An Introduction to Regression Analysis

By John Tyssedal

Sir Francis Galton (1822-1911), a well-known British anthropologist and meteorologist, was the first to introduce the word regression. It appears in a publication in Nature in 1885 where it is used to describe a biological phenomenon, namely that the heights of descendants of tall ancestors tend to regress down towards a normal average, a phenomenon also known as regression towards the mean. To Francis Galton regression had a pure biological meaning, but the term regression soon came to be applied to relationships in situations other than the one from which it originally arose. Today the goal of performing a regression analysis is to find a relationship between a response variable and one or more regression variables or to make models for prediction. Regression analysis is one of the most widely used statistical techniques. You may find that for practitioners of statistics, regression analysis is the method they are most familiar with. In the simplest case there is only one regression variable to explain the response variable. This is called simple linear regression.

Simple linear regression

The model for simple linear regression is $Y = \alpha + \beta x + \varepsilon$ where α and β are constants. The response variable Y (also called dependent variable) is a random variable while the regression variable x (also called independent or explanatory variable) can be deterministic or stochastic. ε is a noise variable assumed to have expected value 0 and constant variance, σ^2. To fit a model, we need to have a certain number, say n, of observation pairs (x_i, y_i) of the regression and the response variable. $(y_1, y_2, \ldots, y_n)$ can then be considered to be realizations of $(Y_1, Y_2, \ldots, Y_n)$, where

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

and $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ are assumed to be uncorrelated. It is important to be aware of that by a linear model we mean linear in the coefficients such that x_i can be substituted by any transformation, for instance $ln(x_i)$, x_i^2 or $sin(x_i)$, as long as the transformation is allowed.

While the response variables, $Y_1, Y_2, \ldots, Y_n$ are random variables, the regression variables $x_1, x_2, \ldots, x_n$ can be either deterministic or stochastic. This has some implications in the interpretation of the model. If $x_1, x_2, \ldots, x_n$ are deterministic we have

$$E[Y_i] = \alpha + \beta x_i.$$

If $x_1, x_2, \ldots, x_n$ are stochastic variables we have

$$E[Y_i | x = x_i] = \alpha + \beta x_i.$$

Hence, the parameter α is the expected response value when $x = 0$. Depending on whether $x_1, x_2, ..., x_n$ are deterministic or stochastic the parameter β is the expected change in the mean or the conditional mean respectively of the response when the regression variable changes with one unit.

Now let us assume we have n observation pairs: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. If these satisfy (1), we should have:

$$\begin{aligned} y_1 &= \alpha + \beta x_1 + \varepsilon_1^* \\ y_2 &= \alpha + \beta x_2 + \varepsilon_2^* \\ &\vdots \\ y_n &= \alpha + \beta x_n + \varepsilon_3^* \end{aligned}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{bmatrix},$$

with ε_i^* being realizations of ε_i, i=1,2,…,n.

With $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ and $\boldsymbol{\varepsilon}^* = \begin{bmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_2^* \end{bmatrix}$, we get the shorter notation

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$.

The least square estimates of (α, β) are the values (a, b) that minimizes

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

with respect to *a* and *b*.

These are known to be the ones that satisfies

$$\frac{dQ}{da} = 0 \Leftrightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \qquad (2)$$

$$\frac{dQ}{db} = 0 \Leftrightarrow -2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)x_i = 0\,. \qquad (3)$$

From (2) and (3) we obtain the two normal equations

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (4)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (5)$$

which gives

$$b = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad \text{and} \quad a = \bar{y} - b\bar{x} \qquad (6)$$

with corresponding least squares estimators

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)Y_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{x}\,. \qquad (7)$$

Estimated model (estimated expected regression line) is then:

$$\hat{y} = a + bx \quad \text{or} \quad \hat{y}_i = a + bx_i\,.$$

The deviations $e_i = y_i - \hat{y}_i$, i=1,2,…,n are called residuals. These are estimates of $\left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)$ and should have approximately the same properties, i.e. uncorrelated with mean equal to zero and constant variance. Figure 1 shows some pattern that may be observed when plotting the residuals.

The plot of e1 against $\hat{y}$ (”yhat”) shows an increase in the residual variance when $\hat{y}$ is increasing. The plot of e2 against $\hat{y}$ is u-shaped indicating the lack of a quadratic term. This is supported by the plot of e2 against x which has the same form. Finally, the plot off e against $\hat{y}$ shows a residual plot, where no specific pattern is observed. Such a plot will support our assumption about uncorrelated residuals with a constant variance. In addition, a normal plot should be used to check if a normal distribution is an appropriate assumption. Most procedures for statistical inference in regression analysis rely on such an assumption.

Transformations for constant variance.

When a pattern as observed when e1 is plotted against $\hat{y}$ occurs, the response should be transformed to obtain constant variance, otherwise our least squares estimators will not be optimal. The following calculations give a motivation for how the response should be transformed. Let $z = g(y)$ be a transformation of y. A first order approximation with a Taylor series gives us:

$$z \approx g(\mu) + g'(\mu)(y - \mu) \text{ where } \mu \text{ is some constant.}$$

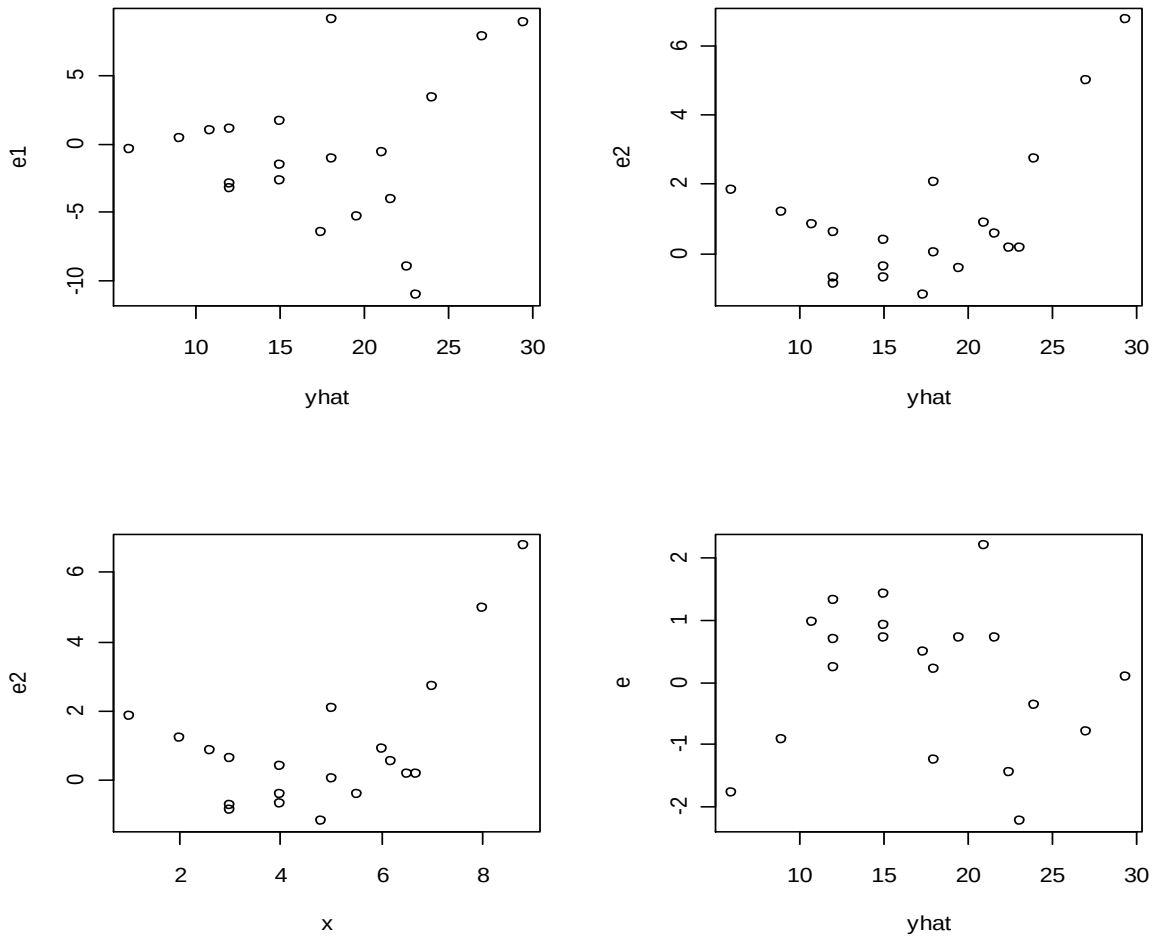


Figure 1. Four residuals plots illustrating: increase in variance with increasing $\hat{y}$, two u-shaped residual plots illustrating lack of a seconds order term in the model and finally a plot where the residuals have no particular pattern and thereby support our choice of model.

Hence as an approximation it should be possible to write a random variable $Z = g(Y)$ as:

$$Z \approx g(\mu) + g'(\mu)(Y - \mu) \quad \text{where } \mu = E(Y).$$

Thereby we get

$$Var(Z) \approx g'(\mu)^2 Var(Y).$$

For the variance of $Z = g(Y)$ to be (approximately) constant we must have:

$$g'(\mu)^2 Var(Y) = k \quad \text{or} \quad g'(\mu) = \frac{\sqrt{k}}{SD(Y)} \text{ where } k \text{ is a constant.}$$

The following situations give raise to the most common transformations:

$SD(Y) \propto \sqrt{\mu} \Rightarrow g'(\mu) = \dfrac{\sqrt{k}}{c\sqrt{\mu}}$ and an appropriate transformation will be $g(Y) = \sqrt{Y}$

$SD(Y) \propto \mu \Rightarrow g'(\mu) = \dfrac{\sqrt{k}}{c\,\mu}$ and an appropriate transformation will be $g(Y) = ln(Y)$

$SD(Y) \propto \mu^2 \Rightarrow g'(\mu) = \dfrac{\sqrt{k}}{c\,\mu^2}$ and an appropriate transformation will be $g(Y) = \dfrac{1}{Y}$

There is a constant missing in the transformations but that is irrelevant since the goal is to obtain constant variance. In general the transformation can be of the form Y^λ with λ arbitrary, where $\lambda = 0$ means the logarithm. If none of the transformations above gives a satisfactory results, one can try the Box-Cox transformation. The steps are as follows:

Suppose $Y_i > 0, \ \forall i$.

1. Consider a value of λ from a grid on a selected range for instance [-2,2], [-1,1].
2. For each value of λ compute new responses:

$$V_i = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda \left(Y^*\right)^{\lambda - 1}}, & \lambda \neq 0 \\ Y^* \ln(Y_i), & \lambda = 0 \end{cases}$$

where $Y^* = \left(\prod_{i=1}^{n} Y_i \right)^{\frac{1}{n}}$ i.e. the geometric mean of $Y_1, \ldots, Y_n$. Hence

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \rightarrow \boldsymbol{V} = (V_1, \ldots, V_n)^T$.

3. Regress $\boldsymbol{V}$ on $\boldsymbol{X}$ and find $SS_E(\lambda) = \sum_{i=1}^{n} (v_i - \hat{v}_i)^2$ for each λ.

4. Plot $SS_E(\lambda)$ versus λ. Draw a smooth curve and find $\lambda = \lambda^*$ that gives the smallest SS_E. If λ^* is close to values like $\ldots, -2, -1.5, -1. -0.5, 0, 0.5, 1, 1.5, 2, \ldots$, use these instead for easier interpretation.
5. Use the transformation Y^{λ^*}.

Transformations can also sometimes transfer a nonlinear model into a linear one as shown for the following three non-linear models below:

$$Y_i = \alpha e^{\beta x_i + \varepsilon_i} \Rightarrow ln(Y_i) = ln\,\alpha + \beta x_i + \varepsilon_i,$$

$$Y_i = \alpha x_i^{\beta} \varepsilon_i \Rightarrow ln(Y_i) = ln\,\alpha + \beta ln(x_i) + ln(\varepsilon_i),$$

$$Y_i = \frac{x_i}{\alpha + (\beta + \varepsilon_i) x_i} \Rightarrow \frac{1}{Y_i} = \frac{\alpha}{x_i} + \beta + \varepsilon_i.$$

After the transformation the model has become linear in the parameters $(ln(\alpha), \beta)$, $(ln(\alpha), \beta)$ and (α, β) respectively. These cases show that it may also be necessary to transform regression variables to obtain a linear model. Variance stabilized data will also normally have a distribution that is close to the normal distribution.

An example

The dataset below consists of 30 observation pairs of density and stiffness for a particular wood product. Since data for density are more easily obtainable than stiffness, it is of interest to find a relationship between stiffness and density such that given a density one is able to estimate or predict the stiffness. Therefore, stiffness will be taken to be the response variable and density to be the regression variable.

i	(x_i, y_i)	i	(x_i, y_i)	i	(x_i, y_i)
1	(9.5, 14184)	11	(17.4, 43243)	21	(25.6, 96305)
2	(8.4, 17502)	12	(15.0, 25319)	22	(23.4, 104170)
3	(9.8, 14007)	13	(15.2, 28028)	23	(24.4, 72594)
4	(11, 19443)	14	(16.4, 41792)	24	(23.3, 49512)
5	(8.3, 7573)	15	(16.7, 49499)	25	(19.5, 32207)
6	(9.9, 14194)	16	(15.4, 25312)	26	(21.2, 48218)
7	(8.6, 9714)	17	(15.0, 26222)	27	(22.8, 70453)
8	(6.4, 8076)	18	(14.5, 22148)	28	(21.7, 47661)
9	(7.0, 5304)	19	(14.8, 26751)	29	(19.8, 38138)
10	(8.2, 10728)	20	(13.6, 18036)	30	(21.3, 53045)

We always start by plotting the response variable against the regression variable. The plot is shown in Figure 2. It seems like stiffness increases with increasing density.

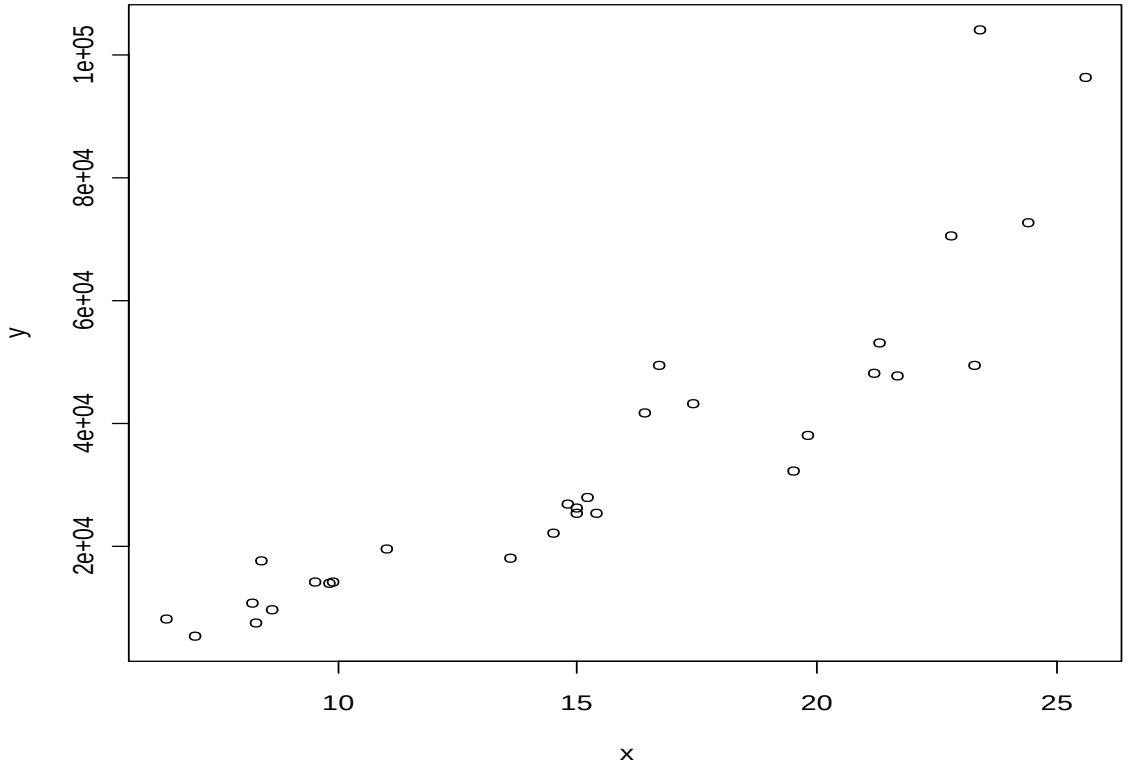


Figure 2. A plot of stiffness against density

Below is the result of a least square analysis performed with R. We shall comment on this analysis in greater detail later. Now we notice that α is estimated to -25513.3 and β to 3888.8 and that both are highly significant. The estimated value for σ is 11620.

```
Coefficients:
             Estimate   Std. Error t value   Pr(>|t|)
(Intercept) -25513.3     6101.6    -4.181   0.000258 ***
x             3888.8      369.8    10.515   3.14e-11 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 ---
Residual standard error: 11620 on 28 degrees of freedom
```

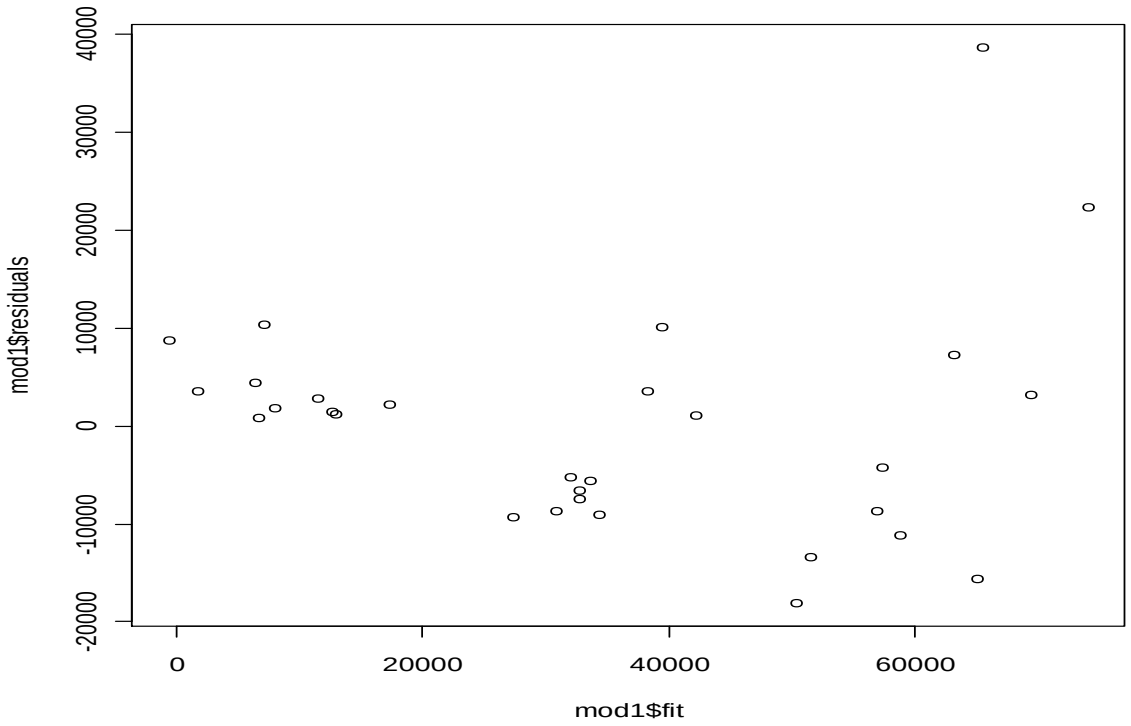


Figure 3. Plot of residuals against $\hat{y}$.

The plot of the residuals shows that the variance is increasing when $\hat{y}$ or the estimated mean is increasing. This calls upon a transformation. A plot of the log transformed stiffness against density is given in Figure 4.

With $\log(y)$ as the response we get the following least squares estimates

```
Coefficients:
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  8.25193    0.12819   64.37   < 2e-16 ***
x            0.12518    0.00777   16.11   1.08e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2441 on 28 degrees of freedom
```

The plot of the residuals against $\hat{y}$ with the log transformed response indicates a far better fit as can be seen in Figure 5, and the normal plot indicates that their distribution is not too far away from a normal distribution.

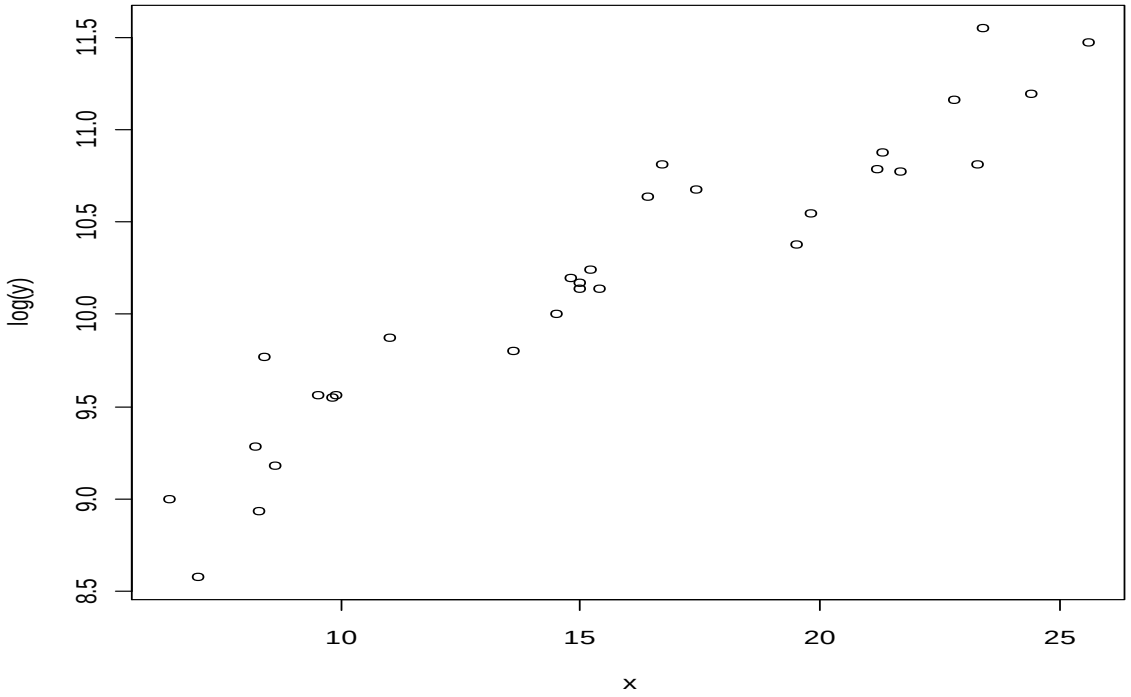


Figure 4. Plot of the logarithm of stiffness against density.

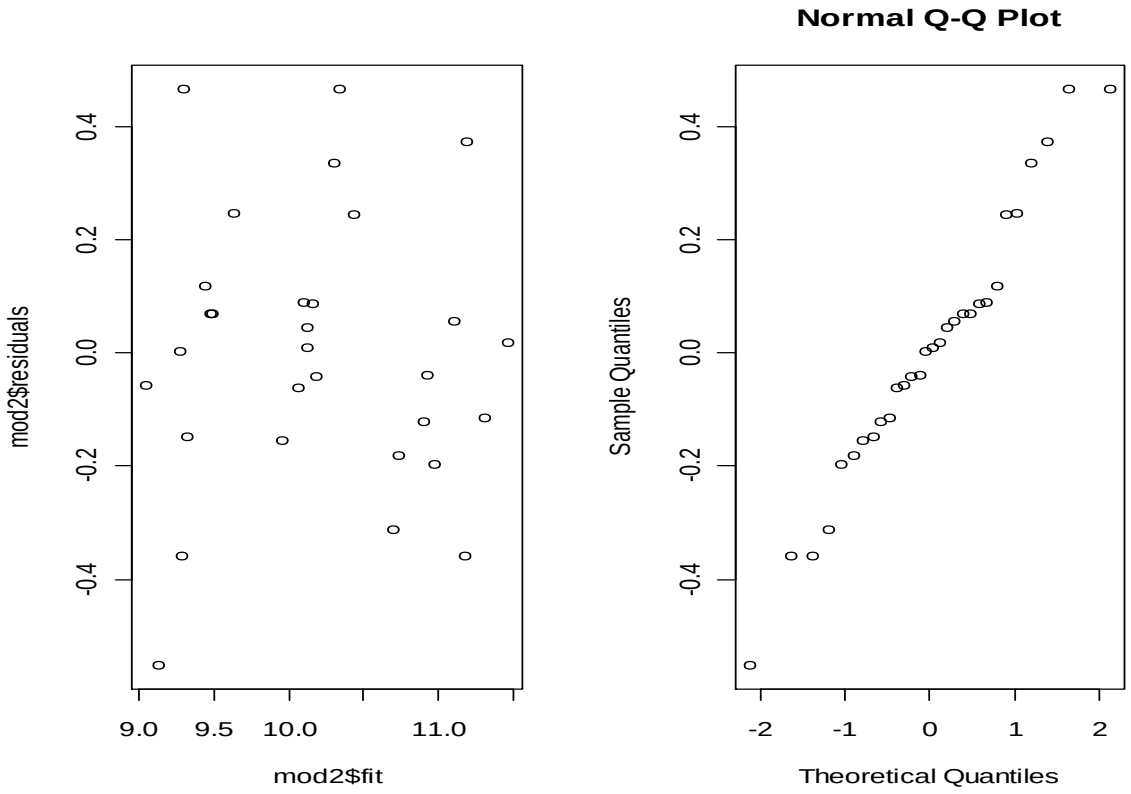


Figure 5. Plot of residuals against $\hat{y}$ for the model with the log transformed response together with a Normal Q-Q plot of the residuals.

Approximation of the expectation and variance for functions of random variables.

The technique used to find appropriate transformations can be used to find an approximation of expectation and variance for non-linear functions of random variable.

An example.

The estimated simple regression model can be written as $\hat{y} = \bar{y} + b(x - \bar{x})$. Suppose we are interested in the *x*-value for which the regression line crosses the *x*-axis. Simple calculations will suggest this value to be estimated with $x = \bar{x} - \frac{\bar{y}}{b}$. Now suppose we are interested in the uncertainty in this estimate. An estimator for the crossing value is $\bar{x} - \frac{\bar{Y}}{\hat{\beta}}$. The random variables here are $\bar{Y}$ and $\hat{\beta}$.

Both these have known variances given as $\frac{\sigma^2}{n}$ and $\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ respectively. They can also be shown to be uncorrelated random variables. However, we have no formal way to find the exact value of $Var\left(\frac{\bar{Y}}{\hat{\beta}}\right)$ by calculations.

For a function $g(x,y)$ a first order Taylor expansion gives:

$$g(x,y) \approx g(a,b) + \frac{dg}{dx}(a,b)(x - a) + \frac{dg}{dy}(a,b)(y - b).$$

Then expanding a random variable $Z = g(X,Y)$ around $\mu_1 = E(X)$ and $\mu_2 = E(Y)$ gives

$$g(X,Y) \approx g(\mu_1,\mu_2) + \frac{dg}{dx}(\mu_1,\mu_2)(X - \mu_1) + \frac{dg}{dy}(\mu_1,\mu_2)(Y - \mu_2) \qquad (8)$$

From this expression we have:

$$Var(g(X,Y)) \approx \frac{dg}{dx}(\mu_1,\mu_2)^2 Var(X) + \frac{dg}{dy}(\mu_1,\mu_2)^2 Var(Y) + 2\left(\frac{dg}{dx}(\mu_1,\mu_2)\frac{dg}{dy}(\mu_1,\mu_2)Cov(X,Y)\right)$$

If the two random variable *X* and *Y* are uncorrelated the last term vanishes.

Applied to our crossing problem we get:

$$Var\left(\frac{\bar{Y}}{\hat{\beta}}\right) \approx \frac{1}{\beta^2}\cdot\frac{\sigma^2}{n} + \frac{(\alpha + \beta\bar{x})^2}{\beta^4}\cdot\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{\beta^2}\left(\frac{1}{n} + \frac{(\alpha + \beta\bar{x})^2}{\beta^2\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

In practice estimates of α and β will be used to get an approximate value for the variance.

The general approximation formula for the variance of a random variable $g(X_1, X_2, ..., X_n)$ is:

$$Var\left(g\left(X_1, X_2, ..., X_n\right)\right) \approx \sum_{i=1}^{n} \frac{dg}{dx_i}\left(\mu_1, \ldots, \mu_n\right)^2 Var\left(X_i\right) + 2\sum_{i<j}^{n} \frac{dg\left(\mu_1, \ldots, \mu_n\right)}{dx_i} \cdot \frac{dg\left(\mu_1, \ldots, \mu_n\right)}{dx_j} Cov\left(X_i, X_j\right)$$

Using the Taylor expansion, it is also possible to find an approximation to the expected value of a function of random variables. By taking the expected value of the linearized expression in (8) we get:

$$E\left[g\left(X,Y\right)\right] \approx g\left(\mu_1, \mu_2\right)$$

and in general

$$E\left[g\left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\right)\right] \approx g\left(\mu_1, \mu_2, \ldots, \mu_n\right).$$

For the approximation formulas to work well the variances of the respective variables should be small. However, simulation shows that the approximation formula for the variance works well in most situations.

The reader should be aware of how many times these approximation formulas are needed when working with measured data. For instance, we measure length, width, height, radius, current, resistance etc. and may want to calculate, area, volume or voltage. In each case we have expressions that are not linear in the variables we measure. It is also often the case that our measurements of the different variables may be assumed to be uncorrelated and hence the covariance terms in the expression for the variance vanish.

One more example

Sometimes measurements are given with a relative uncertainty. Let us assumed that the height of a cylinder is measured with a relative uncertainty of 2% and that the radius is measured with a relative uncertainty of 1%.
As an approximation:

$$ln\left(X\right) \approx ln\left(\mu_{\mathbf{x}}\right) + \frac{1}{\mu_{\mathbf{x}}}\left(X - \mu_{\mathbf{x}}\right).$$

And hence

$\sigma_{ln(X)} \approx \frac{\sigma_X}{\mu_X}$, the coefficient of variation or relative uncertainty.

The volume of a cylinder is given by $V = \pi r^2 h$. Taking the logarithm, we get:

$$ln\left(V\right) = ln\left(\pi\right) + 2ln\left(r\right) + ln\left(h\right),$$

and a direct calculation gives $\sigma_{ln(V)} = \sqrt{4\sigma_{ln(r)}^2 + \sigma_{ln(h)}^2}$.

The relative uncertainty in V is $\frac{\sigma_V}{\mu_V} \approx \sigma_{ln(V)} = \sqrt{4(0.01)^2 + (0.02)^2} = 0.028$ or 2.8%.

Multiple linear regression. Least squares estimators and their covariance matrix.

An example. Sour precipitation.

To study the influence of sour precipitation in Norwegian lakes, SFT in 1986 carried through an investigation where data from 1005 lakes were collected. In this example 26 random lakes, 16 from Telemark and 10 from Trøndelag are chosen out. The variables and the data are given below.

y = Measured pH-value
x1 = Content of SO_4
x2 = Content of NO_3
x3 = Content of Ca
x4 = Content of latent aluminum
x5 = Content of organic material
x6 = Area of water
x7 = Location (0 – Telemark, 1-Trøndelag)

	y	x1	x2	x3	x4	x5	x6	x7
1	5.38	4.9	39	1.54	78	2.02	0.30	0
2	5.68	4.1	75	1.55	17	2.98	1.85	0
3	5.04	3.5	80	0.83	157	3.40	0.25	0
4	4.81	3.8	75	0.53	163	3.42	0.30	0
5	4.92	3.8	90	0.82	105	3.41	0.25	0
6	5.34	2.6	49	0.62	114	1.90	0.65	0
7	5.74	2.7	79	1.08	15	2.53	1.15	0
8	5.17	2.6	90	0.67	90	1.89	0.91	0
9	5.02	2.4	64	0.41	107	0.97	0.60	0
10	5.88	2.8	27	1.15	12	3.04	0.37	0
11	5.36	3.4	13	0.89	93	2.95	0.58	0
12	5.26	2.7	14	0.74	72	2.75	0.15	0
13	5.69	3.2	13	1.03	99	3.34	0.53	0
14	5.51	2.5	79	0.67	80	0.78	0.56	0
15	5.25	1.5	77	0.33	72	0.11	1.08	0
16	6.06	3.7	15	1.94	4	5.53	0.72	0
17	6.08	1.9	16	1.05	0	5.63	0.48	1
18	6.08	1.3	2	0.81	0	3.84	0.25	1
19	6.20	2.2	32	1.40	7	3.77	1.87	1
20	5.64	2.2	21	0.75	13	5.62	0.53	1

```
21 5.75 1.6  7 0.79   2 3.95 0.35  1
22 5.43 2.0 27 0.47  16 2.39 1.03  1
23 5.82 1.8 17 0.74   3 3.08 0.32  1
24 5.50 2.0  6 0.49  10 2.82 0.28  1
25 5.62 1.5  3 0.36   7 1.71 0.10  1
26 5.41 1.7  7 0.54   9 6.48 0.21  1
```

We notice that there are seven potential regression variables that may affect the response. Our modeling goal is to find a functional relationship between the response y and those regression variables that really can explain the variation in y. Let us first find out how we can obtain the least squares estimators in a multiple regression model.

The model for multiple linear regression is:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

where again Y is the response, $x_1,\ldots,x_k$ the regression variables and ε the error term. Alternatively, we have

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

Here $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1,2,\ldots,n$ and $E(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$.

We observe

$$(y_i, x_{i1}, x_{i2}, \ldots, x_{ik}), \quad i = 1,2,\ldots,n$$

Which according to the model must satisfy:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1^* \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2^* \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n^* \end{aligned}$$

where ε_i^* is a realization of ε_i. Written in matrix form this becomes:

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

Estimates for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$, $(b_0, b_1, \ldots, b_k)^t$, are obtained by minimizing

$$Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2$$

with respect to $b_0, b_1, \ldots, b_k$. Note that Q can be shown to be a convex function in $b_0, b_1, \ldots, b_k$.

The estimated model becomes:

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} .$$

Setting the partial derivatives equal to zero gives the normal equations:

$$\begin{aligned}
-2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}\right) &= -2\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0 \\
-2\sum_{i=1}^{n} x_{i1}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}\right) &= -2\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{y}_i\right) = 0 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
-2\sum_{i=1}^{n} x_{ik}\left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}\right) &= -2\sum_{i=1}^{n} x_{ik}\left(y_i - \hat{y}_i\right) = 0
\end{aligned}$$

which can be written as:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 - \hat{y} \\ y_2 - \hat{y} \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \mathbf{0} \text{ or } \mathbf{X}^t\left(\mathbf{y} - \hat{\mathbf{y}}\right) = \mathbf{X}^t\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right) = \mathbf{0}$$

or

$$\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{y} \text{ which implies } \mathbf{b} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} \text{ if } \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \text{ exists.}$$

The least squares estimator is therefore: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$. Notice that the first normal equation guarantees that the sum of the residuals is zero. If there is no constant term in the model, we have no such guarantee.

Expectation and Covariance matrix of the estimators.

The least square estimator is given by:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right)$$

which implies $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}$.

Since $E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$ we have that $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$ i.e. the least square estimator is unbiased.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$Cov\left(\hat{\boldsymbol{\beta}}\right) = E\left\{\begin{bmatrix}\hat{\beta}_0-\beta_0\\ \hat{\beta}_1-\beta_1\\ \vdots\\ \hat{\beta}_k-\beta_k\end{bmatrix}\left[\hat{\beta}_0-\beta_0,\quad \hat{\beta}_1-\beta_1,\quad \cdots\quad ,\hat{\beta}_k-\beta_k\right]\right\}$$

$$= E\begin{bmatrix}\left(\hat{\beta}_0-\beta_0\right)^2 & \left(\hat{\beta}_0-\beta_0\right)\left(\hat{\beta}_1-\beta_1\right) & \cdots & \left(\hat{\beta}_0-\beta_0\right)\left(\hat{\beta}_k-\beta_k\right)\\ \left(\hat{\beta}_1-\beta_1\right)\left(\hat{\beta}_0-\beta_0\right) & \left(\hat{\beta}_1-\beta_1\right)^2 & \cdots & \left(\hat{\beta}_1-\beta_1\right)\left(\hat{\beta}_k-\beta_k\right)\\ \vdots & \vdots & \cdots & \vdots\\ \left(\hat{\beta}_k-\beta_k\right)\left(\hat{\beta}_0-\beta_0\right) & \left(\hat{\beta}_k-\beta_k\right)\left(\hat{\beta}_1-\beta_1\right) & \cdots & \left(\hat{\beta}_k-\beta_k\right)^2\end{bmatrix}$$

$$= \begin{bmatrix}Var\left(\hat{\beta}_0\right) & Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) & \cdots & Cov\left(\hat{\beta}_0,\hat{\beta}_k\right)\\ Cov\left(\hat{\beta}_1,\hat{\beta}_0\right) & Var\left(\hat{\beta}_1\right) & \cdots & Cov\left(\hat{\beta}_1,\hat{\beta}_k\right)\\ \vdots & \vdots & \cdots & \vdots\\ Cov\left(\hat{\beta}_k,\hat{\beta}_0\right) & Cov\left(\hat{\beta}_k,\hat{\beta}_1\right) & \cdots & Var\left(\hat{\beta}_k\right)\end{bmatrix}.$$

We get :

$$E\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t\right] = E\left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right]$$
$$= \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\,E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\right)\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}.$$

An example. Simple linear regression.

The model can be written as:

$$\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{bmatrix} = \begin{bmatrix}1 & x_{11}\\ 1 & x_{21}\\ \vdots & \vdots\\ 1 & x_{n1}\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}$$

We get:

$$\left(\mathbf{X}^t\mathbf{X}\right) = \begin{bmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{21} & \cdots & x_{n1}\end{bmatrix}\begin{bmatrix}1 & x_{11}\\ 1 & x_{21}\\ \vdots & \vdots\\ 1 & x_{n1}\end{bmatrix} = \begin{bmatrix}n & \sum_{i=1}^{n} x_{i1}\\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2\end{bmatrix}$$

$$\mathbf{X}^{\mathrm{t}}\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_{i1} Y_i \end{bmatrix}$$

and $\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1} = \dfrac{1}{n\sum_{i=1}^{n} x_{i1}^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & -\sum_{i=1}^{n} x_{i1} \\ -\sum_{i=1}^{n} x_{i1} & n \end{bmatrix}$

such that

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2}\left(-\bar{x}_1 \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} x_{i1} Y_i\right) = \frac{\sum_{i=1}^{n} Y_i \left(x_{i1} - \bar{x}_1\right)}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2}$$

and

$$\hat{\alpha} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2}\left(\bar{Y}\left(\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2\right) + \bar{x}_1\left(n\bar{Y}\bar{x}_1 - \sum_{i=1}^{n} x_{i1} Y_i\right)\right) = \bar{Y} - \hat{\beta}\,\bar{x}_1 .$$

$$Cov\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \sigma^2 \begin{bmatrix} \dfrac{\sum_{i=1}^{n} x_{i1}^2}{n\left(\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2\right)} & -\dfrac{\bar{x}_1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2} \\ -\dfrac{\bar{x}_1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2} & \dfrac{1}{\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2} \end{bmatrix}.$$

$\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2$ can be written as $\sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2$. This follows because

$$\sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2 = \sum_{i=1}^{n} x_{i1}^2 - 2\bar{x}_1 \sum_{i=1}^{n} x_{i1} + n\bar{x}_1^2 = \sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2 .$$

Partitioning of variation

A measure of the total variance in the response is given by $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$. We can split up this as follows:

First write

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

such that

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

Now

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \sum_{i=1}^{n}(y_i - \hat{y}_i)\bar{y} = 0 .$$

This follows because the two last expressions are zero, since the normal equations have to be fulfilled.

Thereby we have

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

This can also be written as:

$$SS_T = SS_E + SS_R .$$

Or the total sum of squares equals the error sum of squares + the regression sum of squares.

This partitioning of the total variation in the response is only true if the normal equations are fulfilled. In models with no constant term this partitioning is normally not valid.

Some results about idempotent matrices and distribution of quadratic forms and their application to multiple linear regression.

Most of the necessary theory behind procedures in regression analysis and in linear models in general relies upon properties of idempotent matrices and the spectral decomposition of symmetric matrices. In fact, in this course you will find these two mathematical topics involved in most of the theory we are building up.

Idempotent matrices

First define $\boldsymbol{J}_{n\times n}$ to be a $n \times n$ matrix with all entries equal to 1. Let $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ be a vector of random variables. We have:

$$\bar{\boldsymbol{Y}} = \begin{bmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \frac{1}{n} \boldsymbol{J}\boldsymbol{Y} .$$

Now $\left(\boldsymbol{I}_{n\times n} - \frac{1}{n}\boldsymbol{J} \right)$ is obviously symmetric. Therefore $\boldsymbol{Y} - \bar{\boldsymbol{Y}} = \left(\boldsymbol{I}_{n\times n} - \frac{1}{n}\boldsymbol{J} \right)\boldsymbol{Y}$ and

$$\left(\boldsymbol{y} - \bar{\boldsymbol{y}}\right)^t \left(\boldsymbol{y} - \bar{\boldsymbol{y}}\right) = \boldsymbol{y}^t \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \right)\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \right)\boldsymbol{y} = SS_T .$$

Further $\hat{\boldsymbol{Y}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^t\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t$ is the often referred to hat matrix in linear regression theory. $\left(\boldsymbol{I}_{n\times n} - \boldsymbol{H}_{n\times n}\right)$ is symmetric and we obtain

$$\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \left(\boldsymbol{I}_{n\times n} - \boldsymbol{H}_{n\times n}\right)\boldsymbol{Y} \text{ and } \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)^t \left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right) = \boldsymbol{y}^t \left(\boldsymbol{I} - \boldsymbol{H}\right)\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} = SS_E$$

Also $\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} \right)$ is symmetric and $\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}} = \left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} \right)\boldsymbol{Y}$. Hence

$$\left(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\right)^t \left(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\right) = \boldsymbol{y}^t \left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} \right)\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} \right)\boldsymbol{y} = SS_R .$$

For the matrices $\boldsymbol{J}_{n\times n}$ and $\boldsymbol{H}_{n\times n}$ we have:

$$\frac{1}{n}\boldsymbol{J}\frac{1}{n}\boldsymbol{J} = \frac{1}{n}\boldsymbol{J} \text{ and } \boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\,\boldsymbol{X}\left(\boldsymbol{X}^t\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t = \boldsymbol{X}\left(\boldsymbol{X}^t\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t = \boldsymbol{H}$$

Matrices that remain unchanged regardless of how many times we multiply them together are called idempotent.

We further get

$$\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \right)\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \right) = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J} - \frac{1}{n}\boldsymbol{J} + \frac{1}{n}\boldsymbol{J}\frac{1}{n}\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J} - \frac{1}{n}\boldsymbol{J} + \frac{1}{n}\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \text{ and}$$

$$\left(\boldsymbol{I} - \boldsymbol{H}\right)\left(\boldsymbol{I} - \boldsymbol{H}\right) = \boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H} .$$

The following property of the hat matrix is useful:

R1. $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^t\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\,\boldsymbol{X} = \boldsymbol{X}$.

This means that

$$\boldsymbol{H}\begin{bmatrix} 1 & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_k \\ 1 & & & \\ \vdots & & & \\ 1 & & & \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_k \\ 1 & & & \\ \vdots & & & \\ 1 & & & \end{bmatrix}$$ where $\boldsymbol{x}_i$, $i = 1, 2, \ldots, k$, is the column of observed values for the i'th regression variable. In particular we get $\boldsymbol{H}\boldsymbol{1} = \boldsymbol{1}$ or $\sum_{j=1}^{n} h_{ij} = \sum_{i=1}^{n} h_{ij} = 1$ since the hat matrix is symmetric. In other words, the elements in any row and columns in the hat matrix sum to 1.

Then we obtain

$$\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}\right)\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}\right) = \boldsymbol{H}^2 - \frac{1}{n}\boldsymbol{J}\,\boldsymbol{H} - \frac{1}{n}\boldsymbol{H}\,\boldsymbol{J} + \frac{1}{n}\boldsymbol{J}\,\frac{1}{n}\boldsymbol{J} = \boldsymbol{H} - \frac{1}{n}\boldsymbol{J} - \frac{1}{n}\boldsymbol{J} + \frac{1}{n}\boldsymbol{J} = \boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$$ since $\boldsymbol{J}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{J} = \boldsymbol{J}$.

A matrix $\boldsymbol{A}$ that satisfies $\boldsymbol{A}\,\boldsymbol{A} = \boldsymbol{A}$ is called idempotent. If $\boldsymbol{A}$ also is symmetric it is also called a projection matrix.

R2. $\boldsymbol{H}$, $\frac{1}{n}\boldsymbol{J}$, $\boldsymbol{I} - \boldsymbol{H}$, $\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}$ and $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$ *are all idempotent matrices and also projection matrices.*

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices such that their product $\boldsymbol{AB}$ is defined. Then

$\text{rank}(\boldsymbol{AB}) \le \min\left(\text{rank}\left(\boldsymbol{A}\right), \text{rank}\left(\boldsymbol{B}\right)\right)$.

This follows since the columns of $\boldsymbol{AB}$ are linear combinations of the columns of $\boldsymbol{A}$ and the rows in $\boldsymbol{AB}$ are linear combinations of the rows of $\boldsymbol{B}$.

For a symmetric matrix $\boldsymbol{A}_{n\times n}$ we have $\boldsymbol{A} = \boldsymbol{P}\,\boldsymbol{\Lambda}\,\boldsymbol{P}^t$ where $\boldsymbol{P} = \left[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\right]$, the matrix of eigenvectors and $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_1 \end{bmatrix}$ is a diagonal matrix of eigenvalues. This is known as the spectral decomposition of the matrix $\boldsymbol{A}$.

R3. $\boldsymbol{A}_{n\times n}$ *symmetric implies* $\text{rank}\left(\boldsymbol{A}\right) =$ *the number of nonzero eigenvalues*

Proof.

We have

$\text{rank}(\boldsymbol{A})=\text{rank}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t) \le min(\text{rank}(\boldsymbol{P}^t), \text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t)) \le \text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t)$

and $\text{rank}(\boldsymbol{\Delta}\boldsymbol{P}^t)=\text{rank}(\boldsymbol{P}^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t) \le min(\text{rank}(\boldsymbol{P}^t), \text{rank}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t)) \le \text{rank}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t) =$ $\text{rank}(\boldsymbol{A})$. Also $\text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t) \le min(\text{rank}(\boldsymbol{\Lambda}), \text{rank}(\boldsymbol{P}^t)) \le \text{rank}(\boldsymbol{\Lambda}) = \text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{P}) \le \text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t)$.

Hence $\text{rank}(\boldsymbol{A})=\text{rank}(\boldsymbol{\Lambda}\boldsymbol{P}^t) = \text{rank}(\boldsymbol{\Lambda})$.

R4. $\boldsymbol{A}_{n\times n}$ *symmetric and idempotent with rank r implies r eigenvalues are* 1 *and n-r are zero.*

Proof.

$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ and $\lambda\boldsymbol{x}^t\boldsymbol{x} = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \lambda^2\boldsymbol{x}^t\boldsymbol{x}$. Hence $\boldsymbol{x}^t\boldsymbol{x}\lambda(\lambda-1)=0$ which implies $\lambda=1$ or $\lambda=0$. Since the number of nonzero eigenvalues are r, r eigenvalues are 1 and n-r are zero.

R5. $\boldsymbol{A}_{n\times n}$ *symmetric and idempotent implies* $tr(\boldsymbol{A})$ = *rank*($\boldsymbol{A}$).

Proof .

$\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t) = \text{tr}(\boldsymbol{P}^t\boldsymbol{P}\boldsymbol{\Lambda}) = \text{tr}(\boldsymbol{\Lambda}) = \text{rank}(\boldsymbol{A})$.

Distribution of Quadratic forms

R6. *Let* $\boldsymbol{A}_{n\times n}$ *be a symmetric and idempotent matrix of rank r and let* $Y_1, Y_2, \ldots, Y_n$ *be independent normally distributed with expectation* $\mu_1, \mu_2, \ldots, \mu_n$ *respectively and equal variance* σ^2. *Define* $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ *and* $\underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$. *Then* $\dfrac{(\boldsymbol{Y}-\underline{\mu})^t(\boldsymbol{A})(\boldsymbol{Y}-\underline{\mu})}{\sigma^2}$ *is* $\chi^2(r)$.

Proof . $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t$. Let $\boldsymbol{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix}$ where $U_i = \dfrac{Y_i-\mu_i}{\sigma}$, i=1,2,…,n. $U_1, U_2, \ldots, U_n$ are independent random variables. Also $E(U_i)=0$ and $SD(U_i)=1$, i=1,2,…,n. We get:

$$\frac{(\boldsymbol{Y}-\underline{\mu})^t\boldsymbol{A}(\boldsymbol{Y}-\underline{\mu})}{\sigma^2} = \boldsymbol{U}^t\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{U} = \boldsymbol{Z}^t\boldsymbol{\Lambda}\boldsymbol{Z} = \sum_{i=1}^{r} Z_i^2 \text{ where } \boldsymbol{Z} = \boldsymbol{P}^t\boldsymbol{U}.$$

Further $Cov(\boldsymbol{Z}) = Cov(\boldsymbol{P}^t\boldsymbol{U}) = E(\boldsymbol{P}^t\,\boldsymbol{U}\boldsymbol{U}^t\,\boldsymbol{P}) = \boldsymbol{P}^t\,E(\boldsymbol{U}\boldsymbol{U}^t)\boldsymbol{P} = \boldsymbol{P}^t\,\boldsymbol{I}\boldsymbol{P} = \boldsymbol{I}$.

Obviously $E(\boldsymbol{P}^t\,\boldsymbol{U}) = \boldsymbol{P}^t\,E(\boldsymbol{U}) = \boldsymbol{0}$. Therefore, $Z_i \sim N(0,1), i = 1,2,\ldots,n$ and independent which implies that $\sum_{i=1}^{r} Z_i^2$ is $\chi^2(r)$.

R7. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be symmetric and idempotent matrices such that* $\boldsymbol{AB} = \boldsymbol{0}$. *Assume further that* $Y_1, Y_2, \ldots, Y_n$ *are independent normally distributed with expectation* $\mu_1, \mu_2, \ldots, \mu_n$ *respectively and equal variance* σ^2. *Let* $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ *and* $\underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$. *Then* $(\boldsymbol{Y} - \underline{\mu})^t\,\boldsymbol{A}(\boldsymbol{Y} - \underline{\mu})$ *and* $(\boldsymbol{Y} - \underline{\mu})^t\,\boldsymbol{B}(\boldsymbol{Y} - \underline{\mu})$ *are independent.*

Proof.

Let $\boldsymbol{U} = \boldsymbol{Y} - \underline{\mu}$. Then $E(\boldsymbol{U}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{U}) = \sigma^2\,\boldsymbol{I}$. Further $\boldsymbol{U}^t\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^t\boldsymbol{A}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{Z}^t\boldsymbol{Z}$ where $\boldsymbol{Z} = \boldsymbol{A}^t\boldsymbol{U}$, and $\boldsymbol{U}^t\,\boldsymbol{B}\boldsymbol{U} = \boldsymbol{U}^t\,\boldsymbol{B}\boldsymbol{B}\boldsymbol{U} = \boldsymbol{V}^t\boldsymbol{V}$ where $\boldsymbol{V} = \boldsymbol{B}^t\,\boldsymbol{U}$. $\boldsymbol{Z}$ and $\boldsymbol{V}$ are independent if $E(\boldsymbol{Z}\boldsymbol{V}^t) = \boldsymbol{0}$.

$E(\boldsymbol{Z}\boldsymbol{V}^t) = E(\boldsymbol{A}^t\,\boldsymbol{U}\boldsymbol{U}^t\,\boldsymbol{B}) = \boldsymbol{A}\,E(\boldsymbol{U}\boldsymbol{U}^t)\boldsymbol{B} = \boldsymbol{A}\sigma^2\,\boldsymbol{I}\,\boldsymbol{B} = \sigma^2\,\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$.

Some specific results for multippel linear regression models.

Theorem 12.1

In the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ *with* $E(\varepsilon_i\varepsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, \text{elles} \end{cases}$ *an estimator for* σ^2 *is given by* $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n - (k+1)}$ *where* k *is the number of regression variables.*

Proof.

$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^t(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}}$.

$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\,\boldsymbol{X}^t\,\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\beta})$

since $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\,\boldsymbol{\beta} = \boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{X}\,\boldsymbol{\beta} = \boldsymbol{0}$.

Hence $\hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^t(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$,

and $E\left(\hat{\boldsymbol{\varepsilon}}^t\hat{\boldsymbol{\varepsilon}}\right) = E\left(\boldsymbol{\varepsilon}^t(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}\right) = E\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i(\boldsymbol{I}-\boldsymbol{H})_{ij}\boldsymbol{\varepsilon}_j\right) = \sum_{i=1}^{n}(\boldsymbol{I}-\boldsymbol{H})_{ii}E\left(\boldsymbol{\varepsilon}_i^2\right)$

$$= \sigma^2\sum_{i=1}^{n}(\mathrm{I}-\mathrm{H})_{ii} = \sigma^2 tr(\mathrm{I}-\mathrm{H}) = \sigma^2\left(tr(\mathrm{I})-tr(\mathrm{H})\right) = \sigma^2 n - \sigma^2 tr\left(\mathrm{X}\left(\mathrm{X}^t\mathrm{X}\right)^{-1}\mathrm{X}^t\right)$$
$$= \sigma^2 n - \sigma^2 tr\left(X^t\,\mathrm{X}\left(\mathrm{X}^t\mathrm{X}\right)^{-1}\right) = \sigma^2\left(n - tr\left(\mathrm{I}_{(k+1)\times(k+1)}\right)\right) = (n-k-1)\sigma^2.$$

R8. *In the linear regression model* $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} = \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix}$, *with* $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\sigma}^2\boldsymbol{I})$,

$\frac{SS_E}{\sigma^2} = \frac{\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}\right)^t\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}\right)}{\sigma^2}$ *is* $\chi^2(n\text{-}k\text{-}1)$.

Proof

$(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Hence
$\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{I}-\boldsymbol{H})^t(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$

Rank $(\boldsymbol{I}-\boldsymbol{H}) = tr(\boldsymbol{I}-\boldsymbol{H}) = n-(k+1)$.

Hence $\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2}$ is $\chi^2\left(n-(k+1)\right)$.

Now assume $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ i.e. $Y_i = \beta_o + \varepsilon_i$, $i = 1,2,\ldots,n$. Let $\mathbf{1}_{n\times 1} = [1,1,\ldots,1]^t$

Then $\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\bar{\boldsymbol{Y}}+\frac{1}{n}\boldsymbol{J}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Y}-\bar{\boldsymbol{Y}}-\mathbf{1}\beta_0+\mathbf{1}\beta_0 = \boldsymbol{Y}-\bar{\boldsymbol{Y}}$.

and $\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{Y}}-\boldsymbol{X}\boldsymbol{\beta}+\frac{1}{n}\boldsymbol{J}\boldsymbol{X}\boldsymbol{\beta} = \hat{\boldsymbol{Y}}-\bar{\boldsymbol{Y}}$.

Therefore $\frac{SS_T}{\boldsymbol{\sigma}^2} = \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}{\sigma^2} = \frac{\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\right)^t\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\right)}{\sigma^2} = \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^t\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2}$ is $\chi^2(n-1)$

since $\mathrm{tr}\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right)=n-\mathrm{tr}\left(\frac{1}{\mathrm{n}}\boldsymbol{J}\right)=n-1$ and

$$\frac{SS_R}{\boldsymbol{\sigma}^2}=\left(\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\sigma^2}\right)=\frac{\left(\hat{\boldsymbol{Y}}-\bar{\boldsymbol{Y}}\right)^t\left(\hat{\boldsymbol{Y}}-\bar{\boldsymbol{Y}}\right)}{\sigma^2}=\frac{\left(\boldsymbol{Y}-\boldsymbol{X}\,\beta\right)^t\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)\left(\boldsymbol{Y}-\boldsymbol{X}\,\beta\right)}{\sigma^2}\ \text{is}\ \chi^2(k)$$

since $\mathrm{rank}\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}\right)=\mathrm{tr}(\boldsymbol{H})-\mathrm{tr}\left(\frac{1}{\mathrm{n}}\boldsymbol{J}\right)=k+1-1=k$.

We also have $(\boldsymbol{I}-\boldsymbol{H})\left(\boldsymbol{H}-\frac{1}{\mathrm{n}}\boldsymbol{J}\right)=\boldsymbol{0}$ from which it follows that SS_E and SS_R are independent.

R9. *Assume* $\varepsilon \sim N\left(0,\sigma^2\boldsymbol{I}\right)$. *Then* $\hat{\boldsymbol{\beta}}$ *is independent of* $\hat{\sigma}^2$.

Proof. $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}},\boldsymbol{Y}-\boldsymbol{X}\,\boldsymbol{\beta}\right)=E\left(\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right),(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{Y})^t\right)=E\left(\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right),\boldsymbol{\varepsilon}^t\right)$

$=E\left(\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\left(\boldsymbol{Y}-\boldsymbol{X}\,\boldsymbol{\beta}\right)\boldsymbol{\varepsilon}^t\left(\boldsymbol{I}-\boldsymbol{H}\right)\right)=E\left(\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\,\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\left(\boldsymbol{I}-\boldsymbol{H}\right)\right)$

$=\boldsymbol{\sigma}^2\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{I}\left(\boldsymbol{I}-\boldsymbol{H}\right)=\boldsymbol{\sigma}^2\left(\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\right)-\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t)=\boldsymbol{0}$.

The Fisher (F) distribution and the analysis of variance table for linear regression.

R1**0**. Assume the random variable X is chi-squared distributed with ν_1 degrees of freedom, writes $\chi^2(\nu_1)$, and $Y \sim \chi^2(\nu_2)$ and that X and Y are independent. Then $\frac{X/\nu_1}{Y/\nu_2}$ is Fisher distributed with ν_1 and ν_1 degrees of freedom, writes F_{ν_1,ν_2}.

Note

$$F \sim F_{\nu_1,\nu_2} \Rightarrow \frac{1}{F} \sim F_{\nu_2,\nu_1}$$

Now define α by

$$P\left(F_{\nu_1,\nu_2} \leq f_{\alpha,\nu_1,\nu_2}\right) = \alpha$$

and since

$$P\left(f_{1-\frac{\alpha}{2},\nu_1,\nu_2} \leq F \leq f_{\frac{\alpha}{2},\nu_1,\nu_2} \right) = 1-\alpha \Leftrightarrow P\left(\frac{1}{f_{\frac{\alpha}{2},\nu_1,\nu_2}} \leq \frac{1}{F} \leq \frac{1}{f_{1-\frac{\alpha}{2},\nu_1,\nu_2}} \right) = 1-\alpha$$

we have $\frac{1}{f_{\frac{\alpha}{2},\nu_1,\nu_2}} = f_{1-\frac{\alpha}{2},\nu_2,\nu_1}$ and $f_{1-\frac{\alpha}{2},\nu_1,\nu_2} = \frac{1}{f_{\frac{\alpha}{2},\nu_2,\nu_1}}$

The density function in the F distribution is given by:

$$f(x) = k_{\nu_1,\nu_2} \cdot x^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, x > 0$$

where $k_{\nu_1,\nu_2} = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\cdot\Gamma\left(\frac{\nu_2}{2}\right)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}$

$E(F) = \frac{\nu_1}{\nu_2-2}$ and $Var\left(F_{\nu_1,\nu_2}\right) = \frac{2\nu_2^2\left(\nu_1+\nu_2-2\right)}{\nu_1\left(\nu_2-2\right)^2\left(\nu_2-4\right)}$ provided $\nu_2 > 4$.

Some examples of how the density function in the Fisher distribution looks like is given in Figure 6.

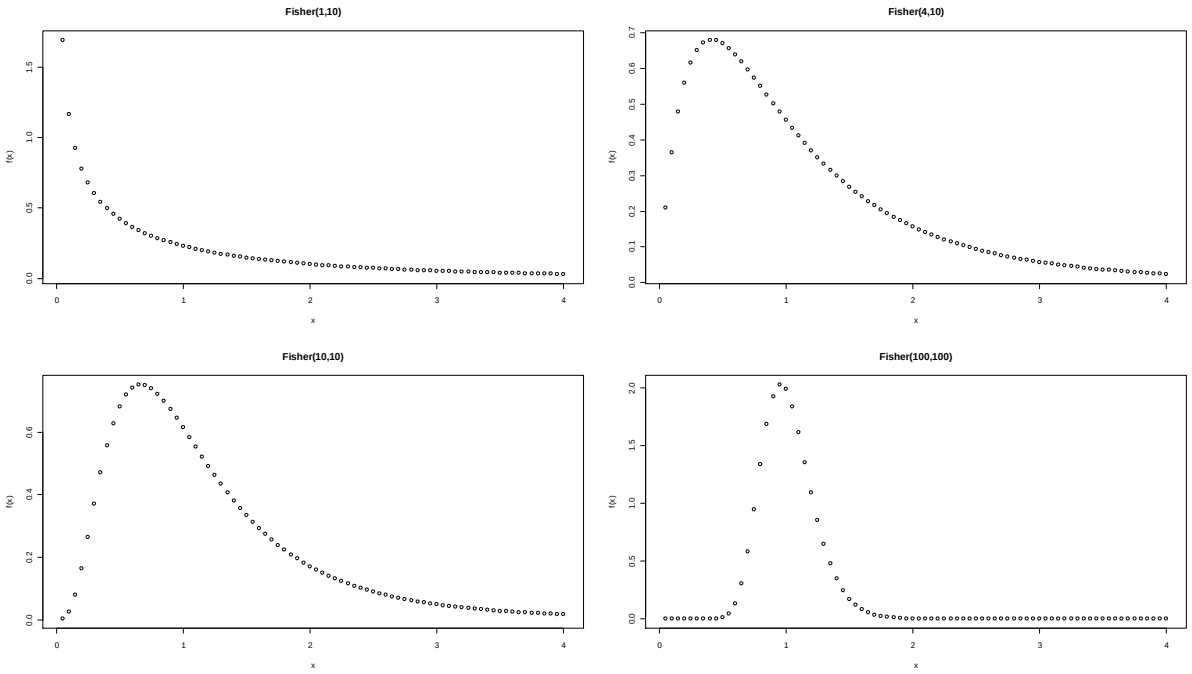


Figure 6.

Inference in multippel linear regression

A test on whether the regression is significant (or if the regression equation differs from a constant) is given by:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ against H_1: at least one β_i, i=1,2,…,k, is different from 0.

With $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, we know that $\frac{SS_E}{\sigma^2}$ *is* $\chi^2(n-k-1)$. Under H_0 we also have that

$\frac{\mathrm{SS_R}}{\sigma^2}$ is $\chi^2(k)$ and since they are independent, we get $\frac{\mathrm{SS_R}}{k} \Big/ \frac{\mathrm{SS_E}}{n-k-1} \sim F_{k,n-k-1}$. Since the denominator is independent of H_0 and $\mathrm{SS_R}$ under H_1 is greater than under H_0 a natural test for H_0 can be built on $F_{k,n-k-1}$.

We reject H_0 if $f \geq f_{\alpha,k,n-(k+1)}$ where f is the observed value of $F_{k,n-k-1}$.

It is normal to collect the sum of squares values from a regression analysis in a variance of analysis table as follows:

Table. The analysis of variance table for regression analysis

Source	Sum of squares	Df	Mean sum of squares	F
Regression	$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	$MSS_R = \frac{SS_R}{k}$	$f = \frac{MSS_R}{MSS_E}$
Error	$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-(k+1)$	$MSS_E = \frac{SS_E}{n-(k+1)}$	
Total	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$		

The partial F-test.

The partial F-test is useful if we want to test the significance of a group of regression variables. Assume we want to test the following hypothesis.

$H_0: \beta_{r+1}, \beta_{r+2}, \ldots, \beta_k = 0$ H_1: at least one of $\beta_{r+1}, \ldots, \beta_k$ are $\neq 0$.

Let $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r, \ldots, \boldsymbol{x}_k]$ and $\boldsymbol{X}_1 = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r]$ and define $\boldsymbol{H}$ and $\boldsymbol{H}_1$ by

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t$ and $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t$. We get for all $\boldsymbol{y}$:

$\boldsymbol{H}\boldsymbol{H}_1 \boldsymbol{y} = \boldsymbol{H}\boldsymbol{X}_1 \boldsymbol{b}_1 = \boldsymbol{H}\boldsymbol{X} \begin{bmatrix} \boldsymbol{b}_1 \\ 0 \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \boldsymbol{b}_1 \\ 0 \end{bmatrix} = \boldsymbol{X}_1 \boldsymbol{b}_1 = \boldsymbol{H}_1 \boldsymbol{y}$. Thereby $\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1$ and by transposing it we get $\boldsymbol{H}_1 \boldsymbol{H} = \boldsymbol{H}_1$.Therefore $(\boldsymbol{H} - \boldsymbol{H}_1)^2 = \boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}_1 - \boldsymbol{H}_1 \boldsymbol{H} + \boldsymbol{H}_1^2 = \boldsymbol{H} - \boldsymbol{H}_1$ i.e. $(\boldsymbol{H}\text{-}\boldsymbol{H}_1)$ is idempotent.

Let $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$ and $\boldsymbol{\beta}_1 = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_r \end{bmatrix}$.

$SS_r(\boldsymbol{\beta}) - SS_r(\boldsymbol{\beta}_1) = SS_T - \boldsymbol{Y}^t(\boldsymbol{I}\text{-}\boldsymbol{H})\boldsymbol{Y} - SS_T + \boldsymbol{Y}^t(\boldsymbol{I}\text{-}\boldsymbol{H}_1)\boldsymbol{Y} = \boldsymbol{Y}^t(\boldsymbol{H}\text{-}\boldsymbol{H}_1)\boldsymbol{Y} = (\boldsymbol{Y}\text{-}\boldsymbol{X}_1\boldsymbol{\beta}_1)^t(\boldsymbol{H}\text{-}\boldsymbol{H}_1)(\boldsymbol{Y}\text{-}\boldsymbol{X}_1\boldsymbol{\beta}_1)$ since $(\boldsymbol{H}\text{-}\boldsymbol{H}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_1\boldsymbol{\beta}_1$.

$\text{Rank}(\boldsymbol{H} - \boldsymbol{H}_1) = \text{tr}(\boldsymbol{H}) - \text{tr}(\boldsymbol{H}_1) = k + 1 - (r+1) = k - r$. Therefore under $\boldsymbol{H}_0$

$\frac{SS_r(\boldsymbol{\beta}) - SS_r(\boldsymbol{\beta}_1)}{\sigma^2} = \frac{(\boldsymbol{Y}\text{-}\boldsymbol{X}_1\boldsymbol{\beta}_1)^t(\boldsymbol{H}\text{-}\boldsymbol{H}_1)(\boldsymbol{Y}\text{-}\boldsymbol{X}_1\boldsymbol{\beta}_1)}{\sigma^2}$ is $\chi^2(k-r)$.

Also $\left(\boldsymbol{H}-\boldsymbol{H}_1\right)\left(\boldsymbol{I}-\boldsymbol{H}\right)=\boldsymbol{H}-\boldsymbol{H}_1-\boldsymbol{H}+\boldsymbol{H}_1=\boldsymbol{0}$ which implies

$\boldsymbol{Y}^t\left(\boldsymbol{H}-\boldsymbol{H}_1\right)\boldsymbol{Y}$ and $\hat{\boldsymbol{\sigma}}^2$are independent.

The test statistic for H_0 is $F=\dfrac{\dfrac{SS_r\left(\boldsymbol{\beta}\right)-SS_r\left(\boldsymbol{\beta}_1\right)}{k-r}}{\dfrac{SS_E\left(\boldsymbol{\beta}\right)}{n-\left(k+1\right)}}\sim F_{(k-r),(n-k-1)}$,

and we reject H_0 if $f_{obs}\geq f_{\boldsymbol{\alpha},(k-r)(n-k-1)}$. Note that a partial F-test can be used to test on single variables given that the rest of the variables are in the model, on groups of variables given that the rest of the variables are in the model and if the regression is significant.

Testing a general linear hypothesis

Suppose $E\left[Y\right]=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$ and we want to test the hypothesis

$H_0:\begin{bmatrix}\beta_1-\beta_2\\ \beta_2-\beta_3\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$ or equivalently $H_0:\beta_1=\beta_2=\beta_3=\beta$.

The hypothesis can be formulated as: $H_0:\boldsymbol{C}\boldsymbol{\beta}=\boldsymbol{0}$ vs $H_1:\boldsymbol{C}\boldsymbol{\beta}\neq\boldsymbol{0}$, where

$\boldsymbol{C}=\begin{bmatrix}0&1&-1&0\\0&0&1&-1\end{bmatrix}$ and $\boldsymbol{\beta}=\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix}$. We have thus r (in this case 2) constraints on $\boldsymbol{\beta}$.

With the assumption of normally distributed errors,

$$\hat{\boldsymbol{\beta}}\sim\boldsymbol{N}\left(\boldsymbol{\beta},\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)\text{ and }\boldsymbol{C}\hat{\boldsymbol{\beta}}\sim\boldsymbol{N}\left(\boldsymbol{C}\boldsymbol{\beta},\sigma^2\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right).$$

Hence, under H_0, $\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)^T\left(\sigma^2\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)\sim\chi^2\left(r\right)$.

Since $\hat{\beta}$ and SS_E (full model) are independent and $\dfrac{SS_E}{\sigma^2}\sim\chi^2\left(n-k-1\right)$, we get

$$F=\frac{\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)\Big/\sigma^2r}{SS_E\Big/\sigma^2\left(n-k-1\right)}=\frac{\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}\right)\Big/r}{SS_E\Big/\left(n-k-1\right)}\sim F_{r,n-k-1}.$$

We reject H_0 if F is large.

If we want to test $H_0:\boldsymbol{C}\boldsymbol{\beta}=\boldsymbol{d}$ vs H_1: $\boldsymbol{C}\boldsymbol{\beta}\neq\boldsymbol{d}$, we just substitute $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ with $\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{d}$ in F .

Test on single coefficients, intervals and regions.

We will now be mainly concerned with how regression analysis can be performed in practice. Assume that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is a sequence of independent, normally distributed random variables with expectation equal to zero and constant variance σ^2.

In multiple linear regression we have:

$$E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k .$$

Thus β_i, $i = 1,2,\ldots,k$ is the change in $E(Y)$ if x_i is changed with one unit and all the rest of the variables are kept unchanged.

We want a test for: Does a variable have a significant impact on the response given that the other variables are in the model.

Such a test for x_j, $j = 1,2,\ldots,k$ is:

$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0$$

We know that $\hat{\boldsymbol{\beta}}$ is independent of $S^2 = \dfrac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-(k+1)}$.

Now let c_{jj} be the $(j+1)th$ diagonal element in $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$. Then

$T = \dfrac{\beta_j}{S\sqrt{c_{jj}}} = \dfrac{\dfrac{\hat{\beta}_j}{\sigma\sqrt{c_{jj}}}}{\dfrac{S}{\sigma}}$ and hence the enumerator is $N(0,1)$ and since $\dfrac{S^2(n-k-1)}{\sigma^2}$ is

$\chi^2(n-k-1)$ the denominator can be written as a variable that is: $\sqrt{\dfrac{\chi^2(n-k-1)}{n-k-1}}$. Therefore T is t-distributed with $(n-k-1)$ degrees of freedom and we reject H_0 if $|t_{obs}| \geq t_{\frac{\alpha}{2},n-(k+1)}$. For the test

$H_0 : \beta_j = \beta_{j0}$ against $H_1 : \beta_j \begin{matrix} < \\ = \\ > \end{matrix} \beta_{j0}$, we use $T = \dfrac{\hat{\beta}_j - \beta_{j0}}{S\sqrt{c_{jj}}}$ the same way except that $\dfrac{\alpha}{2}$ is

substituted by α if the test is one-sided.

Individual confidence intervals for each β_j, $j = 1,2,\ldots,k$, can be deduced using T.

We get $P\left(\hat{\beta}_j - t_{\frac{\alpha}{2},n-k-1} S\sqrt{c_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\frac{\alpha}{2},n-k-1} S\sqrt{c_{jj}}\right) = 1-\alpha$.

The confidence ellipsoid

A confidence ellipsoid may be useful if several parameters are to be studied jointly.

Exploiting that $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0},\sigma^2\mathbf{I}\right) \Rightarrow \hat{\boldsymbol{\beta}} \sim \boldsymbol{N}\left(\boldsymbol{\beta},\sigma^2\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\right)$ we get

$\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t \frac{\boldsymbol{X}^t\boldsymbol{X}}{\sigma^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \sim \chi^2\left(k+1\right)$. We also know that $\frac{SS_E}{\sigma^2} \sim \chi^2\left(n-k-1\right)$ and $\frac{SS_E}{n-k-1} = S^2$. Therefore, $\frac{S^2\left(n-k-1\right)}{\sigma^2} \sim \chi^2\left(n-k-1\right)$. Thereby

$$\frac{\frac{\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t \boldsymbol{X}^t\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}{\sigma^2\left(k+1\right)}}{\frac{\left(n-k-1\right)S^2}{\sigma^2\left(n-k-1\right)}}$$ and a $100\left(1-\alpha\right)\%$ confidence region for $\boldsymbol{\beta}$ is given by those $\boldsymbol{\beta}$ that satisfy $\left(\boldsymbol{b}-\boldsymbol{\beta}\right)^t \frac{\boldsymbol{X}^t\boldsymbol{X}}{\sigma^2}\left(\boldsymbol{b}-\boldsymbol{\beta}\right) \le \left(k+1\right)s^2 f_{\alpha,k+1,n-k-1}$. Here $\boldsymbol{b}$ and s^2 are estimates.

Confidence intervals for expected response and predictions.

We will now construct a confidence interval for the expected response given values $x_{10},x_{20},\ldots x_{k0}$ for $x_1,x_2,\ldots x_k$.

Since $\hat{\boldsymbol{\beta}}$ is independent of S^2 we have that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ also must be independent of S^2.

Further

$$E\left(Y\middle|x_{10},x_{20},\ldots,x_{k0}\right) = \beta_0 + \beta_1 x_{10} + \ldots + \beta_k x_{k0} = \left[1,x_{10},x_{20},\ldots,x_{k0}\right]\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{bmatrix} = \mathbf{x}_0^t\boldsymbol{\beta} = \mu_{Y|x_{10},\ldots x_{k0}}$$. Also

$\hat{\mu}_{Y|x_{10},\mathbf{K}\,x_{k0}} = \mathbf{x}_0^t\hat{\boldsymbol{\beta}} = \hat{Y}_0$.

$$Var\left(\hat{Y}_0\right) = E\left[\boldsymbol{x}_0^t\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t\boldsymbol{x}_0\right] = \boldsymbol{x}_0^t E\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t\right]\boldsymbol{x}_0 = \sigma^2\,\boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0$$

Since $\hat{\boldsymbol{\beta}}$ and S^2 are independent we get:

$T = \dfrac{\hat{Y}_0 - \mu_{Y|x_{10},x_{20},\mathrm{K},x_{k0}}}{S\sqrt{\boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}}$ is t-distributed with $(n-k-1)$ degrees of freedom.

Hence a $100(1-\alpha)\%$ confidence interval for expected response when $\boldsymbol{x} = \boldsymbol{x}_0$ is:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-k-1}\, s\sqrt{\left(\boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right)}$$

Prediction interval for a new value of the response when $\boldsymbol{x} = \boldsymbol{x}_0$.

Let Y_0 be the new value. Then

$$E\left[Y_0 - \hat{Y}_0\right] = E\left[\boldsymbol{x}_0^t\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right] - E\left[\boldsymbol{x}_0^t\boldsymbol{\beta}\right] = \boldsymbol{x}_0^t\boldsymbol{\beta} - \boldsymbol{x}_0^t\boldsymbol{\beta} = 0.$$

$$Var\left[Y_0 - \hat{Y}_0\right] = Var\left[Y_0\right] + Var\left[\hat{Y}_0\right] = \sigma^2 - \sigma^2\,\boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0$$

such that

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + \boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}}$$

is t-distributed with $(n-k-1)$ degrees of freedom. Hence a $100(1-\alpha)\%$ prediction interval for Y given $\boldsymbol{x} = \boldsymbol{x}_0$ is:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-k-1}\, s\sqrt{\left(1 + \boldsymbol{x}_0^t\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right)}.$$

Choice of a fitted model

Let Y be the response and $x_1, x_2, \ldots x_k$ be the explanatory variable. Explanatory variables with small or no influence on the response can give the model a bad prediction ability.

Criteria for evaluating the adequacy of a model

We have

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \text{ or } SS_T = SS_E + SS_R.$$

In a good model $y_i - \hat{y}_i$ should be small and $\hat{y}_i - \bar{y} \approx y_i - \bar{y}, i = 1, 2, \ldots, n$. A measure of how much variation in the data that can be explained by the model is the coefficient of multiple determination:

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}.$$

We have $0 \le R^2 \le 1$. $R^2 = 0.84$ tells us that 84% of the variation in the data can be explained by the model. Note that $R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T}$.

A problem with R^2 is that it will always increase when we increase the number of explanatory variables. Therefore, $R^2_{adjusted}$ has been introduced defined as

$$R^2_{adjusted} = 1 - \frac{\frac{SS_E}{n - (k+1)}}{\frac{SS_T}{n-1}} = 1 - \frac{(n-1)s^2}{SS_T}.$$

Therefore maximizing $R^2_{adjusted}$ is equivalent to minimizing s^2.

An example. Sour precipitation

In order to study the influence of sour precipitation in Norwegian lakes, SFT in 1986 carried through an investigation where data from 1005 lakes were collected. In this example 26 random lakes, 16 from Telemark and 10 from Trøndelag is chosen out.

y = Measured pH-value
x_1 = Content of SO_4
x_2 = Content of NO_3
x_3 = Content of Ca
x_4 =Content of latent aluminum
x_5 = Content of organic material
x_6 = Area of water
x_7 = Location (0 – Telemark, 1-Trøndelag)

The observed data is given below:

```
      y  x1 x2   x3  x4   x5  x6  x7
1  5.38 4.9 39 1.54  78 2.02 0.30  0
2  5.68 4.1 75 1.55  17 2.98 1.85  0
```

```
3  5.04 3.5 80 0.83 157 3.40 0.25  0
4  4.81 3.8 75 0.53 163 3.42 0.30  0
5  4.92 3.8 90 0.82 105 3.41 0.25  0
6  5.34 2.6 49 0.62 114 1.90 0.65  0
7  5.74 2.7 79 1.08  15 2.53 1.15  0
8  5.17 2.6 90 0.67  90 1.89 0.91  0
9  5.02 2.4 64 0.41 107 0.97 0.60  0
10 5.88 2.8 27 1.15  12 3.04 0.37  0
11 5.36 3.4 13 0.89  93 2.95 0.58  0
12 5.26 2.7 14 0.74  72 2.75 0.15  0
13 5.69 3.2 13 1.03  99 3.34 0.53  0
14 5.51 2.5 79 0.67  80 0.78 0.56  0
15 5.25 1.5 77 0.33  72 0.11 1.08  0
16 6.06 3.7 15 1.94   4 5.53 0.72  0
17 6.08 1.9 16 1.05   0 5.63 0.48  1
18 6.08 1.3  2 0.81   0 3.84 0.25  1
19 6.20 2.2 32 1.40   7 3.77 1.87  1
20 5.64 2.2 21 0.75  13 5.62 0.53  1
21 5.75 1.6  7 0.79   2 3.95 0.35  1
22 5.43 2.0 27 0.47  16 2.39 1.03  1
23 5.82 1.8 17 0.74   3 3.08 0.32  1
24 5.50 2.0  6 0.49  10 2.82 0.28  1
25 5.62 1.5  3 0.36   7 1.71 0.10  1
26 5.41 1.7  7 0.54   9 6.48 0.21  1
```

In R we have the possibility of plotting pairs of columns like this:

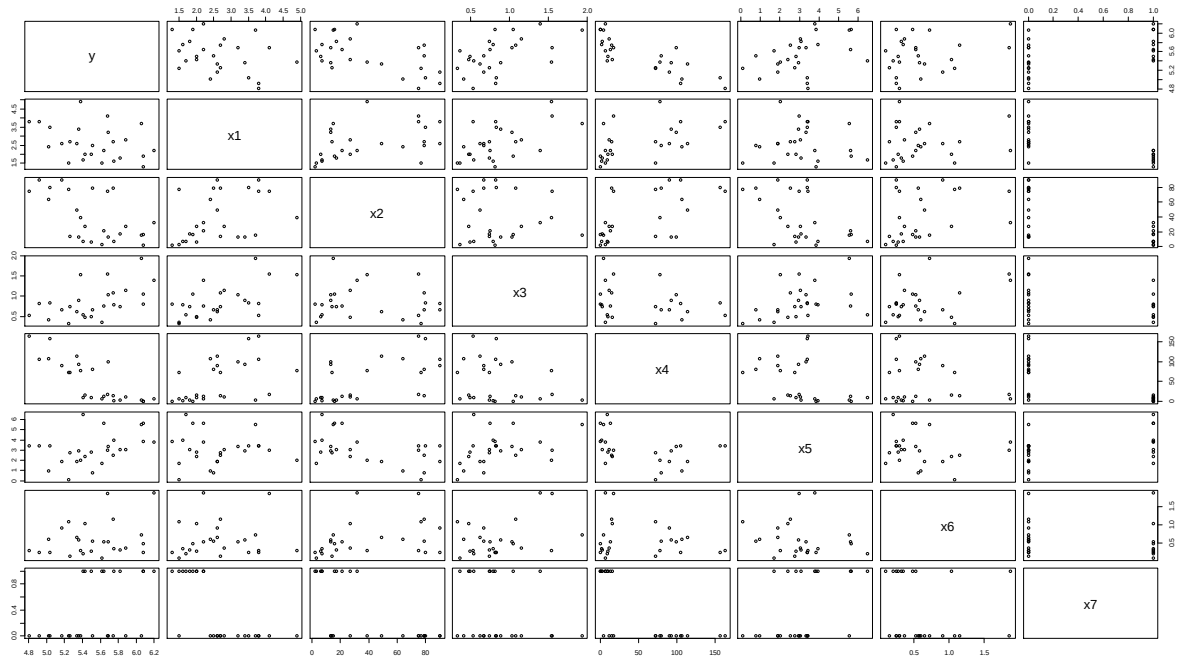


Figure7. In this figure the response is plotted against each regression variable and each regression variable is plotted against other regression variables

From the figure it seems like there might be a relationship between y and x_2, x_3, x_4 and x_5.

We can also calculate the correlation between all pairs of two variables. The correlation between Y and x_1 is calculated as: $r_{Y,x_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{1i} - \bar{x}_1)}{\sqrt{(y_i - \bar{y})^2 (x_{1i} - \bar{x})^2}}$, for instance

	y	x1	x2	x3	x4	x5	x6	x7
y	1.0000000	-0.33530173	-0.5703950	0.5263695	-0.8128681	0.40547721	0.2764476	0.4930767
x1	-0.3353017	1.00000000	0.4090205	0.5830377	0.5144954	-0.02402104	0.0891068	-0.7029792
x2	-0.5703950	0.40902048	1.0000000	-0.0585331	0.5973926	-0.47822779	0.3481284	-0.6470683
x3	0.5263695	0.58303773	-0.0585331	1.0000000	-0.2453771	0.36579209	0.3744328	-0.2292290
x4	-0.8128681	0.51449542	0.5973926	-0.2453771	1.0000000	-0.39443507	-0.1857614	-0.6966124
x5	0.4054772	-0.02402104	-0.4782278	0.3657921	-0.3944351	1.00000000	-0.1519310	0.4444162
x6	0.2764476	0.08910679	0.3481284	0.3744328	-0.1857614	-0.15193098	1.0000000	-0.1045081
x7	0.4930767	-0.70297917	-0.6470683	-0.2292290	-0.6966124	0.44441622	-0.1045081	1.0000000

It is clear that the variable that is the most correlated with the response is x_4, but there is also some correlation between Y and x_2, x_3, x_5 and x_7.

Using x_4 as the single regression variable we get the following output from R:

```
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  5.8259182  0.0619470  94.047   < 2e-16 ***
x4          -0.0058244  0.0008519  -6.837   4.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.222 on 24 degrees of freedom
Multiple R-squared: 0.6608,     Adjusted R-squared: 0.6466
F-statistic: 46.75 on 1 and 24 DF,  p-value: 4.521e-07

With the following analysis of variance table
                     x4 Residuals
Sum of Squares  2.303553  1.182693
Deg. of Freedom        1        24

Residual standard error: 0.2219885
```

According to the t-test, the variable x_4 and the constant term is highly significant. Further we get that $s = 0.222$ or 0.2219885 and $24s^2$ =1.182693= SS_E. $R^2 = \frac{SS_R}{SS_T}$ =0.6608 and R^2_{adj} =0.6466. SS_R =2.303553 and $SS_T = 2.303553 + 1.182693 = 3.486245$

If we try to estimate a model with x_2, x_3, x_4, x_5 and x_7 as the regression variables we get the following results:

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  5.444654   0.205522  26.492  < 2e-16 ***
x2          -0.002140   0.001620  -1.321  0.20139
x3           0.468389   0.139062   3.368  0.00306 **
x4          -0.003517   0.001256  -2.800  0.01107 *
x5          -0.038055   0.033321  -1.142  0.26691
x7           0.164354   0.155003   1.060  0.30163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1815 on 20 degrees of freedom
Multiple R-squared: 0.8111,     Adjusted R-squared: 0.7639
F-statistic: 17.18 on 5 and 20 DF,  p-value: 1.237e-06
```

We observe that with these variables in the model, x_3 and x_4 seem to be significant.
Finally, with all the variables in the model x_1 and x_3 seem to be the significant ones while x_4 is not even close to being significant.

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  5.6764334  0.1389162  40.862  < 2e-16 ***
x1          -0.3150444  0.0587512  -5.362 4.27e-05 ***
x2          -0.0018533  0.0012587  -1.472    0.158
x3           0.9751745  0.1449075   6.730 2.62e-06 ***
x4          -0.0002268  0.0010038  -0.226    0.824
x5          -0.0334242  0.0225009  -1.485    0.155
x6          -0.0039399  0.0724339  -0.054    0.957
x7           0.0888722  0.1025724   0.866    0.398
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1165 on 18 degrees of freedom
Multiple R-squared:  0.93,      Adjusted R-squared: 0.9027
F-statistic: 34.15 on 7 and 18 DF, p-value: 3.904e-09
```

This example shows that it is not an easy task to find the model that best explains the variation in the response. Fortunately, there are other methods than the ones above that are helpful when doing model selection. First, we will show some procedures based on significant tests that are useful to select variables to be in the model.

Variable selection methods

The most common variable selection methods are forward selection, backward elimination, and stepwise regression. Normally these algorithms are performed as partial F-tests, adding or removing one variable at the time. We will now give an algorithmic presentation of these methods.

Forward selection

1. Start with only β_0 in the model.
2. Find $\max_j R(\beta_j) = \max_j SS_R(\beta_j|\beta_0) = \max_j \{SS_R(\beta_0,\beta_j) - SS_R(\beta_0)\}$.
3. If $\max_j \frac{R(\beta_j)}{\frac{SS_E}{n-2}} = \frac{R(\beta_m)}{\frac{SS_E}{n-2}} < f_{\alpha,1,n-2}$ stop, no variable is entered into the model.
4. If $\frac{R(\beta_m)}{\frac{SS_E}{n-2}} \geq f_{\alpha,1,n-2}$ add x_m to the modell.

Find $\max_{j\neq m} R(\beta_j|\beta_m) = \max_{j\neq m} \{SS_R(\beta_0,\beta_m,\beta_j) - SS_R(\beta_0,\beta_m)\}$ and go to step 3. If $\max_{j\neq m} \frac{R(\beta_j|\beta_m)}{\frac{SS_E}{n-3}} < f_{\alpha,1,n-3}$, no more variables are entered into the model. Otherwise proceed in the same way. Note that the degrees of freedom in the partial F-test is reduced by one for each variable that is entered into the model.

Backward elimination

Define $\boldsymbol{\beta} \setminus \beta_j = (\beta_0,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_k)$.

1. Start with all the variables in the model.
2. Find $\min_j R(\beta_j|\boldsymbol{\beta}\setminus\beta_j) = \min_j \{SS_R(\beta_0,\ldots,\beta_k) - SS_R(\beta_0,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_k)\}$.
3. If $\min_j \frac{R(\beta_j|\boldsymbol{\beta}\setminus\beta_j)}{\frac{SS_E}{n-k-1}} = \frac{R(\beta_m|\boldsymbol{\beta}\setminus\beta_m)}{\frac{SS_E}{n-k-1}} > f_{\alpha,1,n-k-1}$ stop, no variable is removed from the model.
4. If $\frac{R(\beta_m|\beta\setminus\beta_m)}{\frac{SSE}{n-k-1}} < f_{\alpha,1,n-k-1}$, remove x_m.

Find $\min_{j\neq m} R(\beta_j|\boldsymbol{\beta}\setminus\{\beta_m,\beta_j\}) = \min_{j\neq m} \{SS_R(\beta_0,\ldots,\beta_{m-1},\beta_{m+1},\ldots,\beta_k) - SS_R(\boldsymbol{\beta}\setminus\{\beta_m,\beta_j\})\}$ and go to step 3.

If $\dfrac{\min\limits_{j\neq m} R\left(\beta_j \middle| \boldsymbol{\beta} \setminus \{\beta_m, \beta_j\}\right)}{\dfrac{SS_E}{n-k}} > f_{\alpha,1,n-k}$ no more variables are taken out of the model.

Otherwise proceed in the same way. Note that the degrees of freedom in the partial F-test is increased by one in each step.

Stepwise regression

This method combines forward selection and backward elimination.

1. Start as with forward selection. Assume x_n and x_m are chosen as the first two variables to enter the model.
2. Find $\min\limits_{j=n,m} R\left(\beta_j \middle| \{\beta_n, \beta_m\} \setminus \beta_j\right)$ and check if one of the variables x_n or x_m can be taken out as with backward selection.
3. Continue as with forward selection but check in each step if variables that are chosen to be in the model in earlier steps can be taken out.

Since an F-test where the enumerator has one degrees of freedom is equivalent to a t-test for evaluating the significance of a parameter, t-tests could be used instead of F-tests. In R these three selection methods are performed based on criteria and not on significance evaluation. These criteria penalize large models. Default in R is the Akaike's information criterion given as:

$$AIC = n\ln\left(\frac{SS_E}{n}\right) + 2\left(k+2\right)$$

and small values for AIC are to be preferred. Note that now also σ^2 is considered a parameter, therefore k+2. An alternative that penalizes large models more is the Bayesian information criterion:

$$BIC = n\ln\left(\frac{SS_E}{n}\right) + \ln\left(n\right)\left(k+2\right).$$

Other methods are best subset regression, where one seek through the set of regression variables and for each s, $s = 1, 2, \ldots, k$ find the subset of s regression variables that gives the best fit to the response measured in terms of criteria such as R^2, $R^2_{adjusted}$, AIC, BIC or Mallow's-C_p.

Mallows-C_p

The idea behind Mallows-C_p is that most (all) models are wrong and therefore $E\left(Y_i\right) \neq E\left(\hat{Y}_i\right)$. Now writing

$$\hat{Y}_i - E\left(Y_i\right) = \hat{Y}_i - E\left(\hat{Y}_i\right) + E\left(\hat{Y}_i\right) - E\left(Y_i\right),$$

we get

$$E\left(\hat{Y}_i - E\left(Y_i\right)\right)^2 = E\left(\hat{Y}_i - E\left(\hat{Y}_i\right)\right)^2 + E\left(E\left(\hat{Y}_i\right) - E\left(Y_i\right)\right)^2 + 2E\left[\left(\hat{Y}_i - E\left(\hat{Y}_i\right)\right)\left(E\left(\hat{Y}_i\right) - E\left(Y_i\right)\right)\right].$$

The last term is zero and we obtain

$$E\left[\left(\hat{Y}_i - E\left(Y_i\right)\right)^2\right] = Var\left(\hat{Y}_i\right) + \left(Bias\hat{Y}_i\right)^2$$

where $Bias\hat{Y}_i = E\left(\hat{Y}_i\right) - E\left(Y_i\right)$.

We want to minimize

$$E\left[\sum_{i=1}^{n}\frac{\left(\hat{Y}_i - E\left(Y_i\right)\right)^2}{\sigma^2}\right] = \sum_{i=1}^{n}\frac{Var\left(\hat{Y}_i\right)}{\sigma^2} + \sum_{i=1}^{n}\frac{\left(Bias\hat{Y}_i\right)^2}{\sigma^2}.$$

Now $\sum_{i=1}^{n} Var\left(\hat{Y}_i\right) = \sigma^2\sum_{i=1}^{n} h_{ii} = \sigma^2\left(k+1\right)$. An estimate for Mallows-C_p is given by:

$$C_p = k + 1 + \frac{\left(s^2 - \hat{\sigma}^2\right)\left(n - k - 1\right)}{\hat{\sigma}^2}.$$

Correct model has $C_p = k + 1 = p$.

s^2 is an estimate for $Var\left(Y_i\right)$ in the model with *p*-parameters and $\hat{\sigma}^2$ is an estimate for σ^2. Often it is the estimate of σ^2 from the full model.

As an example, forward regression is applied to the data set about sour precipitation using the default algorithm in R.

```
>step(lm(y~1),y~1+x1+x2+x3+x4+x5+x6+x7, direction="forward")
Start:  AIC=-50.24
y ~ 1

       Df Sum of Sq    RSS     AIC
+ x4    1   2.30355 1.1827 -76.348
+ x2    1   1.13425 2.3520 -58.474
+ x3    1   0.96592 2.5203 -56.676
+ x7    1   0.84759 2.6387 -55.484
+ x5    1   0.57318 2.9131 -52.911
+ x1    1   0.39195 3.0943 -51.342
+ x6    1   0.26643 3.2198 -50.308
<none>              3.4862 -50.241

Step:  AIC=-76.35
y ~ x4

       Df Sum of Sq     RSS     AIC
```

```
+ x3    1   0.39645 0.78625 -84.963
<none>              1.18269 -76.348
+ x6    1   0.05682 1.12587 -75.628
+ x2    1   0.03898 1.14372 -75.219
+ x7    1   0.03627 1.14642 -75.158
+ x1    1   0.03260 1.15010 -75.075
+ x5    1   0.02973 1.15297 -75.010

Step:  AIC=-84.96
y ~ x4 + x3

       Df Sum of Sq     RSS      AIC
+ x1    1   0.46780 0.31844 -106.463
+ x2    1   0.07316 0.71309  -85.502
<none>              0.78625  -84.963
+ x7    1   0.04413 0.74212  -84.465
+ x6    1   0.00050 0.78575  -82.980
+ x5    1   0.00035 0.78590  -82.975

Step:  AIC=-106.46
y ~ x4 + x3 + x1

       Df Sum of Sq     RSS     AIC
+ x2    1  0.041571 0.27687 -108.10
<none>              0.31844 -106.46
+ x7    1  0.010093 0.30835 -105.30
+ x6    1  0.007048 0.31140 -105.04
+ x5    1  0.003697 0.31475 -104.77

Step:  AIC=-108.1
y ~ x4 + x3 + x1 + x2

       Df Sum of Sq     RSS     AIC
+ x5    1 0.0223623 0.25451 -108.29
<none>              0.27687 -108.10
+ x6    1 0.0025490 0.27432 -106.34
+ x7    1 0.0003829 0.27649 -106.14

Step:  AIC=-108.29
y ~ x4 + x3 + x1 + x2 + x5

       Df Sum of Sq     RSS     AIC
<none>              0.25451 -108.29
+ x7    1 0.0103492 0.24416 -107.37
+ x6    1 0.0002081 0.25430 -106.31

Call:
lm(formula = y ~ x4 + x3 + x1 + x2 + x5)
```

```
Coefficients:
(Intercept)      x4         x3          x1          x2          x5
  5.7668186 -0.0006377  0.9292754  -0.3210916  -0.0021220  -0.0242642
```

In the first step x_4 is added as the regression variable that gives the smallest *AIC*. Then x_3 is the only variable that together with x_4 will reduce the *AIC* criterion, so x_3 is added. In the next step both x_1 and x_2 are candidates to enter the model, but x_1 is the one that reduces the *AIC* criterion the most. x_2 is entered in the next step and after that x_5. The procedure stops with x_1, x_2, x_3, x_4 and x_5 in the model.

Performing backward elimination, we get the result below.

```
> step(mod)
Start:  AIC=-105.37
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7

       Df Sum of Sq     RSS      AIC
- x6    1   0.00004 0.24416 -107.369
- x4    1   0.00069 0.24481 -107.299
- x7    1   0.01018 0.25430 -106.311
<none>              0.24412 -105.373
- x2    1   0.02940 0.27352 -104.416
- x5    1   0.02993 0.27405 -104.366
- x1    1   0.38998 0.63410  -82.555
- x3    1   0.61421 0.85833  -74.682

Step:  AIC=-107.37
y ~ x1 + x2 + x3 + x4 + x5 + x7

       Df Sum of Sq     RSS      AIC
- x4    1   0.00068 0.24484 -109.297
- x7    1   0.01035 0.25451 -108.289
<none>              0.24416 -107.369
- x5    1   0.03233 0.27649 -106.136
- x2    1   0.04480 0.28896 -104.988
- x1    1   0.41435 0.65851  -83.572
- x3    1   0.78768 1.03184  -71.896

Step:  AIC=-109.3
y ~ x1 + x2 + x3 + x5 + x7

       Df Sum of Sq     RSS      AIC
- x7    1   0.01674 0.26158 -109.577
<none>              0.24484 -109.297
- x5    1   0.03582 0.28066 -107.746
```

```
- x2    1   0.04690 0.29174 -106.740
- x1    1   0.67174 0.91657  -76.975
- x3    1   1.94498 2.18982  -54.331

Step:  AIC=-109.58
y ~ x1 + x2 + x3 + x5

       Df Sum of Sq     RSS      AIC
<none>              0.26158 -109.577
- x5    1   0.02188 0.28346 -109.488
- x2    1   0.07908 0.34066 -104.709
- x1    1   1.23395 1.49553  -66.246
- x3    1   1.93493 2.19651  -56.252

Call:
lm(formula = y ~ x1 + x2 + x3 + x5, data = sourprec)

Coefficients:
(Intercept)         x1           x2           x3           x5
    5.77017    -0.35181     -0.00233      0.98990     -0.02400
```

In the first step, taking out x_6 reduces the AIC the most. In the next steps first x_4 and thereafter x_7 are taken out and we are left with x_1, x_2, x_3 and x_5 in the model.

Finally, if we perform stepwise regression we get:

```
> step(lm(y~1),y~1+x1+x2+x3+x4+x5+x6+x7, direction="both")
Start:  AIC=-50.24

y ~ 1
       Df Sum of Sq    RSS     AIC
+ x4    1   2.30355 1.1827 -76.348
+ x2    1   1.13425 2.3520 -58.474
+ x3    1   0.96592 2.5203 -56.676
+ x7    1   0.84759 2.6387 -55.484
+ x5    1   0.57318 2.9131 -52.911
+ x1    1   0.39195 3.0943 -51.342
+ x6    1   0.26643 3.2198 -50.308
<none>              3.4862 -50.241

Step:  AIC=-76.35
y ~ x4

       Df Sum of Sq    RSS     AIC
+ x3    1   0.39645 0.7862 -84.963
<none>              1.1827 -76.348
+ x6    1   0.05682 1.1259 -75.628
```

```
+ x2    1   0.03898 1.1437 -75.219
+ x7    1   0.03627 1.1464 -75.158
+ x1    1   0.03260 1.1501 -75.075
+ x5    1   0.02973 1.1530 -75.010
- x4    1   2.30355 3.4862 -50.241

Step:  AIC=-84.96
y ~ x4 + x3

       Df Sum of Sq     RSS      AIC
+ x1    1   0.46780 0.31844 -106.463
+ x2    1   0.07316 0.71309  -85.502
<none>              0.78625  -84.963
+ x7    1   0.04413 0.74212  -84.465
+ x6    1   0.00050 0.78575  -82.980
+ x5    1   0.00035 0.78590  -82.975
- x3    1   0.39645 1.18269  -76.348
- x4    1   1.73408 2.52033  -56.676

Step:  AIC=-106.46
y ~ x4 + x3 + x1

       Df Sum of Sq     RSS      AIC
+ x2    1   0.04157 0.27687 -108.100
- x4    1   0.02365 0.34210 -106.600
<none>              0.31844 -106.463
+ x7    1   0.01009 0.30835 -105.300
+ x6    1   0.00705 0.31140 -105.044
+ x5    1   0.00370 0.31475 -104.766
- x1    1   0.46780 0.78625  -84.963
- x3    1   0.83165 1.15010  -75.075

Step:  AIC=-108.1
y ~ x4 + x3 + x1 + x2

       Df Sum of Sq     RSS      AIC
- x4    1   0.00659 0.28346 -109.488
+ x5    1   0.02236 0.25451 -108.289
<none>              0.27687 -108.100
- x2    1   0.04157 0.31844 -106.463
+ x6    1   0.00255 0.27432 -106.340
+ x7    1   0.00038 0.27649 -106.136
- x1    1   0.43622 0.71309  -85.502
- x3    1   0.82188 1.09875  -74.262

Step:  AIC=-109.49
y ~ x3 + x1 + x2
```

```
       Df Sum of Sq     RSS      AIC
+ x5    1   0.02188 0.26158 -109.577
<none>              0.28346 -109.488
+ x4    1   0.00659 0.27687 -108.100
+ x6    1   0.00391 0.27955 -107.849
+ x7    1   0.00280 0.28066 -107.746
- x2    1   0.05864 0.34210 -106.600
- x1    1   1.21835 1.50181  -68.137
- x3    1   2.02498 2.30844  -56.960

Step:  AIC=-109.58
y ~ x3 + x1 + x2 + x5

       Df Sum of Sq     RSS      AIC
<none>              0.26158 -109.577
- x5    1   0.02188 0.28346 -109.488
+ x7    1   0.01674 0.24484 -109.297
+ x4    1   0.00707 0.25451 -108.289
+ x6    1   0.00079 0.26079 -107.656
- x2    1   0.07908 0.34066 -104.709
- x1    1   1.23395 1.49553  -66.246
- x3    1   1.93493 2.19651  -56.252

Call:
lm(formula = y ~ x3 + x1 + x2 + x5)

Coefficients:
(Intercept)     x3           x1           x2           x5
    5.77017  0.98990     -0.35181     -0.00233     -0.02400
```

As expected, x_4 is entered in the model in the first step, but with x_1, x_2 and x_3 also in the model x_4 is taken out and we end up with the same model as we did when performing backward elimination.

As should be noted there is no way of removing variable once entered when doing forward selection. Also, in common with stepwise regression, it may in the beginning be difficult for variable to be judged significant when using the F-test since the variation in the response caused by the rest of the regression variables is included in the error variance. Therefore, one often runs the procedure several times with different choice of α. If possible, it may be an advantage to use backward elimination. The advantage of the forward selection and stepwise regression algorithms is that they are more general applicable. To my knowledge there are no algorithms implemented for using F-tests or (t-tests) in R. However, backward elimination can be performed the following way where we first start with the full model.

```
> mod=lm(y~x1+x2+x3+x4+x5+x6+x7,data=sourprec)
> summary(mod)
```

```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = sourprec)
Residuals:
     Min       1Q   Median       3Q      Max
-0.15266 -0.09355  0.01429  0.06801  0.17755
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6764334  0.1389162  40.862  < 2e-16 ***
x1          -0.3150444  0.0587512  -5.362 4.27e-05 ***
x2          -0.0018533  0.0012587  -1.472    0.158
x3           0.9751745  0.1449075   6.730 2.62e-06 ***
x4          -0.0002268  0.0010038  -0.226    0.824
x5          -0.0334242  0.0225009  -1.485    0.155
x6          -0.0039399  0.0724339  -0.054    0.957
x7           0.0888722  0.1025724   0.866    0.398
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1165 on 18 degrees of freedom
Multiple R-squared:  0.93,      Adjusted R-squared: 0.9027
F-statistic: 34.15 on 7 and 18 DF,  p-value: 3.904e-09
```

The least significant one is x_6. Therefore, in the next step we can use the command update:

```
> mod1=(update(mod,.~.-x6))
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x7, data = sourprec)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15269 -0.09218  0.01337  0.06853  0.17636

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6757654  0.1346927  42.139  < 2e-16 ***
x1          -0.3142383  0.0553395  -5.678 1.79e-05 ***
x2          -0.0018918  0.0010132  -1.867   0.0774 .
x3           0.9714256  0.1240784   7.829 2.31e-07 ***
x4          -0.0002239  0.0009757  -0.229   0.8210
x5          -0.0330468  0.0208352  -1.586   0.1292
x7           0.0877409  0.0977709   0.897   0.3807
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1134 on 19 degrees of freedom
Multiple R-squared:  0.93,      Adjusted R-squared: 0.9078
F-statistic: 42.05 on 6 and 19 DF, p-value: 5.683e-10
```

By continuing the updating, we end up with the model:

```
Model 1
Call:
lm(formula = y ~ x1 + x2 + x3, data = sourprec)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19867 -0.09314  0.03027  0.07030  0.18855

Coefficients:
             Estimate   Std.Error  t value  Pr(>|t|)
(Intercept)  5.6990584  0.0697816  81.670  < 2e-16 ***
x1          -0.3489065  0.0358803  -9.724 2.00e-09 ***
x2          -0.0018364  0.0008608  -2.133   0.0443 *
x3           0.9548356  0.0761644  12.537 1.71e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1135 on 22 degrees of freedom
Multiple R-squared: 0.9187,     Adjusted R-squared: 0.9076
F-statistic: 82.86 on 3 and 22 DF, p-value: 3.821e-12
```

From backward elimination and stepwise regression using the AIC-criterion we get:

```
Model 2
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7701749  0.0871028  66.246  < 2e-16 ***
x1          -0.3518082  0.0353467  -9.953 2.10e-09 ***
x2          -0.0023299  0.0009247  -2.520   0.0199 *
x3           0.9899007  0.0794240  12.463 3.61e-11 ***
x5          -0.0239962  0.0181057  -1.325   0.1993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1116 on 21 degrees of freedom
Multiple R-squared: 0.925,      Adjusted R-squared: 0.9107
F-statistic: 64.72 on 4 and 21 DF,  p-value: 1.659e-11
```

And from the forward selection method using AIC

```
Model 3
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7668186  0.0881546  65.417  < 2e-16 ***
x1          -0.3210916  0.0545427  -5.887 9.29e-06 ***
x2          -0.0021220  0.0009753  -2.176   0.0417 *
x3           0.9292754  0.1142848   8.131 9.06e-08 ***
x4          -0.0006377  0.0008557  -0.745   0.4647
```

```
x5          -0.0242642  0.0183040  -1.326   0.1999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1128 on 20 degrees of freedom
Multiple R-squared: 0.927,      Adjusted R-squared: 0.9087
F-statistic: 50.79 on 5 and 20 DF,  p-value: 1.100e-10
```

Ranked according to $R^2_{adjusted}$ model 2 is better than model 3 which again is better than model 1. But model 3 has two non-significant terms and model 2 has one. The two candidate models would in practice be model 1 and 2.

Now regardless of if we include the amount of organic material or not, we have arrived at a model saying that adding SO_4 and NO_3 makes the lakes more sour and adding Ca makes it less sour which should be in agreement with common knowledge. But what about the content of latent aluminum that is so highly correlated with the ph value. The explanation is simply that in sour lakes aluminum will be released from the rock. Therefore, sour lakes cause a high amount of aluminum, not the other way around.

Diagnostic checks of regression models.

The model we arrive at when performing regression analysis depends, strongly on the quality of the data we have i. e. the observations for the response and the regression variables. To thrust that the model is useful, we should check the quality of these. We will now discuss some problems that may arise and some methods that can be used to detect if such problems are present in our data.

Multicollinearity

Multicollinearity arises when two or more columns in the $\boldsymbol{X}$-matrix are strongly correlated (almost linear dependent). Multicollinearity can sometimes be discovered by estimating the correlation between regression variables. The correlation between regression variable x_i and x_j is given by:

$$r_{ij} = \frac{\sum_{k=1}^{n}\left(x_{ik}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ik}-\bar{x}_i\right)^2\sum_{k=1}^{n}\left(x_{jk}-\bar{x}_j\right)^2}}.$$

Multicollinearity often occurs when the variation interval of two or more variables are approximately equal.

An example is if $x_{ki} \in (0.95, 1.05)$, $i=1,2,\ldots,n$, the column with 1's, x_k and x_k^2 are strongly correlated. A good measure of multicollinearity is given by the variance inflation factor, VIF.

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the multiple determination coefficient when regression variable x_j is regressed on the other regression variables. If you ever meet the word tolerance, it is given by 1-R_j^2. VIF should be less than 10. If multicollinearity occurs one should consider:

1. Remove columns.
2. Collect more data.
3. Respecify the model. For instance if multicollinearity occurs between the three regression variables x_1, x_2 and x_3, a new variable $\frac{x_1 + x_2 + x_3}{3}$ or $x_1 x_2 x_3$ may solve the problem.
4. Center variables i.e. use $x_i - \bar{x}$ instead of x_i if one works with polynomial models.

There are also several statistical methods like Principal component regression (PCR), Partial least square (PLS) and Ridge regression for handling such problems. These are, however, beyond the scope of this course.

Observations with influence

Influential points are points that when removed from the dataset could cause a large change in the fit. The performance of variable selection methods may depend heavily on influential observation. Therefore, it is a good idea to check for leverage points and the Cook's distance before starting a variable selection procedure.

Leverage points

We have $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} \Rightarrow \hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$. If h_{ii} is large y_i could have a large influence on $\hat{y}_i$.

Now $\text{rank}(\boldsymbol{H}) = k+1 = \text{tr}(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii}$.

Hence the average value of $h_{ii} = \frac{k+1}{n}$. Observations where $h_{ii} > \frac{2(k+1)}{n}$ are said to be leverage points. Note that if $\boldsymbol{x}_i^t$ is the i-th row in the $\boldsymbol{X}$ -matrix, then $h_{ii} = \boldsymbol{x}_i^t \left(\boldsymbol{X}^t \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i$.

Hence for an observation with a high leverage, the uncertainty in the estimator for the expected value of the response will be relatively high, and a corresponding prediction interval will be relatively wide.

Leverage points are entirely determined by the values in the $\boldsymbol{X}$ -matrix and need not necessarily influence our estimated coefficients. A check for this is done by the Cook's distance, D_i .

First: $\left(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}_{(-i)}\right)^t\left(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}_{(-i)}\right)=\left(\boldsymbol{Xb}-\boldsymbol{Xb}_{(-i)}\right)^t\left(\boldsymbol{Xb}-\boldsymbol{Xb}_{(-i)}\right)=\left(\boldsymbol{b}-\boldsymbol{b}_{(-i)}\right)^t\boldsymbol{X}^t\boldsymbol{X}\left(\boldsymbol{b}-\boldsymbol{b}_{(-i)}\right)$. The $(-i)$ notation means that the *i*-th set of observations $\left(y_i,x_{i1},...,x_{ik}\right)$ is taken out.

The Cook's distances D_i is defined as $D_i=\dfrac{\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}\right)^t\boldsymbol{X}^t\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}\right)}{p\,\hat{\sigma}^2}$, $i=1,2,\ldots,n$.

Usually, we consider an observation $\left(y_i,x_{i1},...,x_{ik}\right)$ to be influential if $D_i>1$.

Study of residuals

We expect the residuals from a good model to be uncorrelated with mean zero and constant variance. Now, if we have the correct model we can argue as follows:

$$\hat{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{Y}=\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{\varepsilon}.$$

Thereby $E\left(\hat{\boldsymbol{\varepsilon}}\right)=\left(\boldsymbol{I}-\boldsymbol{H}\right)E\left(\hat{\boldsymbol{\varepsilon}}\right)=\boldsymbol{0}$,

and $Cov\left(\hat{\boldsymbol{\varepsilon}}\right)=E\left[\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\left(\boldsymbol{I}-\boldsymbol{H}\right)^t\right]=\sigma^2\left(\boldsymbol{I}-\boldsymbol{H}\right)$.

which shows that $\hat{\varepsilon}_1,\hat{\varepsilon}_2,\ldots,\hat{\varepsilon}_n$ are not uncorrelated. Further we get

$$Var\left(\hat{\varepsilon}_i\right)=\sigma^2\left(1-h_{ii}\right).$$

Hence the variances of $\hat{\varepsilon}_1,\hat{\varepsilon}_2,\ldots,\hat{\varepsilon}_n$ need not be equal either.

Since $Var\left(\dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}\right)=1$ we often use $r_i=\dfrac{e_i}{s\sqrt{1-h_{ii}}}$ in the study of the residuals . These residuals are often said to be the standardized residuals. Some books call them studentized.

The residuals $\delta_i=Y_i-\hat{Y}_{i,-i}$, $i=1,2,\ldots,n$ where $\hat{Y}_{i,-i}$ is the estimator for the i-th fitted value when the observation $\left(y_i,x_{i1},...,x_{ik}\right)$ is taken out are called the Press residuals . These can be shown to be equal to $\dfrac{\hat{\varepsilon}_i}{1-h_{ii}}$ and $Var\left(\delta_i\right)=\dfrac{Var\left(\hat{\varepsilon}_i\right)}{\left(1-h_{ii}\right)^2}=\dfrac{\sigma^2\left(1-h_{ii}\right)}{\left(1-h_{ii}\right)^2}=\dfrac{\sigma^2}{1-h_{ii}}$.

Hence $\dfrac{\delta_i}{Sd\left(\delta_i\right)}=\dfrac{\hat{\varepsilon}_i}{\dfrac{\left(1-h_{ii}\right)\sigma}{\sqrt{1-h_{ii}}}}=\dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{\left(1-h_{ii}\right)}}$

Now assume we estimate σ^2 by $S^2_{(-i)}$ which is the estimator for the variance when $(y_i, x_{i1}, ..., x_{ik})$ is taken out. Then $\delta_i = Y_i - \hat{Y}_{i,-i}$ and $S^2_{(-i)}$ are independent and the residuals

$$\frac{\hat{\varepsilon}_i}{S_{-i}\sqrt{1-h_{ii}}}$$

is t-distributed with *n*-*k*-2 degrees of freedom and can be used to investigate if the expected value of a residual is different from zero, i.e. if we have an outlier. The residuals $\frac{e_i}{s_{-i}\sqrt{1-h_{ii}}}$ are called R-studentized, sometimes external studentized.

To summarize. Check of residuals provide information about:

1. Outliers
2. Heterogenity in variance
3. Misspecification of models (or if other variables should be included)
4. Normal distribution

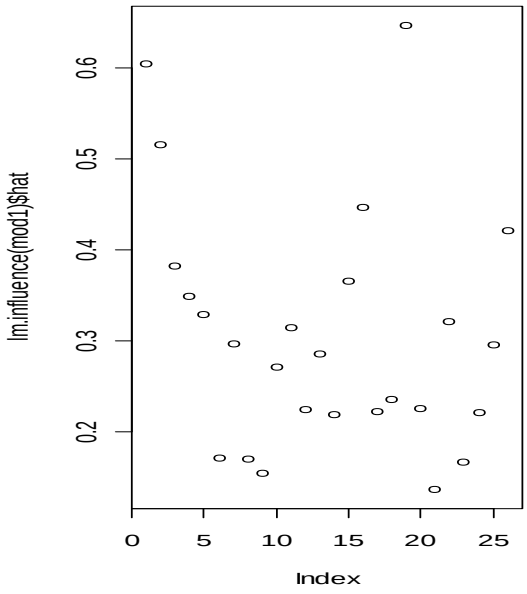
There are extremely many diagnostic tools developed for regression models and one can easily be confused about the importance of the information provided from each of them.
In practice it is most common to check for multicollinearity, check the Cook's distance and perform residuals plots. The most common residual plots are plot of residuals against fitted values to check for heterogeneity in variance, against other regression variables to see if they should be included and normal plot to check for normally distributed data and outliers.

Example. Sour precipitation

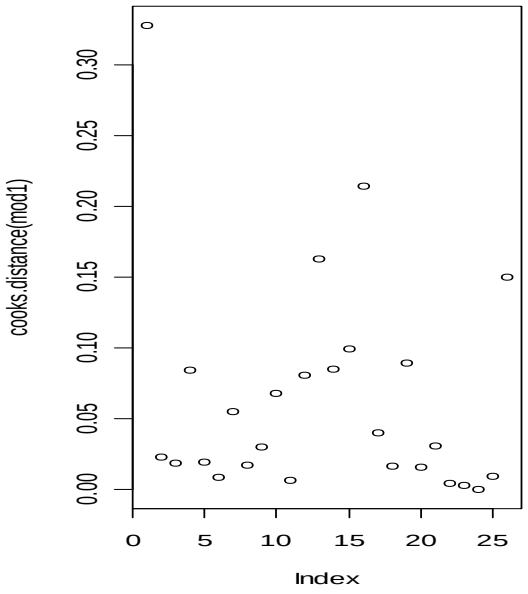
Let us return to our example about sour precipitation. The correlation matrix between pairs of variables indicated some high correlations between x_7 and other regression variables, but no clear sign of multicollinearity. The VIF-factor is in the R-library car and not in the basic R-package. This library car can be downloaded to any personal PC for free.

The critical value for the leverage is for this data set $\frac{2 \cdot 8}{26} = 0.62$. From the plot below there may be one observation with too high leverage. However, from the Cook's distance there

does not seem to be any observation that is really influential on our estimated coefficients.

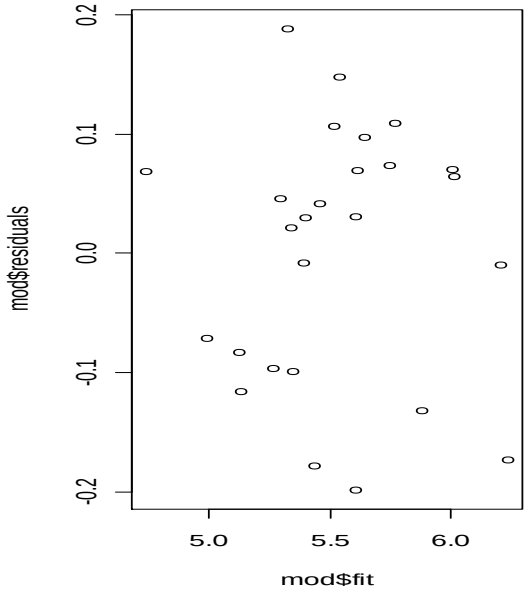


Leverages

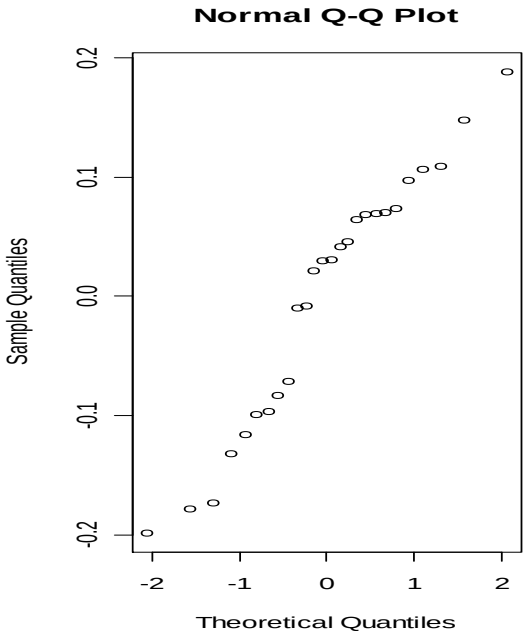


Cook's distances

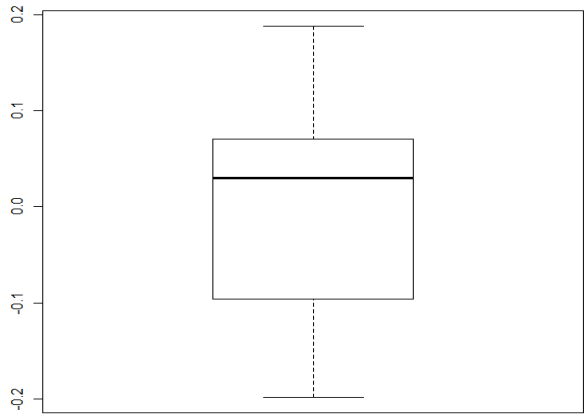
Next is shown a plot of residuals against fitted values and a normal plot. The residuals are obtained using x_1, x_2 and x_3 as regression variables, since these were the only three significant ones. Examining the plots, the variance does not seem to depend on the value of the fitted model and the normal plot is close to a straight line. The box plot reveal that the median of the residuals is above zero, but part from that there is no observation that is close to being an outlier.



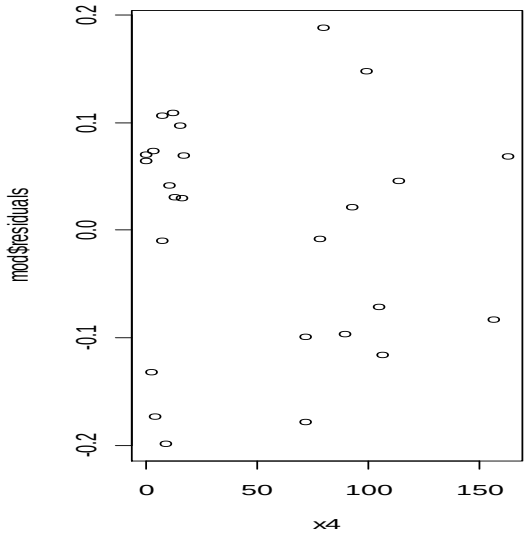
Plot of residuals against fitted values



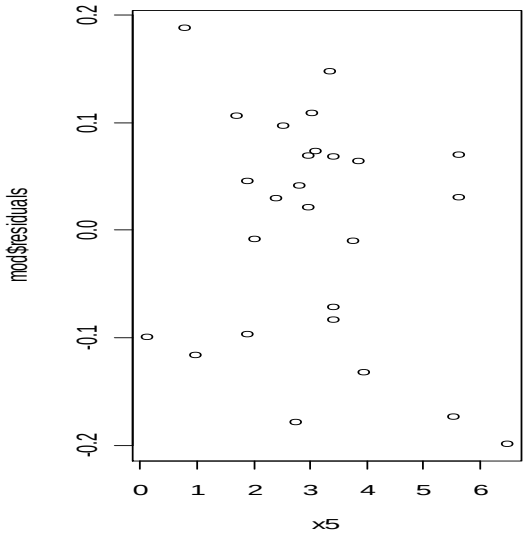
Normal plot of residuals



Boxplot of the residuals



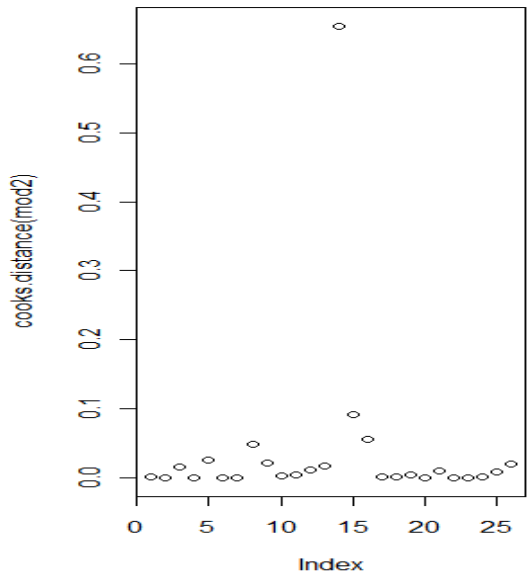
Plot of residuals against x_4



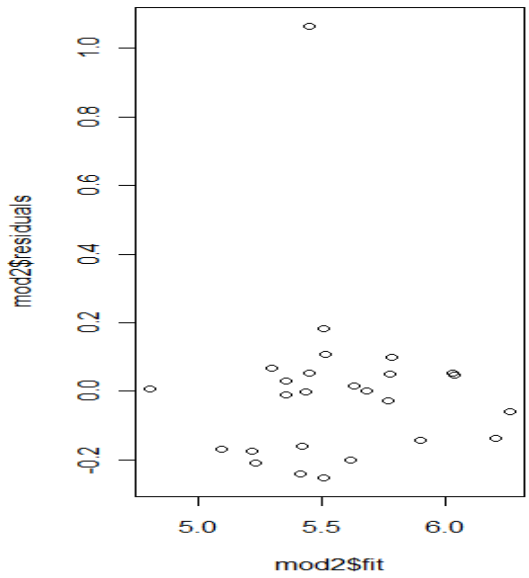
Plot of residuals against x_5

Finally plots of the residuals against x_4 and x_5 (content of aluminum and organic material**)** do not show any particular pattern and there is no obvious reason to include these in the model given that using x_1, x_2 and x_3 already are there. Hence these diagnostic tools give no reason to not trust our model.

To get an impression of what might happen if we had observations of poor quality, we have added 1 to one of the response values, while all the rest is kept unchanged. We estimate the same model.

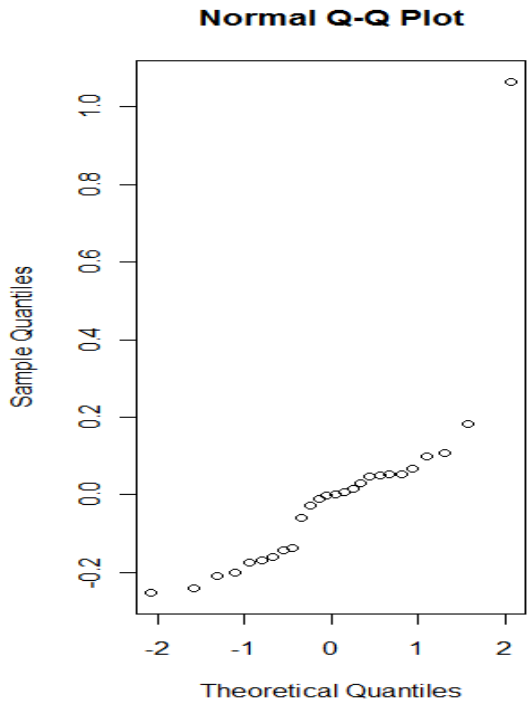


Cook's distances

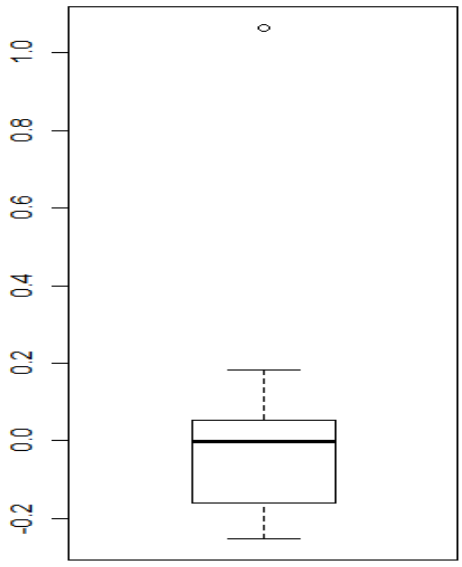


Plot of residuals against fitted values

We observe that the Cook's distance does not classify this observation as being strongly influential, while the normal plot finds one residual that does not fit into the assumption of having mean equal to zero and residuals being normally distributed. The boxplot of the residuals clearly signals an outlier. All the plots signal a very special observation.



Normal plot of residuals



Box plot of residuals

If we look at the estimated models we get:

Original data set:

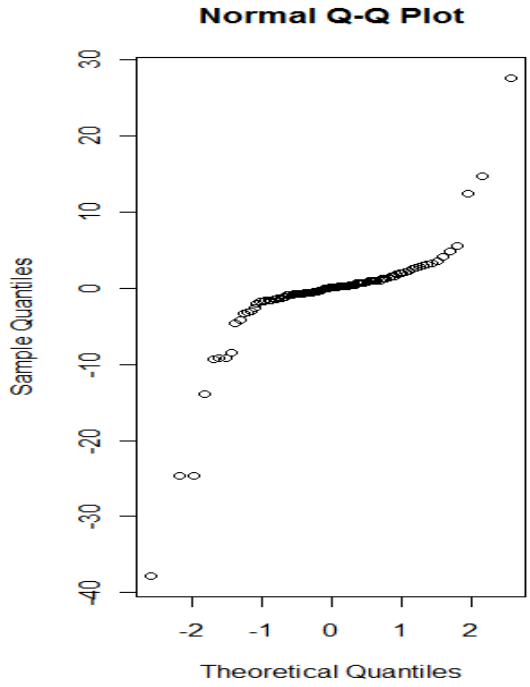
```
lm(formula = y ~ x1 + x2 + x3, data = sour)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6990584  0.0697816  81.670  < 2e-16 ***
x1          -0.3489065  0.0358803  -9.724 2.00e-09 ***
x2          -0.0018364  0.0008608  -2.133   0.0443 *
x3           0.9548356  0.0761644  12.537 1.71e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1135 on 22 degrees of freedom
Multiple R-squared: 0.9187,     Adjusted R-squared: 0.9076
F-statistic: 82.86 on 3 and 22 DF,  p-value: 3.821e-12
```

manipulated data set

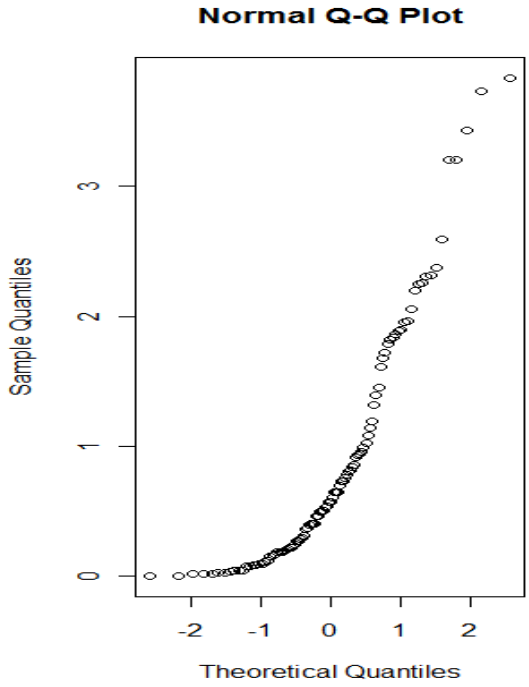
```
lm(formula = y ~ x1 + x2 + x3, data = sour1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7446812  0.1624766  35.357  < 2e-16 ***
x1          -0.3882316  0.0835422  -4.647 0.000124 ***
x2           0.0002539  0.0020043   0.127 0.900338
x3           0.9718510  0.1773381   5.480 1.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2643 on 22 degrees of freedom
Multiple R-squared: 0.6522,     Adjusted R-squared: 0.6048
F-statistic: 13.75 on 3 and 22 DF, p-value: 2.874e-05
```

We observe that the estimated coefficients have not changed much and that is why the observation is not identified as an influential observation. The estimated standard deviations of the coefficients, however, have changed dramatically and x_2 is no longer significant. This illustrates that residual plots may provide information in addition to the leverages and the Cook's distances. A correct procedure would now be to try find out why observation number 14 is an outlier. Is it just a wrong recording? If a good reason is found the data are reanalyzed with the corresponding observation taken out or corrected.

To illustrate some other situations that may occur we show normal plots performed on data from a distribution with heavy tails and data from a skew distribution. If we have data from a distribution with lighter tails than the normal distribution and with expected value equal to 0, the form of the normal-plot would be an approximate mirror-image from the one from a heavy tailed distribution, mirrored around the line y=x.



Data from a distribution with heavy tails.



Data from a skew distribution

Other methods for choosing the best model

Calculating the PRESS-residuals (Predictive sum of squares) $\delta_i^* = y_i - y_{i,-i}$ is an example of doing leave one out cross validation. These residuals measure how good the i-th observation of the response can be estimated or predicted in terms of the other observations.

The model with the "best prediction ability" should be the one that minimizes $\sum_{i=1}^{n}\left|\delta_i^*\right|$ or

$\sum_{i=1}^{n}\left(\delta_i^*\right)^2$ =PRESS, eventually the one that maximizes $R_{pred}^2 = 1 - \dfrac{\sum_{i=1}^{n}\left(\delta_i^*\right)^2}{SS_T}$

s^2 is an estimate for $Var\left(Y_i\right)$ in the model with *p*-parameters and $\hat{\sigma}^2$ is an estimate for σ^2. Often it is the estimate of σ^2 from the full model.

Categorical or Indicator variables

Suppose

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 Z_i + \varepsilon_i$$

Where the variable Z is a categorical variable with l categories. Such regression problems are normally best solved by introducing $l-1$ indicator variables. For instance if we have three different catalyst the regression model can be:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 Z_{1i} + \beta_4 Z_{2i} + \varepsilon_i \quad i = 1, 2, \ldots, n$$

where the columns $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are

$\boldsymbol{Z}_1$	$\boldsymbol{Z}_2$
1	0
1	0
⋮	⋮
1	0
0	1
0	1
⋮	⋮
0	1
0	0
0	0
⋮	⋮
0	0

The $(1,0)$ combination corresponds to catalyst 1, the $(0,1)$ to catalyst 2 and the $(0,0)$ combination to catalyst 3.
The model is then:

$$Y_i = (\beta_0 + \beta_3) + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for catalyst 1}$$

$$Y_i = (\beta_0 + \beta_4) + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for catalyst 2}$$

and

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for catalyst 3 .}$$

If we believe the coefficients for the continuous variables depend on the level of the category variables, we introduce product terms. Suppose we construct the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 Z_{1i} + \beta_4 Z_{2i} + \beta_5 x_{1i} z_{1i} + \beta_6 x_{2i} z_{2i} + \varepsilon_i \quad i = 1, 2, \ldots, n$$

Then we have

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for catalyst 1}$$

$$Y_i = (\beta_0 + \beta_4) + \beta_1 x_{1i} + (\beta_2 + \beta_6) x_{2i} + \varepsilon_i \quad \text{for catalyst 2}$$

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for catalyst 3}$$

Orthogonal columns in the Design matrix

Let the design matrix $\boldsymbol{X} = \left[1, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\right]$.

If $\boldsymbol{x}_p^t \, \boldsymbol{x}_q = 0$ i.e. $\sum_{i=1}^{n} x_{pi} x_{qi} = 0, \; p \neq q$ we say that the columns $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ are orthogonal. If all the columns are orthogonal we have:

$$\left(\boldsymbol{X}^t \boldsymbol{X}\right) = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_1^t \boldsymbol{x}_1 & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_k^t \boldsymbol{x}_k \end{bmatrix} \text{ and}$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t \boldsymbol{Y} = \begin{bmatrix} 1/n & 0 & \cdots & 0 \\ 0 & \left(\boldsymbol{x}_1^t \boldsymbol{x}_1\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\boldsymbol{x}_k^t \boldsymbol{x}_k\right)^{-1} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \boldsymbol{x}_1^t \boldsymbol{Y} \\ \vdots \\ \boldsymbol{x}_k^t \boldsymbol{Y} \end{bmatrix}.$$

We observe that the estimator for β_j only depends upon $\boldsymbol{x}_j$ and $\boldsymbol{Y}$. In addition, we get

$$SS_R = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(b_0 + b_1 x_{1i} + \cdots b_k x_{ki} - b_0\right)^2 = b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2 \text{ or}$$

$$SS_R\left(\beta_0, \beta_1, \cdots, \beta_k\right) = SS_R\left(\beta_1 \middle| \beta_0\right) + SS_R\left(\beta_2 \middle| \beta_0\right) + \cdots + SS_R\left(\beta_k \middle| \beta_0\right) \text{ where}$$

$$SS_R\left(\beta_j \middle| \beta_0\right) = b_j^2 \sum_{i=1}^{n} x_{ji}^2, \; j = 1, 2, \ldots, n.$$

Regression analysis can easily fill a whole course or more. In cases where you find that more knowledge is needed than what is covered here you may benefit from checking out one or more of these books:

Draper, N. R. & Smith, H. Applied Regression Analysis. Third edition.
Montgomery, D. C. Peck, E. A. & Vining, G. G. Introduction to Linear Regression Analysis, Third Edition.
Abraham, B & Ledolter, J. Introduction to Regression Modeling

If you want to to increase your theoretical skills these books may be good choices:

Rencher, A. C. & Schaalje, G. B. Linear Models in Statistics.
Seber, G. A. F. Linear Regression Analysis