Advanced R

Unit 3

Sereina Herzog

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

06.03.2025

Course Content - Advanced R (Unit 3)

- Statistical tests & models
- Simple linear regression
- Statistical models in R
  - R packages
  - Simple linear regression in R

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

Statistical models in R

Statistical models in R (part I)

- For almost all well known statistical models there are R packages which will cover them

Statistical models in R (part I)

- For almost all well known statistical models there are R packages which will cover them
- Questions
  - How to find the correct R package?
  - What if there are several?

Statistical models in R (part II)

- Answers
  - search with the ‘correct’ key words online, ask colleagues, look which R packages are cited in papers
  - R CRAN vs GitHub: CRAN seems to be the more formal of the two
  - Keep in mind that R is a open source - there is no official check of the content, however,
    - big community
    - a lot of the packages on CRAN have accompanying peer-reviewed papers

Statistical models in R (part III)

- About a R package
  - Who did it?
  - Has it survived a couple of R updates?
  - What does the code look like?
  - How widely used is it? How ‘new’ is it?
- Good sources if problems occur
  - read the documentation of functions in the help
  - online: StackOverflow - a question-and-answer website

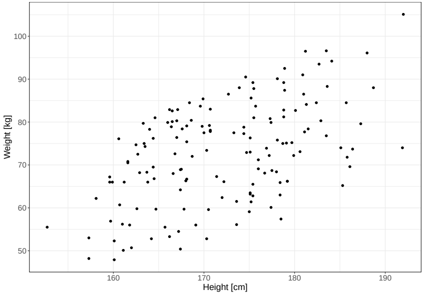
Statistical models in R (part III)

- R packages have often for modelling results
  - plot function
  - summary function

Simple Linear Regression

Example - Height & Weight

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

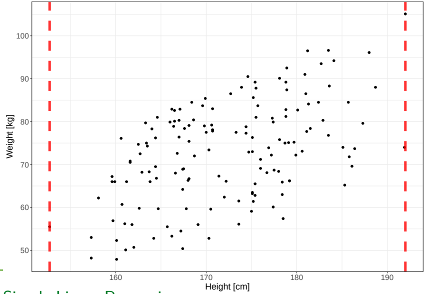


Regression analysis

- 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

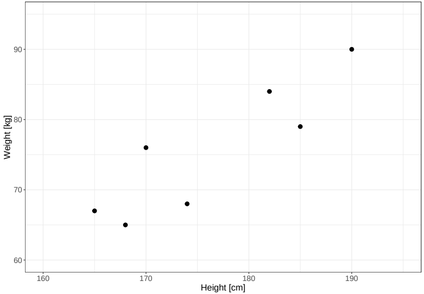
- Dependent variable
  - target variable, response, outcome
  - this variable is to be calculated from the other variable (y-axis)
- Independent variable(s)
  - explanatory variable(s), predictor
  - x-axis
- Aim of the regression analysis
  - prediction, inference of $x \rightarrow y$

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

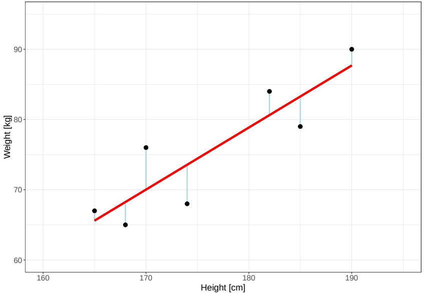
Simple linear regression

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

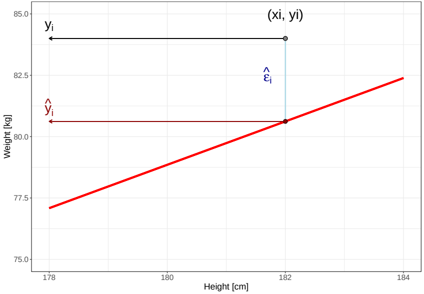
$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$

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.

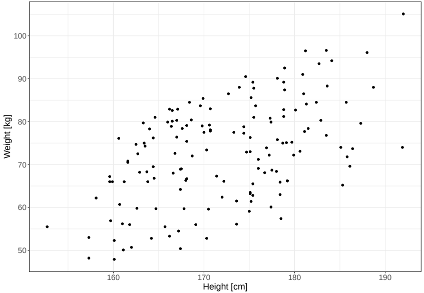
- For 1) study design question
- For 2) scatter plot
- For 3) & 4) looking at residuals

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** ✓

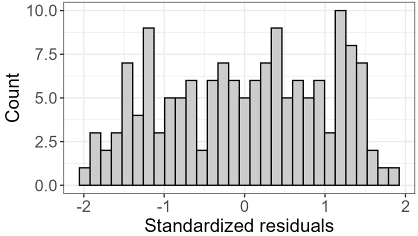
Example - Height & Weight (Residuals)

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

Example - Height & Weight (Residuals)

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

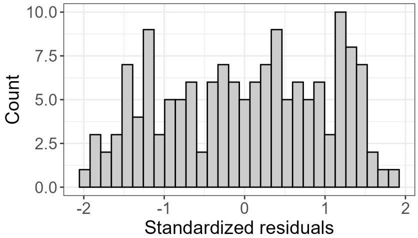
- Normality ✓?



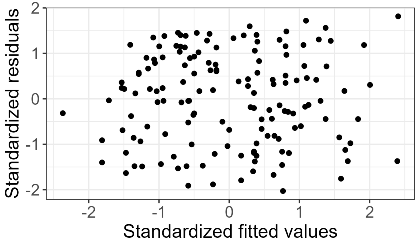
Example - Height & Weight (Residuals)

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

- Normality ✓?



- Homoscedasticity ✓



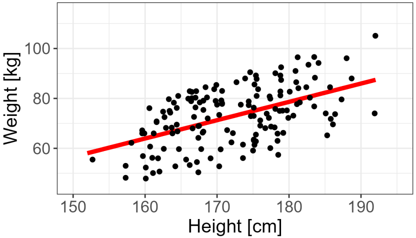
Example - Height & Weight (Residuals)

- 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!

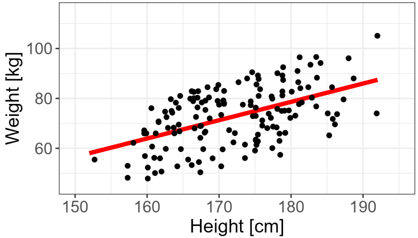
Example - Height & Weight (Residuals)

- 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



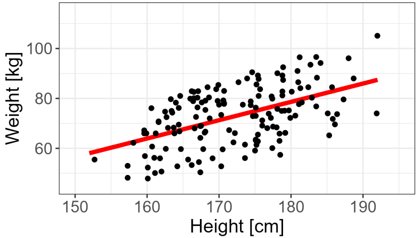
Example - Height & Weight



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

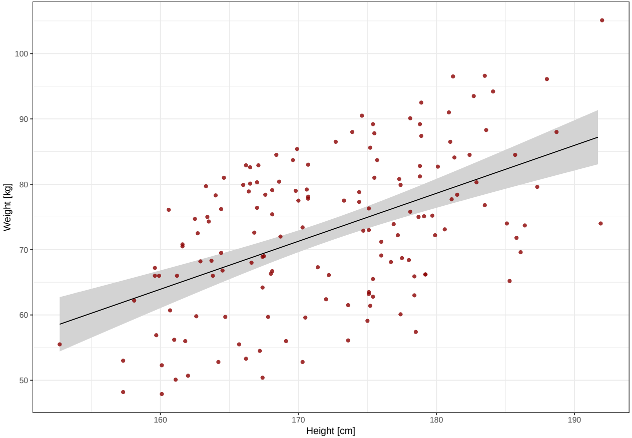
What weight can you expect from a 1.75 m tall person?

Example - Prediction

```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}
  - 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).

Simple linear regression - in R

Simple linear regression - in R (part I)

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

Simple linear regression - in R (part I)

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

- *lm*() from *stats* package
- 'Fitting Linear Models'
- **Description:** *lm* is used to fit linear models, including multivariate ones. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although *aov* may provide a more convenient interface for these).

Simple linear regression - in R (part II)

- **Usage:**

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

- **Value:** *lm* returns an object of class “*lm*” or for multivariate (‘multiple’) responses of class c(“*mlm*”, “*lm*”).

Simple linear regression - in R (part III)

```
res_model
```

```
##
## Call:
## lm(formula = weight ~ height, data = dt_regression)
##
## Coefficients:
## (Intercept)       height
##     -53.495        0.734
```

Simple linear regression - in R (part IV)

```
summary(res_model)
```

```
##
## Call:
## lm(formula = weight ~ height, data = dt_regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.118  -8.445   1.229   8.483  17.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -53.49479   17.10539  -3.127  0.00212 **
## height        0.73396    0.09922   7.397  9.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.967 on 148 degrees of freedom
## Multiple R-squared:  0.2699, Adjusted R-squared:  0.265
## F-statistic: 54.72 on 1 and 148 DF,  p-value: 9.602e-12
```

Simple linear regression - in R (part VI)

```r
coef(summary(res_model))
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -53.4947946 17.10538524 -3.127366 2.124137e-03
## height        0.7339631  0.09922269  7.397130 9.601925e-12
```

Exercise Simple linear regression - in R

- Work through 'Unit 3 - Exercise 1' (no pdf)
  - *UNIT3_ex1_linregression_vYYYYMMDD.Rmd*

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/)