. Hence the derivatives are taken with respect to the βs).

For the minimization, the least squares estimators $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$ of $\beta_0, \beta_1, \cdots, \beta_p$ must satisfy

$$\left. \frac{\partial S}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right) = 0 \tag{2.58}$$

and

$$\left. \frac{\partial S}{\partial \beta_k} \right|_{\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right) x_{ik} = 0, \quad k = 1, 2, \cdots, p \tag{2.59}$$

Simplify (2.58) and (2.59) to obtain the **least-squares normal equations**

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} \\
\sum_{i=1}^{n} x_{i1} y_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{i1} x_{ip} \\
\vdots \quad &= \quad \vdots \\
\sum_{i=1}^{n} x_{ip} y_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ip} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ip} x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip}^2
\end{aligned}$$

If we write

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

then

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \cdots, n$$

can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

With that, the normal equations in matrix form can then be written as

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

where

$$\begin{aligned}
\mathbf{X}^T \mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ip} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1} x_{i2} & \cdots & \sum_i x_{i1} x_{ip} \\ \vdots & \vdots & & \vdots & \\ \sum_i x_{ip} & \sum_i x_{ip} x_{i1} & \sum_i x_{ip} x_{i2} & \cdots & \sum_i x_{ip}^2 \end{bmatrix}
\end{aligned}$$

Now, with

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ \vdots \\ \sum_i x_{ip} y_i \end{bmatrix}$$

It is now easy to see that the matrix equation

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

is equivalent to the set of normal equations we derived. Assuming that $\mathbf{X}^T\mathbf{X}$ is invertible, then we can write

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \tag{2.60}$$

Any statistical software package can be used to compute these coefficient estimates, and in this course we will see how this can be done in R. Fig 2.12 illustrates an example of the least squares fit to a toy data set with $p = 2$ predictors.

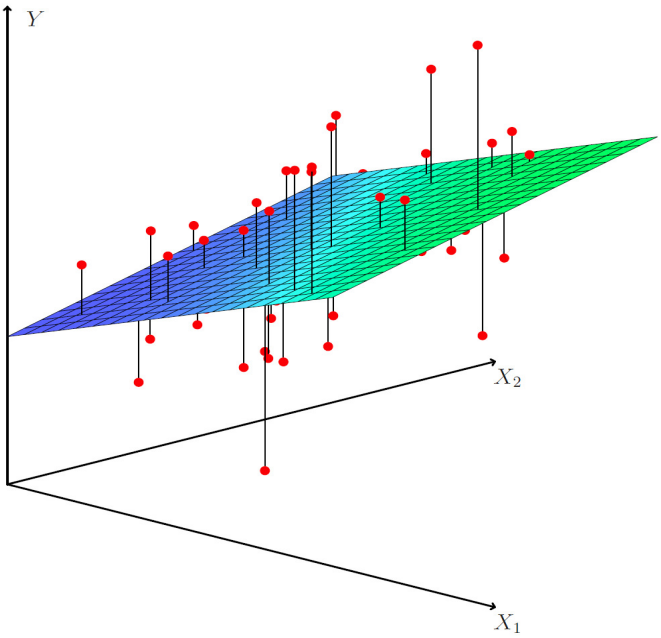


Figure 2.12: In a three-dimenional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of squared vertical distances between each observation (shown in red) and the plane.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Figure 2.13: For the Advertising data, least squared coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.

Figure 2.13 displays the multiple regression coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales using the Advertising data. We interpret these results as follows:

For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising is associated with approximately 189 units of additional sales.

Comparing these coefficient estimates to those displayed in Fig 2.7 and 2.11, we notice that the multiple regression coefficient estimates for TV and radio are pretty similar to the simple linear regression coefficient estimates. However, while the newspaper regression coefficient estimate in Fig 2.11 was significantly non-zero, the coefficient estimate for newspaper in the multiple regression model is close to zero, and the corresponding p-value is no longer significant, with a value around 0.86. This illustrate that the simple and multiple regression coefficients can be quite different. This difference stems from the fact that in the simple regression case, the slope term represents the average increase in product sales associated with $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio. By contrast, in the multiple

regression setting, the coefficient for newspaper represents the average increase in product sales associated with increasing newspaper spending by \$1,000 while holding TV and radio fixed.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Figure 2.14: Above is a summary of information from Fig 2.7, 2.11 and 2.13.

Does it make sense for multiple regression to suggest no relationship between sales and newspaper while the simple regression implies the opposite? In fact it does. Consider the correlation matrix for three predictor variables and response variable, displayed in Fig 2.15. Notice that the correlation between radio and newspaper is 0.35. This indicates that markets with high newspaper advertising tend to also have high radio advertising. Now suppose that the multiple regression is correct and newspaper advertising is not associated with sales, but radio advertising is associated with sales. Then in markets where we spend more on radio our sales will tend to be higher, and as our correlation matrix shows, we also tend to spend more on newspaper advertising in those same markets. Hence, in a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher vales of sales, even though newspaper advertising is not directly associated with sales. So newspaper advertising is a surrogate for radio advertising; newspaper get "credit" for the association between radio on sales.

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Figure 2.15: Correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

This slightly counterintuitive result is very common in many real life situations. Consider an absurd example to illustrate the point. Running a regression of shark attacks versus ice cream sales for data collected at a given beach community over a period of time would show a positive relationship, similar to that seen between sales and newspaper. Of course no one has (yet)

suggested that ice creams should be banned at beaches to reduce shark attacks. In reality, higher temperatures cause more people to visit the beach, which in turn results in more ice cream sales and more shark attacks. A multiple regression of shark attacks onto ice cream sales and temperature reveals that, as intuition implies, ice cream sales is no longer a significant predictor after adjusting for temperature.

2.2.2 Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions listed below:

1. Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting the response?
2. Do all the predictors help explain Y, or is only a subset of predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

We now address each of these questions in turn.

One: Is There a Relationship Between the Response and Predictors?

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether

$$\beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{at least some } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}, \tag{2.61}$$

where, as with simple linear regression, $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ and $\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y})^2$. If the linear model assumptions are correct, one can show that

$$E\left(\frac{\text{RSS}}{n-p-1}\right) = \sigma^2.$$

Note: We have shown this result in the case of simple linear regression model (with $p = 1$), see Theorem 2.1, and that, provided H_0 is true,

$$E\left(\frac{\text{TSS} - \text{RSS}}{p}\right) = \sigma^2.$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $E\left(\frac{\text{TSS}-\text{RSS}}{p}\right) > \sigma^2$, so we expect F to be greater than 1.

Another way to think of (2.61) is the following: Recall that TSS $-$ RSS measures the amount of variability in the response that is explained by performing the regression. While RSS measures the amount of variability that is left unexplained after performing the regression. In the case that H_0 is true, that is the independent variables collectively have no explanatory power, then the term $\frac{\text{TSS}-\text{RSS}}{p} = \text{SS}_R/p$ will be relatively small compared to the mean square due to error/residual (Which we also denote as residual mean square MS_{RES}, see the remark after Theorem 2.1), resulting in a low F-statistic. Conversely, if the independent variables explain a significant portion of the variability in y, then SS_R/p will be large compared to MS_{RES} , resulting in a high F-statistic.

The F-statistic for the multiple linear regression model obtained by regressing sales onto radio, TV, and newspaper is shown in Fig 2.16. In this example, the F-statistic is 570. Since this is far larger than 1, it provides compelling evidence against the null hypothesis H_0. In other words, the large F-statistic suggests that at least one of the advertising media must be related to sales.

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Figure 2.16: More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budget in the Advertising data. Other information about this model was displayed in Fig 2.13

.

However, what if the F-statistic had been closer to 1? How large does the F-statistic need to be before we can reject H_0 and conclude that there is a relationship? It turns out that the answer depends on the values of n and p.

When n is large, an F-statistic that is just a little larger than 1 might still provide enough evidence against H_0. In contrast, a large F-statistic is needed to reject H_0 if n is small. When H_0 is true and the errors ϵ_i have a normal distribution, the F-statistic follows a F-distribution. For any given value of n and p, any statistical software package can be used to compute the p-value associated with the F-statistic using this distribution. Based on this p-value, we can determine whether or not to reject H_0. For the advertising data, the p-value associated with the F-statistic in Fig 2.17 is essentially zero, so we have extremely strong evidence that at least one of the media is associated with increased sales.

In (2.61) we are testing H_0 that all coefficients are zero. Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

where for convenience we have put the variables chosen for omission at the end of the list. In this case we fit a second model that uses all the variables *except* those last q. Suppose that the

```
Call:
lm(formula = df$sales ~ df$TV + df$radio + df$newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
df$TV        0.045765   0.001395  32.809   <2e-16 ***
df$radio     0.188530   0.008611  21.893   <2e-16 ***
df$newspaper -0.001037  0.005871  -0.177     0.86
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figure 2.17: The figure above shows the R-output of regressing sales onto TV, radio and newspaper. Note the p-value of the F-test that is essentially zero, leading to the conclusion of the rejection of H_0.

residual sum of squares for that model is RSS_0. Then the appropriate F-statistic is

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)} \tag{2.62}$$

Notice that in Fig 2.13 (reproduced below), for each individual predictor a t-statistic and a p-value were reported.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

These provide information about whether each individual predictor is related to the response, after adjusting for the other predictors. It turns out that each of these is exactly equivalent to the F-test that omits that single variable from the model, leaving all the others in, i.e. $q = 1$ in (2.62). So it reports the *partial effect* of adding the variable to the model. For instance, as we discussed earlier, these p-values indicate that TV and radio (both have small p-value) are related to sales, but that there is no evidence that newspaper (large p-value) is associated with sales, when TV and radio are held fixed. Let us elaborate more on the discussion above with an example below:

Recall our regression model

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

To test whether newspaper is a significant predictor, we:

1. Fit a full model with all predictors TV, radio, and newspaper.
2. Fit a reduced model without newspaper

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \epsilon$$

3. Perform an F-test comparing these two models. The test evaluates whether the reduction in model fit (i.e. the increase in the sum of squared residuals) when newspaper is removed is statistically significant.

Remark: With fewer predictors, the residuals (the differences between the observed and predicted values) generally become larger. This increase in residuals leads to a higher sum of squared residuals because the model is less effective at capturing the patterns in the data.

Next, the reason why the test for each individual predictor's significance in a multiple regression is the same as the t-test, and how it relates to the F-test, lies in the underlying statistical principles of regression analysis as explained below:

1. The t-test for Individual Predictors:
   - The t-test for each predictor test the null hypothesis that the coefficient is zero ($H_0 : \beta_j = 0$) against the alternative hypothesis that it is not zero ($H_a : \beta_j \neq 0$).
   - The t-statistic is calculated as:
   $$t = \frac{\widehat{\beta}_j}{\text{SE}\left(\widehat{\beta}_j\right)}$$
   where $\widehat{\beta}_j$ is the estimated coefficient and $\text{SE}\left(\widehat{\beta}_j\right)$ is its standard error.
   - If the t-statistic is large (in absolute value), it indicates that the predictor is significantly different from zero, meaning it contributes to the model.
2. The F-test for Overall Model Fit:
   - The F-test in the context of regression is typically used to compare two models: a full model (with all predictors) and a reduced model (with one or more predictors removed).
   - For a single predictor X_j, the F-test compares the full model with the reduced model that excludes X_j.
   - The F-statistic for this comparison can be written as:
   $$F = \frac{(\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}}) \ / \ (df_{\text{reduced}} - df_{\text{full}})}{\text{RSS}_{\text{full}} \ / \ df_{\text{full}}} \tag{2.63}$$
   where RSS is the residual sum of squares and df is the degrees of freedom for the respective models. Compare this version of the equation with (2.62), note that they are the same.
3. Connection Between t-test and F-test:
   - When testing a single predictor in regression model, the t-test and the F-test yield equivalent results.
   - Specifically, for testing the null hypothesis $H_0 : \beta_j = 0$ for a single predictor X_j, the squared t-statistic is equal to the F-statistic:
   $$F = t^2$$
   - This equivalence arises because both tests are based on the same underlying regression model and the same residual sum of squares. The t-test is a special case of the F-test where the F-test has 1 degree of freedom in the numerator (since it is comparing two models that are differ by exactly one predictor).

Let us move on to another common question related to the F-test information given in most statistical analysis package such as the one we see in Fig 2.17. Given these individual t-test p-values for each variable, why do we need to look at the overall F-statistics? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response. However, this logic is flawed, especially when the number of predictors p is large.

For instance, consider an example in which $p = 100$ and $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is true, so no variable is truly associated with the response. Since the p-values are uniformly distributed under the null hypothesis H_0, in this situation, about 5% of the p-values associated with each variable (of the type shown in Fig 2.13) will be below 0.05 by chance. In other words, we expect to see approximately five *small* p-values even in the absence of any true association between the predictors and the response. In fact, it is likely that we will observe at least one p-value below 0.05 by chance! Hence, if we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship. However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors. Hence, if H_0 is true, there is only a 5% chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors or the number of observations.

The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n. However, sometimes we have a very large number of variables. If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them. In this case we cannot even fit the multiple linear regression model using least squares, so the F-statistic cannot be used, and neither can most of the other concepts that we have seen so far in this chapter. When p is large, some of the approaches discussed in the next section, such as *forward selection* can be used. This *higher-dimensional* setting will be discussed in greater detail in later chapter.

Two: Deciding on Important Variables

As discussed in the previous section, the first step in a multiple regression analysis is to compute the F-statistic and to examine the associated p-value. If we conclude on the basis of the p-value that at least one of the predictors is related to the response, then it is natural to wonder *which* are the guilty ones! We could look at the individual p-values as in Fig 2.13, but as discussed, if p is large we are likely to make some false discoveries.

It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors. The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as *variable selection*. The variable selection problem is study extensively later in the course, and so here we will provide only a brief outline of some classical approaches.

Ideally, we would like to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors. For instance, if $p = 2$, then we can consider four models:

1. a model containing no variables,

2. a model containing X_1 only,

3. a model containing X_2 only,

4. a model containing both X_1 and X_2.

We can then select the *best* model out of all of the models that we have considered. How do we determine which model is best? Various statistics can be used to judge the quality of a model. These include

- Mallows' C_p,
- Akaike information criterion (AIC),
- Bayesian information criterion (BIC), and
- adjusted R^2.

These are discussed in more detail later in the course. We can also determine which model is best by plotting various model outputs, such as the residual, in order to search for patterns.

Unfortunately, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible. For instance, we saw that if $p = 2$, then there are $2^2 = 4$ models to consider. But if $p = 30$, then we must consider $2^{30} = 1,073,741,824$ models! Thus is not practical. Therefore, unless p is very small, we cannot consider all 2^p models, and instead we need an automated and efficient approach to choose a smaller set of models to consider. There are three classical approaches for this task:

- *Forward selection.* We begin with the *null model*—a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.

- *Backward selection.* We start with all variables in the model, and remove this variable with the largest p-value—that is, the variable that is the least statistically significant. The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold.

- *Mixed selection.* This is a combination of forward and backward selection. We start this with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variable one-by-one. Of course, as we noted with the Advertising example, the p-values for variables can become larger as new predictors are added to the model. Hence, if at any point the p-value for one of the variables in the model rises above a vertain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Backward selection cannot be used if $p > n$, while forward selection can always be used. Forward selection is a greedy approach, and might include variables early that later become redundant. Mixed selection can remedy this.

Three: Model Fit

Two of the most common numerical measures of model fit are the RSE and R^2, the fraction of variance explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

Recall that in simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals $\text{Cor}(Y, \widehat{Y})^2$, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.

An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable. As an example, we see in Fig 2.18 that for the Advertising data, the model that uses three advertising media to predict sales has an R^2 of 0.897211. On the other hand, the model that uses only TV and radio to predict sales has an R^2 value of 0.897194.

```
Call:
lm(formula = df$sales ~ df$TV + df$radio + df$newspaper)

Residuals:
      Min        1Q    Median        3Q       Max
-8.827687 -0.890814  0.241802  1.189319  2.829223

Coefficients:
                Estimate  Std. Error  t value Pr(>|t|)
(Intercept)   2.93888937  0.31190824  9.42229  < 2e-16 ***
df$TV         0.04576465  0.00139490 32.80862  < 2e-16 ***
df$radio      0.18853002  0.00861123 21.89350  < 2e-16 ***
df$newspaper -0.00103749  0.005871[illegible]01 -0.17671  0.85992
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.68551 on 196 degrees of freedom
Multiple R-squared:  0.897211,  Adjusted R-squared:  0.895637
F-statistic: 570.271 on 3 and 196 DF,  p-value: < 2.22e-16
```

```
Call:
lm(formula = df$sales ~ df$TV + df$radio)

Residuals:
      Min        1Q    Median        3Q       Max
-8.797700 -0.875158  0.242194  1.170770  2.832837

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 2.92109991 0.29448968  9.91[illegible]9 < 2.22e-16 ***
df$TV       0.04575482 0.00139036 32.[illegible]871 < 2.22e-16 ***
df$radio    0.18799423 0.00803997 23.[illegible]8245 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '[illegible]' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.6[illegible]56 on 197 degrees of freedom
Multiple R-squared:  0.897194,  Adjusted R-squared:  0.896151
F-statistic: 859.618 on 2 and 197 DF,  p-value: < 2.22e-16
```

Figure 2.18: Utilizing the R statistical software, the information on R^2 for the two multiple regression models using the Advertising data is presented above. The R-summary on the top is for the model that contains all three advertising media as predictors, while the R-summary at the bottom is for the model that uses only TV and radio as predictors.

In other words, there is a *small* increase in R^2 if we include newspaper advertising in the model that already contains TV and radio advertising, even though we saw earlier that the p-value for newspaper advertising in Fig 2.13 is not significant. It turns out that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This is due to the fact that adding another variable always results in a decrease in the residual sum of squares on the training data (though not necessarily the testing data). Thus, the R^2 statistic, which is also computed on the training data, must increase.

The fact that adding newspaper advertising to the model containing only TV and radio advertising leads to just a tiny increase in R^2 provides additional evidence that newspaper can be dropped from the model. Essentially, newspaper provides no real improvement in the model fit to the training samples, and its inclusion will likely lead to poor results on independent test samples due to overfitting.

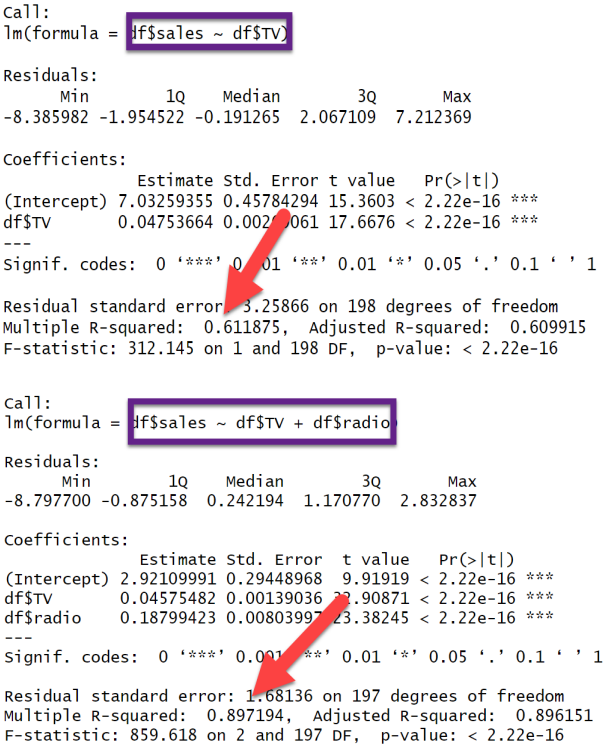


Figure 2.19: Details are as in Fig 2.18. The R-summary on the top is for the model that contains only TV as predictor, while the R-summary at the bottom is for the model that uses TV and radio as predictors.

By contrast, the model containing only TV as a predictor has an R^2 of 0.611875 (Fig 2.19). Adding radio to the model leads to a substantial improvement in R^2, where $R^2 = 0.897194$ (Fig 2.19). This implies that a model that uses TV and radio expenditures to predict sales is substantially better than one that uses only TV advertising. We could further quantify this improvement by looking at the p-value for the radio coefficient in a model that contains only TV and radio as predictors.

The model that contains only TV and radio as predictors has an RSE of 1.68136 (Fig 2.19), and the model that also contains newspaper as a predictor has an RSE of 1.68551 (Fig 2.18). In contrast, the model that contains only TV has an RSE of 3.25866 (Fig 2.19). This corroborates our previous conclusion that a model that uses TV and radio expenditures to predict sales is much more accurate (on training data) than one that only uses TV spending. Furthermore, given that TV and radio expenditures are used as predictors, there is no point in also using newspaper spending as a predictor in the model.

Type of the regression model	RSE for the nodel
$sales \approx \beta_0 + \beta_1 \times TV$	3.25866
$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times radio$	1.68136
$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper$	1.68551

The observant reader may wonder how the RSE can increase when newspaper is added to the model given that RSS must decrease. In general RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}, \tag{2.64}$$

which simplifies to (2.50) for a simple linear regression (with $p = 1$). Thus models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

In addition to looking at the RSE and R^2 statistics just discussed, it can be useful to plot the data. Graphical summaries can reveal problems with a model that are not visible from numerical statistics.

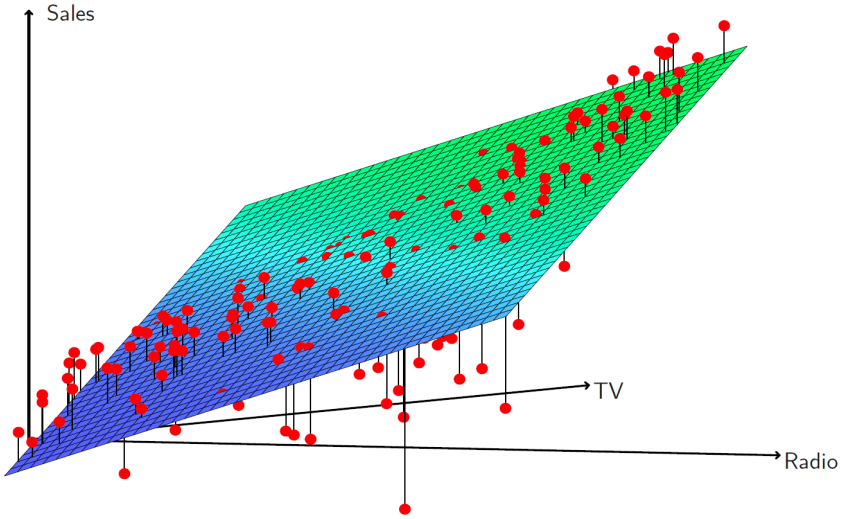


Figure 2.20: For the Advertising data, a linear regression fit to sales using TV and radio as predictors is shown above. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals (most not visible), tend to lie away from this line, where budgets are more lopsided.

For example, Fig 2.20 displays a three-dimensional plot of TV and radio versus sales. We see that some observations lie above and some observations lie below the least squares regression plane. In particular, the linear model seems to overestimate sales for instances in which most of the advertising money was spent exclusively on either TV or radio. It underestimates sales for instances where the budget was split between the two media. This pronounced non-linear pattern suggests a *synergy* or *interaction* effect between the advertising media, whereby combining the media together results in a bigger boost to sales than using any single medium. That is, for example, spending \$50,000 on TV advertising and \$50,000 on radio advertising is associated with higher sales than allocating \$100,000 to either TV or radio individually. In later section, we will discuss extending the linear model to accommodate such synergistic effects through the use of interaction terms.

Remarks: We now provide a more mathematical explanation on why R^2 is non-decreasing when extra predictors are added to the model. The key point of this is that adding an extra predictor cannot increase the residual sum of squares. Formally:

$$\text{RSS}_2 \leq \text{RSS}_1$$

where RSS_1 is for residual sum of squares for the model with p-predictors while RSS_2 is the residual sum of squares for the model with $(p+1)$-predictors. This is because the extra predictor provides additional information that can potentially explain more variance in $\mathbf{y}$, or at the very least, it allows for a more flexible fit that can reduce the residuals. Now, we can see how this can impact the R^2.

Given the relationship between RSS and R^2, we note:

$$R_1^2 = 1 - \frac{\text{RSS}_1}{\text{TSS}}$$

$$R_2^2 = 1 - \frac{\text{RSS}_2}{\text{TSS}}$$

Since $\text{RSS}_2 \leq \text{RSS}_1$, it follows that:

$$\frac{\text{RSS}_2}{\text{TSS}} \leq \frac{\text{RSS}_1}{\text{TSS}}$$

Therefore:

$$R_2^2 \geq R_1^2$$

This shows that adding an extra predictor to the model cannot decrease the R^2 value; it either increases or remains the same. This is a consequence of the residual sum of squares (RSS) either decreasing or staying constant when an additional predictor is included in the model. To address this issue, let's explore the concept of adjusted R^2.

Adjusted R^2 is a modified version of the regular R^2 that adjusts for the number of predictors in a regression model. It does so by penalizing the addition of unnecessary predictors by adjusting based on the number of predictors relative to the number of observations. The formula for adjusted R^2 is given by

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n-1)}{n-p-1} \right)$$

Where:

- R^2 is the regular coefficient of determination.
- n is the number of observations (data points).
- p is the number of predictors in the model.

Notice that

- The term $n-1$ is the total degrees of freedom in the dataset.
- The term $n-p-1$ is the residual degrees of freedom after accounting for the predictors in the model.

Here are some of the key Points on How Adjusted R^2 Works

- Penalty for Adding Variables: The denominator $n-p-1$ increases as more predictors are added to the model. This means that as p increases, the adjustment becomes more significant. If the additional predictors don't contribute meaningful information to the model, the increase in R^2 from adding more variables will not compensate for the penalty, and the adjusted R^2 may actually decrease.

- Prevents Overfitting: The penalty in Adjusted R^2 prevents overfitting, where the model might fit the noise in the data rather than capturing the true underlying relationship. It rewards models that improve fit with meaningful predictors and penalizes those that add unnecessary complexity.
- If a new variable adds value to the model, the increase in R^2 will outweigh the penalty, and adjusted R^2 will increase.
- If the new variable does not add value, the penalty will be larger than the increase in R^2, and adjusted R^2 will decrease.

This way, adjusted R^2 helps in model selection by balancing the trade-off between adding more variables and improving model fit.

Four: Predictions

Once we have fit the multiple regression model, it is straightforward to apply

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p$$

in order to predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \cdots, X_p$. However, there are three sorts of uncertainty associated with this prediction.

1. The coefficient estimates $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$ are estimates for $\beta_0, \beta_1, \cdots, \beta_p$. That is, the *least squares plane*
$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_p X_p$$
is only an estimate for the *true population regression plane*
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$
The inaccuracy in the coefficient estimates is related to the *reducible error* in
$$E(Y - \widehat{Y})^2 = \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}.$$
We can compute a *confidence interval* in order to determine how close $\widehat{Y}$ will be to $f(X)$.
2. Of course, in practice assuming a linear model for $f(X)$ is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*. So when we use a linear model, we are in fact estimating the best linear approximation to the true surface. However, here we will ignore this discrepancy, and operate as if the linear model were correct.
3. Even if we knew $f(X)$—that is, even if we knew the true values for $\beta_0, \beta_1, \cdots, \beta_p$—the response value cannot be predicted perfectly because of the random error ϵ in
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon.$$
In Chapter 1, we referred to this as the *irreducible error*. How much will Y vary from $\widehat{Y}$? We use *prediction intervals* to answer this question. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

We use a *confidence interval* to quantify the uncertainty surrounding the *average* sales over a large number of cities. For example, given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in each city, the 95% confidence interval is $(10,985,\ 11,528)$ (note the length of the interval is 543 units). We interpret this to mean that 95% of intervals of this form will contain the true value of $f(X)$. In other words, if we collect a large number of data set like the Advertising data set, and we construct a confidence interval for the average sales on the basis of each data set (given \$100,000 in TV and \$20,000 in radio advertising), then 95% of these confidence intervals will contain the true value of average sales.

On the other hand, a *prediction interval* can be used to quantify the uncertainty surrounding sales for a *particular* city. Given that \$100,000 is pent on TV advertising and \$20,000 is spent on radio advertising in that city, the 95% prediction interval is $(7,930,\ 14,580)$ (note the length of the interval is 6650 units). We interpret this to mean that 95% of intervals of this form will contain the true value of Y for this city. Note that both intervals are centered at 11,256, but the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many location. We now provide the mathematical details for these concepts. We will start with the simple linear regression model.

Confidence Interval for the Mean (Average) Response $\mu_0 = \beta_0 + \beta_1 x_0$.

Suppose x_0 is given and we would like to find a point estimator for $\mu_0 = \beta_0 + \beta_1 x_0$, then being an unbiased estimator, $\widehat{\mu}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is a natural choice. However, if a confidence interval or hypothesis testing about the mean response is required, then we need to know the probability distribution for the estimator $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Again from

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) y_i$$

and

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

we have

$$\begin{aligned}
\hat{\beta}_0 + \hat{\beta}_1 x_0 &= (\overline{y} - \hat{\beta}_1 \overline{x}) + \hat{\beta}_1 x_0 \\
&= \overline{y} - \hat{\beta}_1 (\overline{x} - x_0) \\
&= \frac{\sum_{i=1}^{n} y_i}{n} - \left(\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) y_i \right) (\overline{x} - x_0) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - c(x_i - \overline{x})(\overline{x} - x_0) \right) y_i
\end{aligned}$$

where $c = 1/S_{xx}$

From

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - c(x_i - \overline{x})(\overline{x} - x_0) \right) y_i$$

where $c = 1/S_{xx}$, we see that $\hat{\beta}_0 + \hat{\beta}_1 x_0$ can be expressed as a linear combination of independent normal random variables y_i and so it itself is also normally distributed (recall that, under the usual assumptions, we have $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$). Since we know

$$E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

we only need to compute $\mathrm{Var}(\hat{\beta}_0+\hat{\beta}_1 x_0)$ to obtain the complete information of the distribution. Note that

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0+\hat{\beta}_1 x_0) &= \sum_{i=1}^{n}\left(\frac{1}{n}-c(x_i-\overline{x})(\overline{x}-x_0)\right)^2 \mathrm{Var}(y_i)\\ &= \sigma^2\left(\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}\right)\end{aligned}$$

Hence

$$\hat{\beta}_0+\hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0+\beta_1 x_0, \sigma^2\left(\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}\right)\right)$$

Recall that if Z and X_n^2 are independent r.v. with $Z\sim \mathrm{N}(0,1)$ and X_n^2 having a chi-square with n degree of freedom, then the r.v. T_n defined below has a t-distribution with n degree of freedom, that is

$$T_n := \frac{Z}{\sqrt{X_n^2/n}} \sim t_n$$

Applying this to

$$\hat{\beta}_0+\hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0+\beta_1 x_0, \sigma^2\left(\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}\right)\right) \quad \text{and} \quad \frac{SS_R}{\sigma^2}\sim \chi^2_{n-2}$$

and $\hat{\beta}_0+\hat{\beta}_1 x_0$ being independent of SS_R/σ^2 it follows that

$$\frac{\hat{\beta}_0+\hat{\beta}_1 x_0-(\beta_0+\beta_1 x_0)}{\sqrt{\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}}\ \sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2} \tag{2.65}$$

Remark: The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by solving the normal equations, which depends on the data points (x_i, y_i). SS_R is a function of the residuals, geometrically, it relates to the orthogonal components of the data, hence $\hat{\beta}_0+\hat{\beta}_1 x_0$ being independent of SS_R/σ^2.

Once we have the result in (2.65), we can then use it to obtain the following confidence interval of $\hat{\beta}_0+\hat{\beta}_1 x_0$. That is, for any significant level $0<\alpha<1$,

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_0+\hat{\beta}_1 x_0-(\beta_0+\beta_1 x_0)}{\sqrt{\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}}\ \sqrt{\frac{SS_R}{n-2}}} < t_{\alpha/2,n-2}\right) = 1-\alpha$$

gives the following confidence interval for the mean response $\mu_0=\beta_0+\beta_1 x_0$, which we write as:

$$\hat{\beta}_0+\hat{\beta}_1 x_0 \pm \sqrt{\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}}\ \sqrt{\frac{SS_R}{n-2}}\ t_{\alpha/2,n-2}, \tag{2.66}$$

Prediction Interval of the Point Prediction $y_0=\beta_0+\beta_1 x_0+\epsilon$.

Let $y_0=\beta_0+\beta_1 x_0+\epsilon$ be the future response whose input level is x_0 and consider the probability distribution of $y_0-\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right)$. For that, we need several useful facts. First, we have

$$y_0 \sim \mathrm{N}\left(\beta_0+\beta_1 x_0, \sigma^2\right)$$

second, we see earlier that

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right).$$

Now, notice that y_0 is independent of the earlier data values $y_1, \cdots, y_n$ that were used to determine $\hat{\beta}_0$ and $\hat{\beta}_1$, hence it follows that y_0 is independent of $\hat{\beta}_0 + \hat{\beta}_0 x_0$, therefore $y_0 - \left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right)$ is normally distributed with the following mean and variance.

$$y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0 \sim \mathrm{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right)$$

This is because of

$$\begin{aligned} E(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= E(\beta_0 + \beta_1 x_0 + \epsilon - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\ &= E(\beta_0 + \beta_1 x_0) + E(\epsilon) - E(\hat{\beta}_0) - E(\hat{\beta}_1 x_0) \\ &= \beta_0 + \beta_1 x_0 + 0 - \beta_0 - \beta_1 x_0 \\ &= 0 \end{aligned}$$

while

$$\begin{aligned} \mathrm{Var}(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= \mathrm{Var}(\beta_0 + \beta_1 x_0 + \epsilon - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\ &= 0 + \mathrm{Var}(\epsilon) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right) \end{aligned}$$

Next, from this distribution

$$y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0 \sim \mathrm{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\right)$$

we obtain

$$\frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}} \sim \mathrm{N}(0, 1)$$

and using the argument laid out before, we get

$$\frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2} \tag{2.67}$$

So, from (2.67), for any $0 < \alpha < 1$, we have the following prediction interval

$$P\left(-t_{\alpha/2,n-2} < \frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0}{\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} < t_{\alpha/2,n-2}\right) = 1 - \alpha,$$

which can also be written as

$$\hat{\beta}_0 - \hat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\sqrt{\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\frac{SS_R}{n-2}} \tag{2.68}$$

Comparing this to confidence interval of the mean response μ_0 (2.66), reproduced here:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm \sqrt{\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}} \quad \sqrt{\frac{SS_R}{n-2}} \quad t_{\alpha/2,n-2},$$

we clearly see that the prediction interval in (2.68) is wider. The case for multiple linear regression model is similar, but with more complicated matrix derivation which we will not include here.

2.2.3 Summary

At the beginning of this Chapter, we have posted a series of questions which some of them have been answered in Section 2.1.5 (when it relates to simple linear regression model). We now answered the rest of the questions which are relevant to multiple linear regression.

- Is there a relationship between advertising budget and sales?

The first goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales. This can be determine by looking at the F-statistic obtained from the following hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_a : \text{at least one } \beta_j \text{ is non-zero.}$$

Fig 2.17 shows that the F-test statistic is 570.3 with the p-value that is essentially zero. Hence we reject H_0 and conclude that there is at least one β_j that is non-zero, which implies that there is an association between advertising expenditure and sales. So it makes sense to spend money on advertising.

- Which media are associated with sales?

Are all three media—TV, radio, and newspaper—associated with sales, or are just one or two of the media associated? In this case the p-vales for the t-test presented in Fig 2.13 can be helpful. Each t-test assesses the null hypothesis that the coefficient of a particular predictor is equal to zero (no effect). Notice from the figure that the p-values for the t-test on the predictor newspaper is 0.8599 which is not statistically significant for all practical significance level. From here, we can conclude that advertising money spend on newspaper does not going to contribute significantly to the sales of the product. Predictors such as TV and radio have very small p-values, in fact they are smaller than all practical significance level. This indicates that they have meaningful impact on sales.

- How large is the association between each medium and sales?

For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase? To determine how large the association between each advertising medium and sales is, we need to look for the regression coefficients for each medium in the following model

$$sales \approx \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper$$

These regression coefficients will give us the expected increase in sales for every dollar spent on each medium. For example, from Fig 2.13 we see that the average increase in product sales associated with increasing TV spending by \$1,000 while holding radio and newspaper fixed is 46 unit. The accuracy of these predictors can be assessed using confidence intervals for the coefficients, p-values, and R^2.

The confidence intervals for the coefficients give a range which we can be confident where the true coefficient lies. Narrower intervals indicate more precise estimates. The p-values for the coefficient test the null hypothesis that the coefficient is zero. A low p-value (typically < 0.05) suggests that the predictor is significantly associated with the response variable. R^2 indicates the proportion of variance in sales explained by the advertising spending. The closer R^2 gets to 1, the better the accuracy. By analyzing these results, you can determine the size and significance of the association between spending on each advertising medium and sales, and how accurately these associations can predict changes in sales.

- How accurately can we predict future sales?

For any given level of TV, radio, or newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction? We can use the regression model to generate predictions and assess the accuracy of these predictions through metrics such as confidence intervals and prediction intervals.

In regression analysis, there is an important distinction between estimating the mean (average) response and making an actual prediction. Both involve using the regression model to generate expected values based on the predictions, but they serve different purposes and come with different levels of uncertainty.

Estimating the Mean Response: Estimating the mean response involves predicting the average value of the response variable (e.g., sales) for a given set of predictor values (e.g., levels of advertising spending). This type of estimate focuses on the central tendency and typically has a narrower confidence interval because it only accounts for the uncertainty in the estimated regression coefficients.

Making an Actual Prediction: Making an actual prediction involves predicting the response variable for a specific instance with given predictor values. This prediction accounts not only for the uncertainty in the regression coefficients but also for the inherent variability (random error) in the response variable. As a result, prediction intervals are typically wider than confidence intervals for the mean response. By understanding these differences, you can better interpret the results of your regression model and the associated predictions, allowing for more informed decision-making and planning.

- Is there synergy among the advertising media?

To determine if there is synergy among the advertising media, you can include interaction terms in your regression model. For example, the interaction between TV and radio spending would be represented as $TV \times radio$. Interaction terms allow you to assess whether the effect of one predictor variable on the response variable depends on the level of another predictor variable. In this context, it helps to identify if spending on one type of advertising medium amplifies or diminishes the effect of spending on another medium. We will discuss how to include interaction terms in regression analysis in the future. In this Chapter, the interaction between TV and radio was discovered by plotting the data, see Fig 2.20.

2.3 Other Considerations in the Regression Model

2.3.1 Qualitative Predictors

In our discussion so far, we have assumed that all variables in our linear regression model are *quantitative*. But in practice, this is not necessarily the case; often some predictors are *qualitative*.

For example, the Credit data set displayed in Fig 2.21 records variables for a number of credit card holders. The response is balance (average credit card debt for each individual) and there are several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). Each panel of Fig 2.21 is a scatterplot for a pair of variables whose identities are given by the corresponding row and column labels. For example, the scatterplot directly to the right of the world "Balance" depicts balance vs age, while the plot directly to the right of "Age" corresponds to age vs cards. In addition to these quantitative variables, we also have four qualitative variables: own (house ownership), student (student status), status (marital status), and region (East, West or South).

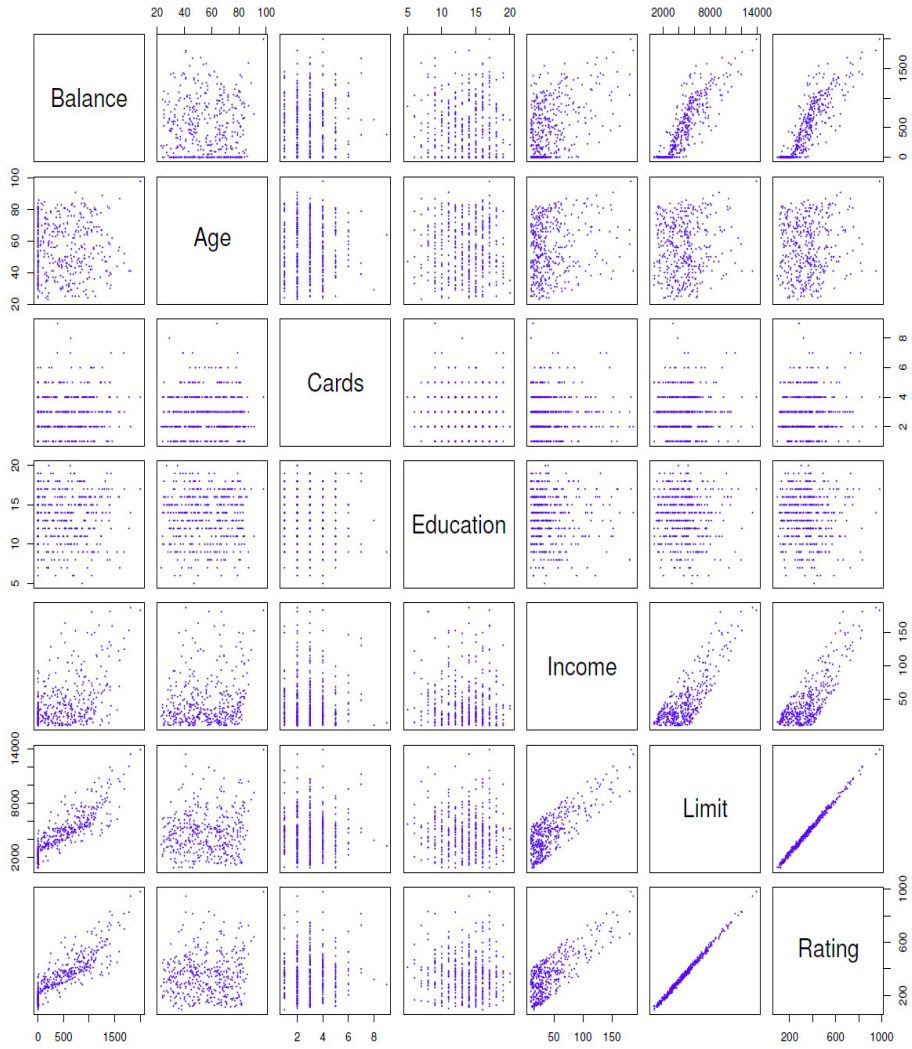


Figure 2.21: The Credit data set contains information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Predictors with Only Two Levels

Suppose that we wish to investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables for the moment. If a qualitative predictor (also known as a factor) only have two *levels*, or possible values, then incorporating it into a regression model is very simple. We simply create an indicator or *dummy variable* that takes on two possible numerical values. (In the machine learning community, the creation of dummy variables to handle qualitative predictors is known as "one-hot encoding"). For example, based on the own variable, we can create a new variable that takes the form

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases} \tag{2.69}$$

and use this variable as a predictor in the regression equation. This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not own a house} \end{cases} \tag{2.70}$$

Now β_0 can be interpreted as the average credit card balance among those who do not own, while $\beta_0 + \beta_1$ as the average credit balance among those who do own their house, and β_1 as the average difference in credit card balance between owners and non-owners.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`own[Yes]`	19.73	46.05	0.429	0.6690

Figure 2.22: Least squares coefficient estimates associated with the regression of balance onto own in the Credit data set. The linear model is given in (2.70) That is, ownership is encoded as a dummy variable, as in (2.69).

.

Figure 2.22 displays the coefficient estimates and other information associated with the model (2.70). The average credit card debt for non-owners is estimated to be \$509.80, whereas owners are estimated to carry \$19.73 in additional debt for a total of \$509.80 + \$19.73 = \$529.53. However, we notice that the p-value for the dummy variable is very high (0.6690). This indicates that there is no statistical evidence of a difference in average credit card balance (of \$19.73) based on house ownership.

The decision to code owners as 1 and non-owners as 0 in (2.70) is arbitrary, and has no effect on the regression fit, but does alter the interpretation of the coefficients. If we had coded non-owners as 1 and owners as 0, then the estimates for β_0 and β_1 would have been 529.53 and -19.73, respectively, leading to once again a prediction of credit card debt of \$529.53 − \$19.73 = \$509.80 for non-owners and a prediction of \$529.53 for owners.

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ -1 & \text{if } i\text{th person does not own a house} \end{cases}$$

and use this variable in the regression equation. This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person does not own a house} \end{cases}$$

Now β_0 can be interpreted as the overall average credit card balance (ignoring the house ownership effect), and β_1 is the amount by which house owners and non-owners have credit card balances that are above and below the average, respectively. To further elaborate on the interpretation of the coefficients β_0 and β_1, let's delve into the details:

Coefficient β_0

β_0 represents the overall average credit card balance. This interpretation stems from the fact that when the effect of house ownership (x_i) is ignored or averaged out, β_0 serves as a central reference point for the credit card balances. To understand this, consider the following:

- When $x_i = 1$ (house owner), the regression equation becomes:

$$y_i = \beta_0 + \beta_1 + \epsilon_i$$

- When $x_i = -1$ (non-owner), the regression equation becomes:

$$y_i = \beta_0 - \beta_1 + \epsilon_i$$

If we take the average of the credit card balances for both groups (house owners and non-owners), β_0 lies at the midpoint because:

$$\text{Average balance for house owners} = \beta_0 + \beta_1$$
$$\text{Average balance for non-owners} = \beta_0 - \beta_1$$

By averaging these two expressions:

$$\frac{(\beta_0 + \beta_1) + (\beta_0 - \beta_1)}{2} = \frac{2\beta_0}{2} = \beta_0$$

Hence, β_0 is the overall average credit card balance, considering both house owners and non-owners.

Coefficient β_1

β_1 captures the difference in credit card balances between house owners and non-owners relative to the overall average β_0. It shows how much the balance deviates due to house ownership status. Here's a detailed interpretation:

- For house owners ($x_i = 1$):

$$y_i = \beta_0 + \beta_1 + \epsilon_i$$

  The term $\beta_0 + \beta_1$ indicates that house owners' credit card balances are β_1 units higher than the overall average β_0.

- For non-owners ($x_i = -1$):

$$y_i = \beta_0 - \beta_1 + \epsilon_i$$

  The term $\beta_0 - \beta_1$ indicates that non-owners' credit card balances are β_1 units lower than the overall average β_0.

Therefore:

- β_0 is the central value (overall average balance) around which the two groups' averages are distributed.

- β_1 indicates the extent of the deviation from this central value due to house ownership status.

By setting up the regression model this way, we can clearly see and interpret how house ownership impacts credit card balances, with β_0 providing a baseline and β_1 showing the directional deviation from this baseline based on whether an individual owns a house or not.

In our example (using the 1 and -1 coding scheme), the estimate for β_0 is \$519.665, halfway between the non-owner and owner averages of \$509.80 and \$529.53 ((\$509.80 + \$529.53)/2 = \$519.665). The estimate for β_1 is \$9.865, which is half of \$19.73, the average difference between owners and non-owners. It is important to note that the final predictions for the credit balances of owners and non-owners will be identical regardless of the coding scheme used. The only difference is in the way that the coefficients are interpreted.

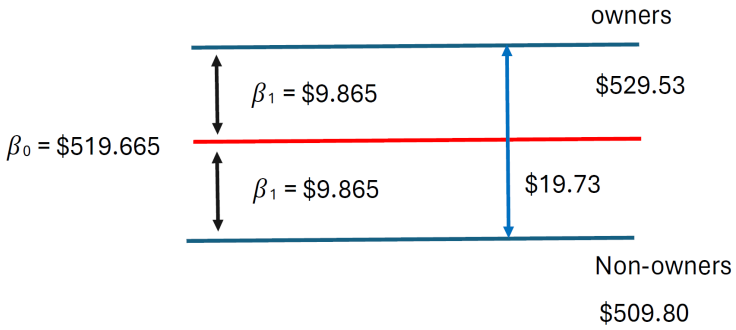


Figure 2.23: The Figure above shows the interpretation of β_0 and β_1 when the dummy variable is coded as 1 and -1. The top level and the bottommost level represent the average credit card debt for the home owners and non-owners at \$529.53 and \$509.80, respectively. The red level represent the value of β_0 in this coding scheme, that is, \$519.665 is interpreted as the central value (overall average balance) among the two groups' averages. Finally, $\beta_1 = \$9.865$ indicates the extend of the derivation from this central value due to the house ownership status.

Qualitative Predictors with More than Two Levels

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables. For example, for the region variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases} \tag{2.71}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases} \tag{2.72}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases} \tag{2.73}$$

Region	x_{i1}	x_{i2}	y_i
East	0	0	$\beta_0 + \epsilon_i$
West	0	1	$\beta_0 + \beta_2 + \epsilon_i$
South	1	0	$\beta_0 + \beta_1 + \epsilon_i$

Now β_0 can be interpreted as the average credit card balance for individuals from the East, β_1 can be interpreted as the difference in the average balance between people from the South versus the East, and β_2 can be interpreted as the difference in the average balance between those from the West versus the East. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable—East in this example—is known as the *baseline*. See the table presented above.

βs	Interpretation of βs
β_0	Average credit card balance for individual from the **East**
β_1	Difference in the average balance between people from the South vs **East**
β_2	Difference in the average balance between people from the West vs **East**

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`region[South]`	−12.50	56.68	−0.221	0.8260
`region[West]`	−18.69	65.02	−0.287	0.7740

Figure 2.24: Least squares coefficient estimates associated with the regression of balance onto region in the Credit data set. The linear model given in (2.73). That is, region is encoded via two dummy variables (2.71) and (2.72).

From Fig 2.24, we see that the estimated balance for the baseline, East, is \$531.00. It is estimated that those in the West will have \$18.69 less debt than those in the East, and that those in the South will have \$12.50 less debt than those in the East. However, the p-values associated with the coefficient estimates for the two dummy variable are very large, suggesting no statistical evidence of a real difference in average credit card balance between South and East, or between West and East.

Once again, the level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of choice. However, the coefficients and their p-values do depend on the choice of dummy variable coding. Rather than rely on the individual coefficients, we can use an F-test to test $H_0 : \beta_1 = \beta_2 = 0$; this does not depend on the coding. This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relation between balance and region.

Using this dummy variable approach presents no difficulties when incorporating both quantitative and qualitative predictors. For example, to regress balance on both a quantitative variable such as income and a qualitative variable such as student, we simply create a dummy variable for student and then fit a multiple regression model using income and the dummy variable as predictors for credit card balance.

There are many different ways of coding qualitative variables besides the dummy variable approach taken here. All of these approaches lead to equivalent model fits, but the coefficients are different and have different interpretations, and are designed to measure particular *contrasts*. This topic is beyond the scope of this course.

2.3.2 Extensions of the Linear Model

The standard linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

provides interpretable results and works quite well on many real-world problems. However, it makes several highly restrictive assumptions that are often violated in practice. Two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*.

The additivity assumption means that the association between a predictor X_j and the response Y does not depend on the values of the other predictors. The linearity assumption states that the change in the response Y associated with a one-unit change in X_j is constant, regardless of the value of X_j.

Later in the course, we examine a number of sophisticated methods that relax these two assumptions. Here, we briefly examine some common classical approaches for extending the linear model.

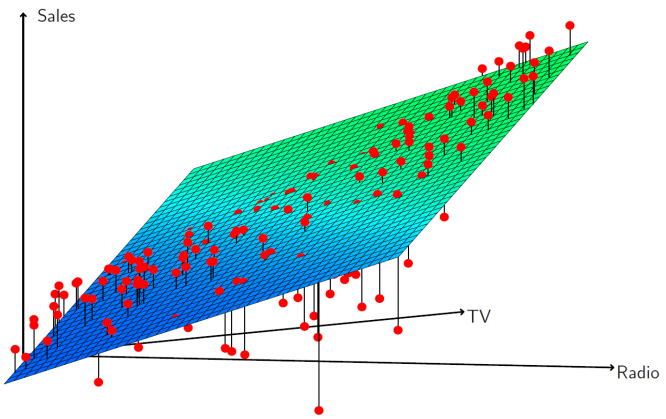
Removing the Additive Assumption

In our previous analysis of the Advertising data, we concluded that both TV and radio seem to be associated with sales. The linear models that formed the basis for this conclusion assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media. For example, the linear model

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

states that the average increase in sales associated with a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

However, this simple model may be incorrect. Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases. In this situation, given a fixed budget of \$100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio. In marketing, this is known as a *synergy* effect, and in statistics it is refereed to as an *interaction* effect. Figure 2.20 (reproduced below) suggests that such an effect may be present in the advertising data.



Notice that when levels of either TV or radio are low, then the true sales are lower than predicted by the linear model. But when advertising is split between the two media, then the model tends to underestimate sales.

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

According to this model, a one-unit increase of X_1 is associated with an average increase in Y of β_1 units. Notice that the presence of X_2 does not alter this statement—that is, regardless of the value of X_2, a one-unit increase in X_1 is associated with a β_1-unit increase in Y. One way of extending this model is to include a third predictor, called an *interaction term*, which is constructed by computing the product of X_1 and X_2. This results in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \tag{2.74}$$

How does inclusion of this interaction term relax the additive assumption? Notice that (2.74) can be written as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{2.75}$$

where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$. Since $\tilde{\beta}_1$ is now a function of X_2, the association between X_1 and Y is no longer constant: a change in the value of X_2 will change the association between X_1 and Y. A similar argument shows that a change in the value of X_1 changes the association between X_2 and Y.

For example, suppose that we are interested in studying the productivity of a factory. We wish to predict the number of units produced on the basis of the number of production lines and the total number of workers. It seems likely that the effect of increasing the number of production lines will depend on the number of workers, since if no workers are available to operate the lines, then increasing the number of lines will not increase production. This suggests that it would be appropriate to include an interaction term between lines and workers in a linear model to predict units. Suppose that when we fit the model, we obtain

$$\begin{aligned} \textit{units} &\approx 1.2 + 3.4 \times \textit{lines} + 0.22 \times \textit{workers} + 1.4 \times (\textit{lines} \times \textit{workers}) \\ &= 1.2 + (3.4 + 1.4 \times \textit{workers}) \times \textit{lines} + 0.22 \times \textit{workers}. \end{aligned}$$

In other words, adding an additional line will increase the number of units produced by $3.4 + 1.4 \times \textit{workers}$. Hence the more workers we have, the stronger will be the effect of lines.

We now return to the Advertising example. A linear model that uses radio, TV, and an interaction between the two to predict sales takes the form

$$\begin{aligned} \textit{sales} &= \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times (\textit{radio} \times \textit{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \textit{radio}) \times \textit{TV} + \beta_2 \times \textit{radio} + \epsilon. \end{aligned} \tag{2.76}$$

We can interpret β_3 as the increase in the effectiveness of TV advertising associated with a one-unit increase in radio advertising (or vice-versa.) The coefficients that result from fitting the model (2.76) are given in Fig 2.25.

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Figure 2.25: For the Advertising data, least squares coefficient estimates associated with the regression of sales onto TV and radio, with an interaction term, as in (2.76).

The results in Fig 2.25 strongly suggest that the model that includes the interaction term is superior to the model that contains only *main effects*. The p-value for the interaction term, $TV \times radio$, is extremely low, indicating that there is strong evident for $H_a : \beta_3 \neq 0$. In other words, it is clear that the true relationship is not additive. The R^2 for the model (2.76) is 96.8%, compared to only 89.7% for the model that predict sales using TV and radio without an interaction term. This means that

$$\frac{(96.8 - 89.7)}{100 - 89.7}\% = 69\%$$

of the variability in sales that remains after fitting the additive model has been explained by the interaction term. (Note that the difference of $(96.8 - 89.7)$ represent the amount of variability explained by adding the interaction term to the additive model. The difference of $(100 - 89.7)$ represents the total variability remaining after fitting the additive model.) The coefficient estimates in Fig 2.25 suggest that an increase in TV advertising of \$1,000 is associated with increase in sales of

$$\left(\widehat{\beta}_1 + \widehat{\beta}_3 \times radio\right) \times 1,000 = 19 + 1.1 \times radio \quad \text{units.}$$

And an increase in radio advertising of \$1,000 will be associated with an increase in sales of

$$\left(\widehat{\beta}_2 + \widehat{\beta}_3 \times TV\right) \times 1,000 = 29 + 1.1 \times TV \quad \text{units.}$$

In this example, the p-values associated with TV, radio, and the interaction term all are statistically significant (Fig 2.25), and so it is obvious that all three variables should be included in the model. However, it is sometimes the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV, and radio) do not. The *hierarchical principle* states that *if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.* In other words, if the interaction between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficients have large p-values.

The rationale for this principle is that including an interaction term (e.g. $X_1 \times X_2$) without the corresponding main effects (e.g. X_1 and X_2) makes the model less interpretable. The interaction term represents how the effect of one predictor depends on the level of another predictor. Without the main effects, it becomes challenging to understand the baseline effects of the predictors individually. Also, $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

In the previous example, we considered an interaction between TV and radio, both of which are quantitative variables. However, the concept of interactions applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables. In fact, an interaction between a qualitative variable and a quantitative variable has a particularly nice interpretation. Consider the Credit data set from the previous section, and suppose that we wish to predict balance using the income (quantitative) and student (qualitative) variables.

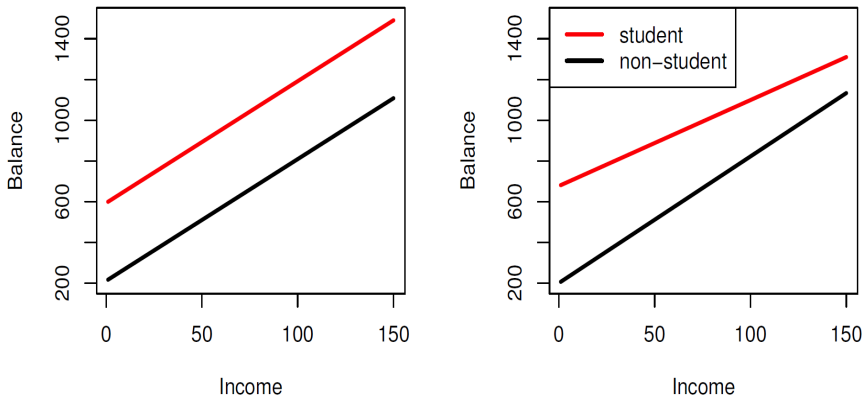


Figure 2.26: For the Credit data, the least squares lines are shown for the prediction of balance from income for students and non-students. Left: The model (2.77) was fit. There is no interaction between income and student. Right: The model (2.78) was fit. There is an interaction between income and student.

In the absence of an interaction term, the model takes the form

$$\begin{aligned} balance_i &\approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases} \end{aligned} \tag{2.77}$$

Notice that this amounts to fitting two parallel lines to the data, one for students and one for non-students. The lines for students and non-students have different intercepts, $\beta_0 + \beta_2$ versus β_0, but the same slope, β_1. This is illustrated in the left-hand panel of Fig 2.26. The fact that the lines are parallel means that the average effect on balance of a one-unit increase in income does not depend on whether or not the individual is a student. This represents a potential serious limitation of the model, since in fact a change in income may have a very different effect on the credit card balance of a student versus non-student.

This limitation can be addressed by adding an interaction variable, created by multiplying income with the dummy variable for student. Our model now becomes

$$\begin{aligned} balance_i &\approx \beta_0 + \beta_1 \times income_i + \beta_2 \times student_i + \beta_3 \times income_i \times student_i \\ &= \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i & \text{if student} \\ 0 & \text{if not a student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i & \text{if student} \\ \beta_0 + \beta_1 \times income_i & \text{if not a student} \end{cases} \end{aligned} \tag{2.78}$$

Once again, we have two different regression lines for the students and the non-students. But now those regression lines have different intercepts, $\beta_0 + \beta_2$ versus β_0, as well as different slopes, $\beta_1 + \beta_3$ versus β_1. This allows for the possibility that changes in income may affect the credit card balances of students and non-students differently. The right-hand panel of Fig 2.26 shows the estimated relationship between income and balance for students and non-students in the model (2.78). We note that the slope for students is lower than the slope for non-students. This suggests that increases in income are associated with smaller increases in credit card balance among students as compared to non-students.

Non-linear Relationships

As discussed previously, the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

assumes a linear relationship between the response and predictors. But in some cases, the true relationship between the response and the predictors may be non-linear. Here we present a very simple way to directly extend the linear model to accommodate non-linear relationship, using *polynomial regression*. In later part of the course, we will present more complex approaches for performing non-linear fits in more general settings.

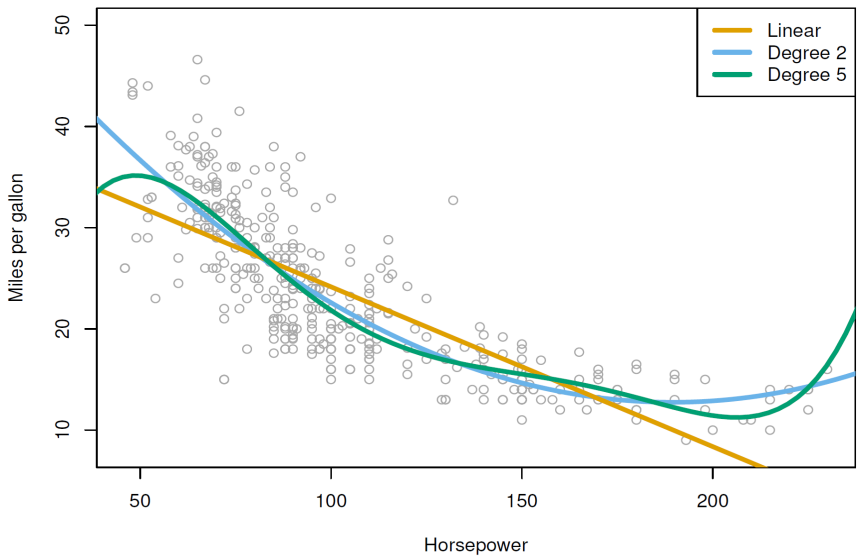


Figure 2.27: The Auto data set. For a number of cars, mpg and horsepower are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes $horsepower^2$ is shown as a blue curve. The linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in green.

Consider Fig 2.27, in which the mpg (gas mileage in miles per gallon) versus horsepower is shown for a number of cars in the Auto data set. The orange line represents the linear regression fit. There is a pronounced relationship between mpg and horsepower, but it seems clear that this relationship is in fact non-linear: the data suggest a curved relationship. A simple approach for incorporating non-linear associations in a linear model is to include transformed version of the predictors. For example, the points in Fig 2.27 seems to have a *quadratic* shape, suggesting that a model of the form

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon \tag{2.79}$$

may provide a better fit. Equation 2.79 involves predicting mpg using a non-linear function of horsepower. *But it is still a linear model!* That is, (2.79) is simply a multiple linear regression model with $X_1 = horsepower$ and $X_2 = horsepower^2$. So we can use standard linear regression software to estimate β_0, β_1 and β_2 in order to produce a non-linear fit.

Remark: In linear regression, the term "linear" refers to the linearity in coefficients, not necessarily the linearity in the predictors.

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	−0.4662	0.0311	−15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Figure 2.28: For the Auto data set, least squares coefficient estimates associated with the regression of mpg onto horsepower and $horsepower^2$.

The blue curve in Fig 2.27 shows the resulting quadratic fit to the data. The quadratic fit appears to be substantially better than the fit obtained when just the linear term is included. The R^2 for the quadratic fit is 0.688, compared to just 0.606 for the linear fit, and the p-value in Fig 2.28 for the quadratic term is highly significant.

If including $horsepower^2$ led to such a big improvement in the model, why not include $horsepower^3$, $horsepower^4$, or even $horsepower^5$? The green curve in Fig 2.27 displays the fit that results from including all polynomials up to fifth degree in the model (2.79). The resulting fit seems unnecessarily wiggly—that is, it is unclear that including the additional terms really has led to a better fit to the data.

The approach that we have just described for extending the linear model to accommodate non-linear relationships is known as *polynomial regression*, since we have included polynomial functions of the predictors in the regression model.

2.3.3 Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

In practice, identifying and overcoming these problems is as much an art as science. Many pages in countless books have been written on this topic. We will provide a brief summary of some key points here.

1. Non-linearity of the Data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.

Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $e_i = y_i - \widehat{y}_i$, versus the predictor x_i. In the case of a multiple regression model, since there are multiple predictors, we instead plot the residual versus the predicted (or fitted) values $\widehat{y}_i$. Ideally, the residual plot will show no discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model.

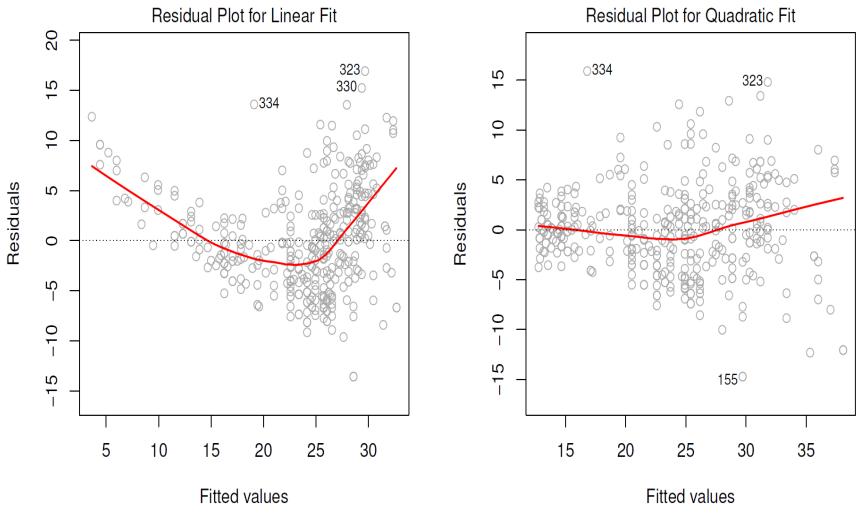


Figure 2.29: Plots of residuals vs predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear model of mpg onto horsepower. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of mpg on horsepower and $horsepower^2$. There is little pattern in the residuals.

The left panel of Fig 2.29 displays a residual plot from the linear regression mpg onto horsepower on the Auto data set that was illustrated in Fig 2.27. The red line is a smooth fit to the residuals, which is displayed in order to make it easier to identify any trends. The residuals exhibit a clear U-shape, which provides a strong indication of non-linearity in the data. In contrast, the right-hand panel of Fig 2.29 displays the residual plot that results from the model (2.79), which contains a quadratic term. There appears to be little pattern in the residuals, suggesting that the quadratic term improves the fit to the data.

Remark: Why the residuals should be randomly scattered?

- Indicates Correct Model Specification: If the residuals are randomly scattered around zero when plotted against the fitted values, it suggests that the model captures the relationship between the predictors and the response variable adequately. Any systematic pattern in the residuals indicates that the model is missing some structure, which might be due to omitted variables, incorrect functional form, or other issues.

- No Systematic Errors: Random scatter implies that the residuals (errors) are not correlated with the fitted values, suggesting that the predictors explain all systematic variation in the response variable.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\log X$, $\sqrt{X}$, and X^2, in the regression model. Note that more advanced non-linear approaches do exist.

2. Correlation of Error Terms

An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated. What does this mean? For instance, if the errors are uncorrelated, then the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.

The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrected error terms. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be (giving us a wrong sense of confidence). For example, a 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter.

In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant. In short, if the error terms are correlated, we may have an unwarranted sense of confidence in our model. Let us see why this phenomenon occur mathematically. Recall from (2.60) that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

We now compute the variance of these least square estimators. For this, we need to remind the reader of a few facts. First, if $A = (a_{ij})$ is a $r \times p$ constant matrix, while Y is a $p \times c$ random matrix, then

$$\begin{aligned} E(AY) &= E\left(\left[\sum_{k=1}^{p} a_{ik}Y_{kj}\right]_{ij}\right) \\ &= \left[E\left(\sum_{k=1}^{p} a_{ik}Y_{kj}\right)\right]_{ij} \\ &= \left[\sum_{k=1}^{p} a_{ik}E(Y_{kj})\right]_{ij} \\ &= AE(Y) \end{aligned}$$

Similarly, we have $E(AYB) = AE(Y)B$, when B is a constant matrix where the multiplication is compatible. Second, for a random vector $\mathbf{y}$ of dimension $n \times 1$, i.e. $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. The variance of $\mathbf{y}$, denoted as $\mathrm{Var}(\mathbf{y})$, is defined as

$$\mathrm{Var}(\mathbf{y}) = E[(\mathbf{y} - E[\mathbf{y}])(\mathbf{y} - E[\mathbf{y}])^T].$$

Finally, analogous to $\mathrm{Var}(ay) = a^2\mathrm{Var}(y)$, we have

$$\begin{aligned} \mathrm{Var}(A\mathbf{y}) &= E[(A\mathbf{y} - A\boldsymbol{\mu})(A\mathbf{y} - A\boldsymbol{\mu})^T] \\ &= E[(A(\mathbf{y} - \boldsymbol{\mu}))(A(\mathbf{y} - \boldsymbol{\mu}))^T] \\ &= E[A(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T A^T] \\ &= AE[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T]A^T \\ &= A\mathrm{Var}(\mathbf{y})A^T, \end{aligned} \tag{2.80}$$

where $\boldsymbol{\mu} = E(\mathbf{y})$ and we have use the fact that $E(A\mathbf{y}B) = AE(\mathbf{y})B$.

Let us now examine the term $\text{Var}(\mathbf{y})$. The variance-covariance matrix $\text{Var}(\mathbf{y})$ for the random vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is given by:

$$\text{Var}(\mathbf{y}) = \begin{pmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) & \cdots & \text{Cov}(y_1, y_n) \\ \text{Cov}(y_2, y_1) & \text{Var}(y_2) & \cdots & \text{Cov}(y_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(y_n, y_1) & \text{Cov}(y_n, y_2) & \cdots & \text{Var}(y_n) \end{pmatrix}$$

where: $\text{Var}(y_i) = E[(y_i - E[y_i])^2]$ and $\text{Cov}(y_i, y_j) = E[(y_i - E[y_i])(y_j - E[y_j])]$.

It is easy to see that when ϵ_i are uncorrelated, then y_i are also uncorrelated, therefore $\text{Cov}(y_i, y_j) = 0$ when $i \neq j$. In this case, we have

$$\text{Var}(\mathbf{y}) = \begin{pmatrix} \text{Var}(y_1) & 0 & \cdots & 0 \\ 0 & \text{Var}(y_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{Var}(y_n) \end{pmatrix}$$

Furthermore, when $\text{Var}(y_i) = \sigma^2$, for $i = 1, \cdots, n$, then

$$\text{Var}(\mathbf{y}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

Back to (2.60), and using (2.80) we then obtain

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{2.81}$$

and under the assumption of ϵ_i being uncorrelated and that $\text{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, \cdots, n$, we can then write,

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 I\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{2.82}$$

In the case of simple linear regression with uncorrelated errors, (2.82) simplifies to (2.21), which is:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

When the error terms ϵ_i is correlated, Eq (2.81) can not be further simplified. However, $\text{Var}(\mathbf{y})$ is a semi-positive definite matrix in the sense that for any vector $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{x}^T\text{Var}(\mathbf{y})\mathbf{x} \geq 0$. This can be seen from

$$\mathbf{x}^T\text{Var}(\mathbf{y})\mathbf{x} = E\left[\left\|(\mathbf{y} - E(\mathbf{y}))^T\mathbf{x}\right\|^2\right] \geq 0.$$

This means that the variance calculation involving $\text{Var}(\mathbf{y})$ will not subtract from the overall variability but will add to it or keep it the same.

This implies:

$$\text{Var}(\hat{\beta}_1) \text{ with correlation} \geq \text{Var}(\hat{\beta}_1) \text{ without correlation}$$

Therefore, the least square estimators that ignores the correlation among errors underestimates the true variance of $\hat{\beta}_1$.

Since the standard error of $\hat{\beta}_1$ is the square root of its variance, underestimating the variance leads to underestimating the standard error. Confidence intervals for $\hat{\beta}_1$ are calculated as:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \sqrt{\text{Var}(\hat{\beta}_1)}$$

If $\text{Var}(\hat{\beta}_1)$ is underestimated, the intervals are narrower than they should be. Similarly, prediction intervals, which depend on the variance of predictions, will also be narrower due to the underestimated standard errors.

In conclusion, correlated error terms result in underestimated standard errors, leading to narrower confidence and prediction intervals than are appropriate, giving a false sense of precision in the estimates and predictions.

As an extreme example, suppose we accidentally doubled our data, leading to observations and error terms identical in pairs (hence y_i are correlated and so are the errors ϵ_i). If we ignored this, our standard error calculations would be as if we had a sample of size $2n$, when in fact we have only n samples. Our estimated parameters would be the same for the $2n$ samples as for the n samples, but the confidence intervals would be narrower by a factor of $\sqrt{2}$. Assuming $\text{Var}(\epsilon_i) = \sigma^2$ for all i, here are the mathematical details.

When we double the data, our dataset becomes

$$(x_1, y_1), (x_1, y_1), (x_2, y_2), (x_2, y_2), \ldots, (x_n, y_n), (x_n, y_n),$$

hence we now have $2n$ observations. The regression model for the pair remains the same, i.e.

$$y_{2i-1} = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{2i} = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the least square estimators of β_0 and β_1 remain unchanged because they are based on the same data points, just repeated. Now, using (2.21) we calculate the variance of $\hat{\beta}_1$ with the doubled data. Take note that (2.21) is derived assuming that we have no correlations in the data. With the application of (2.21) we obtain

$$\text{Var}(\hat{\beta}_1)_{\text{doubled}} = \frac{\sigma^2}{\sum_{i=1}^{2n}(x_i - \bar{x})^2}. \tag{2.83}$$

We rewrite (2.83) as

$$\text{Var}(\hat{\beta}_1)_{\text{incorrect}} = \frac{\sigma^2}{2\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{2}\left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \overset{\text{def}}{=} \frac{1}{2}\text{Var}(\hat{\beta}_1)_{\text{correct}}$$

since, because of the duplicated data, we have

$$\sum_{i=1}^{2n}(x_i - \bar{x})^2 = 2\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Taking the square root to find the standard errors, we get:

$$\text{SE}(\hat{\beta}_1)_{\text{incorrect}} = \sqrt{\frac{1}{2}\text{Var}(\hat{\beta}_1)_{\text{correct}}} = \frac{1}{\sqrt{2}}\text{SE}(\hat{\beta}_1)_{\text{correct}}$$

Since σ^2 is known, the confidence intervals for $\hat{\beta}_1$ are constructed as:

$$\hat{\beta}_1 \pm Z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_1)$$

With the incorrect standard error, the confidence intervals would be:

$$\hat{\beta}_1 \pm Z_{\alpha/2} \cdot \frac{1}{\sqrt{2}}\text{SE}(\hat{\beta}_1)_{\text{correct}}$$

This is narrower by a factor of $\sqrt{2}$ compared to the correct confidence interval:

$$\hat{\beta}_1 \pm Z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_1)_{\text{correct}}$$

As a conclusion, we see that

$$\sqrt{2} \times \underbrace{\text{Width}_{\text{ CI with } 2n\text{-data points}}}_{\text{narrower}} = \text{Width}_{\text{ CI with } n\text{-data points}},$$

due to incorrectly duplicated the data, and still assuming having uncorrelated errors. Notice that, in this example, we did not compute the correct variance of $\widehat{\beta}_1$ when correlation of errors are present. What we've shown is that if we use (2.21) for a correlated data set, then the confidence interval constructed will be narrower, as compare with the confidence interval without the presence of correlation issue. Hence giving us a false sense of confidence.

Why might correlations among the error terms occur? Such correlations frequently occur in the context of *time series* data, which consists of observations for which measurements are obtained at discrete points in time.

In many cases, observations that are obtained at adjacent time points will have positively correlated errors. In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time. If the errors are uncorrelated, then there should be no discernible pattern. On the other hand, if the error terms are positively correlated, then we may see *tracking* in the residuals—that is, adjacent residuals may have similar values. Fig 2.30 provides an illustration. In the top panel, we see the residuals from a linear regression fit to data generated with uncorrelated errors. There is no evidence of time-related trend in the residuals. In contrast, the residuals in the bottom panel are from a data set in which adjacent errors had a correlation of 0.9. Now there is a clear pattern in the residuals—adjacent residuals tend to take on similar values. Finally, the center panel illustrates a more moderate case in which the residuals had a correlation of 0.5. There is still evidence of tracking, but the pattern is less clear.

Many methods have been developed to properly take account of correlations in the error terms in time series data. We will see one such example later. Correlation among the error terms can also occur outside of time series data. For instance, consider a study in which individuals' heights are predicted from their weights. The assumption of uncorrelated errors could be violated if some of the individuals in the study are members of the same family, eat the same diet, or have been exposed to the same environmental factors. In general, the assumption of uncorrelated errors is extremely important for linear regression as well as for other statistical methods, and good experimental design is crucial in order to mitigate the risk of such correlations.

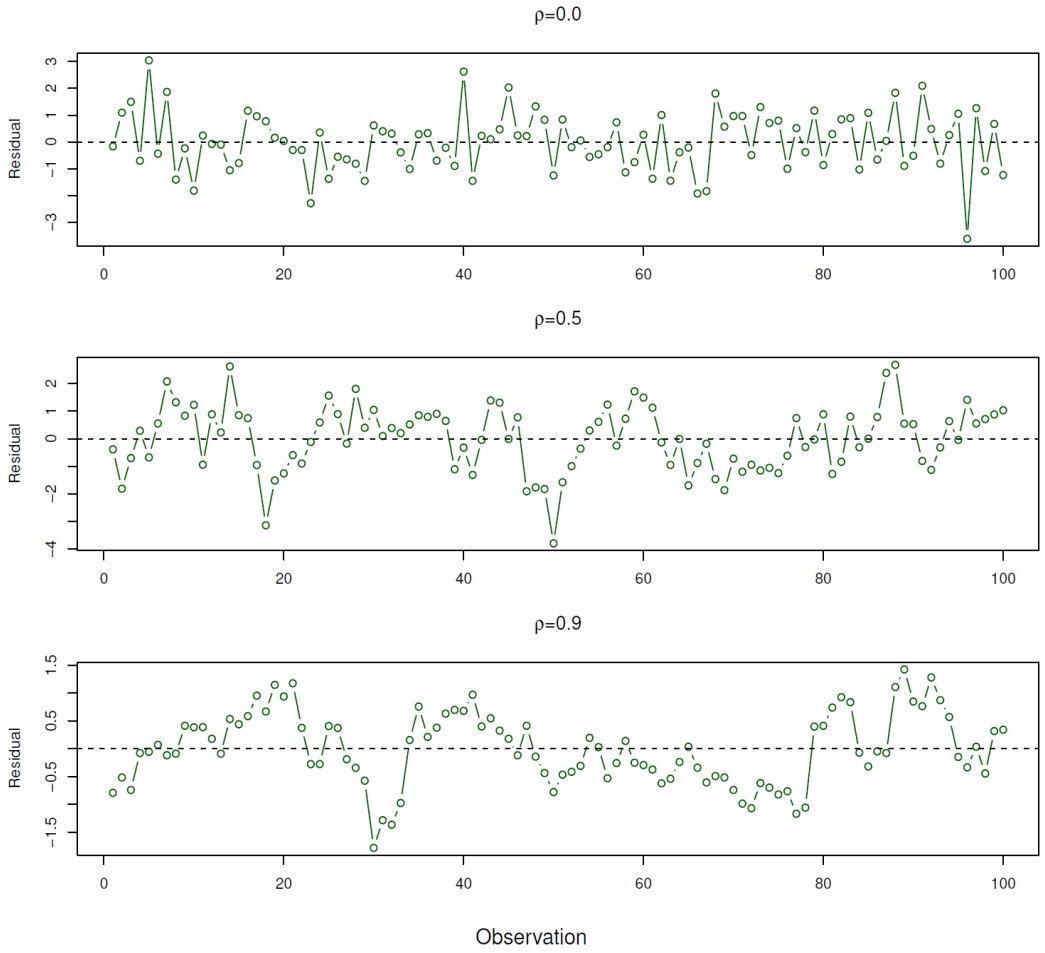


Figure 2.30: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

There are various statistical tests that we can employ to detect the presence of autocorrelation. Here, as an example, we are going to introduce a test developed by Durbin and Watson. The test is based on the assumption that the errors in the regression model are correlated in the sense that:

$$\epsilon_i = \phi\epsilon_{i-1} + a_i. \tag{2.84}$$

Here a_i is normally and independently distributed random variable with mean 0 and variance σ^2, and ϕ is a parameter that defines the relationship between successive values of the model errors ϵ_i, and ϵ_{i-1}. (Note: a_is are independent of each other and of ϵs from earlier period.) We also required that the process is *(weakly) stationary*, that is (i) the mean $E(\epsilon_i)$ and the variance $\text{Var}(\epsilon_i)$, are constant over all time i, as well as, (ii) the covariance between ϵ_i and ϵ_{i-s} depends only on the lag s and not the specific time i. For the process to be weakly stationary, we need $|\phi| < 1$.

Equation (2.84), with all the assumptions mentioned above, describes what we call a first-order autoregressive, AR(1), process. The AR(1) process is a fundamental building block in time series analysis.

Let us now look at the simple linear regression model with first-order autoregressive errors, (AR(1) errors), given by

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{where} \\ \epsilon_i &= \phi\epsilon_{i-1} + a_i. \end{aligned}$$

It can be shown easily (using the weakly stationary property) that the error terms in this model has the following properties:

$$\begin{aligned} E(\epsilon_i) &= 0 \\ \text{var}(\epsilon_i) &= \frac{\sigma^2}{1-\phi^2} \\ \text{cov}(\epsilon_i, \epsilon_{i-1}) &= \frac{\phi\sigma^2}{1-\phi^2} = \phi\text{var}(\epsilon_i) \\ \text{cov}(\epsilon_i, \epsilon_{i-s}) &= \frac{\phi^s\sigma^2}{1-\phi^2} = \phi^s\text{var}(\epsilon_i), \text{ for } s > 0 \end{aligned} \tag{2.85}$$

Therefore, from (2.85) the autocorrelation between two errors that are one period apart, or the *lag one* autocorrelation is given by

$$\begin{aligned} \rho_1 &= \text{corr}(\epsilon_i, \epsilon_{i+1}) \\ &= \frac{\text{cov}(\epsilon_i, \epsilon_{i+1})}{\sqrt{\text{var}(\epsilon_i)}\sqrt{\text{var}(\epsilon_{i+1})}} \\ &= \frac{\phi\sigma^2\left(\frac{1}{1-\phi^2}\right)}{\sqrt{\sigma^2\left(\frac{1}{1-\phi^2}\right)}\sqrt{\sigma^2\left(\frac{1}{1-\phi^2}\right)}} \\ &= \phi \end{aligned}$$

Similarly, the lag k autocorrelation is given by

$$\rho_k = \text{corr}(\epsilon_i, \epsilon_{i+k}) = \phi^k \quad \text{for} \quad k = 1, 2, \cdots . \tag{2.86}$$

Where ϕ is the constant found in (2.84). For an AR(1) process, the equation (2.86), $\rho(k) = \phi^k$, is called the *autocorrelation function*. Notice that when ϕ is positive, all error terms are positively correlated, but the magnitude of the correlation decreases as the errors grow further apart, since

$$\rho_k = \phi^k, \quad \text{and} \quad |\phi| < 1, \quad \text{hence} \quad \lim_{k\to\infty} \rho_k = 0$$

Also, only when $\phi = 0$ are the model errors uncorrelated.

With all the preliminary knowledge stated, we can now turn to the Durbin-Watson Test. The Durbin-Watson test is a hypothesis test for the following null and alternate hypotheses:

$$H_0 : \phi = 0 \quad \text{vs} \quad H_a : \phi > 0,$$

where the Durbin-Watson test statistic is

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}. \tag{2.87}$$

If d is smaller than a certain critical value, we will reject the null hypothesis.

Remark: Most time series regression problems are involve with positive autocorrelation. Hence our H_a is set as $\phi > 0$. We now give an intuitive explanation on why Durbin-Watson test works. Notice from (2.87) that, if we have a lag one autocorrelation, that is, where the i-residual ϵ_i correlates to the previous $(i-1)$-residual ϵ_{i-1}, then the two terms must "looks" alike and hence the distance $(\epsilon_i - \epsilon_{i-1})^2$ is going to be small. Therefore if we find many lag one residuals, then the Durbin-Watson statistic is very small. Hence a small value of d indicate higher possibility of correlation. So if d is smaller than a certain critical value, we will reject the null hypotheses $H_0 : \phi = 0$.

Unlike the usual hypothesis testing that we have seen before, the Durbin-Watson Test has the following decision procedure:

$$\begin{aligned} \text{If } d < d_L \text{ reject } \mathrm{H}_0: \ \phi = 0 \\ \text{If } d > d_U \text{ do not reject } \mathrm{H}_0: \ \phi = 0 \\ \text{If } d_L \leq d \leq d_U \text{ the test is inconclusive} \end{aligned}$$

for some lower d_L and upper d_U critical values. However, it can be shown that $d \approx 2(1-\hat{\phi})$, where $\hat{\phi}$ is the estimate for the autocorrelation parameter ϕ, because of this fact, we have a more convenient way of drawing the test conclusion:

- A value of d close to 2 indicates no significant autocorrelation.
- Values significantly different from 2 suggest the presence of autocorrelation:
  - If d is close to 0, it suggests positive autocorrelation.
  - If d is close to 4, it suggests negative autocorrelation.

Before we move to an example, please see the remark below:

Remark: The Durbin-Watson Test only check for autocorrelation with a lag of one, hence longer lags may not be detected. In such cases, alternative tests and diagnostics should be considered, such as:

- Portmanteau Tests: These tests, such as the Ljung-Box test or the Box-Pierce test, examine autocorrelation across multiple lags simultaneously. They can help detect autocorrelation at various lags, not just lag 1.
- Residual Plot Analysis: Visual inspection of residuals plotted against time or against predicted values can sometimes reveal patterns of autocorrelation at different lags.

As an example, let us look at the simulated data set Autoc1. We first run a regression and then the Durbin-Watson test to obtain the following result shown in Fig 2.31.

```
library(lmtest)
data1<-read.csv("Autoc1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
model1=lm(y~x)
dwtest(model1)
```

```
        Durbin-Watson test

data:  model1
DW = 0.23818, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 2.31: The figure above shows the R-output of the Durbin-Watson test performed on the Autoc1 data set.

From the R-output shown in Fig 2.31 above, we see that the test statistics is $d = 0.23818$ and the p-value is practically 0. We hence reject $H_0 : \phi = 0$ and conclude that the errors ϵ are positively correlated.

In R, the acf() function from the tseries library can be employed to compute and plot the autocorrelation function, ACF, for different value of lag k, see (2.86). Figure 2.32 shows the ACF plot using information in the Autoc1 data set.

```
library(tseries)
data1<-read.csv("Autoc1.CSV", header=TRUE, sep=",")
x<-data1$x
y<-data1$y
model1=lm(y~x)
dwtest(model1)
acf(resid(model1))
```

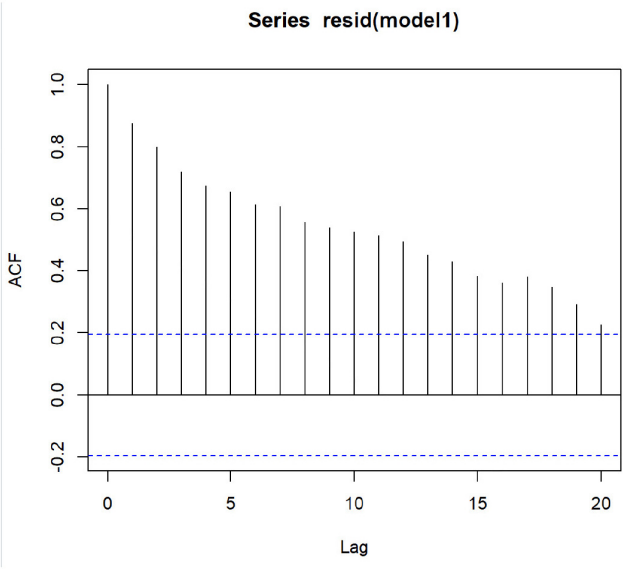


Figure 2.32: The figure shows the ACF plot obtained from Autoc1.

Notice that the x-axis of Fig 2.32 corresponds to the different lags of the residuals, while the y-axis shows the correlation of each lag. In this plot, the first vertical bar corresponds to lag-0, hence it shows the correlation of a residual with itself and therefore is always taken as one. The second vertical bar has a height of about 0.8738 and it is the value of $\rho_1 = \text{corr}(\epsilon_i, \epsilon_{i+1}) = \phi^1$ in an AR(1) process, as indicated by (2.86). In the absence of autocorrelation, the subsequent vertical bars would quickly drop to almost zero, or at least between or near the blue-dashed lines. Hence in our case, we see that autocorrelation is clearly present.

The blue-dashed lines in Fig 2.32 represent the upper and lower bounds of the 95% confidence interval for the autocorrelation coefficients being zero. If an autocorrelation coefficient at a specific lag crosses these dashed lines, it indicates that the autocorrelation at that lag is statistically significant at the 95% confidence level. Points outside these bounds are considered to have statistically significant autocorrelation, suggesting that the observed correlation is unlikely to have occurred by random chance.

Remedying autocorrelation involves several approaches, depending on the nature of the data and the context of the analysis. Here are some common methods to mitigate autocorrelation:

- Look for an improved model. Look for missing predictor variables, since these variables often show up as autocorrelated residuals.
- Improve the measurement/experiment to remove the dependence of time, space, and order.
- Preforming a suitable change of variables.

Here, we would like to introduce a transformation (change of variables) method for the regression model with first-order autoregressive errors given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{and} \quad \epsilon_i = \phi \epsilon_{i-1} + a_i$$

where $a_i \sim \text{NID}(0, \sigma^2)$, i.e. a_i is normally and independently distributed random variable with mean 0 and variance σ^2. For the transform method, let us define new variables y' and x' as follows:

$$y'_i = y_i - \phi y_{i-1}, \quad x'_i = x_i - \phi x_{i-1}$$

We will show how these transformations lead to a new model that would not have autocorrelation. With $y'_i = y_i - \phi y_{i-1}, x'_i = x_i - \phi x_{i-1}$ and $\epsilon_i = \phi\epsilon_{i-1} + a_i$, we see that

$$\begin{aligned} y'_i &= y_i - \phi y_{i-1} \\ &= \beta_0 + \beta_1 x_i + \epsilon_i - \phi(\beta_0 + \beta_1 x_{i-1} + \epsilon_{i-1}) \\ &= \beta_0(1-\phi) + \beta_1(x_i - \phi x_{i-1}) + \epsilon_i - \phi\epsilon_{i-1} \\ &= \beta'_0 + \beta'_1 x'_i + a_i \end{aligned}$$

where $\beta'_0 = \beta_0(1-\phi)$ and $\beta'_1 = \beta_1$. This gives a better transformed model, where a_i is not autocorrelated. Since the errors are from the AR(1) process, the value of ϕ can be obtained from computing $\phi = \rho_1$ using the acf() function in R (set as r in the R code in Fig 2.33 below). Now, with $y'_i = y_i - ry_{i-1}, x'_i = x_i - rx_{i-1}$ coded in R, we will do an "ordinary" regression between y'_i and x'_i. Fig 2.33 below shows how this change of variables is done in R.

```r
library(tseries)
library(lmtest)

# Read data
data1 <- read.csv("Autoc1.CSV", header=TRUE, sep=",")
x <- data1$x
y <- data1$y

# Fit linear model and perform Durbin-Watson test
model1 <- lm(y ~ x)
dwtest(model1)

# Calculate autocorrelation of residuals
rho <- acf(resid(model1), plot = FALSE)
acf(model1$residuals)

# Extract first-order autocorrelation coefficient
# Lag-0 is indexed as 1, so lag-1 is indexed as 2
r <- rho$acf[2]

# Get length of residuals
n <- length(model1$residuals)

# Initialize transformed variables
yprime <- rep(0, n-1)
xprime <- rep(0, n-1)

# Transform the data to remove autocorrelation
for(i in 1:(n-1)) {
  yprime[i] = data1$y[i+1] - r * data1$y[i]
  xprime[i] = data1$x[i+1] - r * data1$x[i]
}

#Fit a new model to the transformed data
#Create the ACF-plot
model2 <- lm(yprime ~ xprime)
acf(model2$residuals)
```

Figure 2.33: The figure shows how the change of variables is performed using R.

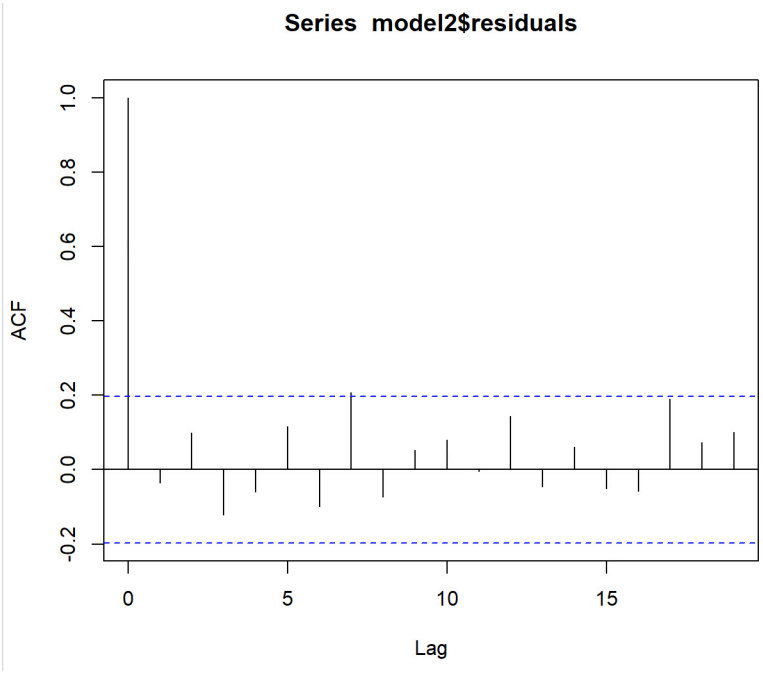


Figure 2.34: The figure shows the new ACF plot for the transformed model. Comparing it with Fig 2.32, it is clear that the issue of autocorrelation has been greatly remediated.

Fig 2.34 shows the ACF-plot for the new transformed model. It is clear from the plot that the correlation of each lag stay reasonably well within the blue-dashed lines. This implies that there is no significant autocorrelation at those lags beyond what would be expected by random chance.

Finally, the correct estimate of the intercept and the slope of the original model y vs x is then calculated as

$$\beta_0 = \frac{\beta_0'}{1-r} \quad \text{and} \quad \beta_1 = \beta_1'$$

3. Non-constant Variance of Error Terms

Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{Var}(\epsilon_i) = \sigma^2$. The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption. In fact, we have the following issues when the errors ϵ_is have non-constant variances (also known as *heteroscedasticity*):

- Model Efficiency: In the presence of heteroscedasticity, Ordinary Least Squares (OLS) estimators are still unbiased and consistent, meaning they still provide estimates that are close to the true population parameters and converge to them as the sample size increases. However, OLS estimators are no longer efficient. Efficient estimators are those that have the smallest variance among all unbiased estimators. Heteroscedasticity increases the variance of OLS estimators, reducing their efficiency compared to what could be achieved with heteroscedasticity-robust methods.

- Incorrect Inferences: Heteroscedasticity can lead to incorrect inferences about the statistical significance of the coefficients. Specifically, standard errors of the coefficients tend to be underestimated in the presence of heteroscedasticity. This means that t-tests and F-tests used for hypothesis testing may indicate significance where none exists, leading to misleading conclusions about the importance of variables in the model.

- Biased Estimates of Variances: When heteroscedasticity is present, the usual estimates of the variances of the coefficients (covariance matrix) can be biased. This affects the

calculation of confidence intervals and hypothesis tests based on the normal distribution assumptions, as these rely on accurate estimates of variances.

Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response. One can identify non-constant variances in the errors, from the presence of a *funnel shape* in the residual plot. An example is shown in the left-hand panel of Fig 2.35, in which the magnitude of the residuals tends to increase with the fitted values.

When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. The right-hand panel of Fig 2.35 displays the residual plot after transforming the response using $\log Y$. The residuals now appear to have constant variance, though there is some evidence of a slight non-linear relationship in the data.

Sometimes we have a good idea of the variance of each response. For example, the ith response could be an average of n_i raw observations. If each of these raw observations is uncorrelated with variance σ^2, then their average has variance $\sigma_i^2 = \sigma^2/n_i$, as an example, see:

$$\begin{aligned}\text{Var}(y_i) &= \text{Var}\left(\frac{1}{n_i}\sum_{k=1}^{n_i} y_k\right) \\ &= \frac{1}{n_i^2}\sum_{k=1}^{n_i}\text{Var}(y_i) \\ &= \frac{1}{n_i^2} n_i\sigma^2 = \frac{\sigma^2}{n_i}\end{aligned}$$

In this case a simple remedy is to fit our model by *weighted least squares*, with weights proportional to the inverse variance—i.e. $w_i = n_i$ in this case. Most linear regression software allows for observation weights. Let us see how this is done mathematically.

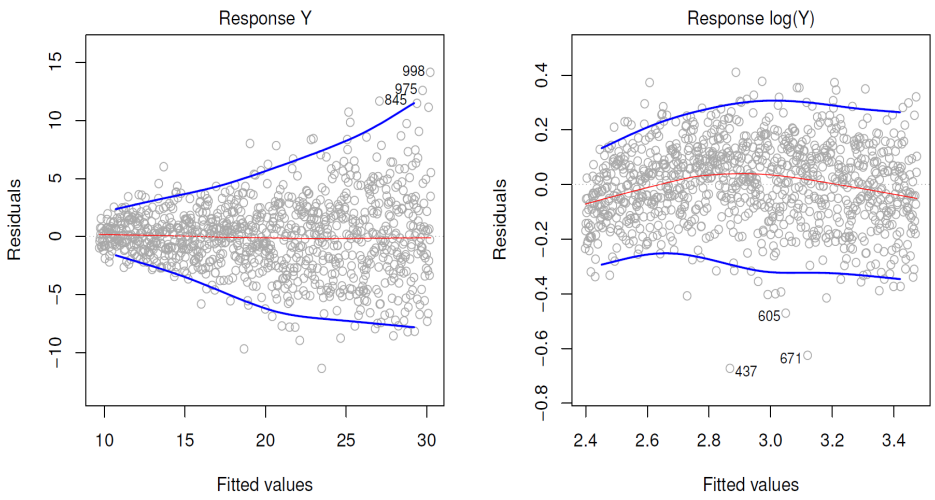


Figure 2.35: Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the of the residuals, and emphasize patterns. Left: The funnel shape indicated heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.

In a standard multiple linear regression, the model is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

where y_i is the dependent variable, x_{ij} (for $j = 1, \ldots, p$) are the predictor variables, β_j are the coefficients to be estimated, and ϵ_i are the error terms assumed to be independently and identically distributed with $\epsilon_i \sim N(0, \sigma^2)$.

In the weighted multiple linear regression (WMLR), each observation i is assigned a weight w_i. The weighted least squares estimators are then obtained by minimizing the weighted sum of squared residuals:

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

Let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights w_i and using the same notation as from the derivation of (2.60), where (2.60) is restated below

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

the weighted least squares estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are now given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \tag{2.88}$$

To show this, let's write the objective function for WLS as:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

where:

- y_i is the observed value of the dependent variable for the i-th observation.
- $\mathbf{x}_i$ is the $(p+1) \times 1$ vector of predictor variables for the i-th observation, namely $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{ip})^T$.
- $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients.
- w_i is the weight for the i-th observation.

In matrix notation, this can be written as:

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

where:

- $\mathbf{y}$ is an $n \times 1$ vector of observed values.
- $\mathbf{X}$ is an $n \times (p+1)$ matrix of predictors (including a column of ones for the intercept).
- $\mathbf{W}$ is an $n \times n$ diagonal matrix of weights with w_i on the diagonal.

Expanding $Q(\boldsymbol{\beta})$ we obtain

$$Q(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{W}\mathbf{y} - \mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$$

Since $\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ is a scalar, it is equal to its transpose:

$$\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = (\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y}$$

Thus, the objective function simplifies to:

$$Q(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{W}\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$$

To find the minimum of $Q(\boldsymbol{\beta})$, we take the derivative with respect to $\boldsymbol{\beta}$:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \left(\mathbf{y}^T\mathbf{W}\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}\right)$$

Notice that we are using the notation $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ as a $(p+1) \times 1$ vector representing

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \cdots, \frac{\partial Q}{\partial \beta_p}\right)^T$$

The first term $\mathbf{y}^T\mathbf{W}\mathbf{y}$ is a constant with respect to $\boldsymbol{\beta}$, so its derivative is zero. For the second term $-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y}$, it is easy to see that

$$\frac{\partial}{\partial \boldsymbol{\beta}}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y}) = -2\mathbf{X}^T\mathbf{W}\mathbf{y}$$

For the third term, using the fact that if $\mathbf{A}$ is a symmetric matrix, then

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}\right) = 2\mathbf{A}\boldsymbol{\beta} \tag{2.89}$$

we see that the derivative of $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$, with respect to $\boldsymbol{\beta}$ is $2\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ (since $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is symmetric). Combining these results, we write

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{W}\mathbf{y} + 2\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}$$

Setting $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ to find the minimum, we get

$$-2\mathbf{X}^T\mathbf{W}\mathbf{y} + 2\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = 0$$

Rearranging gives

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{W}\mathbf{y}$$

Finally, multiplying both sides by $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ to get

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

Therefore the WLS estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

this is what we wanted to show.

Reamrk: Here, we give a proof on

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} \right) = 2\mathbf{A}\boldsymbol{\beta}$$

that was used earlier in (2.89). Consider the quadratic form:

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}^T A \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector and A is a $(p+1) \times (p+1)$ symmetric matrix. Let us write $\boldsymbol{\beta}^T A \boldsymbol{\beta}$ in summation notation:

$$\boldsymbol{\beta}^T A \boldsymbol{\beta} = \sum_{i=1}^{p+1} \sum_{j=1}^{p+1} \beta_i A_{ij} \beta_j$$

The partial derivative of $f(\boldsymbol{\beta})$ with respect to β_k is then given by

$$\begin{aligned} \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \left(\sum_{i=1}^{p+1} \sum_{j=1}^{p+1} \beta_i A_{ij} \beta_j \right) \\ &= \sum_{i=1}^{p+1} \sum_{j=1}^{p+1} \left(\frac{\partial}{\partial \beta_k} (\beta_i A_{ij} \beta_j) \right) \end{aligned}$$

Next, we apply the product rule. The term $\beta_i A_{ij} \beta_j$ depends on β_k in two cases: when $i = k$ and when $j = k$. So we need to consider both cases:

$$\frac{\partial}{\partial \beta_k} (\beta_i A_{ij} \beta_j) = \delta_{ik} A_{ij} \beta_j + \beta_i A_{ij} \delta_{jk}$$

where δ_{ik} and δ_{jk} are the Kronecker delta, which is 1 if $i = k, j = k$ and 0 otherwise. Sum over all i and j to obtain

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^{p+1} A_{kj} \beta_j + \sum_{i=1}^{p+1} \beta_i A_{ik}$$

Since A is symmetric ($A_{ik} = A_{ki}$), this simplifies to:

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^{p+1} A_{kj} \beta_j + \sum_{i=1}^{p+1} \beta_i A_{ki} = 2 \sum_{j=1}^{p+1} A_{kj} \beta_j = 2(A\boldsymbol{\beta})_k$$

Therefore we obtain

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} \right) = 2\mathbf{A}\boldsymbol{\beta}.$$

The selection of weights is critical and varies depending on the context of the data. We conclude this discussion with one more example. Consider the following model

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, x\sigma^2)$$

where the predictor takes only positive values. Taking a sample of n data points (x_i, y_i), we can write the model as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, x_i \sigma^2), \quad i = 1, \cdots, n.$$

In this case, $\text{Var}(y_i) = \text{Var}(\epsilon_i) = x_i\sigma^2$. To stabilize the variance, we let

$$\epsilon_i' = \frac{\epsilon_i}{\sqrt{x_i}} \sim \text{N}(0, \sigma^2), \quad i = 1, \cdots, n$$

We then rewrite the original model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

as follow

$$\frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + \beta_1 \sqrt{x_i} + \epsilon_i', \quad i = 1, \cdots, n$$

where

$$\epsilon_i' = \frac{\epsilon_i}{\sqrt{x_i}} \sim \text{N}(0, \sigma^2), \quad i = 1, \cdots, n$$

The sum of the weighted errors are then given by

$$\sum_{i=1}^{n} (\epsilon_i')^2 = \sum_{i=1}^{n} \left(\frac{y_i}{\sqrt{x_i}} - \beta_0 \frac{1}{\sqrt{x_i}} - \beta_1 \sqrt{x_i} \right)^2 = \sum_{i=1}^{n} \frac{1}{x_i} (y_i - \beta_0 - \beta_1 x_i)^2$$

The weighted least-squares function which we want to minimize is then given by

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^{n} \frac{1}{x_i} (y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

with $w_i = 1/x_i$ as weight that can help stabilized the variance of the error terms.

4. Outliers

An *outlier* is a point for which y_i is far from the value predicted by the model. (Note: sometimes the term *regression outlier* is used to differentiate it from its day-to-day usage.) Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

The red point (observation 20) in the left-hand panel of Fig 2.36 illustrates a typical outlier. The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier. In this case, removing the outlier has little effect on the least squares line: it leads to almost no change in the slope, and a miniscule reduction in the intercept. It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit. However, even if an outlier does not have much effect on the least squares fit, it can cause other problems. For instance, in this example, the RSE is 1.09 when the outlier is included in the regression, but it is only 0.77 when the outlier is removed. Since the RSE is used to compute all confidence intervals and p-values, such a dramatic increase caused by a single data point can have implications for the interpretation of the fit. Similarly, inclusion of the outlier causes the R^2 to decline from 0.892 to 0.805.

Residual plots can be used to identify outliers. In this example, the outlier is clearly visible in the residual plot illustrated in the center panel of Fig 2.36. But in practice, it can be difficult to decide how large a residual needs to be before we consider the point to be an outlier.

To address this problem, instead of plotting the residuals, we can plot the *Studentized residuals*, computed by dividing each residual e_i by its estimated standard error. (We will say more about this later.) Observations whose Studentized residuals are greater than 3 in absolute value are possible outliers. In the right-hand panel of Fig 2.36, the outlier's Studentized residual exceeds 6, while all other observations have Studentized residuals between -2 and 2.

If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation. However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.

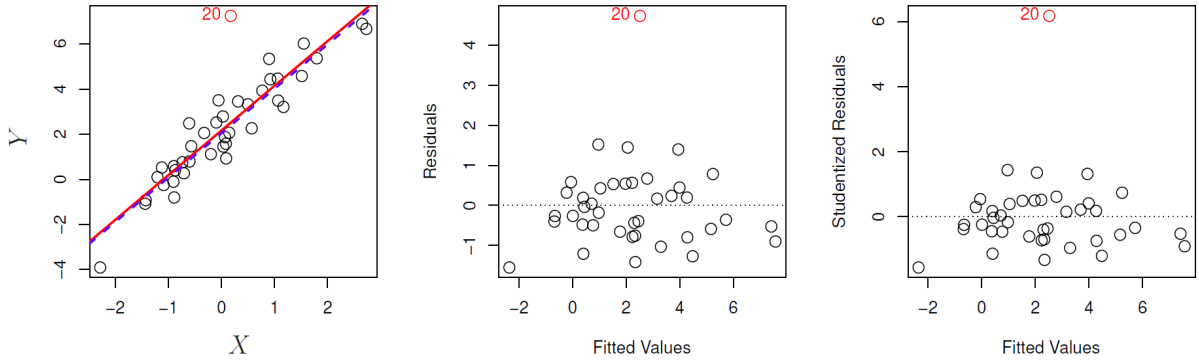


Figure 2.36: Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a Studentized residual of 6; typically we expect values between -3 and 3.

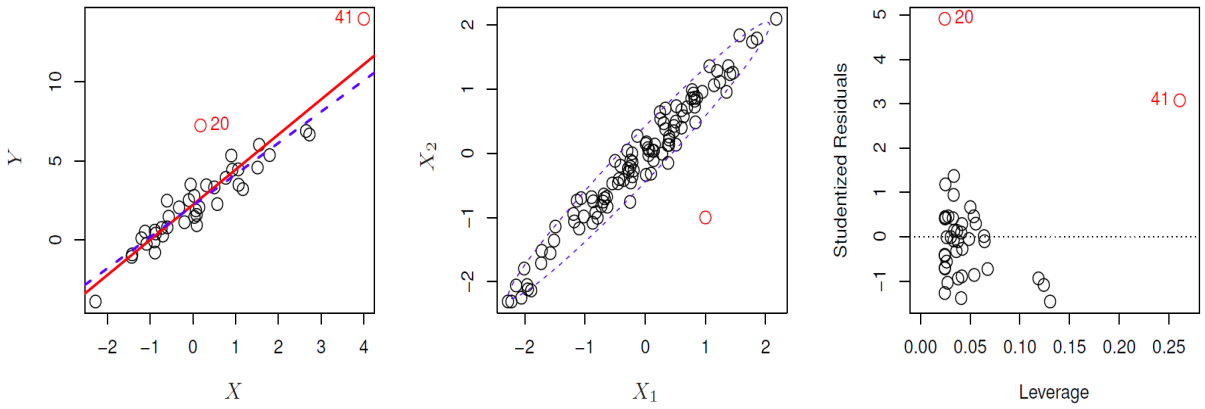


Figure 2.37: Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

5. High Leverage Points

We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i. In contrast, observation with *high leverage* have an usual value for x_i. For example, observation 41 in the left-hand panel of Fig 2.37 has high leverage, in that the predictor value

for this observation is large relative to the other observations. (Note that the data displayed in Fig 2.37 are the same as the data displayed in Fig 2.36 but with the addition of a single high leverage observation.)

The red solid line is the least squares fit to the data, while the blue dashed line is the fit produced when observation 41 is removed. Comparing the left-hand panels of Fig 2.36 and Fig 2.37, we observe that removing the high leverage observation has a much more substantial impact on the least squares line than moving the outlier. In fact high leverage observations tend to have a sizable impact on the estimated regression line. It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problems with these points may invalidate the entire fit. For this reason, it is important to identify high leverage observations.

In a simple linear regression, high leverage observations are fairly easy to identify, since we can simply look for observations for which the predictor value is outside of the normal range of observations. But in a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors. An example is shown in the center panel of Fig 2.37, for a data set with predictors, X_1 and X_2. Most of the observations' predictor values fall within the blue dashed ellipse, but the red observation is well outside of this range. But neither its value for X_1 nor its value for X_2 is unusual. So if we examine just X_1 or just X_2, we will fail to notice this high leverage point. This problem is more pronounced in multiple regression settings with more than two predictors, because then there is no simple way to plot all dimensions of the data simultaneously.

In order to quantify an observation's leverage, we compute the *leverage statistic*. A large value of this statistic indicates an observation with high leverage. For a simple linear regression, the leverage statistic is

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{k=1}^{n} (x_k - \overline{x})^2}. \tag{2.90}$$

It is clear from this equation that h_i increases with the distance of x_i from $\overline{x}$. There is a simple extension of h_i to the case of multiple predictors, which we will discussed later. The leverage statistic h_i is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$. So if a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

The right-hand panel of Fig 2.37 provides a plot of the Studentized residuals versus h_i for the data in the left-hand panel of Fig 2.37. Observation 41 stands out as having a very high leverage statistic as well as a high Studentized residual. In other words, it is an outlier as well as a high leverage observation. This is a particularly dangerous combination! This plot also revels the reason that observation 20 had relatively little effect on the least squares fit in Fig 2.36: it has low leverage.

We now fill in some of the details that we left out earlier. We begin with the concept of *hat matrix* (also known as *projection matrix*). Recall from (2.60) that the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Therefore the vector of fitted values $\widehat{y}_i$ corresponding to the observed values y_i is

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{2.91}$$

where the $n \times n$ matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{2.92}$$

is usually called the *hat matrix*. It maps the vector of observed values into vector of fitted values in the way given by (2.91):

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Hence every predicted value $\widehat{y}_i$ is actually a linear combination of $y_1, y_2, \cdots, y_n$:

$$\widehat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n,$$

where h_{ij} is the (i,j)th element of the matrix $\mathbf{H}$ and it is completely determined by the predictors $\mathbf{X}$ as seen in (2.92). Now

- h_{ij} is the *weight* given to y_j in predicting $\widehat{y}_i$,
- h_{ii} is the *weight* given to y_i in predicting $\widehat{y}_i$, and this is called the *leverage* of the ith observation, for $i = 1, 2, \cdots, n$.

Recall that (2.90) gives the formula for h_{ii} when there is only one predictor in the model. For multiple regression model, we can compute h_{ij} from the matrix multiplication given by (2.92).

The reason why $\mathbf{H}$ is also known as the projection matrix stem from its geometrical interpretation in linear algebra. In this setting, the least squares error, $\mathbf{e}$, is orthogonal to the column space of matrix $\mathbf{X}$, denoted as $\text{Col}(\mathbf{X})$, while $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ is the orthogonal projection of the vector $\mathbf{y}$ onto the column space of matrix $\mathbf{X}$. See Fig 2.38.

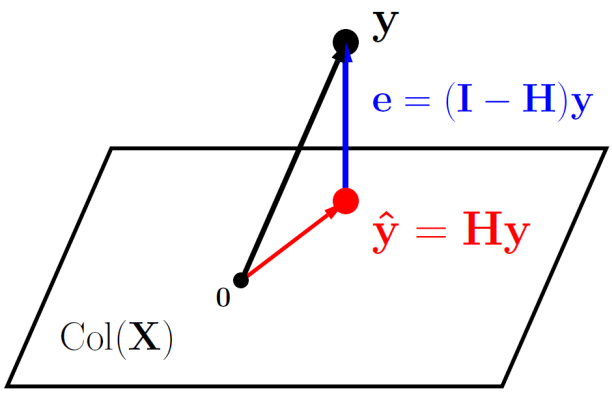


Figure 2.38: The geometrical interpretation of the hat matrix using the language of linear algebra, where $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ is the orthogonal projection of the vector $\mathbf{y}$ onto the column space of matrix $\mathbf{X}$.

As mentioned earlier, in a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictor (see the center panel of Fig 2.37). See also Fig 2.39. This phenomenon is also refer to as an *extrapolation* beyond the region containing the original observations.

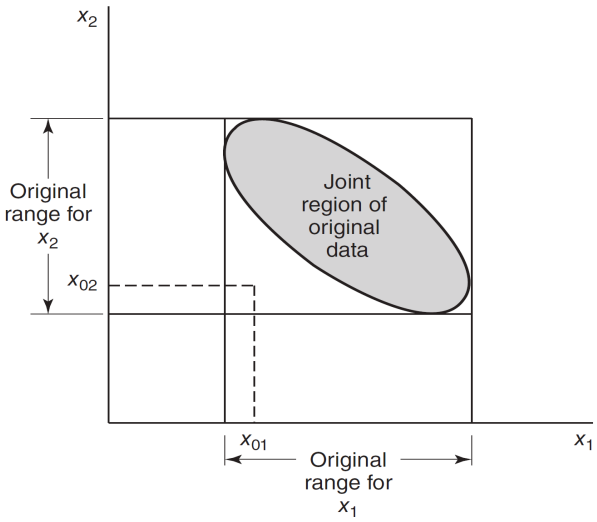


Figure 2.39: An example of extrapolation in multiple regression, where (x_{01}, x_{02}) is unusual, but individually x_{01} and x_{02} are not.

Since simply comparing the levels of x's for a new data point with the ranges of the original x's will not always detect a hidden extrapolation, it would be helpful to have a formal procedure to do so. Let us define the smallest convex set containing all of the original n data points $(x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, 2, \cdots, n$, as the regressor variable hull (RVH).

Now, the diagonal elements h_{ii} of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ are going to be useful in detecting *hidden extrapolation*. The values of h_{ii} depend both on the Euclidean distance of the point $\mathbf{x}_i$ from the centroid and on the density of the points in the RVH. In general, the point that has the largest value of h_{ii}, say $h_{\max}$, will lie on the boundary of the RVH in a region of the x space where the density of the observation is relatively low. The set of points $\mathbf{x}$ (not necessarily data points used to fit the model) that satisfy

$$\mathbf{x}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x} \leq h_{\max}$$

is an ellipsoid enclosing all points inside the RVH. Thus, if we are interested in prediction or estimation at the point $\mathbf{x}_0^T = (1, x_{01}, x_{02}, \cdots, x_{0p})$, the location of that point relative to the RVH is reflected by

$$h_{00} = \mathbf{x}_0^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_0.$$

Points for which $h_{00} > h_{\max}$ are outside the ellipsoid enclosing the RVH and are extrapolation points. Generally the smaller the value of h_{00}, the closer the point $\mathbf{x}_0$ lies to the centroid of the x space. This is why the $\mathbf{H}$ matrix can help us identify leverage points, which are points that are unusual in the x space.

We next look at some methods for scaling residuals, which can help identify an outlying observations. We need an index of the unusualness of Y given the Xs, and it turns out that residuals can help us with this task. Recall we have previously defined the residuals as

$$e_i = y_i - \widehat{y}_i, \quad i = 1, 2, \cdots, n \tag{2.93}$$

where y_i is an observation and $\widehat{y}_i$ is its corresponding fitted value. Note that residuals do not have equal variance, in spite of fact that the errors ϵ_i are assumed to have equal variance. To show this, the concept of the $\mathbf{H}$ matrix is going to be useful again.

Let us first begin the study of *scaled residual* with the following observations: The n residuals $e_i = y_i - \widehat{y}_i, \quad i = 1, \cdots, n$ can be conveniently written in matrix notation as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} \tag{2.94}$$

With (2.91) and (2.94), we write

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{2.95}$$

Substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ into (2.95) yields

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \end{aligned} \tag{2.96}$$

Thus, the residuals can be written as a linear transformation of the observations $\mathbf{y}$ and the errors $\boldsymbol{\epsilon}$, see (2.95) and (2.96).

The hat matrix has several useful properties. It is *symmetric*, that is, $\mathbf{H}^T = \mathbf{H}$, and it is *idempotent*, that is, $\mathbf{H}\mathbf{H} = \mathbf{H}$. Hence, the matrix $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent, that is, $(\mathbf{I} - \mathbf{H})^T = (\mathbf{I} - \mathbf{H})$ and $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$. These properties are going to be useful in the computation of $\mathrm{Var}(\mathbf{e})$ below.

Recall from (2.80) that for a non-random matrix $\mathbf{A}$, we have $\mathrm{Var}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T$. Now since the matrix $\mathbf{H}$ is purely determined by the predictors, it is a non-random matrix. Using $\mathbf{A} = (\mathbf{I} - \mathbf{H})$ and (2.80), we see that the covariance matrix of the residuals is

$$\begin{aligned} \mathrm{Var}(\mathbf{e}) &= \mathrm{Var}\left[(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\right] \\ &= (\mathbf{I} - \mathbf{H})\,\mathrm{Var}(\boldsymbol{\epsilon})\,(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned} \tag{2.97}$$

since $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ and the matrix $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent.

The matrix $\mathbf{I} - \mathbf{H}$ is generally not diagonal, so the residuals have different variances and they are correlated. From (2.97), we see that the variance of the ith residual is

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}) \tag{2.98}$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$. Also from (2.97) we see that the covariance between residuals e_i and e_j is

$$\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij} \tag{2.99}$$

where h_{ij} is the (i, j)th element of the hat matrix. Notice from (2.98) that high leverage observations (those with h_{ii} value close to 1) tend to have small residuals—an intuitively sensible result because these observations can pull the regression surface toward them.

Now, a simple way to scale the residual is as follow:

$$d_i = \frac{e_i}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}}, \quad i = 1, 2, \cdots, n \tag{2.100}$$

Note: Since we do not know the true value of σ^2, we will compare the residuals against its point estimate $\mathrm{MS}_{\mathrm{Res}}$.

The *normalized residuals* d_i have mean zero and approximately unit variance. Consequently, a large normalized residual ($|d_i| > 3$, say) potentially indicates an outlier. However, this is not a correct way to standardize the residuals. To correctly standardize the residuals, we need to obtain the exact standard deviation of the residual e_i, which we manage to do in (2.98).

Studentized residuals:

Taking all these into account, the *standardized* residuals, also called the *(internally) Studentized* residuals are defined as

$$r_i = \frac{e_i}{\sqrt{\mathrm{MS}_{\mathrm{Res}}\,(1 - h_{ii})}}, \quad i = 1, 2, \cdots, n \tag{2.101}$$

The internally Studentized residuals have mean zero and constant variance regardless of the location of x_i when the form of the model is correct. Also for large data set $h_{ii} \approx 0$, so d_i and r_i have little difference in those cases. This measure is slightly inconvenient because it's numerator and denominator are not independent, preventing r_i from following a t-distribution.

The *externally Studentized* residuals are defined as

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}\,(1 - h_{ii})}}, \quad i = 1, \cdots, n \tag{2.102}$$

where $S^2_{(i)}$ is an estimate of σ^2 obtained by fitting a linear regression model to all data but the ith observation. Note that the e_i are still computed from the model on the full data set. The externally Studentized residual has an independent numerator and denominator and follows a t-distribution with $n - p - 2$ degrees of freedom.

Externally Studentized residuals t_i can be calculated from internally studentized residual r_i via

$$t_i = r_i \sqrt{\frac{n - p - 2}{n - p - 1 - r_i^2}} \tag{2.103}$$

If n is large, then the factor under the square root in (2.103) is close to 1, and the distinction between the two Studentized residuals essentially disappear. For outliers detection, it's common use $|r_i|$ and $|t_i|$ greater than 3 as an indication of a potential outlier.

Remark: Here, as elsewhere in statistics, terminology is not wholly standard: t_i is sometimes called a *deleted Studentized residual*, an *externally Studentized residual*, or even a *standardized residual*; likewise, r_i is called an *internally Studentized residual*, or simply a *Studentized residual*. It is therefore helpful, especially in small samples, to determine exactly what is being calculated by a computer program.

Remark: Studentized residuals are named after the statistician William Sealy Gosset, who published under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland, where he developed statistical methods to improve the brewing process. He introduced the concept of the t-distribution and the t-test, which are fundamental in statistical inference when sample sizes are small and the population standard deviation is unknown.

Influence Points

Previously, we discussed the concept of (regression) outliers and high leverage observations separately, while hinted that some combination of these two behaviors can have a huge impact on the regression plane. Here, we would like to make this more precise by making the following definition:

An *influence point* is an observation whose removal from the data set would cause a large change in estimated regression model coefficients.

Recall that:

- *Leverage*: High leverage points are observations with unusual predictor values that can disproportionately affect the regression line. These points are far from the centroid of the predictor variables and have the potential to pull the regression line towards themselves.
- *Outlier*: High residuals indicate observations where the actual value is far from the predicted value. When combined with high leverage, these outliers can indicate influential points.

Combined Effect: The *influential points* are those that not only have high leverage but also large residuals. These points have a significant impact on the estimation of regression coefficients and the fitted values.

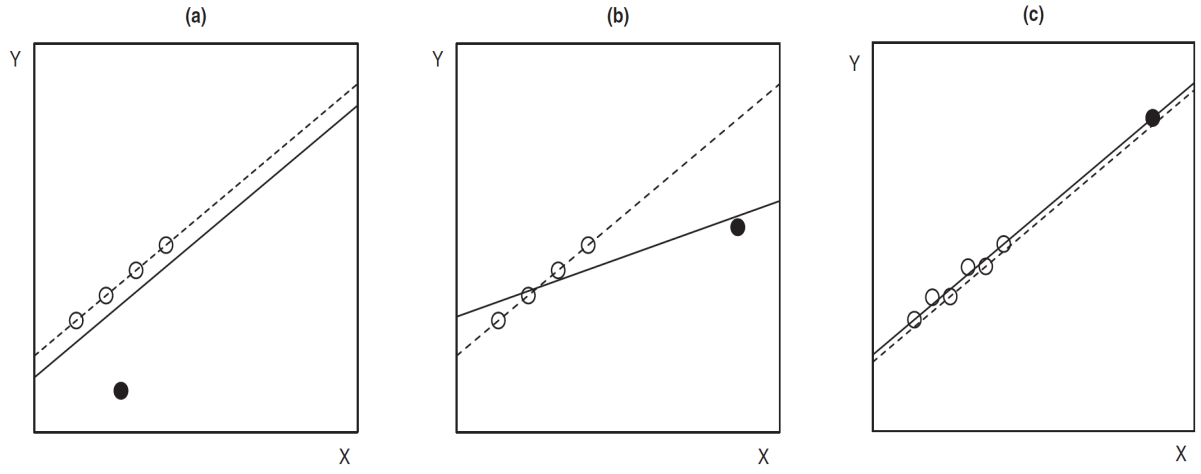


Figure 2.40: In each graph, the solid line gives the least squares regression for all the data, while the broken line gives the least squares regression with the unusual data point (the black circle) omitted. In panel (a), an outlier near the mean of X has low leverage and little influence on the regression coefficients. In panel (b), An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect but are, in fact, coincident.

Fig 2.40 above give three examples on how leverage point and outlier can combine to produce influence point. Panel (c) of Fig 2.40 shows that a leverage point may have no influence if the observation lies close to the regression line. Panel (a) and (b) of Fig 2.40 indicate that a point has to have at least some leverage in order to be influential.

Cook's Distance, introduced by R. Dennis Cook in 1977, is a measure used in regression analysis to identify influential data points.

Cook's Distance for the i-th observation, denoted by D_i, quantifies the influence of removing the i-th observation on the regression coefficients. It is defined as:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\sigma^2}$$

where:

- $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients using all n observations.
- $\hat{\boldsymbol{\beta}}_{(i)}$ is the vector of estimated coefficients with the i-th observation removed.
- $p+1$ is the number of parameters in the model (including the intercept).
- σ^2 is the estimated variance of the errors.

Cook's Distance can also be expressed in terms of the residuals ($e_i = y_i - \hat{y}_i$) and the leverage values (h_{ii}). Using these, Cook's Distance can be rewritten as:

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2} \tag{2.104}$$

where $\hat{\sigma}^2$ is the mean squared error estimate of σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-(p+1)}$$

Using the Studentized residual r_i, see (2.101), we can also write D_i as

$$D_i = \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}} \tag{2.105}$$

Cook's Distance measures how much the regression coefficients change when an observation is removed. If removing an observation leads to a substantial change in the coefficients, it indicates that the observation is influential. The common thresholds are listed below:

- $D_i > \frac{4}{n}$: An observation is often considered influential if its Cook's distance exceeds $\frac{4}{n}$, where n is the number of observations.
- $D_i > 1$: Another common rule of thumb is that observations with Cook's distance greater than 1 are considered to have high influence.

6. Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The concept of collinearity is illustrated in Fig 2.41 using the Credit data set. In the left-hand panel of Fig 2.41, the two predictors limit and age appear to have no obvious relationship. In contrast, the right-hand panel of Fig 2.41, the predictors limit and rating are very highly correlated with each other, and we say that they are *collinear*. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. In other words, since limit and rating tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response balance.

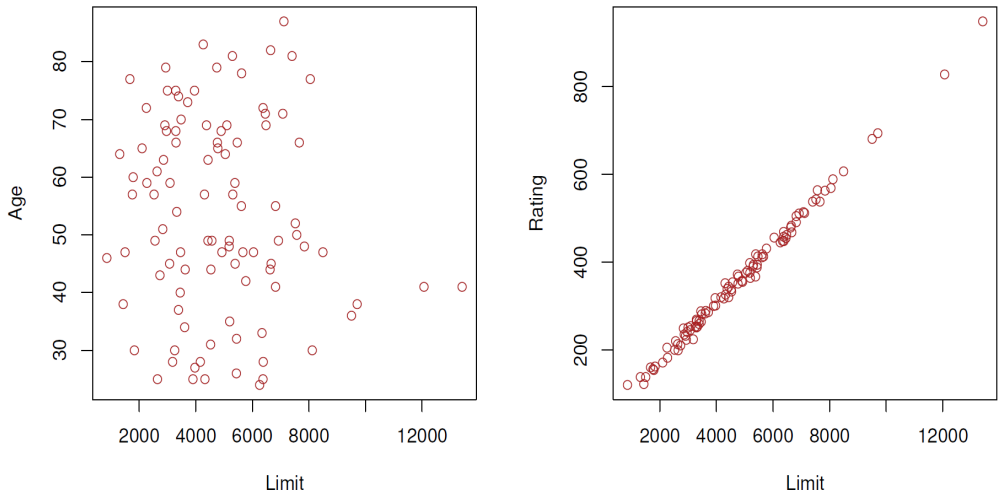


Figure 2.41: Scatterplots of the observations from Credit data set. Left: A plot of age versus limit. These two variables are not collinear. Right: A plot of rating versus limit. There is high collinearity.

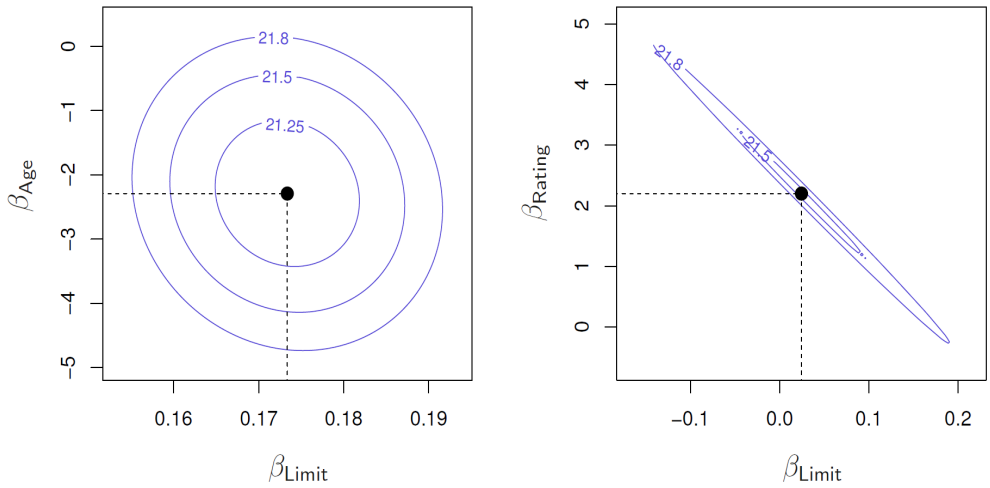


Figure 2.42: Contour plot for the RSS values as a function of the parameters β for various regressions involving the Credit data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of balance onto age and limit. The minimum value is well defined. Right: A contour plot of RSS for the regression of balance onto rating and limit. Because of the collinearity, there are many pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a smimilar value for RSS.

Fig 2.42 illustrates some of the difficulties that can result from collinearity. The left-hand panel of Fig 2.42 is a contour plot of the RSS (stated below)

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

associated with different possible coefficient estimates for the regression of balance onto limit and age. Each ellipse represents a set of coefficients that correspond to the same RSS with ellipses nearest to the center taking on the lowest values of RSS. The black dots and associated dashed lines represent the coefficient estimates that result in the smallest possible RSS—in other words, these are the least squared estimates. The axes for the limit and age have been

scaled so that the plot includes possible coefficient estimates that are up to four standard errors on either side of the least squares estimates. Thus the plot includes all plausible values for the coefficients. For example, we see that the true limit coefficient is almost certainly somewhere between 0.15 and 0.20.

In contrast, the right-hand panel of Fig 2.42 displays contour plots of the RSS associated with possible coefficient estimates for the regression of balance onto limit and rating, which we know to be highly collinear. Now the contours run along a narrow valley; there is a broad range of values for the coefficient estimates that result in equal values for RSS. Hence a small change in the data could cause the pair of coefficient values that yield the smallest RSS—that is, the least squares estimate—to move anywhere along this valley. This results in a great deal of uncertainty in the coefficient estimates. Notice that the scale for limit coefficient now runs from roughly -0.2 to 0.2; that is an eight-fold increase over the plausible range of the limit coefficient in the regression with age. Interestingly, even though the limit and rating coefficients now have much more individual uncertainty, they will almost certainly lie somewhere in this contour valley. For example, we would not expect the true value of the limit and rating coefficients to be -0.1 and 1 respectively, even though such a value is possible for each coefficient individually.

Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\widehat{\beta}_j$ to grow. Recall that the t-statistic for each predictor is calculated by dividing $\widehat{\beta}_j$ by its standard error. Consequently, collinearity results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that the *power* of the hypothesis test—the probability of correctly detecting a *non-zero* coefficient—is reduced by collinearity.

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	-173.411	43.828	-3.957	< 0.0001
	`age`	-2.292	0.672	-3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	-377.537	45.254	-8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

Figure 2.43: The results for two multiple regression models involving the Credit data set are shown. Model 1 is a regression of balance on age and limit, and Model 2 is a regression of balance on rating and limit. The standard error of $\widehat{\beta}_{\text{limit}}$ increases 12-fold in the second regression, due to collinearity.

Fig 2.43 compares the coefficient estimates obtained from two seperate multiple regression models. The first is a regression of balance on age and limit, and the second is a regression of balance on rating and limit. In the first regression, both age and limit are highly significant with very small p-values. In the second, the collinearity between limit and rating has caused the standard error for the limit coefficient estimate to increase by a factor of 12 and the p-value to increase to 0.701. In other words, the importance of the limit variable has been masked due to the presence of collinearity.

Fig 2.44 provide further insight into collinearity, illustrating its effect on estimation when there are two explanatory variables in a regression. The black and gray dots in Fig 2.44 represent the data points (the gray dots are below the regression plane) while the white dots represent fitted values lying in the regression plane; the +s show the projection of the data onto the $\{X_1, X_2\}$-plane.

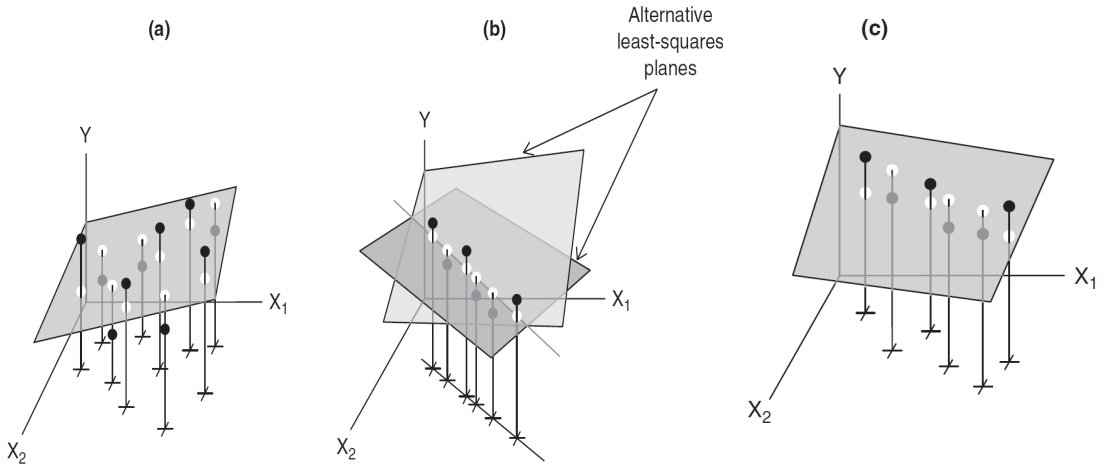


Figure 2.44: The impact of collinearity on the stability of the least squares regression plane. In (a), the correlation between X_1 and x_2 is small, and the regression plane therefore has a broad base of support. In (b), X_1 and X_2 are perfectly correlated; the least squares plane is not uniquely defined. In(c), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2; the least squares plane is uniquely defined, but it is not well "supported" by the data.

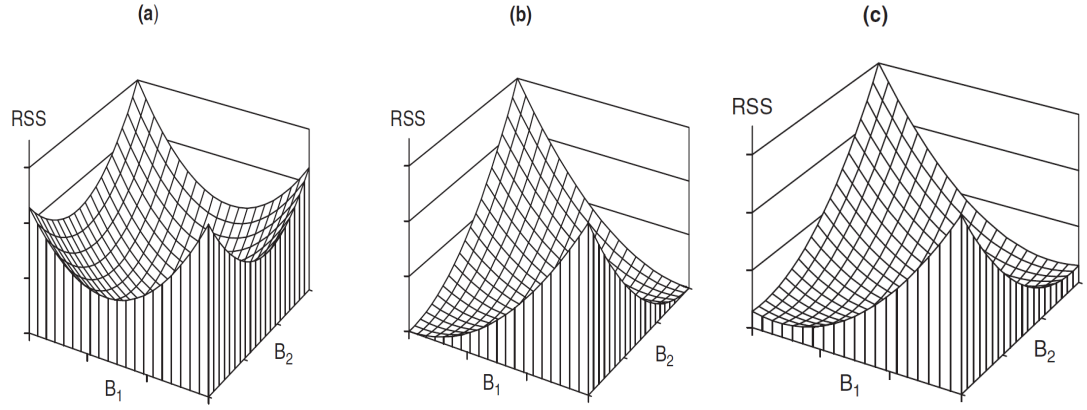


Figure 2.45: The residual sum of squares (RSS) as a function of the coefficients B_1 and B_2. In each graph, the vertical axis is scaled so that the least squares value of the RSS is at the bottom of the axis. When, as in (a), the correlation between the explanatory variables X_1 and X_2 is small, the RSS has a well-defined minimum, much like a deep bowl. When there is a perfect linear relationship between X_1 and X_2, as in (b), the RSS is flat at its minimum, above a line in the $\{B_1, B_2\}$-plane: The least squares values of B_1 and B_2 are not unique. When, as in (c), there is a strong, but less-than-perfect, linear relationship between the X_1 and X_2 , the RSS is nearly flat at its minimum, so values of B_1 and B_2 quite different from the least squares values are associated with the RSS near the minimum.

In Fig 2.44(a), the correlation between the explanatory variables X_1 and X_2 is slight, as indicated by the broad scatter of points in the $\{X_1, X_2\}$-plane. The least squares regression plane, also shown in this figure, therefore has a firm base of support. Correspondingly, Fig 2.44(a) shows that small changes in the regression coefficients are associated with relatively large increases in the residual sum of squares—the sum of squares (RSS) function is like a deep bowl, with steep sides and a well-defined minimum. See Fig 2.45 and also Fig 2.42.

Fig 2.44(b), X_1 and X_2 are perfectly collinear. Because the explanatory variable observations form a line in the $\{X_1, X_2\}$-plane, the least squares regression plane, in effect, also reduces to a line. The plane can tip about this line without changing the residual sum of squares, as Fig 2.45(b) reveals: The sum of squares function is flat at its minimum along a line defining pairs of values for B_1 and B_2—rather like a sheet of paper with two corners raised—and thus there are an infinite number of pairs of coefficients (B_1, B_2) that yield the minimum RSS.

Finally, in Fig 2.44(c), the linear relationship between X_1 and X_2 is strong, although not perfect. The support afforded to the least squares plane is tenuous, so that the plane can be tipped without causing large increases in the residual sum of squares, as is apparent in Fig 2.45(c)—the sum of squares function is like a shallow bowl with a nearly flat bottom and hence a poorly defined minimum.

We next give a simple mathematical example to illustrate the effects of collinearity on regression. Let us consider a multiple linear regression with two predictors x_1 and x_2 in the form of

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \cdots, n$$

where

- The response y is centered ($\overline{y} = 0$), and
- Both predictors have been normalized to have unit length and mean zero:

  $$\overline{x}_1 = \overline{x}_2 = 0$$

  as well as

  $$\sum_{i=1}^{n} x_{i1}^2 = 1 \text{ and } \sum_{i=1}^{n} x_{i2}^2 = 1$$

In this case, $\mathbf{X}^T\mathbf{X}$ is the correlation matrix between x_1 and x_2, since

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\end{aligned}$$

where r_{12} is the correlation between x_1 and x_2

$$r_{12} = \frac{\sum(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2)}{\sqrt{\sum(x_{i1} - \overline{x}_1)^2 \sum(x_{i2} - \overline{x}_2)^2}}.$$

Therefore we see that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

This gives

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{1 - r_{12}^2} = \text{Var}(\hat{\beta}_2)$$

and

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{r_{12}\sigma^2}{1 - r_{12}^2}.$$

From here it follows that

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{\beta}_2) \to \infty$$

as $r_{12} \to 1$ (or -1).

This means that in the case of strong collinearity ($|r_{12}| \approx 1$), a slightly different sample can lead to vastly different estimate of the model parameters.

Another effect of having collinearity in a model is that the parameters tend to be overestimates in magnitude. For this, we look at $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 = \sum_j (\hat{\beta}_j - \beta_j)^2$. Taking the expectation leads to

$$\begin{aligned} E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) &= \sum_j E\left[(\hat{\beta}_j - \beta_j)^2\right] \\ &= \sum_j \text{Var}(\hat{\beta}_j) \\ &= \text{Trace}\left(\text{Var}(\hat{\boldsymbol{\beta}})\right) \\ &= \sigma^2 \text{Trace}\left((\mathbf{X}^T\mathbf{X})^{-1}\right) \\ &= \frac{2\sigma^2}{1 - r_{12}^2} \end{aligned}$$

The equation

$$E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) = \frac{2\sigma^2}{1 - r_{12}^2}$$

indicates that when there is a strong collinearity, $\hat{\boldsymbol{\beta}}$ is far from $\boldsymbol{\beta}$ on average. Also, one can show that

$$E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) = E\left(||\hat{\boldsymbol{\beta}}||^2\right) - ||\boldsymbol{\beta}||^2$$

Therefore

$$E\left(||\hat{\boldsymbol{\beta}}||^2\right) = ||\boldsymbol{\beta}||^2 + \frac{2\sigma^2}{1 - r_{12}^2}$$

This implies that when there is a strong collinearity, the length of the vector $\hat{\boldsymbol{\beta}}$ is on average, much larger than $\boldsymbol{\beta}$.

To avoid such a situation, it is desirable to identify and address potential collinearity problems while fitting the model.

A simple way to detect collinearity is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data. Unfortunately, not all collinearity problems can be detected by inspection of the correlation matrix: it is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation *multicollinearity*.

Instead of inspecting the correlation matrix, a better way to access multicollinearity is to compute the *variance inflation factor* (VIF). The VIF is the ratio of the variance of $\widehat{\beta}_j$ when fitting the full model divided by the variance of $\widehat{\beta}_j$ if fit on its own.

Another way to think about the variance inflation factor is this: In order to detect whether there is a relationship between one predictor with a linear combination of the rest of the predictors, we regress each single predictor $X_j,\ j = 1, \cdots, p$ on the remaining ones, i.e.

$$X_j \sim X_1 + \cdots + X_{j-1} + X_{j+1} + \cdots + X_p \tag{2.106}$$

and compute the corresponding coefficients of determination R_j^2. The value of R_j^2 from (2.106) tells us how well is X_j describable by the other variables. Hence a large value of R_j^2 indicates strong linear dependence of X_j on the other predictors. This then implies that there is multicollinearity of the predictors in the model. Instead of comparing each R_j^2, we define the variance inflation factors (VIF) of the predictors as

$$\text{VIF}(\widehat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors (except X_j). Notice that with VIF_j so defined,

- when X_j is nearly a combination of the other predictors:

$$R^2_{X_j|X_{-j}} \approx 1 \quad \text{hence} \quad \text{VIF}(\widehat{\beta}_j) \ \text{ is large}$$

- when X_j is orthogonal to all the other predictors:

$$R^2_{X_j|X_{-j}} = 0 \quad \text{hence} \quad \text{VIF}(\widehat{\beta}_j) = 1$$

As a summary, the smallest possible value for VIF is 1, which indicates the complete absence of collinearity. Typically in practice there is a small amount of collinearity among the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

In the Credit data, a regression of balance onto age, rating, and limit indicates that the predictors have VIF values of 1.01, 160.67, and 160.59. As we suspected, there is considerable collinearity in the data!

When faced with the problem of collinearity, there are two simple solutions. The first is to drop one of the problematic variables from the regression. This can usually be done without much compromise to the regression fit, since the presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables. For instance, if we regress balance onto age and limit, without the rating predictor, then the VIF values are close to the minimum value of 1, and the R^2 drops from 0.754 to 0.75. So dropping rating from the set of predictors has effectively solved the collinearity problem without compromising the fit. The second solution is to combine the collinearity variables together into a single predictor. For instance, we might take the average of standardized versions of limit and rating in order to create a new variable that measures *credit worthiness*.

2.4 The Marketing Plan

We now briefly return to the seven questions about the Advertising data that we set out to answer at the beginning of this chapter. See also the two summaries in Section 2.1.5 and 2.2.3.

1. *Is there a relationship between sales and advertising budget?*
   This question can be answered by fitting a multiple regression model of sales onto TV, radio, and newspaper, and testing the hypothesis $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$. In Section 2.2.2, we showed that the F-statistic can be used to determined whether or not we should reject this null hypothesis. In this case the p-value corresponding to the F-statistic is practically zero, indicating clear evidence of a relationship between advertising and sales. (See Fig 2.17.)

2. *How strong is the relationship?*
   We discussed two measures of model accuracy in Section 2.1.3. First, the RSE estimates the standard deviation of the response from the population regression line. For the Advertising data, the RSE is 1.69 units (See Fig 2.17) while the mean value for the response can be computed as 14.022, indicating a percentage error of roughly 12%. Since

$$\begin{aligned} \text{Percentage Error} &= \left(\frac{\text{RSE}}{\text{Mean Value}}\right) \times 100\% \\ &= \left(\frac{1.69}{14.022}\right) \times 100\% \\ &\approx 12.05\% \end{aligned}$$

   (Note: RSE measures the standard deviation of the model's predictions. It's often used as a measure of the accuracy of prediction relative to the mean of the response variable.)

   Second, the R^2 statistic records a percentage of variability in the response that is explained by the predictors. The predictors explain almost 90% of the variance in sales. The RSE and the R^2 statistics can be found in and Fig 2.16 and Fig 2.17. See also the discussion on Section 2.1.5.

3. *Which media are associated with sales?*
   To answer this question, we can examine the p-values associated with each predictor's t-statistic (Section 2.1.2). In the multiple linear regression displayed in Fig 2.13, the p-values for TV and radio are low, but the p-value for newspaper is not. This suggests that only TV and radio are related to sales. We will explore this question in greater detail when we discuss model selections.

4. *How large is the association between each medium and sales?*
   We saw in Section 2.1.2 that the standard error of $\widehat{\beta}_j$ can be used to construct construct confidence intervals for β_j. For the Advertising data, we can use the result in Fig 2.13 to compute the 95% confidence intervals for the coefficients in a multiple regression model using all three media budgets as predictors. The confidence intervals are as follows: $(0.043, 0.049)$ for TV, (0.172, 0.206) for radio, and (-0.013, 0.011) for newspaper. The confidence interval for TV and radio are narrow and far from zero, providing evidence that these media are related from sales. But the interval for newspaper includes zero, indicating that the variable is not statistically significant given the value of TV and radio.

   We saw in Section 2.3.3 that collinearity can result in very wide standard errors. Could collinearity be the reason that the confidence interval associated with newspaper is so

wide? The VIF scores are 1.005, 1.145, and 1.145 for TV, radio, and newspaper, suggesting no evidence of collinearity.

In order to assess the association of each medium individually on sales, we can perform three seperate simple linear regressions. Results are shown in Fig 2.7 and Fig 2.11. There is evidence of an extremely strong association between TV and sales and between radio and sales. There is evidence of a mild association between newspaper and sales, when the values of TV and radio are ignored. See also Section 2.2.3.

5. *How accurately can we predict future sales?*
The response can be predicted using

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p.$$

The accuracy associated with this estimate depends on whether we wish to predict an individual response, $Y = f(X) + \epsilon$, or the average response, $f(X)$ (Section 2.2.2). If the former, we use a prediction interval, and if the latter, we use a confidence interval. Prediction intervals will always be wider than confidence intervals because they account for uncertainly associated with ϵ, the irreducible error. See also Section 2.1.5 and Section 2.2.3.

6. *Is the relationship linear?*
In Section 2.3.3, we saw that the residual plots can be used in order to identify non-linearity. If the relationships are linear, then the residual plots should display no pattern. In the case of the Advertising data, we observe a non-linear effect in Fig 2.20, though this effect could also be observed in a residual plot. In Section 2.3.2, we discussed the inclusion of transformations of the predictors in the linear regression model in order to accommodate non-linear relationships.

7. *Is there synergy among the advertising media?*
The standard linear regression model assumes an additive relationship between the predictors and the response. An additive model is easy to interpret because the association between each predictor and the response is unrelated to the values of the other predictors. However, the additive assumption may be unrealistic for certain data sets. In Section 2.3.2, we showed how to include an interaction term in the regression model in order to accommodate non-additive relationships. A small p-value associated with the interaction term indicates the presence of such relationships. Figure 2.20 suggested that the Advertising data may not be additive. Including an interation term in the model results in a substantial increase in R^2, from around 90% to almost 97%.

2.5 Comparison of Linear Regression with K-Nearest Neighbors

As discussed in Chapter 1, linear regression is an example of a *parametric* approach because it assumes a linear functional form for $f(X)$. Parametric methods have several advantages. They are often easy to fit, because one need estimate only a small number of coefficients. In the case of linear regression, the coefficients have simple interpretations, and test of statistical significance can be easily performed. But parametric methods do have a disadvantage: by construction, they make strong assumptions about the form of $f(X)$. If the specified functional form is far from the truth, and prediction accuracy is our goal, then the parametric method will perform poorly. For instance, if we assume a linear relationship between X and Y but the

true relationship is far from linear, then the resulting model will provide a poor fit to the data, and any conclusion drawn from it will be suspect.

In contrast, *non-parametric* methods do not explicitly assume a parametric form for $f(X)$, and thereby provide an alternative and more flexible approach for performing regression. We discuss various non-parametric method in this course. Here we consider one of the simplest and best-known non-parametric methods, *K-nearest neighbors regression* (KNN regression). The KNN regression method is closely related to the KNN classifier discussed in Chapter 1. Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$. In other words,

$$\widehat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

Figure 2.46 illustrates two KNN fits on a data set with $p = 2$ predictors. The fit with $K = 1$ is shown in the left-hand panel, while the right-hand panel corresponds to $K = 9$. We see that when $K = 1$, the KNN fit perfectly interpolates the training observations, and consequently takes the form of a step function. When $K = 9$, the KNN fit still is a step function, but averaging over nine observations results in much smaller regions of constant prediction, and consequently a smoother fit.

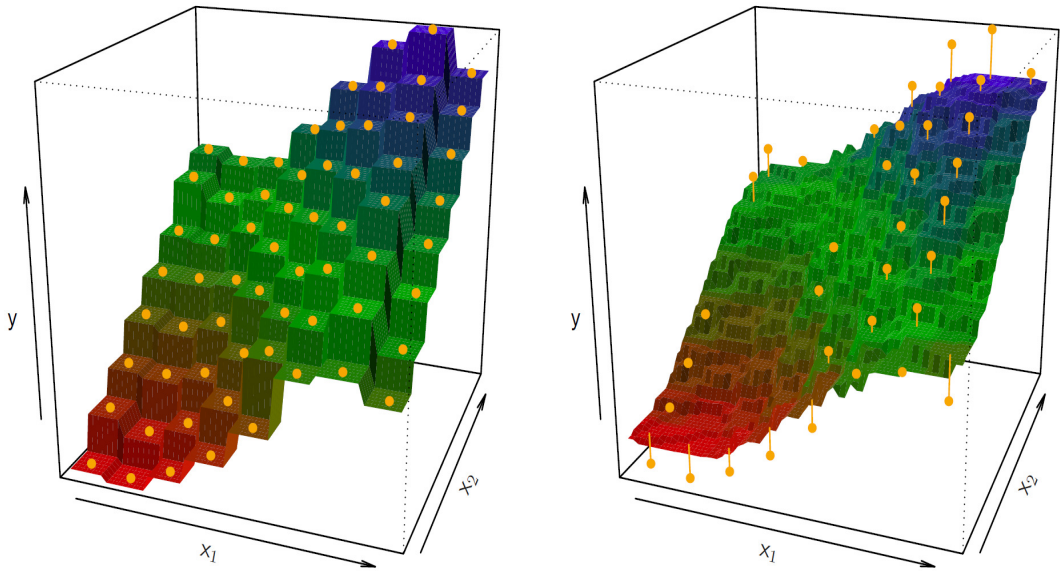


Figure 2.46: Plot of $\widehat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots). Left: $K = 1$ results in a rough step function fit. Right: $K = 9$ produces a much smoother fit.

In general, the optimal value for K will depend on the *bias-variance tradeoff*, which we introduced in Chapter 1. A small value of K provides the most flexible fit, which will have low bias but high variance. This variance is due to the fact that the prediction in a given region is entirely depend on just one observation. In contrast, a larger values of K provide a smoother and less variable fit; the prediction in a region is an average of several points, and changing one observation has a smaller effect. However, the smoothing may cause bias by masking some of the structure in $f(X)$. Later we introduce several approaches for estimating test error rates. These methods can be used to identify the optimal value of K in KNN regression.

In what setting will a parametric approach such as least squares linear regression outperform a non-parametric approach such as KNN regression? The answer is simple: *the parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of* f. Figure 2.47 provides an example with data generated from a one-dimensional linear regression model. The black solid lines represent $f(X)$, while the blue curves correspond to the KNN fit using $K = 1$ and $K = 9$. In this case, the $K = 1$ predictions are too variable, while the smoother $K = 9$ fit is much closer to $f(X)$. However, since the true relationship is linear, it is hard for a non-parametric approach to compete with linear regression: a non-parametric approach incurs a cost in variance that is not offset by a reduction in bias.

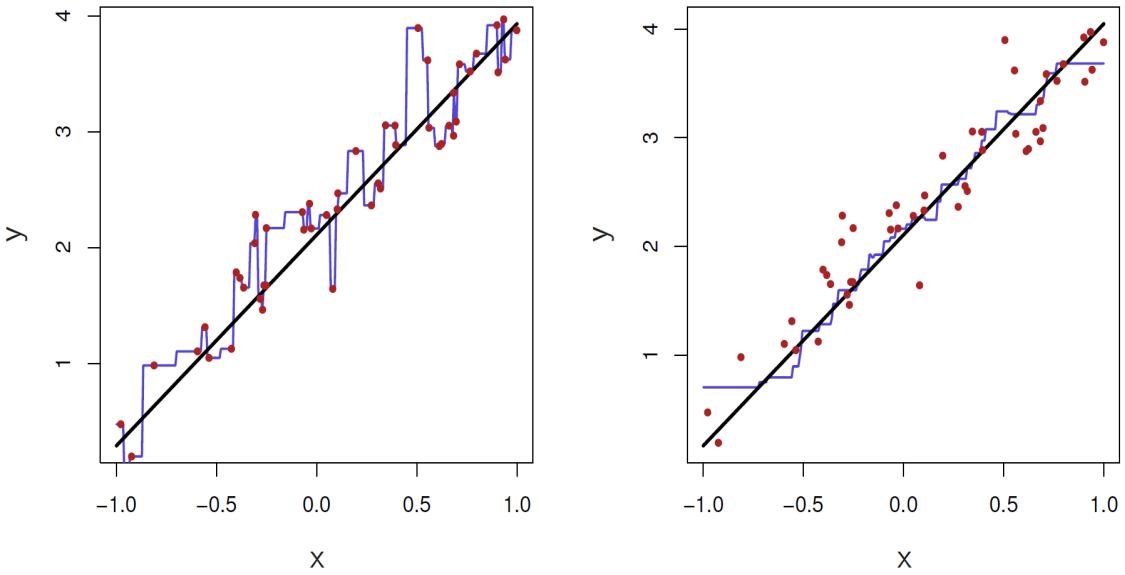


Figure 2.47: Plots of $\widehat{f}(X)$ using KNN regression on a one-dimensional data set with 50 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly through) the training data. Right: The blue curve corresponds to $K = 9$, and represents a smoother fit.

The blue dashed line in the left-hand panel of Figure 2.48 represents the linear regression fit to the same data. It is almost perfect. The right-hand panel of Figure 2.48 reveals that linear regression outperforms KNN for this data. The green solid line, plotted as a function of $1/K$, represents the test set mean squared error (MSE) for KNN. The KNN errors are well above the black dashed line, which is the test MSE for linear regression. When the value of K is large, then KNN performs only a little worst than least squares regression in terms of MSE. It performs far worse when K is small.

In practice, the true relationship between X and Y is rarely exactly linear. Figure 2.49 examines the relative performances of least squares regression and KNN under increasing levels of non-linearity in the relationship between X and Y. In the top row, the true relationship is nearly linear. In this case we see that the test MSE for linear regression is still superior to that of KNN for low values of K. However, for $K \geq 4$, KNN out-performs linear regression. The second row illustrates a more substantial deviation from linearity. In this situation, KNN substantially outperforms linear regression for all values of K. Note that as the extend of non-linearity increases, there is little change in the test set MSE for the non-parametric KNN method, but there is a large increase in the test MSE of linear regression.

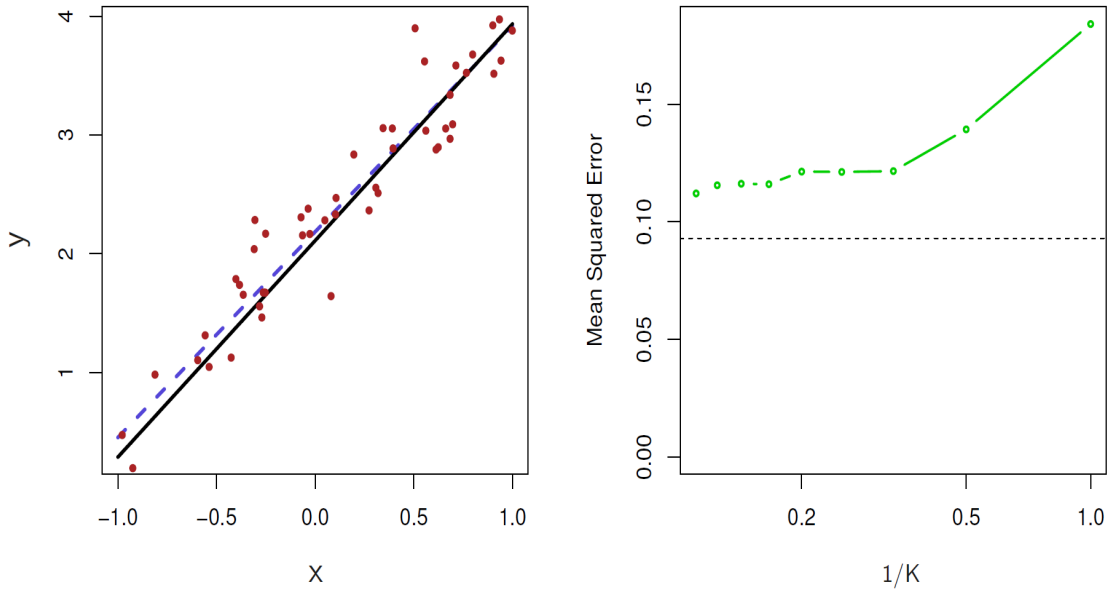


Figure 2.48: The same data set shown in Figure 2.47 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of $f(X)$. Right: The dashed horizontal line represents the least squares test set MSE while the green solid line corresponds to the MSE for KNN as a function of $1/K$ (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of $1/K$.

Figures 2.48 and 2.49 display situations in which KNN performs slightly worst than linear regression when the relationship is linear, but much better than linear regression for non-linear situations. In a real life situation in which the true relationship is unknown, one might suspect that KNN should be favored over linear regression because it will at worst be slightly inferior to linear regression if the true relationship is linear, and may give substantially better results if the true relationship is non-linear. But in reality, even when the true relationship is highly non-linear, KNN may still provide inferior results to linear regression. In particular, both Figures 2.48 and 2.49 illustrate setting with $p = 1$ predictor. But in higher dimensions, KNN often performs worse than linear regression.

Figure 2.50 considers the same strongly non-linear situation as in the second row of Figure 2.49, except that we have added additional *noise* predictors that are not associated with the response. When $p = 1$ or $p = 2$, KNN outperforms linear regression. But for $p = 3$ the results are mixed, and for $p \geq 4$ linear regression is superior to KNN. In fact, the increase in dimension has only caused a small deterioration in the linear regression test set MSE, but it has caused more than a ten-fold increase in the MSE for KNN. This decrease in performance as the dimension increases is common problem for KNN, and results from the fact that in higher dimensions there is effectively a reduction in sample size. In this data set there are 50 training observations; when $p = 1$, this provides enough information to accurately estimate $f(X)$. However, spreading 50 observations over $p = 20$ dimensions results in a phenomenon in which a given observation has no *nearby neighbors*—this is the so called *curse of dimensionality*. That is, the K observations that are nearest to a given test observation x_0 may be very far away from x_0 in p-dimensional space when p is large, leading to a very poor prediction of $f(x_0)$ and hence a poor KNN fit. As a general rule, parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor.

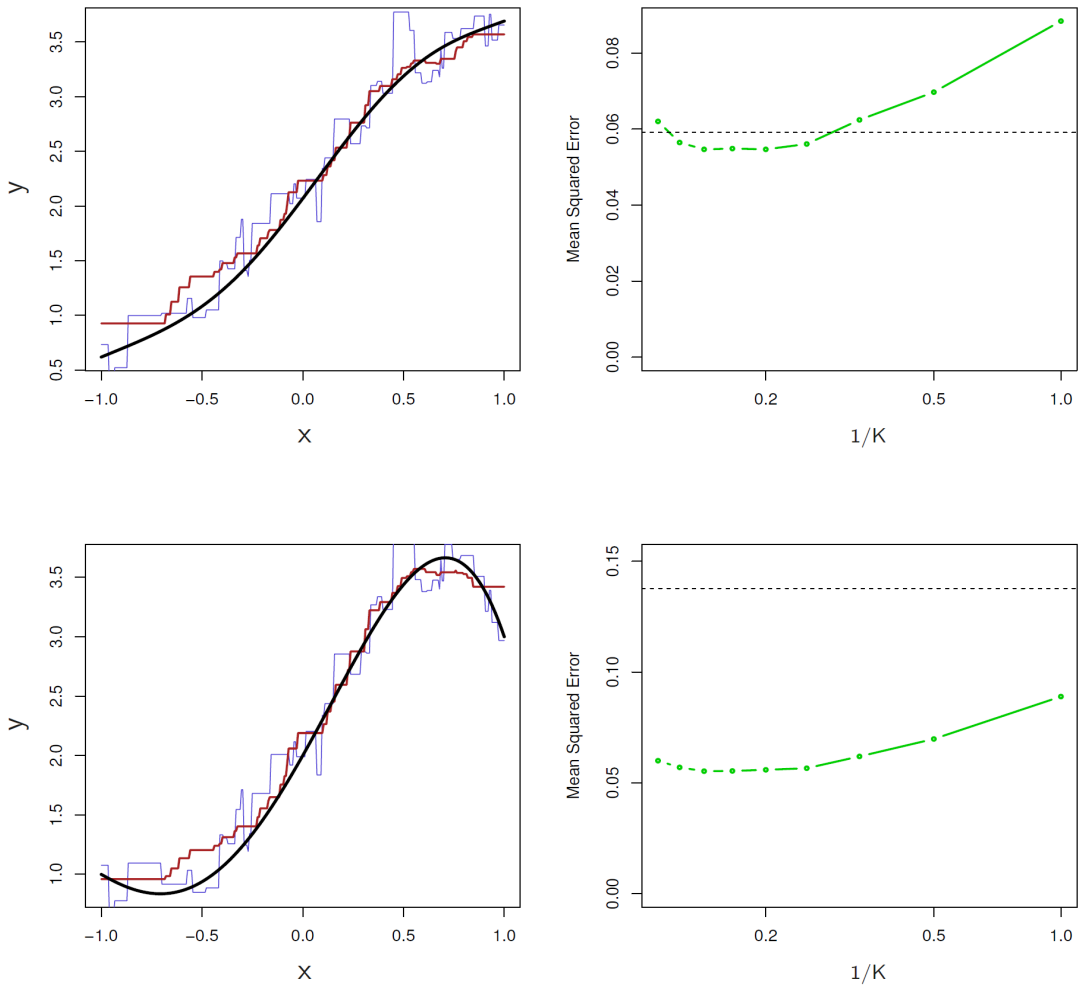


Figure 2.49: Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed. Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of $1/K$ (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y.

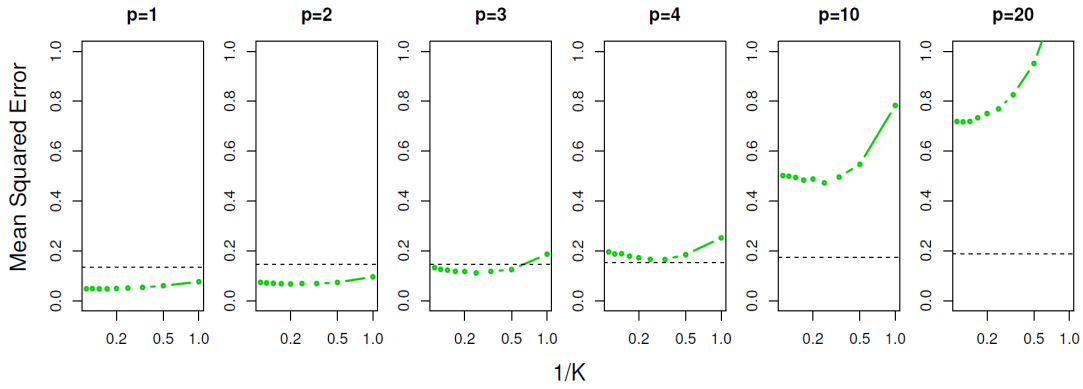


Figure 2.50: Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in Fig 2.49, and does not depend on additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.

Even when the dimension is small, we might prefer linear regression to KNN from an interpretability standpoint. If the test MSE of KNN is only slightly lower than that of linear regression, we might be willing to forego a little bit of prediction accuracy for the sake of a simple model that can be described in terms of just a few coefficients, and for which p-values are available.

2.6 References

1. Chatterjee, Samprit, and Ali S. Hadi. Regression Analysis by Example. 5th ed. Hoboken, NJ: John Wiley & Sons, 2013.

2. Fox, John. Applied Regression Analysis and Generalized Linear Models. 3rd ed., SAGE Publications, Inc., 2015.

3. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning. 2nd ed. Springer Texts in Statistics. New York, Springer, 2021.

4. Kutner, M., Nachtsheim, C., Neter, J., & Li, W. Applied Linear Statistical Models. 5th ed., McGraw-Hill, 2013.

5. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to Linear Regression Analysis. 5th ed. Hoboken, NJ: John Wiley & Sons, 2012.

Chapter 3

Classification

The linear regression model discussed in Chapter 2 assumes that the response variable Y is quantitative. But in many situations, the response variable is instead *qualitative*. For example, eye color is qualitative. Often qualitative variables are referred to as *categorical*; we will use these terms interchangeably. In this chapter, we study approaches for predicting qualitative responses, a process that is known as *classification*. Predicting a qualitative response for an observation can be referred to as *classifying* that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability that the observation belongs to each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.

There are many possible classification techniques, or *classifiers*, that one might use to predict a qualitative response. In this Chapter we discuss some widely-used classifiers: *logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes, and K-nearest neighbors*. The discussion of logistic regression is used as a jumping-off point for a discussion of *generalized linear models*, and in particular, *Poisson regression*. We discuss more computer-intensive classification methods in later chapters: these include generalized additive models; trees, random forests, and boosting; and support vector machines.

3.1 An Overview of Classification

Classification problems occur often, perhaps even more so than regression problems. Some examples include:

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

3. On the basis of DNA sequence data for a number of patients with and without given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Just as in the regression setting, in the classification setting we have a set of training observations $(x_1, y_1), \cdots, (x_n, y_n)$ that we can use to build a classifier. We want our classifier to

perform well not only on the training data, but also on test observations that were not used to train the class

In this chapter, we will illustrate the concept if classification using the simulated Default data set. We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance. The data is displayed in Figure 3.1. In the left-hand panel of Figure 3.1, we have plotted annual income and monthly credit card balance for a subset of 10,000 individuals. The individuals who defaulted in a given month are shown in orange, and those who did not in blue. (The overall defacult rate is about 3%, so we have plotted only a fraction of the individuals who did not default.) It appears that individuals who defaulted tended to have higher credit card balances than those who did not. In the center and right-hand panels of Figure 3.1, two pairs of bloxplots are shown. The first shows the distribution of balance split by the binary default variable; the second is a similar plot for income. In this chapter, we learn how to build a model to predict default (Y) for any given value of balance (X_1) and income (X_2). Since Y is not quantitative, the simple linear regression model of Chapter 2 is not a good choice: we will elaborate this further in Section 3.2.

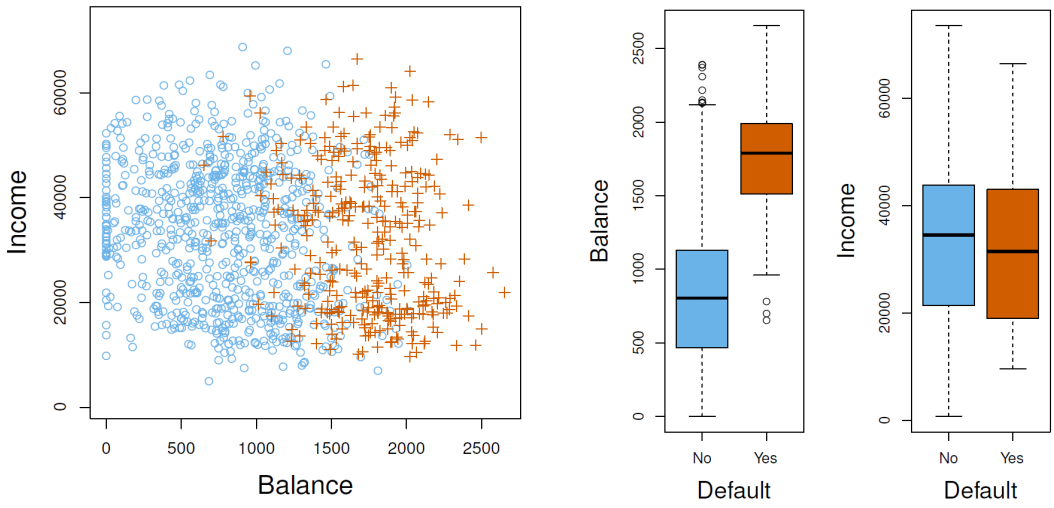


Figure 3.1: The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

It is worth nothing that Figure 3.1 displays a very pronounced relationship between the predictor balance and the response default. In most real applications, the relationship between the predictor and the response will not be nearly so strong. However, for the sake of illustrating the classification procedures discussed in this chapter, we use an example in which the relationship between the predictor and the response is somewhat exaggerated.

3.2 Why Not Linear Regression?

We have stated that linear regression is not appropriate in the case of a quantitative response. Why not?

Suppose that we are trying to predict the medical condition of a patient in the emergency

room on the basis of her symptoms. In this simplified example, there are three possible diagnoses: stroke, drug overdose, and epileptic seizure. We could consider encoding these value as a quantitative response variable, Y, as follows:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors $X_1, \cdots, X_p$. Unfortunately, this coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure. In practice there is no particular reason that this needs to be the case. For instance, one could choose an equally reasonable coding,

$$Y = \begin{cases} 1 & \text{if epileptic seizure} \\ 2 & \text{if stroke} \\ 3 & \text{if drug overdose} \end{cases}$$

which would imply a totally different relationship among the three conditions. Each of these codings would produce fundamentally different linear models that would ultimately lead to different sets of predictions on test observations.

If the response variable's values did take on a natural ordering, such as *mild, moderate*, and *severe*, and we felt the gap between mild and moderate was similar to the gap between moderate and severe, then a $1, 2, 3$ coding would be reasonable. Unfortunately, in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

For a *binary* (two level) qualitative response, the situation is better. For instance, perhaps there are only two possibilities for the patient's medical condition: stroke and drug overdose. We could then potentially use the *dummy variable* approach from Section 2.3.1 to code the response as follows:

$$Y = \begin{cases} 0 & \text{if stroke} \\ 1 & \text{if drug overdose.} \end{cases}$$

We could then fit a linear regression to this binary response, and predict drug overdose if $\widehat{Y} > 0.5$ and stroke otherwise. In the binary case it is not hard to show that even if we flip the above coding, linear regression will produce the same final predictions.

For a binary response with a 0/1 coding as above, regression by least squares is not completely unreasonable: it can be shown that the $X\widehat{\beta}$ obtained using linear regression is in fact an estimate of $\Pr(\text{drug overdose}|X)$ in this special case. However, if we use linear regression, some of our estimates might be outside the $[0, 1]$ interval (see Figure 3.2), making them hard to interpret as probabilities! Nevertheless, the predictors provide an ordering and can be interpreted as crude probability estimates. Curiously, it turns out that the classifications that we get if we use linear regression to predict a binary response will be the same as for the linear discriminant analysis (LDA) procedure we discuss in Section 3.4.

To summarize, there are at least two reasons not to perform classification using a regression method:

(a) a regression method cannot accommodate a qualitative response with more than two classes;

(b) a regression method will not provide meaningful estimates of $\Pr(Y|X)$, even with just two classes.

Thus, it is preferable to use classification method that is truly suited for qualitative response values. In the next section, we present logistic regression, which is well-suited for the case of a binary qualitative response; in later sections we will cover classification methods that are appropriate when the qualitative response has two or more classes.

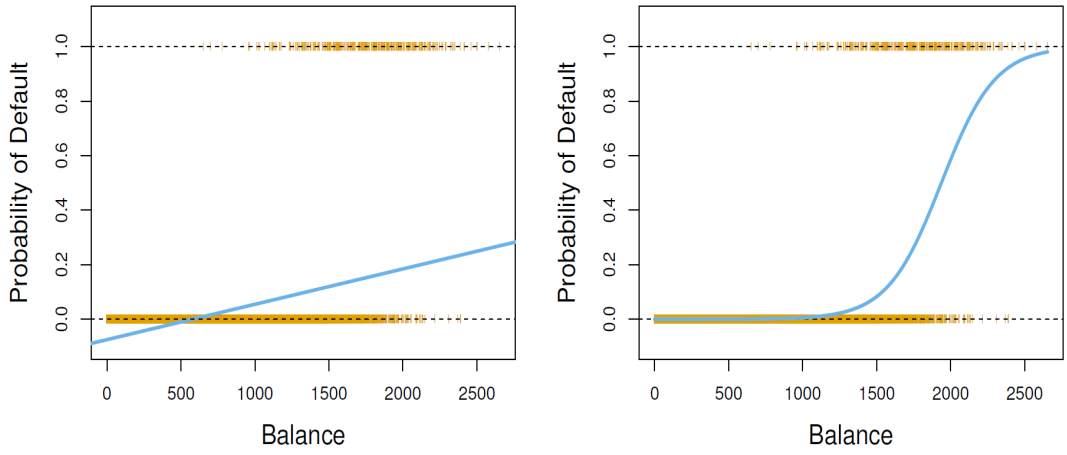


Figure 3.2: Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default (No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

3.3 Logistic Regression

Consider again the Default data set, where the response default falls into one of two categories, Yes or No. Rather than modeling this response Y directly, logistic regression models the *probability* that Y belongs to a particular category.

For the Default data, logistic regression models the probability of default. For example, the probability of default given balance can be written as

$$\Pr(\text{default} = \text{Yes} \mid \text{balance}).$$

The values of $\Pr(\text{default} = \text{Yes} \mid \text{balance})$, which we abbreviate $p(\text{balance})$, will range between 0 and 1. Then for any given value of balance, a prediction can be made for default. For example, one might predict default = Yes for any individual for whom $p(\text{balance}) > 0.5$. Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$.

3.3.1 The Logistic Model

How should we model the relationship between $p(X) = \Pr(Y = 1 \mid X)$ and X? (For convenience we are using the generic 0/1 coding for the response.) In Section 3.2 we considered using a linear regression model to represent these probabilities:

$$p(X) = \beta_0 + \beta_1 X. \tag{3.1}$$

If we use this approach to predict default = Yes using balance, then we obtain the model shown in the left-hand panel of Figure 3.2. Here we see the problem with this approach: for balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1. These predictions are not sensible, since of course the true probability of default, regardless of credit card balance, must fall between 0 and 1. This problem is not unique to the credit default data. Any straight line fit to a binary response that is codded as 0 or 1, in principle we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others (unless the range of X is limited).

To avoid this problem, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X. Many functions meet this description. In logistic regression, we use the *logistic function*,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{3.2}$$

To fit the model (3.2), we use a method called *maximum likelihood*, which we discuss in the next section. The right-hand panel of Figure 3.2 illustrates the fit of the logistic regression model to the Default data. Notice that for low balances we now predict the probability of default as close to , but never below, zero. Likewise, for high balances, we predict a default probability close to, but never above, one. The logistic function will always produce a S-shaped curve of this form, and so regardless of the value of X, we will obtain a sensible prediction. We also see that the logistic model is better able to capture the range of probabilities than is the linear regression model in the left-hand plot. The average fitted probability in both cases is 0.0333 (averaged over the training data), which is the same as the overall proportion of defaulters in the data set.

After a bit of manipulation of (3.2), we find that

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}. \tag{3.3}$$

The quantity $p(X)/[1 - p(X)]$ is called the *odds*, and can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicates very low and very high probabilities of default, respectively.

For example, on average 1 in 5 people with an odds of 1/4 will default, since $p(X) = 0.2$ implies an odds of

$$\frac{0.2}{1 - 0.2} = \frac{1}{4}.$$

Likewise, on average nine out of every ten people with an odds of 9 will default, since $p(X) = 0.9$ implies an odds of

$$\frac{0.9}{1 - 0.9} = 9.$$

Odds are traditionally used instead of probabilities in horse-racing, since they related more naturally to the correct betting strategy.

By taking the logarithm on both side of (3.3), we arrive at

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X. \tag{3.4}$$

The left-hand side is called the *log odds* or *logit*. We see that the logistic regression model (3.2) has a logit that is linear in X.

Recall from Chapter 2 that in a linear regression model, β_1 gives the average change in Y associated with a one-unit increase in X. By contrast, in a logistic regression model, increasing X by one unit changes the log odds by β_1 (3.4). Equivalently, it multiplies the odds by e^{β_1} (3.3).

To see this, let

$$\log\left(\frac{p(X)}{1-p(X)}\right)(x) = \beta_0 + \beta_1 x$$

suppose we have $x = x_0$ then

$$\log\left(\frac{p(X)}{1-p(X)}\right)(x_0) = \beta_0 + \beta_1 x_0$$

$$\log\left(\frac{p(X)}{1-p(X)}\right)(x_0+1) = \beta_0 + \beta_1(x_0+1)$$

Subtracting these two equations we obtain

$$\log\left(\frac{p(X)}{1-p(X)}\right)(x_0+1) - \log\left(\frac{p(X)}{1-p(X)}\right)(x_0) = \beta_1.$$

Hence increasing X by one unit changes the log odds by β_1. Let's write the equation above as

$$\log\left(\text{Odds}\right)(x_0+1) - \log\left(\text{Odds}\right)(x_0) = \beta_1$$

hence we see that

$$\log\left(\frac{\text{Odds}(x_0+1)}{\text{Odds }(x_0)}\right) = \beta_1$$

or equivalently

$$\frac{\text{Odds}(x_0+1)}{\text{Odds }(x_0)} = e^{\beta_1}$$

and

$$\text{Odds}(x_0+1) = e^{\beta_1}\text{Odds}(x_0).$$

That is, increasing X by one unit, it multiplies the odds by e^{β_1}. However, because the relationship between $p(X)$ and X in (3.2) is not a straight line, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in X. The amount that $p(X)$ changes due to a one-unit change in X depends on the current value of X. But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$. The fact that there is not a straight-line relationship between $p(X)$ and X, and the fact that the rate of change in $p(X)$ per unit change in X depends on the current value of X, can also be seen by inspection of the right-hand panel of Figure 3.2. See also Figure 3.3.

3.3.2 Estimating the Regression Coefficients

Before we discuss how the regression coefficients are obtained, let us review the concepts of likelihood function and maximum likelihood estimation. The likelihood function is a fundamental concept in statistical inference, particularly in parameter estimation. Given a statistical model and observed data, the likelihood function quantifies the plausibility of different parameter values that could have generated the observed data.

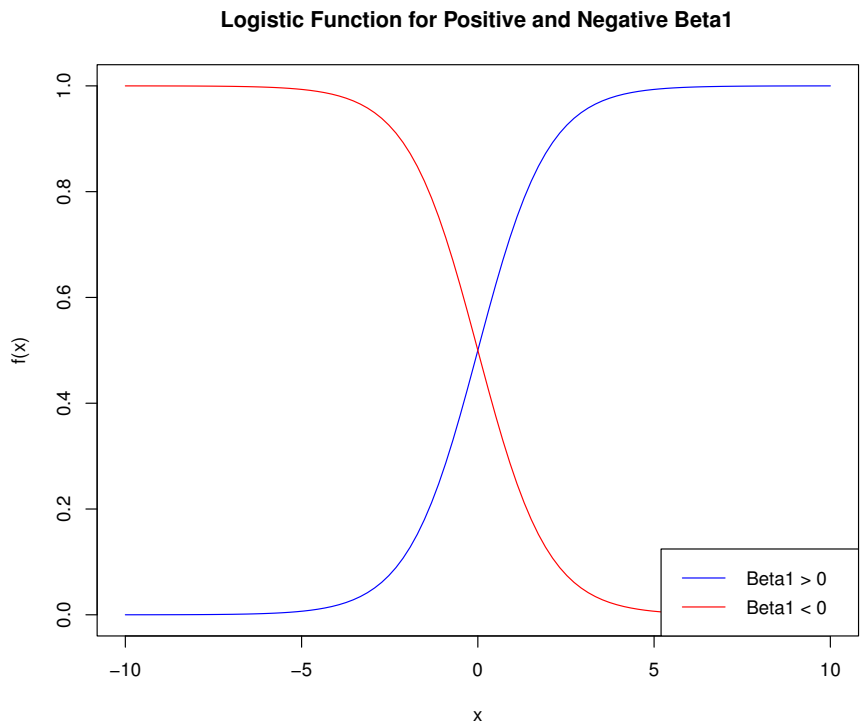


Figure 3.3: The graphs of $\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ for different signs of β_1.

Formally, let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a set of independent and identically distributed, IID, random variables with a probability density function $f(\mathbf{X} \mid \theta)$, where θ is a parameter (or a vector of parameters) of the model. The likelihood function, $L(\theta \mid \mathbf{X})$, is given by:

$$L(\theta \mid \mathbf{X}) = f(\mathbf{X} \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

Here, $f(X_i \mid \theta)$ is the pdf (or probability mass function for discrete variables) of the i-th observation given the parameter θ.

Remarks:

- Likelihood vs. Probability: The likelihood function $L(\theta \mid \mathbf{X})$ is different from the probability $P(\mathbf{X} \mid \theta)$. While probability is a measure of the chance of observing the data given a fixed parameter θ, likelihood is a measure of the plausibility of the parameter θ given the observed data.

- Relative Measure: The likelihood function is typically used to compare different parameter values. A higher likelihood value indicates a parameter value that makes the observed data more plausible.

- Not a Probability: The likelihood function is not a probability distribution over the parameter space. It does not integrate (or sum) to one over all possible parameter values.

The likelihood function is a key component in several parameter estimation methods: The most common use of the likelihood function is in *Maximum Likelihood Estimation* (MLE). The goal of MLE is to find the parameter value(s) $\hat{\theta}$ that maximize the likelihood function. Formally:

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid \mathbf{X})$$

Since the logarithm is a monotonically increasing function, the maximization is often performed on the log-likelihood function, $\ell(\theta \mid \mathbf{X})$:

$$\ell(\theta \mid \mathbf{X}) = \log L(\theta \mid \mathbf{X}) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

We conclude the review with the following example: Consider a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a normal distribution with unknown mean μ and known variance σ^2. The pdf is:

$$f(X_i \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

The likelihood function is:

$$L(\mu \mid \mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

Taking the log-likelihood:

$$\ell(\mu \mid \mathbf{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Using Calculus, maximizing $\ell(\mu \mid \mathbf{X})$ with respect to μ gives:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is the sample mean, which is the MLE for the mean of a normal distribution with known variance.

Returning to our logistic regression. The coefficients β_0 and β_1 in (3.2) are unknown, and must be estimated based on the available training data. In Chapter 2, we used the least squares approach to estimate the unknown linear regression coefficients. Although we could use (non-linear) least squares to fit the model (3.4), the more general method of *maximum likelihood* is preferred, since it has better statistical properties.

The basic intuition behind using the maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β_0 and β_1 such that the predicted probability $\widehat{p}(x_i)$ of default for each individual, using (3.2), corresponds as closely as possible to the individual's observed default status. In other words, we try to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that plugging these estimates into the model for $p(X)$, given in (3.2), yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not. This intuition can be formalized using the *likelihood function* mentioned earlier:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{3.5}$$

The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen to maximize this likelihood function.

Maximum likelihood is a very general approach that is used to fit many of the non-linear

models. In the linear regression setting, the least squares approach is in fact a special case of maximum likelihood (Exercise). The maximum likelihood method can be easily applied using software such as R, so we do not need to concern ourselves with the details of the maximum likelihood fitting procedure here.

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

Figure 3.4: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance is shown. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

Figure 3.4 shows the coefficient estimates and related information that result from fitting a logistic regression model on the Default data in order to predict the probability of default=Yes using balance. We see that $\widehat{\beta}_1 = 0.0055$; this indicates that an increase in balance is associated with an increase in the probability of default. To be precise, a one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

Many aspects of the logistic regression output shown in Figure 3.4 are similar to the linear regression output from Chapter 2. For example, we can measure the accuracy of the coefficient estimates by computing their standard errors. The z-statistic in Figure 3.4 plays the same role as the t-statistic in the linear regression output, for example in Figure 2.7. For instance, the z-statistic associated with β_1 is equal to $\widehat{\beta}_1/\text{SE}(\widehat{\beta}_1)$, and so a large (absolute) value of the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$. This null hypothesis implies that

$$p(X) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} :$$

in other words, the probability of default does not depends on balance. Since the p-value associated with balance in Figure 3.4 is tiny, we can reject H_0. In other words, we conclude that there is indeed an association between balance and the probability of default. The estimates intercept in Figure 3.4 is typically not of interest; its main purpose is to adjust the average fitted probabilities to the proportion of ones in the data (in this case, the overall default rate).

3.3.3 Making Predictions

Once the coefficients have been estimated, we can compute the probability of default for any given credit card balance. For example, using the coefficient estimates given in Figure 3.4, we predict that the default probability of an individual with a balance of $1,000 is

$$\widehat{p}(X) = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X}}{1 + e^{\widehat{\beta}_0+\widehat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1 + e^{-10.6513+0.0055\times 1{,}000}} = 0.00576,$$

which is below 1%. In contrast, the predicted probability of default for an individual with a balance of $2,000 is much higher, or equals 0.586 or 58.6%.

One can use qualitative predictors with the logistic regression model using the dummy variable approach from Section 2.3.1. As an example, the Default data set contains the qualitative variable student. To fit a model that uses student status as a predictor variable, we simply create

a dummy variable that takes on a value of 1 for students and 0 for non-students. The logistic regression model that results from predicting probability of default from student status can be seen in Figure 3.5. The coefficient associated with the dummy variable is positive, and the associated p-value is statistically significant. This indicates that students tend to have higher default probabilities than non-students:

$$\widehat{\Pr}(\text{default=Yes}|\text{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$
$$\widehat{\Pr}(\text{default=Yes}|\text{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Figure 3.5: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable Student[Yes] in the table.

3.3.4 Multiple Logistic Regression

We now consider the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression in Chapter 2, we can generalize (3.4) as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{3.6}$$

where $X = (X_1, \cdots, X_p)$ are p predictors. Equation 3.6 can be written as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{3.7}$$

Just as in Section 3.3.2, we use the maximum likelihood method to estimate $\beta_0, \beta_1, \cdots, \beta_p$.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

Figure 3.6: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.

Figure 3.6 shows the coefficient estimates for a logistic regression model that uses balance, income (in thousands of dollars), and student status to predict probability of default. There

is a surprising result here. The p-values associated with balance and the dummy variable for student status are very small, indicating that each of these variables is associated with the probability of default. However, the coefficient for the dummy variable is negative, indicating that students are less likely to default than non-students. In contrast, the coefficient for the dummy variable is positive in Figure 3.5. How is it possible for student status to be associated with an increase in probability of default in Figure 3.5 and a decrease in probability of default in Figure 3.6?

The left-hand panel of Figure 3.7 provides a graphical illustration of this apparent paradox. The orange and blue solid lines show the average default rates for students and non-students, respectively, as a function of credit card balance. The negative coefficient for student in the multiple logistic regression indicates that for a fixed value of balance and income, a student is less likely to default than a non-student. Indeed, we observe from the left-hand panel of Figure 3.7 that the student default rate is at or below that of the non-student default rate for every value of balance. But the horizontal broken lines near the base of the plot, which show the default rates for students and non-students average over all values of balance and income (see the remark next page), suggest the opposite effect: the overall student default rate is higher than the non-student default rate. Consequently, there is a positive coefficient for student in the single variable logistic regression output shown in Figure 3.5.

The right-hand panel of Figure 3.7 provides an explanation for this discrepancy. The variables student and balance are correlated. Students tend to hold higher level of debt, which in turn associated with higher probability of default. In other words, students are more likely to have large credit card balances, which, as we know from the left-hand panel of Figure 3.7, tend to be associated with high default rates. Thus, even though an individual student with a given credit card balance will tend to have a lower probability of default than a non-student with the same credit card balance, the fact that students on the whole tend to have higher credit card balances means that overall, students tend to default at a higher rate than non-students. This is an important distinction for a credit card company that is trying to determine to whom they should offer credit. A student is riskier than a non-student if no information about the student's credit card balance is available. However, that student is less risky than a non-student with the same credit card balance!

This simple example illustrate the dangers and subtleties associated with performing regressions involving only a single predictor when other predictors may also be relevant. As in the linear regression setting, the results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors. In general, the phenomenon seen in Figure 3.7 is known as *confounding*.

By substituting estimates for the regression coefficients from Figure 3.6 into (3.7), we can make predictions. For example, a student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default of

$$\widehat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058. \tag{3.8}$$

A non-student with the same balance and income has an estimated probability of default of

$$\widehat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105. \tag{3.9}$$

(Here we multiple the income coefficient estimate from Figure 3.6 by 40, rather than by 40,000, because in that table the model was fit with income measures in units of \$1,000.)

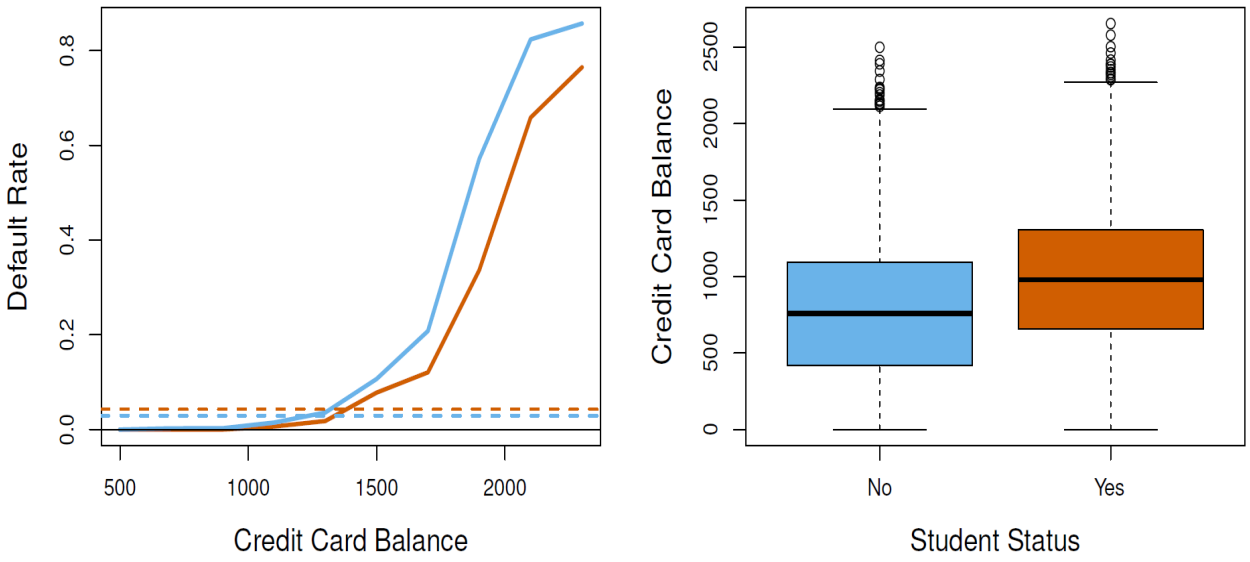


Figure 3.7: Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines displays default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

Remark: The phrase "average over all values of balance and income" means that the default rates for students and non-students are calculated by considering the entire range of balance and income values in the dataset, rather than at specific levels of these variables. The example below shows how this is done.

Example:

Assume you have a dataset with the following structure:

ID	Balance	Income	Student	Default
1	1000	30000	Yes	No
2	2000	40000	No	Yes
3	1500	35000	Yes	No
4	1200	32000	No	No
5	3000	45000	Yes	Yes
6	2200	38000	No	Yes
7	1800	37000	Yes	No
8	1300	34000	No	No

Calculating for Students:

Separate student data:

ID	Balance	Income	Default
1	1000	30000	No
3	1500	35000	No
5	3000	45000	Yes
7	1800	37000	No

then

$$\text{Default rate (students)} = \frac{\text{Number of defaults}}{\text{Total number of students}} = \frac{1}{4} = 0.25$$

Calculating for Non-Students:

Separate non-student data:

ID	Balance	Income	Default
2	2000	40000	Yes
4	1200	32000	No
6	2200	38000	Yes
8	1300	34000	No

then

$$\text{Default rate (non-students)} = \frac{\text{Number of defaults}}{\text{Total number of non-students}} = \frac{2}{4} = 0.50$$

These default rates are computed by averaging over all values of balance and income, meaning that we consider the entire dataset for each group without focusing on specific balance or income values.

3.3.5 Multinomial Logistic Regression

We sometimes wish to classify a response variable that has more than two classes. For example, in Section 3.2 we had three categories of medical condition in the emergency room: stroke, drug overdose, epileptic seizure. However, the logistic regression approach that we have seen in this section only allows for $K = 2$ classes for the response variable.

It turns out that it is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes. This extension is sometimes known as *multinomial logistic regression*. To do this, we first select a single class to serve as the *baseline*; without loss of generality, we select the Kth class for this role. Then we replace the model (3.7) with the model

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{3.10}$$

for $k = 1, \cdots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{3.11}$$

It is not hard to show that for $k = 1, \cdots, K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p. \tag{3.12}$$

Notice that (3.12) is quite similar to (3.6). Equation 3.12 indicates that once again, the log odds between any pair of classes is linear in the features.

It turns out that in (3.10)-(3.12), the decision to treat the Kth class as the baseline is unimportant. For example, when classifying emergency room visit into stroke, drug overdose and epileptic seizure, suppose that we fit two multinomial logistic regression models: one treating stroke as the baseline, another treating drug overdose as the baseline. The coefficient estimates will differ between the two fitted models due to the differing choice of the baseline, but the fitted values (predictions), the log odds between any pair of classes, and other key model outputs will

remain the same.

Nonetheless, interpretation of the coefficients in a multinomial logistic regression model must be done with care, since it is tied to the choice of baseline. For example, if we set epileptic seizure to be the baseline, then we can interpret $\beta_{stroke0}$ as the log odds of stroke versus epileptic seizure, given that $x_1 = \cdots = x_p = 0$. Furthermore, a one-unit increase in X_j is associated with a $\beta_{stroke}j$ increase in the log odds of stroke over epileptic seizure. Stated another way, if X_j increases by one unit, then

$$\frac{\Pr(Y = \text{stroke}|X = x)}{\Pr(Y = \text{epileptic seizure}|X = x)}$$

increases by a factor of $e^{\beta_{stroke}j}$.

Let us return to another example: Suppose we have a dataset where we want to predict the type of vehicle a person owns based on their annual income and age. The outcome variable Y has 3 categories:

1. Sedan (Category 1)
2. SUV (Category 2)
3. Truck (Category 3, which we will use as the reference category)

Data: Let's say we have the following data for a person:

- Annual income (x_1) = \$50,000
- Age (x_2) = 30 years

We have estimated the following coefficients from our multinomial logistic regression model:

For Sedans ($k = 1$):

$$\begin{aligned}\beta_{10} &= -1\\ \beta_{11} &= 0.0001\\ \beta_{12} &= 0.05\end{aligned}$$

For SUVs ($k = 2$):

$$\begin{aligned}\beta_{20} &= -2\\ \beta_{21} &= 0.0002\\ \beta_{22} &= 0.03\end{aligned}$$

The reference category is Trucks, so we set $\beta_{30} = \beta_{31} = \beta_{32} = 0$.

Calculating Linear Predictors:

For Sedans:

$$\begin{aligned}\eta_1 &= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2\\ &= -1 + 0.0001 \cdot 50000 + 0.05 \cdot 30\\ &= -1 + 5 + 1.5\\ &= 5.5\end{aligned}$$

For SUVs:

$$\begin{aligned}\eta_2 &= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 \\ &= -2 + 0.0002 \cdot 50000 + 0.03 \cdot 30 \\ &= -2 + 10 + 0.9 \\ &= 8.9\end{aligned}$$

For Trucks (reference category):

$$\eta_3 = 0$$

Calculating Probabilities:

We now use the linear predictors to calculate the probabilities.

For Sedans:

$$\Pr(Y = 1|X = x) = \frac{e^{\eta_1}}{1 + e^{\eta_1} + e^{\eta_2}} = \frac{e^{5.5}}{1 + e^{5.5} + e^{8.9}}$$

For SUVs:

$$\Pr(Y = 2|X = x) = \frac{e^{\eta_2}}{1 + e^{\eta_1} + e^{\eta_2}} = \frac{e^{8.9}}{1 + e^{5.5} + e^{8.9}}$$

For Trucks (reference category):

$$\Pr(Y = 3|X = x) = \frac{1}{1 + e^{\eta_1} + e^{\eta_2}} = \frac{1}{1 + e^{5.5} + e^{8.9}}$$

Numerical Calculations:

Let's compute the exponentials and probabilities.

$$e^{5.5} \approx 244.691$$
$$e^{8.9} \approx 7334.414$$

Now, sum the terms:

$$1 + e^{5.5} + e^{8.9} = 1 + 244.691 + 7334.414 \approx 7580.105$$

Finally, compute the probabilities:

For Sedans:

$$\Pr(Y = 1|X = x) \approx \frac{244.691}{7580.105} \approx 0.0323$$

For SUVs:

$$\Pr(Y = 2|X = x) \approx \frac{7334.414}{7580.105} \approx 0.9677$$

For Trucks:

$$\Pr(Y = 3|X = x) \approx \frac{1}{7580.105} \approx 0.00013$$

Interpretation:

Given the person's annual income of \$50,000 and age of 30, the model predicts the probabilities of owning each type of vehicle as follows:

- Sedan: 3.23%
- SUV: 96.77%
- Truck: 0.013%

Thus, this person is most likely to own an SUV based on the given predictors.

We now briefly present an alternative coding for multinomial logistic regression, know as the *softmax* coding. The softmax coding is equivalent to the coding just described in the sense that the fitted values, log odds between any pair of classes, and other key model outputs will remain the same, regardless of coding. But the softmax coding is used extensively in some areas of the machine learning literature, so it is worth being aware of it. In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, \cdots, K$,

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{3.13}$$

Thus, rather than estimating coefficients for $K-1$ classes, we actually estimate coefficients for all classes. It is not hard to see that as a result of (3.13), the log odds ratio between the kth and k'th classes equals

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p. \tag{3.14}$$

For the proof, let us write the probability of class k given $X = x$ as:

$$\Pr(Y = k|X = x) = \frac{e^{\eta_k}}{\sum_{l=1}^{K} e^{\eta_l}}$$

where $\eta_k = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p$.

Consider the ratio of the probabilities for classes k and k':

$$\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)} = \frac{\frac{e^{\eta_k}}{\sum_{l=1}^{K} e^{\eta_l}}}{\frac{e^{\eta_{k'}}}{\sum_{l=1}^{K} e^{\eta_l}}} = \frac{e^{\eta_k}}{e^{\eta_{k'}}}.$$

Take the natural logarithm of both sides:

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = \log(e^{\eta_k - \eta_{k'}}) = \eta_k - \eta_{k'}$$

Substitute the expressions for η_k and $\eta_{k'}$:

$$\eta_k = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p$$
$$\eta_{k'} = \beta_{k'0} + \beta_{k'1}x_1 + \cdots + \beta_{k'p}x_p$$

Therefore,

$$\begin{aligned}\eta_k - \eta_{k'} &= (\beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p) - (\beta_{k'0} + \beta_{k'1}x_1 + \cdots + \beta_{k'p}x_p) \\ &= (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p\end{aligned}$$

Thus, we have shown that:

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p$$

This completes the proof.

3.4 Generative Models for Classification

Skip

3.4.1 Linear Discriminant Analysis for $p = 1$

Skip

3.4.2 Linear Discriminant Analysis for $p > 1$

Skip

3.4.3 Quadratic Discriminant Analysis

Skip

3.4.4 Naive Bayes

Skip

3.5 A Comparison of Classification Methods

Skip

3.5.1 An Analytical Comparison

Skip

3.5.2 An Empirical Comparison

Skip

3.6 Generalized Linear Models

In Chapter 2, we assumed that the response Y is quantitative, and explored the use of least squares linear regression to predict Y. Thus far in this chapter, we have instead assume Y is qualitative. However, we may sometimes faced with situations in which Y is neither qualitative nor quantitative, and so neither linear regression from Chapter 2 nor the classification approaches covered in this chapter is applicable.

As a concrete example, we consider the Bikeshare data set. The response is bikers, the number of hourly users of a bike sharing program in Washington, DC. This response value is neither qualitative nor quantitative: instead, it takes on non-negative integer values, or *counts*. We will consider predicting bikers using the covariates mnth (month of the year), hr (hour of the day, from 0 to 23), workingday (an indicator variable that equals 1 if it is neither a weekend nor holiday), temp (the normalized temperature, in Celsius), and weathersit (a qualitative variable that takes on one of four possible values: clear; misty or cloudy; light rain or light snow; or heavy rain or heavy snow.)

In the analyses that follow, we will treat mnth, hr, and weathersit as qualitative variables.

Remark: In the context provided, treating mnth and hr as qualitative variables may seem unusual because they represent numeric values that can be ordered. However, in data analysis and statistical modeling, the classification of variables into qualitative (categorical) or quantitative (numerical) categories depends on how they are used in the analysis. Here, although mnth and hr are numeric, they represent categories (months of the year and hours of the day) rather than continuous measurements. Each value of mnth or hr corresponds to a distinct category rather than a quantity on a numerical scale.

Remark: When using mnth (month) as a categorical variable in a regression model, the approach involves creating a set of indicator (dummy) variables for each level of the month variable. Here's how you would typically include mnth in a regression model and interpret its coefficients:

Suppose you are using a multiple linear regression model, and mnth is one of the categorical predictors. The model formula can be written as:

$$\text{Response} = \beta_0 + \beta_1 \cdot \text{mnth1} + \beta_2 \cdot \text{mnth2} + \cdots + \beta_{11} \cdot \text{mnth11} + \text{other predictors} + \epsilon$$

- mnth1, mnth2, ..., mnth11 are dummy variables representing each of the months except one (which serves as the reference category).
- β_1, β_2, ..., β_{11} are the regression coefficients for these dummy variables.
- β_0 is the intercept of the model.
- other predictors represent other variables in the model.
- ϵ is the error term.

Steps to Construct the Dummy Variables:

1. **Choose a Reference Month**: One month will be chosen as the reference category. This month will not have a corresponding dummy variable and serves as the baseline for comparison.

2. **Create Dummy Variables**: For each month other than the reference month, create a dummy variable that equals 1 if the observation is in that month and 0 otherwise.

   For example, if January is the reference month:

   - mnth2 would be 1 if the month is February and 0 otherwise.
   - mnth3 would be 1 if the month is March and 0 otherwise, and so on.

Coefficient Interpretation

- **Intercept (β_0)**: The intercept represents the expected value of the response variable when all predictors are set to zero, including when mnth is in the reference category (e.g., January).
- **Coefficients for Dummy Variables (β_1, β_2, ..., β_{11})**: Each coefficient for the dummy variables represents the change in the response variable relative to the reference month.

  For instance:

  – If β_1 is the coefficient for mnth2 (February), it indicates the difference in the response variable between February and January. A positive β_1 suggests that February has higher values compared to January, while a negative β_1 suggests lower values.
  – Similarly, β_2 for mnth3 (March) would indicate the difference between March and January.

Example:

Suppose you have a model for predicting bike usage with January as the reference month and the following coefficients:

- Intercept (β_0): 100 (bike rides)
- Coefficient for mnth2 (February) (β_1): 10
- Coefficient for mnth3 (March) (β_2): -5

If the bike usage in January is 100 rides, then:

- In February, bike usage would be $100 + 10 = 110$ rides.
- In March, bike usage would be $100 - 5 = 95$ rides.

The coefficients for the dummy variables tell you how much the bike usage deviates from the baseline month (January) for each other month.

3.6.1 Linear Regression on the Bikeshare Data

To begin, we consider predicting bikers using linear regression. The results are shown in Figure 3.8.

We see, for example, that a progression of weather from clear to cloudy results in, on average, 12.89 fewer bikers per hour; however, if the weather progresses further to rain or snow, then this further results in 53.60 fewer bikers per hour. Figure 3.9 displays the coefficients associated with mnth and the coefficients associated with hr. We see that bike usage is highest in the spring and fall, and lowest during the winter months. Furthermore, bike usage is greatest

around rush hour (9 AM and 6 PM), and lowest overnight. Thus, at first glance, fitting a linear regression model to the Bikeshare data set seems to provide reasonable and intuitive results.

But upon more careful inspection, some issues become apparent. For example, 9.6% of the fitted values in the Bikeshare data set are negative: that is, the linear regression model predicts a *negative* number of users during 9.6% of the hours in the data set. This calls into question our ability to perform meaningful predictions on the data, and it also raises concerns about the accuracy of the coefficient estimates, confidence intervals, and other outputs of the regression model.

	Coefficient	Std. error	*t*-statistic	*p*-value
`Intercept`	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	-12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	-66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15

Figure 3.8: Results for a least squares linear model fit to predict bikers in the Bikeshare data. The predictors mnth and hr are omitted from this table due to space constraints, and can be seen in Figure 3.9. For the qualitative variable weathersit, the baseline level corresponds to clear skies.

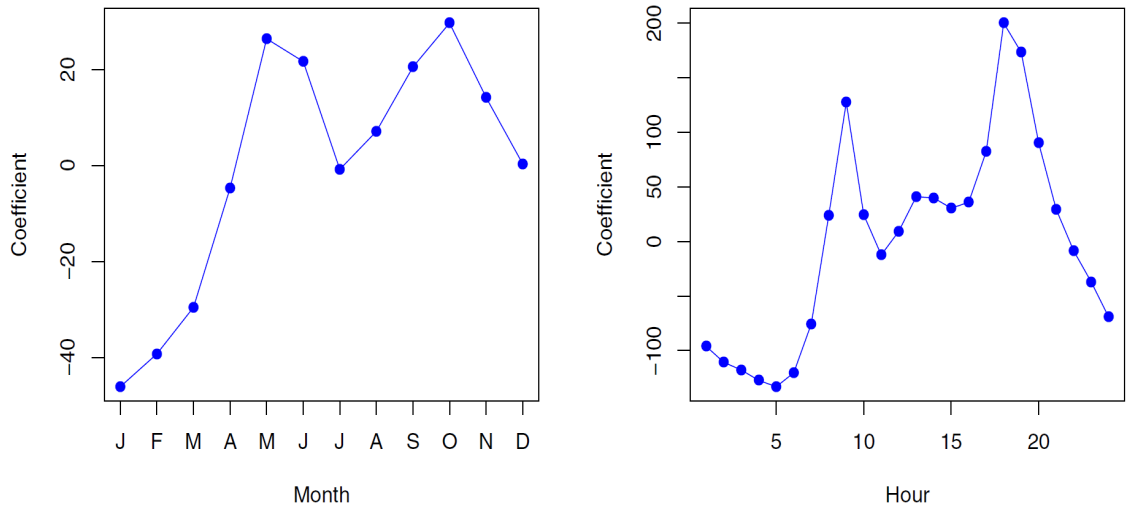


Figure 3.9: A least squares linear regression model was fit to predict bikers in the Bikeshare data set. Left: The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter. Right: The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight.

Furthermore it is reasonable to suspect that when the expected value of bikers is small, the variance of bikers should be small as well. For instance, at 2 AM during a heavy December snow storm, we expect that extremely few people will use a bike, and moreover that there should be little variance associated with the number of users during those conditions. This is borne out in the data: between 1 AM and 4 AM, in December, January, and February, when it is raining, there are 5.05 users, om average, with a standard deviation of 3.73. By contrast,

between 7 AM and 10 AM, in April, May, and June, when skies are clear, there are 243.59 users, on average, with a standard deviation of 131.7. The mean-variance relationship is displayed in the left-hand panel of Figure 3.10. This is a major violation of the assumptions of a linear model, which state that

$$Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon,$$

where ϵ is a mean-zero error term with variance σ^2 that is *constant*, and not a function of the covariates. Therefore, the heteroscedasticity of the data calls into question the suitability of a linear model.

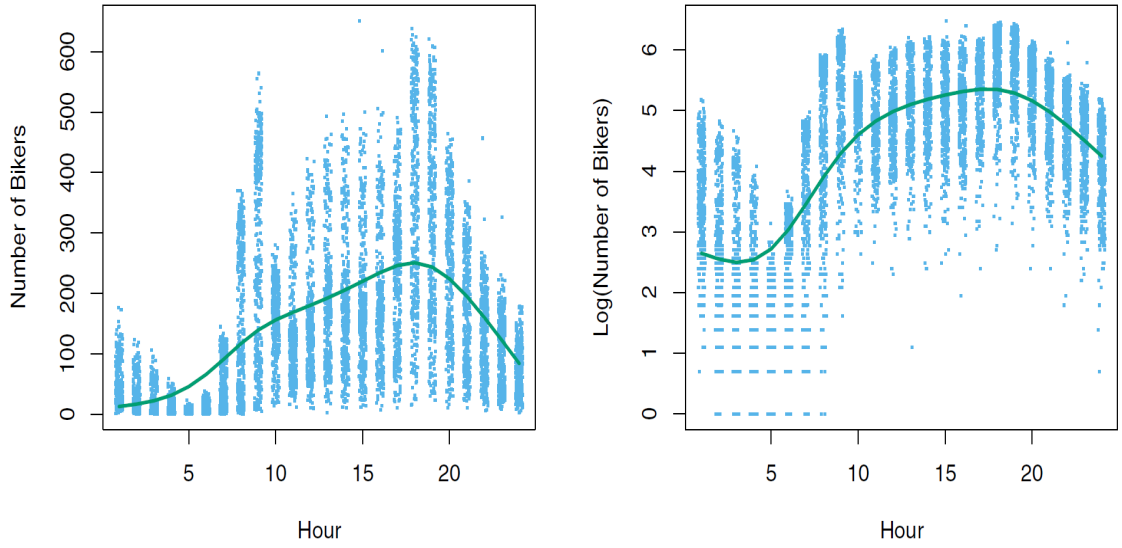


Figure 3.10: Left: On the Bikeshare dataset, the number of bikers is displayed on the y-axis, and the hour of the day is displayed on the x-axis. Jitter was applied for ease of visualization. For the most part, as the mean number bikers increases, so does the variance in the number of bikers. A smoothing spline fit is shown in green. Right: The log of the number of bikers is now displayed on the y-axis.

Finally, the response bikers is integer-valued. But under a linear model,

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon,$$

where ϵ is a continuous-valued error term. This means that in a linear model, the response variable Y is necessarily continuous-valued (quantitative). Thus, the integer nature of the response bikers suggests that a linear regression model is not entirely statisfactory for this data set.

Some of the problems that arise when fitting a linear regression model to the Bikeshare data can be overcome by transforming the response; for instance, we can fit the model

$$\log(Y) = \sum_{j=1}^{p} X_j \beta_j + \epsilon.$$

Transforming the response avoids the possibility of negative predictions, and it overcomes much of the heteroscedasticity in the untransformed data, as is shown in the right-hand panel of Figure 3.10. However, it is not quite a satisfactory solution, since predictions and inference

are made in terms of the log of the response, rather than the response. This leads to challenges in interpretation, e.g. "a one-unit increase in X_j is associated with an increase in the mean of the log of Y by an amount β_j". Furthermore, a log transformation of the response cannot be applied in settings where the response can take on a value of 0. Thus, while fitting a linear model to a transformation of the response may be an adequate approach for some count-valued data sets, it often leaves something to be desired. We will see in the next section that a Poisson regression model provides a much more natural and elegant approach for this task.

3.6.2 Poisson Regression on the Bikeshare Data

To overcome the inadequacies of linear regression for analyzing the Bikeshare data set, we will make use of an alternative approach, called *Poisson regression*. Before we can talk about Poisson regression, we must first introduce the Poisson distribution.

Suppose that a random variable Y takes on nonnegative interger values, i.e. $Y \in \{0, 1, 2, \cdots\}$. If Y follows the Poisson distribution, then

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \cdots, \tag{3.15}$$

Here, $\lambda > 0$ is the expected value of Y, i.e. $E(Y)$. It turns out that λ also equals the variance of Y, i.e. $\lambda = E(Y) = \text{Var}(Y)$. This means that if Y follows the Poisson distribution, then the larger the mean of Y, the larger its variance.

The Poisson distribution is typically used to model counts; this is a natural choice for a number of reasons, including the fact that counts, like the Poisson distribution, take on nonnegative integer values. To see how we might use the Poisson distribution in practice, let Y denote the number of users of the bike sharing program during a particular hour of the day, under a particular set of weather conditions, and during a particular month of the year. We might model Y as a Poisson distribution with mean $E(Y) = \lambda = 5$. This means that the probability of no users during this particular hour is

$$\Pr(Y = 0) = \frac{e^{-5}5^0}{0!} = e^{-5} = 0.0067.$$

The probability that there is exactly one user is

$$\Pr(Y = 1) = \frac{e^{-5}5^1}{1!} = 5e^{-5} = 0.034,$$

the probability of two users is

$$\Pr(Y = 2) = \frac{e^{-5}5^2}{2!} = 0.084,$$

and so on.

Of course, in reality, we expect the mean number of users of the bike sharing program, $\lambda = E(Y)$, to vary as a function of the hour of the day, the month of the year, the weather conditions, and so forth. So rather than modeling the number of bikers, Y as a Poisson distribution with a fixed mean value like $\lambda = 5$, we would like to allow the mean to vary as a function of the covariates. In particular, we consider the following model for the mean $\lambda = E(Y)$, which is we now write as $\lambda(X_1, \cdots, X_p)$ to emphasize that it is a function of the covariates $X_1, \cdots, X_p$:

$$\log\left(\lambda(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.16}$$

or equivalently

$$\lambda(X_1, \cdots, X_p) = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}. \tag{3.17}$$

Here, $\beta_0, \beta_1, \cdots, \beta_p$ are parameters to be estimated. Together, (3.15) and (3.16) define the Poisson regression model.

Notice that in (3.16), we take the log of $\lambda(X_1, \cdots, X_p)$ to be linear in $X_1, \cdots, X_p$, rather than having $\lambda(X_1, \cdots, X_p)$ itself to be linear in $X_1, \cdots, X_p$, in order to ensure that $\lambda(X_1, \cdots, X_p)$ takes on nonnegative values for all values of the covariates.

To estimate the coefficients $\beta_0, \beta_1, \cdots, \beta_p$, we use the same maximum likelihood approach that we adopted for logistic regression in Section 3.3.2. Specifically, given n independent observations from the Poisson regression model, the likelihood takes the form

$$l(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!} \tag{3.18}$$

where $\lambda(x_i) = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}$, due to (3.17). We estimate the coefficients that maximize the likelihood $l(\beta_0, \beta_1, \cdots, \beta_p)$, i.e. that make the observed data as likely as possible.

We now fit a Poisson regression model to the Bikeshare data set. The results are shown in Figure 3.11 and Figure 3.12. Qualitatively, the results are similar to those from linear regression in Section 3.6.1. We again see that bike usage is highest in the spring and fall and during rush hour, and lowest during the winter and in the early morning hours. Moreover, bike usage increases as the temperature increases, and decreases as the weather worsen. Interestingly, the coefficient associated with weathersit is statistically significant under the Poisson regression model, but not under the linear regression model.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

Figure 3.11: Results for a Poisson regression model fit to predict bikers in the Bikeshare data. The predictors mnth and hr are omitted from this table due to space constraints, and can be seen in Figure fig: Book Figure 4.15. For the qualitative variable weathersit, the baseline corresponds to clear skies.

Some important distinctions between the Poisson regression model and the linear regression model are as follows:

- *Interpretation*: To interpret the coefficients in the Poisson regression model, we must pay close attention to (3.17), which states that an increase in X_j by one unit is associated with a change in $E(Y) = \lambda$ by a factor of $\exp(\beta_j)$. For example, a change in weather from clear to cloudy skies is associated with a change in mean bike usage by a factor of $\exp(-0.08) = 0.923$, i.e. on average, only 92.3% as many people will use bikes when it is cloudy relative to when it is clear. If the weather worsens further and it begins to rain,

then the mean bike usage will further change by a factor of $\exp(-0.5) = 0.607$, i.e. on average only 60.7% as many people will use bikes when it is rainy relative to when it is cloudy.

- *Mean-variance relationship*: As mentioned earlier, under the Poisson model, $\lambda = E(Y) = \text{Var}(Y)$. Thus, by modeling bike usage with a Poisson regression, we implicitly assume that mean bike usage in a given hour equals the variance of bike usage during that hour. By contrast, under a linear regression model, the variance of bike usage always takes on a constant value. Recall from Figure 3.10 that in the Bikeshare data, when biking conditions are favorable, both the mean and the variance in bike usage are much higher than when conditions are unfavorable. Thus, the Poisson regression model is able to handle the mean-variance relationship seen in the Bikeshare data in a way that the linear regression model is not.

- *Nonnegative fitted values* There are no negative predictions using the Poisson regression model. This is because the Poisson model itself only allows for nonnegative values; see (3.15). By contrast, when we fit a linear regression model to the Bikeshare data set, almost 10% of the predictions were negative.

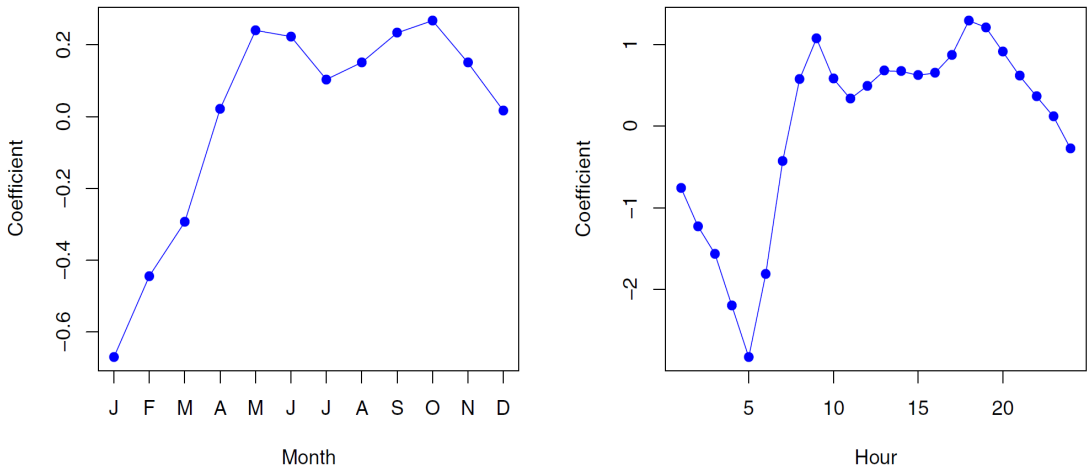


Figure 3.12: A Poisson regression model was fit to predict bikers in the Bikeshare data set. Left: The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and the lowest in the winter. Right: The coefficients associated with the hour of the day. Bike usage is the highest during peak commute times, and lowest overnight.

3.6.3 Generalized Linear Models in Greater Generality

We habe now discussed three types of regression models: linear, logistic and Poisson. These approaches share some common characteristics:

(1) Each approach uses predictors $X_1, \cdots, X_p$ to predict a response Y. We assume that, conditional on $X_1, \cdots, X_p$, Y belongs to a certain family of distributions. For linear regression, we typically assume that Y follows a Gaussian or normal distribution. For logistic regression, we assume that Y follows a Bernoulli distribution. Finally, for Poisson regression, we assume that Y follows a Poisson distribution.

(2) Each approach models the mean of Y as a function of the predictors.

(i) In linear regression, the mean of Y takes the form

$$E(Y|X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{3.19}$$

i.e. it is a linear function of the predictors.

(ii) For logistic regression, the mean instead takes the form

$$E(Y|X_1, \cdots, X_p) = \Pr(Y = 1|X_1, \cdots, X_p) \tag{3.20}$$

$$= \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}, \tag{3.21}$$

(iii) while for Poisson regression it takes the form

$$E(Y|X_1, \cdots, X_p) = \lambda(X_1, \cdots, X_p) = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}. \tag{3.22}$$

Equations (3.19)-(3.22) can be expressed using a *link function*, η, which applies a transformation to $E(Y|X_1, \cdots, X_p)$ so that the transformed mean is a linear function of the predictors. That is,

$$\eta(E(Y|X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{3.23}$$

The link function for linear, logistic and Poisson regression are

(i) $\eta(\mu) = \mu$

(ii) $\eta(\mu) = \log(\mu/1 - \mu)$, and

(iii) $\eta(\mu) = \log(\mu)$,

respectively.

The Gaussian, Bernoulli and Poisson distributions are all members of a wider class of distribution known as the *exponential* family. Other well-known members of this family are the *exponential distribution*, the *Gamma distribution*, and the *negative binomial distribution*. In general, we can perform a regression by modeling the response Y as coming from a particular member of the exponential family, and then transforming the mean of the response so that the transformed mean is a linear function of the predictors via (3.23). Any regression approach that follows this very general recipe is known as a *generalized linear model* (GLM). Thus, linear regression, logistic regression, and Poisson regression are three examples of GLMs. Other examples not covered here include *Gamma regression* and *negative binomial regression*.

Chapter 4

Resampling Nethods

Resampling methods are an indispensable tool in modern statistics. They involve repeatedly drawing samples from a training data set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. For example, in order to estimate the variability of a linear regression fit, we can repeatedly draw different samples from the training data, fit a linear regression to each new sample, and then examine the extend to which the resulting fits differ. Such an approach may allow us to obtain information that would not be avilable from fitting the model only once using the original training sample.

Resampling approaches can be computationally expensive, because they involve fitting the same statistical method multiple times using different subsets of the training data. However, due to recent advances in computing power, the computational requirements of resampling methods generally are not prohibitive. In this chapter, we discuss two of the most commonly used resampling methods, *cross-validation* and *bootstrap*. Both methods are important tools in practical application of many statistical learning procedures. For example, cross-validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility. The process of evaluating a model's performance is known as *model assessment*, whereas the process of selecting the proper level of flexibility for a model is known as *model selection*. The bootstrap is used in several contexts, most commonly to provide a measure of accuracy of a parameter estimate or of a given statistical learning method.

4.1 Cross-Validation

In Chapter 1 we discuss the distinction between the *test error rate* and the *training error rate*. The test error is the average error that results from using a statistical learning method to predict the response on a new observation–that is, a measurement that was not used in training the method. Given a data set, the use of a particular statistical learning method is warranted if it results in a low test error. The test error can be easily calculated if a designated test set is available. Unfortunately, this is usually not the case. In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training. But as we saw in Chapter 1, the training error rate often is quite different from test error rate, and in particular the former can dramatically underestimate the latter.

In the absence of a very large designated test set that can be used to directly estimate the test error rate, a number of techniques can be used to estimate this quantity using the available training data. Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate. Such approaches are discussed in Chapter 5. In this

section, we instead consider a class of methods that estimate the test error rate by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

In Sections 4.1.1-4.1.4, for simplicity we assume that we are interested in performing regression with a quantitative response. In Section 4.1.5 we consider the case of classification with a qualitative response. As we will see, the key concepts remain the same regardless of whether the response is quantitative or qualitative.

4.1.1 The Validation Set Approach

Suppose that we would like to estimate the test error associated with fitting a particular statistical learning method on a set of observations. The *validation set approach*, displayed in Figure 4.1, is a very simple strategy for this task. It involves randomly dividing the avilable set of observations into two parts, a *training set* and a *validation set* or *hold-out* set. The model is fit on the training set, and the fitted model is used to predict the response for the observations in the validation set. The resulting validation set error rate–typically assessed using MSE in the case of a quantitative response–provides an estimate of the test error rate. Recall that the MSE we referring to is given by (1.5) and restated below:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2,$$

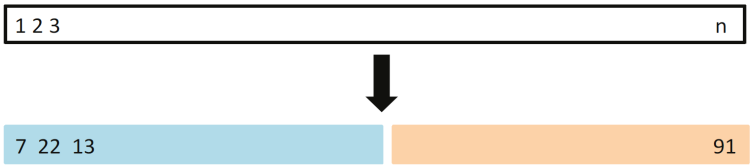


Figure 4.1: A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

We illustrate the validation set approach on the Auto data set. Recall from Chapter 2 that there appears to be a non-linear relationship between mpg and horsepower, and that a model that predict mpg using horsepower and horsepower^2 gives better results than a model that uses only a linear term. It is natural to wonder whether a cubic or higher-order fit might provide even better results. We answer this question in Chapter 2 by looking at the p-values associated with a cubic term and higher-order polynomial terms in a linear regression. But we could also answer this question using the validation method.

We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations. The validation set error rates that result from fitting various regression models on the training sample and evaluating their performance on the validation sample, using MSE as a measure of validation set error, are shown in the left-hand panel of Figure 4.2. The validation set MSE for the quadratic fit is considerably smaller than for the linear fit. However, the validation set MSE for the cubic fit is actually slightly larger than for the quadratic fit. This implies that including a cubic term in the regression does not lead to better prediction than simply using a quadratic term.

Recall that in order to create the left hand panel of Figure 4.2, we randomly divides the data set into two parts, a training set and a validation set. If we repeat the process of randomly splitting the sample set into two parts, we will get a somewhat different estimate for the test MSE. As an illustration, the right-hand panel of Figure 4.2 displays ten different validation set MSE curves from the Auto data set, produced using ten different random splits of the observations into training and validation sets. All ten curves indicate that the model with a quadratic term has a dramatically smaller validation set MSE than the model with only a linear term. Furthermore, all ten curves indicate that there is not much benefit in including cubic or higher-order polynomial terms in the method. But it is worth noting that each of the ten curves results in a different test MSE estimate for each of the ten regression models considered. And there are no consensus among the curves as to which model results in the smallest validation set MSE. Based on the variability among these curves, all that we can conclude with any confidence is that the linear fit is not adequate for this data.

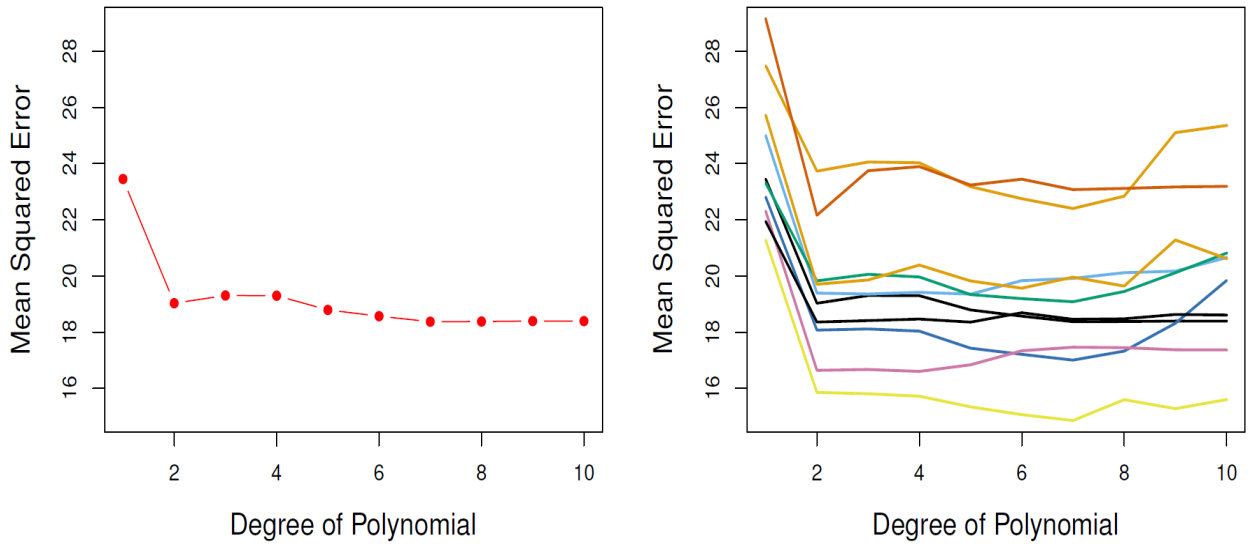


Figure 4.2: The validation set approach was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: Validation error estimates for a single split into training and validation data sets. Right: The validation method was repeated ten times each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.

The validation set approach is conceptually simple and is easy to implement. But it has two potential drawbacks:

(1) As is shown in the right-hand panel of Figure 4.2, the validation estimate of the test set error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

(2) In the validation approach, only a subset of the observations–those that are included in the training set rather than in the validation set–are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire data set.

In the coming subsections, we will present *cross-validation*, a refinement of the validation set approach that addresses these two issues.

4.1.2 Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) is closely related to the validation set approach of Section 4.1.1, but it attempts to address that method's drawbacks.

Like the validation set approach, LOOCV involves splitting the set of observations into two parts, However, instead of creating two subsets of comparable size, a single observation (x_1, y_1) is used for the validation set and the remaining observations $\{(x_2, y_2), \cdots, (x_n, y_n)\}$ make up the training set. The statistical learning method is fit on the $n-1$ training observations, and a prediction $\widehat{y}_1$ is made for the excluded observation, using its value x_1. Since (x_1, y_1) was not used in the fitting process,

$$\text{MSE}_1 = (y_1 - \widehat{y}_1)^2$$

provides an approximately unbiased estimate for the test error. But even though MSE_1 is unbiased for the test error, it is a poor estimate because it is highly variable, since it based upon a single observation (x_1, y_1).

We can repeat the procedure by selecting (x_2, y_2) for the validation data, training the statistical learning procedure on the $n-1$ observations $\{(x_1, y_1), (x_3, y_3), \cdots, (x_n, y_n)\}$, and computing

$$\text{MSE}_2 = (y_2 - \widehat{y}_2)^2.$$

Repeating this approach n times produces n squared errors, $\text{MSE}_1, \cdots, \text{MSE}_n$. The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i \tag{4.1}$$

A scematic of the LOOCV approach is illustrated in Figure 4.3.

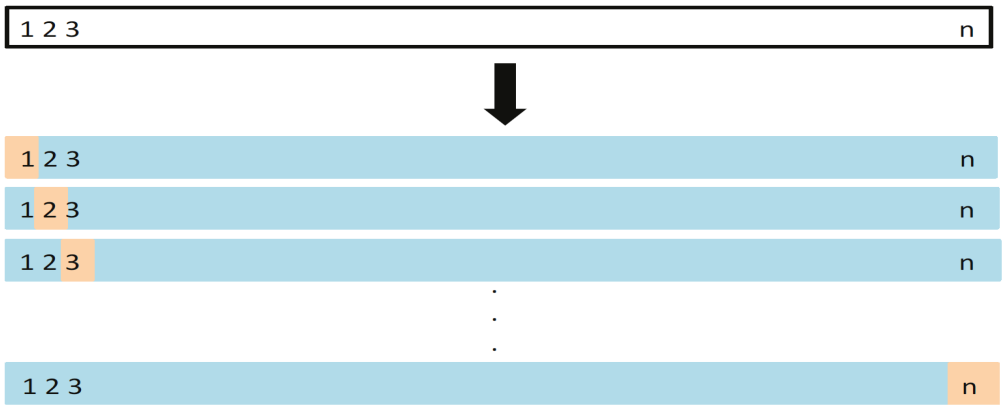


Figure 4.3: A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

LOOCV has a couple of major advantages over the validation set approach. First, it has far less bias. In LOOCV, we repeatedly fit the statistical learning method using training sets that

contain $n-1$ observations, almost as many as are in the entire data set. This is in contrast to the validation set approach, in which the training set is typically around half the size of the original data set. Consequently, the LOOCV approach tends not to overestimate the test error rate as much as the validation set approach does. Second, in contrast to the validation approach which will yield different results when applied repeatedly due to randomness in the training/validation set splits, performing LOOCV multiple times will always yield the same results: there is no randomness in the training/validation set splits.

We used LOOCV on the Auto data set in order to obtain an estimate of the test set MSE that results from fitting a linear regression model to predict mpg using polynomial functions of horsepower. The results are shown in the left-hand panel of Figure 4.4.

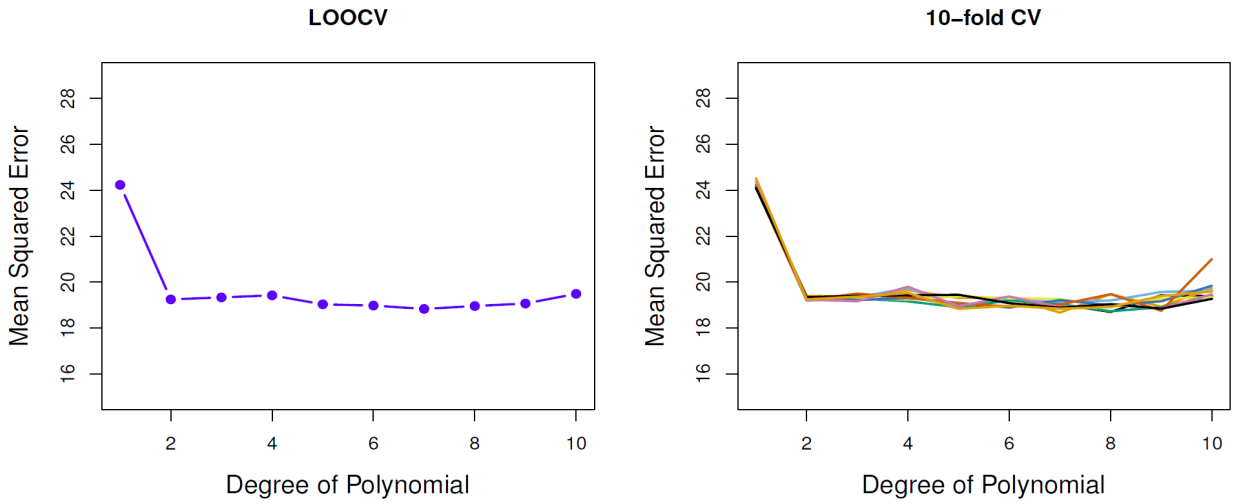


Figure 4.4: Cross-validation was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: The LOOCV error curve. Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.

LOOCV has the potential to be expensive to implement, since the model has to be fit n times. This can be very time consuming if n is large, and if each individual model is slow to fit. With least squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1-h_i}\right)^2, \tag{4.2}$$

where $\widehat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the leverage defined in (2.90)

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}.$$

This is like the ordinary MSE, except the ith residual is divided by $1-h_i$. The leverage lies between $1/n$ and 1, and reflects the amount that an observation influences its own fit. Hence the residuals for high-leverage points are inflated in this formula by exactly the right amount for this equality to hold.

LOOCV is a very general method, and can be used with any kind of predictive modeling. For example we could use it with logistic regression or linear discriminant analysis, or any of

the method discussed in later chapters. The magic formula (4.2) does not hold in general, in which case the model has to be refit n times.

Remark: In the case of multiple linear regression, the leverage takes a slightly more complicated form than (2.90), but (4.2) still holds.

4.1.3 k-Fold Cross-Validation

An alternative to LOOCV is k-fold CV. This approach involves randomly diving the set of observations into k groups, or *folds*, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. The mean squared error, MSE_1, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, $\text{MSE}_1, \text{MSE}_2, \cdots, \text{MSE}_k$. The k-fold CV estimate is computing by averaging these values,

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i. \tag{4.3}$$

Figure 4.5 illustrates the k-fold CV approach.

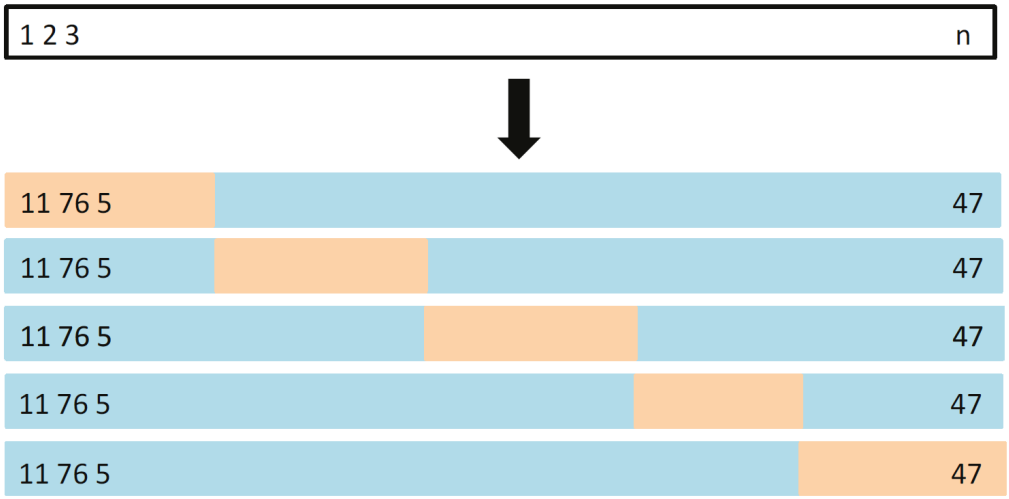


Figure 4.5: A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

It is not hard to see that LOOCV is a special case of k-fold CV in which k is set to equal n. In practice, one typically performs k-fold CV using $k = 5$ or $k = 10$. What is the advantage of using $k = 5$ or $k = 10$ rather than $k = n$? The most obvious advantage is computational. LOOCV requires fitting the statistical learning method n times. This has the potential to be computationally expensive (except for linear fit by least squares, in which case formula (4.2) can be used.) But cross-validation is a very general approach that can be applied to almost any statistical learning method. Some statistical learning methods have computationally intensive fitting procedures, and so performing LOOCV may pose computational problems, especially if n is extremely large. In contrast, performing 10-fold CV requires fitting the learning procedure only ten times, which may be more feasible. As we see in Section 4.1.4, there also can

be other non-computational advantages to performing 5-fold or 10-fold CV, which involve the bias-variance trade-off.

The right-hand panel of Figure 4.4 displays nine different 10-fold CV estimates for the Auto data set, each resulting from a different random split of the observations into ten folds. As we can see from the figure, there is some variability in the CV estimates as a result of the variability in how the observations are divided into ten folds. But this variability is typically much lower than the variability in the test error estimates that results from the validation set approach (right-hand panel of Figure 4.2).

When we examine real data, we do not know the *true* test MSE, and so it is difficult to determine the accuracy of the cross-validation estimate. However, if we examine simulated data, then we can compute the true test MSE, and can thereby evaluate the accuracy of our cross-validation results. In Figure 4.6, we plot the cross-validation estimates and true test error rates that result from applying smoothing splines to the simulated data sets as illustrated in Figures 1.10-1.12 of Chapter 1. The true test MSE is displayed in blue. The black dashed and orange solid lines respectively show the estimated LOOCV and 10-fold CV estimates. In all three plots, the two cross-validation estimates are very similar. In the right-hand panel of Figure 4.6, the true test MSE and the cross-validation curves are almost identical. In the center panel of Figure 4.6, the two sets of curves are similar at the lower degree of flexibility, while the CV curves overestimate the test MSE for higher degrees of flexibility. In the left-hand panel of 4.6, the CV curves have the correct general shape, but they underestimate the true test MSE.

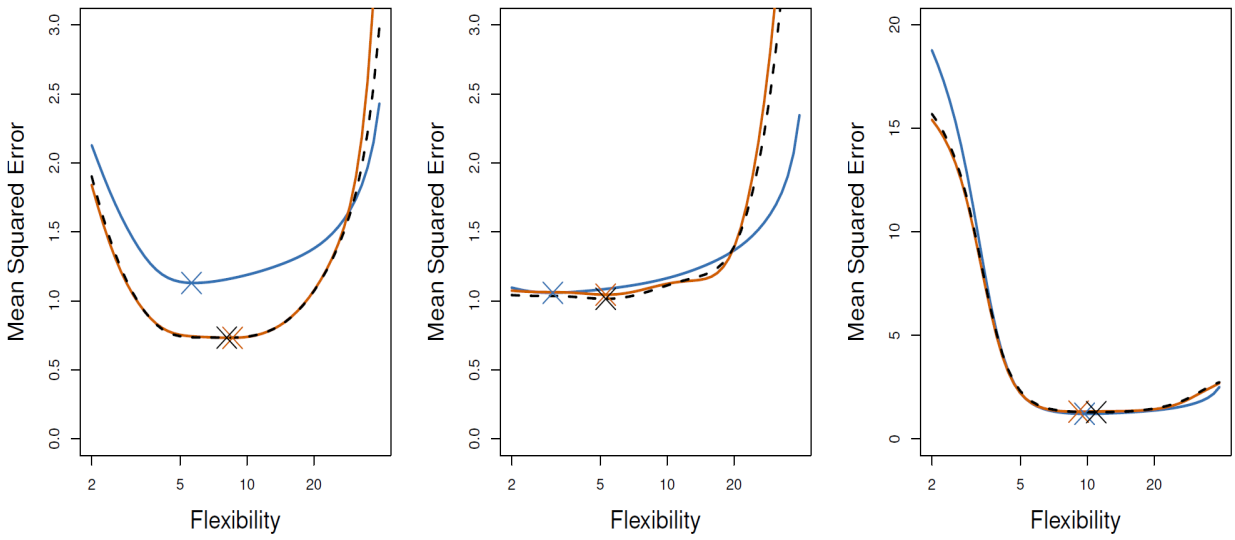


Figure 4.6: True and estimated test MSE for the simulated data sets in Figure 1.10 (left), 1.11 (center), and 1.12 (right). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.

When we perform cross-validation, our goal might be to determine how well a given statistical learning procedure can be expected to perform on independent data; in this case, the actual estimate of the test MSE is of interest. But at other times we are interested only in the location of the *minimum point* in the estimated test MSE curve. This is because we might be performing cross-validation on a number of statistical learning methods, or on a sigle method using different levels of flexibility, in order to identify the method that results in the lowest test error. For this purpose, the location of the minimum point in the estimated test MSE curve is important,

but the actual value of the estimated test MSE is not. We find in Figure 4.6 that dispite the fact that they sometimes underestimate the true test MSE, all of the CV curves come close to identifying the correct level of flexibility—that is, the flexibility level corresponding to the smallest test MSE.

4.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation

We mentioned in Section 4.1.3 that k-fold CV with $k < n$ has a computational advantage to LOOCV. But putting computational issue aside, a less obvious but potentially more important advantage of k-fold CV is that it often gives more accurate estimates of the test error rate than does LOOCV. This has to do with a bias-variance trade-off.

It was mentioned in Section 4.1.1 that the validation set approach can lead to overestimates of the test error rate, since in this approach the training set used to fit the statistical learning methods contains only half of the observations of the entire data set. Using this logic, it is not hard to see that LOOCV will give approximately unbiased estimates of the test error, since each training set contains $n-1$ observations, which is almost as many as the number of observations in the full data set. And performing k-fold CV, for $k = 5$ or $k = 10$ will lead to an intermediate level of bias, since each training set contains approximately $(k-1)n/k$ observations—fewer than in the LOOCV approach, but substantially more than in the validation set approach. Therefore, from the perspective of bias reduction, it is clear that LOOCV is to be preferred to k-fold CV.

However, we know that bias is not the only source of concern in an estimating procedure: we must also consider the procedure's variance. It turns out that LOOCV has higher variance than does k-fold CV with $k < n$. Why is this the case? When we perform LOOCV, we are in effect averaging the output of n fitted models, each of which is trained on an almost identical set of observations: therefore, these outputs are highly (positively) correlated with each other. In contrast, when we perform k-fold CV with $k < n$, we are averaging the outputs of k fitted models that are somewhat less correlated with each other, since the overlap between the training sets in each model is smaller. Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.

To summarize, there is a bias-variance trade-off associated with the choice of k in the k-fold cross-validation. Typically, given these considerations, one performs k-fold cross-validation using $k = 5$ or $k = 10$, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

Remarks: To wrap up this subsection, we'll demonstrate how correlation impacts the variance of the mean with an example.

Consider n random variables $X_1, X_2, \ldots, X_n$. Let the mean of these variables be given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We want to determine the variance of $\bar{X}$, which is:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Using the properties of variance, we have:

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Thus:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{1\le i<j\le n}\text{Cov}(X_i, X_j)\right)$$

Assume each random variable X_i has the same variance σ^2, and let ρ be the correlation coefficient between any two variables X_i and X_j. Thus, $\text{Cov}(X_i, X_j) = \rho\sigma^2$. Substituting these into the formula gives:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\left(n\sigma^2 + 2\sum_{1\le i<j\le n}\rho\sigma^2\right)$$

The number of pairs (i, j) where $1 \le i < j \le n$ is $\frac{n(n-1)}{2}$. Hence:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\left(n\sigma^2 + 2\cdot\frac{n(n-1)}{2}\rho\sigma^2\right)$$

Simplify this to:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\left(n\sigma^2 + (n^2-n)\rho\sigma^2\right)$$

$$\text{Var}(\bar{X}) = \frac{n\sigma^2 + (n^2-n)\rho\sigma^2}{n^2}$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{(n-1)\rho\sigma^2}{n}$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 + (n-1)\rho\right)$$

Let's now examine these two cases:

- *High Correlated Case:* If the variables are highly correlated, ρ is close to 1. For instance, if $\rho = 1$:

  $$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 + (n-1)\cdot 1\right) = \frac{\sigma^2}{n}\cdot n = \sigma^2$$

  The variance of the mean in this case is σ^2.

- *Less Correlated Case:* If the variables are less correlated, ρ is closer to 0. For instance, if $\rho = 0$:

  $$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 + (n-1)\cdot 0\right) = \frac{\sigma^2}{n}$$

  The variance of the mean in this case is $\frac{\sigma^2}{n}$, which is lower than σ^2.

From the calculations above, it is evident that as the correlation ρ between the variables increases, the variance of the mean of the variables increases. When the variables are highly correlated, the variance of their mean approaches σ^2, which is higher than $\frac{\sigma^2}{n}$ when the variables are uncorrelated. Thus, the variance of the mean increases with the correlation between the variables. In our context, LOOCV leads to highly correlated prediction errors because the training sets are very similar. This correlation results in higher variance in the performance estimate, as the test MSE is averaged from these correlated test errors, see (4.1).

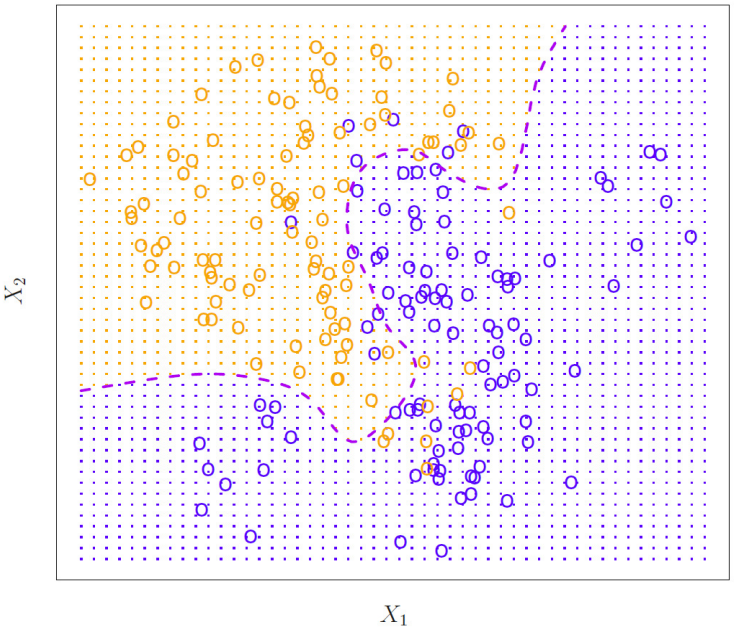
4.1.5 Cross-Validation on Classification Problems

In this chapter so far, we have illustrated the use of cross-validation in the regression setting where the outcome Y is quantitative, and so have used MSE to quantify test error. But cross-validation can also be very useful approach in the classification setting when Y is qualitative. In this setting, cross-validation works just as described earlier in this chapter, except that rather than using MSE to quantify test error, we instead use the number of misclassified observations. For instance, in the classification setting, the LOOCV error rate takes the form

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i, \tag{4.4}$$

where $\mathrm{Err}_i = I(y_i \neq \widehat{y}_i)$. The k-fold CV error rate and validation set error rates are defined analogously.

As an example, we fit various logistic regression models on the two dimensional classification data displayed in Figure 1.14 (reproduced below). In the top-left panel of Figure XXX,



the black solid line shows the estimated decision ... Page 206 of text.