Correlation and regression analyis

We are looking for a relationship between two or more variables.

- Is there a relationship between BMI and time spent with watching tv in children aged 7 to 10 years?

Correlation and regression analyis

We are looking for a relationship between two or more variables.

- Is there a relationship between BMI and time spent with watching tv in children aged 7 to 10 years?

- Is the linear function appropriate to describe this relationship? If yes, what are the parameters of the function? If not, what other function should be used?

Correlation and regression analyis

We are looking for a relationship between two or more variables.

- Is there a relationship between BMI and time spent with watching tv in children aged 7 to 10 years?
- Is the linear function appropriate to describe this relationship? If yes, what are the parameters of the function? If not, what other function should be used?
- Is this picture changing if we consider further explanatory variables? (BMI of the parents, time spent with sport, etc.)

Correlation and regression analyis

We are looking for a relationship between two or more variables.

- Is there a relationship between BMI and time spent with watching tv in children aged 7 to 10 years?

Correlation analysis

- Is the linear function appropriate to describe this relationship? If yes, what are the parameters of the function? If not, what other function should be used?

Regression analysis

- Is this picture changing if we consider further explanatory variables? (BMI of the parents, time spent with sport, etc.)

Correlation and regression analyis

We are looking for a relationship between two or more variables.

- Is there a relationship between BMI and time spent with watching tv in children aged 7 to 10 years?

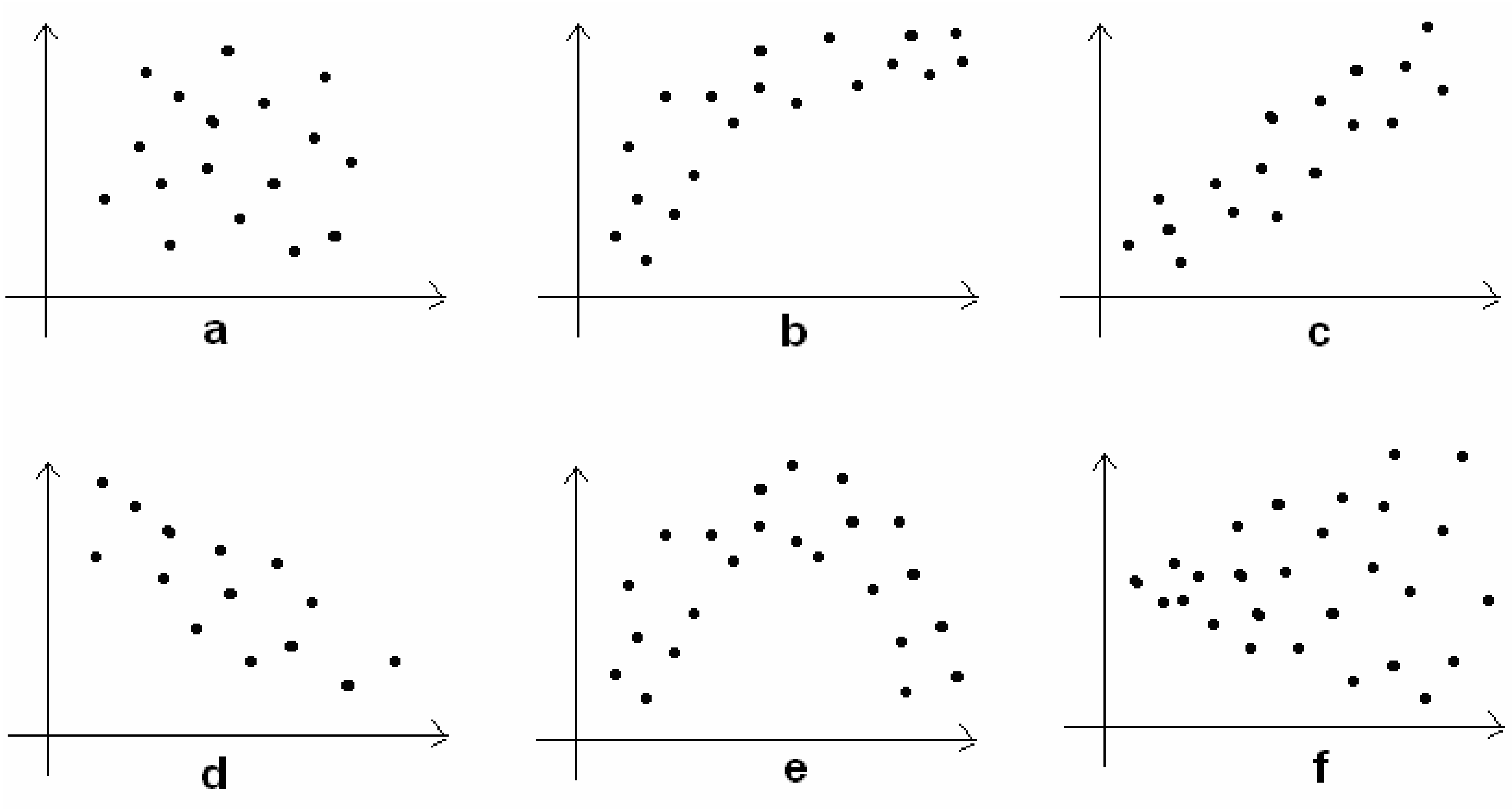
Correlation analysis

- Is the linear function appropriate to describe this relationship? If yes, what are the parameters of the function? If not, what other function should be used?

Testing *Regression analysis* *Estimation*

- Is this picture changing if we consider further explanatory variables? (BMI of the parents, time spent with sport, etc.)

Note that there are relationships other than correlation!

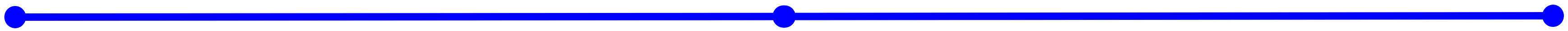


a – no association, b – positive nonlinear correlation,
c – positive linear correlation, d – negative linear correlation,
e, f – non-monotonic relationships (association but no correlation)

Correlation coefficients

…quantify how strong a correlation exists between *X* and *Y*.

The traditional setting:



–1
strongest negative correlation

0
no correlation

1
strongest positive correlation

Ordering of the subjects according to their *X* values is exactly the inverse of their ordering according to their *Y* values

Independence of *X* and *Y* implies zero correlation (but zero correlation does not imply independence)

Ordering of the subjects according to their *X* values is fully identical with their ordering according to their *Y* values

„Correlation between time spent with wathching tv and time spent with sport was –0.23.”

What is missing here?

„Correlation between time spent with wathching tv and time spent with sport was –0.23.”

What is missing here?

Pearson's correlation coefficient is the one that is used most frequently

- Insensitive to nonlinear relationships
- Outliers have serious impact on it
- Even monotonic scale transformations of data may change the correlation (except linear transformations)
- Classical significance test works only for normally distributed variables (distribution-free test is possible by bootstrapping)

Two alternatives of Pearson's correlation coefficient:

Spearman's rank correlation coefficient

- Pearson's coefficient applied to the ranks
- Sensitive to nonlinear relationships as well
- Outliers have less influence on it
- Invariant to monotonic scale transformations

Two alternatives of Pearson’s correlation coefficient:

Spearman’s rank correlation coefficient

- Pearson’s coefficient applied to the ranks
- Sensitive to nonlinear relationships as well
- Outliers have less influence on it
- Invariant to monotonic scale transformations

Kendall’ tau

- Sensitive to nonlinear relationships as well
- Invariant to monotonic scale transformations
- Outliers have even less influence on it

Regression analysis

A causal relationship?

We are looking for a **functional relationship** between one or more **explanatory variables** x_1, x_2, etc. and a **dependent variable** y.

Regression analysis

A causal relationship?

We are looking for a **functional relationship** between one or more **explanatory variables** x_1, x_2, etc. and a **dependent variable** y.

It is assumed that the explanatory variables do not determine y completely, it has a random component as well.

Regression model: f is the function ε is the random component

$$y = f(x_1, x_2, \ldots) \quad + \quad \varepsilon$$

or assuming that the function is linear:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \varepsilon$$

Regression analysis

A causal relationship?

We are looking for a **functional relationship** between one or more **explanatory variables** x_1, x_2, etc. and a **dependent variable** y.

It is assumed that the explanatory variables do not determine y completely, it has a random component as well.

Regression model: *f is the function* *ε is the random component*

$$y = f(x_1, x_2, \ldots) \quad + \quad \varepsilon$$

or assuming that the function is linear:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \varepsilon$$

Explanatory variables are not random variables in this model!

Steps of regression modeling

1. Intuitive, informal model. What depends on what, how to measure them, what are the random components?
2. Aims. Scientific finding (demonstrating significance)? Or prediction? What precision is required?
3. What do we know about the relationship from the literature? Linear? Logarithmic? Monotonic et all?
4. Look at graphical representations of data! Are there outliers? How can we explain them? What type of function do the data suggest?
5. Select variables and function type, and carry out the analysis!
6. Check the results! Is the regression significant? (*p*-values) Satisfied with model fit? (R^2) Do the applicability conditions hold? (residuals) If something is wrong, back to 5!

Differences between correlation and regression analysis

- Correlation analysis assumes a **symmetric relationship** between x and y while regression analysis assumes a **directional relationship** $x \rightarrow y$.

Differences between correlation and regression analysis

- Correlation analysis assumes a **symmetric relationship** between *x* and *y* while regression analysis assumes a **directional relationship** $x \to y$.

- Correlation analysis assumes that **both of *x* and *y* are random variables** while in regression analysis ***x* is not assumed to be a random variable**.

Differences between correlation and regression analysis

- ❖ Correlation analysis assumes a **symmetric relationship** between *x* and *y* while regression analysis assumes a **directional relationship** $x \to y$.

- ❖ Correlation analysis assumes that **both of *x* and *y* are random variables** while in regression analysis ***x* is not assumed to be a random variable**.

Correlation analysis does not make sense if x is set by the experimenter (e.g. doses of a drug)!

Differences between correlation and regression analysis

- Correlation analysis assumes a **symmetric relationship** between *x* and *y* while regression analysis assumes a **directional relationship** $x \rightarrow y$.

- Correlation analysis assumes that **both of *x* and *y* are random variables** while in regression analysis ***x* is not assumed to be a random variable**.

Correlation analysis does not make sense if x is set by the experimenter (e.g. doses of a drug)!

Neither correlation nor regression analysis is appropriate to assess the agreement of two measurement methods!

Why is it important to decide which variable should be the dependent one and which the independent one?

Suppose we study the relationship between the hours spent with learning and the score at statistics midterm test. A naïve and pragmatic idea:

- If we want to predict the expected score given a certain time of learning, we should set up a regression model

  hours spent with learning → *score*

Why is it important to decide which variable should be the dependent one and which the independent one?

Suppose we study the relationship between the hours spent with learning and the score at statistics midterm test. A naïve and pragmatic idea:

- If we want to predict the expected score given a certain time of learning, we should set up a regression model

  hours spent with learning → *score*

- If we want to find out how many hours one needs to reach a certain score, then we should set up a model as

  score → *hours spent with learning*

Why is it important to decide which variable should be the dependent one and which the independent one?

Suppose we study the relationship between the hours spent with learning and the score at statistics midterm test. A naïve and pragmatic idea:

- If we want to predict the expected score given a certain time of learning, we should set up a regression model

  hours spent with learning → *score*

- If we want to find out how many hours one needs to reach a certain score, then we should set up a model as

  score → *hours spent with learning*

But this idea is wrong... Why?

Let us have an example illustrating the problem! Say, the true relation between V_1 and V_2 is

$$V_2 = 0.5 \cdot V_1 + 3,$$

which, by simple algebra, can be rewritten like this:

$$V_1 = 2 \cdot V_2 - 6.$$

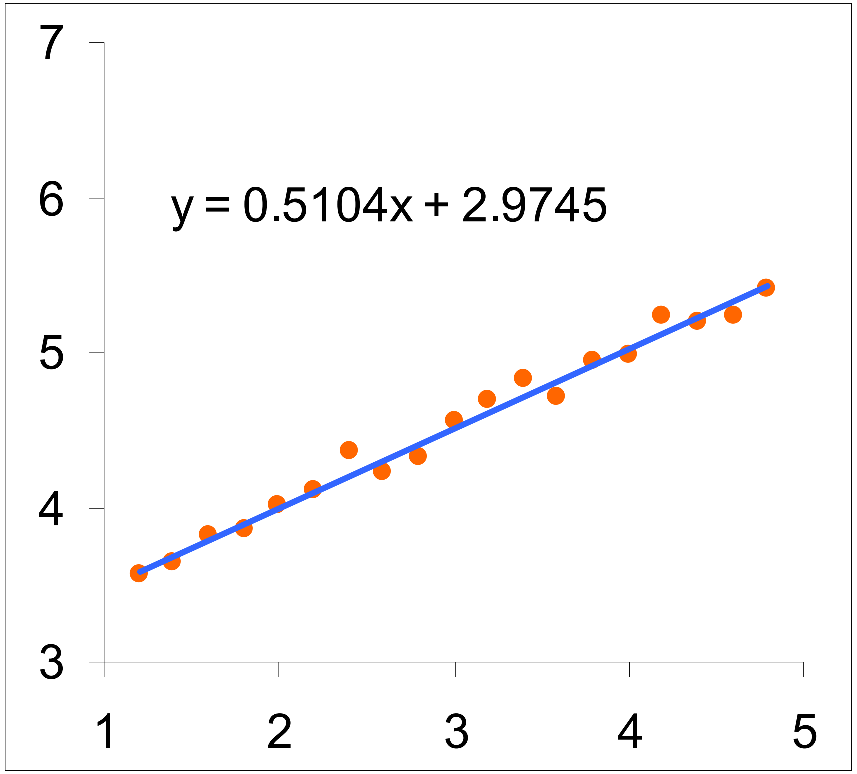
Assume that V_2 depends on V_1, but it also contains a random component ε, that is, the true model is

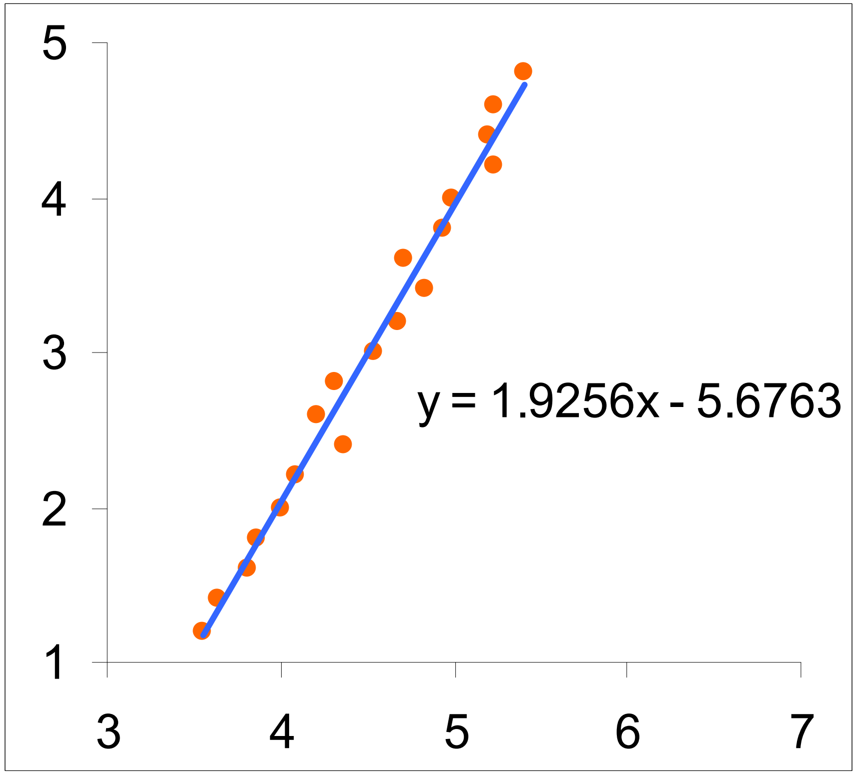
$$V_2 = 0.5 \cdot V_1 + 3 + \varepsilon \, .$$

If the variance of ε is small, both models $V_1 \to V_2$ and $V_2 \to V_1$ give the same results. However, the larger the variance of ε, the worse will be the results from the wrong model $V_2 \to V_1$ while the right one continues giving right results.

Let us fit the regression model

$V_2 = 0.5 \cdot V_1 + 3$ $V_1 = 2 \cdot V_2 - 6$

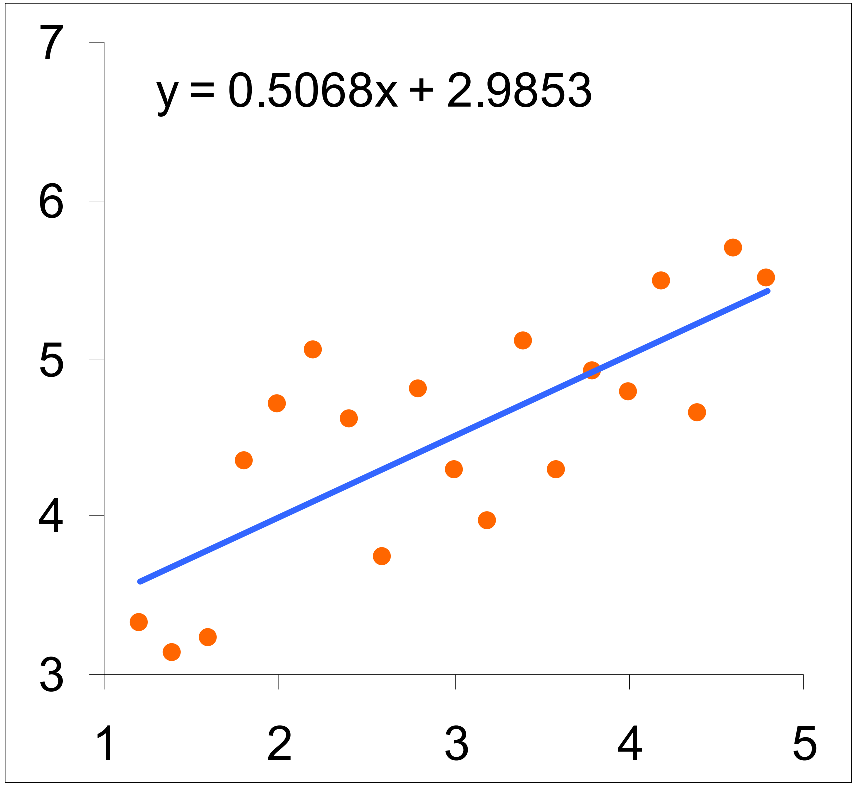


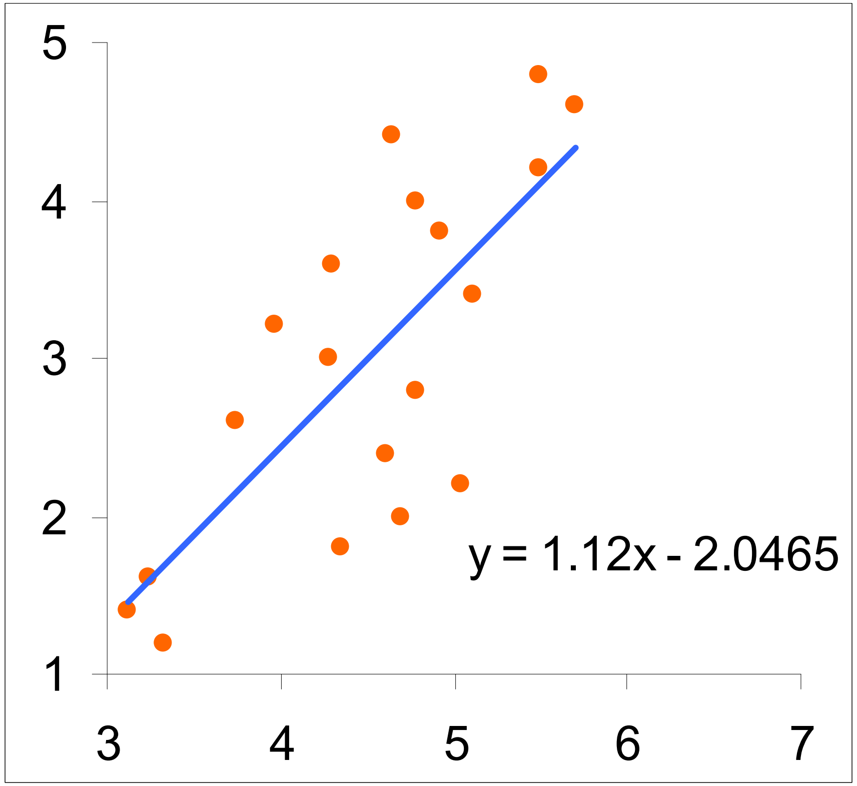


If the variance of ε is small, both models are o.k.

Let us fit the regression model

$V_2 = 0.5 \cdot V_1 + 3$ $\qquad$ $V_1 = 2 \cdot V_2 - 6$



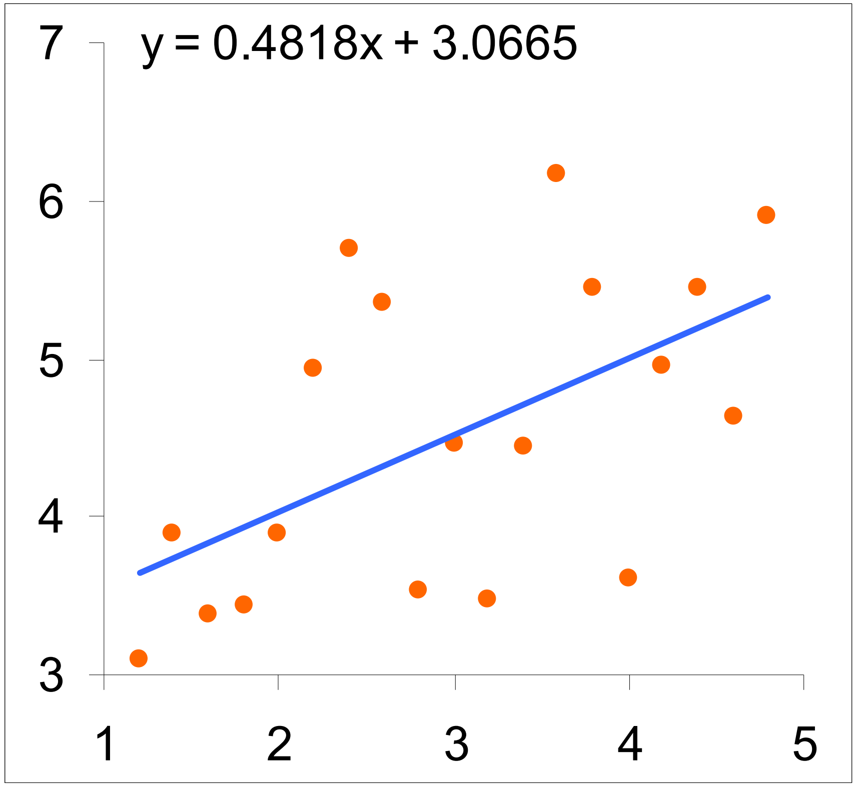


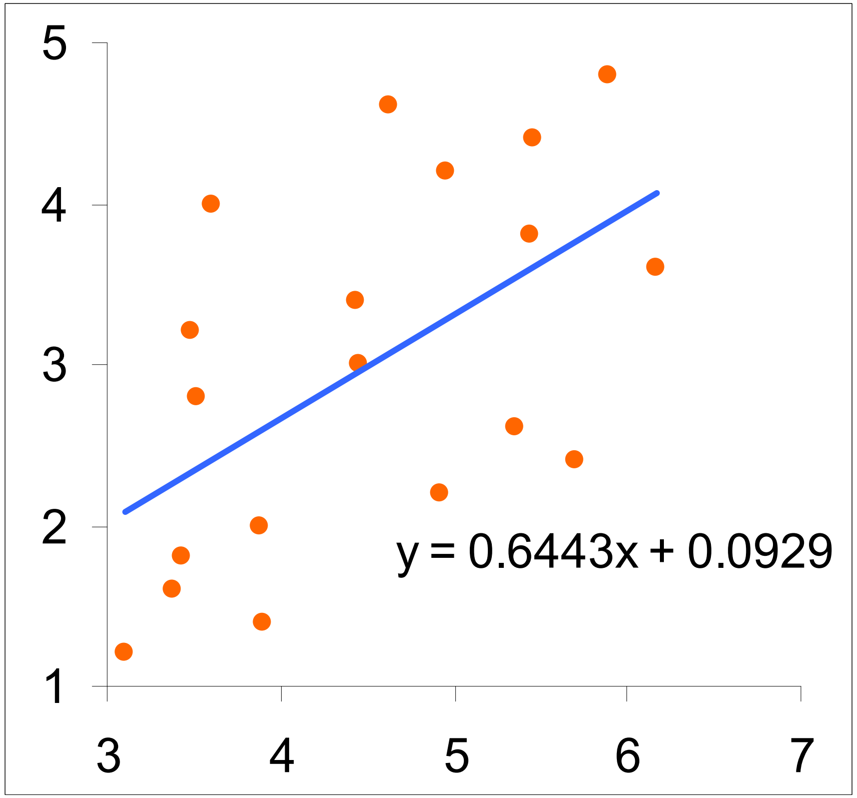
If the variance of ε is larger, results start to deviate.

Let us fit the regression model

$V_2 = 0.5 \cdot V_1 + 3$

$V_1 = 2 \cdot V_2 - 6$



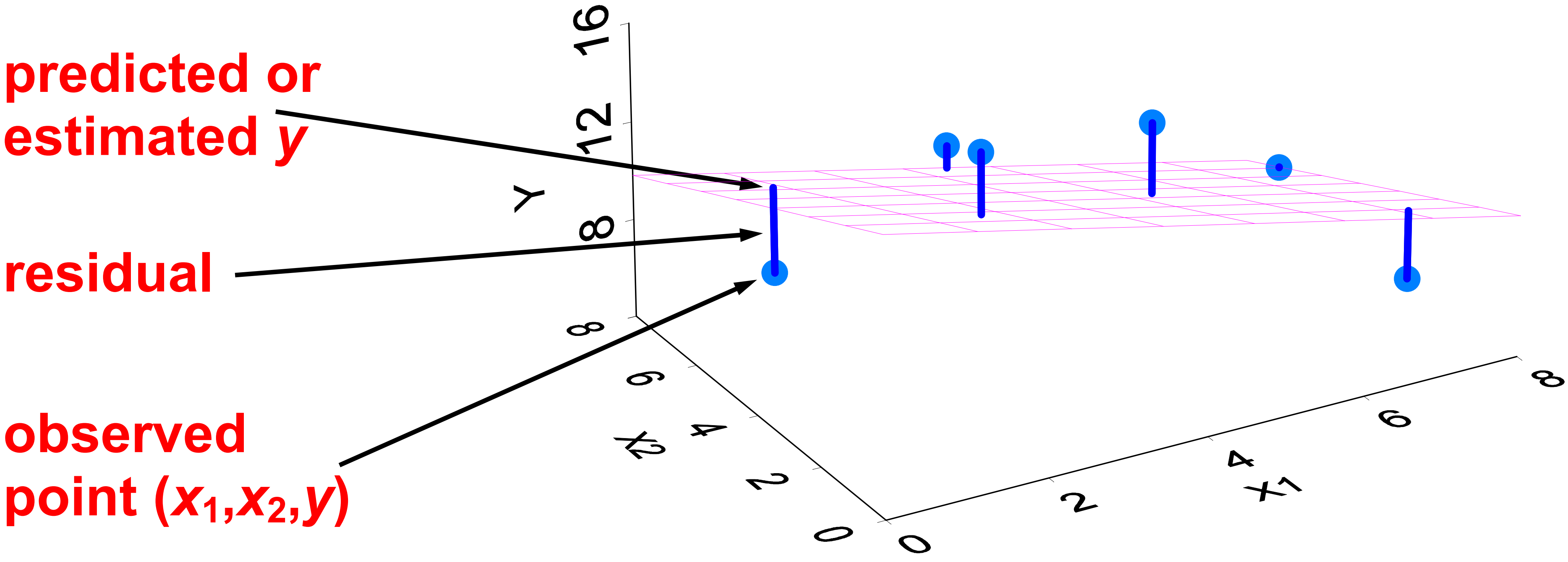


The right model is still fine while the wrong one’s crashed.

Multiple linear regression

There are *n* explanatory variables x_1, x_2, … , x_n, and a single dependent variable *y*. Function is linear. For *n*=2 the regression surface is a plane.

In the figure *y* depends negatively on both x_1 and x_2.



predicted or estimated *y*

residual

observed point (x_1,x_2,y)

Estimation of the parameters: the least squares method

That surface is chosen as the **least square estimate**, which results in the **minimal residual sum of squares**. (in other words: of all possible surfaces that one, from which the squared deviations of the observed points is minimal).

The best fitting surface (curve, line, function) in least squares sense

The statistical programs estimate the parameters of this surface (curve, function). With two explanatory variables there are three parameters: b_0, b_1, b_2 , as the function is

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

The programs also give SEs of the parameter estimates, so we can calculate 95% CIs for the parameters by $b_i \pm 1.96 \cdot \mathrm{SE}$.

Estimation of the parameters: the least squares method

That surface is chosen as the **least square estimate**, which results in the **minimal residual sum of squares**. (in other words: of all possible surfaces that one, from which the squared deviations of the observed points is minimal).

The best fitting surface (curve, line, function) in least squares sense

The statistical programs estimate the parameters of this surface (curve, function). With two explanatory variables there are three parameters: b_0, b_1, b_2 , as the function is

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

The programs also give SEs of the parameter estimates, so we can calculate 95% CIs for the parameters by b_i ± 1.96·SE.

CIs are valid only if the random component ε is normally distributed!

Statistical tests (testing hypotheses on regression)

Typical hypotheses of interest:

„Does y really depend on x_1 (x_2 …)?”

H_0: it does not (what we see has occured by chance)

H_1: it does

If we reject H_0 we say „the regression is significant”.

Statistical tests (testing hypotheses on regression)

Typical hypotheses of interest:

„Does y really depend on x_1 (x_2 …)?”

H_0: it does not (what we see has occured by chance)

H_1: it does *If we reject H_0 we say „the regression is significant”.*

- We can test the **dependence of *y* on each x_i separately** by t-tests (those x_is indicated non-significant should be omitted from the model).

Statistical tests (testing hypotheses on regression)

Typical hypotheses of interest:

„Does y really depend on x_1 (x_2 …)?”

H_0: it does not (what we see has occured by chance)

H_1: it does *If we reject H_0 we say „the regression is significant”.*

- We can test the **dependence of *y* on each x_i separately** by t-tests (those x_is indicated non-significant should be omitted from the model).
- Or we can test the **dependence of *y* on the whole set of x_is** by an F-test.

Statistical tests (testing hypotheses on regression)

Typical hypotheses of interest:

„Does *y* really depend on x_1 (x_2 …)?"

H_0: it does not (what we see has occured by chance)

H_1: it does *If we reject H_0 we say „the regression is significant".*

- We can test the **dependence of *y* on each x_i separately** by *t*-tests (those x_is indicated non-significant should be omitted from the model).
- Or we can test the **dependence of *y* on the whole set of x_is** by an *F*-test.

Necessary condition of all tests is that ε is normally distributed and its variance is constant (=independent of the xs).

To what extent do the x_is determine y?

This is quantified by the coefficient of determination (R^2).

Its value ranges from 0 to 1 (like the association measures).

$R^2=0$ $R^2=1$

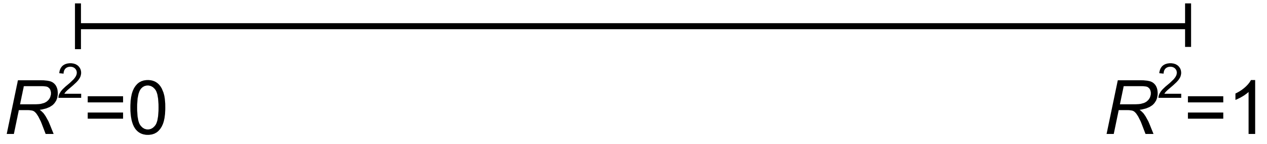
It isn't worth knowing the x_is because they tell us nothing about y (at least nothing in the current regression model).

The x_is completely determine y, it has no random component at all (all points lie exactly on the regression surface).

To what extent do the x_is determine y?

This is quantified by the coefficient of determination (R^2).

Its value ranges from 0 to 1 (like the association measures).



R^2=0

R^2=1

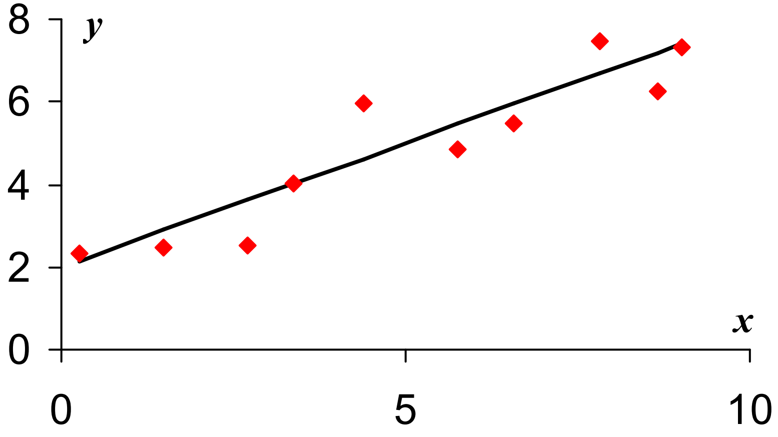
It isn't worth knowing the x_is because they tell us nothing about y (at least nothing in the current regression model).

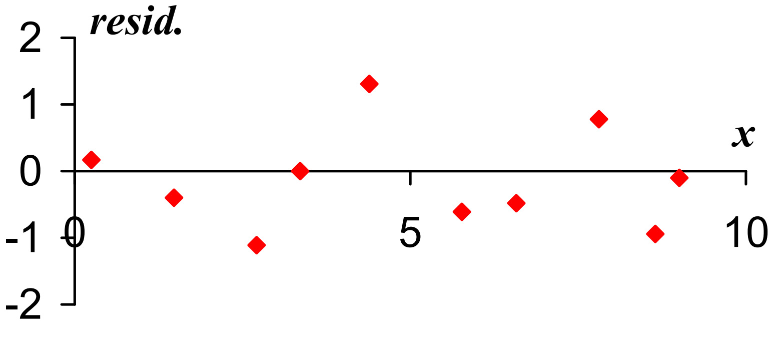
The x_is completely determine y, it has no random component at all (all points lie exactly on the regression surface).

Don't want by all means the highest R^2! (It is better if you have a meaningful model with smaller R^2.)

Is the model all right? (regression diagnostics)

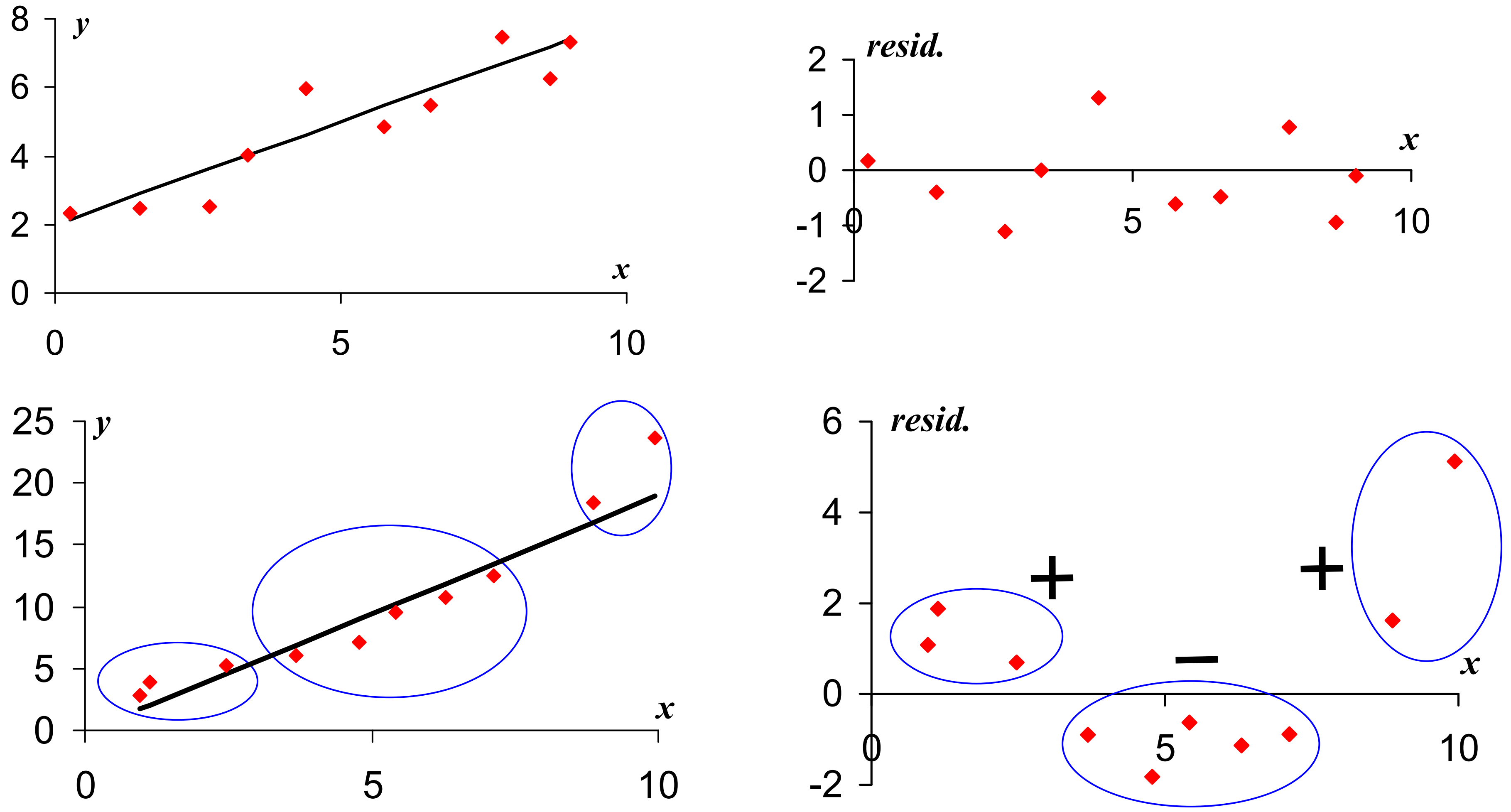
If the model is all right, residuals show a random pattern.





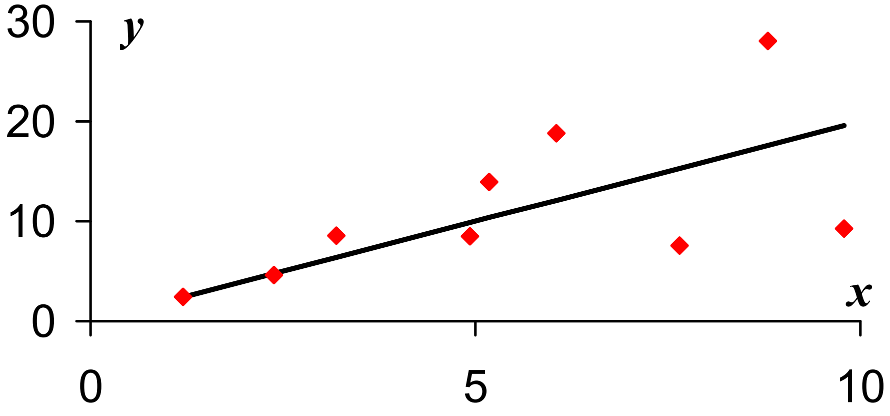
Is the model all right? (regression diagnostics)

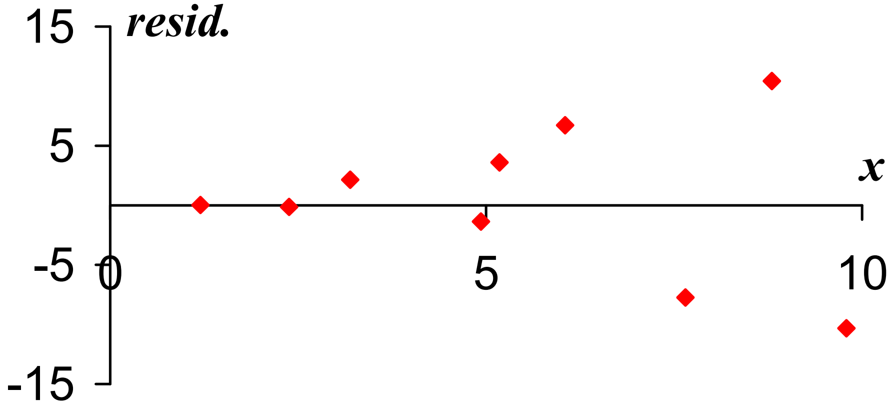
If the model is all right, residuals show a random pattern.



If we want to test significance or to construct confidence intervals, also the followings must hold.

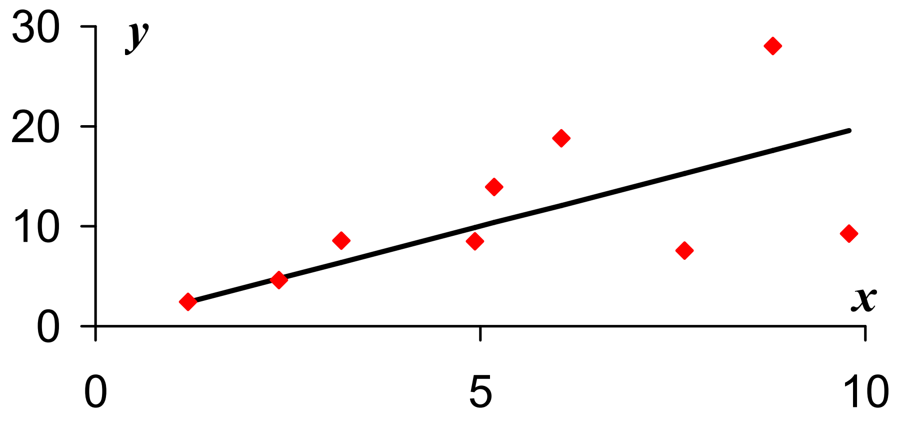
- Variance of the residuals should be constant. If this is not the case, use the weighted least squares (WLS) method.

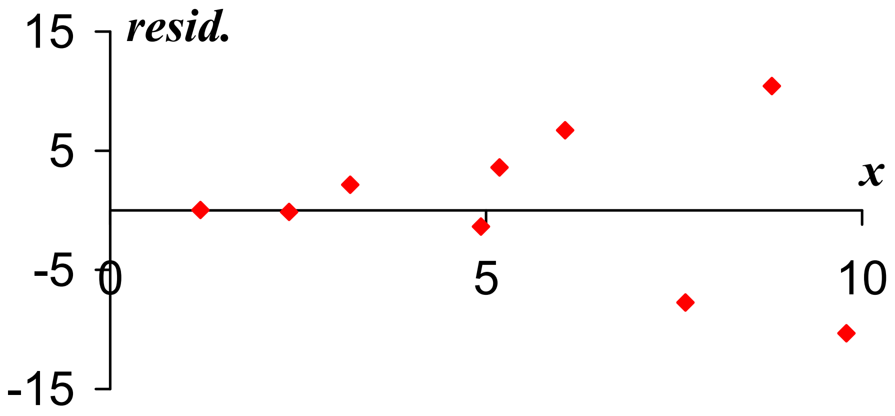




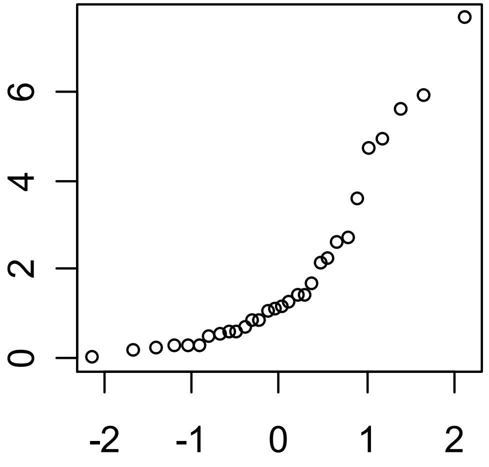
If we want to test significance or to construct confidence intervals, also the followings must hold.

- Variance of the residuals should be constant. If this is not the case, use the weighted least squares (WLS) method.





- Residuals should (at least approximately) follow the normal distribution. This can be checked graphically by a QQ plot.



Check for outliers and judge how strongly they influence the parameter estimates! (That is, the regression line or surface)

Check for high correlations between the explanatory variables (**multicollinearity** or **collinearity**)! If there are high correlations between them, tests can be corrupted (e.g. nothing turns out to be significant).

Check for outliers and judge how strongly they influence the parameter estimates! (That is, the regression line or surface)

Check for high correlations between the explanatory variables (**multicollinearity** or **collinearity**)! If there are high correlations between them, tests can be corrupted (e.g. nothing turns out to be significant).

Check for the diagnostic tools (plots, statistics) offered by the statistical program you are using! They are different in each program.

The general linear model

If you have **categorical as well as continuous explanatory variables**, then you can apply ANCOVA (analysis of variance-covariance): this offers an ANOVA-like analysis of the categorical variables combined with linear regression on the continuous **covariates** at the same time. The mathematical solution is regression analysis by generating **binary dummy variables for the categorical explanatory variables**.

General linear model is the name of this model class.

The general linear model

If you have **categorical as well as continuous explanatory variables**, then you can apply ANCOVA (analysis of variance-covariance): this offers an ANOVA-like analysis of the categorical variables combined with linear regression on the continuous **covariates** at the same time. The mathematical solution is regression analysis by generating **binary dummy variables for the categorical explanatory variables**.

General linear model is the name of this model class.

There is another class of models, the generalized linear models, in which even normality of the random component of y is released.