Chapter 6

Simple linear models

6.1 Objectives

1. To state, fit, check, and interpret simple linear models (comprising only one explanatory variable).
2. To use R in the fitting and checking of models.

6.2 Background

At the end of the last chapter, we looked at creating estimates and tests for the purposes of comparing two populations. This chapter places those ideas inside a larger framework: inference for the distribution of a random variable, say Y, given that we know the value of another, hopefully related variable, say X.

6.3 One numerical explanatory variable

In this section we consider inference associated with the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i; \quad e_i \sim \mathrm{N}(0, \sigma)$$

We could also write this $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma)$. We will denote the predictions from this model by $\hat{y}_i$, that is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

In lab 2 we used α and β to denote the parameters. Here we use β_0 and β_1 because it generalises easily to more complex models.

6.3.1 Data Display

Example **Fuel consumption of cars**

This data set, which was extracted from the RACV website, is listed fully in section 2.4.2 of lab 2. First, we load the car data by:

```
> cars <- read.csv("../data/racv.csv")
```

Graph the variables of interest using the following code (Figure 6.1).

```
> plot(lp100km ~ mass.kg, data=cars,
+      xlab="Mass (kg)", ylab="Fuel consumption (l/100km)")
```

Note the use of `xlab` and `ylab` to create meaningful axis labels.

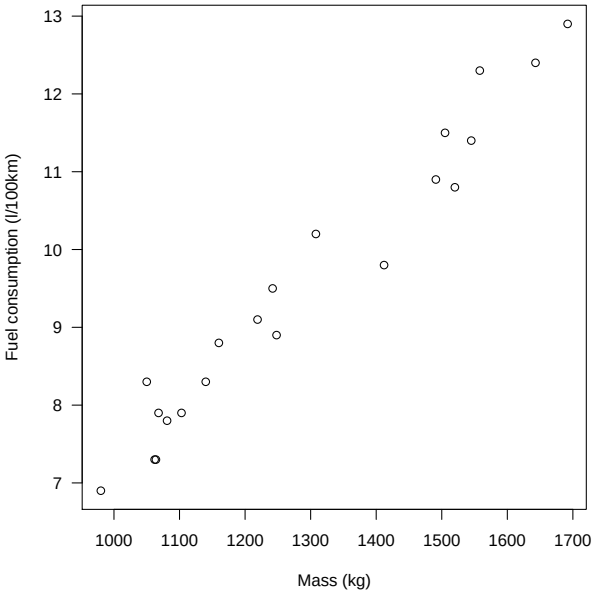


Figure 6.1: Scatterplot of variables of interest in the RACV data.

6.3.2 Estimation of parameters

The parameters β_0 and β_1 are (usually) estimated by the method of least squares, and σ is estimated by

$$\hat{\sigma}^2 = \frac{\text{sum of squared residuals}}{n-2}$$

The denominator is $n-2$ because there are two parameters to estimate (besides σ). Using calculus, it is possible to find the following expressions for these estimates:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}
\end{aligned}$$

These estimates can readily be obtained using a scientific calculator or spreadsheet, but it is even easier to use a statistical package which provides a lot more as well.

6.3.3 Fitting and checking the model in R

The `lm` function (`lm` = linear model) provides us with least-squares estimates of the parameters of the model.

```
> cars.lm <- lm(lp100km ~ mass.kg, data = cars)
```

Notice that there is no output. For that we need the `summary` function, which we consider shortly.

An obvious visual check of the model is seeing how the predicted line looks in relationship to the scatterplot.

```
> abline(cars.lm, col = "blue")
```

However, in the process of modelling we have made a number of explicit assumptions, which should be checked. We need to check that the straight line form is reasonable, and examine the distribution of the errors. Beyond the graph above, which is excellent for this simple design but not available for more complex designs, the residuals are the best tool. We can obtain useful diagnostic information using the following code (which produces Figure 6.3):

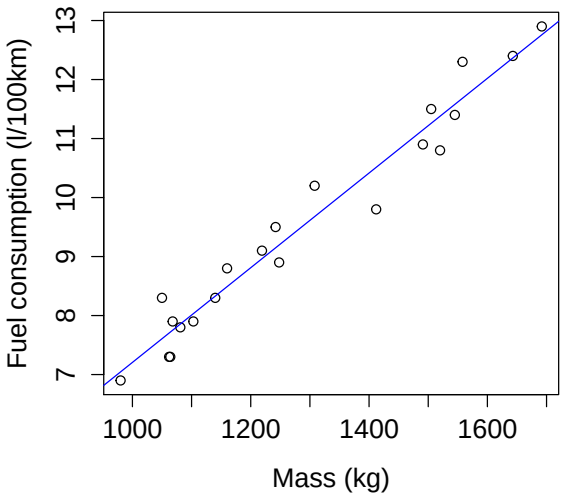


Figure 6.2: Scatterplot of RACV data with fitted line..

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

The `plot` statement by default produces four graphs when applied to a linear model (`lm`).

- The top left panel shows a plot of the residuals against the fitted values, with a smooth red curve superimposed. Here we're looking for evidence of curvature, non-constant variance (heteroscedasticity), and outliers. Our plot shows no problems, as the points are consistent with a random scatter.
- The top right panel shows a quantile plot of the standardized residuals (each residual is divided by its standard error) against the normal distribution. Here the ideal plot is a straight line, although modest departures from a straight line are usually acceptable (due to the central limit theorem and other large-sample theory). Departures from a straight line in this plot can indicate non-normality of the residuals *or* non-constant variance. Our plot is sufficiently linear.
- The bottom left panel shows the square root of the absolute residuals against the fitted values, along with a smooth red line. Departures from a horizontal line signify heteroscedasticity. While the smoothed curve isn't perfectly straight, this is minor and in part due to the small number of observations around $y = 10$; it doesn't suggest any problems with our model.
- The bottom right panel shows a plot of the "leverage" of the observations against the standardized residuals. These are the two components of *Cook's Distance*, a statistic that reports the overall impact of the observations on the parameter estimates (note that being a large residual or a high leverage point alone is no guarantee of having a substantial impact upon the parameter estimates). A reasonably well accepted rule of thumb is that Cook's Distances greater that 1 should attract our attention. Contours of these distances at 0.5 and 1.0 are added to the graph to assist interpretation. All of our observations are within the 0.5 contour.

Overall we have no reason to doubt the validity of our regression assumptions, so the fitted model can be used confidently. We proceed to interpret it.

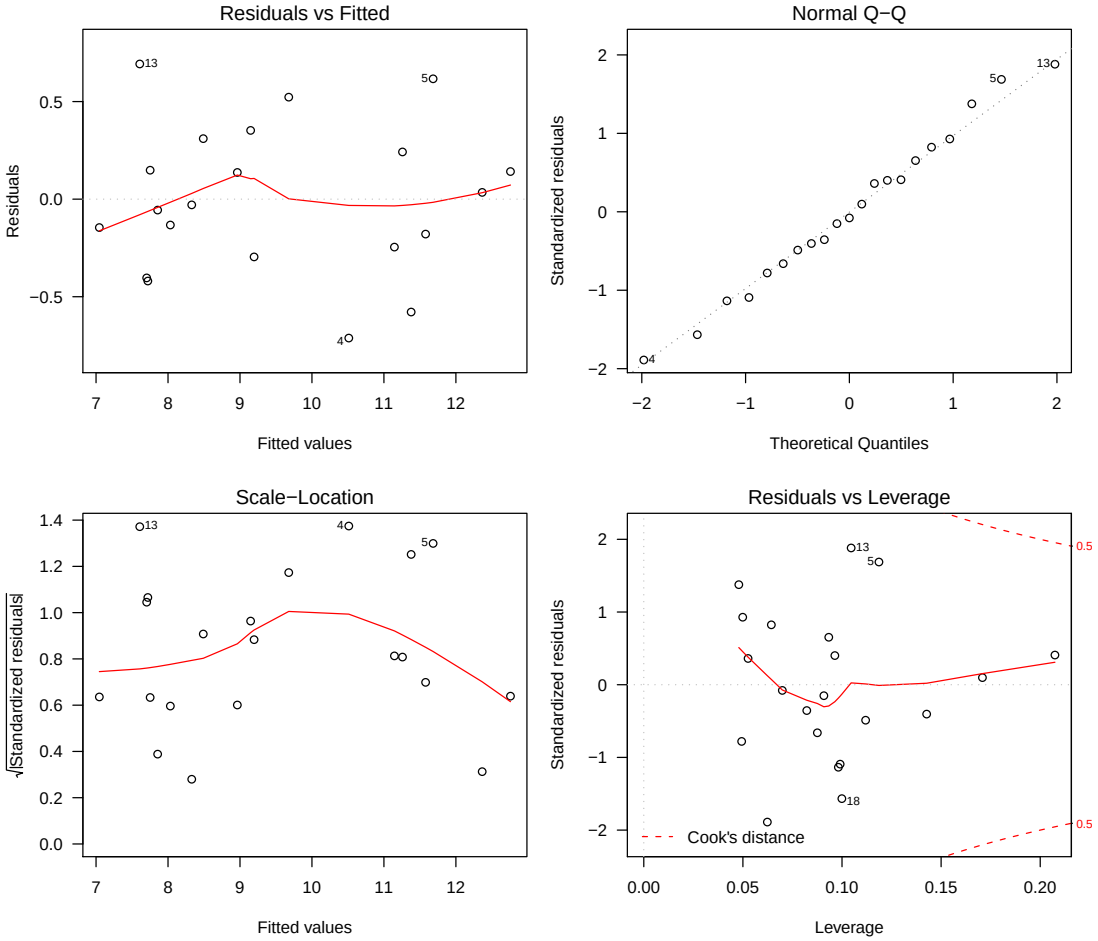


Figure 6.3: Diagnostic information for the linear regression of car mass on efficiency.

6.3.4 Parameter estimates, standard errors and confidence intervals

To examine the fitted model, we now type:

```
> summary(cars.lm)

Call:
lm(formula = lp100km ~ mass.kg, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71186 -0.24574 -0.02938  0.24193  0.69276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.817768   0.506422  -1.615    0.123
mass.kg      0.008024   0.000387  20.733 1.65e-14 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 0.3891 on 19 degrees of freedom
Multiple R-squared: 0.9577,        Adjusted R-squared: 0.9554
F-statistic: 429.9 on 1 and 19 DF,  p-value: 1.653e-14
```

R first confirms the model that it has fitted, and then provides a 5-figure summary of the distribution of the residuals (which we would like to be normal, or at least symmetrical. Symmetry appears very well

satisfied here.

The next section, `Coefficients`, summarizes the parameter estimation for the model. The parameter estimates are in the column labelled `Estimate`. The entries in the `Std Error` column are the *standard errors* of the estimates, those in the `t value` column are the values of the estimates divided by their standard error, and the values in the `Pr(>|t|)` column are the P-values for testing the hypothesis that the parameter is zero, against the 2-sided alternative. Here, the P-value for H_0: $\beta_0 = 0$ is 0.123, which means that the (true) line could plausibly pass through the origin, whereas the P-value for H_0: $\beta_1 = 0$ is extremely small which means that β_1 is almost certainly different from zero.

The slope or rate of increase (β_1) is the more important parameter here because of its interpretation. The estimate 0.008 indicates that fuel consumption is predicted to increase by 0.008 l/100km for each additional kg of mass, or about 1 l/100km for each extra 125kg.

The standard errors can be used to produce confidence intervals for the parameters

6.3.5 Overall summary statistics s and R^2

The output also has values for `Residual Standard Error`, `Multiple R-squared`, and `Adjusted R-squared`.

Residual standard error, s (also called residual standard deviation)
The value of s is a measure of the variability of individual points as they vary from the fitted model.

$$s_{residual} = \hat{\sigma} = \sqrt{\frac{\text{sum of (residuals)}^2}{\text{no. study units} - \text{no. parameters estimated}}}$$

In this case

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^{n}[(\text{fuel consumption})_i - (\text{predicted consumption})_i]^2}{n-2}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-2}}.\end{aligned}$$

The square of this quantity, $\hat{\sigma}^2$, is the "residual mean square".

R^2
The quantity R^2 (which R prints as `Multiple R-squared`) is calculated as $SS_{regression}/SS_{total}$ and is a measure of the variation explained by the model relative to the natural variation in the response values.

The SS (sums of squares) quantities are computed as follows. Recall that the predictions from the model are denoted $\hat{y}_i$.

$$SS_{regression} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

We note in passing that there is a third such quantity.

$$SS_{residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and that for reasons that are not at all obvious, but are fundamentally Pythagorean,

$$SS_{total} = SS_{residual} + SS_{regression}.$$

In this case $R^2 = 0.958$ ($= 95.8\%$), indicating that the straight line model with car mass explains a large proportion of the variation in fuel consumption. R^2 is widely used as a summary measure of the explanatory effectiveness of a straight line model.

Adjusted R^2
R also gives the value of the so-called adjusted R^2, which is an alternative measure of the proportion of the variance that can be attributed to the model, or "explained by" the model.

It is defined by: $1 - R^2_{\text{adj}} = \dfrac{s^2}{s_y^2}$,

where s^2 denotes the residual variance and s_y^2 the (total) variance of the observations.

Many regard this as a preferable measure as it adjusts for the number of degrees of freedom in the model.

R^2 and r

R^2 can be calculated for any linear model, but when the model is a linear regression with just one explanatory variable, R^2 is equal to the square of the correlation coefficient r.

We can thus find the correlation between weight and fuel consumption by

`Multiple R-squared = 0.958` $\Rightarrow r = \pm\sqrt{0.958} = \pm 0.979$

We choose between plus or minus by making sure the sign of r matches the sign of the slope. In this case the slope is positive (there is a positive association between weight and fuel consumption) so $r = 0.979$. The correlation coefficient r can also be obtained from R using:

```
> cor(cars[,2:4])
```

```
           Make   lp100km   mass.kg List.price
             lp100km   mass.kg List.price
lp100km    1.0000000 0.9786067   0.825725
mass.kg    0.9786067 1.0000000   0.772601
List.price 0.8257250 0.7726010   1.000000
```

F -statistic

One final element of the regression output is the `F-statistic` presented on the final line. That output is a summary of the test of the whole model, $y_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma)$, against a default null model $y_i \sim \text{N}(\mu, \sigma)$ or $y_i \sim \text{N}(\beta_0, \sigma)$, which has a slope of 0.

Note that for this simple model with one numerical explanatory variable, the F-statistic is the square of the t-statistic and the P-value is the same for both.

6.4 One categorical explanatory variable

Consider the simplest situation possible—the comparison of two populations. The usual way to formulate the problem is to assume that they differ only in their means, i.e.

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \sim \text{N}(0, \sigma) \quad i = 1, 2. \tag{6.1}$$

We can reformulate this model, or *reparameterise* it, as follows:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad e_i \sim \text{N}(0, \sigma) \quad i = 1, 2. \tag{6.2}$$

where β_1 is the mean of the first population, β_2 is the difference between the means of the populations, and x_i is a predictor variable that takes on values 0 for all the observations in population 1 and 1 for all the observations in population 2, i.e. $x_1 = 0$ and $x_2 = 1$. Both of these parameterisations are useful depending on the context, and on the statistical package used.

In the context of a categorical explanatory variable, it turns out to be easier to consider hypothesis testing before confidence intervals.

6.4.1 Hypothesis testing

Example **Potato yields (revisited)**

The following data were obtained from an experiment (using a completely randomized design) to investigate the effectiveness of four fertiliser treatments on the yield of King Edward potatoes, measured as crop weight in kilograms.

```
> potatoes <- data.frame(treatment = factor(rep(c(1:4), rep(4,
+     4))), crop_wt.kg = c(752, 762, 686, 787, 621, 637, 670, 575,
+     642, 667, 655, 660, 645, 627, 596, 576))
```

Note that we explicitly call the `factor` function to be certain that R interprets the treatment as a categorical variable, rather than as a numerical variable.

The question being addressed here is whether there is a difference between the mean yields with different treatments. That is, can some of the variation in yields be attributed to the different treatments? This question can be expressed formally in the hypothesis testing context as:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \ (= \mu \text{ say}) \text{ versus } H_1 : H_0 \text{ is not true.}$$

If H_0 is true, then the model given by (6.1) simplifies to

$$y_{ij} = \mu + e_{ij} \tag{6.3}$$

First we look at the summary statistics:

```
> summary(potatoes)
```

```
 treatment          crop_wt.kg
 1:4               Min.    :575.0
 2:4               1st Qu. :625.5
 3:4               Median  :650.0
 4:4               Mean    :659.9
                   3rd Qu. :674.0
                   Max.    :787.0
```

We now fit a linear model using R:

```
> potato.lm <- lm(crop_wt.kg ~ treatment, data = potatoes)
> par(mfrow=c(2,2))
> plot(potato.lm)
```

This gives us Figure 6.4. The most important plot is the graph of residuals vs fitted values (top left). It indicates that the assumption of constant variance is reasonable. The other plots don't show anything alarming.

To check if we should reject H_0 (the hypothesis that our four treatment groups have the same mean weights) we use the F test, which R calculates as part of the summary statistics:

```
> summary(potato.lm)
```

```
Call:
lm(formula = crop_wt.kg ~ treatment, data = potatoes)

Residuals:
    Min      1Q  Median      3Q     Max
-60.750 -14.250   4.625  15.437  44.250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   746.75      16.74  44.611 1.05e-14 ***
treatment2   -121.00      23.67  -5.111 0.000257 ***
treatment3    -90.75      23.67  -3.834 0.002380 **
treatment4   -135.75      23.67  -5.734 9.39e-05 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 33.48 on 12 degrees of freedom
Multiple R-squared:  0.7678,        Adjusted R-squared:  0.7097
F-statistic: 13.22 on 3 and 12 DF,  p-value: 0.0004117
```

According to the F test, we can reject H_0 at the 0.05 significance level since we have a P-value of 0.0004, and so we suspect that the different treatment groups do have different means.

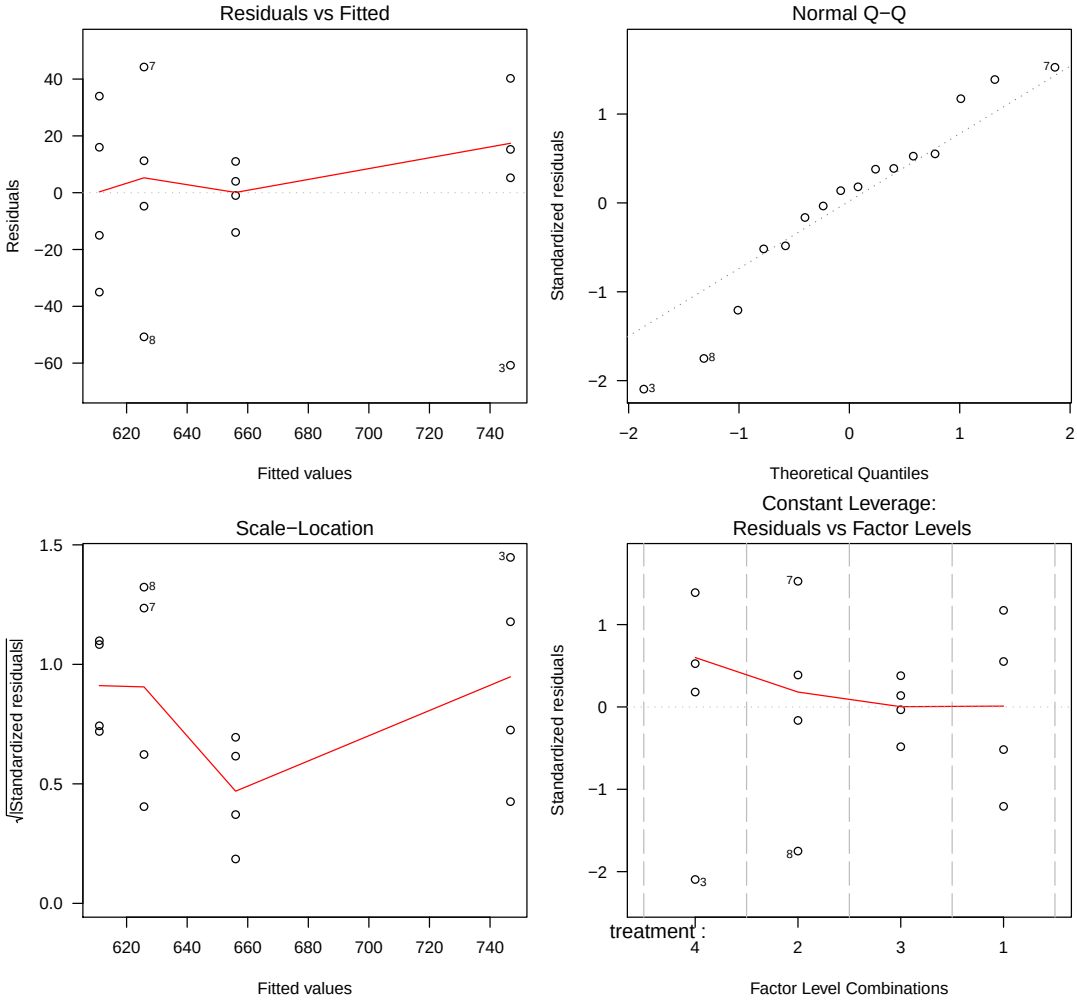


Figure 6.4: Diagnostic plots for potato linear model.

6.4.2 Point estimation

Having rejected $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, it is appropriate to estimate the μ_is. This is quite easy by hand, but a bit tricky in R because by default it uses the parameterisation shown at the start of section (6.3). R assumes that the goal will be to estimate differences between certain parameters.[1]

So, the "intercept" is taken as the first level of the categorical explanatory variable. All subsequent coefficients are deviations from the intercept. Consider this small example:

```
> y <- c(5, 5, 12, 12, 15, 15, 21, 21)
> x <- factor(c(0, 0, 1, 1, 2, 2, 3, 3))
> lm(y ~ x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x1           x2           x3
          5            7           10           16
```

Here we see that the first factor is treated as the intercept of all of the lines, and each of the others

[1] This choice is known as choosing a *contrast*, and the particular choice that R makes by default is the *treatment contrast*.

are treated as different dimensions for the data. In fact, R has fit three different lines:

$$\begin{aligned}\text{line}_1(x_1) &= 5 + 7x_1\\ \text{line}_2(x_2) &= 5 + 10x_2\\ \text{line}_3(x_3) &= 5 + 16x_3\end{aligned}$$

and so we see that R has output the gradients of the lines in each of the three directions.

To recover the original data we move one step along each line

$$\begin{aligned}\text{line}_1(1) &= 5 + 7 = 12\\ \text{line}_2(1) &= 5 + 10 = 15\\ \text{line}_3(1) &= 5 + 16 = 21\end{aligned}$$

Althernatively, we can think of the results from R as a 3D function

$$f(x_1, x_2, x_3) = 5 + 7x_1 + 10x_2 + 16x_3$$

so

$$f(0,0,0) = 5, \quad f(1,0,0) = 5 + 7 = 12, \quad f(0,1,0) = 5 + 10 = 15, \quad f(0,0,1) = 5 + 16 = 21.$$

This may seem quite confusing, but it is the default behaviour of R since it allows us to easily compare (or *contrast*) each factor with the first one in the list.

Other software packages may use different default settings, but the final model will be the same and so it is good to be aware of this.

The estimates from R of the parameters in the potato yield experiment are as follows:

```
> summary(potato.lm)$coef
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   746.75   16.73911 44.611084 1.048431e-14
treatment2   -121.00   23.67268 -5.111377 2.569058e-04
treatment3    -90.75   23.67268 -3.833533 2.380249e-03
treatment4   -135.75   23.67268 -5.734458 9.389787e-05
```

The estimates of the four means are:

1. 746.75
2. $746.75 + (-121.00) = 625.75$
3. $746.75 + (-90.75) = 656.00$
4. $746.75 + (-135.75) = 611.00$

We can see that this does in fact give us the correct means by comparing with the direct calculation:

```
> tapply(potatoes$crop_wt.kg, potatoes$treatment, mean)
```

```
     1      2      3      4
746.75 625.75 656.00 611.00
```

6.5 Making predictions

Consider again the case of one numerical explanatory variable — simple linear regression. The fitted equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

can be used for predicting a future observation for specified values of x_i. For example, we may want to predict the fuel consumption of a car of mass 1200kg which was not part of the original data set. The predicted consumption is $-0.818 + 0.00802 \times 1200 = 8.81$ l/100km. But what is the uncertainty around this prediction?

A confidence interval for a single future observation such as this is known as a *prediction interval*. It has the general form

$$\text{predicted values} \pm t \text{ value} \times \sqrt{s^2 + \text{s.e.}^2(\text{predicted value})}$$

This formula includes s.e.(predicted value) to allow for the uncertainty of the fitted equation, and an additional s^2 to allow for the individual observation fluctuating around the predicted value.

It can be obtained from R, using the `predict` function.

The t values are obtained from the t_{n-p} distribution.

Example **Nursery plants**

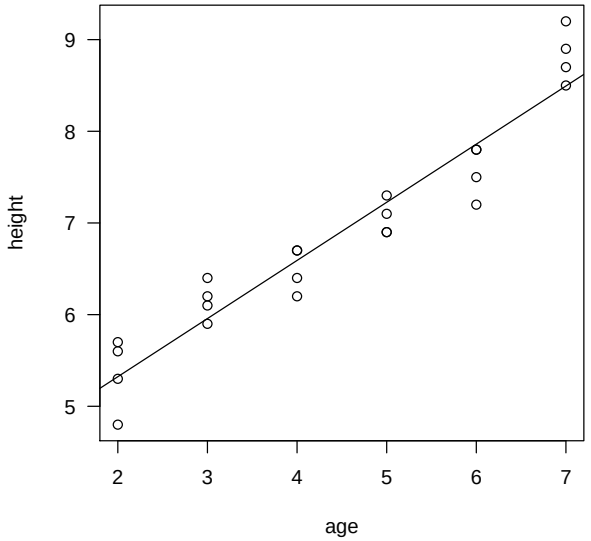
A plant nursery is interested in the relationship between the age and height of a variety of arborvitae. Twenty-four bushes aged 2, 3, 4, 5, 6 or 7 years old (4 of each) are selected and their heights recorded with the following results (in feet):

```
> plants <- data.frame(age = rep(2:7, rep(4, 6)),
+                      height = c(
+                        5.6, 4.8, 5.3, 5.7, 6.2, 5.9, 6.4, 6.1,
+                        6.2, 6.7, 6.4, 6.7, 7.1, 7.3, 6.9, 6.9,
+                        7.2, 7.5, 7.8, 7.8, 8.9, 9.2, 8.5, 8.7))
```

Construct a 95% prediction interval for the height of a 4 year old plant.

```
> plants.lm <- lm(height ~ age, data = plants)

> plot(height ~ age, data = plants)
> abline(plants.lm)
```



The relationship looks like it may be curved rather than linear, but we will ignore this for now.

This is how we use R to obtain prediction intervals:

```
> predict(plants.lm,
+         newdata = data.frame(age = 4),
+         interval = "prediction")

      fit      lwr      upr
1 6.59119 5.877015 7.305366
```

Also try this with interval = "confidence" to compare to the confidence interval.

If you bought a plant from the nursery and you wanted to know how tall it will be when it is 4 years old, you would predict a height of 6.59 feet, and would be 95% confident that the height would be between 5.88 and 7.31 feet.

6.6 Exercises

1. **Survival of green tree frogs**
   The following data give estimates of the population size of green tree frogs observed over a 10 year period (1985 to 1994) in a catchment in Victoria (adult females only):

   $$477, 303, 413, 408, 327, 299, 255, 281, 378, 240$$

   (a) Plot the number of frogs against year and fit a simple linear model with diagnostic plots.

   (b) Has there has been a decline in the green tree frog population over the years?

   (c) Give a point estimate for when the green tree frogs are likely to become extinct.

2. **Chicks on feeds**
   Here are the weight gains of chicks on four feeds. Formulate, estimate, check and interpret an appropriate model.

Feed A	42	68	85			
Feed B	42	97	81	95	61	103
Feed C	61	112	30	89	63	
Feed D	169	137	169	111	154	

3. **Newspaper prices**
   A morning newspaper lists the following prices for a particular make of used car with age x measured in years and selling price y measured in dollars:

x (years)	1	2	3	4
y (price)	\$24 500	\$20 000	\$20 000	\$17 000

   (a) A friend suggests that the best model for the data is

   $$y_i = 25000 - 2500x_i + error$$

   where Y_i is the random variable denoting the ith observation (y is a realisation of Y) and $error$ is a normally distributed random variable with mean 0 and standard deviation σ. Another friend suggests that a better model is given by

   $$y_i = 26000 - 2600x_i + error.$$

   Justifying your reasoning, determine which of your friends' models is the better.

   (b) An article in the paper suggests that it is possible to think of this make of car as "new" if it is less than or equal to two years old and "used" if it is more than two years old. Moreover, the article then says that car's price does not depend on its actual age, but just on whether it is "new" or "used".

   i. Assuming that the paper's statement is true, write down a model which describes this situation. This model should involve three parameters, two "mean" parameters and one "standard deviation" parameter. Give an explanation of each of the parameters.

   ii. Calculate an estimate of each of the three parameters.

   iii. Is this model better or worse than the straight line models in (a)?

4. **Smoking**
   The following data were collected on ten emphysema patients: the number of years x the patient smoked and inhaled, and a physician's evaluation y of the patient's diminution in lung capacity (measured on a scale of 0 to 100).

Patient	1	2	3	4	5	6	7	8	9	10
Years smoking (x)	25	36	22	15	48	39	42	31	28	33
Diminution in lung capacity (y)	55	60	50	30	75	70	70	55	30	35

   (a) Make a scatterplot of y against x and plot the regression line.

(b) What do you conclude from the plots about the relationship between duration of smoking and diminution in lung capacity? Can you conclude that smoking causes diminution in lung capacity? Give reasons for your answer.

(c) Fit a linear regression model, with diagnostic plots, to confirm your conclusion (or otherwise). What proportion of the variation of y is explained by the variation in x?

(d) Find the sample correlation coefficient between x and y.

(e) Give the meaning of the slope of the regression line and find a 95% confidence interval for the slope.

(f) Predict a person's diminution in lung capacity after 5 years or 30 years of smoking with 95% prediction intervals.

5. **Nematodes and tomato plant growth**
To demonstrate the effect of nematodes (microscopic worms) on plant growth, an agricultural scientist prepared 16 identical planting pots and then introduced different numbers of nematodes into the pots. A tomato seedling was transplanted into each plot. Here are the data on the increase in height of the seedlings, 16 days after the planting.

Nematodes	Seedling growth (cm)			
0	10.8	9.1	13.5	9.2
1000	11.1	11.1	8.2	11.3
5000	5.4	4.6	7.4	5.0
10000	5.8	5.3	3.2	7.5

(a) Investigate the possibility of fitting a straight line graph to these data. Analyse fully.

(b) Another perspective on these data is to forget about the numerical value of *number of nematodes*, but take it to be a treatment at four different categorical levels. It is as though *no nematodes* is treatment 1, *1000 nematodes* is treatment 2, etc. Analyse this model fully.

(c) Which of these ways is the better way to understand the data? Are there other models which may be better?

6. **Profitability versus stocking rate for sheep farms**
An observational study of ten farms recorded net profit and stocking rate, and gave these results.

	Net profit of ten Merino farms									
Stocking rate (ewes/ha)	4.0	4.8	5.2	5.5	6.1	7.5	8.0	8.3	8.7	9.0
Net profit ($/ha)	139	144	166	230	193	261	255	200	187	150

Plot an appropriate graph of the data. Fit a simple linear model with diagnostics, and assess its usefulness.

7. **Thread strength**
A laboratory technician measures the breaking strength of each of five kinds of linen threads, and obtains the following results:

thread 1	thread 2	thread 3	thread 4	thread 5
20.6	24.7	25.2	24.5	19.3
20.7	26.5	23.4	21.5	21.5
20.0	27.1	21.6	23.6	22.2
21.4	24.3	23.9	25.2	20.6

What can you conclude about the differences between the mean breaking strengths of the different types of thread?