Advanced R

Day 2

Sereina Herzog

Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

05.03.2024

Course Content - Advanced R (Day 2)

- Statistical tests & models

Course Content - Advanced R (Day 2)

- Statistical tests & models
- Simple linear regression

Course Content - Advanced R (Day 2)

- Statistical tests & models
- Simple linear regression
- Data import & preparation

Statistical tests & models

Statistical tests & models

Which statistical tests and models are suitable for your research questions?

Statistical tests & models

Which statistical tests and models are suitable for your research questions?

Depends on a lot of factors

- measuring level
  - nominal, ordinal, ...

Statistical tests & models

Which statistical tests and models are suitable for your research questions?

Depends on a lot of factors

- measuring level
  - nominal, ordinal, ...
- number of variables
  - types: independent (predictor), dependent (outcome)

Statistical tests & models

Which statistical tests and models are suitable for your research questions?

Depends on a lot of factors

- measuring level
  - nominal, ordinal, ...
- number of variables
  - types: independent (predictor), dependent (outcome)
- type of relationship between variables
  - e.g., difference between ...

Statistical tests & models

Which statistical tests and models are suitable for your research questions?

Depends on a lot of factors

- measuring level
  - nominal, ordinal, ...
- number of variables
  - types: independent (predictor), dependent (outcome)
- type of relationship between variables
  - e.g., difference between ...
- study design, ...

Statistical tests & models

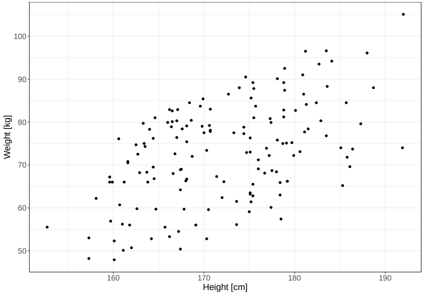
Which statistical tests and models are suitable for your research questions?

$\Rightarrow$ not easy to give an answer

Simple Linear Regression

Example - Height & Weight

What is the relationship between height and weight, respectively can height explain weight?

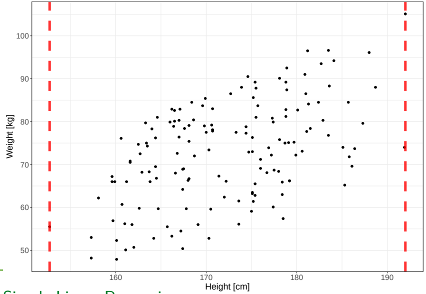


Regression analysis

- Regression analysis is used to describe the nature of a relationship using a mathematical equation

Regression analysis

- Regression analysis is used to describe the nature of a relationship using a mathematical equation
- Possibility of prognosis/prediction for an individual patient (incl. CI) within the value range of the predictor



Regression analysis

- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)

Regression analysis

- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis

Regression analysis

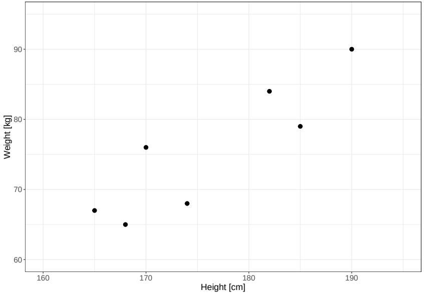
- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis
- Aim of the regression analysis
  - prediction, inference of $x \rightarrow y$

Regression analysis

- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis
- Aim of the regression analysis
  - prediction, inference of x $\rightarrow$ y
- method
  - e.g. minimize deviation squares of the observed values from the regression line

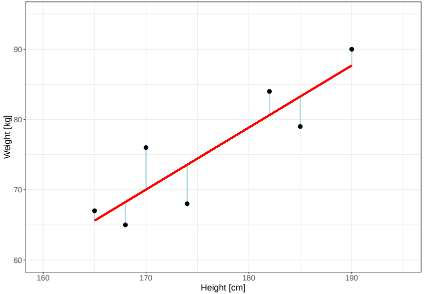
Simple linear regression

Find a straight line



Simple linear regression

- Problem: Find a straight line so that the vertical distance (residuals) between the data points and the straight line is minimized.
- Method, e.g., least squares method



Simple linear regression

As a statistical model

$Y = \beta_0 + \beta_1 * X$

Simple linear regression

As a statistical model

$Y = \beta_0 + \beta_1 * X$

As an empirical model with data

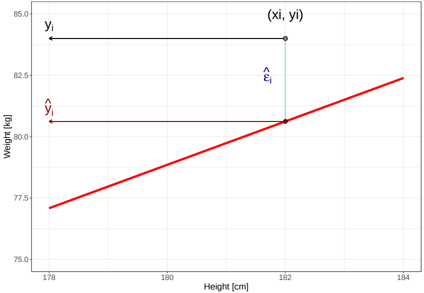
$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$

where ϵ_i describes the error (residual)

Simple linear regression

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$ are the predicted values of the regression

$\hat{\epsilon}_i = \hat{y}_i - y_i$ are the residuals of the regression



Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.
3) **Normality** - For any fixed value of X, Y is normally distributed.

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.
3) **Normality** - For any fixed value of X, Y is normally distributed.
4) **Homoscedasticity** - The variance of residual is the same for any value of X.

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.
3) **Normality** - For any fixed value of X, Y is normally distributed.
4) **Homoscedasticity** - The variance of residual is the same for any value of X.

- For 1) study design question

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.
3) **Normality** - For any fixed value of X, Y is normally distributed.
4) **Homoscedasticity** - The variance of residual is the same for any value of X.

- For 1) study design question

- For 2) scatter plot

Simple linear regression - Assumptions

1) **Independence** - Observations are independent of each other.
2) **Linearity** - The relationship between X and the mean of Y is linear.
3) **Normality** - For any fixed value of X, Y is normally distributed.
4) **Homoscedasticity** - The variance of residual is the same for any value of X.

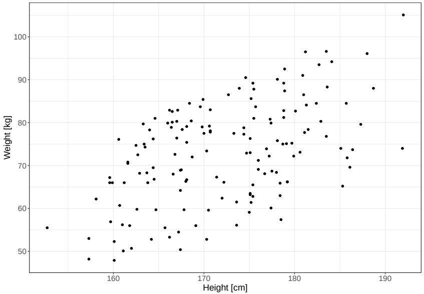
- For 1) study design question
- For 2) scatter plot
- For 3) & 4) looking at residuals

Coefficient of Determination R^2

R^2 specifies the proportion of variance in the data that is explained by the model

$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ and $0 \leq R^2 \leq 1$

Example - Height & Weight



1) **Independence** ✓
2) **Linearity** ✓

Example - Height & Weight (Residuals)

```r
res_model <- lm(weight ~ height, data = dt_regression)
```

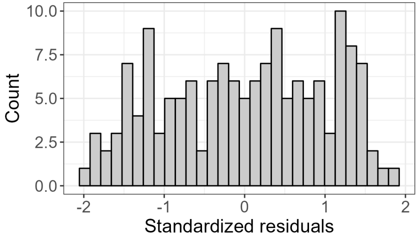
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Example - Height & Weight (Residuals)

```
res_model <- lm(weight ~ height, data = dt_regression)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

- Normality ✓?

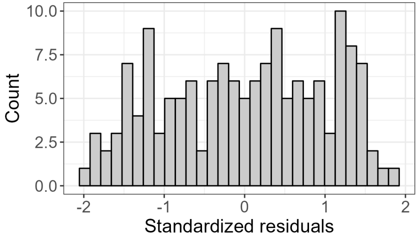


Example - Height & Weight (Residuals)

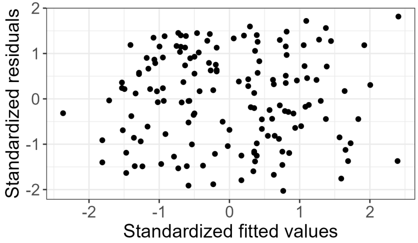
```r
res_model <- lm(weight ~ height, data = dt_regression)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

- Normality ✓?



- Homoscedasticity ✓



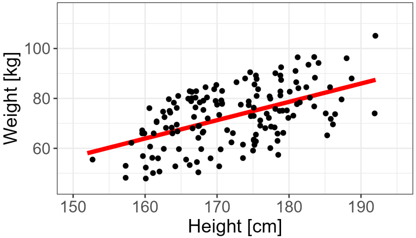
Example - Height & Weight (Residuals)

- If a model accurately captures the structure in the data, then all that should remain after the model is through making its predictions is random noise!

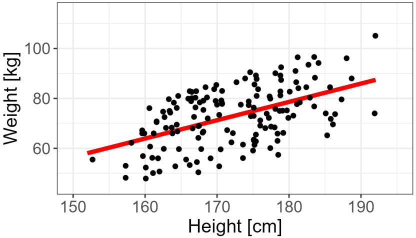
Example - Height & Weight (Residuals)

- If a model accurately captures the structure in the data, then all that should remain after the model is through making its predictions is random noise!
- Why plot residuals vs. fitted values, and not observations?
  - Because residuals and fitted values are uncorrelated by construction
  - Residuals and observations may be correlated—they both depend on observations — which would make such plots harder to interpret

Example - Height & Weight



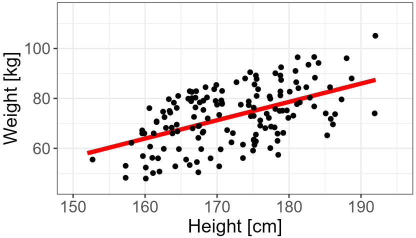
Example - Height & Weight



- intercept -53.49 (95% CI -87.3 to -19.69)
- slope 0.73 (95% CI 0.54 to 0.93)

- $R^2 = 0.27$
- $R^2_{adj} = 0.265$

Example - Height & Weight



- intercept -53.49 (95% CI -87.3 to -19.69)
- slope 0.73 (95% CI 0.54 to 0.93)

- $R^2 = 0.27$
- $R^2_{adj} = 0.265$

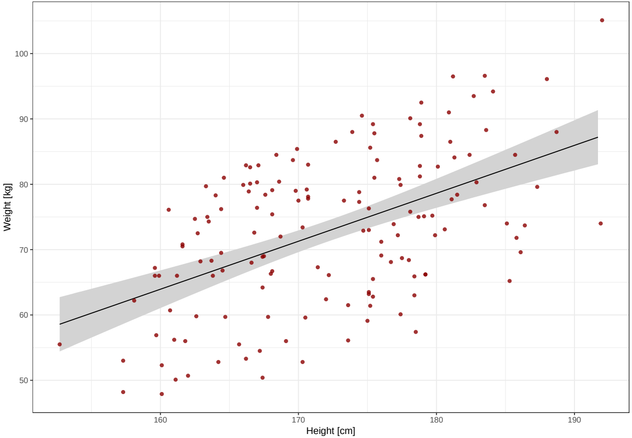
What weight can you expect from a 1.75 m tall person?

Example - Prediction

```r
predict(res_model, newdata = tibble(height = 175),
        interval = "confidence", level = 0.95)
```

```
##      fit   lwr   upr
## 1 74.95 73.25 76.65
```

Example - Uncertainty



Remarks (I)

- mathematical relationship $\neq$ causality

Remarks (I)

- mathematical relationship $\neq$ causality
- R^2 vs. R^2_{adj}

Remarks (I)

- mathematical relationship $\neq$ causality
- R^2 vs. R^2_{adj}
- R^2 tends to increase as more variables are added to the model (even if they don't improve the model significantly)
  - R^2_{adj} penalizes the addition of unnecessary variables:
    - $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$
    - n = number of samples
    - p = number of predictors

Remarks (I)

- mathematical relationship $\neq$ causality
- R^2 vs. R^2_{adj}
- R^2 tends to increase as more variables are added to the model (even if they don't improve the model significantly)
  - R^2_{adj} penalizes the addition of unnecessary variables:
    - $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$
    - n = number of samples
    - p = number of predictors
- R^2, R^2_{adj}
  - does not indicate whether the model was specified correctly
  - low/high coefficient of determination $\neq$ bad/good model

Remarks (II)

Assumptions not fulfilled - What then?

Remarks (II)

Assumptions not fulfilled - What then?

- Transform X (e.g. $Z = X^2$, $Y = \beta_0 + \beta_1 * Z$)
  - if linearity condition is violated

Remarks (II)

Assumptions not fulfilled - What then?

- Transform X (e.g. $Z = X^2$, $Y = \beta_0 + \beta_1 * Z$)
  - if linearity condition is violated

- Transform Y (e.g. log-transformation of Y)
  - in case of violation of variance homogeneity and/or normal distribution

Remarks (II)

Assumptions not fulfilled - What then?

- Transform X (e.g. $Z = X^2$, $Y = \beta_0 + \beta_1 * Z$)
  - if linearity condition is violated
- Transform Y (e.g. log-transformation of Y)
  - in case of violation of variance homogeneity and/or normal distribution
- Apply more complex or robust estimation methods
  - e.g. weighted least squares estimation, sandwich estimator, bootstrapping,...

Remarks (II)

Assumptions not fulfilled - What then?

- Transform X (e.g. $Z = X^2$, $Y = \beta_0 + \beta_1 * Z$)
  - if linearity condition is violated
- Transform Y (e.g. log-transformation of Y)
  - in case of violation of variance homogeneity and/or normal distribution
- Apply more complex or robust estimation methods
  - e.g. weighted least squares estimation, sandwich estimator, bootstrapping,...
- Multiple regression: further conditions must be checked (multicollinearity).

Data import

R packages for import

- **readr**
  - to read rectangular data (like csv, tsv, and fwf)
  - is a core package
  - https://readr.tidyverse.org/

R packages for import

- **readr**
  - to read rectangular data (like csv, tsv, and fwf)
  - is a core package
  - https://readr.tidyverse.org/
- **readxl**
  - to read data from Excel (xls, xlsx)
  - is not a core package
  - https://readxl.tidyverse.org/

R packages for import

- **readr**
  - to read rectangular data (like csv, tsv, and fwf)
  - is a core package
  - https://readr.tidyverse.org/
- **readxl**
  - to read data from Excel (xls, xlsx)
  - is not a core package
  - https://readxl.tidyverse.org/
- ... just search, e.g., import to R dat

Data import

- file type $\rightarrow$ R package and function

Data import

- file type $\rightarrow$ R package and function
- work with arguments in functions
  - e.g., first row contains column headers
  - e.g., type of parameters
  - e.g., which strings to interpret as missing values
  - ...

Data preparation with tidyverse

What is tibble?

*"A **tibble**, or **tbl_df** , is a modern reimagining of the data.frame, keeping what time has proven to be effective, and throwing out what is not. Tibbles are data."*

https://tibble.tidyverse.org/

What is tidyverse?

- tidyverse is a collection of R packages designed for data science

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
    - they share an underlying design philosophy, grammar, and data structure

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning
    - ...
  - all core packages can be loaded at once: *library(tidyverse)*

What is tidyverse?

- tidyverse is a collection of R packages designed for data science
  - they share an underlying design philosophy, grammar, and data structure
    - *ggplot2* for data visualization
    - *readr* for data importation from various file sources
    - *tidyr* and *dplyr* useful for data cleaning
    - ...
  - all core packages can be loaded at once: *library(tidyverse)*
  - 'R for Data Science' (see slide with links)

Useful functions for data preparation

- **select()** extracts columns and returns a tibble

Useful functions for data preparation

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows

Useful functions for data preparation

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows
- **filter()** picks cases based on their values

Useful functions for data preparation

- **select()** extracts columns and returns a tibble
- **arrange()** changes the ordering of the rows
- **filter()** picks cases based on their values
- **mutate()** adds new variables that are functions of existing variables

What is %>% in Tidyverse?

%>% is used to emphasis a sequence of actions, rather than the object that the actions are being performed on

What is %>% in Tidyverse?

%>% is used to emphasis a sequence of actions, rather than the object that the actions are being performed on

```
dt_example %>%
  mutate(bmi = weight/(height^2)) %>%
  select(pat_id, sex, bmi)
```

What will we cover

- We will look at
  - importing data (example: .xlsx)
  - useful function for data preparation
  - save R environment (.Rdata)
- Data import and preparation in R file (.R)
  - input: dataset (e.g., .xlsx)
  - output: .Rdata
- We will work with .Rdata in a Rmarkdown file

Example supraclavicular

This data set contains 103 patients who were scheduled to undergo an upper extremity procedure suitable for supraclavicular anesthesia. Patients were randomly assigned to either

(1) combined group-ropivacaine and mepivacaine mixture; or
(2) sequential group-mepivacaine followed by ropivacaine.

A number of demographic and post-op pain medication variables (fentanyl, alfentanil, midazolam) were collected. The primary outcome is time to 4-nerve sensory block onset.

[Source: R package medicaldata]

Links

Links (I)

- Introduction to R
  - R for Data Science (https://r4ds.hadley.nz/)
- Plots using ggplot
  - Overview with further links to course material: https://ggplot2.tidyverse.org/
- Display tables using flextable
  - flextable bool https://ardata-fr.github.io/flextable-book/
  - Function references https://davidgohel.github.io/flextable/reference/index.html
- knit_child()
  - link (https://bookdown.org/yihui/rmarkdown-cookbook/child-document.html)

Links (II)

- Download R
  - CRAN (https://cran.r-project.org/)
- Download RStudio
  - RStudio Desktop (https://posit.co/download/rstudio-desktop/)