Linear Regression

Janez Stare

Faculty of Medicine, Ljubljana, Slovenia

Ljubljana, 2017

Simple (bivariate) linear regression

When we want to describe a relationship between two numerical variables we usually use regression analysis (name explained later). The data consist of pairs of observations (x_i, y_i), where x_i are values of variable X and y_i are values of variable Y.

This is the standard terminology

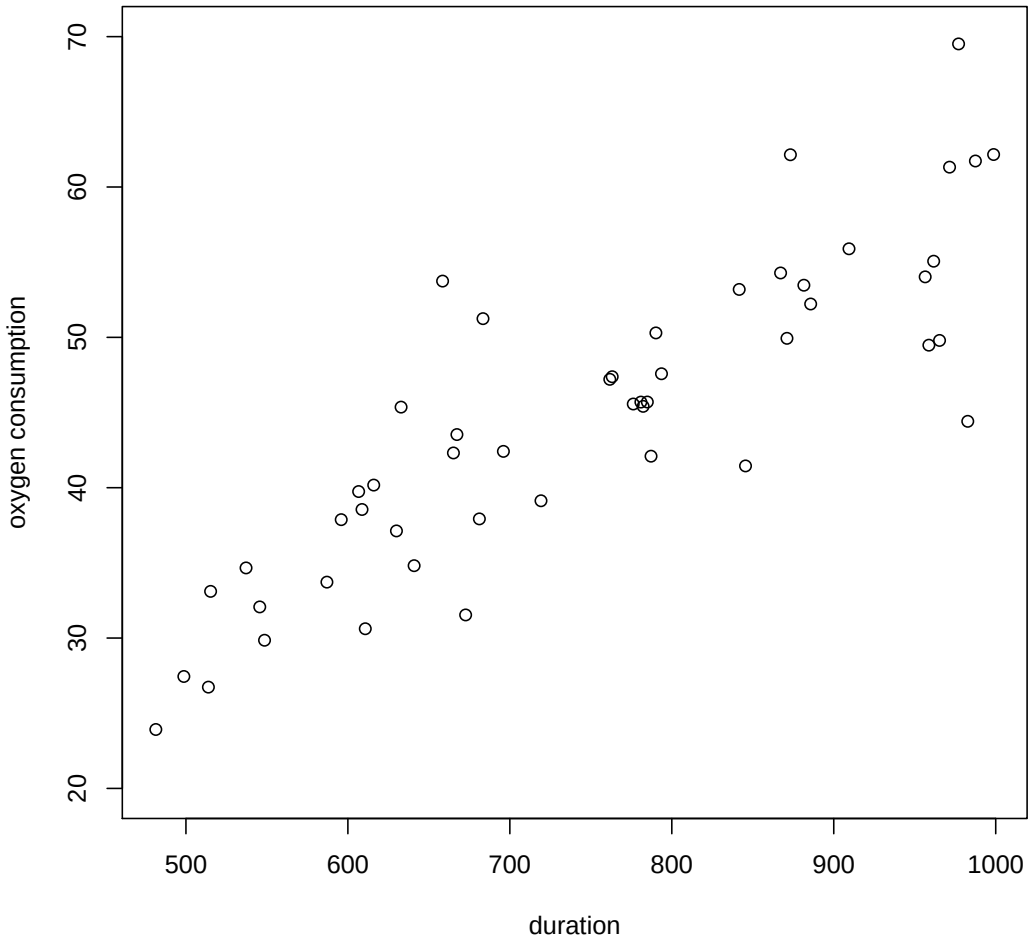
$$X = \text{independent variable (prognostic variable, covariate, } \ldots)$$
$$Y = \text{dependent variable (outcome)}$$

Scattergram (or scatterplot)

It is always a good idea to plot the points.

Example: In a treadmill test 50 men were asked to run until exhaustion and the **duration** of the test and the **oxygen consumption** (per minute and kilogram body weight) were measured. On a scattergram on the next slide we see that oxygen consumption increases with duration approximately linearly.

Scattergram (or scatterplot)



Goals of statistical analysis

1. Describe and evaluate association between the variables.
2. Predict Y, if we know X.

Since we will discuss only linear associations, we will talk about **linear regression** (and still wonder why we say regression).

Statistical model for association is then

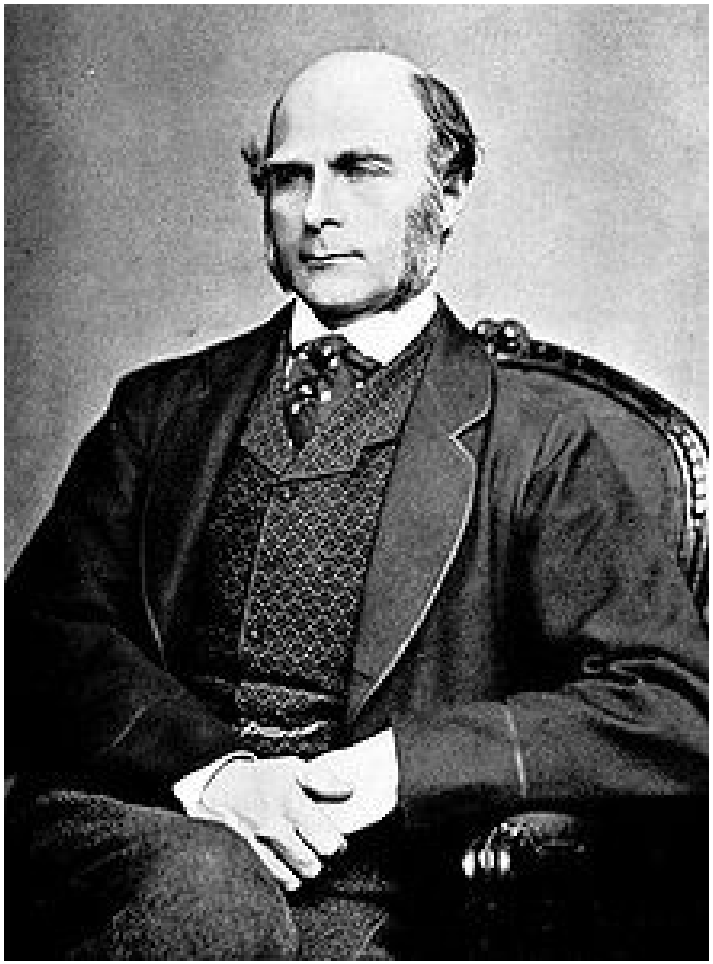
$$Y = \alpha + \beta X + \epsilon,$$

where 'the error' ϵ represents random variation around the line.

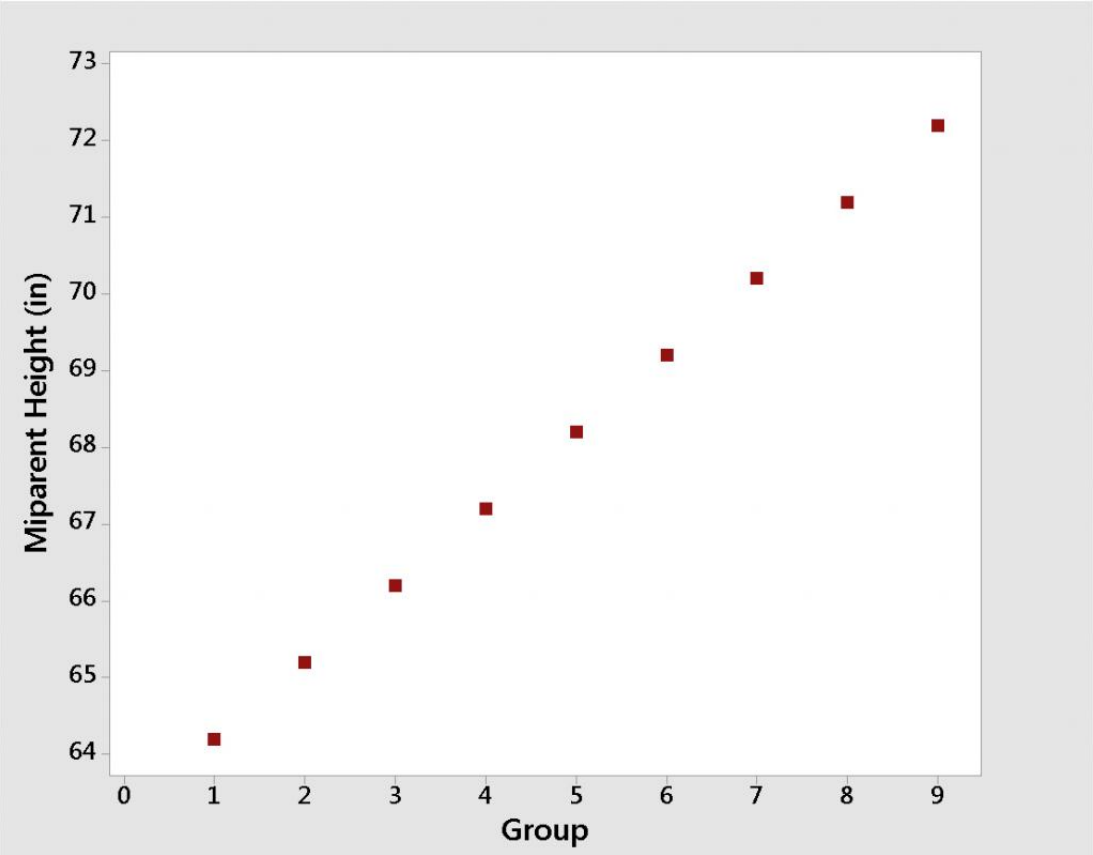
Note: Limiting ourselves to linear model is not really a restriction!

How Regression Got Its Name

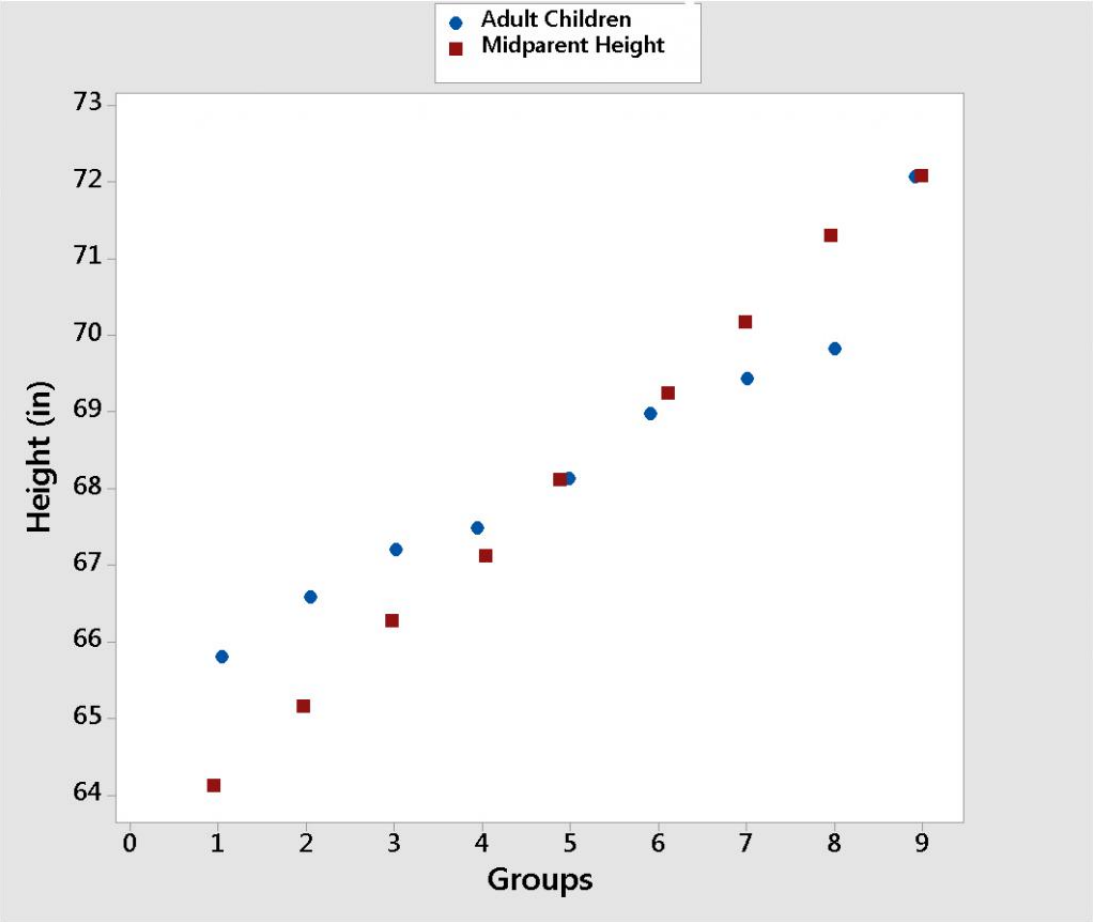
Sir Francis Galton



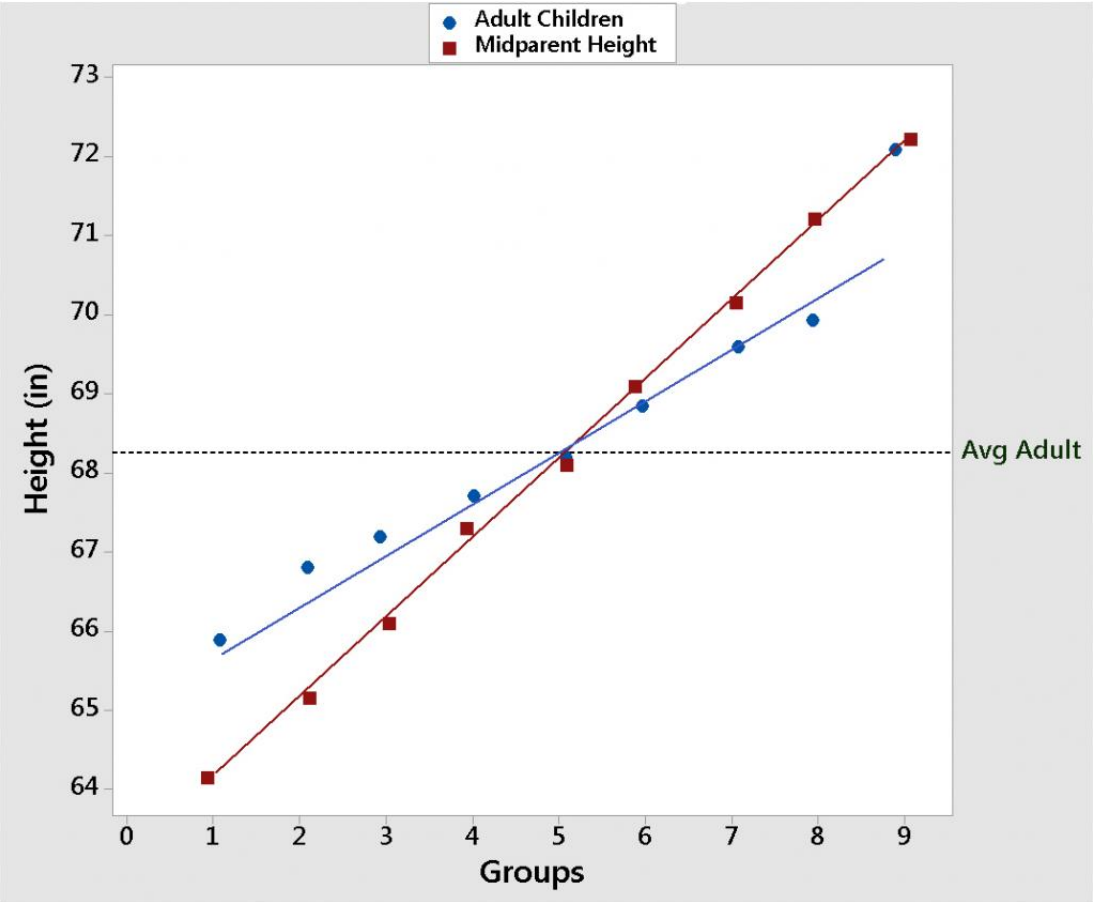
How Regression Got Its Name



How Regression Got Its Name



How Regression Got Its Name



Assumptions

We will require that the observations are independent and errors are normally distributed with mean 0 and a given variance, so

$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$

Putting it differently, our model assumes that for a given x the outcome Y is normally distributed with (conditional) mean

$$E(Y|x) = \alpha + \beta x$$

and (conditional) variance

$$Var(Y|x) = \sigma^2.$$

Assumptions cont.

Since the variance does NOT depend on x, we are actually assuming that the variability around the line is the same everywhere. This property is called **homoscedasticity**. If the condition is not satisfied, we talk about heteroscedasticity.

Assumptions summary

Our statistical model has **four assumptions**. Here they are again:

1. Observations are **independent**.
2. The regression function $E(Y|x)$ is **linear**.
3. The values of Y vary around the line with a **constant variance** (homoscedasticity).
4. The values of Y for a given x are **normally distributed**.

The appropriatness of these assumptions should always be **checked**.

Estimation of parameters

Linear regression model has three unknown parameters: the constant α, the coefficient β and the variance σ^2. Our first goal after collecting the data is to estimate the regression line, and here we have the following question:

How do we choose α and β, which line is the best?

Estimation of parameters - criterium

There are different criteria for the ‘best’ line, the most common is this one:

For all values of x we look at the differences between the predicted (lying on the line) and observed values of Y, and we require that **the sum of squares of those differences** be minimal.

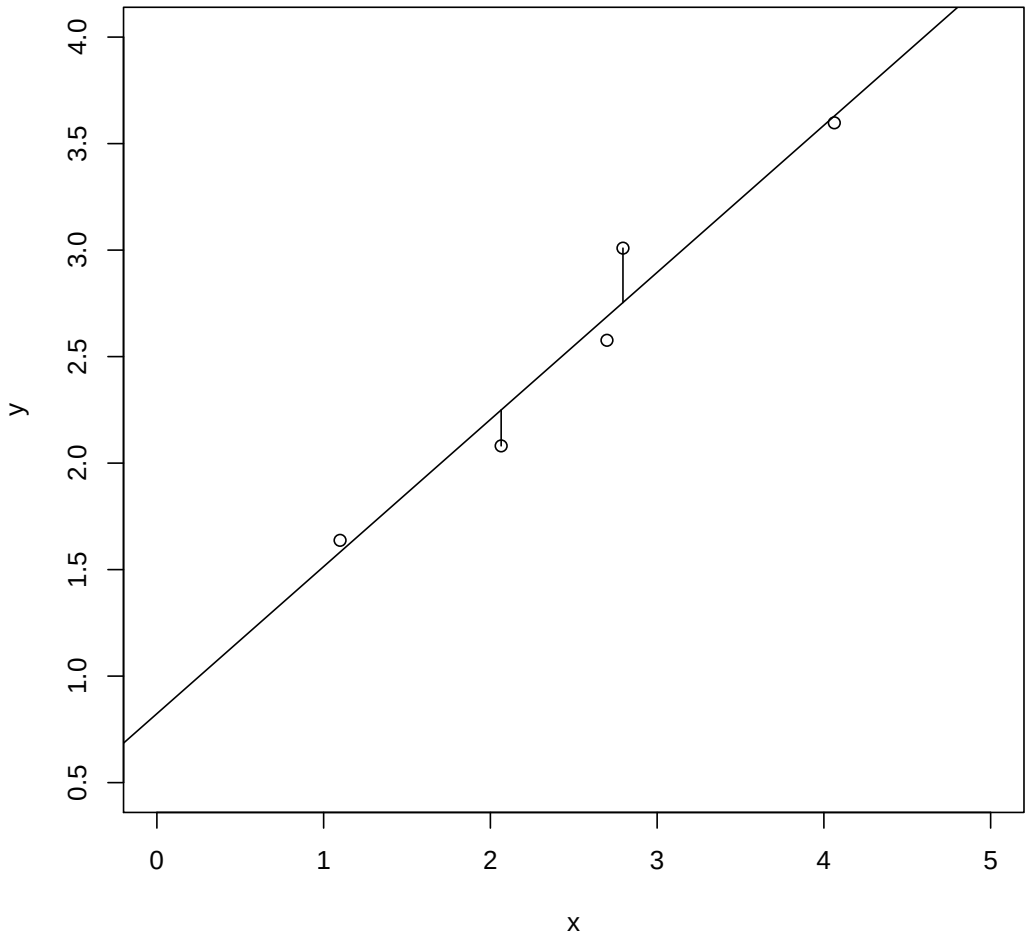
Estimation of parameters - criterium

We then have to minimize the sum

$$SS(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Here *SS* stands for sum-of-squares.

Estimation of parameters - criterium illustration



Maximum likelihood estimation

Under our assumptions, the density of Y for a given value x is

$$f(y,x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(y-\alpha-\beta x)^2}{\sigma^2})$$

The likelihood of given data is then

$$L(\alpha,\beta) = \frac{1}{\sqrt{2\pi}\sigma} \prod_{i=1}^{n} \exp(-\frac{1}{2}\frac{(y_i-\alpha-\beta x_i)^2}{\sigma^2})$$

We see that maximizing ln L is the same as minimizing $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$

Estimates

We need to minimize a function of two variables (α and β), which is easily done if we know enough mathematics. We get:

$$\hat{\beta} = \frac{\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sum_i (x_i - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

The estimated line is

$$y = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x}),$$

from where we see that the line passes through the point $(\bar{x},\bar{y})$.

Estimation of the variance σ^2

The observed values y_i vary around $\hat{y}_i$, the differences

$$r_i = y_i - \hat{y}_i$$

are called the **residuals**. These are in fact **estimated errors**. The variance of the residuals is estimated by

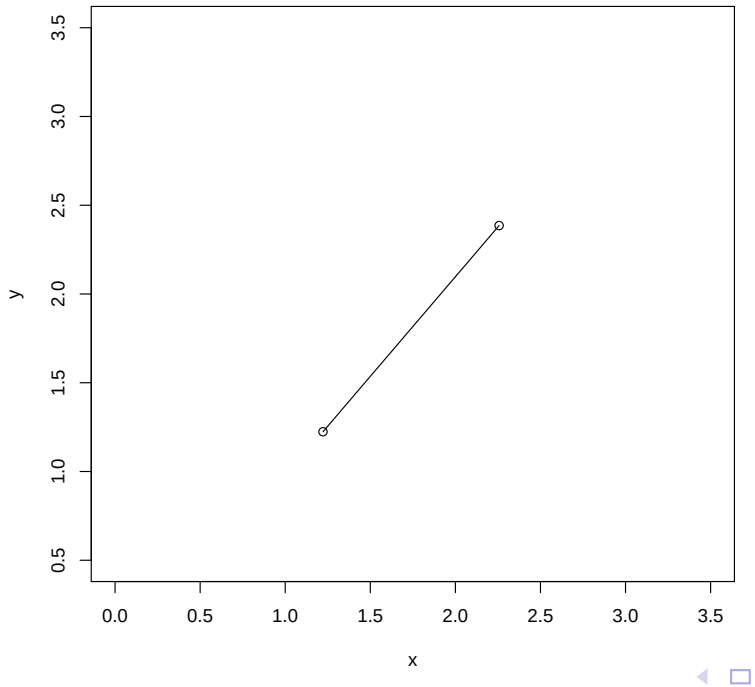
$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_{Res}}{n-2},$$

where SS_{Res} denotes the sum of squared residuals. This is of course the same as the sum that we minimized above to obtain $\hat{\alpha}$ and $\hat{\beta}$, except that we did not speak of residuals then.

Estimation of the variance σ^2 - why division by $n - 2$?

Formally we can say that this way the estimator is unbiased (and we can even prove it).

Intuition: if we only have two points, the regression line will go through them and no estimation of variance will be possible. We therefore need more points and only those can be used to estimate the variance (not quite true, but I hope you get the point).



A look at the residuals

$$\begin{aligned} r_i = y_i - \hat{y}_i &= y_i - \hat{\alpha} - \hat{\beta}x_i = y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i \\ &= (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}) \end{aligned}$$

Taking squares

$$\begin{aligned} (y_i - \hat{y}_i)^2 &= (y_i - \bar{y})^2 + \hat{\beta}^2(x_i - \bar{x})^2 - 2\hat{\beta}(x_i - \bar{x})(y_i - \bar{y}) \\ &= (y_i - \bar{y})^2 + \hat{\beta}(x_i - \bar{x})[\hat{\beta}x_i - \hat{\beta}\bar{x} - 2y_i + 2\bar{y}] \\ &= (y_i - \bar{y})^2 + \hat{\beta}(x_i - \bar{x})[\hat{\beta}x_i - \hat{\beta}\bar{x} - 2(\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i) + 2\bar{y}] \\ &= (y_i - \bar{y})^2 - \hat{\beta}^2(x_i - \bar{x})^2 \end{aligned}$$

Example: fit to treadmill test data

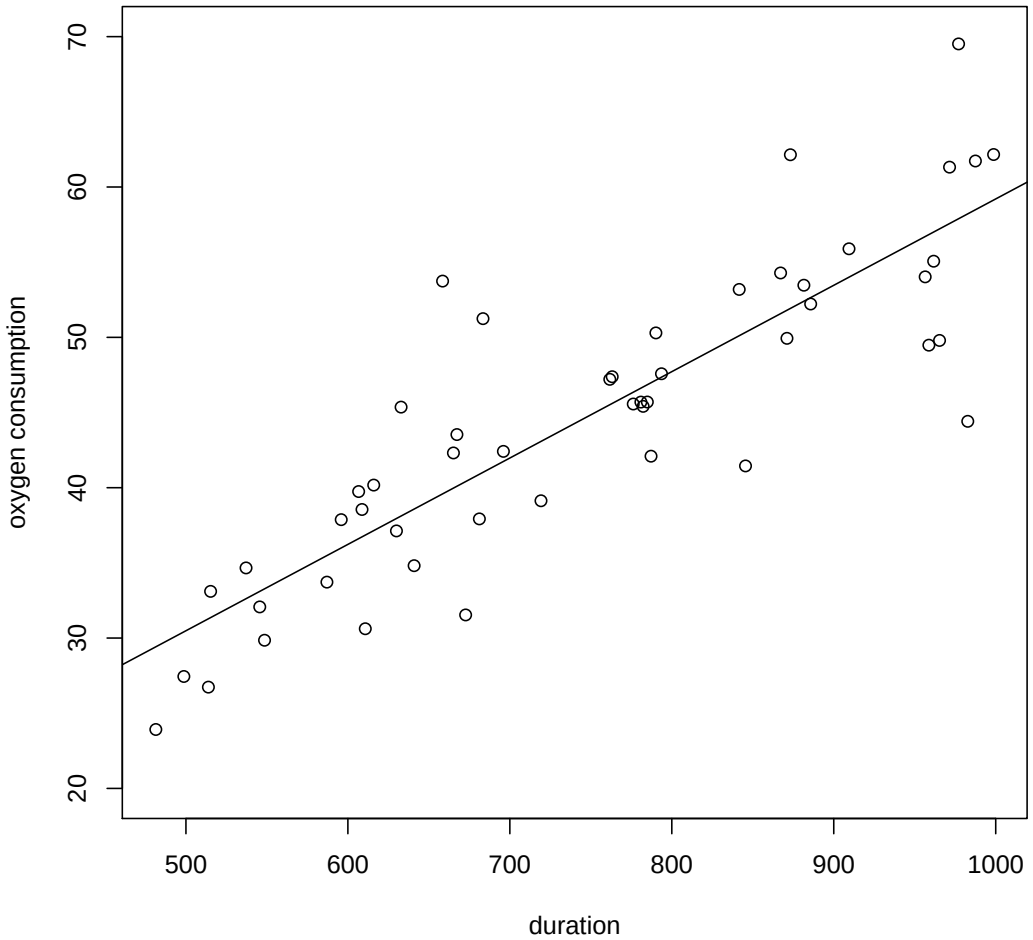
Applying the formulas for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}$ to data on treadmill test, we get

$$\hat{\alpha} = 1{,}765, \quad \hat{\beta} = 0{,}057, \quad \hat{\sigma} = 5{,}348.$$

We can calculate the estimated values $\hat{y}_i$ for given x_i using the equation

$$\hat{y}_i = 1{,}765 + 0{,}057 \cdot x_i.$$

Example: graph of the fit to treadmill test data



Interpretation

1. α is of course the value on the line when x (duration in our case) is equal to 0. Such values rarely make sense which is why α is usually not very interesting. But we need it to calculate y for a given x.
2. The coefficient β is much more important. Let's calculate the estimated $\hat{y}$s for two x values which are 1 unit apart.

$$\hat{y}(x) = \alpha + \beta \cdot x$$

$$\hat{y}(x+1) = \alpha + \beta \cdot (x+1) = \hat{y}(x) + \beta$$

So

$$\hat{y}(x+1) - \hat{y}(x) = \beta.$$

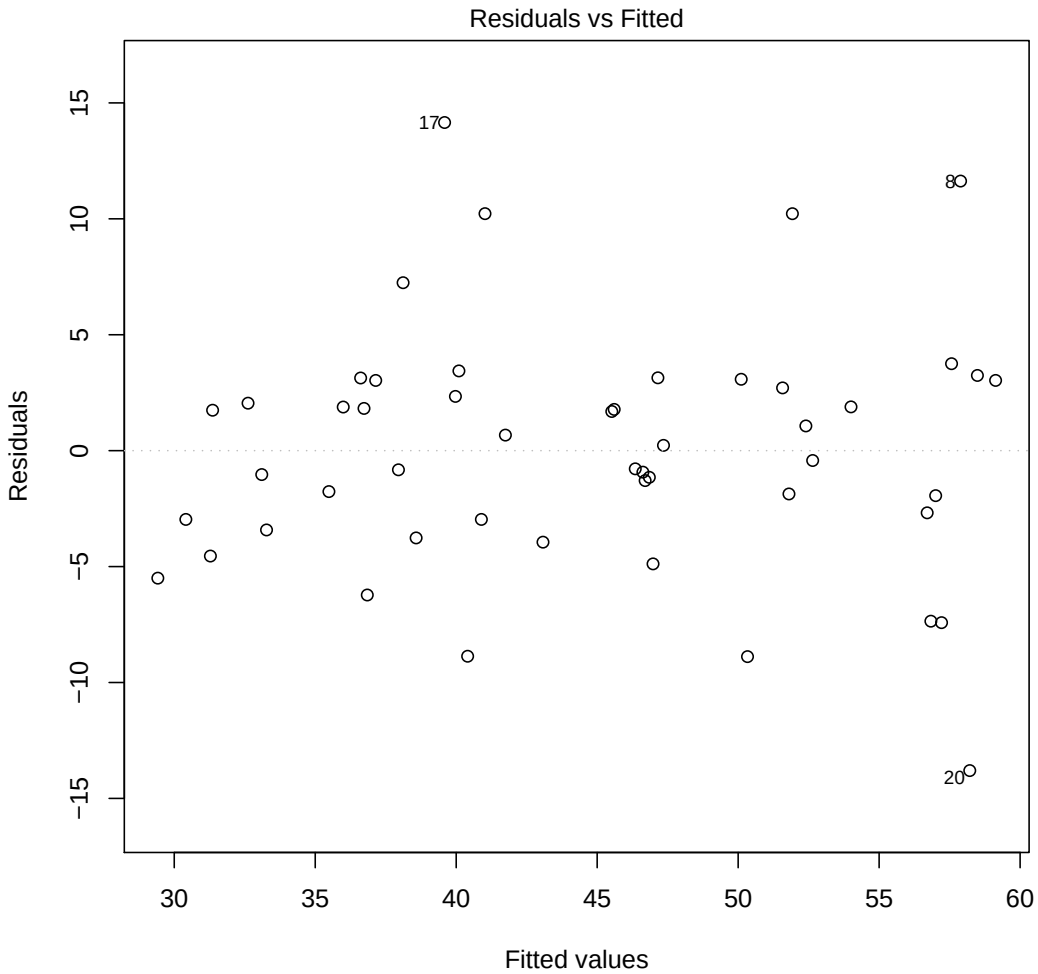
Interpretation cont.

3. In our example X (duration) is measured in seconds so that β represents a change in oxygen consumption (Y) when the duration changes for one second. It is not surprising that β is small. It would make more sense if we knew what is the increase in Y if duration increases for a minute. We get $60 \cdot \beta = 3{,}45$ ml/kg/min.
4. The standard deviation σ describes variability around the regression line. Since we assumed the normal distribution, we can calculate that in our example 95% of all the values of the oxygen consumption falls in the interval $(-1{,}96 \cdot 5{,}348, 1{,}96 \cdot 5{,}348)$ $(-10{,}48, 10{,}48)$ ml/kg/min around the line.

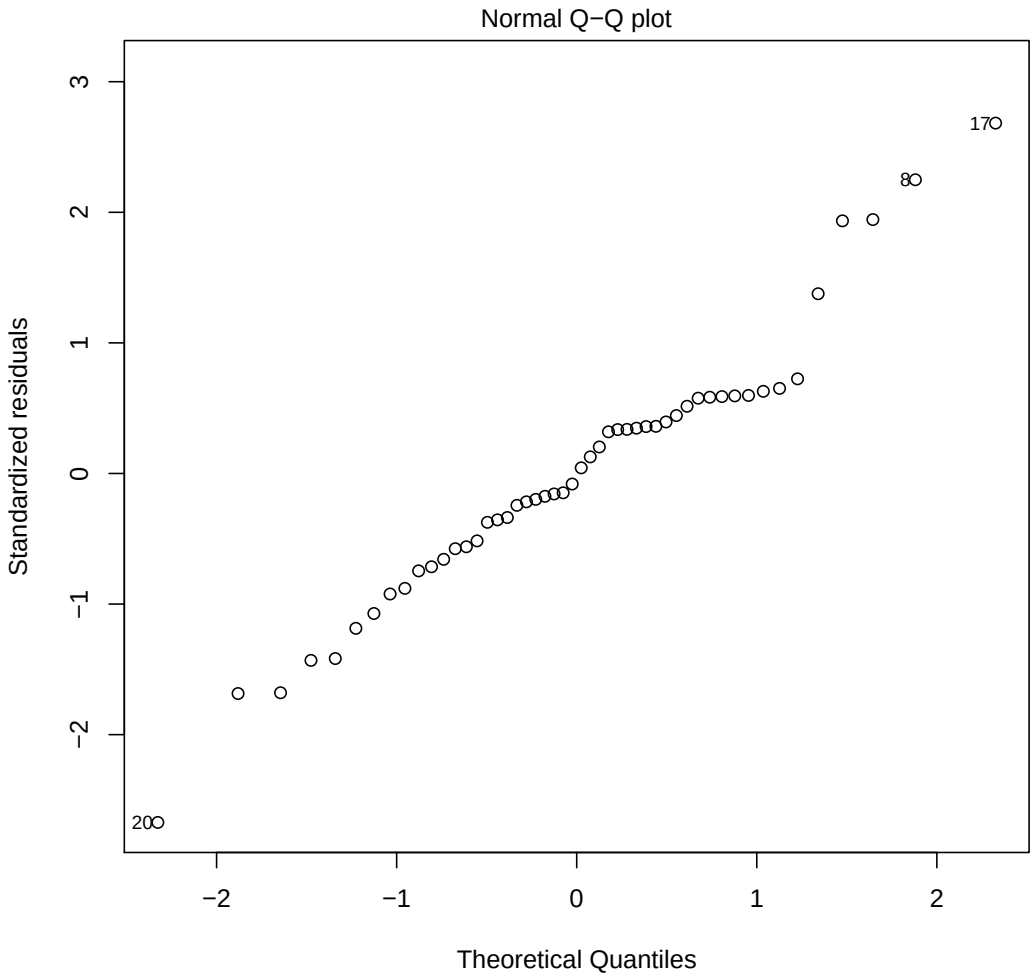
Checking assumptions of the model

1. If the model is correct **the residuals should be symmetrically distributed around the regression line with a constant variance**. The graph of residuals r_i with respect to the predicted values $\hat{y}_i$ should reflect this.
2. The normality of residuals can be checked in different ways, Q-Q plots are one option.

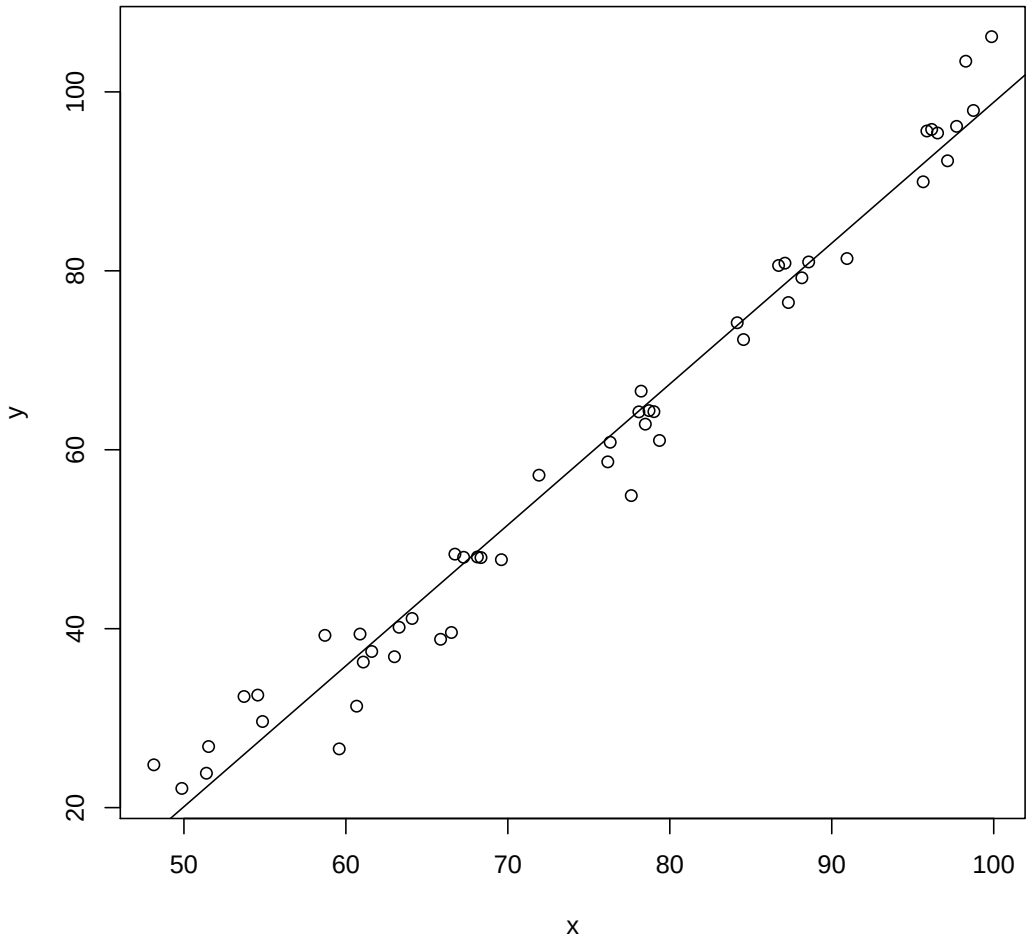
Graph of the residuals



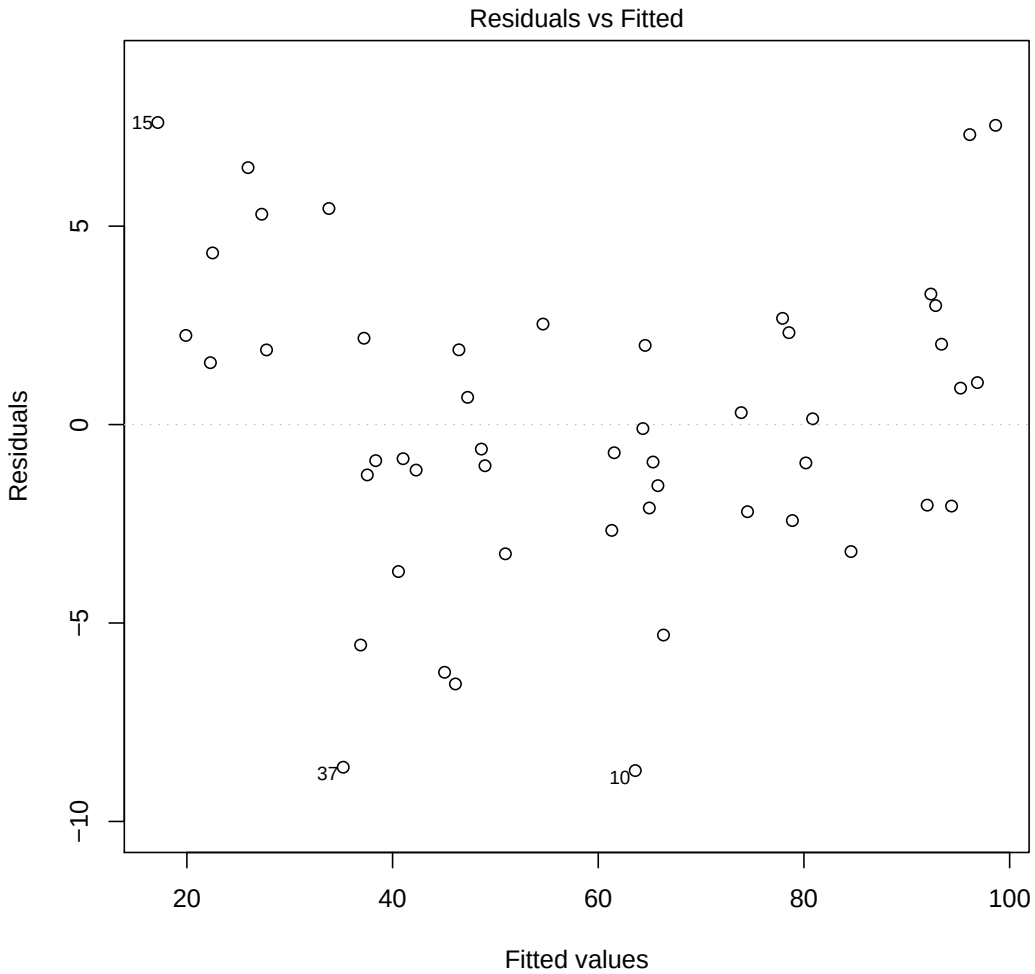
Q-Q plot


Normal Q−Q plot
Standardized residuals
Theoretical Quantiles
3
2
1
0
−1
−2
−2
−1
0
1
2
170
8
20

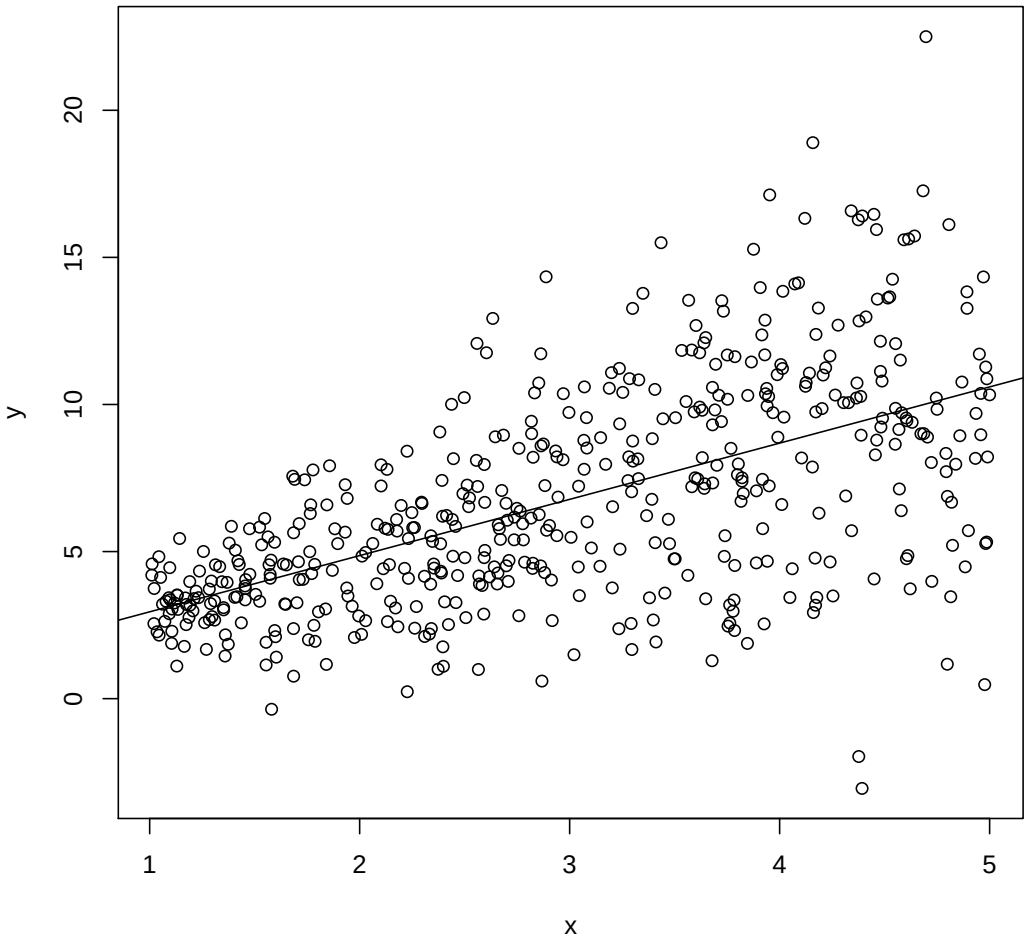
Nonlinear association



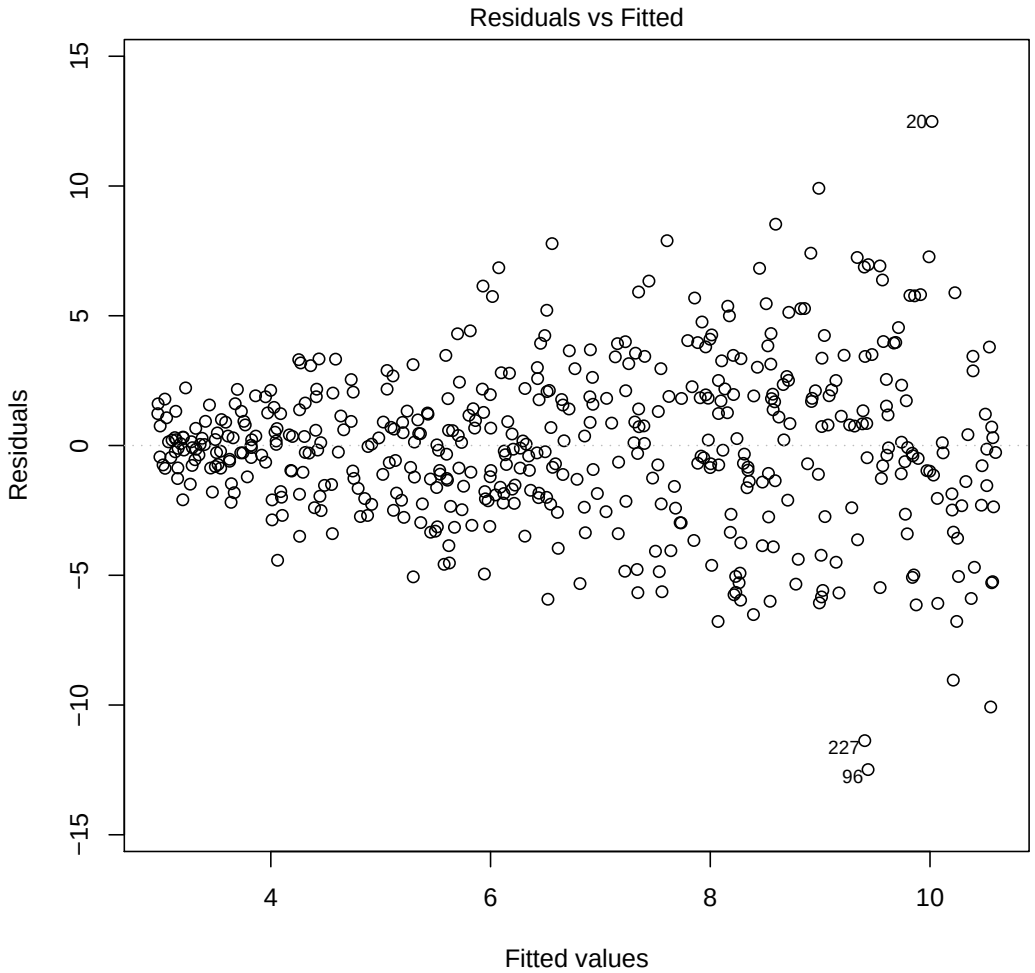
Nonlinear association - graph of the residuals



Heteroscedasticity



Heteroscedasticity - graph of the residuals



Sampling error of the regression coefficient

Different samples will give us different estimates of the regression coefficient and we must ask how they vary. We are especially interested in the slope which tells us if there is any association between X and Y.

Remember that

$$\hat{\beta} = \frac{\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sum_i (x_i - \bar{x})^2}$$

The numerator can be rewritten like this

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i y_i (x_i - \bar{x}) - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i y_i (x_i - \bar{x}),$$

since $\sum_i (x_i - \bar{x}) = 0$.

Sampling error of the regression coefficient

Now assume that sampling is repeated with x_i fixed, so that only values of Y randomly. Then we have

$$\begin{aligned} var(\hat{\beta}) &= var\left[\frac{\sum_i Y(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 var(Y)}{\left[\sum_i (x_i - \bar{x})^2\right]^2} \\ &= \frac{var(Y)}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}. \end{aligned}$$

Under the assumptions of the distribution of the residuals we then have

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

Testing the hypothesis about the regression line

To calculate the variance of $\hat{\beta}$ from the previous formula, we need to know σ. And since we usually don't, we replace by the estimate $\hat{\sigma}$. This changes the distribution of $\hat{\beta}$ from the normal into t with $n-2$ degress of freedom. Test of the hypothesis

$$H_0 : \beta = \beta_0$$

is then

$$t_{\beta=\beta_0} = \frac{\hat{\beta} - \beta_0}{\sqrt{var(\hat{\beta})}}, \qquad \text{sp} = n - 2.$$

By far the most often tested hypothesis is that $\beta = 0$.

Example cont.: fit to treadmill test data

For the null hypothesis $H_0 : \beta = 0$ we get $t = 11{,}593$ with 48 degrees of freedom, p value is $1{,}613 \cdot 10^{-16}$. The null hypothesis is easily rejected.

Decomposition of the total variation

Say we measured the outcome Y and the covariate X on n units. **Total variation** of the outcome can be described by the sum

$$SS_{tot} = \sum_i (y_i - \bar{y})^2,$$

where SS stands for (**S**um of **S**quares). This sum represents variation due to biological diversity as well as due to different values of X.

Decomposition of the total variation

The question is:

What proportion of the total variation of Y is due to variation of X?

Or

How much of a variation we would see if all units had the same values of X?

Decomposition of the total variation

We already know

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2$$

or, after rewriting

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{\beta} x_i - \hat{\beta}\bar{x})^2.$$

And since

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

and

$$\bar{y} = \hat{\beta}\bar{x} + \hat{\alpha}$$

we can write the second term on right as $\sum_i (\hat{y}_i - \bar{y})^2$ and we have

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2.$$

Decomposition of the total variation

We usually write

$$SS_{tot} = SS_{res} + SS_{reg},$$

and it means that total variation was decomposed into residual variation and variation due to regression. It should be rather obvious that the ratio

$$\frac{SS_{reg}}{SS_{ost}}$$

will be small if the **null hypothesis**, that $\beta = 0$, holds.

Testing the null hypothesis - again

What is small and what is large (meaning that it goes against the null hypothesis) is a question to which a theory has the answer. And the theory says that the ratio

$$F = \frac{SS_{reg}/1}{SS_{ost}/(n-2)} \tag{1}$$

has a $\mathcal{F}$ distribution with 1 and $n-2$ degrees of freedom. The p value is then

$$p \text{ value} = P(\mathcal{F}(1,n-2) \geq F)$$

We will reject the null hypothesis when the variation due to regression will be large compared to the variation of the residuals.

Testing the null hypothesis - again

Results are usually presented in the so called analysis of variance table (ANOVA).

Source	df	SS	MS	F	Significance
Regression	1	SS_{reg}	$SS_{reg}/1$	F	p
Residuals	$n-2$	SS_{res}	$SS_{res}/(n-2)$		
Total	$n-1$	SS_{tot}			

Primer: Treadmill test

Source	df	SS	MS	F	Significance
Regression	1	3843.5	3843.5	134.40	1.613e-15
Residuals	48	1372.7	28.6		
Total	49	5216.2			

A measure of explained variation in linear regression

We had

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2.$$

The statistics

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_{reg}}{SS_{totl}},$$

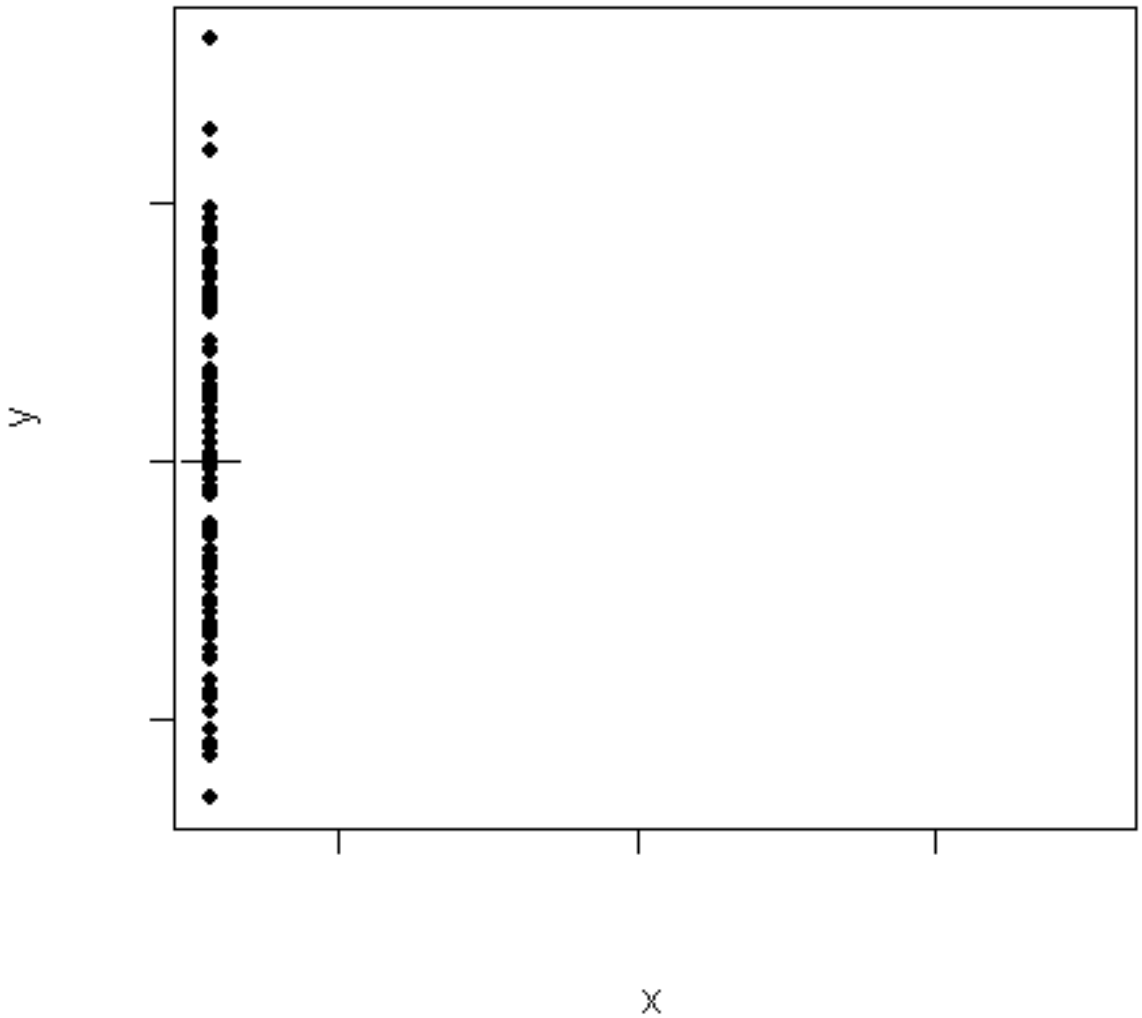
is then a proportion of the total variation explained by the model.

Example: Treadmill test-cont

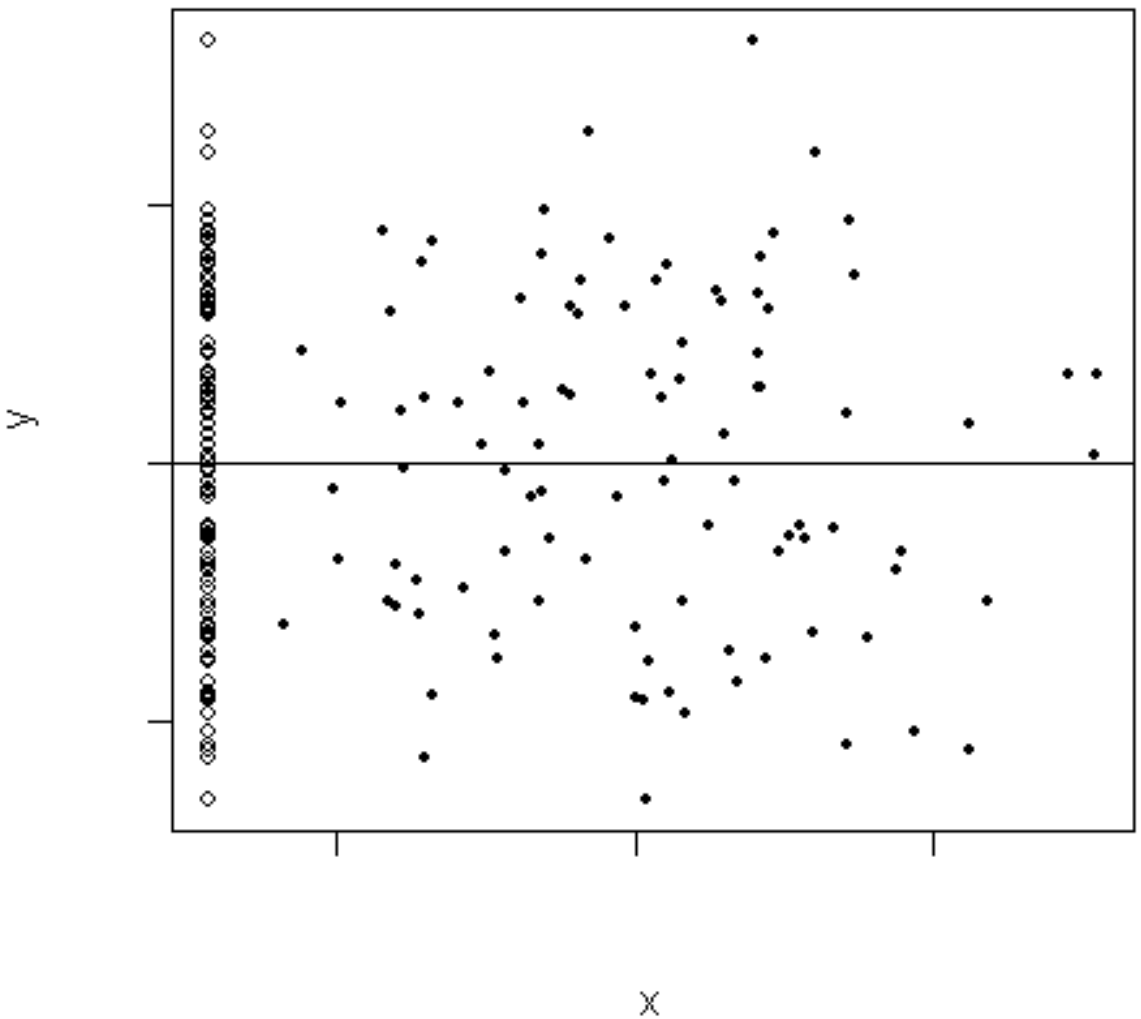
R^2 is 0,737, meaning that approximately 74% of the variation in oxygen consumption was explained by duration of the test.

The next 4 slides illustrate the meaning of R^2

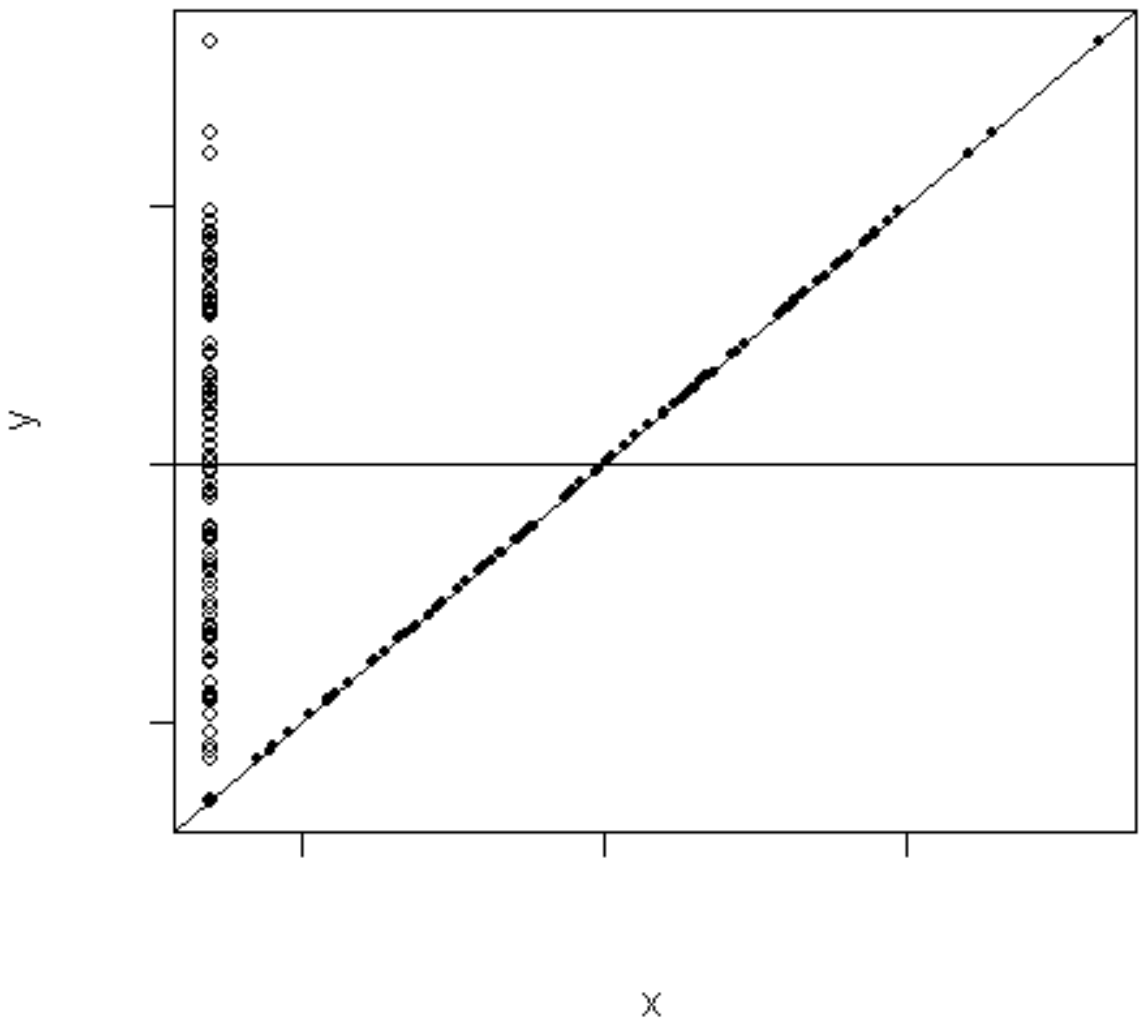
We know this about Y if we know nothing about X



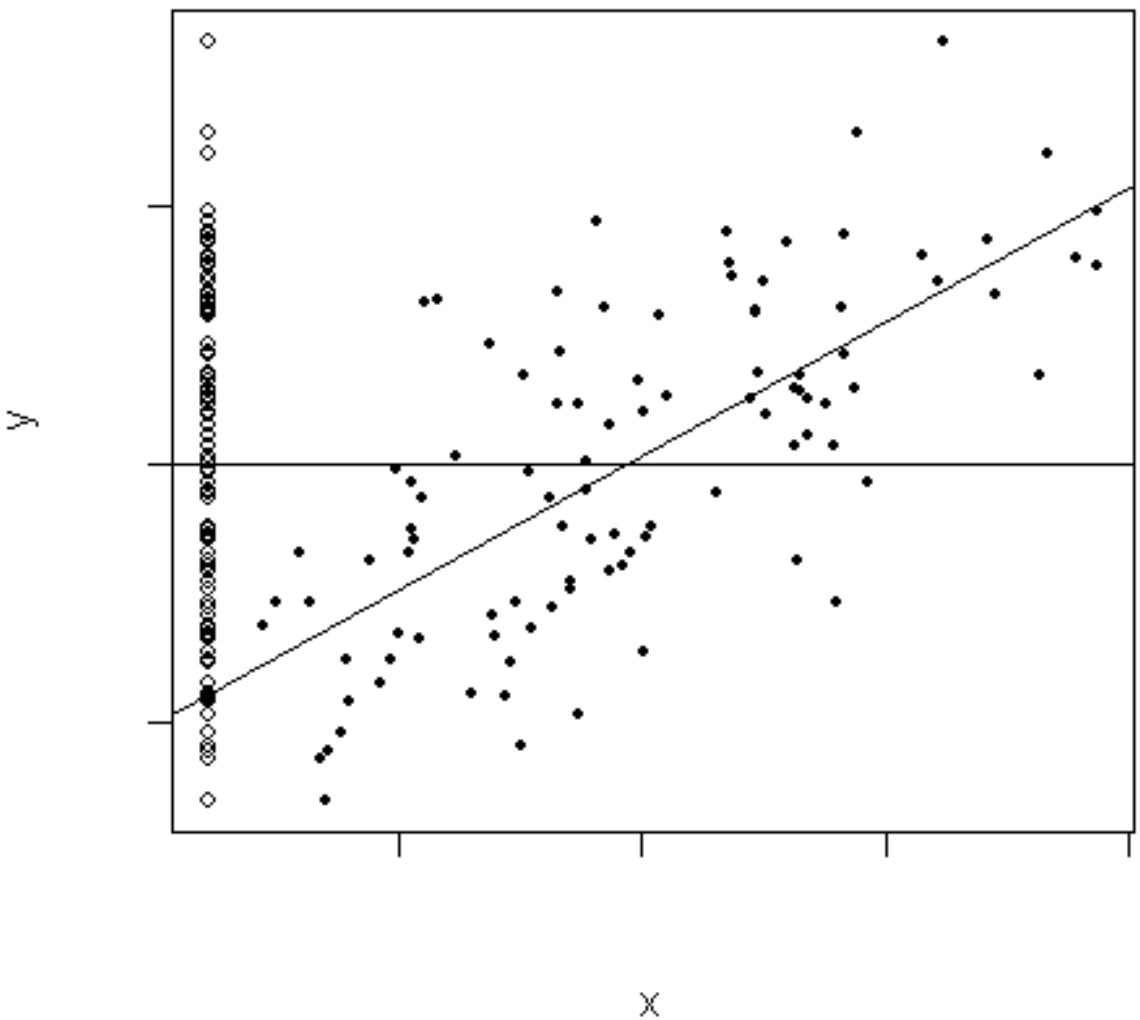
We know this about Y, if knowing X doesn't change anything



We know this about Y, if there is a perfect correlation with X



And we know this about Y, if knowing X tells something about Y



Multiple linear regression - just a few lines

Often we want to have more than one independent variable. A natural extension of the bivariate regression model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where again the 'the error' ϵ represents random variation around the hyper plane.

Multiple linear regression - just a few lines

All the assumptions remain the same, and the method of estimation also does not change (except that we now have a larger system of equations to solve).

An unbiased estimator of the variance of the error term is given by

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - k - 1} = \frac{SS_{Res}}{n - k - 1}$$

Multiple linear regression - just a few lines

Interpretation of coefficients in the model is the same as in bivariate case.

But there is a difference regarding the null hypothesis. We can now look at different null hypotheses.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_k + \beta_k X_k + \epsilon,$$