UNIVERSITY OF BERGEN

Regression analyses

Sebastian Jentschke

UNIVERSITY OF BERGEN

Welcome to the lecture on linear regression analysis.

Agenda

- **introduction**
- **principles and background**: mathematical background, choosing IVs, required sample sizes
- **how to conduct a linear regression?**
- **assumptions for linear regressions**: initial checks, checks within a regression model
- **regression types and model building**: standard, hierarchical, statistical

The lecture will start with a brief introduction that "embeds" regression analyses among other statistical analyses introduced in the course.

Then, we dig deeper in principles and background for regression analyses, covering the mathematical background (and how it compares to a correlation), criteria that we may use to choose independent variables, and what sample sizes are recommended for regression analyses.

Afterwards, we will conduct a simple regression in jamovi.

There are multiple assumptions to be met in order to be sure that the regression model we estimated is valid. This begins with initial checks: e.g., normality and outliers in the variables that we want to include in the model, and continues with assumption checks after estimating the model, among them checks for collinearity and for characteristics of the residuals.

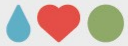
Agenda

- **introduction**
- **principles and background**: mathematical background, choosing IVs, required sample sizes
- **how to conduct a linear regression?**
- **assumptions for linear regressions**: initial checks, checks within a regression model
- **regression types and model building**: standard, hierarchical, statistical

The fourth part introduces several approaches to build regression models: standard, hierarchical, and statistical. Within hierarchical, a practical introduction is given how to carry that out in jamovi.

Introduction

The next part embeds and contextualizes regression analyses in relation to other statistical analyses introduced in the course.



Categorical vs. continuous vars.

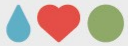
- categorical variables contain a limited number of steps (e.g., male – female, experimentally manipulated or not, level of education)
- continuous variables have a (theoretically unlimited) number of steps (e.g., body height, weight, IQ)
- ANOVA (session in two weeks) is for categorical predictors, regression analyses (this session) and correlation (refresher) are for continuous predictors

I will (again) use the distinction between categorical and continuous variables to contextualize correlation and regression analyses.

Categorical variables encompass the variables from the two measurement levels nominal and ordinal (for an more extensive overview on measurement levels, see the introduction). Strictly speaking is ordinal a hybrid since non-parametric methods (i.e., non-parametric correlations) are possible with ordinal variables. However, for more complex regression models there is no non-parametric choice.

Continuous variables encompass the two variable levels interval and ratio.

For regression analyses (and correlations) we use continuous predictor (independent) variables and continuous outcome (dependent) variables. For ANOVAs we have (mainly) categorical predictors.



Categorical vs. continuous vars.

	independent	
dependent	***categorical***	***continuous***
categorical	chi-squared	logistic regression
continuous	t-test, ANOVA (incl. ANCOVA) *experimental design*	correlation, **linear regression** (incl. moderation, mediation) *survey design*

Both classes of methods (linear regression and ANOVA) are based upon the General Linear Model (I will say more on that later). Linear regression is covered quite extensively in this lecture, ANOVA in another lecture in two weeks time. In between, there will be a lecture on mediation and moderation which is based upon regression models but extends them.

Both, ANOVA and lineare regression, are quite central to our methods repertoire. ANOVAs are typically used to analyse data from experiments (where we manipulated one or more factors, representing categorical variables), whereas linear regression models are often used to analyse data acquired with questionnaires.

Relation vs. difference hypotheses

- ***relation hypotheses*** explore whether there is a relation between one (or more) independent and a dependent variable (*functional form*)
- ***difference hypotheses*** explore whether there is a difference between the steps of one (or more) independent and a dependent variable (*parameter*)
- the distinction between IV and DV is blurred for relation hypotheses
  → causality can only be inferred if the independent variable was experimentally manipulated

Another way of describing the distinction between categorical and continuous predictor (independent) variables is that between relation vs. difference hypotheses.

With relation hypotheses we explore the relationship between one or more independent (predictor) variable and a continuous dependent (outcome) variable. Regression analyses explore relation hypotheses.

For categorical variables as predictor (independent) variables we are often more interested in whether results from two (or more) categories significantly differ from each other, most typically to ask whether our experimental manipulation made a difference (i.e., had an effect).

Relation vs. difference hypotheses

- ***relation hypotheses*** explore whether there is a relation between one (or more) independent and a dependent variable (*functional form*)
- ***difference hypotheses*** explore whether there is a difference between the steps of one (or more) independent and a dependent variable (*parameter*)
- the distinction between IV and DV is blurred for relation hypotheses
  → causality can only be inferred if the independent variable was experimentally manipulated

One thing to be careful about in the context of relation hypotheses is that causality can typically not be claimed for relational hypotheses.

Causality can only be inferred if (1) the dependent variable precedes the independent variable in time and (2) if the independent variable was manipulated in order to measure the effect on the dependent variable. Often the precedence can't be so easily established for relational hypotheses and regression analyses exploring such hypotheses. An example where a manipulation of a continuous variable could be imagined (but still would be quite costly to acquire) is where the dosage of a treatment is manipulated in an experiment. However, often we compare two to four steps of dosage (which makes the variable categorical) since sampling over the whole range would require too many measurements.

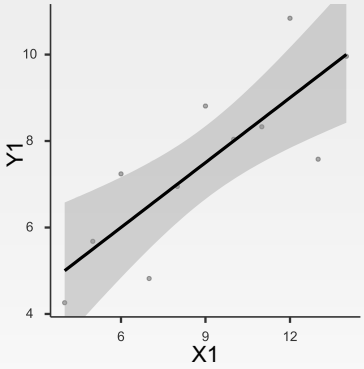
Principles and background

The next part is speaking about the background and the principles behind correlation and regression analyses. This first half of that part is mainly theoretical and covers the mathematics behind the method, criteria for choosing independent variables, and considerations regarding sample sizes.

That will be followed by a more practical part with a demonstration of a simple regression in jamovi.

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

$R = r_{Y\hat{Y}}$ (r_{xy}) $Y_i = \hat{Y}_i + \varepsilon_i$

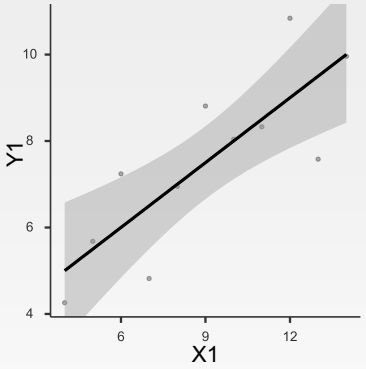
When we use a correlation, our aim is to determine the size and the direction of a linear relationship between pairs of variables. Regression, however, is a little more complex and has slightly different aims.

First of all, regressions can include several independent (predictor) variables in order to predict one dependent (outcome) variable. In contrast, in correlations, the distinction in independent and dependent variables is often rather blurred.

However, within regression models, we still can not claim causality in most cases, even though we might think of the relationship between independent and dependent variables as the independent influencing the dependent variables.

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

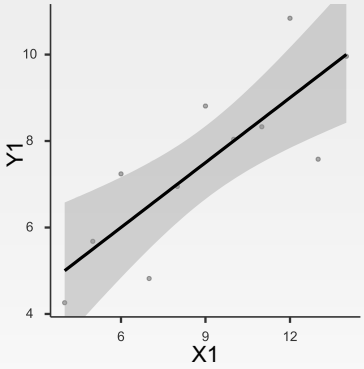
$R = r_{Y\hat{Y}}$ (r_{xy}) $Y_i = \hat{Y}_i + \varepsilon_i$

When we think about correlations, they are most easily imagined as a scatter plot with a regression line. The term regression line already makes a reference that a correlation follows the same principles as a regression.

To mathematically describe such a regression line in a correlation, we use the green formula on the right hand side: ŷ = a + bx. It means that the value we predict for y is composed of multiplying the value from the x-variable with a certain weight b and then adding a. b describes the slope of the regression line, i.e., how much increase in y results from increasing x by 1. a describes what value ŷ has if x = 0 (that is the point where we cut the y-axis; in the current example that value is 3). The hat (^) we put on y denotes that this is an estimate that differs (more or less) from the real value of y that we measured. More on that in a second.

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

$R = r_{Y\hat{Y}} (r_{xy})$ $Y_i = \hat{Y}_i + \varepsilon_i$

If we take a closer look at the formula, then we see that the formula on the left is just a little more complex way to write the formula that we used for the correlation. Whereas we only deal with one variable for x in the correlation formula (right, starting with ŷ), we can include multiple predictors in a linear regression.

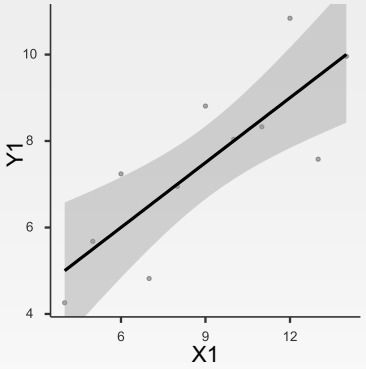
The reason why all the letters are uppercase is that those variables are now arranged as vectors (Ŷ, B) and matrices (X).

I mentioned that both regression analyses and ANOVAs are based upon the General Linear Model which uses matrix algebra. You are lucky that you won't see much of all the mathematics behind that. Software is doing all that for you. I had to do a matrix inversion (on paper) to manually estimate the B-vector in my statistics exam (some 25 years ago). Tiresome, but it helped understanding what happened behind the scenes.

UNIVERSITY OF BERGEN

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

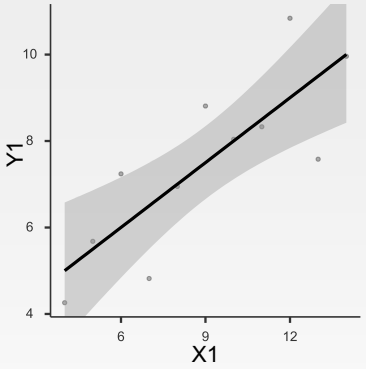
$R = r_{Y\hat{Y}} (r_{xy})$ $Y_i = \hat{Y}_i + \varepsilon_i$

Within the X matrix, we have one participant per row and one variable per column. In addition, we have a column at the very beginning (X_0; containing "1" for all participants). This column permits to calculate B_0 which is equivalent to the "a" in the simple form and takes care of the mean, i.e., which value y would have if all values in X were 0).

Given that we deal with matrices, the whole equation could even be simplified to $\hat{Y} = BX$. This means that $\hat{Y}$ is estimated by multiplying our independent variables (assembled in the X matrix) by certain weights B (slopes, i.e., information about how much Y is going to increased if the independent variables were to increase by 1). This means, we calculate $\hat{Y}$ by first taking B_0 multiplied by the "1" in the first column, add the value of the first predictor X_1 multiplied with its slope B_1, add X_2 multiplied by B_2 and so on to arrive at our estimate $\hat{Y}$.

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y}_i = a + bx_i)$

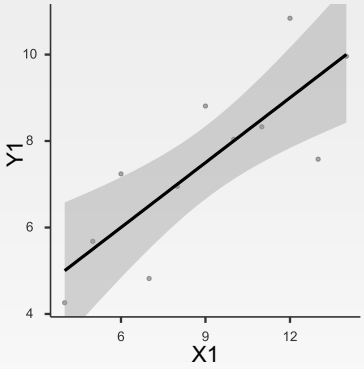
$R = r_{Y\hat{Y}} (r_{xy})$ $Y_i = \hat{Y}_i + \varepsilon_i$

There are two caveats with those formulas. Sometimes (as in the jamovi book) you may not find the additional index / subscript i in the formula, making it look like $\hat{Y} = B_0 + B_1X_1 + B_2X_2 + \ldots$ This i indicates a counter for each individual included in the analysis. A nice feature of adding the index is that it makes clear which variables can vary and which are fixed. Another difference may be that b might occasionally be written in lowercase.

The other caveat is the we have our estimate $\hat{Y}$ on the one hand and the real value of Y (that we measured) on the other hand. They are not the same and therefore we have a further vector ε that contains the difference between our prediction and the real value. This is shown in the formula at the bottom left ($Y = \hat{Y} + \varepsilon$). Our aim is to make our predictions as exactly as possible. That is typically achieved by trying to minimize the values in ε.

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

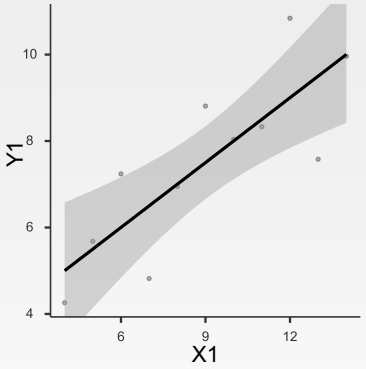
$R = r_{Y\hat{Y}} (r_{xy})$ $Y_i = \hat{Y}_i + \varepsilon_i$

Actually, we aim to minimize the squared values that are contained in ε. The reason for that is (at least) twofold: If we square those values, (1) larger differences get a higher weight, and (2) all values become positive (which makes handling them easier). These ***squared deviations*** contained in ***ε*** are then ***summed up*** (over rows, i.e., participants) and this ***sum of squared deviations*** is going to be ***minimized***. The method for that is called ***ordinary least squares (OLS) regression***.

A visual impression about how this is done can be got from Figure 12.11 and Figure 12.12 in the jamovi-book (Navarro & Foxcroft, 2022; p. 295 / 296).

Principles and background

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)



$\hat{Y}_i = B_0 (= a) + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik}$ $(\hat{y} = a + bx)$

$R = r_{Y\hat{Y}} (r_{xy})$ $Y_i = \hat{Y}_i + \varepsilon_i$

One last comment on our matrices: I already mentioned that X (one row per participant, one column per variable) is multiplied by B (one column per variable). When doing a matrix multiplication, each cell is multiplied and then added up (according to the formula above: $B_0 [\cdot 1] + B_1 \cdot X_{1i} + B_2 \cdot X_{2i} \ldots.$) This time, I added the $_i$ to indicate that this is done per participant. We end up with a vector $\hat{Y}$ (one row per participant). This vector differs from what we measured in Y (also one row per participant). When subtracting the two, we get ε (again, one row per participant).

Taking the whole formula, Y (our measured values) is composed of the part we predict (BX) and the part we can't predict (ε). Setting those two into relation is central for the evaluation of statistical significance for our model.

Principles and background

- Parameter estimation: Minimize the squared error
- $Y_i = B_0 + B_1 \cdot X_{i1} + \ldots + B_k \cdot X_{ik} + \varepsilon_i$

  $Y = BX + \varepsilon \qquad [\hat{B} = (X'X)^{-1} X'Y]$

  Y, y = dependent variable
  X, $[x_1 \ldots x_k]$ = predictor variable
  B, $[b_0 \ldots b_k]$ = predictor weights (b_0: intercept; $b_1 \ldots b_n$: slope)
  E, [e] = error term

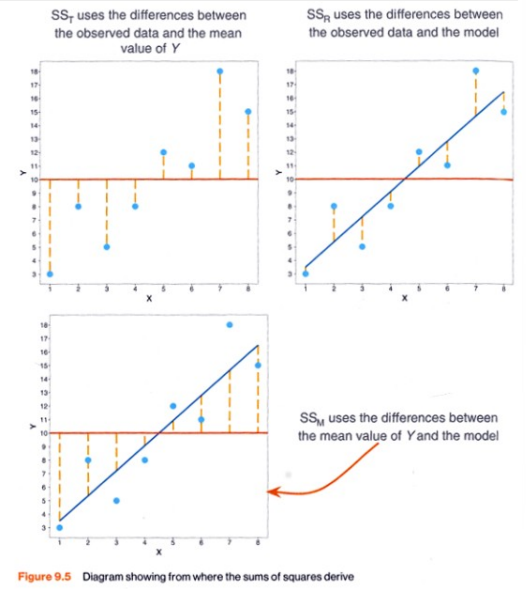


Figure 9.5 Diagram showing from where the sums of squares derive

The current slide provides a bit of a graphical visualization of how we assess significance. We see again the formulas for the general linear model, and three figures on the right.

The first figure (top-left) shows the situation if we had no independent (predictor) variables. In that situation all participants are assessed relative to the mean.

If we fit a regression line (top-right), we add a certain independent variable to "explain" the data. As a consequence, our prediction gets more precise and the yellow lines (representing the error ε) get shorter.

What we do for significance testing is to take the predictions that are due to the model (bottom-left; i.e., the distances from the mean to our regression line) and set those into relation to the part that we can't explain (the short yellow line in the top-right figure).

Principles and background

- Parameter estimation: Minimize the squared error
- $Y_i = B_0 + B_1 \cdot X_{i1} + \ldots + B_k \cdot X_{ik} + \varepsilon_i$

  $Y = BX + \varepsilon \qquad [\hat{B} = (X'X)^{-1} X'Y]$

  Y, y = dependent variable
  X, $[x_1 \ldots x_k]$ = predictor variable
  B, $[b_0 \ldots b_k]$ = predictor weights (b_0: intercept; $b_1 \ldots b_n$: slope)
  E, [e] = error term

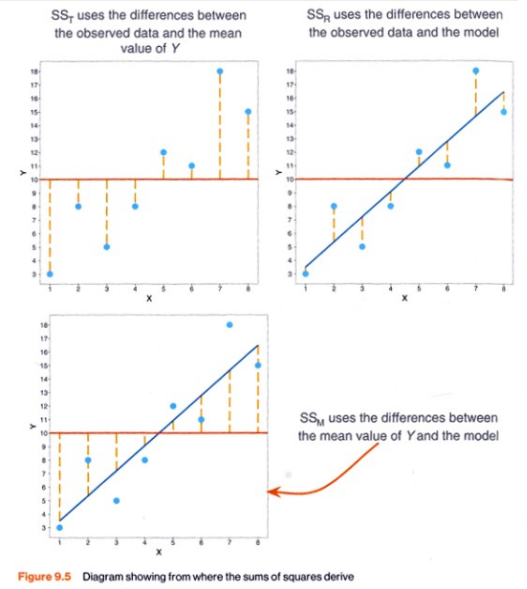


Figure 9.5 Diagram showing from where the sums of squares derive

We then compare a null hypotheses where we assume that the mean is a reasonable predictor to our alternative hypothesis claiming that we make a better prediction if we fit regression lines to include one (or more) independent variables.

H_0: $Y_i = b_0 + \varepsilon_i$

H_1: $Y_i = b_0 + \left(\sum_{k=1}^{K} b_k \cdot X_{ik} \right) + \epsilon_i$

When we square the length of the yellow lines in the figure and sum them up. This is called sum of squares. We have three of those:

SS_{tot} (top left) = SS_{mod} (bottom left) – SS_{res} (top right)

Principles and background

- Parameter estimation: Minimize the squared error
- $Y_i = B_0 + B_1 \cdot X_{i1} + \ldots + B_k \cdot X_{ik} + \varepsilon_i$

  $\mathbf{Y = BX + \varepsilon}$ $\quad [\hat{B} = (X'X)^{-1} X'Y]$

  Y, y = dependent variable
  X, $[x_1 \ldots x_k]$ = predictor variable
  B, $[b_0 \ldots b_k]$ = predictor weights (b_0: intercept; $b_1 \ldots b_n$: slope)
  E, [e] = error term

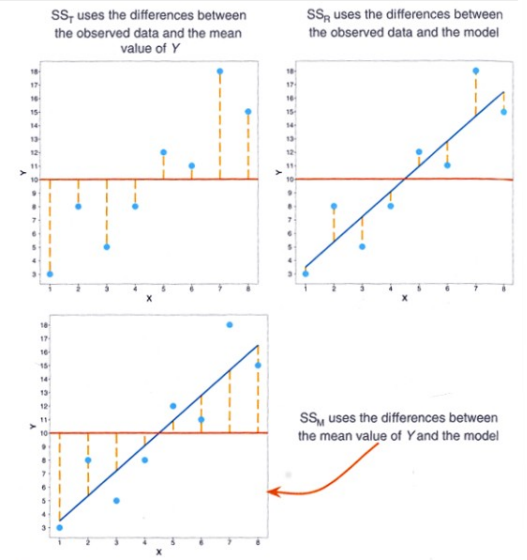


Figure 9.5 Diagram showing from where the sums of squares derive

This square sums are then „standardized" by how many degrees of freedoms we needed to calculate them. The core behind the idea of the degrees of freedom (df) is that each sum of squares (SS) is standardized or weighed by how many independent sources of variation contributed to calculating it.

Originally, the number of participants (N) was the number of sources. If you calculate the mean, we have one parameter fixed. Each independent variable takes up on degree of freedom for the model. The remaining degrees of freedom that go to the residuals are therefore df_{res} = N (original sources of variation) – K (independent variables in the model) – 1 (mean).

We use df_{res} = N – K – 1 and df_{mod} = K to „standar-dize" our sums of squares and to arrive at what is called the mean sum of squares: $MS_{mod} = SS_{mod} / df_{mod}$ and $MS_{res} = SS_{res} / df_{res}$ Those are set into relation to calculate $F = MS_{mod} / MS_{res}$

Principles and background

regression techniques:

- standard, hierarchical, statistical

typical research questions for using regression analysis:

- ***investigate*** a ***relationship*** between one DV and several IV
- ***investigate*** a ***relationship*** between one DV and some IVs with the effect of other IVs ***statistically eliminated***
- ***compare*** the ability of several ***competing sets of IVs*** to ***predict a DV***
- (ANOVA as a special case with dichotomous IVs)

We distinguish three different techniques for doing a regression: standard (where all independent variables are entered in the regression model at once), hierarchical (where we add independent predictors in a certain succession based upon how we theoretically assess the influence of that variable), and statistical (where the statistics software does the variable selection for you).

I won‘t cover too much of that statistical approach. The practical reason is that it is not implemented in jamovi (it can be easily done in R from within jamovi if you really want it). There are also reasons why jamovi doesn‘t implement it: the statistical approach is most suited for generating hypotheses but often abused for testing them. This often happens without proper hypotheses regarding which independent variable contributes to the effect on the dependent variable and how. Overfitting is another problem of that method.

Principles and background

regression techniques:

- standard, hierarchical, statistical

typical research questions for using regression analysis:

- ***investigate*** a ***relationship*** between one DV and several IV
- ***investigate*** a ***relationship*** between one DV and some IVs with the effect of other IVs ***statistically eliminated***
- ***compare*** the ability of several ***competing sets of IVs*** to ***predict a DV***
- (ANOVA as a special case with dichotomous IVs)

We have three (plus one) main purposes (i.e., classes of research questions) that are typically evaluated using regression analyses.

(1) We aim to evaluate the relationship between several independent variables onto one dependent variables.

(2) Within the independent variables, we might have some variables we are genuinely interested in and some other variables where we want to statistically control for their influence (nuisance variables) thereby excluding or removing that influence.

(3) Quite central within regression also stands that we can collect data from one sample and after-wards making predictions for another sample. Let's assume we measure one personality characteristic that is difficult to assess with test X,Y, and Z in one sample. Afterwards we could estimate this characteristic in another sample, just by taking their test results in X, Y, and Z.

Principles and background

regression techniques:

- standard, hierarchical, statistical

typical research questions for using regression analysis:

- ***investigate*** a ***relationship*** between one DV and several IV
- ***investigate*** a ***relationship*** between one DV and some IVs with the effect of other IVs ***statistically eliminated***
- ***compare*** the ability of several ***competing sets of IVs*** to ***predict a DV***
- (ANOVA as a special case with dichotomous IVs)

(4) Finally, given that both linear regression and ANOVA are based upon the same mathematical model, the General Linear Model, both of them represent “special cases” of each other and can be converted into one another. Regression analyses and ANOVAs are mathematically quite similar. What differs is whether the mean focus is on exploring relation-hypotheses or difference-hypotheses.

For example, if we were to include categorical (dichotomous) independent variables in a regression, we can quite easily include elements of an ANOVA. Vice versa, including a continuous predictor (making the ANOVA an ANCOVA) includes elements of a regression into the ANOVA.

Principles and background

predicting scores for members of a new sample:

- regression coefficients (B) can be applied to new samples
- generalizability should be checked with cross-validation (e.g., 50/50, 80/20 or boot-strapping)

changing IVs:

- squaring IVs (or raising to higher power) to explore curvilinear relationships

I mentioned that we can use the coefficients (B) estimated from one sample where we measured both independent and dependent variables to another sample where we only know the values for the independent variables.

When we do the estimation, we should employ a technique called cross-validation in order to ensure that the data can really be generalized to new samples. The idea behind cross-validation is that we split our sample (e.g., using 80% of the participants for estimating the model, applying it to the remaining 20%) and see how exactly that estimate fits with the real dependent variable. Boot-strapping describes that this process (split, estimate, check) is repeated numerous times.

Principles and background

predicting scores for members of a new sample:

- regression coefficients (B) can be applied to new samples
- generalizability should be checked with cross-validation (e.g., 50/50, 80/20 or boot-strapping)

changing IVs:

- squaring IVs (or raising to higher power) to explore curvilinear relationships

Another opportunity to go even beyond that is that regression analyses could in principle even employ independent variables raised to higher power (e.g., by squaring them) in order to explore curvilinear relationships. If we can see that our dependent variable is curvilinear, we can raise one or more independent variables to higher power. The decision which variables are raised can be either made based on theoretical assumptions: there might be one independent variable where we expect that it might contribute to the curvilinear slope in the dependent variable. We could as well try it out for all variables (which possibly isn't really recommended given that we still have to interpret those results).

Given that such situations are rather special, the method is not covered in detail in this lecture.

Principles and background

considerations for which IVs to choose:

- implied causality
- further considerations (or lack of) regd. inclusion of variables
  theoretical: if the goal is the manipulation of a DV, include some IVs that can be manipulated as well as some who can't
  practical: include «cheaply obtained» IVs (existing data; SSB)
  statistical: IVs should correlate strongly with the DV but weak with other IVs (goal: predict the DV with as few as possible IVs); remove IVs that degrade prediction (check residuals)
  choose IVs with a high reliability

There are several aspects to consider when deciding which variables to include in our regression model. The first case is that we assume that a certain independent variable might be responsible for a certain effect on the dependent variable (that is, we imply a certain causal relationship between the two variables).

There are further aspects that may influence which independent variables we choose to include in our model. Those aspects fall into three categories:

"Theoretical" indicates that we make the decision about inclusion based upon that we theoretically expect them that independent variable to have a relation to the dependent variable AND to be capable of being manipulated easily. Such considerations might especially apply if we want to manipulate the dependent variable (e.g., in the context of an intervention).

Principles and background

considerations for which IVs to choose:

- implied causality
- further considerations (or lack of) regd. inclusion of variables
  theoretical: if the goal is the manipulation of a DV, include some IVs that can be manipulated as well as some who can't
  practical: include «cheaply obtained» IVs (existing data; SSB)
  statistical: IVs should correlate strongly with the DV but weak with other IVs (goal: predict the DV with as few as possible IVs); remove IVs that degrade prediction (check residuals)
  choose IVs with a high reliability

We may make the decision based upon practical considerations, such as that the data can easily be obtained from a register or the SSB (Statistisk Sentralbyrå). Collecting data is always costly, so a rather "inexpensive" option is preferable where this is possible.

Finally, there are statistical considerations. Ideally, independent variable are all relatively highly correlated with the dependent variable and relatively minimal among each other. If we think of the variation of the dependent variable as a cake, each independent variable should help to explain a different piece or bit of the cake. If the independent variables are relatively highly correlated, these pieces would overlap to a considerable degree. That is, even though both variables could contribute to explaining that piece only one is considered and the other would be "wasted".

Principles and background

ratio of cases to IVs (m = IVs):

N ≥ 50 + 8m for multiple correlation (standard / hierarchical)
N ≥ 40m for multiple correlation (statistical)
N ≥ 104 + m for individual predictors
(assuming α = .05, β = .20 and medium effect size;
higher numbers are needed if the DV is skewed, small effect size is anticipated or substantial measurement error is expected)
N ≥ (8 / f²) + (m – 1) [f = .02, .15, .35 for small, medium, large eff.]

strategies for insufficient N: exclude IVs, create composite meas.

Finally, a couple of recommendations for sample sizes. All these recommendations are based upon the assumption that we are willing to accept to make a type-I-error (rejecting the H_0 if it were true; α) in 5% of all cases (0.05), a type-II-error (retain the H_0 if it were false; β) in 20% of all cases and a medium effect size (equating to a correlation of at least 0.3).

For the two cases covered in that lecture, standard and hierarchical, the number of participants should be 50 plus an additional 8 for each independent variable we want to include (with 3 IVs: 50 + 3 · 8 = 74). Statistical selection has a higher danger for overfitting therefore the sample size there should be 40 for each independent variable. If we wan't to assess coefficients for individual predictors, we need sample sizes of 104 plus one for each independent variable.

Principles and background

ratio of cases to IVs (m = IVs):

N ≥ 50 + 8m for multiple correlation (standard / hierarchical)
N ≥ 40m for multiple correlation (statistical)
N ≥ 104 + m for individual predictors
(assuming α = .05, β = .20 and medium effect size;
higher numbers are needed if the DV is skewed, small effect size is anticipated or substantial measurement error is expected)
N ≥ (8 / f²) + (m – 1) [f = .02, .15, .35 for small, medium, large eff.]

strategies for insufficient N: exclude IVs, create composite meas.

This might mean that if we had a sample size of 80, we could evaluate a model as a whole if it had up to three predictors, but we shouldn't report the significance for individual coefficients. These sample sizes are recommendations so you might still do otherwise but you should be aware that it maybe a bit shaky grounds that you are standing on (which may call into question whether your results can be generalized).

Another caveat is that effect sizes and the character of your variable might play in. These recommendations might be too low if your dependent variable were skewed or if your effect size were not medium but small (vice versa could the required sample sizes be reduced if you effect sizes were large).

Strategies to deal with too low sample size could be to exclude independent variables or to calculate composites (by summing or averaging scores).

How to conduct a linear regression?

The next part is speaking about the background and the principles behind correlation and regression analyses. This first half of that part is mainly theoretical and covers the mathematics behind the method, criteria for choosing independent variables, and considerations regarding sample sizes.

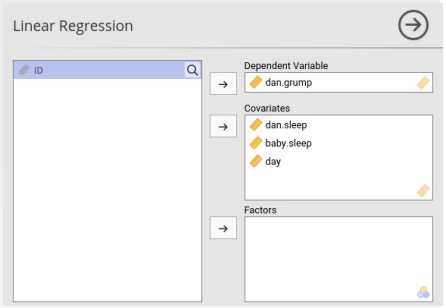
That will be followed by a more practical part with a demonstration of a simple regression in jamovi.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- select Regression → Linear Regression
- assign dan.grump to “Dependent variable” and dan.sleep, baby.sleep, and day to “Covariates”
- the results indicate a strong prediction, based mainly upon dan.sleep



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p
Intercept	126.279	3.242	38.945	< .0001
dan.sleep	−8.969	0.560	−16.016	< .0001
baby.sleep	0.016	0.273	0.058	0.9541
day	−0.004	0.015	−0.288	0.7736

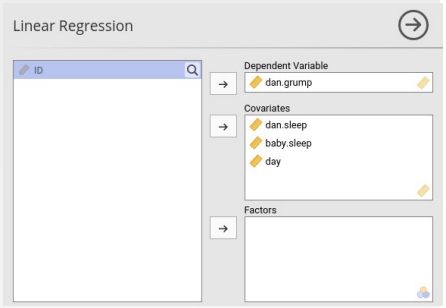
We can conduct a linear regression by clicking on the “Regression”-button and then selecting the second option “Linear regression” from the menu that opens. On the left hand side of the screen there is a window / grey area where we make our selections. In the most simple case, we assign one variable that we want to predict as “Dependent variable”, in our case dan.grumpy. Then we assign one or more variables that we want to use to predict the dependent variable to “Covariates”, those variables are dan.sleep, baby.sleep, and day.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- select Regression → Linear Regression
- assign dan.grump to “Dependent variable” and dan.sleep, baby.sleep, and day to “Covariates”
- the results indicate a strong prediction, based mainly upon dan.sleep



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p
Intercept	126.279	3.242	38.945	< .0001
dan.sleep	−8.969	0.560	−16.016	< .0001
baby.sleep	0.016	0.273	0.058	0.9541
day	−0.004	0.015	−0.288	0.7736

Then, we can have a look at the output: The first table “Model Fit Measures” tells us a R is 0.903. This value can be treated like a correlation coefficient, meaning that we can interpret it like one. “0.903” is close to 1 and therefore fairly substantial. Whereas we used one “predictor” to determine the correlation coefficient, we now use multiple predictors to “compile” R (i.e., each independent variable contributes to R).

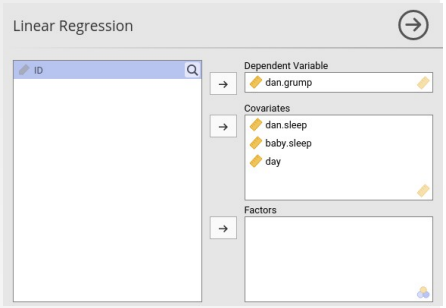
There is another aspect which is different from the correlation coefficient. R is always positive (whereas r can range between -1 and 1). The reason is that several independent variables contribute to R. They might influence the dependent variable in different directions. If we look under “Estimate” we see a negative coefficient for dan.sleep. This tells us that the less Dan sleeps the more grumpy Dan gets.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- select Regression → Linear Regression
- assign dan.grump to “Dependent variable” and dan.sleep, baby.sleep, and day to “Covariates”
- the results indicate a strong prediction, based mainly upon dan.sleep



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p
Intercept	126.279	3.242	38.945	< .0001
dan.sleep	−8.969	0.560	−16.016	< .0001
baby.sleep	0.016	0.273	0.058	0.9541
day	−0.004	0.015	−0.288	0.7736

Now let’s assess the coefficients. We can only interpret the coefficient for dan.sleep that became significant (according to the p in the last column; $p < 0.001$). That is, baby.sleep and day don’t really seem to contribute to the prediction and we may decide to remove them from the independent variable (i.e., the variable list in “Covariates”).

Let’s have a closer look at how it is assessed whether a coefficient is significant or not. We, as always start with two hypotheses: H_0: $b = 0$ and H_1: $b \neq 0$. In order to arrive at the t-value, we divide that coefficient (or strictly speaking the estimate for it $\hat{b}$) by it’s standard error $SE(\hat{b})$: $t = \hat{b} / SE(\hat{b})$

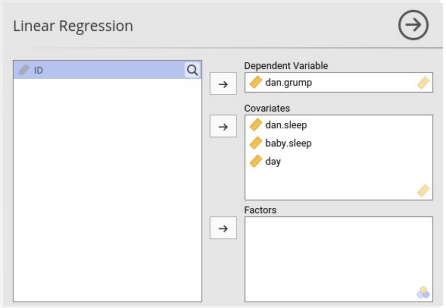
If that variation in $\hat{b}$ is considerably larger than the standard error when measuring it $SE(\hat{b})$ that variation is likely not just due to chance and we can assume that its influence is significant.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- select Regression → Linear Regression
- assign dan.grump to “Dependent variable” and dan.sleep, baby.sleep, and day to “Covariates”
- the results indicate a strong prediction, based mainly upon dan.sleep



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p
Intercept	126.279	3.242	38.945	< .0001
dan.sleep	−8.969	0.560	−16.016	< .0001
baby.sleep	0.016	0.273	0.058	0.9541
day	−0.004	0.015	−0.288	0.7736

What is important is that the standard error of the estimated regression coefficient SE($\hat{b}$) depends on both the predictor and outcome variables, and it is somewhat sensitive to violations of the homogeneity of variance assumption.

You may skip the rest of the slide if you don’t like formulas.

To calculate the standard error we start with the residuals: $\varepsilon = y - X\hat{b}$

Those residuals are then used to calculate the estimated residual variance $\hat{\sigma}^2 = \varepsilon'\varepsilon / (N - K - 1)$

This variance is then multiplied by the inverse of our data matrix and its transposition: $\hat{\sigma}^2 (X'X)^{-1}$

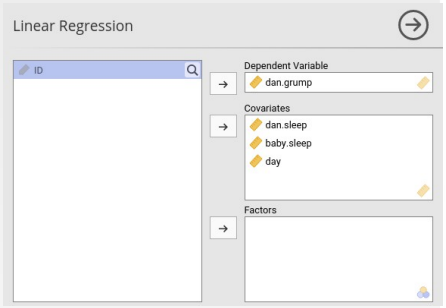
The main diagonal of the resulting matrix is SE($\hat{b}$).

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- select Regression → Linear Regression
- assign dan.grump to “Dependent variable” and dan.sleep, baby.sleep, and day to “Covariates”
- the results indicate a strong prediction, based mainly upon dan.sleep



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p
Intercept	126.279	3.242	38.945	< .0001
dan.sleep	−8.969	0.560	−16.016	< .0001
baby.sleep	0.016	0.273	0.058	0.9541
day	−0.004	0.015	−0.288	0.7736

There are two further things to mention: (1) As I said one aim within regression models is to control for nuisance variables. Those might be categorical. If so, they can be added to the variable list under “Factors”.

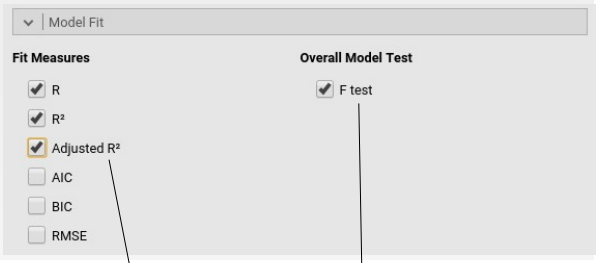
(2) Remembering what was said about sample sizes before, we should be cautious about interpreting our coefficients. In principle, the demanded sample size would be 104 + 3 = 107. We only have 100 cases / measurements. Given that the model has a pretty substantial effect size (R ~ 0.9), this reduces the required sample size. Therefore, we still might be safe to report these coefficients.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- we would like to know how large the adj. R^2 is and whether the model as a whole is significant
  → tick “Adjusted R^2” and “F test”



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Linear Regression

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.903	0.816	0.811	142.164	3	96	< .0001

In addition to R^2 there is a so called adjusted R^2 which can be selected in the drop-down-menu “Model Fit”. The motivation behind calculating the adjusted R^2 is the observation that adding more predictors into the model will always cause the R^2 value to increase (or at least not to decrease). For a regression model with K predictors, fitted to a data set containing N observations, the adjusted R^2 is:

$$adj\ R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}} \times \frac{N-1}{N-K-1}\right)$$

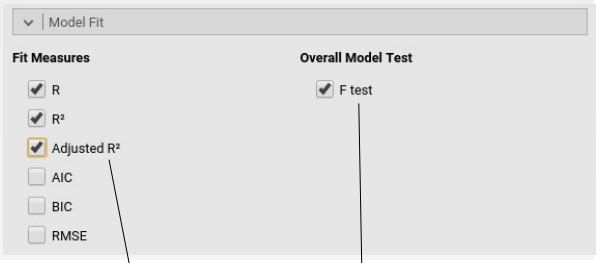
The big disadvantage is that the adjusted R^2 value can’t be interpreted in the elegant way that R^2 can. R^2 has a simple interpretation as the proportion of variance in the outcome variable that is explained by the regression model. If you care more about interpretability, then better report R^2, if correcting for bias is your main concern, then adjusted R^2 is probably better reported.

UNIVERSITY OF BERGEN

Principles and background

how to conduct a linear regression:

- we would like to know how large the adj. R^2 is and whether the model as a whole is significant
  → tick “Adjusted R^2” and “F test”



Linear Regression

Model Fit Measures

Model	R	R^2
1	0.903	0.816

Linear Regression

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.903	0.816	0.811	142.164	3	96	< .0001

Finally, we would like to know whether the model as a whole is significant. When checking the coefficients, we already explored which of the independent variables makes a significant contribution and to what degree.

Now, we are interested in whether the whole model is significant. As said earlier in the theoretical introduction, we are dealing here with a comparison between a null hypothesis based upon a model containing only the mean to an alternative hypothesis where we claim that our model as a whole (i.e., based upon the contribution of all independent variables) makes a better prediction than the one based just on the mean.

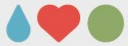
$$F = \frac{SS_{mod}/K}{SS_{res}/(N-K-1)}$$

The whole model is highly significant, indicating that we can make a much better prediction with the help of our independent variables.

Assumption for linear regressions

If something was unclear in that or a previous part, remember that you can ask questions in the discussion for that lecture on MittUiB.

The next part is dealing with assumptions we have to obey in order to ensure that our models are valid.

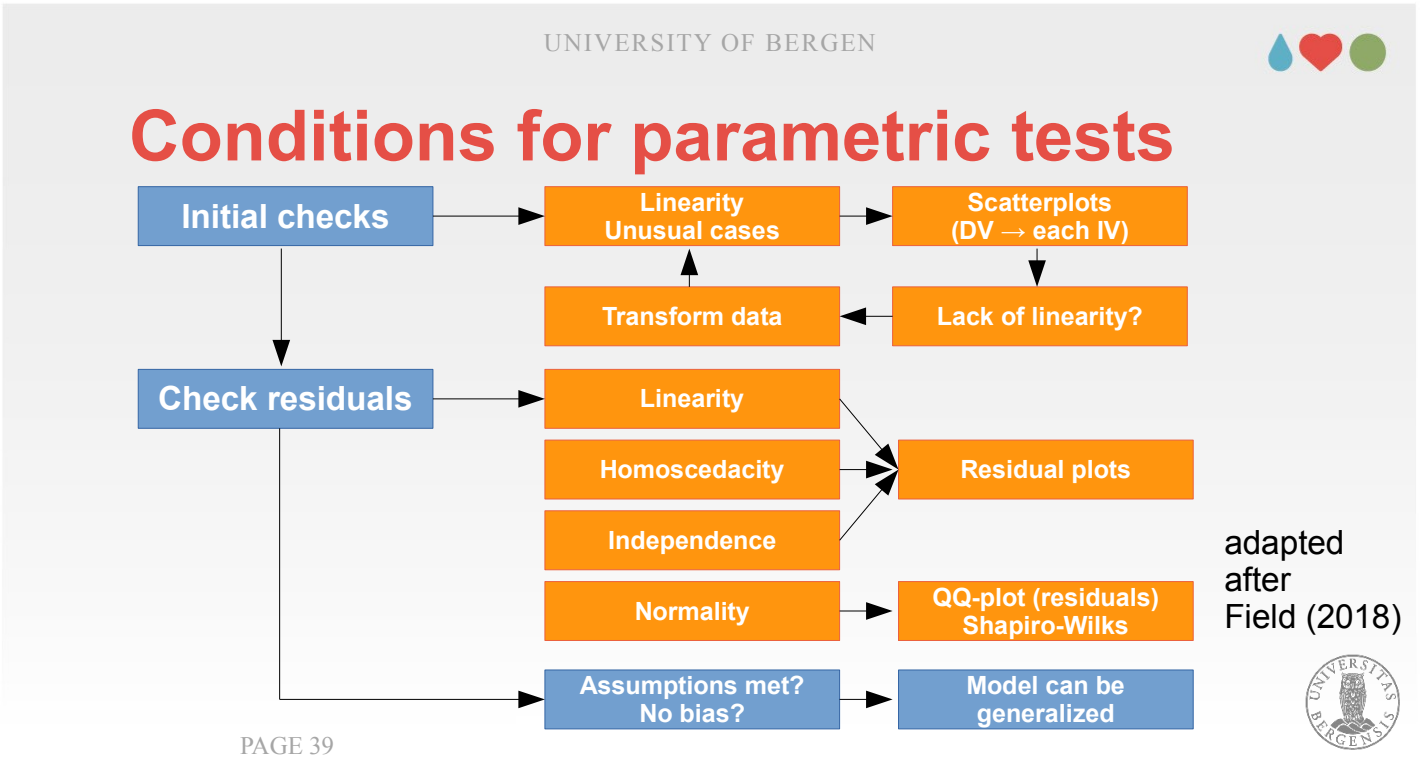


Conditions for parametric tests

- conditions for using parametric tests (such as correlation, regression, t-test, ANOVA)
- if one of these conditions is violated, non-parametric tests have to be used
- robustness against violation of certain assumptions (relatively robust against deviation from **normality**; deviations from **linearity** and **homoscedacity** do not invalidate an analysis but weaken it)

Generally, all parametric methods we use have certain assumptions that are required to be met in order to use those methods. If one of these assumptions (detailed on the next slides) is violated we should consider using non-parametric methods.

That said, most parametric methods have a certain “robustness” against violations of these assumptions. That means we may still use them. However, we should be cautious with our interpretations since our analyses are weakened by violating these assumptions. Three main assumptions are normality, linearity, and homoscedacity.



This overview provides a kind of map where we should go and what we should do.

Generally, the measures fall into two main categories:

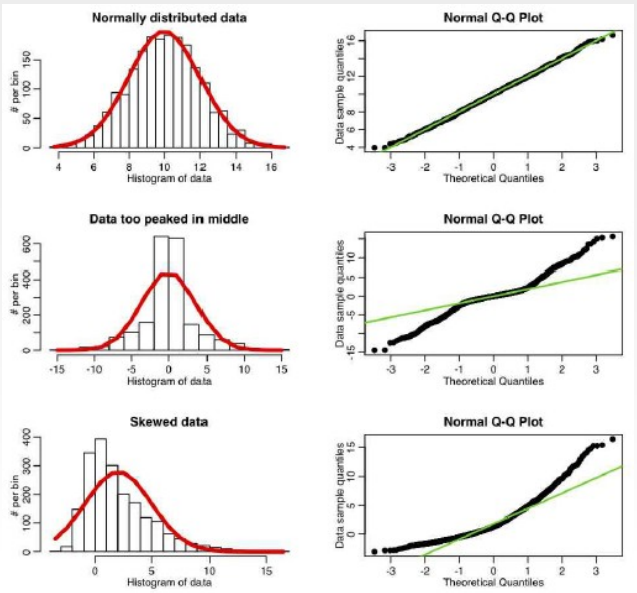
(1) Assumptions that we check before subjecting our variables to our regression analyses.

(2) Assumptions we check within an regression analysis: The latter assumptions chiefly concern the residuals (ε from the formula, i.e., the difference between what we predicted and the real data). A lot takes the form of residual plots (which are scatter plots) where you can visually assess if certain assumptions are violated (leading to a characteristic appearance of the points in the scatter plots). The other aspect is that we have to ensure that the residuals are normally distributed, which can be assessed using the Shapiro-Wilk test and visually in a QQ-plot for those residuals.

If the assumptions are met, we can be reasonably sure that our model is valid and can be generalized.

Conditions for parametric tests

- normality and possible causes for normality violations

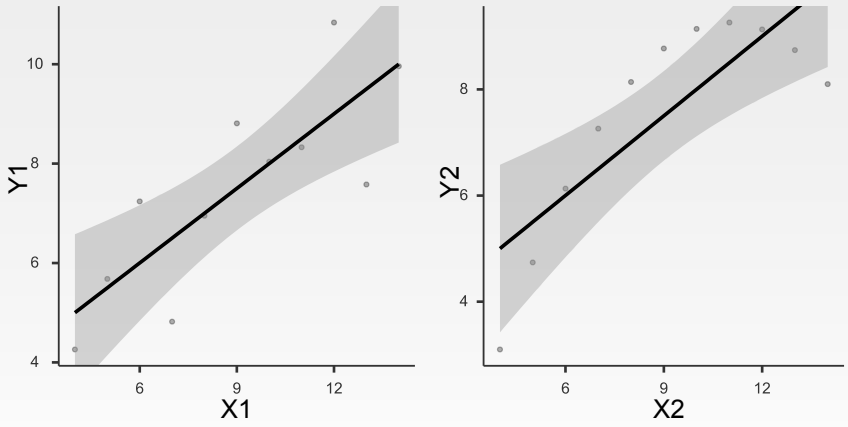


Both the independent and the dependent variables have to be normally distributed. A way to check this statistically is with using the Shapiro-Wilk test (which can be found under “Exploration” → “Descriptives” in the drop-down-menu “Statistics” at bottom right). If this test is not significant, the tested variable doesn’t significantly deviate from a normal distribution and we are safe to use it.

Visually, we can use QQ-plots (also to be found under “Exploration” → “Descriptives” in the drop-down-menu “Plots”, bottom left). If the data are normally distributed, they fall (more or less) on the diagonal line. If they deviate visibly, as in the bottom two examples above, something is wrong, e.g., because our distribution is to “peaky” (i.e., contains an over-proportional amount of data around the mean) or if the distribution is skewed.

Conditions for parametric tests

- linearity (non-linear models are available, but not introduced here)

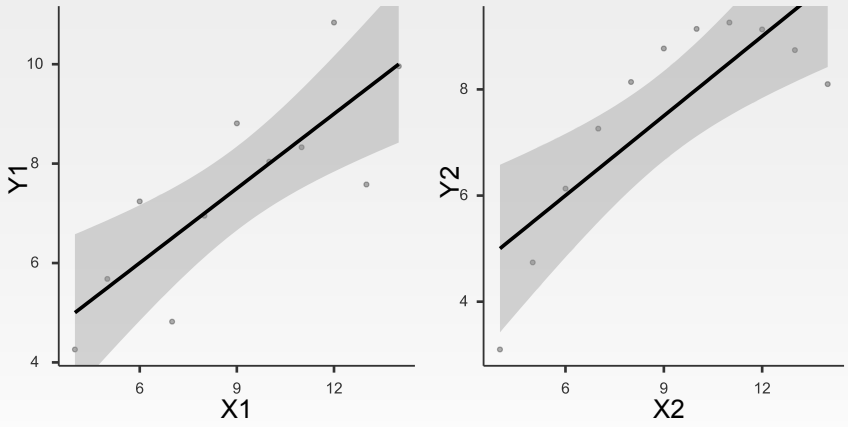


Assessing those relationships is done via scatter plots. The perhaps easiest way to do this is via “Regression” → “Correlation Matrix” and then tick “Correlation matrix” under “Plot”.

Check all combinations of the dependent variable with each independent variable you intend to include in the model. If they show a non-linear relationship (as the right example figure), you should consider not including that independent variable in your prediction.

Conditions for parametric tests

- linearity (non-linear models are available, but not introduced here)

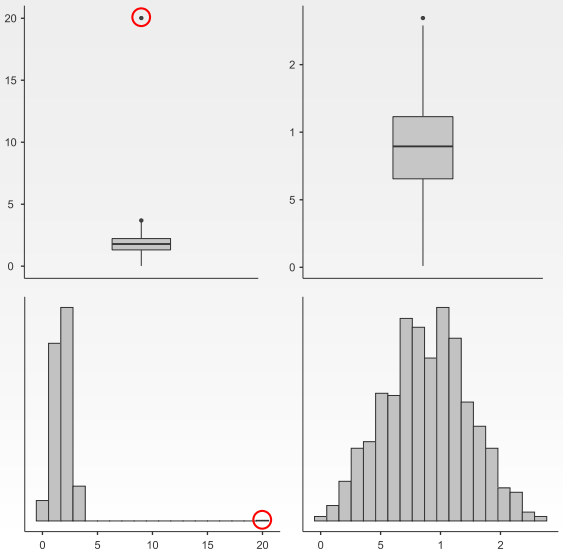


The requirement for linearity (i.e., linear relationships between independent and dependent variables) results from that we use the General Linear Model for our estimation, and the that model describes our predicted value of the dependent variable ($\hat{Y}$) as a linear combination of independent variables (X) multiplied by their weights (B).

A certain robustness can also be seen from the example on the right: The linear regression line that was fitted there does a reasonable job in describing that non-linear relationship. However, violating that assumption also clearly means that our model is weakened because a non-linear model would be much better suited to describe this relation.

Conditions for parametric tests

- consequences of not removing outliers on the skewness (and in consequence the normality) of a distribution

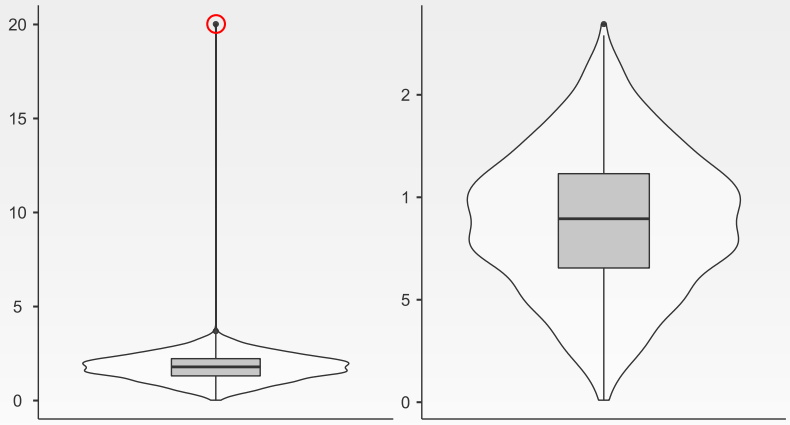


From within “Exploration” → “Descriptives” in jamovi, we can also choose two other plots that help us detect outliers. The plots can be obtained from “Exploration” → Descriptives with ticking “Histogram” and “Box plot” in the drop-down-menu “Plots”

The example on the slides clearly show the effect of one single (and extreme) outlier. In comparison to the right side, where this outlier has been removed and where the data look as if they were pretty much in accordance with a normal distribution is the histogram on the right hand side quite skewed to the left. The outlier is marked in the box plot and the histogram.

Conditions for parametric tests

- a violon- / box-plot combination even allows assessing outliers within one figure

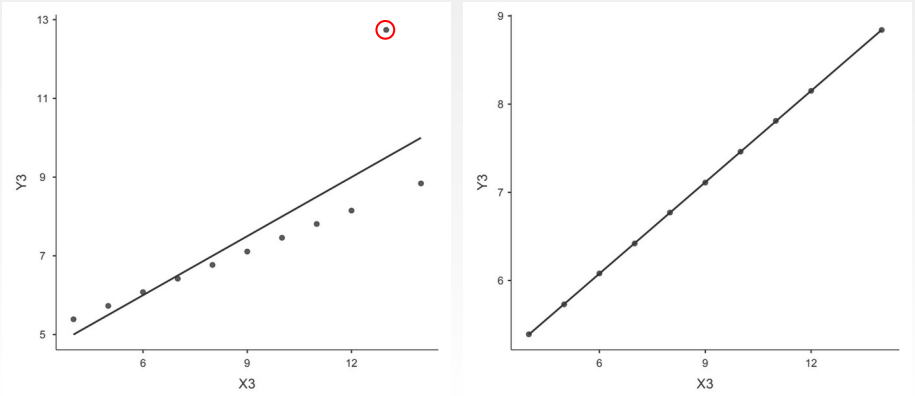


An even more elegant way of displaying outliers is the combination of a box and a violin plot (at least in my opinion). It can be obtained by unticking “Histogram” and ticking “Violin” (under Box plots). In principle is the violin plot a histogram which is turned 90° to the left and smoothed. It makes assessing outliers and the normality of the variable possible within a single figure.

Again, left shows the version with the outlier still contained in the data, right where it has been removed.

Conditions for parametric tests

- consequences of not removing outliers on the slope of a correlation / regression

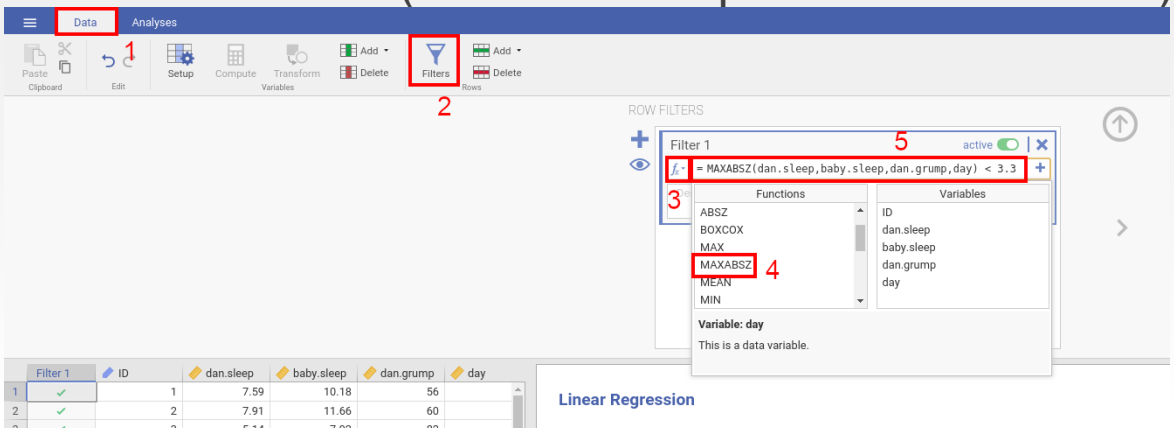


For showing the effect the removal of an outlier has on our regression model, I will use Anscombe's quartet (this time example 3). You can see on the left, that all except one point falls on an (imaginary) regression line that you could think going through all points except the outlier. However, the regression line is "tilted" because the regression model is based upon minimizing the squared deviation over ALL variables. Therefore, the one outlier gets are relative large influence in comparison to the rest of the variables (as those all fall quite close to the regression line). If the outlier is removed, the remaining points all become located perfectly on the regression line (at the same time, the correlation coefficient is increased from 0.816 to 1.000).

Conditions for parametric tests

strategies for removing outliers:

extreme z-values (for 1/1000 → p = .001 → z = ±3.3)



We can adopt a number of different strategies in order to remove outliers. The first one is based upon z-values (i.e., assumptions about how likely the occurrence of certain outliers is given a normal distribution). To ensure that we only exclude cases that are reasonably unlikely (i.e., shouldn't have occurred), we choose 1 / 1000, leading to a p-value of 0.001 and a z-value of 3.3 (on either side, i.e., values below z = -3.3 or above z = 3.3).

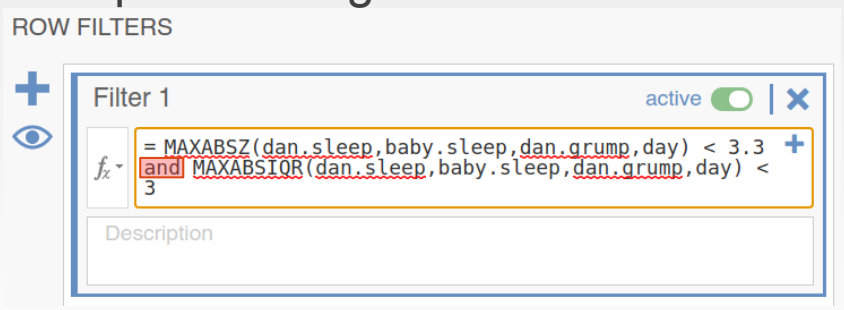
To select and remove such outliers, select the "Data"-tab. You now see the spreadsheet with your data. Select "Filter". This adds a variable and opens an input field. Press f_x and select MAXABSZ(). It takes the maximum of the absolute value of the z-scores of several variables. Include all variables inside the parentheses, separated with commas, and add < 3.3 outside the parentheses:

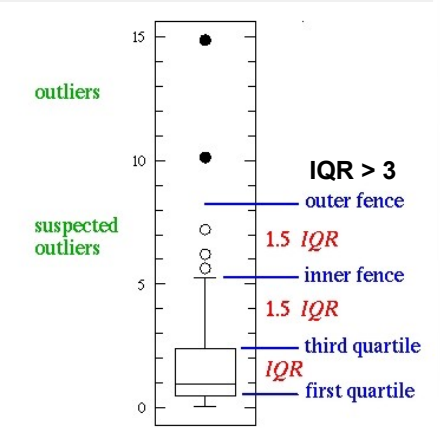
```
MAXABSZ(dan.sleep,baby.sleep,dan.grump,
day) < 3.3
```

Conditions for parametric tests

strategies for removing outliers:

interquartile range





The second strategy is based on how outliers are defined in a box plot. You have the central box, called interquartile range. The interquartile range is from the first (25%) to the third quartile (75%) of your data (i.e., if you would sort your variable and had N = 100; it would be the 26th to the 75th value). This IQR is multiplied with 1.5 and added as whiskers above and below the box. Anything outside that whisker is regarded an outlier (inner fence). We, however, only would like to remove extreme outliers (outside the outer fence; IQR > 3).

Write "and" after MAXABSZ, press f_x and select MAXABSIQR(). It takes the maximum of the absolute value of the IQR of several variables. Include all variables inside the parentheses, separated with commas, and add < 3 outside the parentheses. The whole added part looks like:

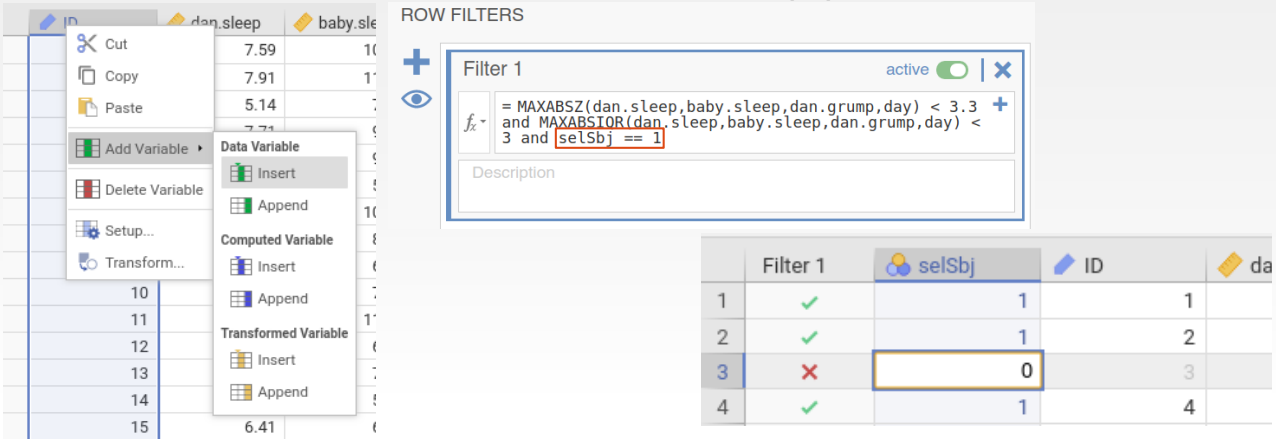
```
and MAXABSIQR(dan.sleep,baby.sleep,
dan.grump,day) < 3
```

UNIVERSITY OF BERGEN

Conditions for parametric tests

strategies for removing outliers:

multivariate: Mahalanobis distance (1)



Before we can do it, we have to add a selection variable. In order to do so, we choose the tab “Data”, go to the header line in our spreadsheet and right-click on ID. There we choose “Add variable” and “Insert” (under “Data variable”). This creates a variable named “A” which we change into “selSbj”. Then we go to the “Filter 1” variable (most to the left) and double click on the header line. This opens the field where we wrote other filter-commands. We append “`and selSbj == 1`”. The filter expression is now:

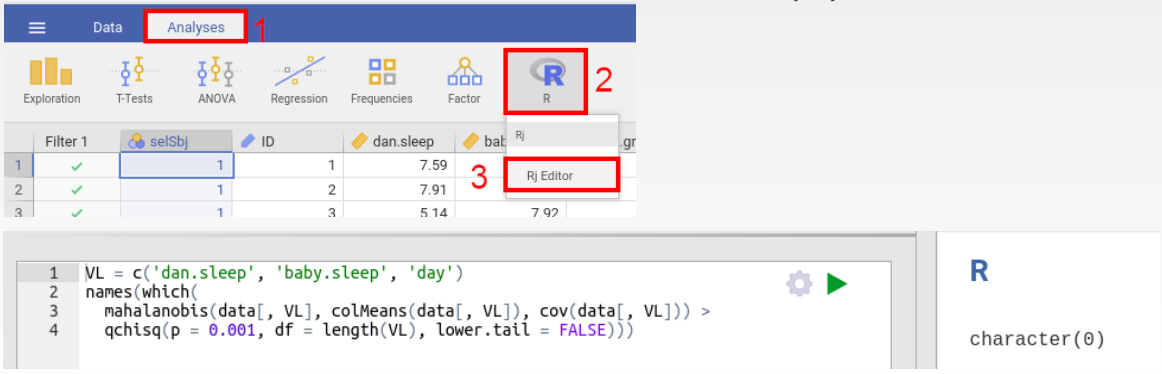
```
MAXABSZ(dan.sleep,baby.sleep,dan.grump,day) < 3.3
and MAXABSIQR(dan.sleep,baby.sleep,dan.grump,day) <
3 and selSbj == 1
```

selSbj is still empty (hence all the red “X”). We don’t want to write “1” one hundred times. We therefore just create an empty computed variable, assign “= 1“ in the input field, copy it’s content to selSbj and atferwards delete that variable. You can now deselect case by changing “1” into “0”.

Conditions for parametric tests

strategies for removing outliers:

multivariate: Mahalanobis distance (2)



The final method for removing outliers is using a measure called Mahalanobis-distance. It considers whether a combination of your independent variables are outliers. Think, e.g., of a 180 cm tall person weighing 48 kg. The Mahalanobis distance is aiming to find those cases.

The code is actually a one-liner split into 3. It begins with “names(which”. “names(which” is a command showing you the line number in you data set.

“pchisq” reports the p-value for the chi-squared value that the mahalanobis-function returns. To hit the threshold, p has to be < 0.001. The Mahalanobis-function is the distance between the variable-values of a participant from the mean for each variable (squared and summed up), divided by the covariance between the variables. The code is on MittUIB → Syntax_Outliers_Mahalanobis.R

You can copy-paste it from there.

Conditions for parametric tests

strategies for removing outliers:

multivariate: Mahalanobis distance (2)


```
1  VL = c('dan.sleep', 'baby.sleep', 'day')
2  names(which(
3    mahalanobis(data[, VL], colMeans(data[, VL]), cov(data[, VL])) >
4    qchisq(p = 0.001, df = length(VL), lower.tail = FALSE)))
```

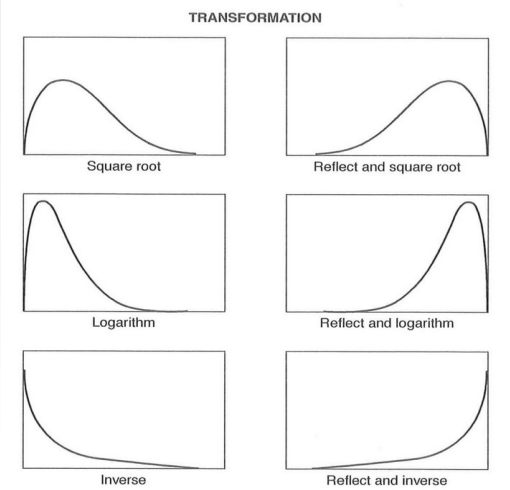
R

```
character(0)
```


Again, we did not find any cases that were selected as outliers. Therefore, we can leave selSbj as it is and don't have to set any values to "0".

Conditions for parametric tests

transforming your data:



Moderate positive skewness	NEWX=SQRT(X)
Substantial positive skewness	NEWX=LG10(X)
With zero	NEWX=LG10(X + C)
Severe positive skewness	NEWX=1/X
L-shaped With zero	NEWX=1/(X + C)
Moderate negative skewness	NEWX=SQRT(K – X)
Substantial negative skewness	NEWX=LG10(K – X)
Severe negative skewness J-shaped	NEWX=1/(K – X)

C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

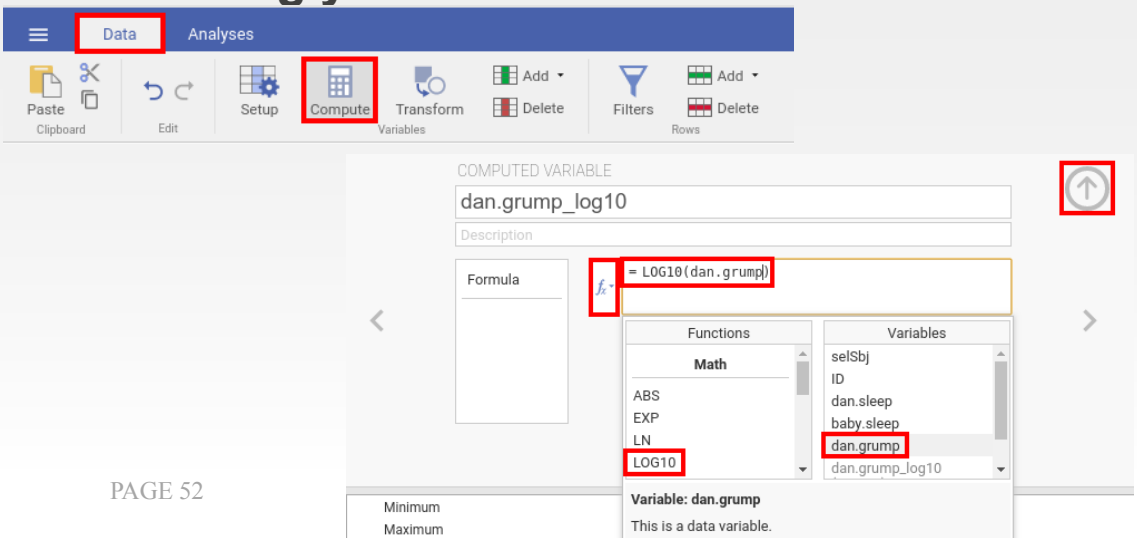
LG10 = LOG10

In certain cases, removing outliers is not sufficient because the values over the whole group of participants are skewed. On the right hand side, you can see examples for how such skewed distributions would look like and on the right hand side, which transformation could be conducted as possible remedy.

Please note that the LG10-function is named LOG10 in jamovi.

Conditions for parametric tests

transforming your data:



In order to do such a transformation, select the "Data"-tab, go to some place in your spreadsheet. Once you press Compute, a new variable is added and you will find the window shown at the bottom.

There you press the "f_x"-button and select the transformation you wish to carry out. I chose LOG10 as an example, but SQRT is also available. You could also write transformations like 1 / dan.grump (or click that in from the variable list).

Once you are satisfied, close the window with the arrow in the top-right corner.

Conditions for parametric tests

Assumption checks (within the regression model):

- Cook's distance
- Autocorrelation test
- Collinearity statistics: VIF
- Normality: Shapiro-Wilk, QQ-plot (of the residuals)
- Residual plots: Residuals vs. Fitted and all DVs and IVs

In addition to the assumptions we did check in preparation for conducting a regression analysis (what we did up to now), we can also check further assumptions from "within" the model.

Conditions for parametric tests

Assumption checks (within the regression model):

- **Cook's distance**
- Autocorrelation test
- Collinearity statistics: VIF
- Normality: Shapiro-Wilk, QQ-plot (of the residuals)
- Residual plots: Residuals vs. Fitted and all DVs and IVs

The first two I won't cover in too much detail. Cook's distance describes for each participant the difference between the prediction if the participant is included vs. excluded.

This means, first, a regression model is build with all participants included. Then, for each participant, one after another, that participant is excluded from estimating the weights for that model, but that model (in which the participant did not contribute to the estimation) is used to predict the value for that participant. If the result from that model differs considerably from those of the first model, the participant likely contains outliers. Typically, finding values above 1 under max should raise red flags. Then you should check for outliers again.

Conditions for parametric tests

Assumption checks (within the regression model):

- Cook's distance
- **Autocorrelation test**
- Collinearity statistics: VIF
- Normality: Shapiro-Wilk, QQ-plot (of the residuals)
- Residual plots: Residuals vs. Fitted and all DVs and IVs

The autocorrelation test (Durbin-Watson) particularly applies when we are dealing with time series. Let's assume we measured job satisfaction over several years (using the same questionnaire). The years are listed in the rows of our data sheet. We expect that those values are correlated (which is denoted as autocorrelated). If the Durbin-Watson test gets significant we have such a situation. It primarily means that our statistics are affected and we should mention it as a limitation when reporting our results.

Conditions for parametric tests

multicollinearity and singularity:

- regression is impossible if IVs are singular (i.e., a linear combination of other IVs) or unstable if they are multicollinear
- collinearity describes a linear association between explanatory variables (i.e. the degree to which one explanatory variable can be predicted by a combination of one or more other explanatory variables)

Collinearity or multicollinearity concerns the correlation between the independent variables.

I said earlier that regression analyses requires independent variables that each are correlated relatively high with the dependent variable but relatively low among themselves.

In the most extreme case, a variable is singular. That situation describes if that variable can be predicted as a combination of the other variables. That typically makes our estimation fail because certain matrix operations are not possible with singular matrices.

In a less serious case, multicollinearity, variables are very highly correlated among themselves. This has a tendency to make our regression models un-stable (i.e., it will change considerably depending on whether certain independent variables are included or not and in which order the are included).

Conditions for parametric tests

multicollinearity and singularity:

- tolerance: 1 – R_j^2 (R_j^2: what degree of variance of variable j is explained by the other predictor variables) variance inflation factor (VIF): 1 / tolerance
- (a) VIF < 5 and tolerance > 0.2; (b) the average of the VIF of all variables should be close to 1
- variable removal should consider also reliability and cost of acquisition

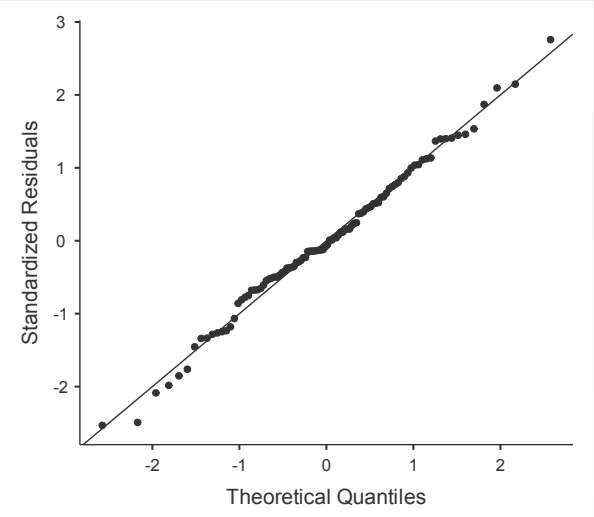
Collinearity Statistics

	VIF	Tolerance
dan.sleep	1.674	0.597
baby.sleep	1.658	0.603
day	1.014	0.986

We can assess the amount of multicollinearity by using the tolerance and the variance inflation factor (VIF). A VIF of above 5 should raise red flags and make us consider to exclude that variable. If the average of the VIF is much above 1, this should also elicit you to further check what variable might be the reason for the collinearity.

The VIF assesses the common variance among the variables. As a consequence, we may have several choices which variable we remove. When deciding about which variable we should remove, we should also consider how reliable the variable is (and rather remove unreliable ones) or how costly the acquisition was (we possibly rather would remove a variable that did not take so much time and money to collect).

Conditions for parametric tests

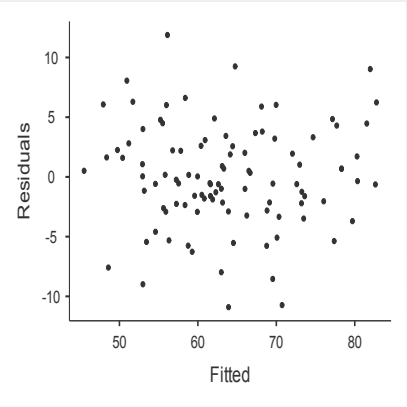


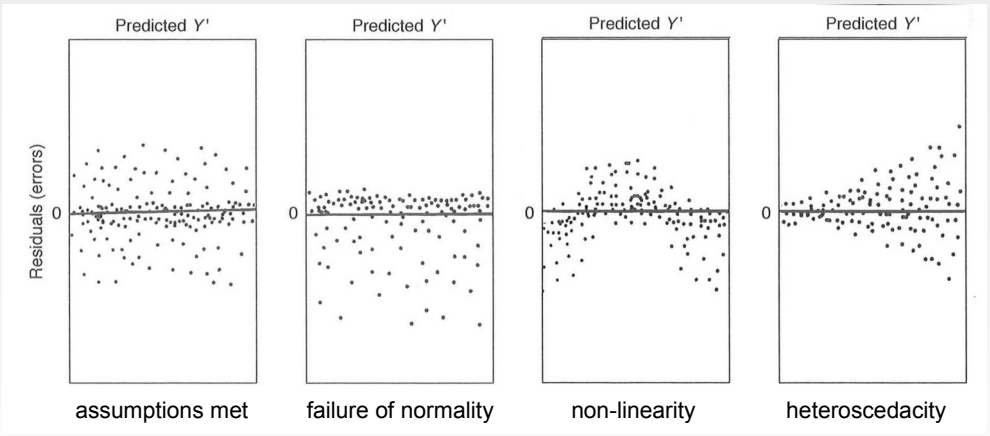
Normality test (Shapiro-Wilk)	
statistic	**p**
0.992	0.8352

The remaining part of the assumption checks deals with the residuals, i.e., the difference between our prediction and the real value.

One assumption is that the residuals should be normally distributed. That can be checked using the Shapiro-Wilk test or visually using the QQ-plot. In the plot, the majority of data points should fall on (or close to) the main diagonal. Both measures indicate that there are no problems with the normality of the residuals for the current analysis.

Conditions for parametric tests





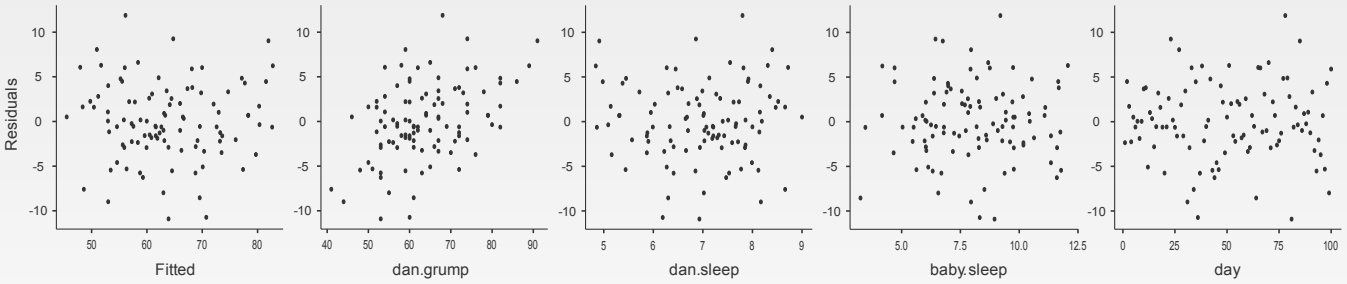
The next task is to check the scatter plots for the residuals. There is a considerable number of them: One for the predicted (Fitted) values vs. the residuals, shown most to the left. The plots should look like a cloud of dots. Our example looks similar to the plot described as “assumptions met”. Our example on the left doesn’t have a line at where the Residuals have the value 0 but the position is roughly aligned so that is should be relatively easy to compare.

If the residuals were not normally distributed, they rather would scatter either above or below the zero-line instead of being (relatively) equally distributed on both sides.

If they were non-linear, we likely would obtain a curved shape in the plot.

If homoscedacity (equality of variances) isn’t given they would have a wider spread at a certain point along that line (it can be left, right, or in the middle).

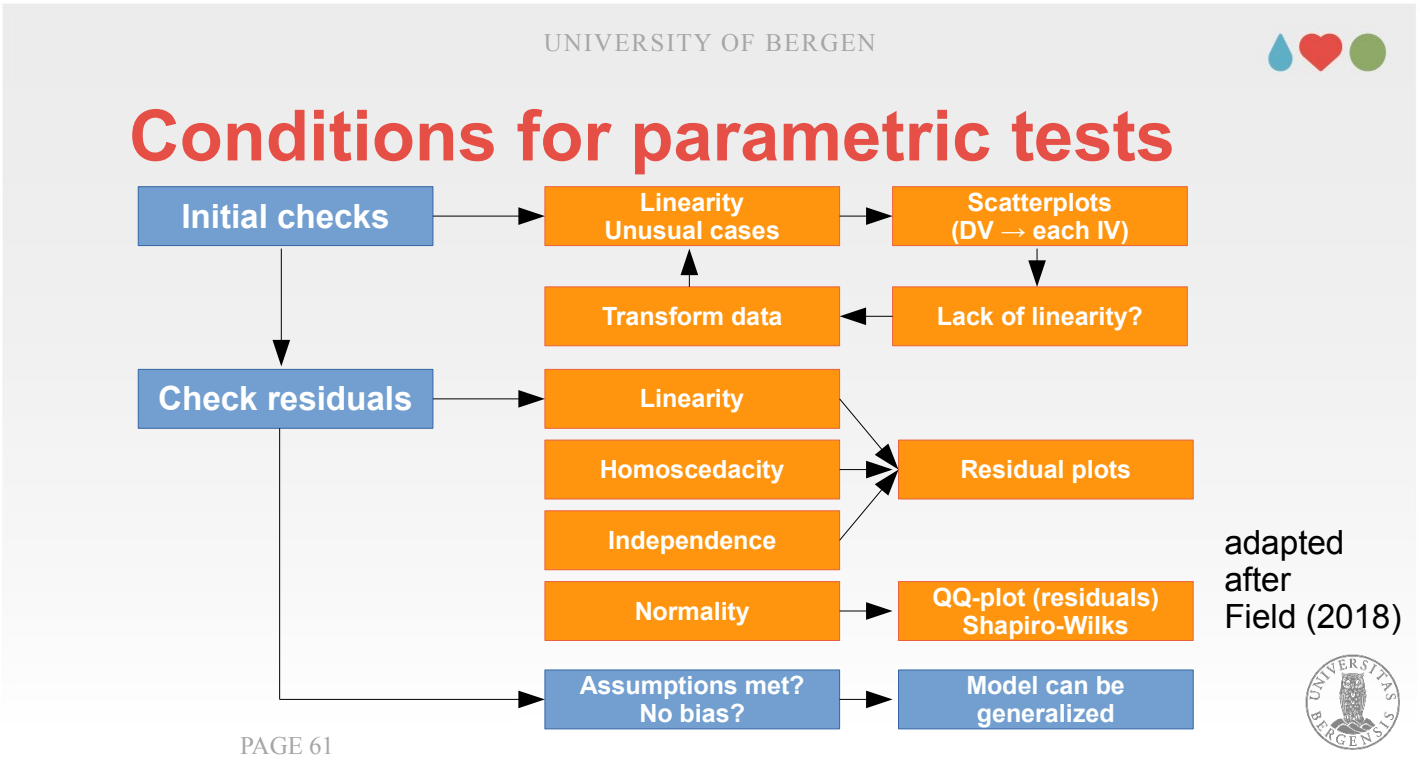
Conditions for parametric tests



all scatter plots should look as if they were randomly distributed and not like of the three figures on the right side of the previous slide

After we assessed the relation between the residuals and our prediction, we do the same for the remaining figures (always keeping the examples on the previous page in our head in order to detect abnormalities).

Left, we see the scatter plot for Residuals vs. Fitted again (already checked on the previous page), then Residuals vs. the “real” Dependent variable (dan.grump), and then the Independent variables (dan.sleep, baby.sleep, day). None of the plots looks as to raise concerns. All are looking like clouds with correlations near 0 and there seem to be not any pattern in those clouds. The only combination that looks a bit more regular is residuals vs. dan.grump. But even that is no reason for concern: It is quite normal that we see a bit of a linear relation in that plot. The values around the mean a typically estimated better whereas the residuals get larger towards the ends.



That was quite an extensive overview. Let‘s come back to the overview what we have to check.

Before we start, we should assess normality and find and remove outliers. If the variable still doesn’t follow a normal distribution we may consider transforming it.

Then, we start assembling our regression model. All options for possible assumption checks are collected within the drop-down-menu named “Assumption checks”. The most important of those checks are Collinearity diagnostics, Shapiro-Wilk, QQ-plot and residual plots where we can assess how much our model gives reason for concern. There are certain measures (e.g., removing variables) to provide remedy.

Either we manage to fulfil all assumptions and can whole-heartedly claim that the model can be generalized. Otherwise, we have to report what gives reason for concern.

Regression types and model building

A statistician and a good looking fellow are having small talk at a party. “What are you doing?”, the statistician asks. “Modelling”, the good looking guy replies. “Oh, that‘s interesting”, says the statistician, “I am modelling too.”

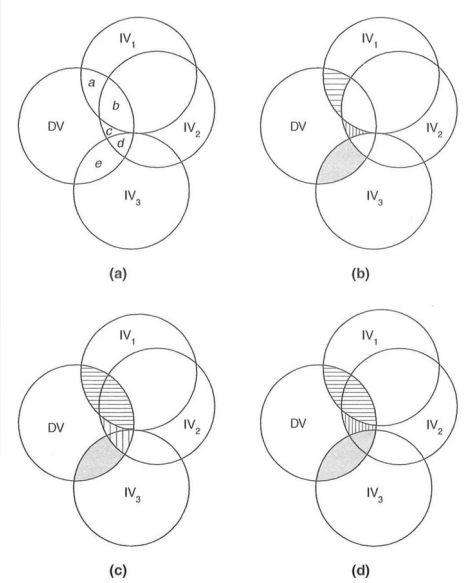
Before we proceed to the next part, you can take a breath and check whether only statisticians find that joke funny.

Major types of multiple regression

three analytic strategies:

- standard (b)
- hierarchical (c)
- statistical (d)

differ in how the IVs contribution to the prediction is weighed



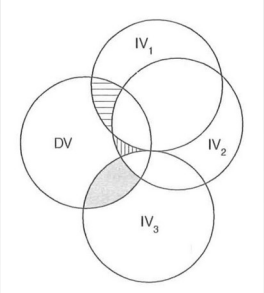
When doing regression analyses we have the choice of three strategies that differ in how they "distribute" the contribution of the different independent variables. Depending on how many variables you have and how much they correlate, you may end up with a considerable number of these sections denoted with the small letters in the top-left figure.

You will also see that the areas that are marked grey differ between the figures denoted (b), (c) and (d). That indicates that those three analytic strategies differently weigh the contributions of the individual independent variables.

Major types of multiple regression

standard regression:

- enters all IVs at once in the equation
- only unique contributions are considered (may make the contribution of a variable look unimportant due to the correlation with other IVs, e.g., IV_2)



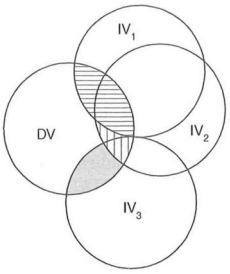
When using “standard”, which is the default in jamovi, only the parts where the overlap is unique between a independent variable and the dependent variable are considered.

Other parts where two independent variables correlate are not considered. We could also say that only those contributions of one variable are taken into account that are not shared with another variable.

Major types of multiple regression

hierarchical regression:

- enters IVs in an order specified
  can be entered separately or in blocks according to logical or theoretical considerations, e.g. experimentally manipulated variables before nuisance variables, the other way round, or comparing different sets
- additional contribution of each IV is considered



Hierarchical regression is the method we will describe in more detail in the later part of this section.

How the independent variables are added to and weighed in the model is subject to theoretical considerations.

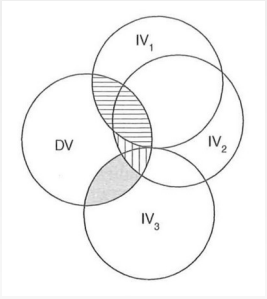
That is, the independent variable which is added first, should have been selected on theoretical grounds to be the one which is regarded most explanatory. The ones that are added afterwards should follow in the order of their assumed importance.

Typically, the question asked when adding a new variable takes the form of: Does it make a difference (i.e. does it result in better prediction) if I add variable X as predictor.

Major types of multiple regression

hierarchical regression:

- enters IVs in an order specified
  can be entered separately or in blocks according to logical or theoretical considerations, e.g. experimentally manipulated variables before nuisance variables, the other way round, or comparing different sets
- additional contribution of each IV is considered



By ordering variables after the relative “importance” you give them in your hypotheses, you typically are most interested to see whether your data support your main hypothesis, afterwards you would assess additional hypotheses and in the end check whether there were any further variables that you wished to control for because you regarded them nuisance variables exerted an influence.

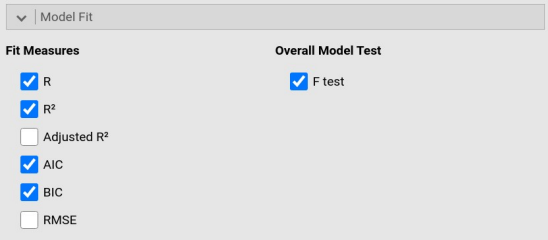
Mathematically, this is reflected that the contribution of the first variable is weighted highest, the second one only contributes aspects that are not already considered by the first variable and so on.

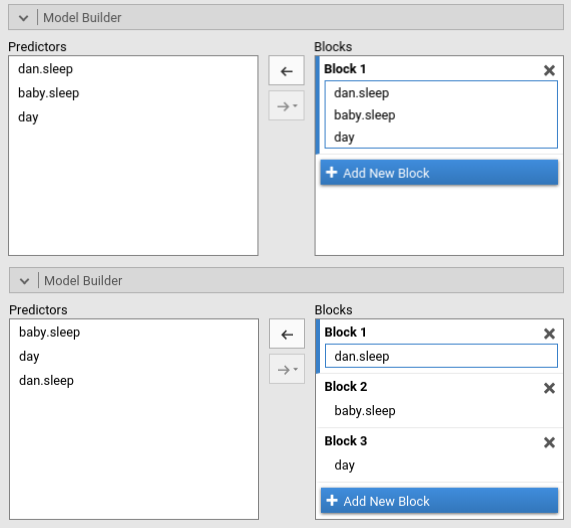
UNIVERSITY OF BERGEN

Major types of multiple regression

hierarchical regression:

- top: all variables entered at once in the equation
- bottom: define the order with which the variables are included in the model (based upon theor. consid.)





An example for how such a hierarchical regression is conducted in jamovi can be seen on that slide. We use the dataset with the relation between Dan's grumpiness, the amount of Dan sleep, of the babies sleep and how old the baby was.

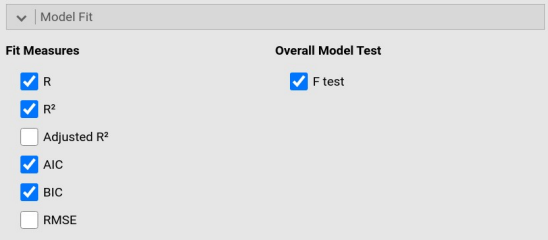
It would be logically to hypothesize that for Dan's grumpiness, the amount Dan is sleeping is most predictive (in the end, it happens within the same person). Afterwards, it could be assumed that the amount to which the baby is sleeping might have the second largest influence, with the age of the baby coming last. In order to test those hypotheses, we would assign dan.grumpy to "Dependent variables" and all variables that we expected to be possibly help to explain Dan's grumpiness to "Covariates" (i.e., the variables dan.sleep, baby.sleep and day).

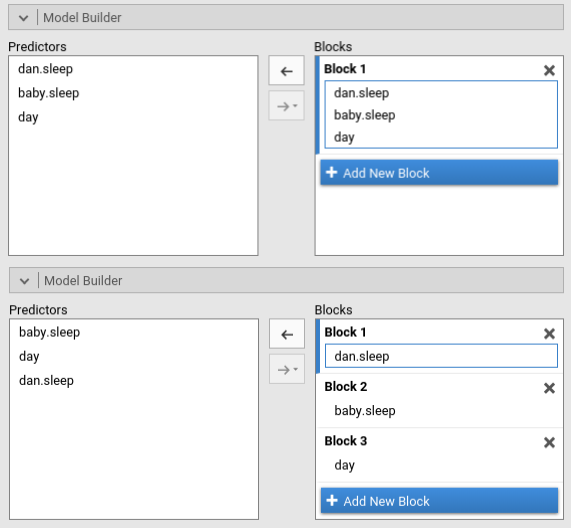
UNIVERSITY OF BERGEN

Major types of multiple regression

hierarchical regression:

- top: all variables entered at once in the equation
- bottom: define the order with which the variables are included in the model (based upon theor. consid.)





Then we go to the drop-down-menu “Model Builder”. Per default, all variables are assigned to one block, assessing the contribution of all variables at the same time (but only considering the unique contributions of each variable, see the figure under “standard regression” three slides back).

We remove the variables that we expect not to make the highest contributions (i.e., we remove baby.sleep and day). Once we did that, only dan.sleep is assigned to “Block 1”. We choose “Add New Block”, and assign baby.sleep to “Block 2”, and “Add New Block” again, assigning day to “Block 3”.

Then open the drop-down-menu “Model Fit” and tick / switch on “AIC”, “BIC”, and “Overall Model Test”.

Major types of multiple regression

hierarchical regression:

- models defined in the model builder are compared against each other
- the bottom table indicates whether a new model increases the quality of prediction

Linear Regression

Model Fit Measures

Model	R	R^2	AIC	BIC	Overall Model Test F	df1	df2	p
1	0.903	0.816	580.953	588.768	434.906	1	98	< .0001
2	0.903	0.816	582.951	593.372	215.238	2	97	< .0001
3	0.903	0.816	584.865	597.890	142.164	3	96	< .0001

Model Comparisons

Comparison Model		Model	ΔR^2	F	df1	df2	p
1	-	2	2.858e-6	0.002	1	97	0.9691
2	-	3	1.593e-4	0.083	1	96	0.7736

After these preparations, we can now (theoretically) describe three different possible criteria to select which of the different blocks or (models as they are called in the output) makes an impact with the variables in it. I will say a little about WHY we use these criteria before continuing with taking them into praxis for our parenthood data set.

Typically the decision is made by weighing three factors: (1) The model as a whole has to be significant – you will find that information as F- and associated p-value in the table “Model Fit Measures”. If the whole model (listed with the model number in the table) is not significant, then there is no need for further considerations.

Major types of multiple regression

hierarchical regression:

- models defined in the model builder are compared against each other
- the bottom table indicates whether a new model increases the quality of prediction

Linear Regression

Model Fit Measures

Model	R	R^2	AIC	BIC	Overall Model Test F	df1	df2	p
1	0.903	0.816	580.953	588.768	434.906	1	98	< .0001
2	0.903	0.816	582.951	593.372	215.238	2	97	< .0001
3	0.903	0.816	584.865	597.890	142.164	3	96	< .0001

Model Comparisons

Comparison Model		Model	ΔR^2	F	df1	df2	p
1	-	2	2.858e-6	0.002	1	97	0.9691
2	-	3	1.593e-4	0.083	1	96	0.7736

(2) Among the models that are significant, it is appropriate to choose the model which is most informative and parsimonious. The claim to be informative and parsimonious is denoted as Ockham's razor, saying "Entities should not be multiplied without necessity". Applied to a regression model that means that you should not add predictors just because they lead to a small increase in your R^2. There are measures that help to assess the amount of information contained in your models: jamovi implements AIC and BIC (Akaike and Bayes Information Criterion; both can be found and selected in the drop-down menu "Model Fit").

Major types of multiple regression

hierarchical regression:

- models defined in the model builder are compared against each other
- the bottom table indicates whether a new model increases the quality of prediction

Linear Regression

Model Fit Measures

Model	R	R^2	AIC	BIC	Overall Model Test F	df1	df2	p
1	0.903	0.816	580.953	588.768	434.906	1	98	< .0001
2	0.903	0.816	582.951	593.372	215.238	2	97	< .0001
3	0.903	0.816	584.865	597.890	142.164	3	96	< .0001

Model Comparisons

Comparison Model		Model	ΔR^2	F	df1	df2	p
1	-	2	2.858e-6	0.002	1	97	0.9691
2	-	3	1.593e-4	0.083	1	96	0.7736

cont. (2): A problem of regression models is that it is possible to increase the likelihood (or the goodness of fit) for that model by adding variables. However, adding variables may result in overfitting (i.e., the model works very well for that specific set of data, but doesn't generalize because it is so much "tailored" to that specific data set). Another matter is that the more predictors you include in your model (i.e., the more complex your model is) the more difficult it is to interpret it or to explain it. Both AIC and BIC try to resolve that by introducing a penalty term for the number of parameters in the model (the penalty term is larger in BIC). For both of them is true: The LOWER the value, the more appropriate is the model. Typically, both of them point in the same direction.

Major types of multiple regression

hierarchical regression:

- models defined in the model builder are compared against each other
- the bottom table indicates whether a new model increases the quality of prediction

Linear Regression

Model Fit Measures

Model	R	R^2	AIC	BIC	Overall Model Test F	df1	df2	p
1	0.903	0.816	580.953	588.768	434.906	1	98	< .0001
2	0.903	0.816	582.951	593.372	215.238	2	97	< .0001
3	0.903	0.816	584.865	597.890	142.164	3	96	< .0001

Model Comparisons

Comparison Model		Model	ΔR^2	F	df1	df2	p
1	-	2	2.858e-6	0.002	1	97	0.9691
2	-	3	1.593e-4	0.083	1	96	0.7736

(3) Finally, there are F- and associated p-values that indicate whether the changes made from the first (simpler) to the second (more complex) model are significant. Those values are listed in the table “Model Comparisons”. If, e.g., the comparison of Model 2 – Model 1 is significant then Model 2 makes some contribution to explain your data IN ADDITION to what already was covered by Model 1.

Major types of multiple regression

hierarchical regression:

- models defined in the model builder are compared against each other
- the bottom table indicates whether a new model increases the quality of prediction

Linear Regression

Model Fit Measures

Model	R	R^2	AIC	BIC	Overall Model Test F	df1	df2	p
1	0.903	0.816	580.953	588.768	434.906	1	98	< .0001
2	0.903	0.816	582.951	593.372	215.238	2	97	< .0001
3	0.903	0.816	584.865	597.890	142.164	3	96	< .0001

Model Comparisons

Comparison Model		Model	ΔR^2	F	df1	df2	p
1	-	2	2.858e-6	0.002	1	97	0.9691
2	-	3	1.593e-4	0.083	1	96	0.7736

We now apply these principles to our parenthood data set. (1) When checking the table “Model Fit Measures”, we see that all three models are highly significant ($p < 0.001$ for all). No model already has to be excluded based on that it was not significant

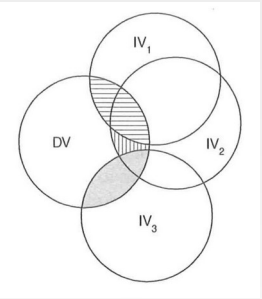
(2) When checking the values for AIC and BIC, both indicate the same: model 1 is the most parsimonious model (and receives the lowest value).

(3) When we look at the comparison between the models in the table “Model Comparisons”, it turns out that neither the comparison of Model 2 to Model 1 is significant ($p = 0.969$), nor is the comparison between Model 2 and Model 3 ($p = 0.774$). Both fail statistical significance by quite some margin: Neither does adding baby.sleep provide further power to predict Dan’s grumpiness (in addition to what is already explained by Dan’s amount of sleep), nor does age in addition to the previous model (with dan.sleep and baby.sleep as predictors).

Major types of multiple regression

statistical regression:

- controversial; order of entry (or possibly removal) specified by statistical criteria
- three versions: forward selection, backward deletion, stepwise regression
- tendency for overfitting → requires large and representative sample; should be cross-validated (R^2 discrepancies indicate lack of generalizability)



A final type of regression approach is called statistical. It is controversial given that you don‘t analyse your data based on hypotheses but use statistical criteria to determine which independent variables are chosen and why.

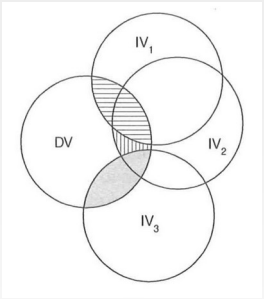
jamovi does not implement automated variable selection methods even though most other statistical programmes (e.g., SPSS) offer it and even though it would have been relatively simple to implement.

Statistical regression is a bit of a poisoned chalice as you are made believe that these methods objectively select appropriate models. However, these methods are quite often used as an excuse for thoughtlessness, caused by not having to think carefully about which predictors you add to the model and what the theoretical basis for their inclusion might be.

Major types of multiple regression

statistical regression:

- controversial; order of entry (or possibly removal) specified by statistical criteria
- three versions: forward selection, backward deletion, stepwise regression
- tendency for overfitting → requires large and representative sample; should be cross-validated (R^2 discrepancies indicate lack of generalizability)



There are three selection approaches: *Forward selection*, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

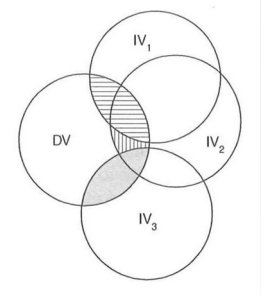
Backward elimination involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically insignificant loss of fit.

Stepwise regression is a combination of the above, testing at each step for variables to be included or excluded.

Major types of multiple regression

statistical regression:

- controversial; order of entry (or possibly removal) specified by statistical criteria
- three versions: forward selection, backward deletion, stepwise regression
- tendency for overfitting → requires large and representative sample; should be cross-validated (R^2 discrepancies indicate lack of generalizability)

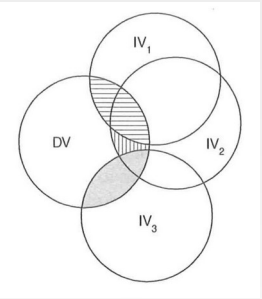


There are two central issues with statistical regression: One is that it may result in overfitting (i.e., the model works very well for a specific set of data, but doesn't generalize because it is so much "tailored" to that data set).

One way to address this issue is to cross-validate your model. For doing so, you would (ideally) collect two different data sets and then use either of these to estimate your regression model. If there aren't any huge discrepancies in the R^2-values of these two estimated models (and if you on top of that got a similar selection of variables with similar regression coefficients included in your model) then you can be reasonably sure that the model you got would also generalize to other cases.

Major types of multiple regression

statistical regression:



- controversial; order of entry (or possibly removal) specified by statistical criteria
- three versions: forward selection, back-ward deletion, stepwise regression
- tendency for overfitting → requires large and representative sample; should be cross-validated (R^2 discrepancies indicate lack of generalizability)

The other issue is that very little agreement exists on what represents appropriate model selection criteria. Available methods include: F-tests, AIC, BIC, NML (Normalized Maximum Likelihood) or (if you're a Bayesian) posterior odds ratios.

When using the hierarchical regression, we weighed two F-tests and the AIC / BIC for making our decision. Combining and weighing them reflects that neither is sufficient as a single criterion.

Major types of multiple regression

choosing regression strategies:

- ***standard***: simply assess relationships (atheoretical)
  what is the size of the overall relationship between IVs and DV?
- ***hierarchical***: testing theoretical assumptions or explicit hypotheses (IVs can be weighted by importance)
  how much does each variable uniquely contribute?
- ***statistical***: model-building (explorative, generating hypotheses) rather than model-testing
  can be very misleading unless based on large, representative samples
  can be helpful for identifiying multicollinear / singular vars.
  what is the best linear combination of variables / best prediction?

The three approaches differ a little with regards to what questions you can answer with them and what possible pitfalls are.

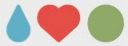
The first approach, standard, isn't based upon any theoretical considerations but simply answers the question of which size the overall relationship between independent and dependent variables is.

With the second approach, hierarchical, you can directly explore hypotheses about the relative contribution of each variable. This approach is driven by theories that you have about the effect each independent variable has on the outcome.

In the last approach, your model building is based upon your data. It can serve explorative purposes and generate hypotheses but isn't really suitable for hypothesis testing. It requires large samples and is subject to possible overfitting. However, it can be helpful when it is required to identify multicollinear or singular variables that should be removed.

Summary and literature

We will end with a brief summary and an overview of the literature for this topic.



Summary

- **introduction**
- **principles and background**: mathematical background, choosing IVs, required sample sizes
- **how to conduct a linear regression?**
- **assumptions for linear regressions**: initial checks, checks within a regression model
- **regression types and model building**: standard, hierarchical, statistical

We started with introducing regression analyses in relation to other statistical methods introduced in the course.
Then, we discussed the mathematical background, described criteria for choosing independent variables and recommended sample sizes.
Afterwards, we turned to a demonstration how to conduct a linear regression.
Next we discussed assumptions for regression analyses and how they can be checked.
Then, we spoke about three regression approaches, and introduced one – hierarchical – more extensive with a practical example.

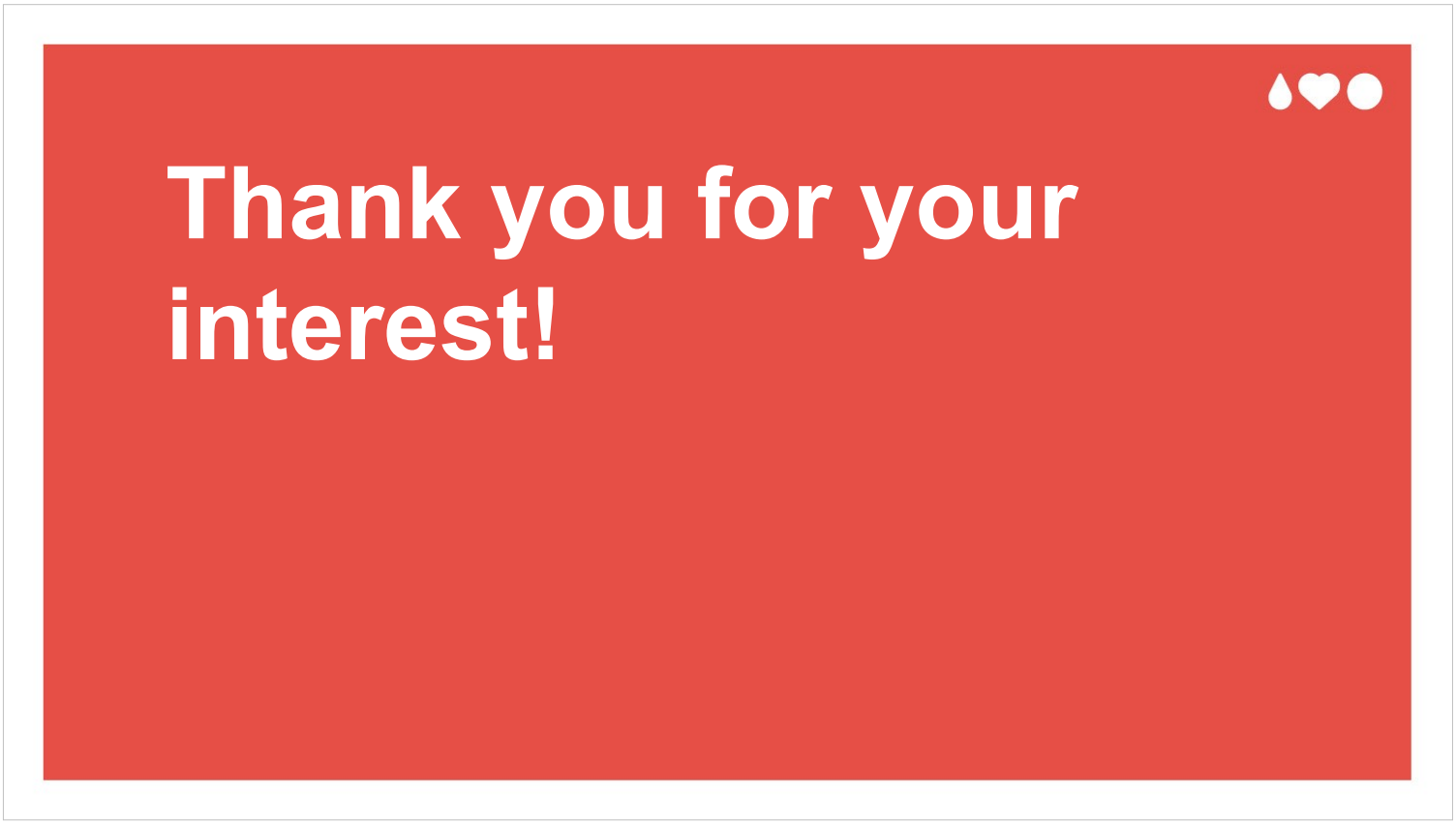
Literature

Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with jamovi*. https://doi.org/10.24384/hgc3-7p15 (Ch. 12; p. 281 – 326)

Aron, A., Coups, E. J., & Aron, E. (2013). *Statistics for psychology* (6th ed). Pearson. (Ch. 11, 12; p. 487 – 5964

Main literature for the lecture was chapter 12 of the jamovi-book (Navarro & Foxcroft, 2022).

For an alternative way of explaining or if you are interested in linear regression analyses would be conducted in SPSS, you can read chapter 11 and 12 in Aron, Aron and Coups (2013).



Thank you very much for you interest! If you have any questions, you can use the discussion for this lecture on MittUiB.

UNIVERSITY OF BERGEN