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

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

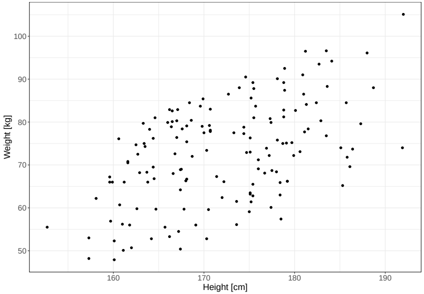
Which statistical tests and models are suitable for your research questions?

$\Rightarrow$ not easy to give an answer

Simple Linear Regression

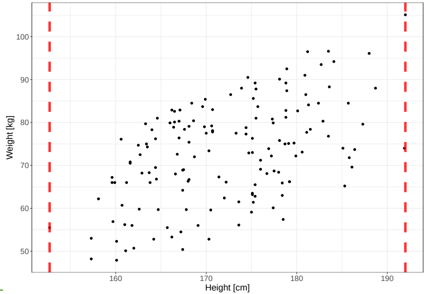
Example - Height & Weight

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
- 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

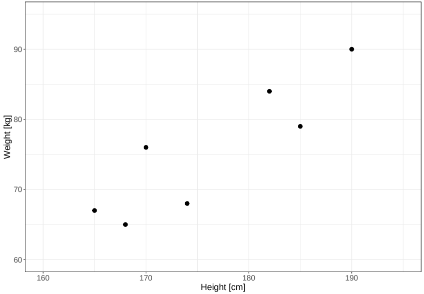
- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis
- Aim of the regression analysis
  - prediction, inference of $x \rightarrow y$

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 $x \to y$
- method
  - e.g. minimize deviation squares of the observed values from the regression line

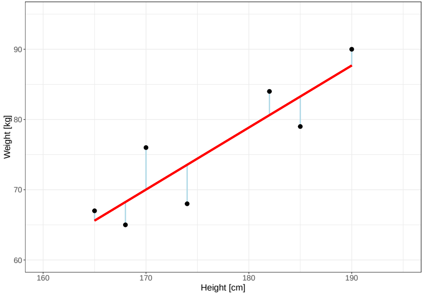
Simple Linear Regression

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

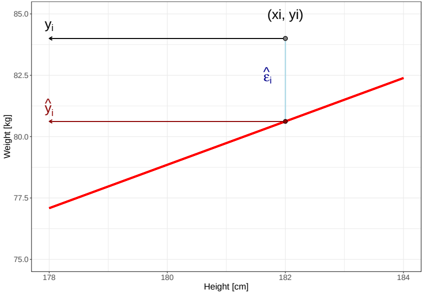
$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$

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.

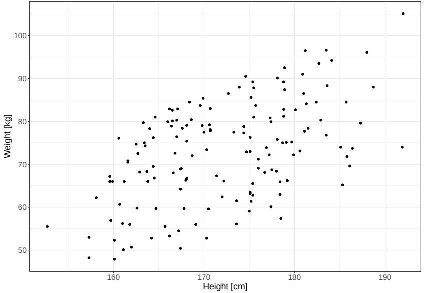
- For 1): study design question
- For 2) scatter plot
- For 3) & 4) looking at residuals

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** ✓

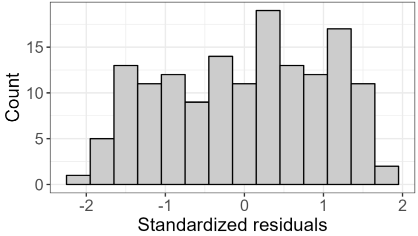
Example - Height & Weight

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

Example - Height & Weight

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

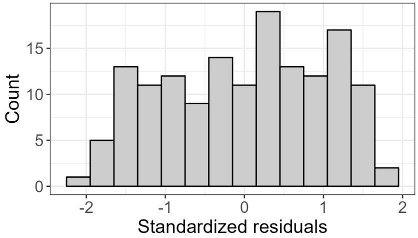
- Normality ✓



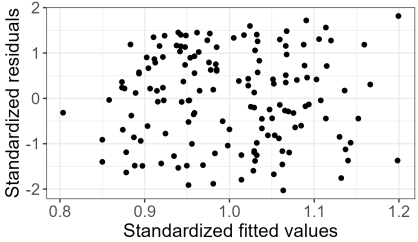
Example - Height & Weight

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

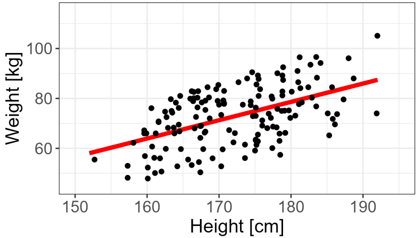
- Normality ✓



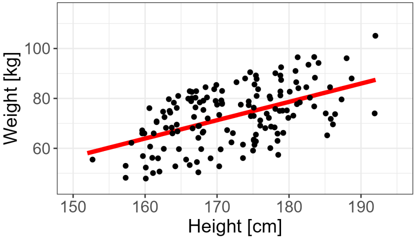
- Homoscedasticity ✓



Example - Height & Weight



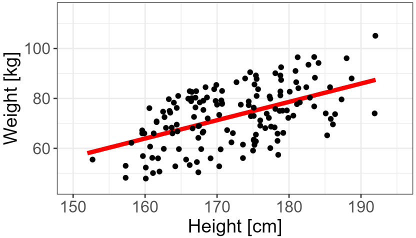
Example - Height & Weight



- 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?

Notes

- R^2 vs. adjusted R^2
  - R^2 tends to increase as more variables are added to the model (even if they don't improve the model significantly)
  - adjusted R^2 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