5 Regression

Many statistical investigations are concerned with relationships between two (or more) variables, e.g. height and weight. In many important cases, one variable (known as the *explanatory variable*[1]) can be measured without error (or with negligible error), whereas the other variable (known as the *response variable*[2]) is random. In general, the distribution of the response variable (Y, say) when the explanatory variable (x, say) takes a given value depends on the value of x. Thus we can think of x as 'explaining' the corresponding observed value y of Y. Alternatively, we can imagine that if x varies then y 'responds'. The relationship between the response variable and the explanatory variable is known (for historical reasons) as *regression*.

Simple linear regression involves finding the regression relationship between the response variable Y and the explanatory variable x, on the basis of n pairs of observations (x_1, y_1), ..., (x_n, y_n) on (x, Y). We say that this is a regression of Y on x.

5.1 Linear regression

The simplest form of regression is *linear regression*, in which the response variable Y is related to the explanatory variable x by

$$E(Y) = \alpha + \beta x, \tag{9}$$

where α and β are (unknown) parameters. Note that (9) says that $E(Y)$ depends *linearly* on x. The line

$$y = \alpha + \beta x \tag{10}$$

is often called the *(population) regression line*, and α and β are called the *regression parameters*. It is useful to write Y_i for the random variable Y associated with the value x_i of x, for $i = 1, \ldots, n$. Then (9) gives

$$E(Y_i) = \alpha + \beta x_i, \qquad i = 1, \ldots, n. \tag{11}$$

This can be re-written as

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{12}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables or *errors* with mean zero and constant variance σ^2. Because the parameters α and β (and the error term) enter the model (12) in a linear way, this model for $Y_1, \ldots, Y_n$ is an example of a *linear model*. Such models are considered in detail in the Honours module *Generalized Linear Models and Data Analysis*.

5.1.1 Least-squares estimation

We need to estimate the values of the parameters α and β, i.e. to fit the model to data. One way to do this is by *least squares*. Consider the vertical distances

$$\epsilon_i = y_i - (\alpha + \beta x_i), \qquad i = 1, \ldots, n, \tag{13}$$

between the observed values $y_1, \ldots, y_n$ and the corresponding values $E(Y_1) = \alpha + \beta x_1, \ldots, E(Y_n) = \alpha + \beta x_n$ given by the model. A mathematically convenient way of measuring the difference between the data and the model is by the sum of squares

$$\begin{aligned} S(\alpha, \beta) &= \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2 \\ &= \sum_{i=1}^{n} \epsilon_i^2. \end{aligned} \tag{14}$$

The *method of least squares* identifies the values $\hat{\alpha}$ and $\hat{\beta}$ of α and β that minimise $S(\alpha, \beta)$.

Obtain expressions for the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$. (Hint: First differentiate $S(\alpha, \beta)$ w.r.t. α, to get an expression for α in terms of β. Then substitute this into the expression which results from differentiating $S(\alpha, \beta)$ w.r.t. β.)

[1] Explanatory variables are also known as 'covariates', 'predictor variables', 'independent variables' or 'risk factors'.
[2] Response variables are sometimes known as 'dependent variables', 'outcomes' or 'phenotypes'.

$$\frac{\partial l}{\partial \alpha} = -\sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\frac{\partial l}{\partial \beta} = -\sum_{i=1}^{n} 2x_i(y_i - \alpha - \beta x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2.$$

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

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i/n = \hat{\beta}(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i/n}{(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n)}$$

The least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are often expressed as

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \tag{15}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{16}$$

where

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{17}$$

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{18}$$

$$S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 . \tag{19}$$

(Although S_{YY} is not needed in (15) and (16), it will be used later.)

Note that estimation by least squares requires no assumptions about the distributions of $Y_1, \ldots, Y_n$. Since $Y_1, \ldots, Y_n$ are random variables, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are also random variables. We can calculate their expectations using (16)–(18) and (11), and by noting that $x_1, \ldots, x_n$ are *fixed*. Show that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β, respectively.

$$\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\frac{1}{S_{XX}}\sum_{i=1}^{n}(x_i - \bar{x})\left(Y_i - \bar{Y}\right)\right] \\ &= \frac{1}{S_{XX}}\sum_{i=1}^{n}(x_i - \bar{x})\,E\left(Y_i - \bar{Y}\right) \\ &= \frac{1}{S_{XX}}\sum_{i=1}^{n}(x_i - \bar{x})\,\beta\,(x_i - \bar{x}) \\ &= \beta. \end{aligned} \tag{20}$$

Similarly,

$$\begin{aligned} E(\hat{\alpha}) &= E\left(\bar{Y} - \hat{\beta}\bar{x}\right) \\ &= (\alpha + \beta\bar{x}) - \beta\bar{x} \\ &= \alpha. \end{aligned} \tag{21}$$

Exercise

1. Show that $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} y_i(x_i-\bar{x})$
2. Hence show that $\text{var}(\hat{\beta}) = \frac{\sigma^2}{S_{XX}}$
3. Given that $\bar{y}$ and $\hat{\beta}$ are independent, show that $\text{var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)$

5.1.2 The normal linear regression model

In general, we need to do more than just find (point) estimates of the parameters α and β; we need to test hypotheses about them and to construct confidence intervals for them. In order to do this, we need to make assumptions about the distributions of $Y_1,\ldots,Y_n$. We shall suppose that $Y_1,\ldots,Y_n$ are independent, normally distributed with the same variance, and satisfy (11), i.e.

$$Y_i \sim N(\alpha+\beta x_i,\sigma^2), \qquad i=1,\ldots,n \qquad \text{with } Y_1,\ldots,Y_n \text{ independent.} \tag{22}$$

Write down the likelihood function of α and β.

$L(\alpha,\beta;(x_1,y_1),\ldots,(x_n,y_n)) =$

$$L(\alpha,\beta;(x_1,y_1),\ldots,(x_n,y_n)) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-(y_i-\mu_i)^2/(2\sigma^2)] = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right)$$

Hence obtain the log-likelihood function

$l(\alpha,\beta;(x_1,y_1),\ldots,(x_n,y_n)) =$

$$l = \log(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i-(\alpha-\beta x_i)\}^2$$

Note two things about this log-likelihood:

(i) it can be maximised w.r.t. α and β when σ^2 is unknown,

(ii) maximising the log-likelihood is equivalent to minimising $S(\alpha,\beta)$.

Property (ii) means that, *for the normal linear regression model* (22), the maximum likelihood estimates are equal to the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$.

Since (from (16)–(18)) $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of the normal random variables $Y_1,\ldots,Y_n$, $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. It follows from this, together with (20)–(21) and the results from the above exercise for $\text{var}(\hat{\beta})$ and $\text{var}(\hat{\alpha})$, that

$$\hat{\beta} \sim N\left(\beta,\frac{\sigma^2}{S_{XX}}\right) \tag{23}$$

$$\hat{\alpha} \sim N\left(\alpha,\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)\right). \tag{24}$$

Since the variance σ^2 in (22) is usually unknown, it must be estimated from the data. The quantity s^2 is defined by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \tag{25}$$

where

$$\hat{y}_i = \hat{\alpha}+\hat{\beta}x_i \tag{26}$$

is the *fitted value* corresponding to x_i. Calculation shows that s^2 is an unbiased estimator of σ^2. Further calculation (postponed to Honours) shows that

$$(n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2} \tag{27}$$

and

$$s^2 \text{ is independent of } \hat{\alpha} \text{ and } \hat{\beta} \tag{28}$$

[**Warning**: $\hat{\alpha}$ and $\hat{\beta}$ are *not* independent!]

Putting (23)–(24) and (27) together, we obtain

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{s^2}{S_{XX}}}} \sim t_{n-2} \tag{29}$$

$$\frac{\hat{\alpha}-\alpha}{\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}} \sim t_{n-2}. \tag{30}$$

Results (29)–(30) are the basis of inference on α and β. In particular, they enable us to (use R to) calculate confidence intervals for α and β. For example, 95% confidence intervals for β and α are

$$\hat{\beta} \pm t_{n-2;0.025}\sqrt{\frac{s^2}{S_{XX}}} \tag{31}$$

$$\hat{\alpha} \pm t_{n-2;0.025}\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)} \tag{32}$$

The quantities

$$\text{s.e.}\left(\hat{\beta}\right) = \sqrt{\frac{s^2}{S_{XX}}} \tag{33}$$

$$\text{s.e.}\left(\hat{\alpha}\right) = \sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)} \tag{34}$$

are provided by R. These quantities are the *standard errors* of $\hat{\beta}$ and $\hat{\alpha}$, respectively.

5.2 Regression using R

Linear regression can be carried out in R using the `lm` (linear modelling) command, which is designed for fitting quite general linear models. Recall that the linear regression model (12) is a linear model, in that the parameters α and β (and the error term) enter the model in a linear way.

The use of `lm` can be illustrated by the following example (which has been chosen partly because it is simple enough that the calculations could easily be done with a pocket calculator).

Example

The following measurements give the concentration of chlorine (in parts per million) in a swimming pool at various times after chlorination treatment.

time (hours)	2	4	6	8	10	12
chlorine (p.p.m.)	1.8	1.5	1.4	1.1	1.1	0.9

Consider the following R session.

```
> pool
        [,1] [,2] [,3] [,4] [,5] [,6]
time     2.0  4.0  6.0  8.0 10.0 12.0
```

```
chlorine  1.8  1.5  1.4  1.1  1.1  0.9
> swim<-lm(chlorine~time)
> swim

Call:
lm(formula = chlorine ~ time)

Coefficients:
(Intercept)         time
    1.90000     -0.08571
```

First the data (in `pool`) are inspected. Then the object `swim` is defined as the linear model in which `chlorine` is regressed on `time`, i.e. the linear regression equation

$$\text{chlorine} = \alpha + \beta\,\text{time}$$

is fitted to the data by least squares. Then `swim` contains the parameter estimates $\hat{\alpha} = 1.9$ and $\hat{\beta} = -0.0857$.

More detail of this regression is given by applying the `summary` command to `swim`, as follows.

```
> summary(swim)

Call:
lm(formula = chlorine ~ time)

Residuals:
       1        2        3        4        5        6
 0.07143 -0.05714  0.01429 -0.11429  0.05714  0.02857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.900000   0.074642  25.455 1.41e-05 ***
time        -0.085714   0.009583  -8.944 0.000864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08018 on 4 degrees of freedom
Multiple R-Squared: 0.9524, Adjusted R-squared: 0.9405
F-statistic:    80 on 1 and 4 DF,  p-value: 0.0008642
```

The *residuals* $r_1, \ldots, r_n$ are the differences between the data $y_1, \ldots, y_n$ and the fitted values $\hat{y}_i$ given by (26), i.e.

$$r_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n. \tag{35}$$

The residuals have already appeared in the definition (25) of s^2. As we shall see later, they play an important role in assessing how well the model fits the data.

The above section which is labelled ‘`Coefficients:`’ has the following columns:

(i) a name for the parameter (`(Intercept)` for α; `time` for β);

(ii) the estimates ($\hat{\alpha}$ and $\hat{\beta}$) of α and β;

(iii) the standard errors of $\hat{\alpha}$ and $\hat{\beta}$;

(iv) the values of the t-statistics given by the left hand sides of (29) and (30);

(v) the p-values of these t-statistics in tests of $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$ and of $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, respectively (note that both of these alternative hypotheses are 2-sided);

(vi) a ‘star rating’ of these p-values.

The `Residual standard error` is s, where s^2 is defined in (25).

The quantity `Multiple R-Squared` is the squared (sample) correlation coefficient r^2 between x and Y. It is defined by

$$\begin{aligned} r^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} \\ &= \frac{\left\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right\}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}. \end{aligned} \tag{36}$$

It can be shown that

$$r^2 \leq 1. \tag{37}$$

An important interpretation of r^2 is as the proportion of the variation in $y_1, \ldots, y_n$ which is explained by the model. The closer r^2 is to 1, the better the model explains the variation in the data.

In R, the fitted values $\hat{y}_1, \ldots, \hat{y}_n$ and the residuals $r_1, \ldots, r_n$ can be obtained using `fitted` and `residuals`, respectively, e.g. for the chlorine data

```
> swimfit<-fitted(swim)
> swimfit
        1         2         3         4         5         6
1.7285714 1.5571429 1.3857143 1.2142857 1.0428571 0.8714286
> swimres<-residuals(swim)
> swimres
          1           2           3           4           5           6
 0.07142857 -0.05714286  0.01428571 -0.11428571  0.05714286  0.02857143
```

obtains and displays the fitted values and then the residuals.

Example:

In the chlorine example, we can use R to obtain 95% confidence intervals for α and β based on (31)–(32) as follows.

```
> summary(swim)

Call:
lm(formula = chlorine ~ time)

Residuals:
       1        2        3        4        5        6
 0.07143 -0.05714  0.01429 -0.11429  0.05714  0.02857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.900000   0.074642  25.455 1.41e-05 ***
time        -0.085714   0.009583  -8.944 0.000864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08018 on 4 degrees of freedom
Multiple R-Squared: 0.9524, Adjusted R-squared: 0.9405
F-statistic:    80 on 1 and 4 DF,  p-value: 0.0008642

> alphalimits <-c(1.900000+qt(0.025,4)*0.074642,1.900000-qt(0.025,4)*0.074642)
> alphalimits
[1] 1.692761 2.107239
> betalimits <-c(-0.085714+qt(0.025,4)*0.009583,-0.085714-qt(0.025,4)*0.009583)
```

```
> betalimits
[1] -0.11232067 -0.05910733
```

giving 95% confidence intervals of $(1.693, 2.107)$ for α and $(-0.1123, -0.0591)$ for β.

5.2.1 Regression through the origin

Sometimes we may wish to assume that the regression line passes through the origin, i.e. that $\alpha = 0$ in (11). The way to do this in the R `lm` command is to use `-1` to show that the constant term (i.e. the coefficient of 1) should be removed. Here is an illustration using the chlorine data (where it would not be sensible to fit a line through the origin!).

```
> swimnoc<-lm(chlorine~time-1)
> swimnoc

Call:
lm(formula = chlorine ~ time - 1)

Coefficients:
  time
0.1335
```

(Note that, in this case, forcing $\alpha = 0$ has changed $\hat{\beta}$ considerably.)

5.3 Confidence intervals and prediction intervals

One of the main purposes in fitting a regression line (of Y on x, say) to data $(x_1, y_1), \ldots, (x_n, y_n)$ is to be able to say something about the values Y_0 of the response variable corresponding to any given value x_0 of the predictor variable. Note that x_0 need not be one of $x_1, \ldots, x_n$. You may find it helpful to think of $y_1, \ldots, y_n$ as observations taken in the *past*, which we use (in the present) to fit the sample regression line

$$y = \hat{\alpha} + \hat{\beta}x, \tag{38}$$

whereas Y_0 is the random variable of *future* observations of Y with $x = x_0$.

There are two types of interval which are of interest:

(i) confidence intervals for $E(Y_0)$,

(ii) prediction intervals for Y_0.

(i) **Confidence intervals for $E(Y_0)$:**
The regression equation (9) gives

$$E(Y_0) = \alpha + \beta x_0,$$

so that $E(Y_0)$ is *fixed* and depends on the unknown parameters α and β. It is sensible to estimate $E(Y_0)$ by

$$\hat{E}(Y_0) = \hat{\alpha} + \hat{\beta}x_0.$$

We know $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, so that

$$\hat{E}(Y_0) = \bar{y} + \hat{\beta}(x_0 - \bar{x}).$$

It can be shown that $\bar{y}$ and $\hat{\beta}$ are independent, and given the distribution of $\hat{\beta}$ from (23), it follows that

$$\text{var}\left[\hat{E}(Y_0)\right] = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]$$

Estimating σ^2 by s^2 and standardizing, we therefore obtain

$$\frac{\hat{E}(Y_0) - E(Y_0)}{\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}} \sim t_{n-2},$$

from which a 95% confidence interval for $E(Y_0)$ is

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}.$$

You do not need to memorise this. However, you should note its key feature:

confidence intervals for $E(Y_0)$ become wider as x_0 moves away from $\bar{x}$.

Thus estimates of $E(Y_0)$ become less reliable, the further x_0 is from $\bar{x}$. This is one reason why it is unwise to extrapolate Y outside the range of $x_1, \ldots, x_n$.

(ii) **Prediction intervals for Y_0:**
Whereas a confidence interval for $E(Y_0)$ considered above is a (random) interval which we would like to contain the (fixed) mean of Y_0, a prediction interval for Y_0 is a (random) interval which we would like to contain the (random) value of a (future) observation y_0 of Y_0. A prediction interval for Y_0 is centred on the fitted value $\hat{\alpha} + \hat{\beta}x_0$ corresponding to x_0. We can express a future estimate of Y_0 as

$$\hat{Y_0} = \left(\hat{\alpha} + \hat{\beta}x_0\right) + \epsilon_0$$

where the error ϵ_0 captures the additional randomness in predicting a single future observation, rather than the expected value, $E(Y_0)$; of course, we take $\hat{\epsilon} = 0$, so that $\hat{Y_0} = \hat{E}(Y_0)$, but their variances differ. Since the ('future') random variable Y_0 is independent of the ('past') random variables $Y_1, \ldots, Y_n$ which are used to fit the model, ϵ_0 is independent of $\hat{\alpha} + \hat{\beta}x_0$, and so

$$\begin{aligned} \operatorname{var}(\hat{Y_0}) = \operatorname{var}\left(\left[\hat{\alpha} + \hat{\beta}x_0 + \epsilon_0\right]\right) &= \operatorname{var}\left(\hat{\alpha} + \hat{\beta}x_0\right) + \operatorname{var}(\epsilon_0) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right) + \sigma^2 \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right). \end{aligned} \tag{39}$$

It follows (after some calculation) that a 95% prediction interval for Y_0 is

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}.$$

You do not need to memorise this. However, you should note its key features:

(a) prediction intervals for Y_0 become wider as x_0 moves away from $\bar{x}$,

(b) prediction intervals for Y_0 are wider than the corresponding confidence intervals for $E(Y_0)$.

The intuitive explanation of (b) is that prediction intervals take into account the variability of the 'future' observation Y_0, whereas confidence intervals are concerned only with the mean value $E(Y_0)$. This is made rigorous in (39).

Confidence intervals and prediction intervals can be produced in R using the `predict.lm` command. For example, the following portion of an R session used the chlorine data to give such intervals for the chlorine residual at the times (2,4,6,8,10,12 hours after chlorination) in the data set. (Recall that `swim` contains the results of the linear regression.)

```
> predict.lm(swim,interval=c("confidence"))
        fit       lwr      upr
1 1.7285714 1.5674575 1.889685
2 1.5571429 1.4361854 1.678100
3 1.3857143 1.2910190 1.480410
4 1.2142857 1.1195904 1.308981
5 1.0428571 0.9218997 1.163815
```

```
6 0.8714286 0.7103147 1.032542
> predict.lm(swim,interval=c("prediction"))
        fit       lwr      upr
1 1.7285714 1.4537746 2.003368
2 1.5571429 1.3037927 1.810493
3 1.3857143 1.1437994 1.627629
4 1.2142857 0.9723709 1.456201
5 1.0428571 0.7895070 1.296207
6 0.8714286 0.5966318 1.146225
```

To obtain confidence intervals and prediction intervals at other times (e.g. 9 and 11 hours after chlorination), we first put these times into a dataframe (e.g. `newswim`).

```
> newswim<-data.frame(time=c(9,11))
> predict.lm(swim,newswim,interval=c("confidence"))
        fit      lwr      upr
1 1.1285714 1.023258 1.233885
2 0.9571429 0.817192 1.097094
> newswim<-data.frame(time=c(9,11))
> predict.lm(swim,newswim,interval=c("prediction"))
        fit       lwr      upr
1 1.1285714 0.8823061 1.374837
2 0.9571429 0.6941945 1.220091
```

5.4 Checking the assumptions

The linear regression model (22) is

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \ldots, n \qquad \text{with } Y_1, \ldots, Y_n \text{ independent.}$$

Thus when fitting the linear regression model (22), we are making the assumptions

(i) $Y_1, \ldots, Y_n$ are independent,

(ii) $Y_1, \ldots, Y_n$ are normally distributed,

(iii) $E(Y_i) = \alpha + \beta x_i$, i.e. the mean of Y_i is a linear function of x_i,

(iv) $Y_1, \ldots, Y_n$ have the *same* variance.

It is important to check that these assumptions hold.

Before fitting the linear regression model, it is sensible to plot the data.

Example (Chemical data):

The following measurements give the masses of desired product produced (from 100 g of reagent) by a certain chemical reaction at various temperatures.

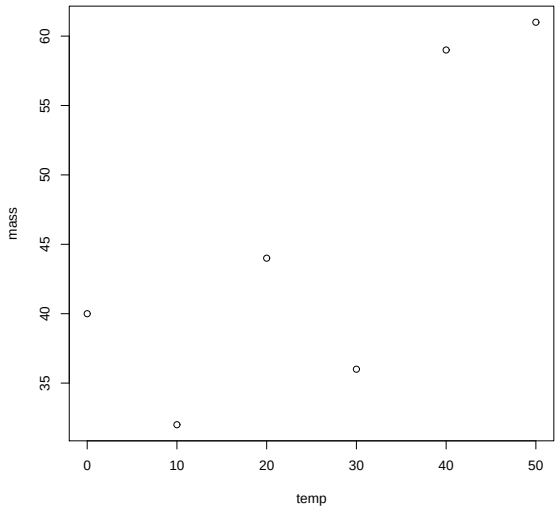
mass (g)	40	32	44	36	59	61
temperature (° C)	0	10	20	30	40	50

These measurements have been placed in the object `chemdata` in R.

```
> chemdata
     [,1] [,2] [,3] [,4] [,5] [,6]
temp    0   10   20   30   40   50
mass   40   32   44   36   59   61
```

The R command

```
> plot(temp,mass,xlab="temp",ylab="mass")
```



gives

The plot is roughly linear, so that the linear regression model is plausible for these data. Once we have fitted the model to the data, we should check that assumptions (i)–(iv) above are reasonable. The key to such checking lies in the residuals $r_1, \ldots, r_n$, which were defined in (35) by

$$r_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n,$$

and measure the discrepancy between the model and the data.

Various residual plots should be examined to check the model assumptions of independence and equal variance. Applying the R command `plot` to the object created by `lm` produces 4 plots. The first 3 are particularly useful. They are

(i) a plot (labelled '`Residuals vs Fitted`' in R) of residuals $r_1, \ldots, r_n$ against fitted values $\hat{y}_1, \ldots, \hat{y}_n$;

(ii) a Q-Q plot (labelled '`Normal Q-Q plot`' in R) of the standardised residuals $r_1^*, \ldots, r_n^*$;

(iii) a plot (labelled '`Scale-Location plot`' in R) of $\sqrt{|r_1^*|}, \ldots, \sqrt{|r_n^*|}$ against fitted values $\hat{y}_1, \ldots, \hat{y}_n$.

Here

$$r_i^* = \frac{r_i}{\sqrt{\widehat{\mathrm{var}}(r_i)}}$$

is the ith *standardised residual*, which is obtained by scaling r_i to have variance 1. If the assumptions hold, then

(i) the plot of residuals against fitted values should show no obvious systematic pattern — the residuals should be well scattered above and below zero, with vertical spread (indicating variance) which does not depend much on the fitted value;

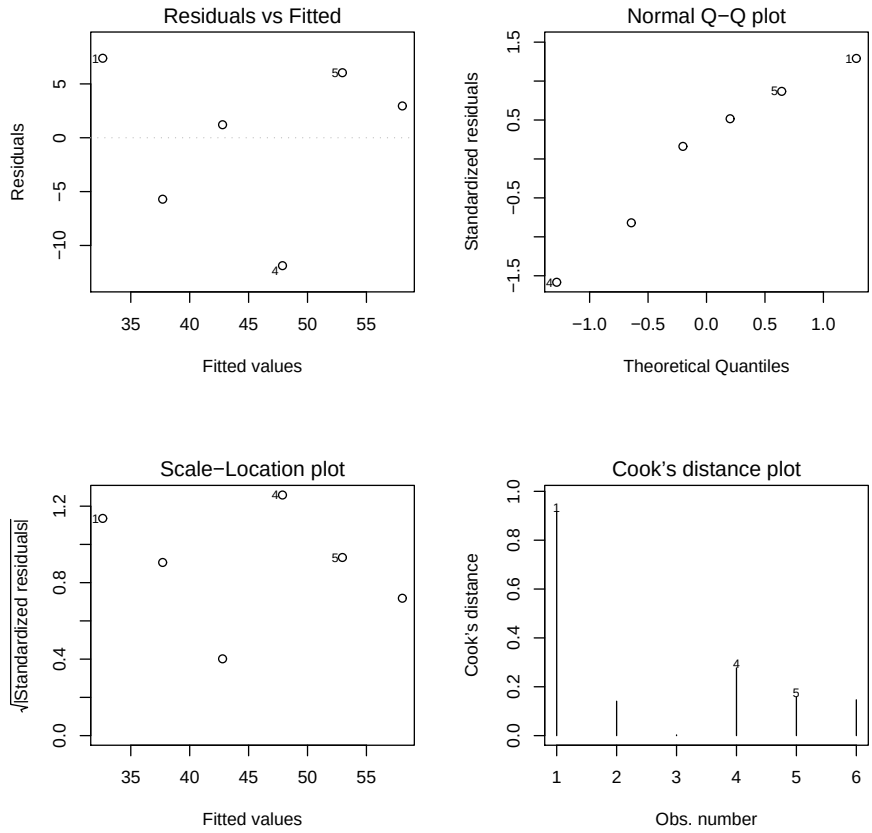
(ii) the Q-Q plot of the standardised residuals should be approximately linear (reflecting the fact that $r_1, \ldots, r_n$ should be *approximately* like a random sample from some normal distribution with mean zero);

(iii) the plot of $\sqrt{|r_1^*|}, \ldots, \sqrt{|r_n^*|}$ against fitted values $\hat{y}_1, \ldots, \hat{y}_n$ should show no trend.

Example (Chemical data):

The R commands

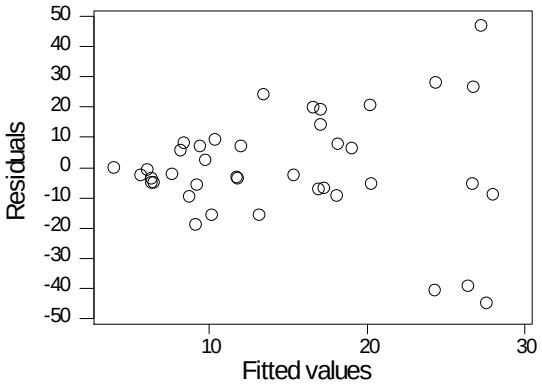
```
> chem<-lm(mass~temp)
> par(mfrow=c(2,2))
> plot(chem)
```



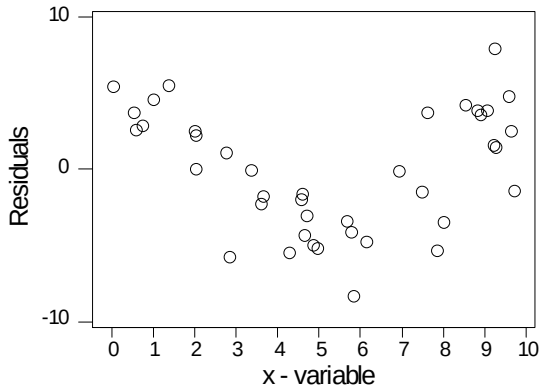
give

None of these plots gives grounds for doubting that the assumptions of the linear regression model hold.

Here are two examples of problematic plots:



This shows a clear relationship between variance and fitted values, indicating that the assumption (numbered (iv) above) of constant variance is violated.



This plot of residuals against the explanatory variable x shows a clear systematic pattern. This indicates that the assumption of linearity (assumption (iii) above) is not valid for this data set. The plot indicates that a quadratic model might be better for these data:

$$E(Y_i) = \alpha + \beta_1 x_i + \beta_2 x_i^2, \qquad i = 1, \ldots, n.$$

5.4.1 Transformation

If residual plots indicate that the linear regression model is not appropriate, then sometimes transformation of the response variable Y or the explanatory variable x can linearise the relationship between Y and x. Three common transformations which it is worth trying are square root ($t \mapsto \sqrt{t}$ for $t \geq 0$), exponential ($t \mapsto \exp t$) and logarithm ($t \mapsto \log t$ for $t > 0$).

Example 1 (Log transform):
Suppose

$$y = Ax^\beta$$

How can you transform x and/or y to turn this into a linear model?
Taking logarithms gives

$$\log y = \alpha + \beta \log x,$$

where $\alpha = \log A$, so that $\log y$ is a linear function of $\log x$.

Example 2:
Suppose

$$y = \frac{\alpha}{1 + \beta x}$$

How can transformation be used to turn this into a linear model?
Taking the reciprocal of each side:

$$\frac{1}{y} = \frac{1}{\alpha} + \frac{\beta}{\alpha} x,$$

which expresses $1/y$ as a linear function of x.

The choice of transformation is a matter of experience and experiment. The scatter plot of $(x_1, y_1), \ldots, (x_n, y_n)$ and the residual plots may suggest a suitable transformation.

Note that transformation of the response variable has an effect on the validity of the assumption of normality, e.g. if Y is normally distributed then $\log Y$ is not. As a *very rough* guide,

(i) if the residual plots indicate normality with constant variance then transform the explanatory variable,

(ii) if the residual plots do not indicate normality with constant variance then transform the response variable.