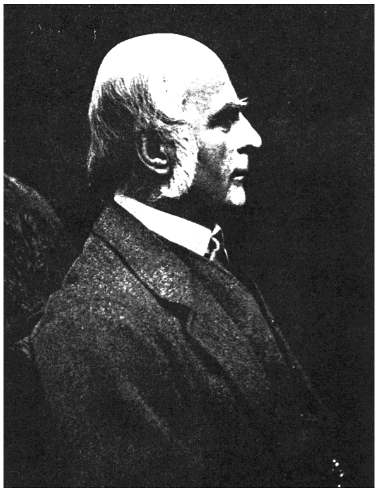
5 Linear regression

Regression is all about relationships between variables. In its simplest form, it constitutes a technique for modelling a relationship between two variables.



Sir Francis Galton 1822 - 1911

It was the pioneering work of Sir Francis Galton in the 1880s that gave rise to the technique, the original idea being the direct result of an experiment on sweet peas. He noticed that the seeds of the progeny of parents with seeds heavier than average were also heavier than average, but the difference was not as pronounced; the same effect was true for the seeds of the progeny of parents with light seeds, where again the differences from the average were not as great. He called this phenomenon *reversion* and wrote that the mean weight "reverted, or regressed, toward mediocrity".

5.1 Data and Questions

Data set 5.1 *Divorces in England and Wales*

The table below gives the annual number of divorces recorded in England and Wales between 1975 and 1980 (*Marriage and Divorce Statistics*, Office of Population Censuses and Surveys, HMSO).

Table 5.1 Divorces in England and Wales

Year	1975	1976	1977	1978	1979	1980
Divorces (thousands)	120.5	126.7	129.1	143.7	138.7	148.3

They are plotted below.

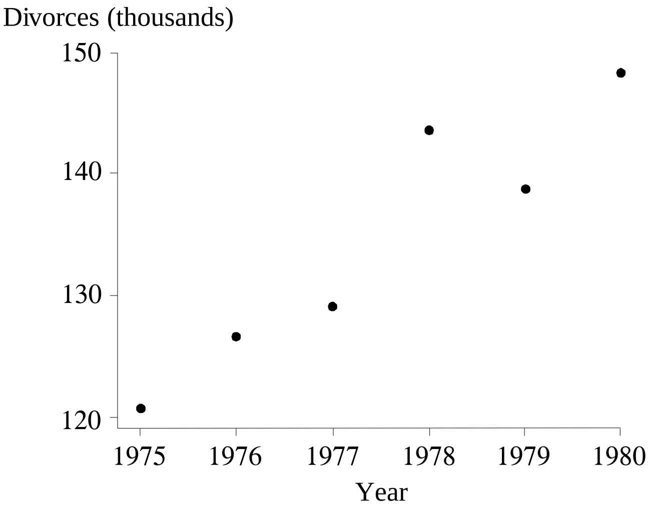


Figure 5.1 Divorces in England and Wales

The plot suggests a roughly linear trend with some random scatter.
□

Data set 5.2 *Olympic Sprint Times*

Table 5.2 gives the times in seconds recorded by the winners in the finals of the men's sprint events (100, 200, 400, 800 and 1500 metres) at each of the 21 Olympic Games from 1900 to 2004 along with the heights above sea level of the different venues. Obviously the years 1916, 1940, 1944 are missing since the Olympic games were not held during the World Wars.

Table 5.2 Winning times for Olympic running events

Year	*Venue*	*100m*	*200m*	*400m*	*800m*	*1500m*	*Altitude*
1900	Paris	10.80	22.20	49.40	121.40	246.00	25
1904	St Louis	11.00	21.60	49.20	116.00	245.40	455
1908	London	10.80	22.40	50.00	112.80	243.40	8
1912	Stockholm	10.80	21.70	48.20	111.90	236.80	46
1920	Antwerp	10.80	22.00	49.60	113.40	241.80	3
1924	Paris	10.60	21.60	47.60	112.40	233.60	25
1928	Amsterdam	10.80	21.80	47.80	111.80	233.20	8
1932	Los Angeles	10.30	21.20	46.20	109.80	231.20	340
1936	Berlin	10.30	20.70	46.50	112.90	227.80	115
1948	London	10.30	21.10	46.20	109.20	225.20	8
1952	Helsinki	10.40	20.70	45.90	109.20	225.20	25
1956	Melbourne	10.50	20.60	46.70	107.70	221.20	3
1960	Rome	10.20	20.50	44.90	106.30	215.60	66
1964	Tokyo	10.00	20.30	45.10	105.10	218.10	45
1968	Mexico City	9.95	19.83	43.80	104.30	214.90	7349
1972	Munich	10.14	20.00	44.66	105.90	216.30	1699
1976	Montreal	10.06	20.23	44.26	103.50	219.20	104
1980	Moscow	10.25	20.19	44.60	105.40	218.40	497
1984	Los Angeles	9.99	19.80	44.27	103.00	212.50	340
1988	Seoul	9.92	19.75	43.87	103.45	215.96	111
1992	Barcelona	9.96	20.01	43.50	103.66	220.12	3
1996	Atlanta	9.84	19.32	43.49	102.58	215.78	1026
2000	Sydney	9.87	20.09	43.84	105.08	212.07	3
2004	Athens		19.79	44.00	104.45	214.20	505

We would like to construct a predictive model for the winning times at the 2008 games in Beijing (and the 2012 games in London). Clearly there is now a possibility of more than one explanatory variable.

□

5.2 Fitting the model

5.2.1 Notation

The equation of a regression line takes the form

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

and gives rise to the regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $Y_1, Y_2, \ldots, Y_n$ are observable random variables, the values $x_1, x_2, \ldots, x_n$ are specified and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ non-observable random variables. The data are, therefore, in the form of n ordered pairs (x_i, Y_i), $i = 1, \ldots, n$.

x_i is called an *explanatory variable* or, sometimes, a *predictor*;
Y_i is usually referred to as the *response*;
ε_i is called a *departure* or *residual.*

Each ε_i has distribution $N(0, \sigma^2)$ and they are pairwise uncorrelated. Clearly

$$E\left[Y_i\right] = \alpha + \beta x_i, \quad V\left[Y_i\right] = \sigma^2, \quad i = 1, 2, \ldots, n.$$

We are going to be concerned with linear model in its more general form involving several explanatory variables. The most convenient way of doing that is to write down the model in matrix notation. Equation (1) is a set of simultaneous equations which can be written more concisely as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Data set 5.3 *AIDS data for the USA*

These data are for AIDS incidence in the USA, adjusted for reporting delays. The data are taken from Rosenberg, P.S. and Gail, M.H. (1991): Backcalculation of flexible linear models of the Human Immunodeficiency Virus infection curve. *Applied Statistics*, **40**, 269-282.

Newly reported cases are recorded quarterly and the variable *Time* therefore counts 3-monthly periods, staring with the first quarter in 1982.

Table 5.3 Reported AIDS cases in the USA

Quarter	*Time*	*Cases*	*Quarter*	*Time*	*Cases*	*Quarter*	*Time*	*Cases*
1982:1	1	185	1984:2	10	1369	1986:2	18	4321
1982:2	2	200	1984:3	11	1563	1986:3	19	4863
1982:3	3	293	1984:4	12	1726	1986:4	20	5192
1982:4	4	374	1985:1	13	2142	1987:1	21	6155
1983:1	5	554	1985:2	14	2525	1987:2	22	6816
1983:2	6	713	1985:3	15	2951	1987:3	23	7491
1983:3	7	763	1985:4	16	3160	1987:4	24	7726
1983:4	8	857	1986:1	17	3819	1988:1	25	8483
1984:1	9	1147						

The scatterplot shows that the trend is not linear.

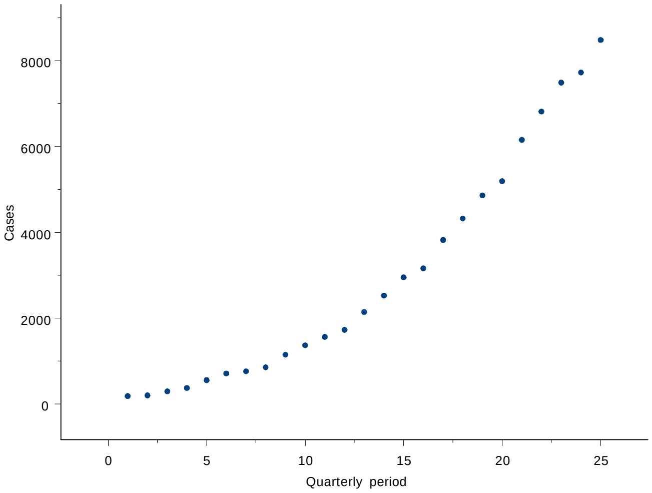


Figure 5.2 Incidence of AIDS cases in the USA against time

The plot has all the appearance of showing a functional relationship between AIDS incidence and time. One might, for example, try to fit an exponential function or some kind of power law to model the growth curve. For example

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

□

At this point you should be careful to note that in general, a model is **not** called a linear model because of fitting linear functions of x. It is called a linear model because it is *linear in the unknown parameters* $\boldsymbol{\theta}$. Thus

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

still classes as a linear model because it is linear in α, β_1, β_2. Furthermore, it is clear that all such linear models can be written in the form of Equation (2) by suitably defining $\mathbf{X}$ and $\boldsymbol{\theta}$. In this notation

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}, \quad V(\mathbf{Y}) = \sigma^2 \mathbf{I},$$

where $\mathbf{I}$ is the identity.

Example 5.1

The model above,

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

can be written in the form of Equations (2) by defining

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

■

5.2.2 The likelihood function

The likelihood function has the simple form

$$L\left(\alpha, \beta, \sigma^2; \mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right] = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right].$$

Maximisation of L corresponds to minimisation of $\sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$. For this reason, the method of optimisation is also known as the *method of least squares.*

$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ can be minimised whatever the underlying distribution; it is only optimal in the sense of maximum likelihood when the underlying distribution is normal.

We need to minimise $(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Differentiating with respect to $\boldsymbol{\theta}$ and equating to $\mathbf{0}$ gives

$$2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{0},$$

so the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}^T\mathbf{Y}. \qquad (3)$$

Provided $\mathbf{X}^T\mathbf{X}$ is non-singular, this can be written

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

It must be a minimum in $(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$, and hence a maximum in the likelihood, because $\boldsymbol{\theta}$ is unconstrained and therefore $(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ could be arbitrarily large.

Equation (3) is a set of linear simultaneous equations called the *normal equations* for the linear model.

Note that, before you can go ahead and estimate these coefficients, you need to be extra careful about validating the assumptions which underlie the method of fitting used. These are

- the variance of the response is constant (i.e. the variance does not depend upon x);
- the distribution of the response is normal;
- the ε_i are independent;
- the model you are trying to fit is a valid one.

The first three of these assumptions are encapsulated by the statement $\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2\mathbf{I}\right)$. Under such assumptions the method of least squares and the method of maximum likelihood are equivalent.

5.2.3 Simple linear regression

The matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, so that

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}$$

so that

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \\ -\sum x_i \sum Y_i + n\sum x_i Y_i \end{pmatrix}.$$

It is more usual to write $n\sum x_i^2 - \left(\sum x_i\right)^2$ as $n\sum\left(x_i - \overline{x}\right)^2$ and $-\sum x_i \sum Y_i + n\sum x_i Y_i$ as $n\sum\left(x_i - \overline{x}\right)Y_i$. The expressions for $\widehat{\alpha}$ and $\widehat{\beta}$ then simplify to

$$\begin{aligned} \widehat{\beta} &= \frac{\sum\left(x_i - \overline{x}\right)Y_i}{\sum\left(x_i - \overline{x}\right)^2}, \\ \widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{x}. \end{aligned}$$

Example 5.3 (Example 5.1 revisited) *Divorces*

Using the data given in Table 5.1,

x_i	1975	1976	1977	1978	1979	1980
y_i	120.50	126.70	129.10	143.70	138.70	148.30
$(x_i - \overline{x})$	−2.50	−1.50	−0.50	0.50	1.50	2.50
$(x_i - \overline{x})^2$	6.25	2.25	0.25	0.25	2.25	6.25
$(x_i - \overline{x})\,y_i$	−301.25	−190.05	−64.55	71.85	208.05	370.75

The totals for the last two rows are 17.5 and 94.8 respectively. We therefore obtain

$$\begin{aligned} \widehat{\beta} &= 5.417, \\ \widehat{\alpha} &= -10577.9 \end{aligned}$$

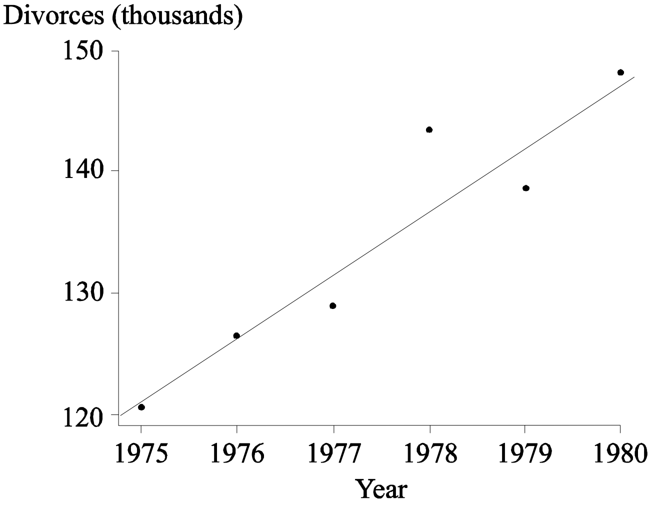


Figure 5.3 Divorces in England and Wales with fitted line

□

5.2.4 Linear transformations - some general results

We have seen that the normal equations for the linear model,(3)

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}^T\mathbf{Y}$$

may be written in the form

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

when $\mathbf{X}^T\mathbf{X}$ is non-singular. $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is a linear transformation of the random vector $\mathbf{Y}$ and it is worth obtaining some general results for such transformations.

The results we shall need are:

If $\mathbf{W} = \mathbf{L}\mathbf{Y}$, then

(i)

$$E(\mathbf{W}) = \mathbf{L}E(\mathbf{Y})$$

This a direct result of expectation being a linear operator.

(ii)

$$V(\mathbf{W}) = \mathbf{L}V(\mathbf{Y})\mathbf{L}^T$$

To see where this comes from, we first need to be clear in our understanding of $V(\mathbf{W})$, the covariance matrix of a random vector.

The covariance matrix, $V(\mathbf{W})$, is

$$\begin{pmatrix} V(W_1) & C(W_1, W_2) & \cdots & C(W_1, W_n) \\ C(W_2, W_1) & V(W_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ C(W_n, W_1) & \cdots & \cdots & V(W_n) \end{pmatrix}$$

or, in other words,

$$V(\mathbf{W})_{ij} = C\left(W_i, W_j\right) = E\left[\left(W_i - E\left(W_i\right)\right)\left(W_j - E\left(W_j\right)\right)\right].$$

The matrix may be written in the form

$$V(\mathbf{W}) = E\left[\left(\mathbf{W} - E(\mathbf{W})\right)\left(\mathbf{W} - E(\mathbf{W})\right)^T\right].$$

Thus

$$\begin{aligned} V(\mathbf{W}) &= E\left[\left(\mathbf{W} - E(\mathbf{W})\right)\left(\mathbf{W} - E(\mathbf{W})\right)^T\right] \\ &= E\left[\mathbf{L}\left(\mathbf{Y} - E(\mathbf{Y})\right)\left(\mathbf{Y} - E(\mathbf{Y})\right)^T \mathbf{L}^T\right] \\ &= \mathbf{L}E\left[\left(\mathbf{Y} - E(\mathbf{Y})\right)\left(\mathbf{Y} - E(\mathbf{Y})\right)^T\right]\mathbf{L}^T \\ &= \mathbf{L}V(\mathbf{Y})\mathbf{L}^T. \end{aligned}$$

As we shall see in the next section, these results make it easy to deduce some of the basic properties of least-squares estimators.

5.2.5 Properties of the estimators

When $\mathbf{X}^T\mathbf{X}$ is non-singular, the estimators have useful properties.

Property (i) $\widehat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$.
□

Proof

$$\begin{aligned} E\left[\widehat{\boldsymbol{\theta}}\right] &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E\left[\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \\ &= \boldsymbol{\theta} \end{aligned}$$

■

Note that this proof does not require the departures $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ either to have equal variances or to be uncorrelated; it simply uses the fact that expectation is a linear operator.

Property (ii) The covariance matrix of $\widehat{\boldsymbol{\theta}}$ is $\boldsymbol{\Sigma}_\theta = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.
□

Proof

For simplicity, let $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$; then $\widehat{\boldsymbol{\theta}} = \mathbf{BY}$. Thus

$$\begin{aligned}
\boldsymbol{\Sigma}_\theta &= V\left(\mathbf{BY}\right) \\
&= \mathbf{B}V(\mathbf{Y})\mathbf{B}^T \\
&= \mathbf{B}\boldsymbol{\Sigma}_Y\mathbf{B}^T \\
&= \sigma^2\mathbf{BIB}^T \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]^T \\
&= \sigma^2\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]^T = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1},
\end{aligned}$$

since $\mathbf{X}^T\mathbf{X}$ $\left(\text{and therefore } \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$is symmetric.

■

This is an important result because it shows that the diagonal entries of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ provide information about the precision of the estimates.

Property (iii) $\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ is a linear transformation of the normally distributed random vector $\mathbf{Y}$, so that

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

□

5.3 An alternative derivation of the estimating equations

We have already seen that the model can be written in the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ can be a $n \times p$ matrix of known constants and $\boldsymbol{\theta}$ can be a vector of p unknown parameters. We have also seen that we need to minimise

$$\sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) = \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|^2.$$

It follows, therefore, that the least squares estimator for $\boldsymbol{\theta}$ is a statistic $\widehat{\boldsymbol{\theta}} = \mathbf{h}(\mathbf{Y})$ such that $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|$ achieves its minimum when $\boldsymbol{\theta} = \mathbf{h}(\mathbf{y})$.

Let $V_{\mathbf{X}}$ be the vector subspace spanned by columns of $\mathbf{X}$, and let $\operatorname{rank}(\mathbf{X}) = r \leq p$ so that $\dim(V_{\mathbf{X}}) = r$. Then

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta} \in V_{\mathbf{X}}$$

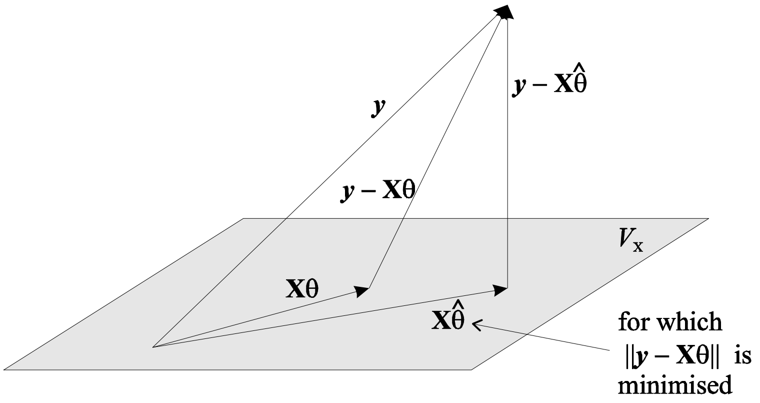


Figure 5.4

Clearly $\|\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\|$ achieves its minimum over $\boldsymbol{\theta} \in \mathbb{R}^p$ when $\mathbf{X}\boldsymbol{\theta}$ is the orthogonal projection of $\mathbf{y}$ onto $V_\mathbf{X}$. Thus

$$\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right)=\mathbf{0}$$

or

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}}=\mathbf{X}^T\mathbf{Y}$$

is the equation satisfied by the least squares estimator $\widehat{\boldsymbol{\theta}}$.

5.4 Estimating the variance

We shall need to estimate the covariance matrix. Given an estimate S^2 of σ^2 this is

$$\widehat{\mathbf{\Sigma}}_\theta = S^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

Now, since

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=\sum_{i=1}^n\varepsilon_i^2$$

and $\sum_{i=1}^n\varepsilon_i^2$ has expected value $n\sigma^2$, we might expect to construct an estimator of σ^2 based on

$$Q^2=\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right)^T\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right)=\mathbf{Y}^T\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right).$$

It is time for a bit of jargon; $\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\theta}}$ is called the *vector of residuals* and Q^2 is called the *residual sum of squares* (or *RSS* for short).

Remember that the r-dimensional subspace of $\mathbb{R}^n$ spanned by columns of $\mathbf{X}$ was denoted by $V_\mathbf{X}$. Therefore construct an orthonormal basis $\{\mathbf{e}_1,\mathbf{e}_{2,}\ldots,\mathbf{e}_r\}$ for $V_\mathbf{X}$ and extend it to $\{\mathbf{e}_1,\ldots,\mathbf{e}_r,\mathbf{e}_{r+1},\ldots\mathbf{e}_n\}$ for $\mathbb{R}^n$. Since $\mathbf{y}\in\mathbb{R}^n$, it follows that there exist random variables $Z_1,Z_2,\ldots,Z_n$ such that

$$\mathbf{Y}=\sum_{i=1}^n Z_i\mathbf{e}_i.$$

Since $\mathbf{X}\widehat{\boldsymbol{\theta}}$ is the orthogonal projection of $\mathbf{Y}$ onto $V_{\mathbf{X}}$,

$$\mathbf{X}\widehat{\boldsymbol{\theta}}=\sum_{i=1}^{r} Z_i\mathbf{e}_i \quad \text{and} \quad \mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\theta}}=\sum_{i=r+1}^{n} Z_i\mathbf{e}_i.$$

Hence

$$Q^2=\left(\sum_{j=r+1}^{n} Z_j\mathbf{e}_j\right)^T\left(\sum_{i=r+1}^{n} Z_i\mathbf{e}_i\right)=\sum_{i=r+1}^{n} Z_i^2,$$

since $\mathbf{e}_j^T\mathbf{e}_i=\delta_{ij}$.

Now

$$\begin{aligned} Z_j=\mathbf{e}_j^T\mathbf{Y} \;\Rightarrow\; E(Z_j) &= \mathbf{e}_j^T E(\mathbf{Y}) \\ &= \mathbf{e}_j^T\mathbf{X}\boldsymbol{\theta}=0 \quad \text{for} \quad j=r+1,\ldots,n, \end{aligned}$$

so that

$$E(Z_j^2)=V(Z_j)=\sigma^2.$$

Z_j is a linear function of the normally distributed Y_i, $i=1,\ldots,n$, so

$$\frac{Z_j}{\sigma}\sim N\left(0,1\right), \quad \text{and therefore} \quad \left(\frac{Z_j}{\sigma}\right)^2\sim\chi^2\left(1\right).$$

Hence

$$E(Q^2)=\sum_{i=r+1}^{n} E(Z_i^2)=\sum_{i=r+1}^{n} V(Z_i)=(n-r)\sigma^2.$$

and

$$S^2=\frac{Q^2}{(n-r)}$$

is unbiased for σ^2. Furthermore, the $Z_j's$ are independent, so:

$$\frac{Q^2}{\sigma^2}\sim\chi^2(n-r) \quad\Rightarrow\quad \frac{(n-r)\,S^2}{\sigma^2}\sim\chi^2(n-r).$$

Finally, provided $\operatorname{rank}(\mathbf{X})=p$, then

$$\widehat{\boldsymbol{\theta}}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\widehat{\boldsymbol{\theta}})=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\sum_{i=1}^{r} Z_i\mathbf{e}_i\right)$$

is independent of S^2.

5.5 Testing the coefficients

You have already seen that

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

so that

$$\widehat{\theta}_i \sim N(\theta_i, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{ii})$$

and

$$\frac{\widehat{\theta}_i - \theta_i}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim N(0,1).$$

Using the result that

$$\frac{(n-r)S^2}{\sigma^2} \sim \chi^2(n-r)$$

and remembering the definition of a t-distribution we conclude that

$$\frac{\widehat{\theta}_i - \theta_i}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t(n-r).$$

This enables us to carry out hypothesis tests or calculate confidence intervals for coefficients.

Example 5.4 (Example 5.3 revisited) *Divorces*

Our estimate of the rate of increase of divorces was $\widehat{\beta} = 5.417$ and we would like to answer the question "Is the divorce rate changing?" In other words, we would like to test the null hypothesis $H_0 : \beta = 0$.

Using lower case letters for realisations, the RSS is $q^2 = 65.175$ and $n = 6$, $r = 2$ so that

$$s^2 = \frac{65.175}{4} = 16.294.$$

We have already obtained

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\frac{1}{n}\sum x_i \\ -\frac{1}{n}\sum x_i & 1 \end{pmatrix}}{\sum\left(x_i - \overline{x}\right)^2}$$

and

$$\sum\left(x_i - \overline{x}\right)^2 = 17.5.$$

Under H_0

$$\frac{\widehat{\beta}}{S(17.5)^{-1/2}} = T \sim t(4)$$

giving

$$t = \frac{5.417}{\sqrt{16.294\,/17.5}} = 5.614.$$

The p-value for the two-sided test is

$$2P(T \geq 5.614) = 0.005.$$

We therefore have strong evidence to reject the null hypothesis that the divorce rate is not changing; that is, there is strong evidence of an increasing divorce rate.
■

Example 5.5 *AIDS data for the USA*

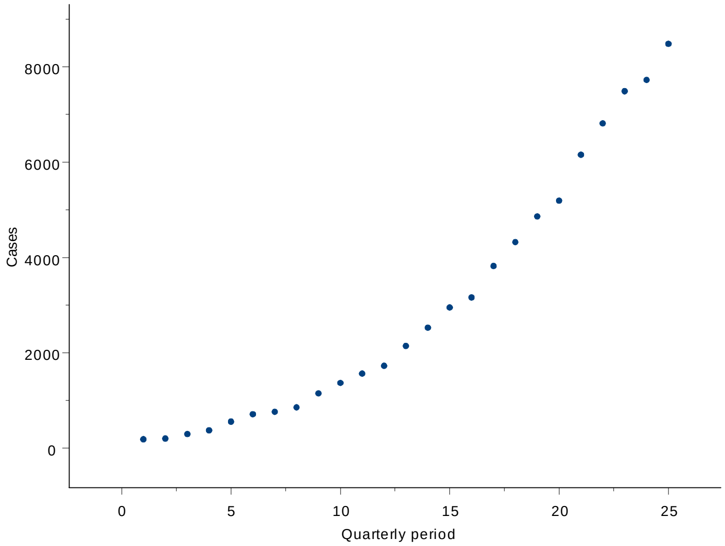
These data are for AIDS incidence in the USA, adjusted for reporting delays. The data are taken from Rosenberg, P.S. and Gail, M.H. (1991): Backcalculation of flexible linear models of the Human Immunodeficiency Virus infection curve. *Applied Statistics*, **40**, 269-282.

Newly reported cases are recorded quarterly and the variable *Time* therefore counts 3-monthly periods, staring with the first quarter in 1982.

Table 5.3 Reported AIDS cases in the USA

Quarter	*Time*	*Cases*	*Quarter*	*Time*	*Cases*	*Quarter*	*Time*	*Cases*
1982:1	1	185	1984:2	10	1369	1986:2	18	4321
1982:2	2	200	1984:3	11	1563	1986:3	19	4863
1982:3	3	293	1984:4	12	1726	1986:4	20	5192
1982:4	4	374	1985:1	13	2142	1987:1	21	6155
1983:1	5	554	1985:2	14	2525	1987:2	22	6816
1983:2	6	713	1985:3	15	2951	1987:3	23	7491
1983:3	7	763	1985:4	16	3160	1987:4	24	7726
1983:4	8	857	1986:1	17	3819	1988:1	25	8483
1984:1	9	1147						

The scatterplot shows that the trend is not linear.



Incidence of AIDS cases in the USA against time

The plot has all the appearance of showing a functional relationship between AIDS incidence and time. One might, for example, try to fit an exponential function or some kind of power law to model the growth curve.

The plot suggests that the incidence of AIDS against time in the USA is not linearly related to time. Try fitting a model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

and test the coefficients β_1, β_2.
□

This time we have 3 coefficients, and the design matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \text{with} \quad \boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

We can use a standard computer package to carry out the regression.

Linear Model
Response: Cases

Variable		Coefficient	s.e.	t-value	p-value
Intercept	α	343.5913	87.7446	3.9158	0.0007
Time	β_1	−60.1380	15.5514	−3.8671	0.0008
Time^2	β_2	15.6277	0.5806	26.9158	0.0000

$r^2 = 0.9976$ $\quad d.f. = 22$ $\quad s = 134.7$ $\quad RSS = 399155$

This is a typical computer package output, and we need to say a few words about what some of these figures mean.

The values for the coefficients α, β_1 and β_2 are simple enough to interpret - the fitted model is

$$Cases = 343.5913 - 60.1380 \times Time + 15.6277 \times Time^2.$$

The column labelled *s.e.* gives the standard errors of the coefficient estimates *i.e.* values calculated from the expression you saw earlier, namely $s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}$.

The next column gives the calculated t-statistic and the last column gives p-values for tests of the coefficients being zero. There is extremely strong evidence for rejecting the null hypothesis that either β_1 or β_2 is zero. We can conclude that $Time$ and $Time^2$ are highly significant in modelling the number of cases.
■

Some of the numbers given in the last example have not, as yet, been explained. *RSS* is the residual sum of squares (as you might expect), but what is r^2?

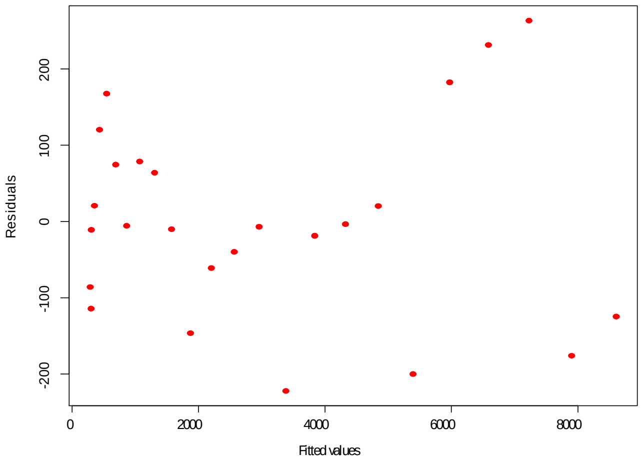
r^2 is a measure of the squared correlation. It can be interpreted as a measure of the strength of the linear relationship. You should think of r^2 as a way of quantifying the reduction in variability of the response variable brought about by taking account of the linear relationship. The value of 0.9976 above means that 99.76% of the variability is explained by the model, whilst the remaining 0.24% is the unexplained or *residual* variability. Clearly, in Example 5.5 the fit looks extremely good, but can this model be relied upon? The crucial question now is “How can we assess the quality of our model?”

5.6 Assessing the model

The first step in looking at the adequacy of a model is to check the assumptions on which it is based, so let us recall what they were.

- the variance of the response is constant (i.e. the variance does not depend upon x);
- the trend is described by the equation;
- the distribution of the response is normal.

Well, we are fairly happy about the first of these because the scatterplot we looked at first shows comparatively little scatter anyway. But the proper way to perform a preliminary check is to examine the *residuals*, and the easiest way to check the variance assumption is to look at a plot of residuals versus fitted values and see whether their spread has any pattern to it.



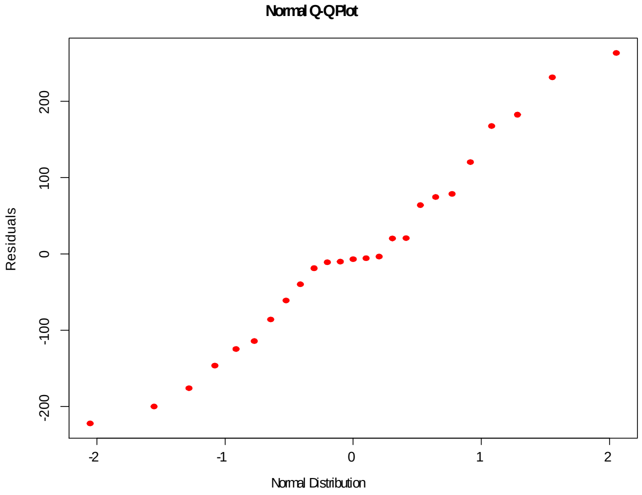
Plot of residuals against fitted values for AIDS data

- The points to the right seem a little more spread out but there is no indication that spread is a function of fitted value.

- There does seem to be some curvature – re-think the assumption about the model?

- We also need to check up on the third assumption. If the residuals show no dependence upon the explanatory variables and are plausibly normal, we shall be happy.

We have checked the first two assumptions by examining a plot of residuals against fitted values – we check the third with a normal probability plot.



Normal probability plot of AIDS residuals

This doesn't look too bad. We can therefore conclude that we have a reasonable model which could perhaps be improved, but which fits the data pretty well for the most part.

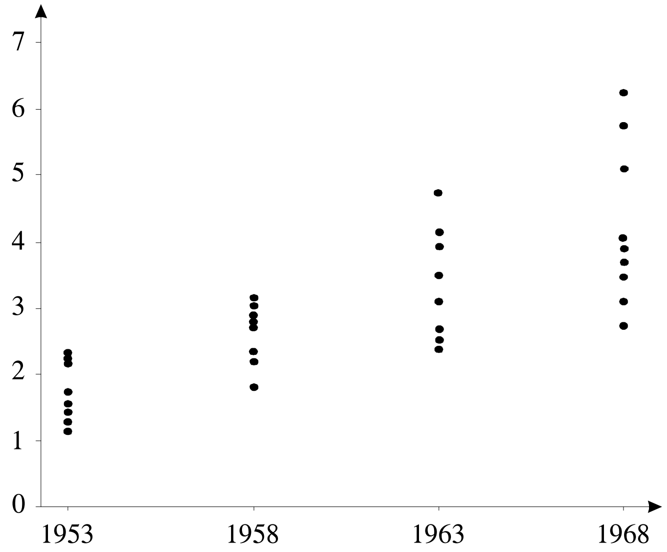
Example 5.6 *Testicular cancer*

The table below comprises data from Lee, Hitosugi and Peterson (1973): Rise in mortality from tumors of the testis in Japan, 1947-70. *J. Nat. Cancer Inst.*, **51**, 1485-90. It gives the populations and numbers of deaths from testicular cancer in 5-year age groups and 5-year periods in Japan. The ages refer to the lowest age in each group and the populations are expressed in millions of persons.

Table 5.4 Deaths in Japan from testicular cancer

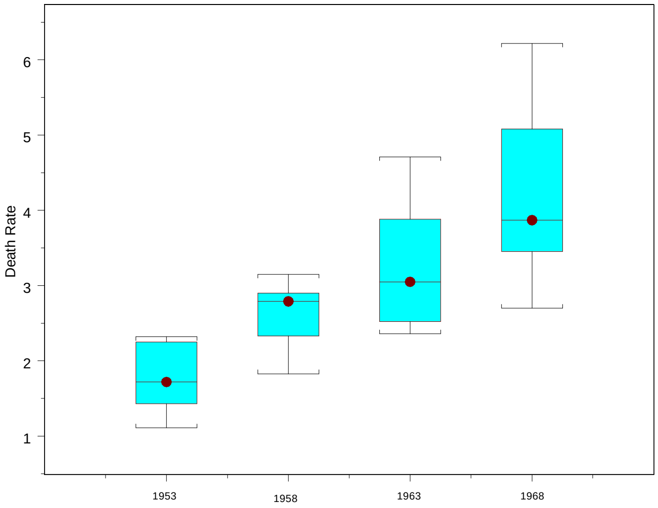
	1951-55		1956-60		1961-65		1966-70	
Age	Popn.	Dths	Popn.	Dths	Popn.	Dths	Popn.	Dths
20	20.4	27	21.3	39	22.2	56	24.0	83
25	17.2	40	20.0	58	20.6	97	21.8	125
30	12.6	18	17.1	54	19.9	77	20.8	129
35	11.7	13	12.5	36	17.0	70	19.9	101
40	11.5	26	11.5	32	12.2	29	16.8	67
45	10.3	16	11.2	26	11.1	34	12.0	37
50	9.3	16	9.8	27	10.7	27	10.7	29
55	7.6	17	8.7	19	9.2	32	10.1	39
60	5.9	13	6.8	21	7.9	21	8.4	31

The scatterplot shows that the mean death rate from cancer of the testis in Japan has been rising steadily since 1951.



Death rate against year for cancer of the testis in Japan

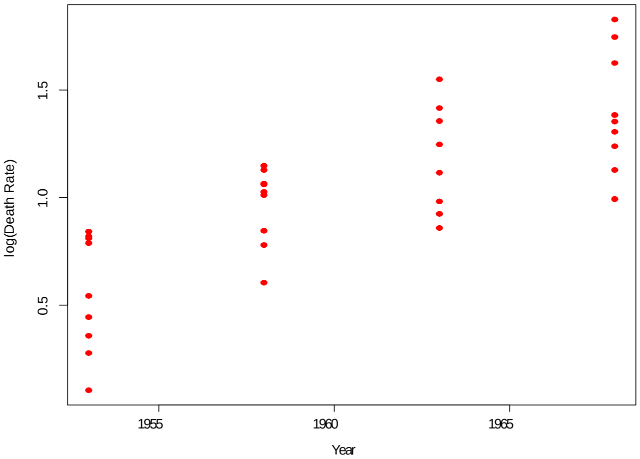
Note that there is no variability in the explanatory variable, but there is marked variability in the response. This is clearly shown in the boxplots.



Boxplots of death rates

It is clear that the variability in the data increases as the year variable increases. What should one do about this?

The answer is to look for a transformation which will stabilise the variance. Here we need a transformation which compresses large values of the response more than it compresses smaller values; something like a square root or a cube root or possibly even a log transformation. Taking the log of the death rate results in the next scatterplot.



Testicular cancer: plot of log(Death rate) against year

This looks reasonable and we could now go ahead and fit a model of the form

$$\log(Y_i) = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

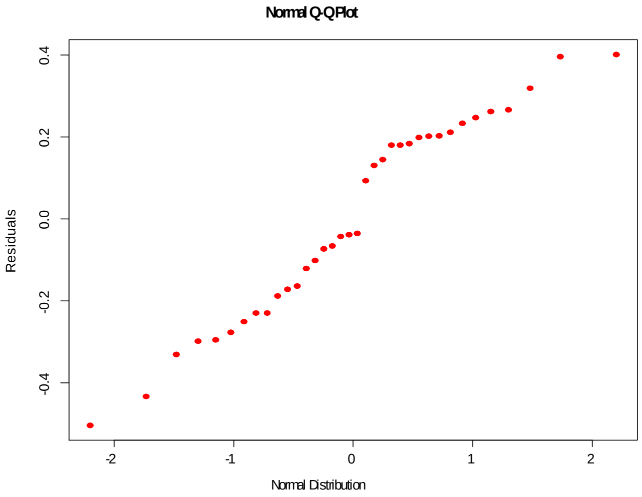
With the log transformed data, the fitted model turns out to be as given below.

Linear Model
Response: log(Death rate)

Variable		Coefficient	s.e.	t-value	p-value
Intercept	α	-105.9198	14.4887	-7.3105	0.0000
Year	β	0.0545	0.0074	7.3808	0.0000

$r^2 = 0.6157$ $\quad d.f. = 34$ $\quad s = 0.2479$ $\quad RSS = 2.0891$

Of course stable variance alone is not enough because the residuals also need to be normally distributed. This can be checked with a normal q-q plot of the residuals.



Normal probability plot of testicular cancer residuals

The plot seems to show a very rough straight line, but it is not entirely convincing.
□

It is worth noting that, in doing Example 5.6, we fitted a model with more than two coefficients. How is this different from fitting more than one explanatory variable? How can the technique of fitting the model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

be any different from fitting, say

$$Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where z_i is some other variable which has been recorded (*e.g.* annual expenditure on AIDS clinics)? The answer is, of course, that mathematically they are exactly the same.

5.7 Multiple regression

With more than one explanatory variable, the model takes the form

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

or, as we have seen,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\theta} = (\alpha, \beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $\mathbf{X}$ is the $n \times (1+p)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Let us see not only how to apply the general theory but also how to set about the whole modelling process by looking at a data set.

Example 5.7 *Olympic Sprint Times*

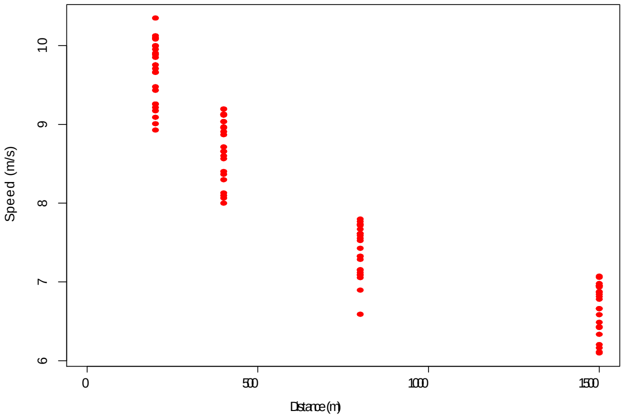
Table 5.2 gives the times in seconds recorded by the winners in the finals of the men's sprint events (100, 200, 400, 800 and 1500 metres) at each of the 21 Olympic Games from 1900 to 2004 along with the heights above sea level of the different venues. Obviously the years 1916, 1940, 1944 are missing since the Olympic games were not held during the World Wars.

Table 5.2 Winning times for Olympic running events

Year	*Venue*	*100m*	*200m*	*400m*	*800m*	*1500m*	*Altitude*
1900	Paris	10.80	22.20	49.40	121.40	246.00	25
1904	St Louis	11.00	21.60	49.20	116.00	245.40	455
1908	London	10.80	22.40	50.00	112.80	243.40	8
1912	Stockholm	10.80	21.70	48.20	111.90	236.80	46
1920	Antwerp	10.80	22.00	49.60	113.40	241.80	3
1924	Paris	10.60	21.60	47.60	112.40	233.60	25
1928	Amsterdam	10.80	21.80	47.80	111.80	233.20	8
1932	Los Angeles	10.30	21.20	46.20	109.80	231.20	340
1936	Berlin	10.30	20.70	46.50	112.90	227.80	115
1948	London	10.30	21.10	46.20	109.20	225.20	8
1952	Helsinki	10.40	20.70	45.90	109.20	225.20	25
1956	Melbourne	10.50	20.60	46.70	107.70	221.20	3
1960	Rome	10.20	20.50	44.90	106.30	215.60	66
1964	Tokyo	10.00	20.30	45.10	105.10	218.10	45
1968	Mexico City	9.95	19.83	43.80	104.30	214.90	7349
1972	Munich	10.14	20.00	44.66	105.90	216.30	1699
1976	Montreal	10.06	20.23	44.26	103.50	219.20	104
1980	Moscow	10.25	20.19	44.60	105.40	218.40	497
1984	Los Angeles	9.99	19.80	44.27	103.00	212.50	340
1988	Seoul	9.92	19.75	43.87	103.45	215.96	111
1992	Barcelona	9.96	20.01	43.50	103.66	220.12	3
1996	Atlanta	9.84	19.32	43.49	102.58	215.78	1026
2000	Sydney	9.87	20.09	43.84	105.08	212.07	3
2004	Athens		19.79	44.00	104.45	214.20	505

We would like to construct a predictive model for the winning times at the 2008 games in Beijing (and the 2012 games in London). Clearly there is now a possibility of more than one explanatory variable.

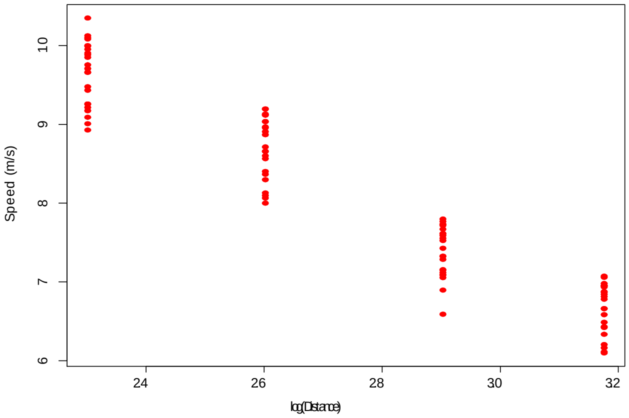
Since the 100m event is largely about getting a fast start rather than keeping up a sustained speed, we shall leave that event out of the analysis. Let us start by looking at a plot of average speed in metres per second against distance of event.



Plot of *Speed* against *Distance*

Clearly the relationship is not linear and we need a transformation.

You can see that the standard deviations of the speeds do not differ much between events. Therefore the response should *not* be transformed. Try taking the log of the explanatory variable.



Plot of *Speed* against log(*Distance*)

That's better, but it still may not be all that good. However, we can try a preliminary fit and see where that gets us.

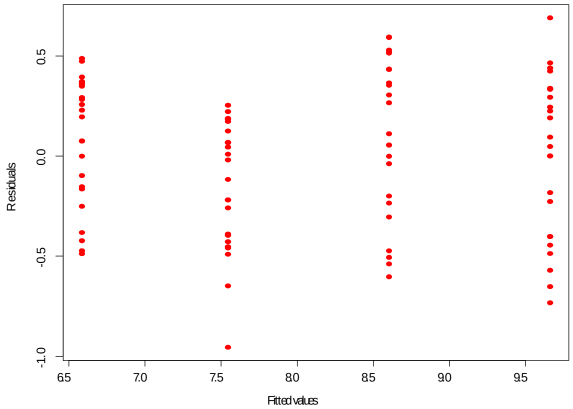
The computer produces the following

Linear regression analysis
Response: Speed

Variable		Coefficient	s.e.	t-value	p-value
Constant	α	17.7837	0.3233	55.01	$< 2 \times 10^{-16}$
Log(distance)	β	-1.5296	0.0508	-30.12	$< 2 \times 10^{-16}$

$r^2 = 0.9061$ $d.f. = 94$ $s = 0.3749$ $RSS = 12.6108$

On the face of it this does not look too bad with highly significant coefficients and $r^2 = 0.9061$. However, we need to check the residuals.



Plot of residuals against fitted values for log(distance) model

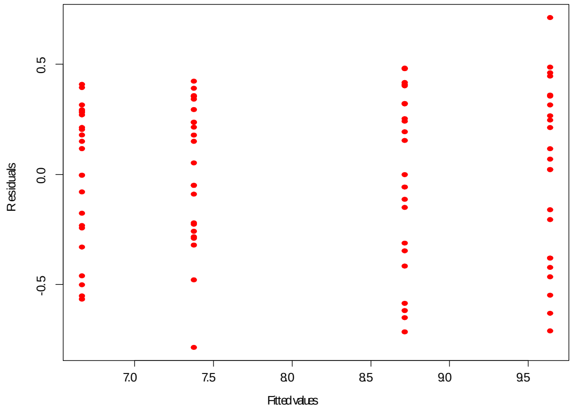
It is clear that our model will not do. A clear pattern is present and the log transformation has not worked.

The fit can be substantially improved by, instead of taking the log of distance, adding a quadratic term to it. In order to keep the numbers sensible, divide *distance* by 100 and re-fit as a quadratic.

Linear regression analysis
Response: Speed

Variable	Coefficient		s.e.	t-value	p-value
Constant	α	10.7571	0.1229	87.53	$< 2 \times 10^{-16}$
Distance	β_1	-0.5918	0.0378	-15.68	$< 2 \times 10^{-16}$
Distance^2	β_2	0.0213	0.0021	10.03	$< 2 \times 10^{-16}$
$r^2 = 0.9134$	$d.f. = 93$		$s = 0.375$	$RSS = 11.6601$	

The value of r^2 has improved to 0.9134, all of the coefficients have values significantly different from zero and a look at a plot of residuals against fitted values shows that the pattern in the residuals has been largely removed.



Plot of residuals against fitted values with a quadratic term included

We should now go on and construct a normal probability plot of the residuals, but there is little point because it is hard to believe that this is going to be our final model. There are obvious things we haven't taken into account.

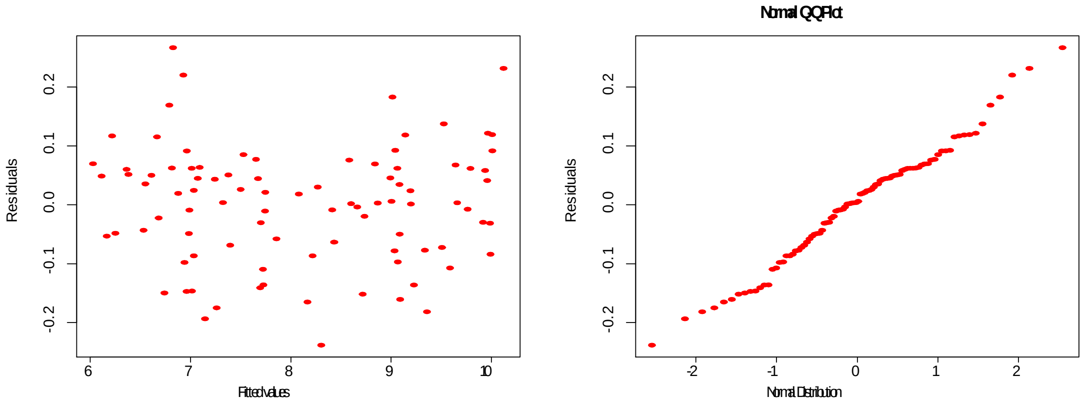
- Every Olympics, runners go faster, though the improvement has started to slow down recently. Obviously any proper treatment must include *year* as a variable (or rather some appropriate functions of *year*).
- We also need to include height above sea-level of each Olympic venue. This made a big difference at the Mexico Olympics.

It turns out that we need to fit *speed* against *distance*, *distance*^{2}, *year*, *year*^{2}, and $\log(altitude)$.

Linear regression analysis
Response: Speed

Variable		Coefficient	s.e.	t-value	p-value
Constant	α	10.001	0.0471	212.394	$< 2 \times 10^{-16}$
Distance	β_1	-0.5918	0.0109	-54.570	$< 2 \times 10^{-16}$
Distance^2	β_2	0.0213	0.0006	34.912	$< 2 \times 10^{-16}$
Year	β_3	0.0161	0.0013	12.375	$< 2 \times 10^{-16}$
Year^2	β_4	-0.000059	0.000012	-4.869	4.8×10^{-6}
log(Altitude)	β_5	0.0254	0.0050	5.010	2.7×10^{-6}

$r^2 = 0.9931 \quad d.f. = 90 \quad s = 0.1040 \quad RSS = 0.9736$



Residuals against fitted values and normal probability plot for the Olympic sprints model

The plots are satisfactory and the model looks pretty good. We now have something we can use to make predictions about the times in the 2008 Olympic Games in Beijing. However, there is more to using such models for prediction than at first meets the eye, as you will see next.
□

5.8 Making predictions

Once the coefficients have been estimated, the fitted equation can be used to obtain values of y for any given values $\mathbf{a} = (1, a_1, \ldots, a_p)$ of the explanatory variables $x = (1, x_1, \ldots, x_p)$. There is a population of such y-values and, for a random observation from this population, $Y \sim N(\mathbf{a}\boldsymbol{\theta}, \sigma^2)$.

First consider estimating the mean of the population, $E(Y \mid \mathbf{a})$. Clearly

$$\widetilde{Y} = \mathbf{a}\widehat{\boldsymbol{\theta}},$$

where $\widehat{\boldsymbol{\theta}}$ is a random vector of estimators, is an unbiased estimator of $\mathbf{a}\boldsymbol{\theta}$.

Furthermore, $\widetilde{Y}$ must be normally distributed and

$$E(\widetilde{Y}) = \mathbf{a}\boldsymbol{\theta}, \quad V(\widetilde{Y}) = \mathbf{a}V\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{a}^T = \sigma^2\mathbf{a}\left(\mathbf{X^TX}\right)^{-\mathbf{1}}\mathbf{a}^T.$$

Example 5.8 *Simple linear regression*

Suppose we wish to predict $E\left(Y\right)$ at the point x_0. Then $\mathbf{a} = (1, x_0)$ so

$$\begin{aligned} E(\widetilde{Y}) &= \alpha + \beta x_0 \\ V(\widetilde{Y}) &= V(\widehat{\alpha}) + x_0^2 V(\widehat{\beta}) + 2x_0 C(\widehat{\alpha}, \widehat{\beta}) \end{aligned}$$

As a point of technique, with the simple linear model, it is usually better to fit

$$Y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i$$

Then

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \begin{pmatrix} n & 0 \\ 0 & \sum(x_i-\overline{x})^2 \end{pmatrix}, \\ \left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-\mathbf{1}} &= \begin{pmatrix} 1/n & 0 \\ 0 & 1/\sum(x_i-\overline{x})^2 \end{pmatrix}\end{aligned}$$

so that $C(\widehat{\alpha},\widehat{\beta})=0$ and $V(\widehat{\alpha})$, $V(\widehat{\beta})$ are straightforward.
□

Fact 5.9
Suppose the two random variables W and Z are jointly normally distributed. (In other words, W and Z have a bivariate normal distribution.)

Then if W and Z are uncorrelated, i.e. they satisfy $C(W,V)=0$, they are independent.

Applying Fact 5.9 to the previous example, $\widehat{\alpha}$ and $\widehat{\beta}$ are both linear combinations of the normal random variables Y_i, so have a joint normal distribution. Further, the new parameterization gives $C(\widehat{\alpha},\widehat{\beta})=0$, so the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are *independent.*

In general it is easy to construct a confidence interval for $E(Y \mid \mathbf{a})=\mathbf{a}\boldsymbol{\theta}$. Since $\text{rank}(\mathbf{X})=p+1$,

$$\frac{\mathbf{a}\widehat{\boldsymbol{\theta}}-\mathbf{a}\boldsymbol{\theta}}{\widehat{\sigma}\sqrt{\mathbf{a}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{a}^T}} \sim t(n-p-1)$$

and an inequality can be constructed and re-arranged for $\mathbf{a}\boldsymbol{\theta}$ in the usual way.

Example 5.10 *Simple linear regression*

For a given value of x_0,

$$E(Y \mid x_0)=\alpha+\beta(x_0-\overline{x}),$$

so

$$\frac{\widehat{\alpha}+\widehat{\beta}(x_0-\overline{x})-(\alpha+\beta(x_0-\overline{x}))}{\widehat{\sigma}\sqrt{(1/n+(x_0-\overline{x})^2/\sum(x_i-\overline{x})^2}} \sim t(n-2).$$

If m is such that $P\left(|T_{n-2}|\geq m\right)=\gamma$,

$$\widehat{\alpha}+\widehat{\beta}(x_0-\overline{x})\pm m\widehat{\sigma}\sqrt{(1/n+(x_0-\overline{x})^2\Big/\sum(x_i-\overline{x})^2}$$

is a $100(1-\gamma)\%$ confidence interval for $E(Y \mid x_0)$.
□

But suppose we wanted to make a prediction of an as-yet unobserved random variable Y (*i.e.* not the expectation, but the variable itself). We are after a *predictive interval* on a random variable Y yet to be observed.

Now $Y \sim N(\mathbf{a}\boldsymbol{\theta}, \sigma^2)$ for a vector $\mathbf{a}$ of given values, and $Y - \mathbf{a}\widehat{\boldsymbol{\theta}}$ has a normal distribution with mean

$$E\left(Y - \mathbf{a}\widehat{\boldsymbol{\theta}}\right) = \mathbf{a}\boldsymbol{\theta} - \mathbf{a}\boldsymbol{\theta} = 0$$

and, since Y and $\widehat{\boldsymbol{\theta}}$ are independent, with variance

$$\begin{aligned} V\left(Y - \mathbf{a}\widehat{\boldsymbol{\theta}}\right) &= V(Y) + V(\mathbf{a}\widehat{\boldsymbol{\theta}}) \\ &= \sigma^2 + \sigma^2 \mathbf{a}\left(\mathbf{X^T X}\right)^{\mathbf{-1}} \mathbf{a}^T. \end{aligned}$$

Note the extra term added to the variance.

Using the independence of $Y, \widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$,

$$\frac{Y - \mathbf{a}\widehat{\boldsymbol{\theta}}}{\widehat{\sigma}\sqrt{1 + \mathbf{a}\left(\mathbf{X^T X}\right)^{\mathbf{-1}} \mathbf{a}^T}} \sim t(n - p - 1). \tag{4}$$

Example 5.11 *Simple linear regression*

For a given value of x_0,

$$\frac{Y - (\widehat{\alpha} + \widehat{\beta}(x_0 - \overline{x}))}{\widehat{\sigma}\sqrt{(1 + 1/n + (x_0 - \overline{x})^2 / \sum (x_i - \overline{x})^2}} \sim t(n - 2).$$

If m is such that $P\left(|T_{n-2}| \geq m\right) = \gamma$,

$$\widehat{\alpha} + \widehat{\beta}(x_0 - \overline{x}) \pm m\widehat{\sigma}\sqrt{(1 + 1/n + (x_0 - \overline{x})^2 \Big/ \sum (x_i - \overline{x})^2}$$

is a $100(1 - \gamma)\%$ predictive interval for Y.
□

Example 5.12 *Predictions for the 2008 Olympic Games.*

The model we have arrived at makes prediction of the times for the Olympic sprint events in 2008 comparatively straightforward. Beijing is 120 feet above sea level so the predicted winning times for the events are as given below.

	200m	*400m*	*800m*	*1500m*
Speed	10.09	9.16	7.81	7.11
Time	19.83	43.68	102.38	211.11

Confidence intervals for the predicted average speeds can be obtained by using Equation 4; these can then be converted back to times for the events. In practice a computer can be used to do the matrix multiplications, although most statistical packages will do the calculations directly as one of the options provided with their regression software.

	200m		*400m*		*800m*		*1500m*	
Speed	10.09		9.16		7.81		7.11	
95% c.i.	9.87	10.30	8.94	9.38	7.59	8.03	6.89	7.32
Time	19.83		43.68		102.38		211.11	
95% c.i.	19.41	20.27	42.67	44.74	99.58	105.33	204.78	217.86

□

Results from Beijing 2008 and predictions for London 2012

The actual times were all inside the appropriate confidence intervals, apart from the 200m:

	200m	*400m*	*800m*	*1500m*
Time	19.30	43.75	104.65	212.94

You might like to try modifying the model by including the Beijing results and trying to make predictions for London, which is 50 feet above sea level.You should get the following hostages to fortune:

	200m		*400m*		*800m*		*1500m*	
Time	19.83		43.71		102.58		211.40	
95% c.i.	19.39	20.28	42.66	44.80	99.70	105.63	204.88	218.35

Out of interest, the world record times in these events are 19.30, 43.18, 101.11 and 206.00 seconds respectively.