# 1 Simple Linear Regression

This and the next chapters are about linear regression. In particular, linear regression is a useful tool for predicting a quantitative response. Linear regression has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern statistical modelling approaches described in later chapters of this notes, linear regression is still a useful and widely used statistical learning method. Moreover, it serves as a good jumping-off point for newer approaches: as we will see in later chapters, many fancy statistical approaches can be seen as generalizations or extensions of linear regression. Consequently, the importance of having a good understanding of linear regression before studying more complex statistical methods cannot be overstated. In this and the next chapters, we review some of the key ideas underlying the linear regression model, as well as the least squares approach that is most commonly used to fit this model.

In order to motivate our study of linear regression, we begin with a simple example. Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. The `Advertising` data set consists of the `sales` of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: `TV`, `radio`, and `newspaper`. The data are displayed in Figure 1. It is not possible for our client to directly increase sales of the product. On the other hand, 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 increasing 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 setting, the advertising budgets are input variables while sales is an output variable. The input variables are typically denoted 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 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 is typically denoted using the symbol $Y$. Throughout this notes, we will use all of these terms interchangeably.

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response $Y$ on the basis of a 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

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

You might read "$\approx$"as "is approximately modeled as". We will sometimes describe (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

$$\texttt{sales} \approx \beta_0 + \beta_1 \texttt{TV}.$$

In equation 1, $\beta_0$ and $\beta_1$ are two unknown constants that represent the intercept and slope terms in the linear model. Together, $\beta_0$ and $\beta_1$ are known as the model coefficients parameters. We can use the data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, and predict future sales on the basis of a particular value of `TV` advertising by computing

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2}$$

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

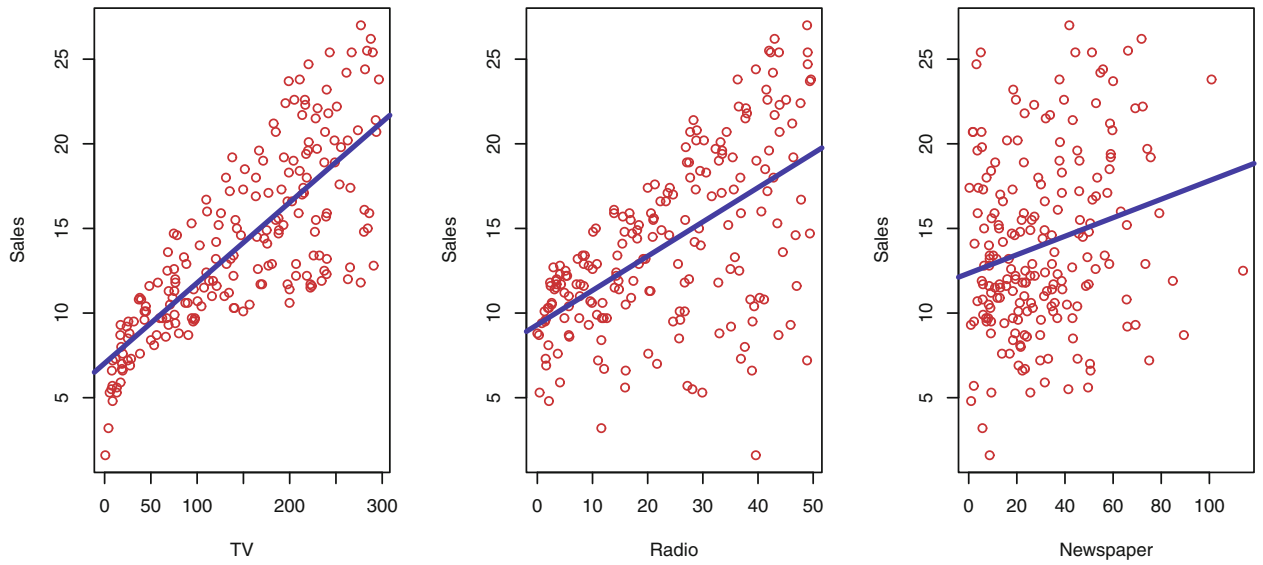


Figure 2: The `Advertising` data set. The plot displays `sales`, in thousands of units, as a function of `TV`, `radio`, and `newspaper` budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable. In other words, each blue line represents a simple model that can be used to predict `sales` using `TV`, `radio`, and `newspaper`, respectively.

### 1.1 Estimating the Coefficients

In practice, $\beta_0$ and $\beta_1$ are unknown. So before we can use (1) to make predictions, we must use data to estimate the coefficients. Let

$$(x_1, y_1), \quad (x_2, y_2), \ldots, \ (x_n, y_n)$$

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 $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model (1) fits the available data well - that is, so that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \ldots, n$. In other words, we want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ 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.

Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $Y$ based on the $i$th value of $X$. Then $e_i = y_i - \hat{y}_i$ represents the $i$th residual - this is the difference between the $i$th observed response value and the $i$th response value that is predicted by our linear model. We define the residual sum of squares (RSS) as

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

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \tag{3}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means. In other words, (4) defines the least squares coefficient estimates for simple linear regression.

Figure 3 displays the simple linear regression fit to the `Advertising` data, where $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.04753$. In other words, according to this approximation, an additional \$1,000 spent on `TV` advertising is associated with selling approximately 47.5 additional units of the product. In Figure 4, we have computed RSS for a number of values of $\beta_0$ and $\beta_1$, 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 squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$ given by (4). These values clearly minimize the RSS.

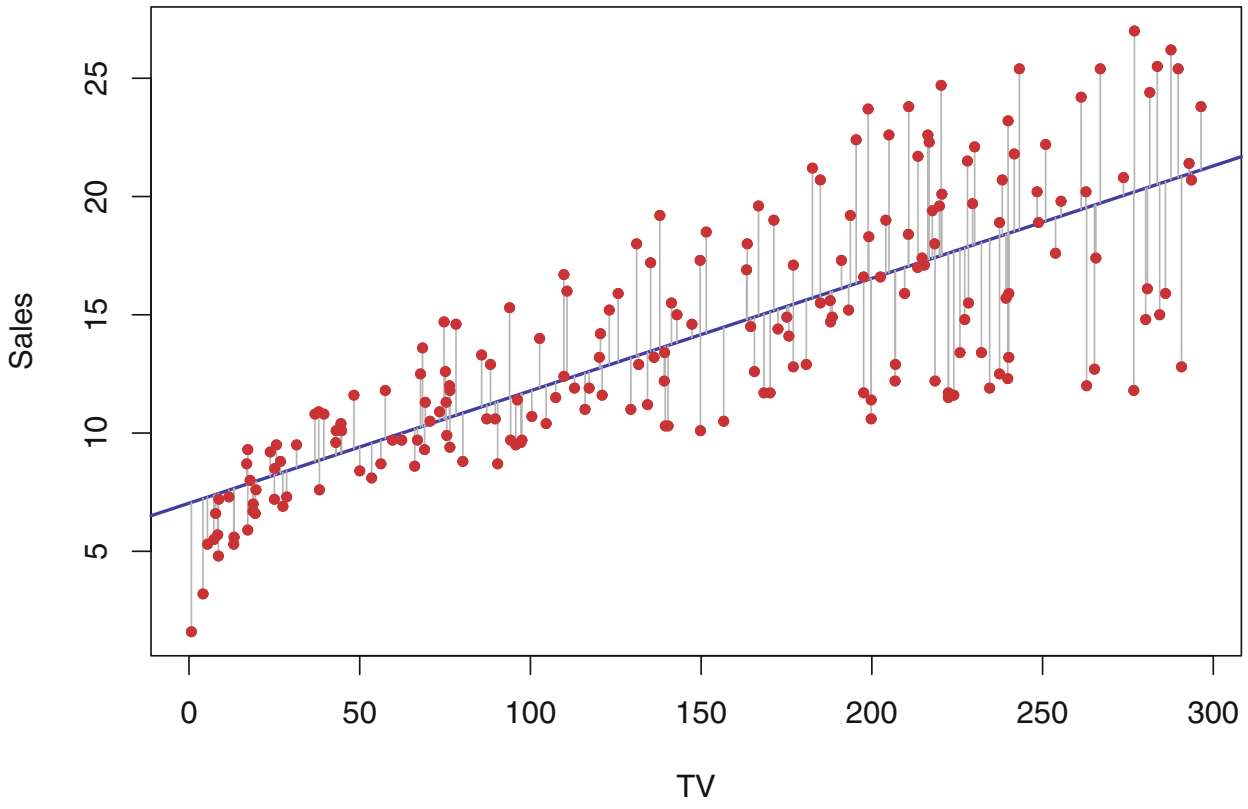


Figure 3: For the `Advertising` data, the least squares fit for the regression of `sales` onto `TV` is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

## 1.2 Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between $X$ and $Y$ takes the form

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad (5)$$

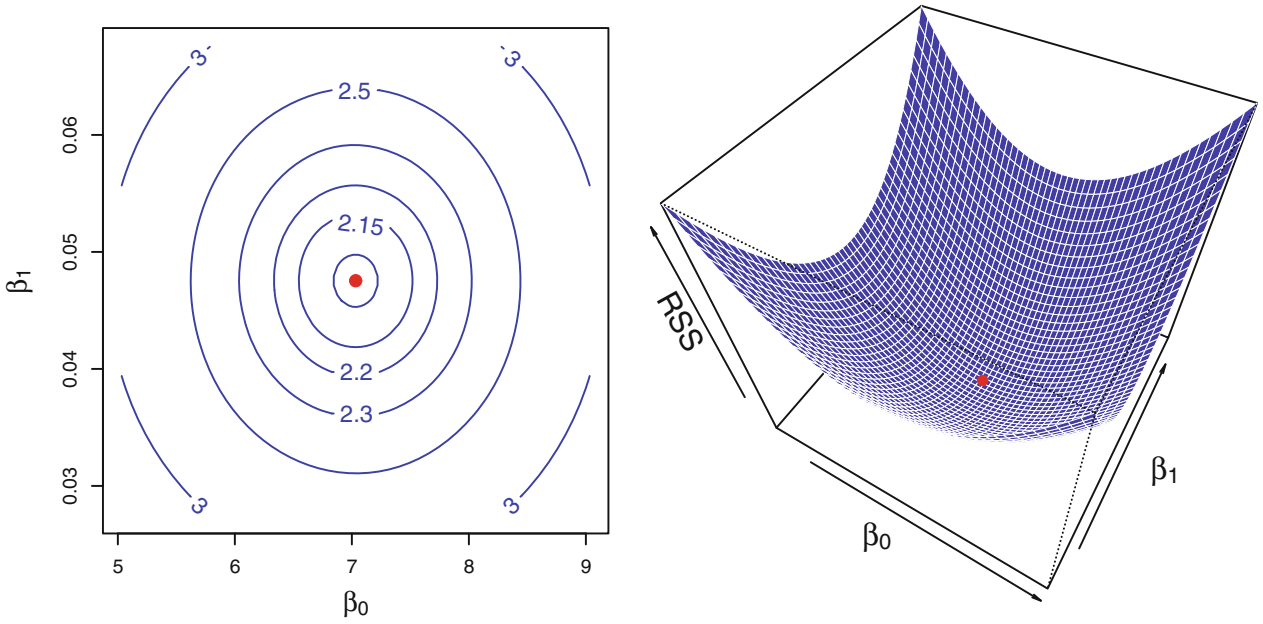


Figure 4: Contour and three-dimensional plots of the RSS on the `Advertising` data, using `sales` as the response and `TV` as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, given by (4).

where $\epsilon$ is a mean-zero random error term. Here $\beta_0$ is the intercept term - that is, the expected value of $Y$ when $X = 0$, and $\beta_1$ is the slope - the average increase in $Y$ associated with a one-unit increase in $X$. The error term is a catch-all for what we miss with this simple model: 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 (5) defines the population regression line, which is the best linear approximation to the true relationship between $X$ and $Y$. The least squares regression coefficient estimates (4) characterize the least squares line (2). The left-hand panel of Figure 5 displays these two lines in a simple simulated example. We created 100 random $X$s, and generated 100 corresponding $Y$s from the model

$$Y = 2 + 3X + \epsilon, \tag{6}$$

where $\epsilon$ was generated from a normal distribution with mean zero. The red line in the left-hand panel of Figure 5 displays the true relationship, $2 + 3X$, 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 given in (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 Figure 5 we have generated ten different data sets from the model given by (6) 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. 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 lines 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

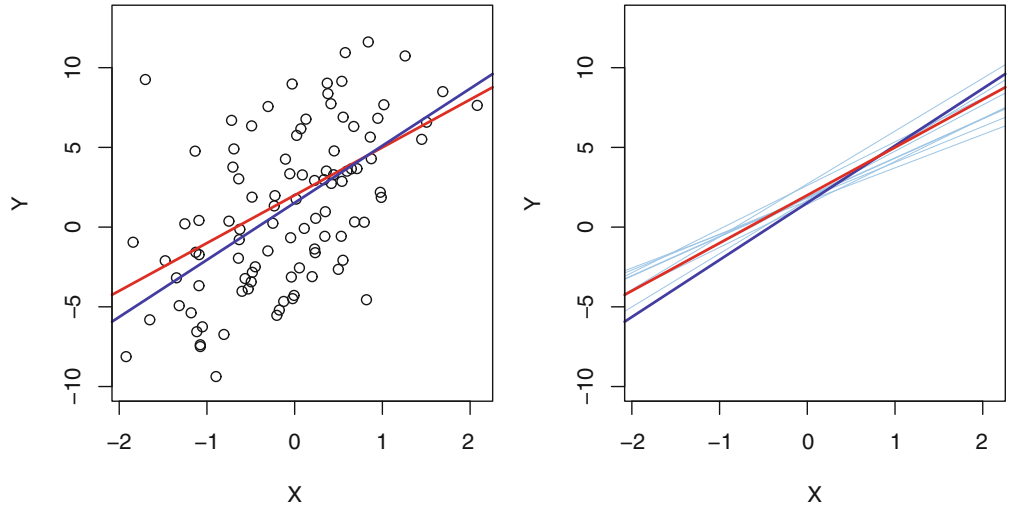


Figure 5: A simulated data set. Left: The red line represents the true relationship, $2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $2 + 3X$ based on the observed data, shown in black. Right: The population regression 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 lines are quite close to the population regression line.

in knowing the population mean $\mu$ of some random variable $Y$. Unfortunately, $\mu$ is unknown, but we do have access to $n$ observations from $Y$, which we can write as $y_1, \ldots, y_n$, and which we can use to estimate $\mu$. A reasonable estimate is $\hat{\mu} = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 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 the same way, the unknown coefficients $\beta_0$ and $\beta_1$ in linear regression define the population regression line. We seek to estimate these unknown coefficients using $\hat{\beta}_0$ and $\hat{\beta}_1$ in (4). These coefficient estimates 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 $\hat{\mu}$ to estimate $\mu$, this estimate is unbiased, in the sense that on average, we expect $\hat{\mu}$ to equal $\mu$. What exactly does this mean? It means that on the basis of one particular set of observations $y_1, \ldots, y_n$, $\hat{\mu}$ might overestimate $\mu$, and on the basis of another set of observations, $\hat{\mu}$ might underestimate $\mu$. But if we could average a huge number of estimates of $\mu$ obtained from a huge number of sets of observations, then this average would exactly equal $\mu$. Hence, an unbiased estimator does not systematically over- or under-estimate the true parameter. The property of unbiasedness holds for the least squares coefficient estimates given by (4) as well: if we estimate $\beta_0$ and $\beta_1$ on the basis of a particular data set, then our estimates won't be exactly equal to $\beta_0$ and $\beta_1$. 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 Figure 5 that the average of many least squares lines, each estimated from a separate data set, is pretty close to the true population regression line.

We continue the analogy with the estimation of the population mean $\mu$ of a random variable $Y$. A natural question is as follows: how accurate is the sample mean $\hat{\mu}$ as an estimate of $\mu$? We have established that the average of $\hat{\mu}$'s over many data sets will be very close to $\mu$, but that a single estimate $\hat{\mu}$ may be a substantial underestimate or overestimate of $\mu$. How far off will that single

estimate of $\hat{\mu}$ be? In general, we answer this question by computing the standard error of $\hat{\mu}$, written as $\text{SE}(\hat{\mu})$. We have the well-known formula

$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}, \tag{7}$$

where $\sigma$ is the standard deviation of each of the realizations $y_i$ of $Y$. Roughly speaking, the standard error tells us the average amount that this estimate $\hat{\mu}$ differs from the actual value of $\mu$. Equation (7) also tells us how this deviation shrinks with $n$ - the more observations we have, the smaller the standard error of $\hat{\mu}$. In a similar vein, we can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values $\beta_0$ and $\beta_1$. To compute the standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the following formulas:

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{8}$$

where $\sigma^2 = \text{Var}(\epsilon)$. For these formulas to be strictly valid, we need to assume that the errors $\epsilon_i$ for each observation are uncorrelated with common variance $\sigma^2$. Notice in the formula that $\text{SE}(\hat{\beta}_1)$ is smaller when the $x_i$ are more spread out. We also see that $\text{SE}(\hat{\beta}_0)$ would be the same as $\text{SE}(\hat{\mu})$ if $\bar{x}$ were zero (in which case $\hat{\beta}_0$ would be equal to $\bar{y}$). In general, $\sigma^2$ is not known, but can be estimated from the data. The estimate of $\sigma$ is known as the residual standard error, and is given by the formula $\text{RSE}=\sqrt{\text{RSS}/(n-2)}$. Strictly speaking, when $\sigma^2$ is estimated from the data we should write $\widehat{\text{SE}}(\hat{\beta}_1)$ to indicate that an estimate has been made, but for simplicity of notation we will drop this extra "hat".

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 of data. For linear regression, the 95% confidence interval for $\beta_1$ approximately takes the form

$$\hat{\beta}_1 \pm 2\text{SE}(\hat{\beta}_1). \tag{9}$$

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

$$\left[ \hat{\beta}_1 - 2\text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2\text{SE}(\hat{\beta}_1) \right] \tag{10}$$

will contain the true value of $\beta_1$. Approximately here is for several reasons. Equation (10) relies on the assumption that the errors are Gaussian. Also, the factor of 2 in front of the $\text{SE}(\hat{\beta}_1)$ term will vary slightly depending on the number of observations $n$ in the linear regression. To be precise, rather than the number 2, (10) should contain the 97.5% quantile of a t-distribution with $n-2$ degrees of freedom. Details of how to compute the 95% confidence interval precisely in `R` will be provided later in this chapter.

Similarly, a confidence interval for $\beta_0$ approximately takes the form

$$\hat{\beta}_0 \pm 2\text{SE}(\hat{\beta}_0). \tag{11}$$

In the case of the advertising data, the 95% confidence interval for $\beta_0$ is [6.130, 7.935] and the 95% confidence interval for $\beta_1$ is [0.042, 0.053]. Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 units. Furthermore, for each \$1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

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

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

Table 3: 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).

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

versus the alternative hypothesis

$$H_a : \text{There is some linear relationship between } X \text{ and } Y.$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model (5) reduces to $Y = \beta_0 + \epsilon$, and $X$ is not linearly associated with $Y$. To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for $\beta_1$, is sufficiently far from zero that we can be confident that $\beta_1$ is non-zero. How far is far enough? This of course depends on the accuracy of $\hat{\beta}_1$ - that is, it depends on $\text{SE}(\hat{\beta}_1)$. If $\text{SE}(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$, and hence that there is a linear relationship between $X$ and $Y$. In contrast, if $\text{SE}(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis. In practice, we compute a t-statistic, given by

$$\text{t} = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}.$$

If there really is no liner relationship between $X$ and $Y$, then we expect that t will have a t-distribution with $n - 2$ degrees of freedom. The t-distribution has a bell shape and for values of $n$ greater than approximately 30 it is quite similar to the normal distribution. Consequently, it is a simple matter to compute the probability of observing any number equal to |t| or larger in absolute value, assuming $\beta_1 = 0$. We call this probability the p-value.

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 a linear association between the predictor and the response. We reject the null hypothesis - that is, we declare a linear 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%. When $n = 30$, these correspond to t-statistics of around 2 and 2.75, respectively.

Table 3 provides details of the least squares model for the regression of number of units sold on `TV` advertising budget for the `Advertising` data. Notice that the coefficients for $\hat{\beta}_0$ and $\hat{\beta}_1$ are very large relative to their standard errors, so the t-statistics are also large; the probabilities of seeing such values if $H_0$ is true are virtually zero. Hence we can conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$. In Table 3, a small p-value for the intercept indicates that we can reject the null hypothesis that $\beta_0 = 0$, and a small p-value for `TV` indicates that we can reject the null hypothesis that $\beta_1 = 0$. Rejecting the latter null hypothesis allows us to conclude that there is a linear relationship between `TV` and `sales`. Rejecting the former allows us to conclude that in the absence of `TV` expenditure, `sales` are non-zero.

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

Table 4: For the `Advertising` data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

## 1.3 Assessing the Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, 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. Table 4 displays the RSE, the $R^2$ statistic, and the F-statistic (to be described in the next sections) for the linear regression of number of units sold on TV advertising budget.

### 1.3.1 Residual Standard Error

Recall from the model (5) that associated with each observation is an error term $\epsilon$. Due to the presence of these error terms, even if we knew the true regression line (i.e. even if $\beta_0$ and $\beta_1$ were known), we would not be able to perfectly predict $Y$ from $X$. The RSE is an estimate of the standard deviation of $\epsilon$. Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$\text{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2}\text{RSS}}, \tag{12}$$

where RSS is the residual sum of squares.

In the case of the advertising data, we see from the linear regression output in Table 4 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 $\beta_0$ and $\beta_1$ 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 3,260 units is an acceptable prediction error depends on the problem context. In the advertising data set, the mean value 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 (5) to the data. If the predictions obtained using the model are very close to the true outcome values - that is, if $\hat{y}_i \approx y_i$ for $i = 1, \ldots, n$ - then the RSE will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ is very far from $y_i$ for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

### 1.3.2 $R^2$ Statistic

The RSE provides an absolute measure of lack of fit of the model (5) 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, and is independent of the scale of $Y$.

To calculate $R^2$, we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{13}$$

where $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$ is the total sum of squares. TSS measures the total variance in the response $Y$, and can be thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, TSS-RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and $R^2$ measures the proportion of variability in $Y$ that can be explained using $X$. An $R^2$ statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error $\sigma^2$ is high, or both. In Table 4, 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 (13) has an interpretational advantage over the RSE (12), since unlike the RSE, it always lies 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 truly 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 (5) is at best an extremely rough approximation to the data, and 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.

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

$$\widehat{\text{Cor}}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{14}$$

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 assess 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 chapter 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 larger number of variables. We will see that $R^2$ fills this role.

## 1.4 Lab: Simple Linear Regression

### 1.4.1 Libraries

The `library()` function is used to load libraries, or groups of functions and data sets that are not included in the base `R` distribution. Basic functions that perform least squares linear regression and other simple analyses come standard with the base distribution, but more exotic functions require additional libraries. Here we load the `MASS` package, which is a very large collection of data sets and functions. We also load the `ISLR` package, which includes some of the data sets used in this notes.

```
library(MASS)
library(ISLR)
```

If you receive an error message when loading any of these libraries, it likely indicates that the corresponding library has not yet been installed on your system. Some libraries, such as `MASS`, come with `R` and do not need to be separately installed on your computer. However, other packages, such as `ISLR`, must be downloaded the first time they are used. This can be done directly from within `R`. For example, on a Windows system, select the `Install package` option under the `Packages` tab. After you select any mirror site, a list of available packages will appear. Simply select the package you wish to install and `R` will automatically download the package. Alternatively, this can be done at the `R` command line via `install.packages("ISLR")`. This installation only needs to be done the first time you use a package. However, the `library()` function must be called each time you wish to use a given package.

### 1.4.2 Simple Linear Regression I

The `MASS` library contains the `Boston` data set, which records `medv` (median house value) for 506 neighborhoods around Boston. We will seek to predict `medv` using 13 predictors such as `rm` (average number of rooms per house), `age` (average age of houses), and `lstat` (percent of households with low socioeconomic status).

```
fix(Boston)
names(Boston)
```

To find out more about the data set, we can type `?Boston`. We will start by using the `lm()` function to fit a simple linear regression model, with `medv` as the response and `lstat` as the predictor. The basic syntax is `lm(y ~ x, data)`, where `y` is the response, `x` is the predictor, and `data` is the data set in which these two variables are kept.

```
lm.fit <- lm(medv ~ lstat)
```

The command causes an error because `R` does not know where to find the variables `medv` and `lstat`. The next line tells `R` that the variables are in `Boston`. If we attach `Boston`, the first line works fine because `R` now recognizes the variables.

```
lm.fit <- lm(medv ~ lstat, data = Boston)
attach(Boston)
lm.fit <- lm(medv ~ lstat)
```

If we type `lm.fit`, some basic information about the model is output. For more detailed information, we use `summary(lm.fit)`. This gives us p-values and standard errors for the coefficients, as well as the $R^2$ statistic and F-statistic for the model.

```
lm.fit
summary(lm.fit)
```

We can use the `names()` function in order to find out what other pieces of information are stored in `lm.fit`. Although we can extract these quantities by name - e.g. `lm.fit$coefficients` - it is safer to use the extractor functions like `coef()` to access them.

```
names(lm.fit)
coef(lm.fit)
```

In order to obtain a confidence interval for the coefficient estimates, we can use the `confint()` command.

```
confint(lm.fit)
```

The `predict()` function can be used to produce confidence intervals and prediction intervals for the prediction of `medv` for a given value of `lstat`.

```
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
        interval = "confidence")
```

```
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
         interval = "prediction")
```

For instance, the 95% confidence interval associated with a lstat value of 10 is (24.47, 25.63), and the 95% prediction interval is (12.83, 37.28). As expected, the confidence and prediction intervals are centered around the same point (a predicted value of 25.05 for medv when lstat equals 10), but the latter are substantially wider.

We will now plot `medv` and `lstat` along with the least squares regression line using the `plot()` and `abline()` functions.

```
plot(lstat, medv)
abline(lm.fit)
```

There is some evidence for non-linearity in the relationship between `lstat` and `medv`. The `abline()` function can be used to draw any line, not just the least squares regression line. To draw a line with intercept `a` and slope `b`, we type `abline(a, b)`. Below we experiment with some additional settings for plotting lines and points. The `lwd = 3` command causes the width of the regression line to be increased by a factor of 3; this works for the `plot()` and `lines()` functions also. We can also use the `pch` option to create different plotting symbols.

```
abline(lm.fit, lwd = 3)
abline(lm.fit, lwd = 3, col = "red ")
plot(lstat, medv, col = "red ")
plot(lstat, medv, pch = 20)
plot(lstat, medv, pch = "+")
plot(1:20, 1:20, pch = 1:20)
```

Next we examine some diagnostic plots, (these diagnostics will be explained in detail in the next chapter). Four diagnostic plots are automatically produced by applying the `plot()` function directly to the output from `lm()`. In general, this command will produce one plot at a time, and hitting *Enter* will generate the next plot. However, it is often convenient to view all four plots together. We can achieve this by using the `par()` function, which tells `R` to split the display screen into separate panels so that multiple plots can be viewed simultaneously. For example, `par(mfrow=c(2,2))` divides the plotting region into a $2 \times 2$ grid of panels.

```
par(mfrow = c(2,2))
plot(lm.fit)
```

Alternatively, we can compute the residuals from a linear regression fit using the `residuals()` function. The function `rstandard()` will return the standardised residuals, and we can use this function to plot the residuals against the fitted values.

```
plot(predict(lm.fit), residuals(lm.fit))
plot(predict(lm.fit), rstandard(lm.fit))
```

On the basis of the residual plots, there is some evidence of non-linearity. Leverage statistics can be computed for any number of predictors using the `hatvalues()` function.

```
plot(hatvalues(lm.fit))
which.max(hatvalues(lm.fit))
```

The `which.max()` function identifies the index of the largest element of a vector. In this case, it tells us which observation has the largest leverage statistic.

### 1.4.3 Simple Linear Regression II

In Sunflowers Apparel (a chain of upscale clothing stores for women) the business objective of the director of planning is to forecast annual sales for all new stores, based on store size. To examine the relationship between the store size in square feet and its annual sales, data were collected from a sample of 14 stores. Figure 6 displays the scatter plot for the Sunflowers Apparel data. Observe the increasing relationship between square feet $X$ and annual sales $Y$. As the size of the store increases, annual sales increase approximately as a straight line. Thus, you can assume that a straight line provides a useful mathematical model of this relationship. Now we need to determine the specific straight line that is the best fit to these data.

Using `R`, we obtain:

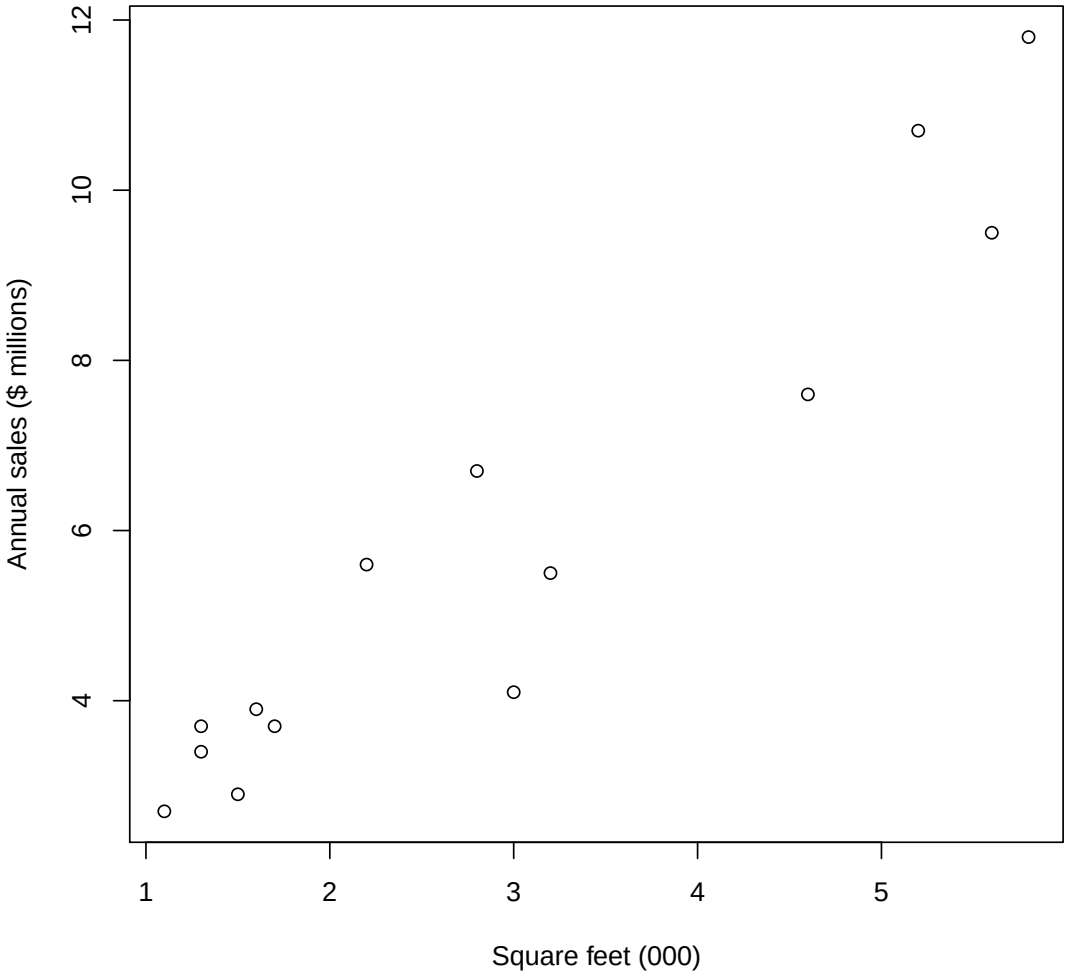


Figure 6: Scatter plot for the Sunflowers Apparel data.

```
site <-read.table("site.txt", header = T)

plot(site$Square_Feet, site$Annual_Sales, xlab = "Square feet (000)",
     ylab = "Annual sales ($ millions)")


site.lm <- lm(Annual_Sales ~ Square_Feet, data = site)
summary(site.lm)
```

The estimated coefficients are: $\hat{\beta}_0 = 0.9645$ and $\hat{\beta}_1 = 1.6699$ and the prediction line:

$$\hat{y}_i = 0.9645 + 1.6699 x_i.$$

The value of the slope means that for each increase of 1 unit in $x$, the average value of $Y$ is estimated to increase by 1.6699 units. In other words, for each increase of 1.0 thousand square feet in the size of the store, the average annual sales are estimated to increase by 1.6699 millions of dollars. (The slope is significantly different from zero.)

The value of the intercept is the predicted value of $Y$ when $x$ equals 0. Because the square footage of the store cannot be 0, this intercept has little or no practical interpretation. Also, the intercept for this example is outside the range of the observed values of the $X$ variable, and therefore interpretations of the value of should be made cautiously.

The coefficient of determination is 90.42%. This means that 90.42% is the variation in annual sales explained by the variability in the size of the store as measured by the square footage. This large value indicates a strong linear relationship between these two variables because the regression model has explained 90.42% of the variability in predicting annual sales.

When using a regression model for prediction purposes, we should consider only the relevant range of the independent variable in making predictions. This relevant range includes all values from the smallest to the largest $x$ used in developing the regression model. Hence, when predicting $Y$ for a given value of $x$, you can interpolate within this relevant range of the $x$ values, but you should not extrapolate beyond the range of $x$ values. When one uses the square footage to predict annual sales, the square footage (in thousands of square feet) varies from 1.1 to 5.8. Therefore, one should predict annual sales only for stores whose size is between 1.1 and 5.8 thousands of square feet.

For example, we could predict annual sales when store size is equal to 5.0 thousands of square feet

```
x <- data.frame(Square_Feet = 5.0)
predict(site.lm, x)
```

The predicted value for annual sale is 9.313785 millions of dollars. Any prediction of annual sales for stores outside this range assumes that the observed relationship between sales and store size for store sizes from 1.1 to 5.8 thousand square feet is the same as for stores outside this range. For example, you cannot extrapolate the linear relationship beyond 5,800 square feet.

To assess the assumptions of the model (see next chapter), we can have a look at the residual plots. The `R` command used to obtain the residual plots is `plot()`:

```
par(mfrow=c(2,2))
plot(site.lm)
```

The top left plot of Residuals vs Fitted, shows a slight indication of increasing variance with meanfitted values. However it does not seem to be a concern (this means that the constant variance assumption holds here). The other feature to look for is a pattern in the average value of the residuals as the fitted values change. The solid curve shows a running average of the residuals to help judging this: there is a slight pattern here, which is not extreme however. The lower left plot shows the square root of the absolute value of the standardized residuals against fitted value (again with a running average curve). If all is well the points should be evenly spread with respect the vertical axis here, with no trend in their average value. A trend in average value is indicative of a problem with the constant variance assumption, and is not a concern in this case. The top right plots the ordered standardized residuals against quantiles of a standard normal: a systematic deviation from a straight line is not present here indicating no departure from normality in the residuals. The lower right plot is looking at leverage and influence of residuals, by plotting standardized residuals against a measure of leverage. A combination of high residuals and high leverage indicates a point with substantial influence on the fit. A standard way of measuring this is via Cook distance (which measures the change in all model fitted values on omission of the data point in question). It turns out that Cook distance is a function of leverage and the standardized residuals, so contours of Cook distance values are shown on the plot. Cook distances over 0.5 are considered borderline problematic, while over 1 is usually considered highly influential, so points to the right of these contours warrant investigation. Although a couple of points have rather high leverage, their actual influence on the fit is not unduly large.

Finally, using the `R` code

```
sres<- rstandard(site.lm)
plot(site$Square_Feet, sres, xlab="Square feet", ylab="Standardized residuals")
```

we show plots of the standardized residuals against the explanatory variable. There are no obvious patterns against it, so that the adequacy of the model is not questioned.