Advanced R

Unit 1

Sereina Herzog

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

05.03.2025

Course Content - Advanced R (Unit 1)

- Short repetition
  - Reproducibility - Rmarkdown for reports
  - Project structure
  - Visualization with ggplot

Repetition

What is reproducibility in science?

- Ability to reproduce results by a peer
- Requires data, methods, and procedures
- Increasingly, science is supposed to be reproducible

Be nice to your future selves!

Reproducibility with RStudio & R

- R with RMarkdown can be used to produce different types of documents [see: http://rmarkdown.rstudio.com/gallery.html]
  - standardised reports (`html`, `pdf`)
  - word documents (`.docx`)
  - slides for presentations (`html`, `pdf`, `powerpoint`)
  - journal articles. using the `rticles` package (`.pdf`)
  - ...

Reproducibility with RStudio & R

- R with RMarkdown can be used to produce different types of documents [see: http://rmarkdown.rstudio.com/gallery.html]
  - standardised reports (`html`, `pdf`)
  - word documents (`.docx`)
  - slides for presentations (`html`, `pdf`, `powerpoint`)
  - journal articles. using the `rticles` package (`.pdf`)
  - ...

⇒ **making transparent and reproducible analysis**

Folder structure

Suggestion how to structure your project folder

- project1
  - literature
  - reports
  - ...
  - R

Folder structure

Suggestion how to structure your project folder

- project1
  - literature
  - reports
  - ...
  - R
    - orig
    - Rdata
    - Rfiles
    - Rmarkdown
    - Routput

Folder structure

Suggestion how to structure your project folder

- project1
  - literature
  - reports
  - ...
  - R
    - orig
    - Rdata
    - Rfiles
    - Rmarkdown
    - Routput

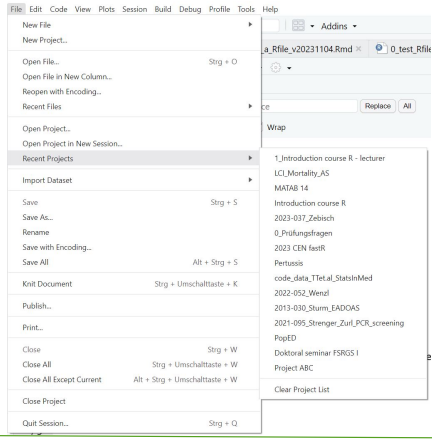
Hint: never touch the original data!

R project

- An R project
  - is a way to organize files and folders related to a specific analysis or project
    - easy to switch different projects
    - the working directory is the project's root folder

R project

- An R project
  - is a way to organize files and folders related to a specific analysis or project
    - easy to switch different projects
    - the working directory is the project's root folder



Create folder structure & R project

1) Download prepared folder structure
   - download 'projectstructure_for_students.zip' from GitHub
   - unzip the file
   - put folder 'Course Advanced R' wherever you want to have it
2) Generate a 'R project' (together)
   - File → New Project... → Existing Directory

What is *ggplot*?

- powerful data visualization package in R
  - wide range of high-quality plots and graphics
  - provides a consistent syntax
  - a layered approach to building plots

What is *ggplot*?

- powerful data visualization package in R
  - wide range of high-quality plots and graphics
  - provides a consistent syntax
  - a layered approach to building plots
- consists of three main components:

What is *ggplot*?

- powerful data visualization package in R
  - wide range of high-quality plots and graphics
  - provides a consistent syntax
  - a layered approach to building plots
- consists of three main components:
  - **data**
    - represents the dataset being visualized

What is *ggplot*?

- powerful data visualization package in R
  - wide range of high-quality plots and graphics
  - provides a consistent syntax
  - a layered approach to building plots
- consists of three main components:
  - **data**
    - represents the dataset being visualized
  - **aesthetics** (aes)
    - define how variables are mapped to visual properties (e.g., x-axis, y-axis, color)

What is *ggplot*?

- powerful data visualization package in R
  - wide range of high-quality plots and graphics
  - provides a consistent syntax
  - a layered approach to building plots
- consists of three main components:
  - **data**
    - represents the dataset being visualized
  - **aesthetics** (aes)
    - define how variables are mapped to visual properties (e.g., x-axis, y-axis, color)
  - **geometric objects** (geom)
    - determine the type of plot (e.g., points, lines, bars)

Example - Iris

A famous iris data set gives the measurements in centimeters of the variables

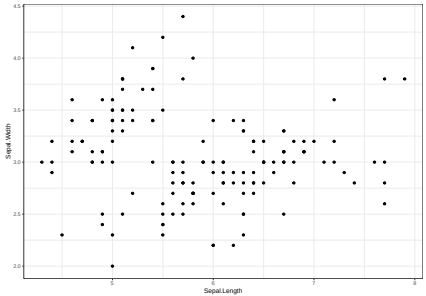
- sepal length
- sepal width
- petal length
- petal width

for 50 flowers from each of 3 species of iris (*Iris setosa*, *versicolor*, and *virginica*).

Iris Versicolor **Iris Setosa** **Iris Virginica**

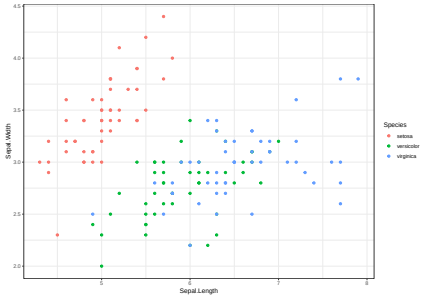
Example - Iris

```
ggplot(data = iris,
       aes(x = Sepal.Length, y = Sepal.Width)) +
    geom_point() +
    theme_bw()
```



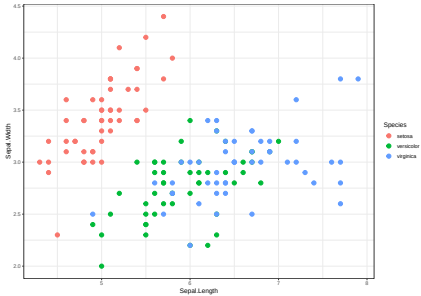
Example - Iris: including species as colour

```
ggplot(data = iris,
       aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
    geom_point() +
    theme_bw()
```



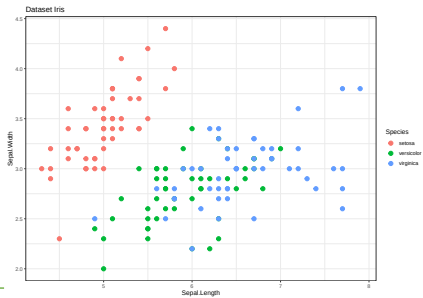
Example - Iris: increase point size

```
ggplot(data = iris,
       aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
    geom_point(size = 3) +
    theme_bw()
```

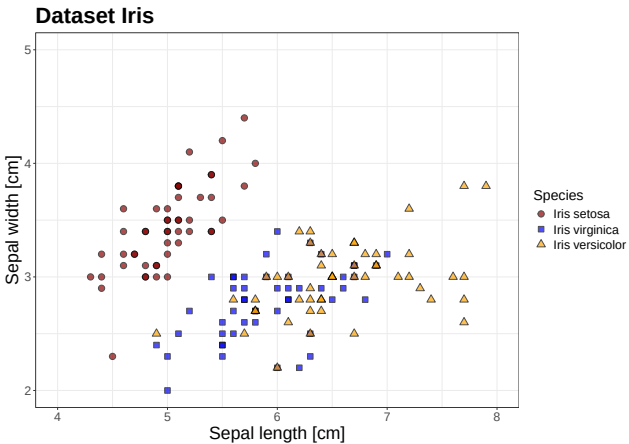


Example - Iris: adding title

```
ggplot(data = iris,
       aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
    geom_point(size = 3) +
    labs(title = "Dataset Iris") +
    theme_bw()
```

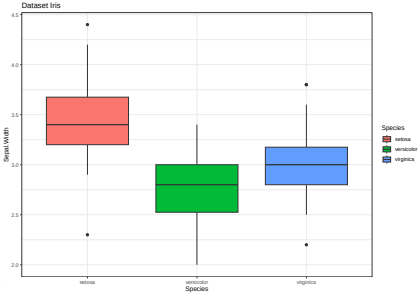


Example - Iris



Example - Iris: using another geom

```
ggplot(data = iris,
       aes(x = Species, y = Sepal.Width, fill = Species)) +
    geom_boxplot() +
    labs(title = "Dataset Iris") +
    theme_bw()
```



Saving ggplots

```
plot_iris <-
  ggplot(data = iris,
         aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
      geom_point() +
      theme_bw()

ggsave(filename = "../Routputs/example_iris.png", plot = plot_iris,
       units = "cm", width = 12, height = 7)
```

Exercise repetition

- Work through ‘Unit 1 - Exercise 1’

Splitting Rmd files

Why split Rmd files?

- If R markdown document is too long
  - split it into shorter documents, and include them as child documents

Why split Rmd files?

- If R markdown document is too long
  - split it into shorter documents, and include them as child documents
- If you want to use the same R markdown document again
  - include this R markdown document as a child document

What is knit_child()?

It is a function which knits a child document and returns a character string to input the result into the main document.

How to use knit_child()?

- It is designed to be used in the chunk option *child*
  - link (https://bookdown.org/yihui/rmarkdown-cookbook/child-document.html)

How to use knit_child()?

- It is designed to be used in the chunk option *child*
  - link (https://bookdown.org/yihui/rmarkdown-cookbook/child-document.html)
- It can be used in combination with chunk option *results* and function *cat()*

knit_child() with cat()

- within header of chunk: *{r, results = 'asis'}*

```
cat(knit_child("0_subRmds/subRmd_example.Rmd"), sep = '\n')
```

knit_child() with cat()

- within header of chunk: *{r, results = 'asis'}*

```
cat(knit_child("0_subRmds/subRmd_example.Rmd"), sep = '\n')
```

- Suggestions
  - in folder Rmarkdown have folder '0_subRmds'
  - start filenames of these subroutine Rmds with *subRmd_**XXX**_v**YYYYMMDD**.Rmd*

knit_child() with cat()

- within header of chunk: *{r, results = 'asis'}*

```
cat(knit_child("0_subRmds/subRmd_example.Rmd"), sep = '\n')
```

- Suggestions
  - in folder Rmarkdown have folder '0_subRmds'
  - start filenames of these subroutine Rmds with *subRmd_**XXX**_v**YYYYMMDD**.Rmd*
- Important
  - if you reuse a 'subRmd' several times → use no chunk name

Example - *UNIT2_ex0_exampleA.Rmd*

```r
dt_analysis <- iris
var_int <- "Sepal.Length"
```

```r
cat(knit_child("0_subRmds/subRmd_exampleA.Rmd"), sep = '\n')
```

- within header of chunk with 'knit_child': *{r, results = 'asis'}*

Example - *subRmd_exampleA.Rmd*

```
ggplot(data = dt_analysis,
       aes(x = get(var_int))) +
  geom_histogram(fill = "grey80", color = "black") +
  theme_bw()
```

- important - subRmd has no header
  - open new Rmd file
    - File/new File/R Markdown ...
    - delete the suggested content

Example - *UNIT2_ex0_exampleB.Rmd*

```
for(i in 1:lenght(parameters)){
  var_int <- parameters[i]
  cat(knit_child("0_subRmds/subRmd_exampleB.Rmd"), sep = '\n')
}
```

Example - *subRmd_exampleB.Rmd*

````
## Parameter `r i`: `r var_int`

```{r}
l <- sum(!is.na(dt %>%  pull(var_int)))
```

* has class: `r class(dt %>%  pull(var_int))`
* has `r l ` valid observations

```{r}
rm(l)
```
````

Let's run exampleA and exampleB

- We work through
  - example A
  - example B

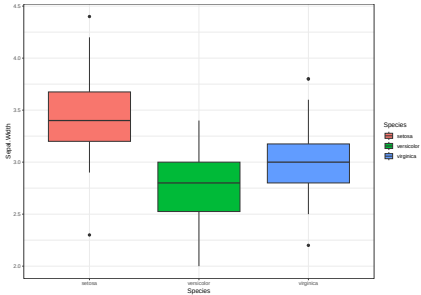
Exercise subRmd

- Work through ‘Unit 2 - Exercise 1’

Placeholders

Example - Iris

```
ggplot(data = iris,
       aes(x = Species, y = Sepal.Width, fill = Species)) +
    geom_boxplot() +
    theme_bw()
```



Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = Species, y = Sepal.Width, fill = Species)) +
    geom_boxplot() +
    theme_bw()
```

Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = Species, y = Sepal.Width, fill = Species)) +
    geom_boxplot() +
    theme_bw()
```

```
var_int <- "Sepal.Width"
group_int <- "Species"
```

Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = Species, y = Sepal.Width, fill = Species)) +
    geom_boxplot() +
    theme_bw()
```

```
var_int <- "Sepal.Width"
group_int <- "Species"
```

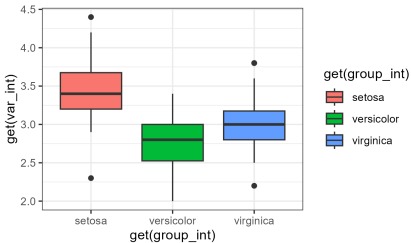
```
ggplot(data = iris,
       aes(x = get(group_int), y = get(var_int), fill = get(group_int))) +
    geom_boxplot() +
    theme_bw()
```

Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = get(group_int), y = get(var_int), fill = get(group_int))) +
  geom_boxplot() +
  theme_bw()
```

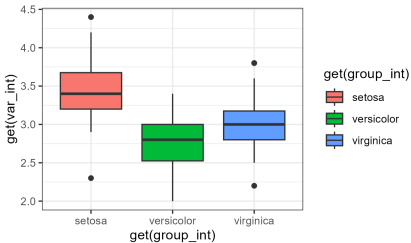
Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = get(group_int), y = get(var_int), fill = get(group_int))) +
  geom_boxplot() +
  theme_bw()
```



Working with variables as placeholders

```
ggplot(data = iris,
       aes(x = get(group_int), y = get(var_int), fill = get(group_int))) +
  geom_boxplot() +
  theme_bw()
```



Problem: axis labels and legend title → need to adapt them too

Working with variables as placeholders

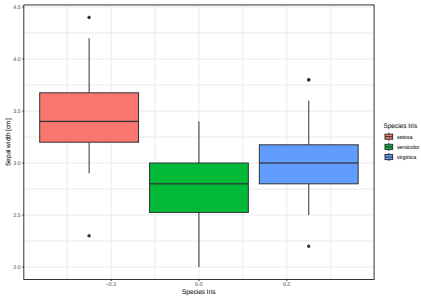
```
var_int <- "Sepal.Width"
var_int_lab <- "Sepal width [cm]"

group_int <- "Species"
group_int_lab <- "Species Iris"
```

Working with variables as placeholders

```
var_int <- "Sepal.Width"
var_int_lab <- "Sepal width [cm]"

group_int <- "Species"
group_int_lab <- "Species Iris"
```

```
ggplot(data = iris,
       aes(x = get(group_int), y = get(var_int), fill = get(group_int))) +
  geom_boxplot() +

  guides(fill = guide_legend(group_int_lab)) +

  xlab(group_int_lab) +
  ylab(var_int_lab) +

  theme_bw()
```

Working with variables as placeholders



Working with variables as placeholders

Advantage - can reuse same code for plots and only need to change things at one place

Working with variables as placeholders

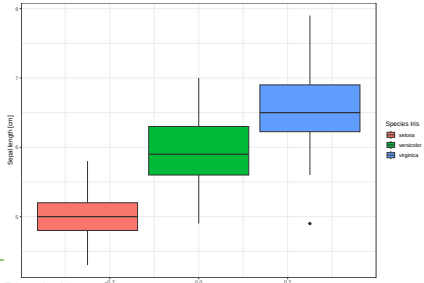
Advantage - can reuse same code for plots and only need to change things at one place

```
var_int <- "Sepal.Length"
var_int_lab <- "Sepal length [cm]"

group_int <- "Species"
group_int_lab <- "Species Iris"
```

Working with variables as placeholders

Advantage - can reuse same code for plots and only need to change things at one place

```
var_int <- "Sepal.Length"
var_int_lab <- "Sepal length [cm]"

group_int <- "Species"
group_int_lab <- "Species Iris"
```



Exercise placeholders

- Work through ‘Unit 1 - Exercise 2’

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

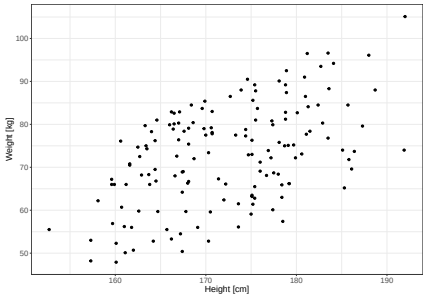
Which statistical tests and models are suitable for your research questions?

$\Rightarrow$ 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?

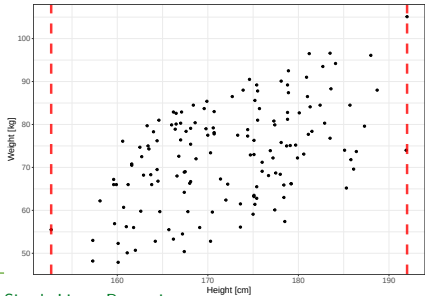


Regression analysis

- 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

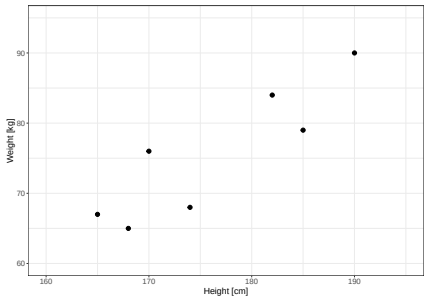
- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis
- Aim of the regression analysis
  - prediction, inference of $x \rightarrow y$

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

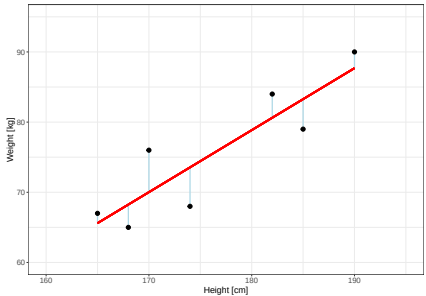
Simple linear regression

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

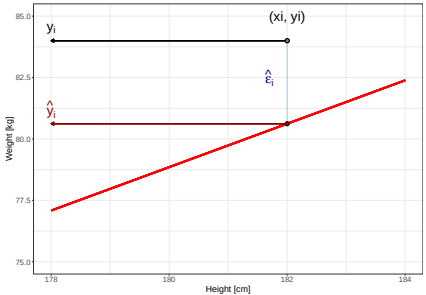
$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

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.

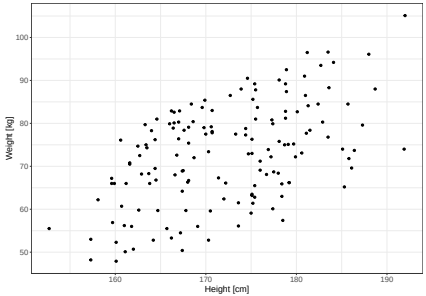
- For 1) study design question
- For 2) scatter plot
- For 3) & 4) looking at residuals

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** ✓

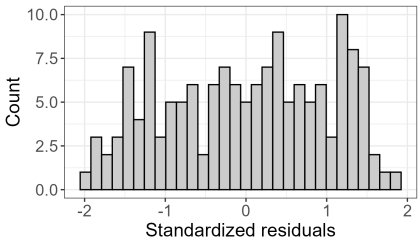
Example - Height & Weight (Residuals)

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

Example - Height & Weight (Residuals)

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

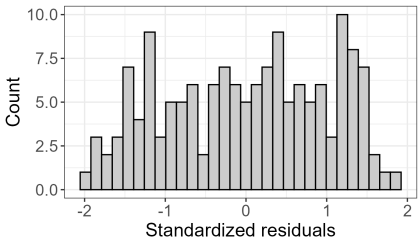
- Normality ✓?



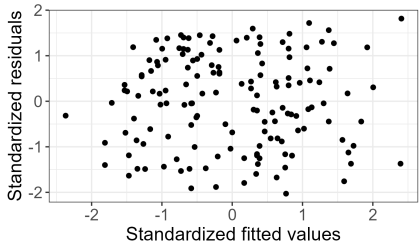
Example - Height & Weight (Residuals)

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

- Normality ✓?



- Homoscedasticity ✓



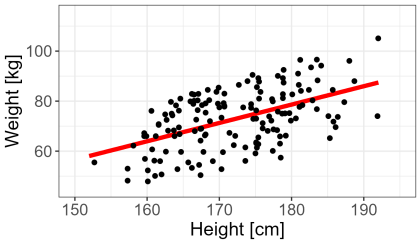
Example - Height & Weight (Residuals)

- 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!

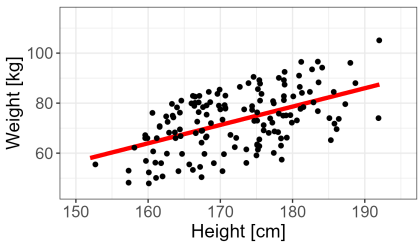
Example - Height & Weight (Residuals)

- 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



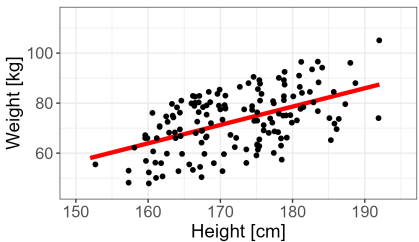
Example - Height & Weight



- 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)

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

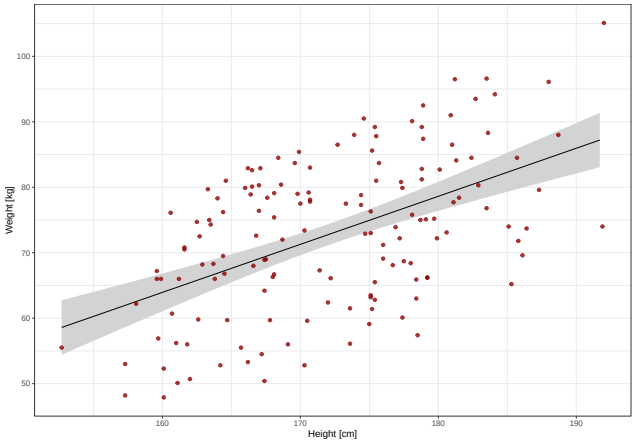
What weight can you expect from a 1.75 m tall person?

Example - Prediction

```
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}
  - 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).

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/)