Linear Models

The Basics of a Linear Model

Motivation

We consider the situation where there is one response variable, and it is continuous. Further, we will assume that the response variable, or some function of it, is *linearly related* (we will discuss later the definition of linearly related) to the explanatory variable(s). With this setup, it is common to use a ***Linear Model*** or ***Linear Regression*** to model the relationship between the response and explanatory variable(s). If there is only one explanatory variable, then the modelling process is referred to as ***simple linear regression***; when there are two or more explanatory variables, then the modelling process is referred to as ***multiple linear regression***. Generally, there is no restriction placed on the explanatory variables: they can be continuous, ordinal, or nominal. However, in this introduction, we will temporarily assume that they are continuous for the purpose of illustration.

As a simple example, consider the following scatter plot:

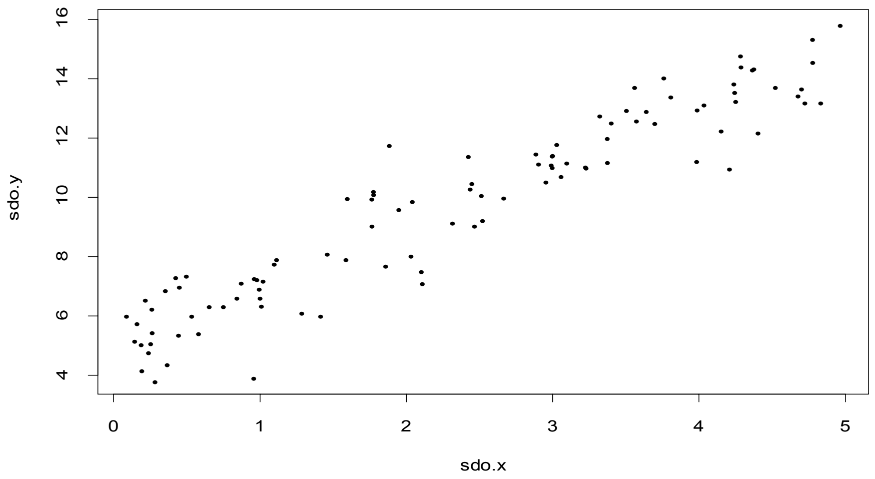


Figure 1: *Scatterplot for continuous response variable vs. continuous explanatory variable*

We observe from this scatter plot that there is a strong positive correlation between the variable *sdo.x* (the explanatory variable) and *sdo.y* (the response variable). It also looks like it would be possible to draw a straight line through these points which would summarise the relationship between *sdo.x* and *sdo.y*, although we are not exactly sure what the best straight line would be. This goal of finding or estimating the best fitting line for a particular set of variables is, in essence, what drives linear regression.

It is worth noting that all the points in the above scatter plot do not fall exactly on a straight line. If they did, we would not have the problem of deciding which was the best line; estimation of the line would be simple (basic mathematical theory tells us how to estimate a straight line when we have at least two points). However, statistics (as opposed to just mathematics) comes into play in linear regression because, in real life, we never observe processes that give exact answers even if the true (unobserved) relationship is an exact one; in real life, we always have some error associated with the process. For this reason, a linear model assumes that there is an exact (but unknown) line between the *mean* of the response variable and the explanatory variable, but that this relationship is not exact (i.e., there is some error in it) for any particular observation of the response variable and explanatory variable. Thus, linear regression is our attempt to estimate this unknown linear relationship for the underlying *population* by finding the line that best fits the observations in our *sample* data.

This idea of trying to find the line that best describes the relationship between the response variable and one explanatory variable can be extended to the situation where there is more than one explanatory variable. For instance, suppose that there are two explanatory variables, and that we have recorded 100 observations of the response and two explanatory variables. Figure 2 demonstrates how we might picture these observations in 3D space. With two explanatory variables, linear regression amounts to finding the best fitting *plane* through the scatter of points, again assuming that there is a true (but unknown) plane that exactly describes the relationship between the response variable and the two explanatory variables.

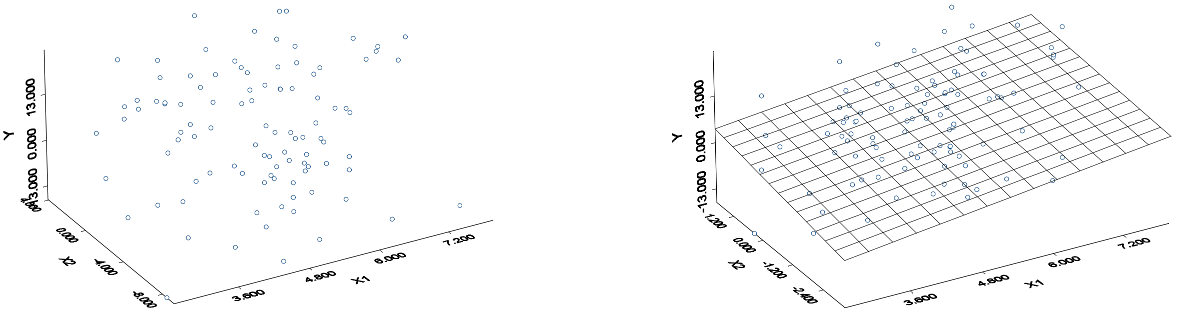


Figure 2: *Linear regression for two (continuous) explanatory variables*

We now have the basic idea of linear regression: fitting lines or planes to scatters of data in an attempt to estimate the true but unknown underlying line or plane that summarises the relationship between the explanatory variable(s) and the response variable. However, we now need to consider how to use linear regression in real life examples and also how to analyse the output of regression functions in various statistical packages (e.g. SAS, SPSS, Stata, S-Plus, Microsoft Excel).

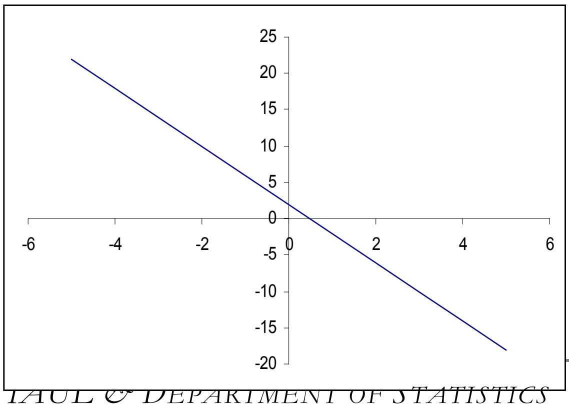
Setup

It would be impossible to discuss linear regression without using at least a little bit of mathematics. First, we simply write down what it means for a set of explanatory variables to be linearly related to a response variable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon . \qquad (1)$$

This means that a linear combination of the *X*'s can be used as a predictor of the response variable, *Y*. The term β_0 in this linear combination is often referred to as the ***intercept***, whereas β_1 , ..., β_p are often referred to as the ***slopes***; together, the *β*s are referred to as the regression ***coefficients*** or regression ***parameters***. As we will see below, the regression parameters allow us to make quantitative statements about the effect that the *X*s have on *Y*, particularly about how a change in each *X* variable will effect *Y*. The term *ε* is often referred to as the ***error term*** and relates to the error which was discussed above; if *ε* were 0, then the straight line would be uniquely and exactly determined, but as indicated, this is unlikely to be the case with actual data. Here, we will *assume* that the only error in the model has to do with the response variable: we will assume that the *X* variables are measured without error (***Assumption 1***).

Before proceeding, we should note that, in the above equation, *Y* may actually be a (known) function (e.g., log) of the outcome variable in its original form; similarly, any of the explanatory variables may be a (known) function of one or more of the design variables or covariates (in their original form), such as Age^2 or Age*Smoking. However, even if *Y* is actually the log of the original outcome variable or if X_1 is actually the square of one of the original covariates, we still refer to this equation as linear. This is because *linear* refers to the coefficients rather than to the variables in Equation (1): for an equation to be linear in the coefficients, each coefficient must be in its original form (not squared, cubed, etc...) and no additive term can involve more than one coefficient. Further, even if the response and/or explanatory variables are defined as functions of the data set's original variables, Equation (1) is linear in terms of these *newly defined* variables (although it may not be linear in terms of the original response variable and original covariates and design variables). The relationship between these newly defined variables might be summarised by the relationships



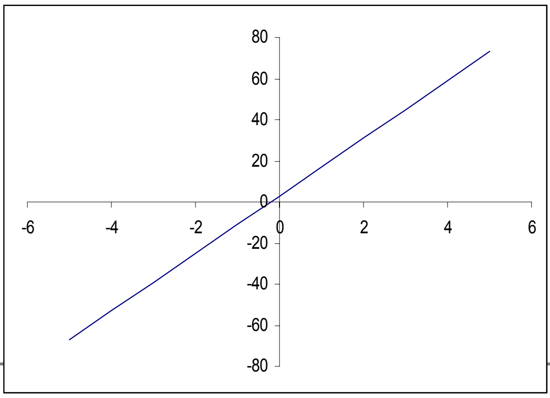


***Figure 3a:** Two linear relationships between a response variable and an explanatory variable*

but not by either of the following relationships

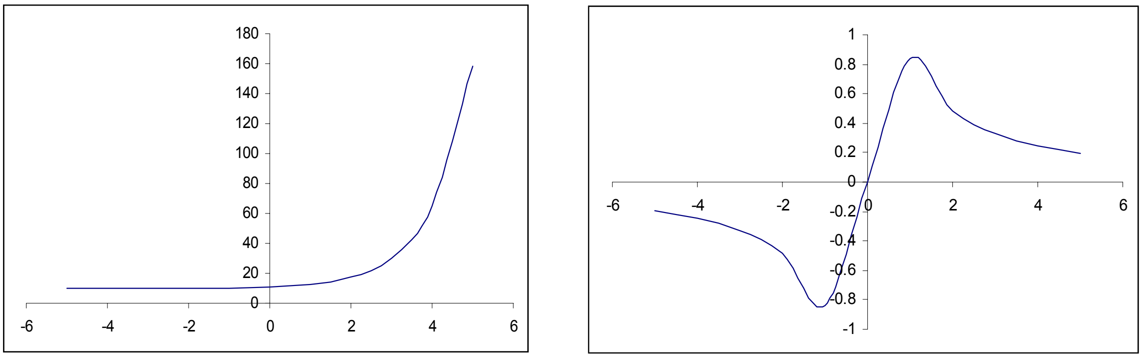


Figure 3b*: Two non-linear relationships between a response variable and an explanatory variable*

If we were to make a scatterplot of the newly defined response variable vs. the newly defined explanatory variables, then the scatter of points should appear linear. For instance, suppose that *Y* in Equation (1) is actually the log of the original outcome variable (called *Y′*) and that there is only one explanatory variable, X_1. Well, a plot of *Y*=log(*Y′*) versus X_1 should appear linear, although a plot of the outcome variable on its original scale, *Y′*, versus X_1 would not.

Returning to Equation (1), we will make the assumption that it *correctly specifies* the relationship between *Y* and *X* (i.e., that the relationship between *Y* and the explanatory variables is truly linear and that all explanatory variables that have an effect on *X* have been included in the equation). Thus, we *assume* that ε has mean zero (***Assumption 2***); further, we *assume* that ε is normally distributed (***Assumption 3***) with a constant (but unknown) variance, σ^2 (***Assumption 4***). Because of our assumption that ε has a zero mean, the conditional mean of *Y* (i.e., conditional on X) can be written as a linear combination of the *X* variables:

$$E[Y \mid X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p .^1 \qquad (2)$$

Further, as a result of our constant variance and normality assumptions, we can say that, conditional on *X*, *Y* is normally distributed with a variance of σ^2 (for this reason, we will often refer to σ^2 as the ***conditional variance***). In other words, these three assumptions about ε are equivalent to saying that there is a normally distributed subpopulation of responses for each combination of the explanatory variables, that the means of the subpopulation fall on a straight-line (or a straight-plane) function of the explanatory variables, and that the subpopulation variances are all equal to σ^2. Figure 4 below is a picture of these three statements for an example in which the 'number of flowers per plant' (*Y*) is a linear combination of the 'light intensity' (X_1) to which the

[1] Comparing Equation (2) above to Equation (1) in the "Overview of Modelling" lecture, which expresses the general form of most statistical models, we see that a linear model is a special case of the general model in which the function *η()* is the identity function.

plant was exposed while growing and the 'time' (X_2) at which the plant was exposed to the light:

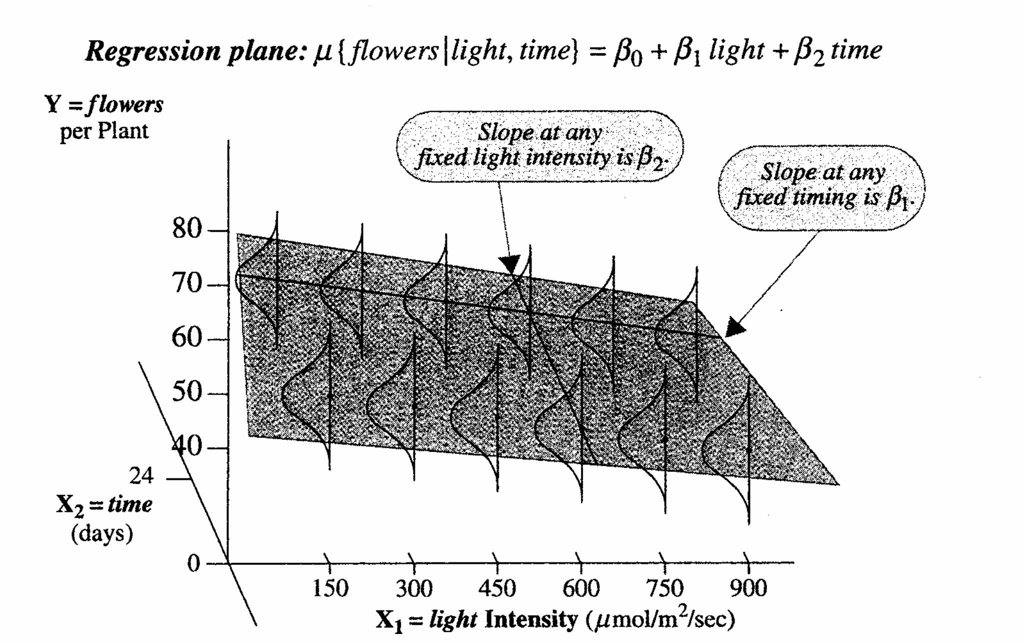

Regression plane: μ{flowers|light, time} = β0 + β1 light + β2 time
Y = flowers per Plant
Slope at any fixed light intensity is β2.
Slope at any fixed timing is β1.
80
70
60
50
40
24
X2 = time (days)
0
150
300
450
600
750
900
X1 = light Intensity (μmol/m²/sec)

Figure 4: Visual representation of the ε ~ N(0, σ²) assumption for a linear model of Light Intensity and Time as predictors for Flowers per plant (taken from Ramsey and Schafer, 2001)

For a particular unit (say unit *i*), the above equations are written as

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \varepsilon_i \tag{3a}$$

and

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}\,, \tag{3b}$$

where Y_i is the value of the response variable for unit *i*, ε_i is the error term for that unit, and $X_{1,i}, \ldots, X_{p,i}$ are the values of the explanatory variables for unit *i*. Since our data set contains *n* units from the underlying population of interest, we have a realisation of the response variable for *n* (possibly non-unique) combinations of the explanatory variables. Here, we will state a fifth assumption, which is that the units are independent of each other: more specifically, we assume that the error terms for the units are independent of each other (in a statistical sense) (***Assumption*** 5).

Treatment of Non-continuous Explanatory Variables

As mentioned in the introduction, the explanatory variables in a linear model need not be continuous, but can be nominal or ordinal.

If a particular explanatory variable is nominal, then the matter of its treatment is straightforward: the nominal variable should be treated as a ***factor***. If a nominal variable with *k* different levels is treated as a factor, *k – 1* different terms should be included in the regression equation. There are a number of different ways in which these *k – 1* terms might be *parameterised*, but the ***treatment parameterisation*** is a commonly used and intuitive parameterisation (this is the default, and perhaps only, option in SPSS). In this parameterisation, one level of the variable is treated as a ***baseline level*** or ***reference level***, and the nominal variable is replaced by *k – 1* ***indicator variables***, each of which corresponds to one of the non-baseline levels and takes a value of 1 for units that have that particular non-baseline level and 0 for units that don't. In the treatment parameterisation, every level except for the baseline level has its own regression slope. For instance, if we want to include the variable Country of Origin (with levels USA, UK, France, and Germany) in the linear model and we decide to use a treatment parameterisation with USA as the baseline level, then the Country of Origin variable is replaced with three indicator variables (X_{UK}, X_{FR}, and X_{GE}), which indicate origins in the UK, France, and Germany, respectively. Then, the UK, France, and Germany levels would each have their own slope (β_{UK}, β_{FR}, and β_{GE}, respectively). The terms $\beta_{UK} X_{UK} + \beta_{FR} X_{FR} + \beta_{GE} X_{GE}$ are then included in Equation (1).

If a particular explanatory variable is ordinal, then its treatment is a more complicated question. One option is to pretend that the variable is actually nominal and to ignore the ordering of the variable's levels: if this is done, then the aforementioned *factor approach* may be adopted and *k - 1* terms added to the linear model. At the other end of the spectrum, another option is adopt a *linear approach* and pretend that the ordinal variable is actually continuous by assigning a number to each of its levels (e.g., *1* for the lowest level, *2* for the second lowest level, . . ., and *k* for the highest level). Note that the factor option means that there will be *k – 1* regression slopes pertaining to the ordinal variable, but in the second option, there will be only one slope pertaining to the variable. To see the difference between these two extreme approaches, consider the ordinal variable Position (with levels Analyst, Associate, Vice President, and President) in a linear model where Salary is the response variable. If the factor approach is adopted (with a treatment parameterisation and Analyst as the baseline level), then the difference between Associates' and Analysts' salaries, the difference between Vice Presidents' and Analysts' salaries, and the difference between Presidents' and Analysts' salaries may all be different. On the other hand, if the linear approach is adopted with the four levels numbered *1*, *2*, *3*, and *4*, then the difference between Vice Presidents' and Analysts' salaries will be forced to be twice the difference between Associates' and Analysts' and the difference between Presidents' and Analysts' salaries will be three times the difference between Associates' and Analysts' salaries.

The treatment of an ordinal explanatory variable can be addressed more elaborately using ***orthogonal polynomials***. A discussion of this topic is beyond the scope of this course. However, let us simply state that for a variable with *k* ordered levels, orthogonal polynomials with degrees anywhere between *1* and *k – 1* can be used: at the two ends of this spectrum, using an orthogonal polynomial with degree *1* is identical to the linear approach, and using an orthogonal polynomial with *k – 1* degrees is identical to the factor approach.

Model Parameters

In Equations (1) and (2) above, the regression coefficients, or βs, pertain to the *population*: in other words, they belong to the true but unknown line/plane that summarises the relationship between the response and explanatory variables. By estimating the βs from our data, we are finding the line/plane that "best" fits our data points. Hopefully, this line/plane is a good representation of the true unknown line/plane; identically, our estimates of the βs, which we will refer to as $\hat{\beta}$ and which pertain to our *data sample*, are hopefully a good representation of the βs.

When the $\hat{\beta}$s replace the βs in Equation (3b), the result is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_p X_{p,i}, \tag{4}$$

where $\hat{Y}_i$ is the (estimated) conditional (on unit *i*'s explanatory variable values) predicted *Y* value for unit i. The difference between the observed *Y* value for unit *i* and this predicted *Y* value can be thought of as an estimate of the error term for that unit and is often referred to as the ***residual***, e_i, for unit *i:*

$$\hat{\varepsilon}_i = e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_p X_{p,i}). \tag{5}$$

Estimation

There are a number of different ways in which we might estimate the βs from our data. However, one of the most intuitive and popular approaches is referred to as ***least squares***: in least squares estimation, we seek to select $\hat{\beta}$s that minimise the sum of the squared residual terms for the *n* units in our sample. In other words, we seek, in some sense, to make the estimated error terms as small as possible. This procedure can be visualised for the scenario in which we have only one explanatory variable, as in Figure 5: essentially, finding the least squares estimates of the βs amounts to finding the line where the sum of the squared vertical distances between the points and that line are smallest.

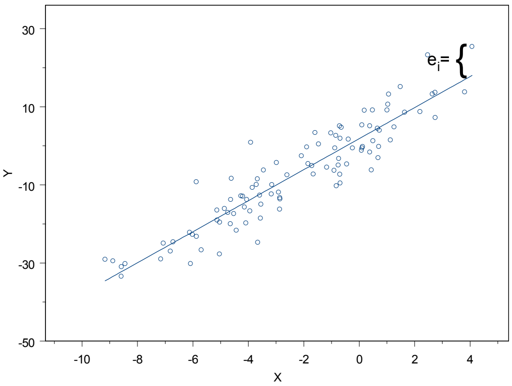


Figure 5: *The geometry of least squares regression, which seeks to minimise the sum of the e_i^2s*

If the aforementioned five error assumptions are satisfied, then the least squares estimates of the βs have a number of nice properties. To be more specific, if the regression equation is, in fact, correctly specified (and the error terms have mean zero), then the least squares estimates are unbiased estimates of the true βs. Further, if the conditional variance is indeed always equal (to σ^2) and the errors are truly independent of each other, as assumed, then the least squares estimates are the 'best' (linear) unbiased estimates, in the sense that they have the minimum possible amount of variance (from the true βs). Lastly, if the assumption about the normality of the errors is true, then the least squares estimates of the parameters will have a normal sampling distribution (around their true values). (Here, it is important to recall that the sampling distribution of a particular estimator is a property of the estimator across all the samples that could have been drawn from the underlying population.)

As a final note, under the assumption of normally distributed error terms, the least squares estimates of the βs are identical to the ***maximum likelihood estimates*** for these parameters (if you are unfamiliar with maximum likelihood estimation, don't worry).

Example 1

Now that we know something about what a linear model is and how its parameters are estimated, we will consider an example. Consider the following, fairly simplistic, data set that we will use as a running example. In this example of a simple (one explanatory variable) linear model, we are trying to relate the age of Toyota Land Cruisers to their Mileage[2]:

Age **(*X*)**	**Mileage in 1000 of miles** **(*Y*)**
3	68.22
5	59.53
5	73.43

[2] This data was taken from a lecture course given by Charles Moser of Clarkson College, USA who collected it from the WWW.

5	82.55
6	67.67
6	93.61
9	86.94
10	120.94
11	112.27
14	129.42
15	38.22

Before fitting a linear model to the data, we should always examine a scatterplot (or scatterplots if there is more than one explanatory variable) of the data:

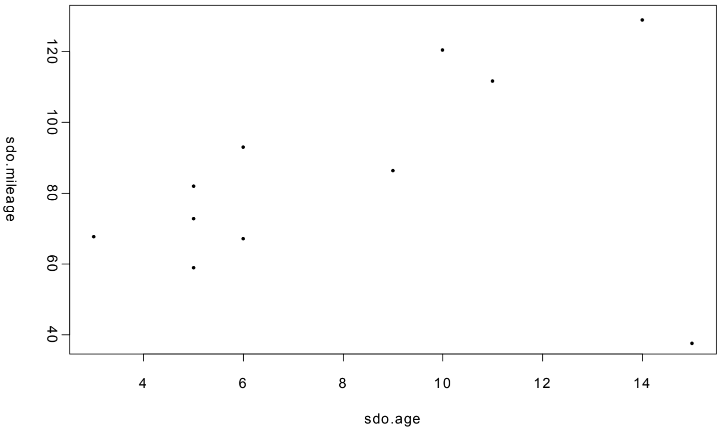


Figure 6: *Accumulated Mileage vs. Age for 11 Toyota Land Cruisers*

We observe that there appears to be a strong linear relationship between the Age of the vehicle and its Mileage, excluding the observation in the lower right which we might consider to be an ***outlier***. (An outlier is an observation that doesn't appear to fit into the general pattern exhibited by the rest of the data; we will discuss outliers in greater detail later.) It would therefore appear to be reasonable to fit this data with a linear model of the form

$$\text{Mileage} = \beta_0 + \beta_1 \text{ x Age}$$

where, as usual, β_0 and β_1 are referred to as the regression coefficients or parameters. If we fit this model using a statistical package, we obtain $\hat{\beta}_0 = 68.78$ and $\hat{\beta}_1 = 1.98$ ■

Standard Errors for Parameter Estimates

Least squares gives us the best possible estimates for the regression coefficients in the sense that they define a line/plane that minimises the sum of the squared residuals for the *n* units. Further, if the error assumptions are satisfied, then the least squares estimates are also best in the sense that they are the linear unbiased estimators with the least variance (around the true population values). However, since our *n* units are but a sample from the underlying population whose true regression coefficients we are trying to estimate, it is unlikely that the estimates will be equal to their true values (particularly for small samples!). Therefore, for every regression parameter that is estimated, there is an associated ***standard error*** for its estimate. Most regression functions in software packages will produces an estimate of the standard error for each parameter; we will refer to the estimated standard error pertaining to coefficient β_j as $se(\hat{\beta}_j)$.

Although we will not describe how to calculate the standard error for a given parameter, we will say that the standard error for a *slope* parameter tends to decrease as the variance of the corresponding explanatory variable increases, that it tends to decrease the better the linear model fits the data (as the sum of squared residuals decreases), and that it tends to decrease as the correlation between the corresponding explanatory variable and the other explanatory variables decreases. Lastly, standard errors of intercept and slope parameters tend to decrease as the sample size, *n*, increases. (Try to think about why each of these statements is intuitively true.)

These estimated standard errors can be used to gain some feel for how variable the estimates of the coefficients are. If a standard error is very large, then we would be less certain that the estimate is a good representation of the parameter's true value.

Example 1 (continued)

For the car data example, we obtain the following error estimates:

```
Coefficients:
              Value    Std. Error
          β0  68.7769    19.8070
          β1   1.9804     2.2158
```

This, in effect, says that the variability of β_0 is moderate (variability of regression coefficients must be viewed in relation to their magnitude) and that variability of β_1 is very high ■

Hypothesis Tests for Parameters

We now have all the parts that we need to test hypotheses about the regression parameters.

Test for Individual Parameters

Suppose you want to test whether one single model parameter (say β_j) is equal to a particular value (say *f*). The most common hypothesis to test is that $\beta_j = 0$ since, for a slope parameter, this is essentially a statement about the strength of the corresponding explanatory variable: if the parameter value is *0*, then this suggests that a model without that particular explanatory variable would fit equally well. More intuitively, in a simple linear model with one explanatory variable, if the regression slope were *0*, this would say that a linear relationship between *Y* and X_1 simply does not exist or, equivalently, that the plain sample mean of the response variable would be as good a predictor of the response variable. If a regression slope is shown to be non-zero (i.e., we reject the null hypothesis that it is zero at the *a*% level), then we say that the slope is *statistically significant* and that there is a statistically significant relationship between the response variable and the corresponding explanatory variable. If the null hypothesis is not rejected, we may want to consider taking the corresponding explanatory variable out of the model since it does not appear to have an effect on the response variable.

A formal statement of this particular hypothesis test is as follows:

$$\mathrm{H}_0 : \beta_j = 0$$
$$\mathrm{H}_1 : \beta_j \neq 0.$$

If the five error assumptions that we made are valid, then the following test statistic, which can be used to test the above hypotheses, will have a *t* distribution with $n - (p{+}1)$ degrees of freedom (here, we subtract $p + 1$ since we are fitting $p + 1$ regression parameters—*p* slopes and *1* intercept):

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} .$$

Hence, you will often find that the output from the regression function of a statistical package will give you something called a *t*-value along with each regression coefficient estimate: this *t-value* is the value of the test statistic given above. A number of packages may also give you the p-value for this test but, if not, then you can use look-up tables.

Example 1 (continued):

Returning to our example on the relationship between Age and Mileage of Toyota Land Cruisers, we obtain the following values after fitting a linear model:

```
Coefficients:
              Value     Std. Error  t value     p-value
          β0  68.7769   19.8070     3.4724      0.0070
          β1  1.9804    2.2158      0.8938      0.3947
```

We are interested in a test of the hypothesis that $\beta_1 = 0$. As commented above, the standard error for this coefficient is relatively large (compared to β_1's estimated value), and therefore we are somewhat unsure that the estimated coefficient is very reliable. If we also look at the *t* statistic, we see that it is 0.89 and its corresponding p-value is 0.3947. This implies that we are not willing to reject the null hypothesis that $\beta_1 = 0$, which says that Age may not be a useful predictor of Mileage ■

A Warning about Multiple Tests

The computer output will typically contain a *t-value* and, possibly, a corresponding p-value for every regression slope. However, in the case where there is a large number of explanatory variables, one must be careful when using these values since multiple tests are being performed at the same time. Essentially, even if none of the explanatory variables in the linear model has an effect on the response variable in truth, if there is a large number of explanatory variables, chances are that at least one of the regression slopes will be (spuriously) significant, leading us to believe that the corresponding explanatory variable actually does have an effect on the response variable. This problem is an artefact of the way in which statistical hypothesis tests are performed. One way around this problem is to decide, before looking at the regression output, on a small number (ideally, *1*) of regression slopes for which you are interested in testing the hypothesis that it equals zero, and then to perform the above test for only these slopes. However, if you are actually interested in testing the above null hypothesis for each explanatory variable slope, then another solution is to use one of the procedures that corrects for multiple tests.

A Note Regarding Categorical Explanatory Variables

In the above, we tested whether a particular regression slope is *0*, which is equivalent, in many cases, to testing whether the corresponding explanatory variable should be in the linear model. However, for an explanatory variable that was originally categorical (ordinal or nominal) and is being treated as a factor, whether the *original* categorical explanatory variable should be in the model involves testing whether all $k - 1$ regression slopes that pertain to it are *0*. Unfortunately, our current tools do not allow us to test whether sets of regression slopes are zero, but a procedure for doing so will be outlined later when we discuss model selection.

It is often of scientific interest to specify a different null hypothesis: we may have a prior belief that the regression coefficient should be a certain value other than *0*. For example, in medical studies, we are often testing to see if a particular difference has been met in accordance with government protocol for drug development. If we are interested in testing the hypotheses

$$\begin{aligned} H_0 &: \beta_j = f \\ H_1 &: \beta_j \neq f \end{aligned}$$

we can use the test statistic

$$t = \frac{\hat{\beta}_j - f}{se(\hat{\beta}_j)}$$

which again has a t distribution with $n - (p+1)$ degrees of freedom if the five error assumptions are valid.

Combined Test for All Slopes

In addition to a *t-value* and (possibly) a p-value for every regression parameter, computer regression output will also typically contain the value of an another test statistic, the *F statistic*, which pertains to the overall linear model rather than to just one parameter/term in it. This *F-value* is used to test the null hypothesis that *all* of the regression *slopes* equal zero (i.e., that $\beta_1 = \beta_2 = \ldots = \beta_p = 0$); this test is often referred to as the *F-test for the overall regression*. If this null hypothesis is not rejected, then it appears that none of the explanatory variables have an effect on the response variable and that they are all bad predictors: in other words, the linear model would not be a very good one! A p-value will often accompany the value of this statistic and can be used to see if the null hypothesis can be rejected at a particular α (significance) level.

Confidence Intervals for Parameters

Returning to individual regression parameters, it is often of interest to calculate a confidence interval for a parameter so that we can have some idea about the possible spread of values that the true value of the parameter is likely to take. What's more, a confidence interval shows whether, for a particular confidence/significance level, the interval contains particular values of the regression coefficient: for example, if a *95%* confidence interval contained the value *f*, then we would not reject the null hypothesis that its true value is *f* at the 5% level. The previous statement follows from the duality between confidence intervals and hypothesis tests.

If our five error assumptions are valid, the expression for a *two-sided* 100*(1-α)% confidence interval for regression parameter β_j is:

$$\hat{\beta}_j - t_{\frac{\alpha}{2},n-(p+1)} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2},n-(p+1)} se(\hat{\beta}_j),$$

where $t_{\frac{\alpha}{2},n-(p+1)}$ is the α/2 percentile of the t distribution with $n - (p + 1)$ degrees of freedom. Normally, we consider setting $\alpha = 0.05$ to form a 95% confidence interval.

Example 1 (continued)
For the automobile example we have been considering, we already have all the values that we need for this calculation and hence the confidence interval is of the form:

$$1.980 - 2.262 \times 2.216 \leq \beta_1 \leq 1.980 + 2.262 \times 2.216$$

which is the same as

$$-3.032 \leq \beta_1 \leq 6.992 .$$

Clearly, this interval tells us that we are fairly unsure about what the true value of the regression coefficient is, since in 95% of samples, we would expect it to lie anywhere between -3 and 7, which is a pretty large range. We also note that this interval contains the value *0*, which implies that a hypothesis test of the parameter being *0* would not be rejected; in fact, this was confirmed above ■

Interpretation of Parameters

How do we interpret what is meant by the value of a particular regression parameter?

Well, let's start with the regression intercept. If the treatment factor approach is used for all explanatory categorical variables, the regression is typically interpreted as the *predicted response when all continuous variables are zero and all categorical variables are at their baseline level*. Here, we should note that this interpretation of the intercept may be nonsensical for a particular model. For instance, suppose we have a simple linear model in which height predicts weight. Well, in that model, the intercept would correspond to the predicted weight for individuals with zero height; since no one has zero height, the intercept is not particularly interesting for answering questions in this example. As another note on the intercept, some theory or previous knowledge may lead us to believe that the intercept should be zero in a particular linear model. For instance, in our weight-height example, it would make sense for the intercept to be *0* since a non-existent person (in terms of height) shouldn't weigh anything. If we believe that the intercept should be zero, then we could fit a linear model ***without an intercept***; most statistical packages allow the user to specify whether or not an intercept should be included. If the "no intercept" option is specified as a result of prior knowledge, then β_0 will be set equal to zero (rather than being estimated from the data), and the model slopes will be estimated using a slightly different procedure than if an intercept were included in the model. If the assumption of a zero intercept is borne out by the data (i.e., if the actual estimate of the intercept from the data is close to zero), then using the "no intercept" option probably won't change the estimates of the model slopes much.

As for a regression slope, the interpretation depends on whether the corresponding explanatory variable is continuous or categorical. If the explanatory variable is continuous, then its slope parameter is *the amount by which the mean response increases or decreases when the explanatory variable increases by 1 unit*. In other words, if the value of the slope is positive, then the mean response increases when the explanatory variable decreases, but if the slope's value is negative, then the mean response decreases when the explanatory variable increases. If the explanatory variable is categorical and the factor approach with a treatment parameterisation is used, the slope value for each (non-baseline) level of the categorical variable is *the difference between the mean response at that level and the mean response at the baseline level.* It is important to note that for a multiple linear model with more than one explanatory variable, these slope interpretations are *conditional on the other explanatory variables being the same*. This is the case because, by definition, the multiple linear model looks at the effect that an explanatory variable has on the response variable with the values of the other explanatory variables fixed. Thus, when we are interpreting the values of slope parameters in a model with more than one explanatory variable, the previous interpretations should always be followed by the clause *when the other explanatory variables are held constant.* However, except in experimental studies, we are typically unable to hold other explanatory variables constant while allowing one to change (typically, the explanatory variables will all vary together, out of our control). Thus, it might make more sense to instead add the clause *for units with approximately the same values of the other explanatory variables*.

If a regression slope corresponds to an explanatory variable that is a function of an original design variable or covariate (e.g., the slope belongs to an interaction term or a higher order term in a hierarchical model), then the interpretation of the slope parameter value can be somewhat more complicated and will be addressed later in our discussion of model selection.

Example 1 (continued)

In the cars example, the intercept means that when cars are new (i.e., Age=0), we expect them to have done 68780 miles. Here, this interpretation is not nonsensical, as cars can actually be new. The slope tells us that for each year the car ages, you expect them to have approximately 1980 more miles ■

Predicted Values: Estimates and Confidence Intervals

Often, we will be interested in predicting the *Y* value for a particular combination of explanatory variable values. For example, in the Toyota Land Cruise example, we may be interested in predicting how many miles a car Aged 8 would have done.

In some instances, we may want to make such a prediction for a unit contained in our data set or, identically, for a particular combination of explanatory variable values contained in the data set. (As a note, the predicted *Y* values for the units contained in the data set are referred to as the model's ***predicted values***). In other cases, we may want to make such a prediction for a new unit or, identically, for a new combination of explanatory variables. In this latter case, we must distinguish between two different scenarios. In the first, which we refer to as ***interpolation***, the new combination of explanatory variables lies *within* the range of the explanatory variable combinations in the actual data set; in the second scenario, which is termed ***extrapolation***, the new combination of explanatory variables lies *outside* the range of the explanatory variable combinations in the actual data set. We must be very wary of extrapolation because we do not know if the fitted linear relationship holds outside the observed range of *X* values. The parameters of our linear model are estimated for the range of explanatory variables that exists in our data set and the validity of the linear model is checked for this range; thus, our particular linear model may not be valid outside of this range or, worse yet, the relationship between the response and explanatory variables may not be even linear outside the range of the explanatory variables in our data set. For instance, in the example we have been considering, we should only think about predicting the Mileage if the Age of the vehicle is between 0 and 15 years, since outside of this range we can not be sure that the relationship remains linear. To get an idea of the problems that can occur with extrapolation, consider the following simple (one explanatory variable) example:

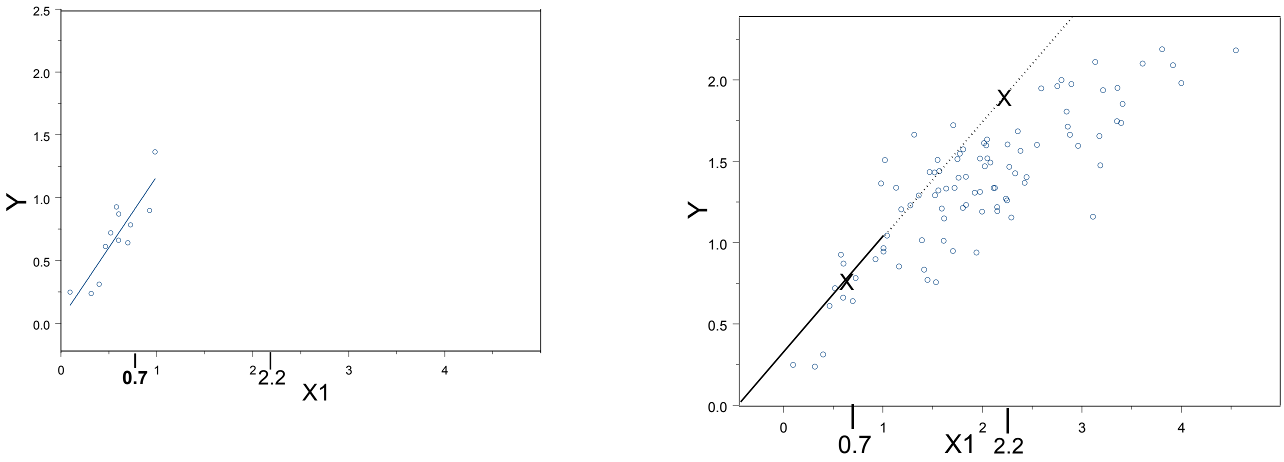


Figure 7: *Interpolation and extrapolation, with predicted values marked by an* X

Looking at the left-hand figure, suppose that a line is fit to the set of points with X_1 values between *0* and *1*. Further, suppose that, from this estimated linear model, we desire to predict the *Y* value for $X_1 = 0.7$ and for $X_1 = 2.2$, which are indicated on the x-axis of the left-hand figure. The first prediction would be an instance of interpolation, and the second would be an instance of extrapolation. Now, we turn to the right-hand figure, where we see a sample of points with X_1 values between *0* and *5*. Clearly, the relationship between *Y* and X_1 that holds (and is estimated) for X_1 values between *0* and *1* is not an adequate representation of the relationship between *Y* and X_1 for X_1 values larger than *1*. As a result, we can see from the right-hand figure that, although the (interpolated) predicted *Y* value for $X_1 = 0.7$ is a good prediction, the (extrapolated) predicted *Y* value for $X_1 = 2.2$ is not particularly representative, giving us a good example of the dangers of extrapolation.

Once we have estimated the *β*s in our linear model, if we want to predict *Y* for the combination of *X* values, $X_{1,s}, \ldots, X_{p,s,}$, then our prediction would be

$$\hat{Y}_s = \hat{\beta}_0 + \hat{\beta}_1 X_{1,s} + \hat{\beta}_2 X_{2,s} + \cdots + \hat{\beta}_p X_{p,s} ,$$

where $\hat{Y}_s$ is the predicted value for that particular combination of explanatory variables. Note that this predicted *Y* value is based on the *estimated* values of the linear model parameters.

Example 1 (continued)

With the estimated regression coefficients, we can use the model to predict what *Y* (Mileage) would be given for the various observed values of *X* (Age) in the data set:

$$(\text{Predicted Mileage})_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Age}_i = 68.78 + 1.98\text{Age}_i$$

The predicted values are:

Age in years (*X*)	Mileage in 1000 of miles (*Y*)	Predicted Mileage in 1000 of miles
3	68.22	74.72
5	59.53	78.68
5	73.43	78.68
5	82.55	78.68
6	67.67	80.66
6	93.61	80.66
9	86.94	86.60
10	120.94	88.58
11	112.27	90.56
14	129.42	96.50
15	38.22	98.48

Looking at this table, it is clear that the observed Mileage for the vehicle Aged 15 does not fit in well with the estimated linear model. However, excluding this observation, it would appear that the model fits the general pattern of the data set well. We could represent all of this information graphically by producing a scatter plot of the two variables along with the predicted values from the model and the suggested straight line from the regression procedure (which is called the ***predicted regression equation***). Doing this we get the following picture:

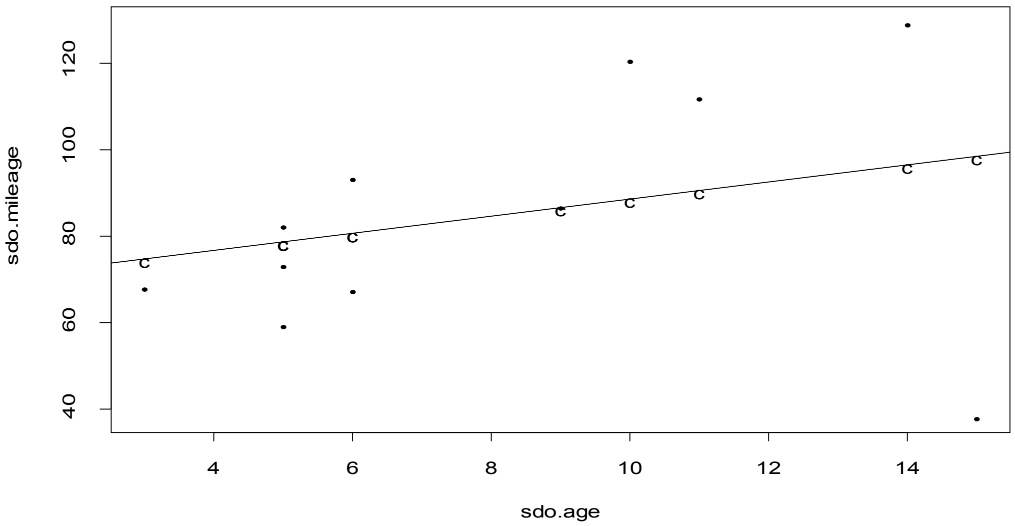


Figure 8: *Predicted Regression Equation for Toyota Land Cruiser Example*

In the above figure, the plotting character *c* represents the predicted values that, by definition, have to fall directly on the predicted regression line ■

In addition to producing a predicted value for a particular combination of explanatory variables, we may also be interested in obtaining a confidence interval for our prediction as well. This interval will then allow us to have some handle on how variable our prediction is. It is important to note that the confidence interval will depend on whether we are interested in predicting a *mean* response for the particular combination of explanatory variables or an *individual unit's* response. There is more uncertainty involved in the latter case. In the former case, we are uncertain about the true regression equation; however, once this equation is known for sure, the mean response for the particular combination of explanatory variables is merely equal to its predicted value, without any error, as shown in Equation (2). Thus, in the former case, the only source of uncertainty in our prediction concerns the parameters of the linear model equation. On the other hand, if we are interested in predicting the response for an individual unit, there is an additional source of uncertainty: not only do we not know the true regression equation that allows us to calculate a mean response, but we also do not know where the particular unit's response is in relation to that mean (since we have assumed that the responses vary around the means with variance σ^2). As a result, the confidence interval for a mean response is always narrower than the confidence interval for an individual unit's response, reflecting the greater amount of uncertainty in the latter case.

We will not present the formulas for calculating the confidence intervals for mean or individual responses. However, you should know that both types of confidence intervals are wider when the linear model fits the data worse (as indicated by the sum of squared residuals being larger) or when the explanatory variable combination of interest is farther from the centre of the range of explanatory variables in the data set. Lastly, you should know that the validity of these confidence intervals for predicted values is particularly sensitive to the validity of the five error assumptions (compared to hypothesis tests and confidence intervals for the model parameters, confidence intervals for predicted values are more likely to be invalid when the error assumptions break down).

Example 1 (continued)

In the car example described previously, consider predicting the Mileage for a car Aged 8, along with the associated confidence interval. The prediction is

$$68.78 + 1.98 \text{ x } 8 = 84.6.$$

Further, a statistical package would tell us that a 95% confidence interval for the *mean* Mileage for Age = 8 is (69.1, 101.1) and that a 95% confidence interval for an *individual* car's Mileage at Age 8 is (21.4, 147.8). Note that the second confidence interval is considerably wider ■

Overall Measures of Fit

As stated previously, the linear model is never an exact fit to all of the units in the data set, as acknowledged by the inclusion of an error term in Equation (1). Thus, we will often seek to measure exactly how well the estimated linear model fits our particular data.

Conditional Variance (Residual Standard Error)

Here, we will recall that the residual for unit *i*, termed e_i, is the difference between its observed and predicted values: $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_p X_{p,i})$. One might use the collection of residuals for the *n* units in the data to measure the estimated linear model's goodness of fit to the data: the greater the differences between the *Y* values predicted by the model and the actual observed *Y* values, the worse we might say that the model fits. More specifically, we might use the sum of these residuals, squared (because squaring prevents negative and positive residuals from cancelling each other out),

$$RSS = \sum_{i=1}^{n} e_i^2 ,$$

which is often referred to as the ***Residual Sum of Squares (RSS)***.

The RSS can be used to estimate the conditional variance, σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{RSS}{n-(p+1)} .$$

Thus, the preceding value is often referred to as the ***(estimated) conditional variance***; the square root of this number, $\hat{\sigma}$, is often referred to as the ***residual standard error***. The estimated conditional variance (or, identically, the residual standard error) can be thought of as a measure of the estimated linear model's goodness of fit to the data: the larger it is, the worse the model fits.

Example (continued)
The computer output for the cars linear regression will probably contain the following information:

```
Residual standard error: 27.93 .
```

However, to decide how well our model fits the cars data, we need to decide whether this residual standard error is large ■

One way to determine whether the conditional variance is large or not is to compare it to the (estimated) unconditional variance of the response variable, which, as one learns in an introductory statistics course, is simply

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1} = \frac{TSS}{n-1},$$

where $\overline{Y}$ is the sample mean of the response variable. Note that the numerator in the preceding equation is often referred to as the ***Total Sum of Squares (TSS).*** By definition, $\hat{\sigma}^2$ will always be less than or equal to s^2. However, how much smaller $\hat{\sigma}^2$ is than s^2 is a measure of how well the model fits the data: if the model fits the data well, $\hat{\sigma}^2$ will be very small relative to s^2. (Identically, $\hat{\sigma}$ will be small relative to s.)

Example (continued)
Compared to the value of s for the Mileage variable, which is about 28, the residual standard error (27.93) is very high, suggesting that the model does not fit the data particularly well ■

R^2

Another popular measure of model fit, and one that is almost always part of computer regression output, is R^2. This statistic is defined as

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}.$$

Note that the difference between Total Sum of Squares and Residual Sum of Squares is often referred to as the ***Explained Sum of Squares (ESS)***. ESS can be calculated by subtracting RSS from TSS, or, alternatively, using the formula

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 .$$

As a result of the way in which R^2 is defined, it can only take values between *0* and *1*. An R^2 value of *0* would indicate that the model does not fit the data well (i.e., the predicted responses, the $\hat{Y}$s, from the estimated linear model are not any better than just using the sample mean, $\overline{Y}$, as a predictor). On the other hand, an R^2 value of *1* would indicate a perfect fit (i.e., the predicted responses from the estimated linear model, the $\hat{Y}$s, are exactly equal to the observed responses, the *Y*s, for all *n* units). R^2 is often thought of as *the percent of the variance in the response variable that is explained by the explanatory variables*.

One interesting point about R^2 is that the hypothesis test of $R^2 = 0$ is the same as the overall *F* test for the regression (which tests whether all the regression slopes are zero). We can demonstrate that this is true mathematically. However, it is also easy to see why it is true intuitively. If R^2 is equal to zero, the predicted responses, the $\hat{Y}$s, from the estimated linear model are not any better than just using the sample mean, $\overline{Y}$, as a predictor. This statement is the same as saying that the explanatory variables in the model do not seem to have an effect on *Y*, in which case all of their slopes would be zero.

As a final note, one should be careful when using R^2 to compare models with different numbers of explanatory variables: since R^2 always decreases (or at least stays the same) when an explanatory variable is taken out of the linear model, R^2 will always favour the model with more explanatory variables (the one that contains all of the explanatory variables of a smaller model, plus a few more).

A Final Example

The data set that we are going to consider is from Lock (1993)[3] and gives information on 93 different cars on sale in the US in 1993. We hope to predict Fuel Consumption in City using a subset of the recorded variables.

The variables in the data set are:

Variable	Description
Fuel Consumption in City	Continuous response
Number of Engine Cylinders	Ordinal explanatory, taking values 3,4,5,6,8 and rotary
Type of Car	Nominal explanatory, taking values compact, large, midsize, small, sporty and van
Weight of Car	Continuous explanatory
Engine Size	Continuous explanatory
Drive Type	Nominal explanatory taking values 4WD, front and rear

Note that Fuel Consumption in City is not in miles per gallon but is a standardised measure in which high levels of the response correspond to low fuel consumption and vice-versa.

Prior knowledge leads us to believe that this collection of explanatory variables can be used to linearly predict the fuel consumption measure. If we go ahead and fit the data with a linear model in which a treatment factor approach is used for all ordinal and nominal explanatory variables, we get the following results:

Coefficients:

	Value	Std. Error	t value	p value
(Intercept)	20.1636	5.6535	3.5665	0.0006
Weight	0.0046	0.0022	2.0948	0.0394
Cylinders:4	7.6687	2.5019	3.0651	0.0030
Cylinders:5	12.5628	4.1145	3.0533	0.0031
Cylinders:6	11.5798	3.2588	3.5533	0.0006
Cylinders:8	12.9720	4.1776	3.1051	0.0027
Cylinders:rotary	25.2770	4.8178	5.2465	0.0000
Type:Large	-1.8174	1.9063	-0.9534	0.3433
Type:Midsize	0.8122	1.4397	0.5641	0.5743
Type:Small	-3.8317	1.5755	-2.4320	0.0173
Type:Sporty	-0.2469	1.4725	-0.1676	0.8673
Type:Van	3.4614	2.2091	1.5669	0.1212
EngineSize	2.4555	1.1285	2.1758	0.0326
DriveType:Front	-3.2119	1.4992	-2.1423	0.0353
DriveType:Rear	-3.0184	1.8550	-1.6272	0.1077

The regression output above can be interpreted in the following manner, beginning with the estimated values of the model parameters:

[3] This analysis comes from an MSc in Applied Statistics practical set by Prof. BD Ripley MT97 and the data is available on disk with "Modern Applied Statistics with S-Plus, Venables W.N. & Ripley B.D., Springer-Verlag 1994".

1. *Weight:* This says that as the weight of the vehicle increases we see a decrease in fuel consumption (which is the same as an increase in the Fuel Consumption in City measure).
2. *Cylinders:* This is a categorical variable with 6 levels and therefore has 5 coefficients associated with it. The coefficients say that as you increase the number of cylinders in the vehicle then the fuel consumption decreases compared to a vehicle with 3 cylinders, and if you have a vehicle with rotary cylinders then the consumption is the worst possible.
3. *Type:* A nominal variable with 6 levels and hence 5 coefficients. The estimates basically say that having a large, small or sporty vehicle improve fuel consumption, whereas having a midsize vehicle or a van decreases consumption, all in comparison to a compact vehicle.
4. *Engine Size:* This says that as you increase engine size, you increase fuel consumption.
5. *Drive Type:* These coefficients tell us that in comparison to a 4WD vehicle, other types of vehicles cause a reduction in fuel efficiency of approximately equal levels.

The statistical package used also automatically calculated the *t*-statistics and associated p-values. For the two continuous explanatory variables, the fact that the p-values in the above table are small indicates that it is probably a good idea to include the variables in the model. For the categorical variables, it is not always obvious that they should be included in the model as some levels have significant p-values and some levels have non-significant p-values, which indicates that it would be appropriate to leave out the indicator variables for those levels. However, as discussed earlier, whether or not to include a categorical variable (treated as a factor) is an issue of model selection and will be discussed later.

It is also possible, although not described here, to calculate confidence intervals for estimated parameters and predictions and the associated confidence intervals for various values of the explanatory variables.

A. Roddam (2000), K. Javaras (2002), and W. Vos (2002)