Chapter 11. Linear Regression Analysis

11.1 Introduction and models

Regression analysis: one of the most frequently used statistical techniques. It aims to build up an explicit relationship between one dependent variable, often denoted as y, and one or several regressors (also called independent variables or predictors), often denoted as $x_1, \cdots, x_p$.

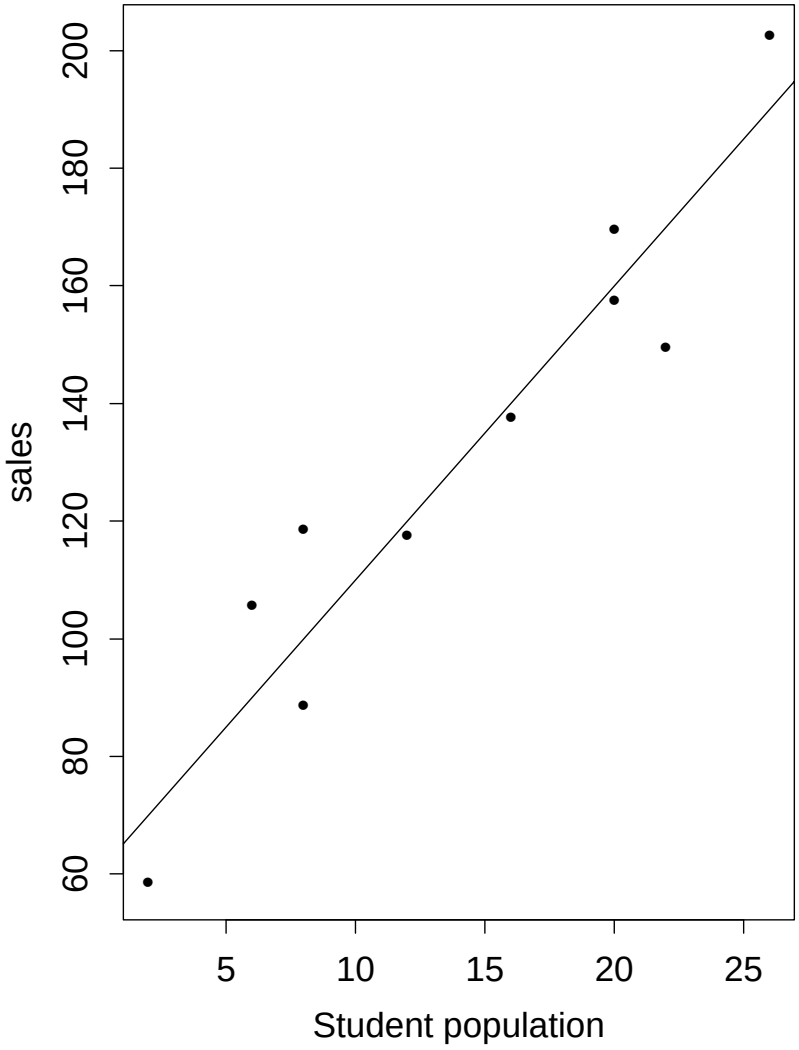
Goal of Regression analysis: to understand how y depends on $x_1, \cdots, x_p$, to predict or control unobserved y based on observed $x_1, \cdots, x_p$.

Statistical Model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$, where ε denotes the random noise satisfying the conditions: $E\varepsilon = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 > 0$.

In a university town, the sales (y) of 10 Armand's Pizza Parlour restaurants are closely related to the student population (x) in their neighbours. The data file 'Armand.txt' contains the sales (in thousand euros) in a period of three months together with the numbers of students (in thousand) in their neighbours.
We plot y against x, and **draw** a straight line through the middle of data points:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε stands for a random error, β_0 is the intercept and β_1 is the slope of the straight line.



For a given population x, the predict sales is $\widehat{y} = \beta_0 + \beta_1 x$.

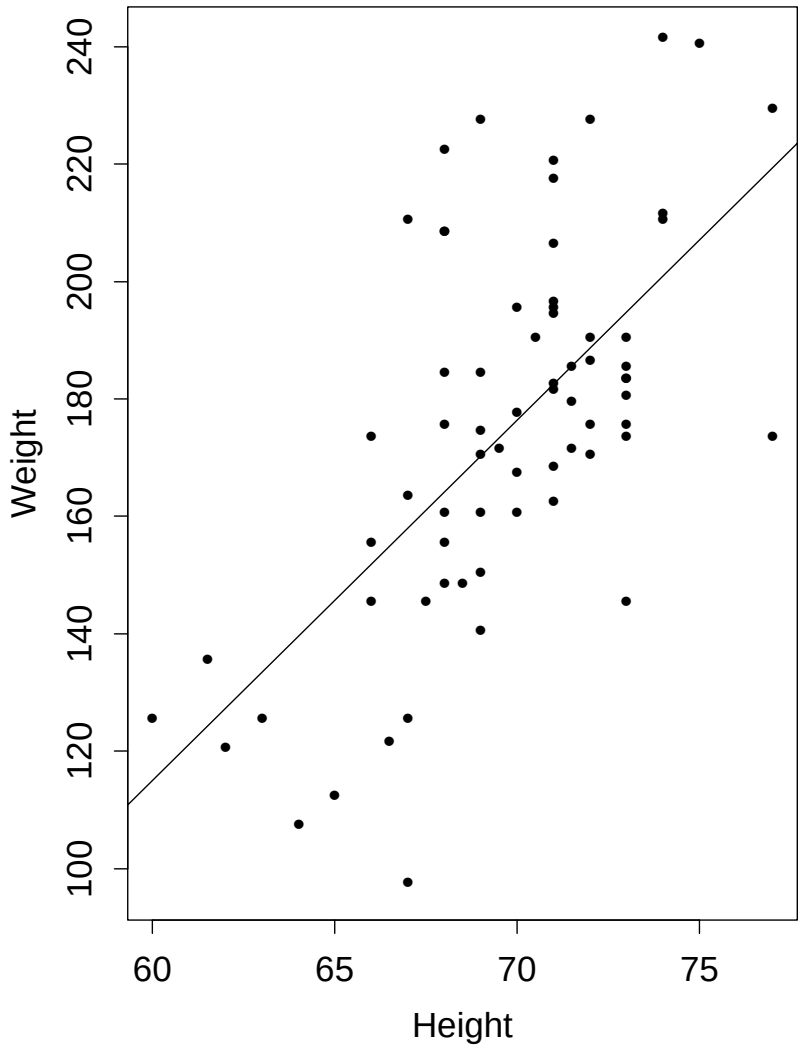
The data file 'weightHeight.txt' contains the heights (x) and the weights (y) from the 69 students in an EMBA class.

We plot y against x, and **draw** a straight line through the middle of the data crowds:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε stands for a random error, β_0 is the intercept and β_1 is the slope of the straight line.

For a given height x, the predict value $\widehat{y} = \beta_0 + \beta_1 x$ may be viewed as a kind of 'standard weight'.



Other possible examples:

y	x
sales	price
supply	demand
weight gain	protein in diet
present FTSE index	past FTSE index
spending	income
salary	service time
son's height	father's height
. . . . .	. . . . .

In most cases, there are more than one x involved.

Questions:

- How to draw a line through data crowds (i.e. to estimate β_0 and β_1)?
- How accurate is the fitted line?
- What is the error in predicting a future y?

In linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

the regressors $x_1, \cdots, x_p$ may be

- quantitative inputs
- transformations of quantitative inputs, such as log, square-root etc
- interactions between variables, e.g. $x_3 = x_1 x_2$
- basis expansions, such as $x_2 = x_1^2, x_3 = x_1^3, \cdots$
- numeric or "dummy" coding of the levels if qualitative variables.

Therefore

- the capacity of the model is large, and
- the model is linear wrt unknown coefficients $\beta_0, \beta_1, \cdots, \beta_p$.

In practice, we typically have n observations $\{(y_i, x_{i1}, \cdots, x_{ip}),\ i = 1, \cdots, p\}$. Using matrix multiplication, we may write

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ & \cdots & \cdots & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

or simple

$$\mathbf{y} \ = \ \mathbf{X}\,\boldsymbol{\beta} \ + \ \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is a $n \times (p+1)$ matrix, $\mathbf{y}$, $\boldsymbol{\varepsilon}$ are $n \times 1$-vectors, and $\boldsymbol{\beta}$ is $(p+1) \times 1$ vector.

Assumptions:

$$E\mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters.

Remark. (i) $E(\boldsymbol{\varepsilon}) = 0$. Note that ε_i represents the (random) noise in the i-th observation. We assume that those noises in different observations are uncorrelated (or even independent). Sometimes we assume ε_i's are independent $N(0, \sigma^2)$ variables, i.e. $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$.

(ii) In regression analysis, we treat x_{ij} as deterministic variables for technical convenience. Note that $E\, y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.

The observations $y_1, \cdots, y_n$ are assumed to be uncorrelated (or even independent) but not identically distributed.

Goals:

- Statistical inference for $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ and σ^2, i.e. point estimation, testing hypotheses, interval estimation.
- Prediction for future values of y

11.2 Estimation

We introduce the LSE for $\boldsymbol{\beta}$, which is also the MLE when ε_i's are normal.

Example 1. First consider a simple regression through the origin:

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1, \cdots, n.$$

The LSE $\widehat{\beta}$ for β minimizes

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 = ||\mathbf{y} - \mathbf{x}\beta||^2 = (\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta).$$

Note that

$$\begin{aligned} & ||\mathbf{y} - \mathbf{x}\beta||^2 = ||\mathbf{y} - \mathbf{x}\widehat{\beta} + \mathbf{x}(\widehat{\beta} - \beta)||^2 \\ = \quad & ||\mathbf{y} - \mathbf{x}\widehat{\beta}||^2 + ||\mathbf{x}(\widehat{\beta} - \beta)||^2 + 2(\widehat{\beta} - \beta)\mathbf{x}'(\mathbf{y} - \mathbf{x}\widehat{\beta}). \end{aligned}$$

If we choose $\widehat{\beta}$ such that $\mathbf{x}'(\mathbf{y} - \mathbf{x}\widehat{\beta}) = 0$,

$$||\mathbf{y} - \mathbf{x}\beta||^2 = ||\mathbf{y} - \mathbf{x}\widehat{\beta}||^2 + ||\mathbf{x}(\widehat{\beta} - \beta)||^2 \geq ||\mathbf{y} - \mathbf{x}\widehat{\beta}||^2$$

for all β. Hence LSE $\widehat{\beta}$ is the solution of equation

$$\mathbf{x}'(\mathbf{y} - \mathbf{x}\widehat{\beta}) = 0, \quad i.e. \quad \widehat{\beta} = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x} = \sum_i x_i y_i / \sum_i x_i^2.$$

Example 2. For simple linear regression

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i, \quad i = 1, \cdots, n.$$

The LSEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - x_i\beta_1)^2.$$

$$\begin{aligned}
&\sum_i (y_i - \beta_0 - x_i\beta_1)^2 = \sum_i \{y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1 + (\widehat{\beta}_0 - \beta_0) + x_i(\widehat{\beta}_1 - \beta_1)\}^2 \\
&= \sum_i (y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1)^2 + \sum_i \{(\widehat{\beta}_0 - \beta_0) + x_i(\widehat{\beta}_1 - \beta_1)\}^2 \\
&+2(\widehat{\beta}_0 - \beta_0)\sum_i (y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1) + 2(\widehat{\beta}_1 - \beta_1)\sum_i x_i(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1).
\end{aligned}$$

If we choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that

$$\begin{cases} \sum_i (y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1) = 0 \\ \sum_i x_i(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1) = 0 \end{cases} \tag{1}$$

we have

$$\sum_i (y_i - \beta_0 - x_i\beta_1)^2 \geq \sum_i (y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1)^2$$

for any β_0 and β_1. Hence the solution of (1) is the LSE.

The first equation can be written as

$$\widehat{\beta}_0 = \bar{y} - \bar{x}\widehat{\beta}_1.$$

Substituting this into the second equation, we have

$$\sum_i x_i\{(y_i - \bar{y}) - (x_i - \bar{x})\widehat{\beta}_1\} = 0.$$

Hence

$$\widehat{\beta}_1 = \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i(x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

Now we back to the general model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix. The LSE $\widehat{\boldsymbol{\beta}}$ is defined to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2.$$

$$\begin{aligned} & ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 \\ = \quad & ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2 + ||\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 + 2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \end{aligned}$$

Now choosing $\widehat{\boldsymbol{\beta}}$ such that

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0, \tag{2}$$

we have

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2 + ||\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 \geq ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2$$

for any $\boldsymbol{\beta}$. Thus the solution of (2) is the LSE.

Note that (2) can be written as $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$. Hence the LSE for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Write $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)'$, the estimated model is

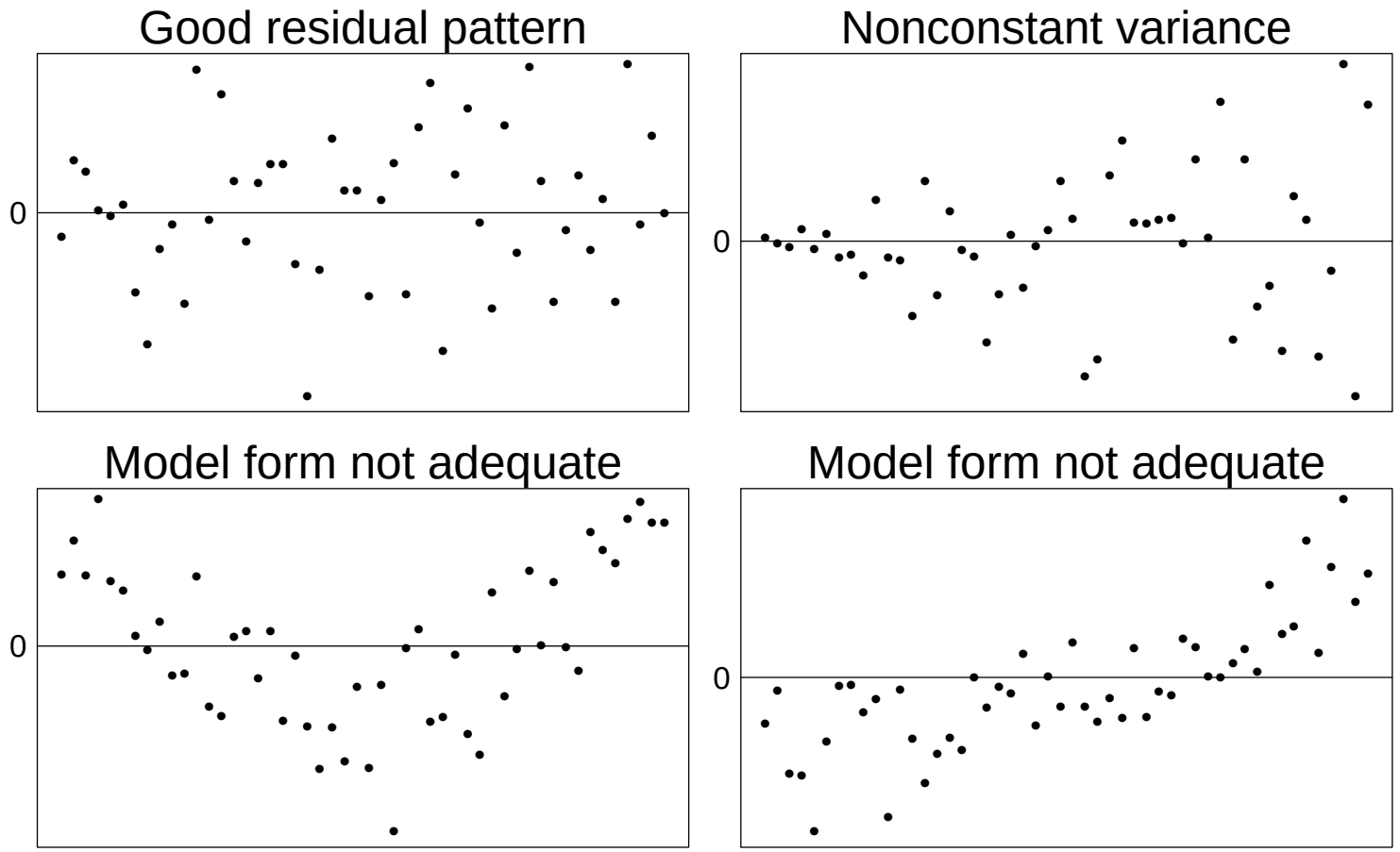
$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta} x_p.$$

For any given $(x_1, \cdots, x_p)$, we may predict the corresponding y using this model.

Residuals: $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$, $i = 1, \cdots, n$. Note $\widehat{y}_i = \widehat{\beta}_0 + \sum_{1 \le j \le p} \widehat{\beta}_j x_{ij}$.

Remark. If the model is correct, the residuals should behave like random noise!

An effective way for checking: plot $\widehat{\varepsilon}_i$ against y_i, or against each of x_{ij} for $j = 1, \cdots, p$.

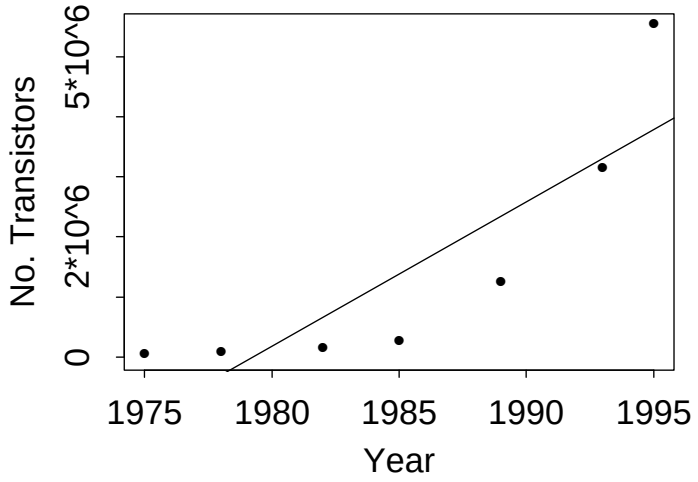


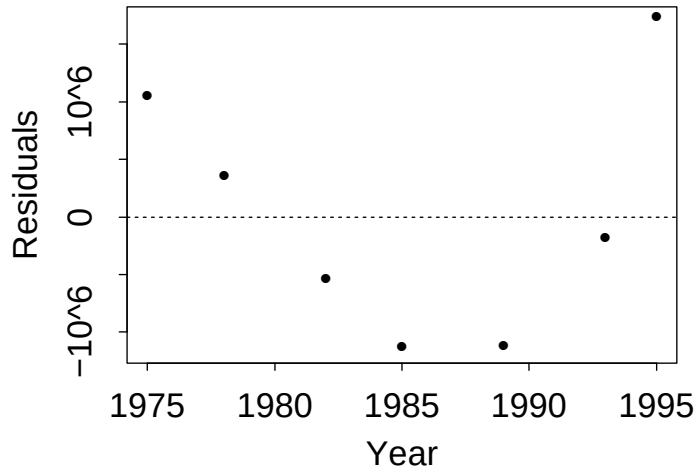
Example 3. *Computers steadily decrease in size as they grown in power*

Year (x)	1975	1978	1982	1985	1989	1993	1995
No. transistors per chip (y)	4.5K	29K	90K	229K	1200K	3100K	5500K

Fitting a linear regression of No. transistors (y) on Year (x):

$$\widehat{y} = -476185683 + 240588.1x, \quad \widehat{\sigma}^2 = 1.421 \times 10^{12}.$$

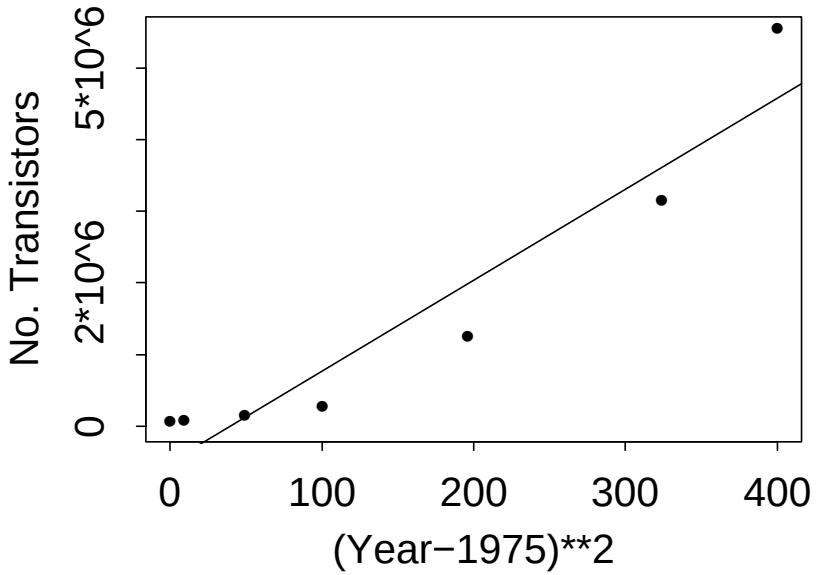


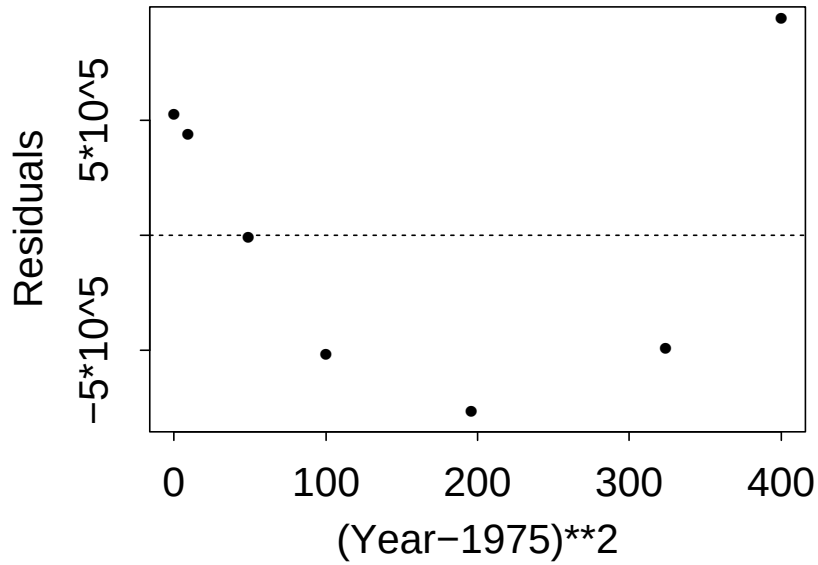


Residuals do not look like random noise!

Fitting a linear regression of No. transistors (y) on $u = (x - 1975)^2$:

$$\widehat{y} = -505893.6 + 12702u, \quad \widehat{\sigma}^2 = 4.909 \times 10^{11}.$$





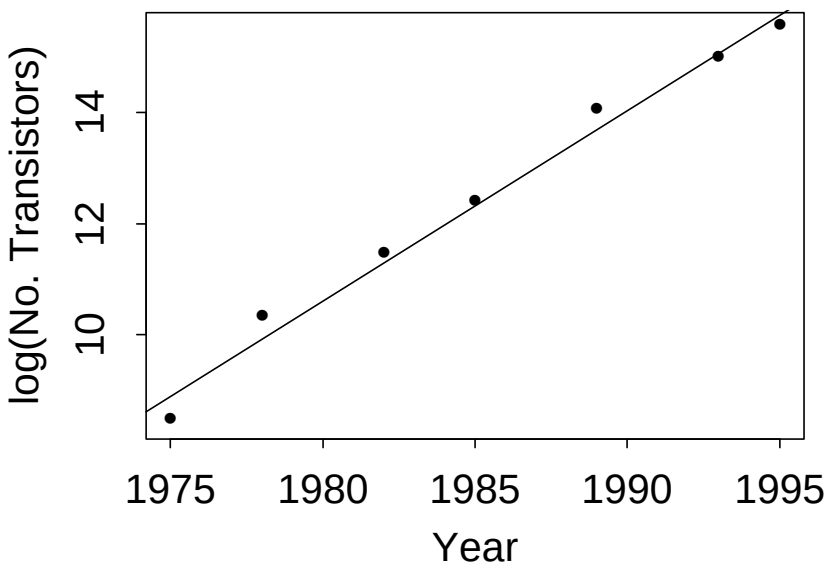
Residuals still do not look like random noise!

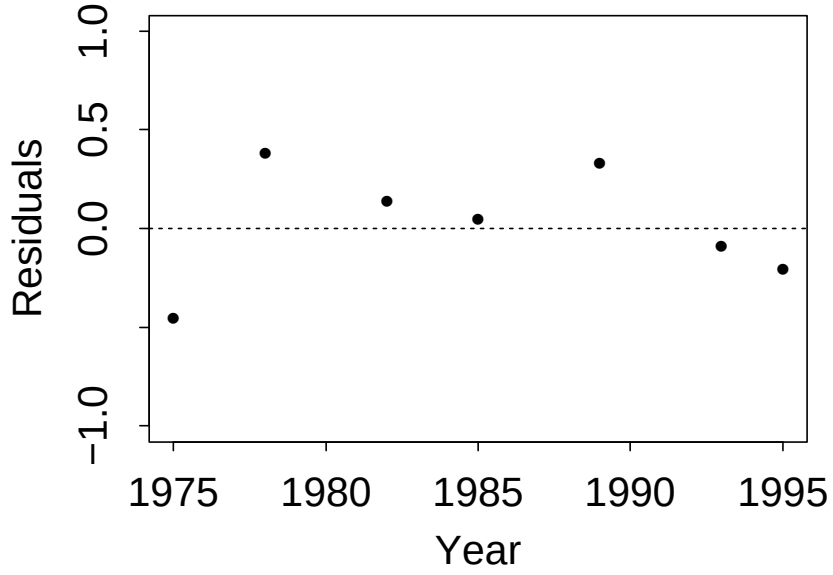
Fitting a linear regression of $z = \log(y)$ on x (Year):

$$\widehat{z} = -668.162 + 0.343x, \quad \widehat{\sigma}^2 = 0.106,$$

which implies

$$\widehat{y} = \exp(-668.162 + 0.343x) = 7247.189 \exp\{0.343(x - 1975)\}.$$





Residuals look like random noise now.

We use the above model to predict the No. of transistors in the 'future'.

Year x	2000	2003	2006	2010
True values of y	42,000K	105,900K	291,000K	1,117,000K
Predicted values $\widehat{y}$	38,392K	107,432K	300,623K	1,185,427K

For this data set, using the transformation $z = \log y$ is the key in achieving a good analysis.

Remark. (i) LSE is derived from an empirical role, making no use of distribution properties.

(ii) If $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$, the likelihood is

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\Big\{ -\frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \Big\}.$$

It is easy to see that

$$L(\boldsymbol{\beta}, \sigma^2) \leq L(\widehat{\boldsymbol{\beta}}, \sigma^2),$$

where $\widehat{\boldsymbol{\beta}}$ is the LSE, therefore is also the MLE.

(iii) Since $E\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$,

$$E\widehat{\boldsymbol{\beta}} = E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Now consider the special case of $p = 1$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} = \frac{1}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}.$$

For the simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the LSE are

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}, \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Furthermore $E\widehat{\beta}_0 = \beta_0$, $E\widehat{\beta}_1 = \beta_1$, and

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}, \quad \mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}, \quad \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\sum_i (x_i - \bar{x})^2}$$

$$\mathrm{SE}(\widehat{\beta}_0) = \sqrt{\frac{\widehat{\sigma}^2 \sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}}.$$

In the above, $\widehat{\sigma}^2 = \frac{1}{2} \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$.

(iv) Under the assumption $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$, the MLE for σ^2 is $||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2/n$. But in practice, we usually use the unbiased estimator

$$\widehat{\sigma}^2 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2/(n-p-1).$$

Proof. With estimated coefficients, the fitted model is

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_x\mathbf{y},$$

where $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It is easy to see that

$$\mathbf{P}_x = \mathbf{P}_x' = \mathbf{P}_x^2, \quad \mathbf{P}_x\mathbf{X} = \mathbf{X}.$$

(Actually $\mathbf{P}_x$ is the project matrix into the linear space spanned by the column vectors of $\mathbf{X}$.) Hence

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-p-1}||\mathbf{y} - \widehat{\mathbf{y}}||^2 = \frac{1}{n-p-1}||(\mathbf{I}_n - \mathbf{P}_x)\mathbf{y}||^2 \\
&= \frac{1}{n-p-1}\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_x)^2\mathbf{y} = \frac{1}{n-p-1}\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_x)^2\boldsymbol{\varepsilon} \\
&= \frac{1}{n-p-1}\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_x)\boldsymbol{\varepsilon}.
\end{aligned}$$

$$
\begin{aligned}
E(\widehat{\sigma}^2) &= \frac{1}{n-p-1} E\{\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_x)\boldsymbol{\varepsilon}\} = \frac{1}{n-p-1} \text{trace} E\{(\mathbf{I}_n - \mathbf{P}_x)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\} \\
&= \frac{1}{n-p-1} \text{trace}\{(\mathbf{I}_n - \mathbf{P}_x)\text{Var}(\boldsymbol{\varepsilon})\} = \frac{\sigma^2}{n-p-1} \text{trace}(\mathbf{I}_n - \mathbf{P}_x) \\
&= \frac{\sigma^2}{n-p-1}(n - \text{trace}(\mathbf{P}_x)) = \frac{\sigma^2}{n-p-1}\{n - \text{trace}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X})\} \\
&= \frac{\sigma^2}{n-p-1}\{n - \text{trace}(\mathbf{I}_{p+1})\} = \frac{\sigma^2}{n-p-1}\{n-p-1\} = \sigma^2.
\end{aligned}
$$

In the above derivation, we have used following facts:

(i) $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$ for any matrices $\mathbf{A}$, $\mathbf{B}'$ of the same size.

(ii) $E(\mathbf{AZ}) = \mathbf{A}E(\mathbf{Z})$, where $\mathbf{A}$ is a constant matrix, $\mathbf{Z}$ is a random matrix.

(iii) $\text{Cov}(\mathbf{A}\boldsymbol{\xi}, \mathbf{B}\boldsymbol{\eta}) = \mathbf{A}\text{Cov}(\boldsymbol{\xi}, \boldsymbol{\eta})\mathbf{B}'$, where $\mathbf{A}, \mathbf{B}$ are constant matrices, and $\boldsymbol{\xi}, \boldsymbol{\eta}$ are random vectors.

(v) Residual vector: $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_x)\mathbf{y}$. Then $E\widehat{\boldsymbol{\varepsilon}} = 0$, and $\text{Var}(\widehat{\boldsymbol{\varepsilon}}) = (\mathbf{I}_n - \mathbf{P}_x)\text{Var}(\mathbf{y})(\mathbf{I}_n - \mathbf{P}_x) = \sigma^2(\mathbf{I}_n - \mathbf{P}_x)$.

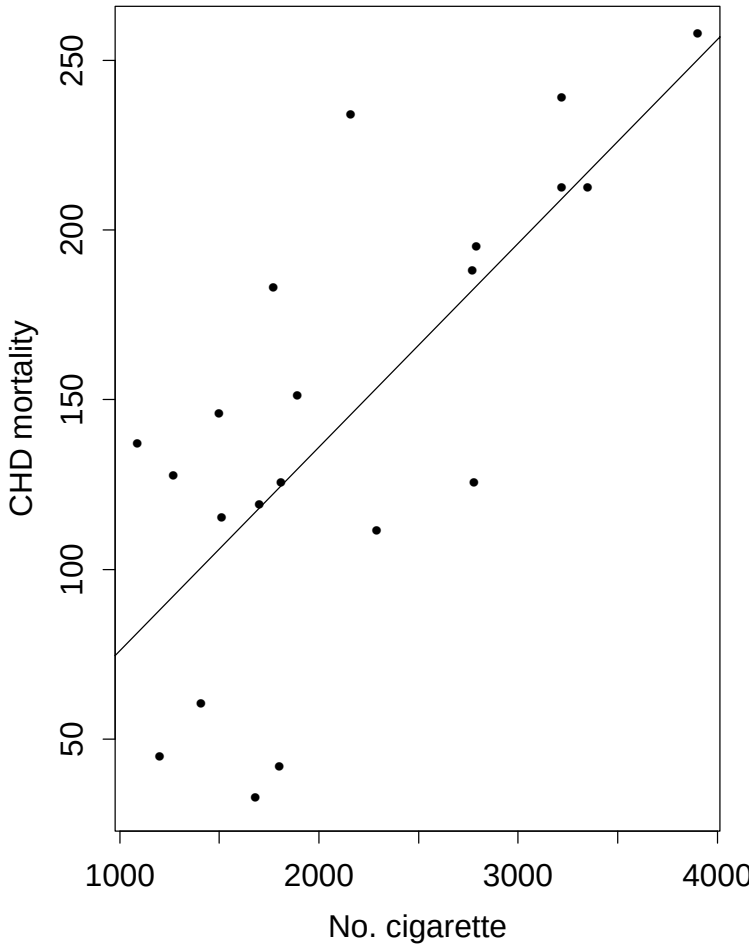
Example 4. The data set 'cigarette.txt' contains the annual cigarette consumption (x), and the corresponding mortality rate (y), due to coronary heart disease (CHD) of 21 countries. Do these data support the suspicion that smoking contributes the CHD mortality?

Country	No. Cigarette per adult per year	CHD mortality per 100,000 (age 35-64)
USA	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
NewZealand	3220	211.8
UK	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
.....	.....	

Note. The assertion "Smoking is harmful for health" is largely based on statistical, rather than laboratory, evidence.

1. Plot y against x. More powerful (than the direct inspection on data) to reveal the pattern on how y and x are related with each other.
   It also indicates if transformation on y or/and x may be explored.
2. Fitting regression model $y = \beta_0 + \beta_1 x + \varepsilon$.
   Note $n = 21$, and

$$\sum_i x_i = 45110, \quad \sum_i y_i = 3042.2,$$
$$\sum_i x_i^2 = 109957100, \quad \sum_i y_i^2 = 529321.58,$$
$$\sum_i x_i y_i = 7319602.$$



$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i x_i y_i - \sum_i x_i \sum_j y_j / n}{\sum_i x_i^2 - (\sum_i x_i)^2 / n}$$
$$= (7319602 - 45110 \times 3042.2/21) / \{109957100 - (45110)^2/21\} = 0.06$$

$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = (3042.2 - 0.06 \times 4511)/21 = 15.77$, and
$\widehat{\sigma}^2 = (n-2)^{-1} \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = 2181.66$.

Superimpose the regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ into the plot. It should go through the middle of the data crowds.

3. Testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$.
   If indeed smoking contributes the CHD mortality, $\beta_1 > 0$. Since $\widehat{\beta}_1 = 0.06$, is it significant? — Apply the Wald test now.

   Under H_0, $T = \widehat{\beta}_1 / \mathrm{SE}(\widehat{\beta}_1) \sim N(0,1)$ approximately.

   $\mathrm{SE}(\widehat{\beta}_1) = \widehat{\sigma} / \{\sum_i (x_i - \bar{x})^2\}^{1/2} = 0.01475$.

   Since $T = 0.06/0.01475 = 4.068 > z_{0.01} = 2.326$, we reject the hypothesis $\beta_1 = 0$ at the 1% significance level.

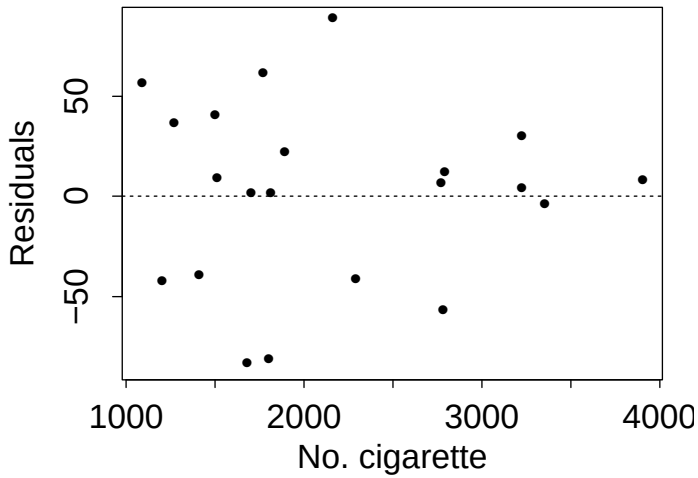
Remark. The magnitude of $\widehat{\beta}_1$ itself is not important in determining if $\beta_1 = 0$ or not: changing the scale of x may make $\widehat{\beta}_1$ arbitrarily small or large.

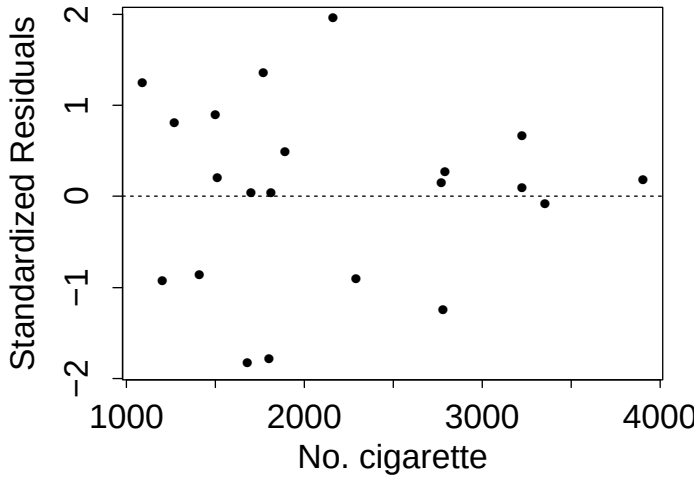
4. **Residuals analysis** (a preliminary): If the model is correct, the residuals

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i = \varepsilon_i + \beta_0 - \widehat{\beta}_0 + (\beta_1 - \widehat{\beta}_1)x_i, \quad i = 1, \cdots, n$$

should behave like independent (or uncorrelated) noise, or even normal if $\varepsilon_i \sim N(0,\ \sigma^2)$. Note $\sum_i \widehat{\varepsilon}_i = 0$.

Standardized residuals: $\widehat{\varepsilon}_i / \sqrt{\widehat{\sigma}^2(1 - h_{ii})}$, which should be approximately independent $N(0,\ 1)$, where h_{ii} is the (i, i)-th element of $\mathbf{P}_x$.





11.3 Gauss-Markov Theorem

Key message. The LSE $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$ (i.e. $\widehat{\beta}$ has the minimum variance among all the linear unbiased estimators for $\boldsymbol{\beta}$.)

Theorem [Gauss-Markov] For the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where } E\boldsymbol{\varepsilon} = 0 \text{ and } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

$\widehat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$ in the sense that

$$\text{Var}(\mathbf{By}) - \text{Var}(\widehat{\boldsymbol{\beta}})$$

is a positive semi-definite matrix for any $p \times n$ constant matrix $\mathbf{B}$ for which $E(\mathbf{By}) = \boldsymbol{\beta}$.

Remark. (i) For $j = 1, \cdots, p$, $\widehat{\beta}_j$ is the minimum variance unbiased linear estimator for β_j.

(ii) Suppose that $Y_1, Y_2, \ldots, Y_n$ are a random sample from a distribution with mean μ and variance σ^2. Then the the sample mean is a minimum variance unbiased linear estimator of μ.

Proof. First note

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.$$

Thus $\text{Var}(\widehat{\boldsymbol{\beta}}) = \text{Var}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\}$. Since $E(\mathbf{By}) = \mathbf{BX}\boldsymbol{\beta} = \boldsymbol{\beta}$, $\mathbf{BX} = \mathbf{I}_p$. Thus

$$\mathbf{By} = \mathbf{B}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \mathbf{B}\boldsymbol{\varepsilon}.$$

Therefore

$$\begin{aligned} \text{Var}(\mathbf{By}) &= \text{Var}(\mathbf{B}\boldsymbol{\varepsilon}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + \{\mathbf{B} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\varepsilon}] \\ &= \text{Var}(\widehat{\boldsymbol{\beta}}) + \text{Var}[\{\mathbf{B} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\varepsilon}] + \mathbf{R} + \mathbf{R}', \end{aligned}$$

where

$$\begin{aligned} \mathbf{R} &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\{\mathbf{B} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}'] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\{\mathbf{B}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} \\ &= \sigma^2\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}' - (\mathbf{X}'\mathbf{X})^{-1}\} = 0. \end{aligned}$$

Thus

$$\text{Var}(\mathbf{By}) - \text{Var}(\widehat{\boldsymbol{\beta}}) = \text{Var}[\{\mathbf{B} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\varepsilon}]$$

which is positive semi-definite.

11.4 Normal linear regression models

Note $(\mathbf{I}_n - \mathbf{P}_x)\mathbf{X} = 0$. Thus $\text{Cov}(\mathbf{P}_x\mathbf{y}, (\mathbf{I}_n - \mathbf{P}_x)\mathbf{y}) = 0$.

Therefore, $\mathbf{P}_x\mathbf{y}$ and $(\mathbf{I}_n - \mathbf{P}_x)\mathbf{y}$ are independent under the condition $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$.

Consequently, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independent, as $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_x\mathbf{y}$. Note

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_x(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{I}_n - \mathbf{P}_x)\mathbf{y} = \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \widehat{\boldsymbol{\varepsilon}},$$

$$\begin{aligned} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 &= ||\mathbf{P}_x(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||^2 + ||(\mathbf{I}_n - \mathbf{P}_x)\mathbf{y}||^2 \\ &= ||\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 + (n - p - 1)\widehat{\sigma}^2. \end{aligned}$$

Theorem. If $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$, then

(i) $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, $\frac{n-p-1}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2(n - p - 1)$, and

(ii) $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independent.

The t-Tests for β_j

By the theorem above,

$$(\widehat{\beta}_j - \beta_j)/\mathrm{SE}(\widehat{\beta}_j) \sim t_{n-p-1}, \quad \mathrm{SE}(\widehat{\beta}_j) = (\widehat{\sigma}^2 v_{jj})^{1/2},$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)'$, and v_{jj} is the $(j+1, j+1)$-element of $(\mathbf{X}'\mathbf{X})^{-1}$.

For the normal regression model, let $T = (\widehat{\beta}_j - \beta_{j0})/\mathrm{SE}(\widehat{\beta}_j)$. We reject the null hypothesis $H_0 : \beta_j = \beta_{j0}$ at the level α

against $H_1 : \beta_j \neq \beta_{j0}$ if $|T| > t_{\alpha/2, n-p-1}$, or

against $H_1 : \beta_j > \beta_{j0}$ if $T > t_{\alpha, n-p-1}$, or

against $H_1 : \beta_j < \beta_{j0}$ if $T < -t_{\alpha, n-p-1}$,

where $t_{\alpha,k}$ is the top α-point of the t_k-distribution.

The $(1-\alpha)$ confidence interval for β_j is $\widehat{\beta}_j \pm t_{\alpha/2, n-p-1}\mathrm{SE}(\widehat{\beta}_j)$

Remark. When n is large, t-test $\approx$ the Wald test.

F-distribution

Definition. Let U, V be two independent r.v.s, and $U \sim \chi^2_p$ and $V \sim \chi^2_k$. Then the distribution of

$$Z = \frac{U/p}{V/k}$$

is called the F-distribution with degrees of freedom $(p,\ k)$, denoted as $F_{p,\,k}$ or $F(p,\ k)$.

(i) If $Z \sim F_{p,\,k}$, $Z > 0$, and $Z^{-1} \sim F_{k,\,p}$.

(ii) For $Z \sim F_{p,\,k}$ with $k > 2$, $E(Z) = \frac{k}{k-2}$. Furthermore if $k > 4$,

$$\text{Var}(Z) = \frac{2k^2(p + k - 2)}{p(k-2)^2(k-4)}.$$

(iii) If $T \sim t_k$, $T^2 \sim F_{1,k}$.

F-Tests

Consider the hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$ against H_1 : not all $\beta_1, \cdots, \beta_p$ are o.

We introduce the decomposition:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

The term on the LHS is the total variation in data $\{y_i\}$ and is also called *the total sum of squares*, the two terms on the RHS are called, respectively, *the regression sum of squares* and *the residual sum squares*.

When $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n)$, the residual SS and the regression SS are independent. Furthermore, under H_0 above,

$$\text{Residual SS} \sim \chi^2_{n-p-1}, \qquad \text{Regression SS} \sim \chi^2_p$$

F-test: Let

$$T = \frac{\sum_i (\widehat{y}_i - \bar{y})^2 / p}{\sum_i (y_i - \widehat{y}_i)^2 / (n-p-1)} = \frac{(n-p-1)}{p} \frac{\text{Regression SS}}{\text{Residual SS}}.$$

We reject

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

at the level α if $T > F_{\alpha,p,n-p-1}$, where $F_{\alpha,p,n-p-1}$ is the top α-point of the $F_{p,n-p-1}$-distribution.

ANOVA (Analysis of Variances) Table

Source	DF	Sum Sq	Mean Sq	F-statistic	p-value
Regressors	p	$\sum_i (\widehat{y}_i - \bar{y})^2$	$\frac{\sum_i (\widehat{y}_i - \bar{y})^2}{p}$	$\frac{\sum_i (\widehat{y}_i - \bar{y})^2 / p}{\sum_i (y_i - \widehat{y}_i)^2 / (n-p-1)}$	p-value
Residual	$n-p-1$	$\sum_i (y_i - \widehat{y}_i)^2$	$\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-p-1}$		
Total	$n-1$	$\sum_i (y_i - \bar{y})^2$			

11.5 List of notation

Model: Let $\{(y_i, x_{i1}, \cdots, x_{ip}),\ i = 1, \cdots, n\}$ be observations from

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where $\{\varepsilon_i\}$ are uncorrelated with common mean o and variance σ^2.

LSE: $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$ are obtained by minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

Fitted regression model: $\widehat{y} = \widehat{\beta}_0 + x_1\widehat{\beta}_1 + \cdots + x_p\widehat{\beta}_p$.

Residuals: $\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j, j = 1, \cdots, n$.

An unbiased estimator for σ^2: $\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{\text{Residual SS}}{n-p-1}$.

Standardized residuals: $\widehat{\varepsilon}_i / \sqrt{\widehat{\sigma}^2(1 - h_{ii})}$**,** and h_{ii} is the (i, i)-th element of matrix $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$**,** $j = 1, \cdots, n$**.**

Regression SS: $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$, **Residual SS**: $\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$

Total SS: $\sum_{i=1}^{n}(y_i - \bar{y})^2$ (= Regression SS + Residual SS).

Regression correlation coefficient:

$$R = \left(\frac{\text{Regression SS}}{\text{Total SS}}\right)^{\frac{1}{2}} = \left(1 - \frac{\text{Residual SS}}{\text{Total SS}}\right)^{\frac{1}{2}} \in [0,\ 1]$$

$100R^2$ is the percentage of the total variation in $\{y_i\}$ explained by all the regressors.

Adjusted regression correlation coefficient:

$$R_{adj} = \left\{1 - \frac{(\text{Residual SS})/(n-p-1)}{(\text{Total SS})/(n-1)}\right\}^{\frac{1}{2}}$$

Testing for single coefficient: $H_0: \beta_i = 0$ vs $H_1: \beta_i \neq 0$. Let $T = \widehat{\beta}_i / \mathrm{SE}(\widehat{\beta}_i)$, reject H_0 at the level α if $|T| > t_{\alpha/2,\, n-p-1}$.

Note. Under H_0, $T \sim t_{n-p-1}$ and $T^2 \sim F_{1,n-p-1}$.

$(1-\alpha)$ **confidence intervals for** β_j: $\widehat{\beta}_i \pm t_{\alpha/2,\, n-p-1}\mathrm{SE}(\widehat{\beta}_i)$, or $\widehat{\beta}_i \pm z_{\alpha/2}\mathrm{SE}(\widehat{\beta}_i)$ when n is large.

Testing for all zero-regression coefficients:

$$H_0 : \ \beta_1 = \cdots = \beta_p = 0.$$

Let

$$T = \frac{(\text{Regression SS})/p}{(\text{Residual SS})/(n-p-1)}.$$

We reject H_0 at the α significance if $T > F_{\alpha,\, p, n-p-1}$.

Interpretation of β_j: the effect of changing x_j on y, *holding other x's fixed* — this is unfortunately not always practical.

11.6 Regression analysis with *R*

`lm` is an *R*-function for linear regression analysis, it returns an object of a special '`lm`-class' which contains all outputs from the analysis. You may use `summary, anova, resid, fitted` etc to view various outcomes of the analysis. It also provides some useful plots for diagnostic checking for the goodness-of-fit of the model.

We provide an illustration below by analyzing the data set 'foods.txt'.

There are observations on the 4 variables (in 4 columns):

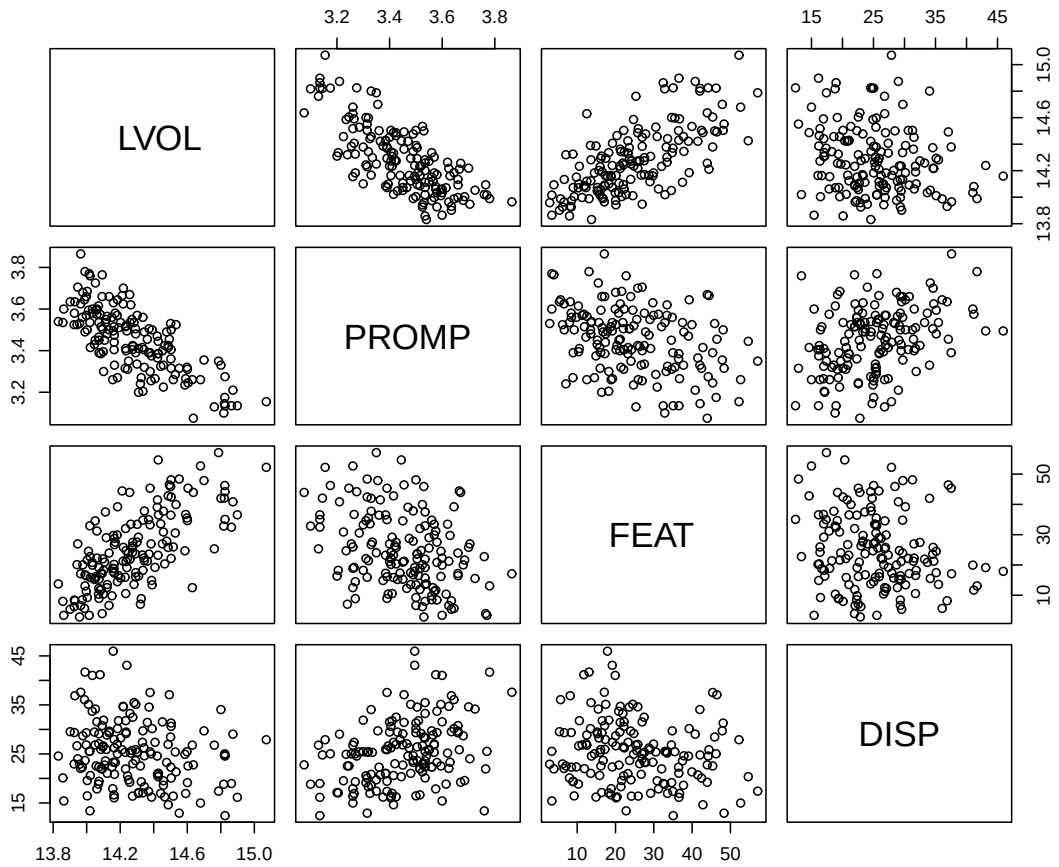
y: LVOL — logarithms of weekly sales volume
x_1: PROMP — promotion price
x_2: FEAT — feature advertising
x_3: DISP — display

We are interested in the impact on the sales (LVOL) of the other 3 features.

(i) Exploratory analysis.

```
> foods <- read.table("foods.txt", header=T)
> foods[1:5,]  # print out the first 5 rows of the data
     LVOL PROMP   FEAT  DISP
1 14.5334 3.525 39.920 21.35
2 14.2185 3.700 25.792 34.62
3 14.3330 3.415 23.280 27.37
4 14.2702 3.550 25.544 25.70
5 14.1843 3.645 39.248 30.15
> dim(foods)
[1] 156   4  # there are 156 observations on 4 variables
> attach(foods) # make the names for each variables accessible in R
> pairs(foods) # scatter plot for each pair of the 4 variables
```

"`pairs`" provides scatter plots of each pair among the 4 variables. They are particularly revealing on the linear correlations among different variables.



Negative correlation between LVOL and PROMP.

Positive correlation between LVOL and FEAT.

Little or no correlation between LVOL and DISP. However, positive correlation between DISP and PROMP.

```
> summary(foods)
```

```
      LVOL            PROMP            FEAT            DISP       
 Min.   :13.83   Min.   :3.075   Min.   : 2.84   Min.   :12.42  
 1st Qu.:14.08   1st Qu.:3.330   1st Qu.:15.95   1st Qu.:20.59  
 Median :14.24   Median :3.460   Median :22.99   Median :25.11  
 Mean   :14.28   Mean   :3.451   Mean   :24.84   Mean   :25.31  
 3rd Qu.:14.43   3rd Qu.:3.560   3rd Qu.:33.49   3rd Qu.:29.34  
 Max.   :15.07   Max.   :3.865   Max.   :57.10   Max.   :45.94  
```

We note that the values of FEAT and DISP are much bigger than the values LVOL.

(ii) Fitting regression model

```
> lm3foods <- lm(LVOL ~ PROMP + FEAT + DISP)
       # record the output as lm3foods
> summary(lm3foods)  # display the main results
Call:
lm(formula = LVOL ~ PROMP + FEAT + DISP)
Residuals:
      Min        1Q    Median        3Q       Max
-0.333630 -0.082033 -0.002723  0.079273  0.338120
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.2372251  0.2490226  69.220   <2e-16 ***
PROMP       -0.9564415  0.0726777 -13.160   <2e-16 ***
FEAT         0.0101421  0.0008728  11.620   <2e-16 ***
DISP         0.0035945  0.0016529   2.175   0.0312 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1253 on 152 degrees of freedom
Multiple R-squared: 0.7633,   Adjusted R-squared: 0.7587
F-statistic: 163.4 on 3 and 152 DF,  p-value: < 2.2e-16
```

Calling `summary(lm3foods)` prints out the main results.

The fitted model is

$$\text{LVOL} = 17.2372 - 0.9564\text{PROMP} + 0.0101\text{FEAT} + 0.0036\text{DISP}.$$

All the coefficients have the right sign.

The t-test for testing $H_0 : \beta_j = 0$ are extremely significant for $j = 0, 1$ and 2 with the p-values smaller than 2×10^{-16}. The t-statistic $\widehat{\beta}_j/\text{SE}(\beta_j) = 69.22, -13.16, 11.62$, respectively, for $j = 0, 1, 2$, indicating

- the intercept term 17.2372 is necessary in the model,
- both PROMP and FEAT have significant impact on LVOL

  LVOL $\nearrow$ as PROMP $\searrow$ and LVOL $\nearrow$ as FEAT $\nearrow$

The t-test for $H_0 : \beta_3 = 0$ yields the p-value 0.0312, indicating DISP also has a positive impact on LVOL (since $\widehat{\beta}_3 = 0.0036$), but less significant than the other two variables. (We might consider the option of fitting the model with two predictors only.)

The estimated σ (not σ^2!): $\widehat{\sigma} = 0.1253$ with d.f. $n - p - 1 = 156 - 3 - 1 = 152$.

$R^2 = 0.7633$ and $R^2_{adj} = 0.7587$, i.e. the 76.33% of the total variation on LVOL is explained by the 3 variables.

The null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is rejected by the F-test with p-value $< 2^{-16}$. The F-test statistic = 163.4. It has F-distribution under H_0 with the d.f. $(p,\ n - p - 1) = (3,\ 152)$.

```
> anova(lm3foods)
Analysis of Variance Table
Response: LVOL
           Df Sum Sq Mean Sq F value  Pr(>F)
PROMP       1 5.5269  5.5269 352.305 < 2e-16 ***
FEAT        1 2.0903  2.0903 133.243 < 2e-16 ***
DISP        1 0.0742  0.0742   4.729 0.03120 *
Residuals 152 2.3845  0.0157
```

`anova(lm3foods)` prints out the ANOVA table. Note that *R* lists the decomposition for each regressor separately. We can also find directly $\widehat{\sigma}^2 = 0.0157$ from the ANOVA table.

The contribution from DISP is small: $\frac{0.0742}{5.5269+2.0903+0.0742} < 1\%$

You may extract the fitted values $\{\widehat{y}_i\}$ by calling `fitted(lm3foods)`, or the residuals $\{\widehat{\varepsilon}_i\}$ by `resid(lm3foods)`.

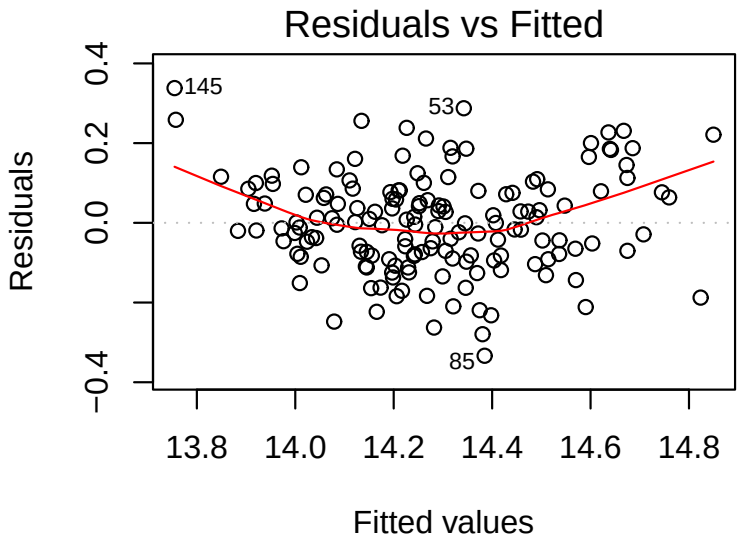
(iii) Diagnostic checking

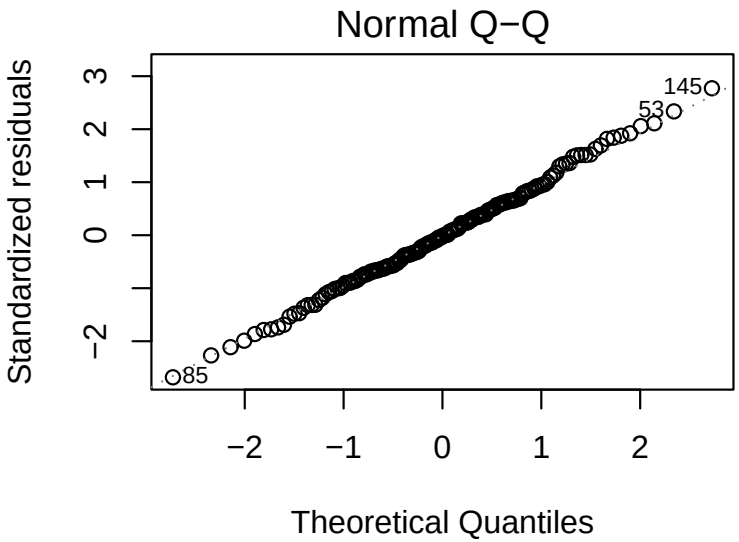
By default, *R* also produces some plots on residuals for diagnostic checking of the fitted model.

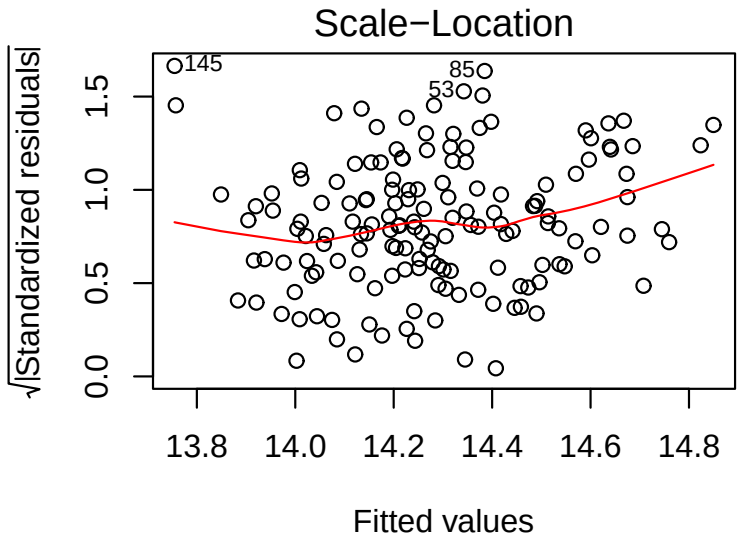
```
> par(mfrow=c(2,2))   # put the 4 plots in a 2x2 panel
> plot(lm3foods)
```

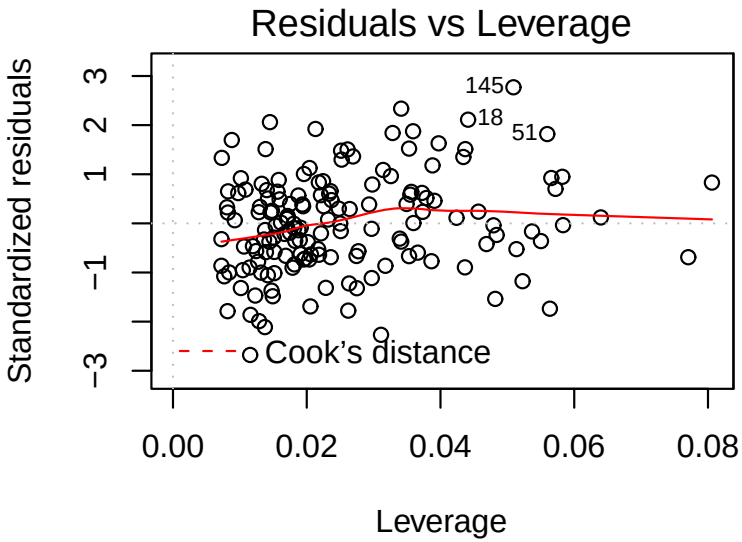
Residuals vs Fitted — a plot of $\widehat{\varepsilon}_i$ against $\widehat{y}_i$. If the model is adequate, this plot should be patternless, as $\widehat{\varepsilon}_i$ should behave like random noise. The plot is helpful in detecting outliers, i.e. those y_i far away from $\widehat{y}_i$.

Normal Q-Q — a plot of the quantiles of standardized residuals against the $N(0, 1)$ quantiles. It shows if residuals are normally distributed or not, and is particularly effective in highlighting heavy tails.


Residuals vs Fitted
Residuals
Fitted values
145
53
85
−0.4
0.0
0.2
0.4
13.8
14.0
14.2
14.4
14.6
14.8


Normal Q−Q
Standardized residuals
Theoretical Quantiles
145
53
85
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
145
85
53
0.0
0.5
1.0
1.5
13.8
14.0
14.2
14.4
14.6
14.8


Residuals vs Leverage
Standardized residuals
Leverage
145
18
51
Cook's distance
−3
−1
1
2
3
0.00
0.02
0.04
0.06
0.08

Scale-Location — a plot of $\sqrt{|\widetilde{\varepsilon}_i|}$ against $\widehat{y}_i$, where $\widetilde{\varepsilon}_i$ denotes the standardized residual. This plot should be patternless too if the fitting is adequate. It is powerful in detecting inhomogeneous variances among different observations. Note that for $Z \sim N(0, \sigma^2)$, $|Z|$ is heavily skewed to the left, and $\sqrt{|Z|}$ is much less skewed.

Residuals vs Leverage – a plot of $\widetilde{\varepsilon}_i$ against leverage h_{ii}, where h_{ii} is the (i, i)-th element of the hat matrix $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$.

A **leverage point** is an observation which has a great influence on the analysis. The amount of the leverage of the i-th observation is reflected by h_{ii}. It is easy to see that

$$\sum_{i=1}^{n} h_{ii} = \text{trace}(\mathbf{P}_x) = p + 1.$$

Therefore the average leverage for each observations is $\frac{p+1}{n}$.

A rule of thumb: if $h_{ii} > \frac{2(p+1)}{n}$, the i-th observation is a leverage point.

Note that the leverage depends on **X** only.

A leverage point is called a **good leverage point** if the corresponding y is close to $\widehat{y}$. It is called a **bad leverage point** if the corresponding y is an outlier.

For the food data set, $\frac{2(p+1)}{n} = \frac{8}{156} = 0.0513$. The figure shows the 51th and perhaps also 145th observations are bad leverage points. We may consider to remove them from the analysis, since they have great influence on the fitted model.

(iv) Prediction

We may use the fitted model to predict the future sales LVOL, given the values for PROMP, FEAT and DISP. The new values of the predictors must be in the format of `data.frame` (i.e. the collection of p vectors). Check the help manual with `?predict.lm`.

```
> PROMP <- c(3, 3.2, 3.6, 4)
> FEAT <- c(15, 40, 35, 26)
> DISP <- c(37, 20, 15, 56)
> newx <- data.frame(PROMP,FEAT,DISP)
> predict(lm3foods, newx, se.fit=T, interval="prediction", level=0.95)
       fit      lwr      upr
1 14.65303 14.38867 14.91738  # point predict 14.57395, and
                      # predictive interval (14.31591, 14.83199)
2 14.65418 14.40352 14.90484
3 14.20292 13.94953 14.45632
4 13.87644 13.60663 14.14625
```

```
> se.fit
         1          2          3          4
0.04706611 0.02021486 0.02759204 0.05442495
           # standard errors for the point predicts
```

Remark. If we change the flag

```
interval="prediction"
```

to

```
interval="confidence"
```

in the `predict`-command line, *R* will return the confidence intervals for $E\,y$ instead of the predictive intervals for y.

11.7 Model selection

When p is large, there may exist significant correlations among p predictors. Therefore not all p variables are required in the model. Therefore we need to select a model with all but only necessary variables.

A model with fewer predictors improves the interpretability, also saves from collecting useless data. More importantly it leads to more stable statistical inference (i.e. estimation, test and prediction).

Two issues in model selection:

- assign a 'score' to each possible model which measures its 'goodness' in relation to the data,
- search through all the models to find the one with the best score.

Note. There are 2^p possible models!

Let $\mathcal{J} \subset \{1, \cdots, p\}$ be a subset of indices, $|\mathcal{J}|$ be the number of elements in $\mathcal{J}$. Let $\mathbf{X}_{\mathcal{J}}$ be $n \times (|\mathcal{J}| + 1)$ matrix consisting of the first column (with all elements being 1) and the columns of $\mathbf{X}$ corresponding to the indices in $\mathcal{J}$, and $\boldsymbol{\beta}_{\mathcal{J}}$ be $(|\mathcal{J}| + 1)$-vector consisting of the components of $\boldsymbol{\beta}$ corresponding to the indices in $\mathcal{J}$.

If we regress $\mathbf{y}$ on $\mathbf{X}_{\mathcal{J}}$ only, the fitted value for $\mathbf{y}$ is

$$\widehat{\mathbf{y}}(\mathcal{J}) = \mathbf{X}_{\mathcal{J}}(\mathbf{X}'_{\mathcal{J}}\mathbf{X}_{\mathcal{J}})^{-1}\mathbf{X}'_{\mathcal{J}}\mathbf{y}.$$

The residual sum of squares is

$$R(\mathcal{J}) = ||\widehat{\mathbf{y}}(\mathcal{J}) - \mathbf{y}||^2 = \sum_{i=1}^{n} \{\widehat{y}_i(\mathcal{J}) - y_i\}^2.$$

When we add any element in $\{1, \cdots, p\} - \mathcal{J}$ into $\mathcal{J}$, $R(\mathcal{J})$ will only decrease. Adding more variables into the model will eventually lead to *overfitting*.

Therefore we have to penalize the complexity of the model appropriately. We introduce two criteria below.

Mallow's C_p: Let $R_m(\mathcal{J}) = R(\mathcal{J}) + 2|\mathcal{J}|\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the estimator for σ^2 obtained from the full model with all p predictors. We select the model which minimizes $R_m(\mathcal{J})$.

AIC criteria: Let $\text{AIC}(\mathcal{J}) = \log\{R(\mathcal{J})/n\} + 2|\mathcal{J}|/n$. We select the model which minimizes $\text{AIC}(\mathcal{J})$.

Remark. Each of the above two criteria consist of two terms: the first term measures the goodness-of-fit of the model, and the second term penalize the complexity of the model. When adding more variables into the model, the first term decreases while the second term increases. The selected model is the trade-off between the two.

When p is moderately large, the searching over all the 2^p possible models is a time-consuming (if ever possible). In practice, some stepwise schemes are often employed in searching for an optimum model in a subset of the 2^p models.

A Stepwise Addition and Deletion Scheme:

Step 1. start with an initial model (e.g. with only intercept and no predictors)

Step 2. **(a)** specify one variable such that by adding it to the model, the decrease of the residual SS is maximized among all candidate variables.

(b) only add the variable specified in (a) to the model if the value of the AIC (or Mallow's C_p) decreases.

Step 3. **(i)** specify one variable among all the variables in the model, such that by deleting it from the model, the increase of the residual SS is minimized.

(ii) only delete the variable specified in (i) if the value of the AIC (or Mallow's C_p) decreases.

(iii) Repeat (i) and (ii) above such that no more variables can be deleted

Step 4. Repeat Steps 2 & 3 above until no more variables can be added to or deleted from the model.

The *R*-function `step` implements the above scheme using the AIC. We illustrate it using the foods data.

```
> foods <- read.table("data/foods.txt", header=T)
> attach(foods)
```

```
> lm1foods<-lm(LVOL~1)   # fitting a simple model
                         # with the intercept term only
> lm3foods <- lm(LVOL ~ PROMP + FEAT + DISP) # fitting the full
                                  # model with 3 predictors
> stepFoods <- step(lm1foods, scope=list(upper=lm3foods))
            # stepwise search starting with lm1foods, lm3foods
            # is the largest model to be selected
Start:  AIC=-425.39
LVOL ~ 1

        Df Sum of Sq     RSS     AIC
+ PROMP  1    5.5269  4.5490 -547.45 # AIC-value from adding PROMP
+ FEAT   1    4.8915  5.1844 -527.06 # AIC-value from adding FEAT
+ DISP   1    0.4327  9.6432 -430.24 # AIC-value from adding DISP
<none>              10.0759 -425.39

Step:  AIC=-547.45
LVOL ~ PROMP

        Df Sum of Sq     RSS     AIC
+ FEAT   1    2.0903  2.4587 -641.43
```

```
<none>                  4.5490 -547.45
+ DISP   1    0.0461   4.5029 -547.04
- PROMP  1    5.5269 10.0759 -425.39 # AIC-value from deleting PROMP

Step:  AIC=-641.43
LVOL ~ PROMP + FEAT

        Df Sum of Sq    RSS     AIC
+ DISP   1   0.07419 2.3845 -644.21
<none>               2.4587 -641.43
- FEAT   1   2.09028 4.5490 -547.45
- PROMP  1   2.72563 5.1844 -527.06

Step:  AIC=-644.21
LVOL ~ PROMP + FEAT + DISP

        Df Sum of Sq    RSS     AIC
<none>               2.3845 -644.21
- DISP   1   0.07419 2.4587 -641.43
- FEAT   1   2.11837 4.5029 -547.04
- PROMP  1   2.71691 5.1015 -527.57
```

```
Call:
lm(formula = LVOL ~ PROMP + FEAT + DISP)

Coefficients:
(Intercept)        PROMP         FEAT         DISP
  17.237225    -0.956442     0.010142     0.003594
```

For the data set, the AIC chooses the full model with the 3 predictors. The more detailed information on the selected model may be viewed by calling `summary(stepFoods)`.