Biostatistics & Epidemiological Data Analysis using R

8

Linear regression I

Stefan Konigorski

Health Intervention Analytics Group, HPI

December 16, 2021

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2021.10.28
	2	First steps in data analysis using R	2021.11.04
	3	Second steps in data analysis using R	2021.11.11
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2021.11.18
	5	Estimation	2021.11.25
	6	Hypothesis testing & study planning	2021.12.02
	7	Missing data	2021.12.09
Data analysis w/ regression models	8	Linear regression I	2021.12.16
	9	Linear regression II	2022.01.13
	10	Regression models for binary and count data	2022.01.20
	11	Analysis of variance & Linear mixed models I	2022.01.27
	12	Linear mixed models II & Meta analysis	2022.02.03
	13	Survival analysis	2022.02.10
	14	Causal inference & Data analysis challenge	2022.02.17

(see full schedule online)

Review: Parameter estimation and hypothesis testing

Non-formal description

- Aim: infer from your sample back onto the population
- For example:
  - give estimate for unknown parameter in the population
  - test hypothesis about unknown parameter in the population

Review: Parameter estimation

Statistical set-up

- Consider an unknown theoretical distribution F that contains some parameter(s) θ.
- Consider samples $x_1, \ldots, x_n$ (of rv's $X_1, \ldots, X_n \sim F$).

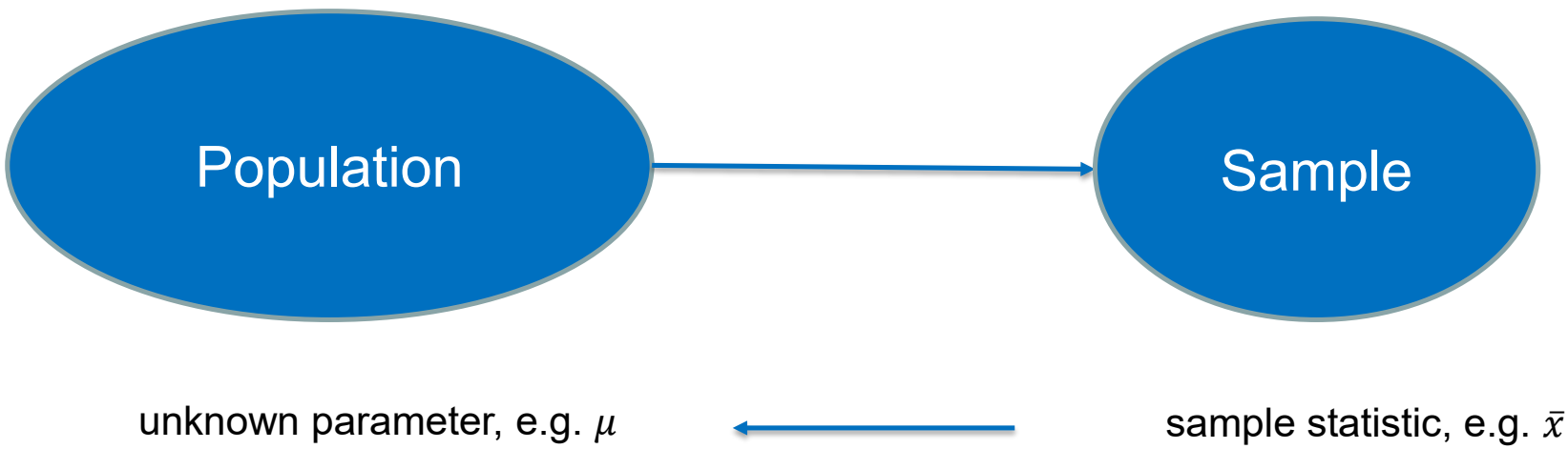
Review: Parameter estimation

Statistical set-up

- Consider an unknown theoretical distribution F that contains some parameter(s) θ.
- Consider samples $x_1, \ldots, x_n$ (of rv's $X_1, \ldots, X_n \sim F$).

Aim of point estimation

Based on $x_1, \ldots, x_n$, give a best guess $\hat{\theta}$ for the unknown θ.



Review: Hypothesis testing

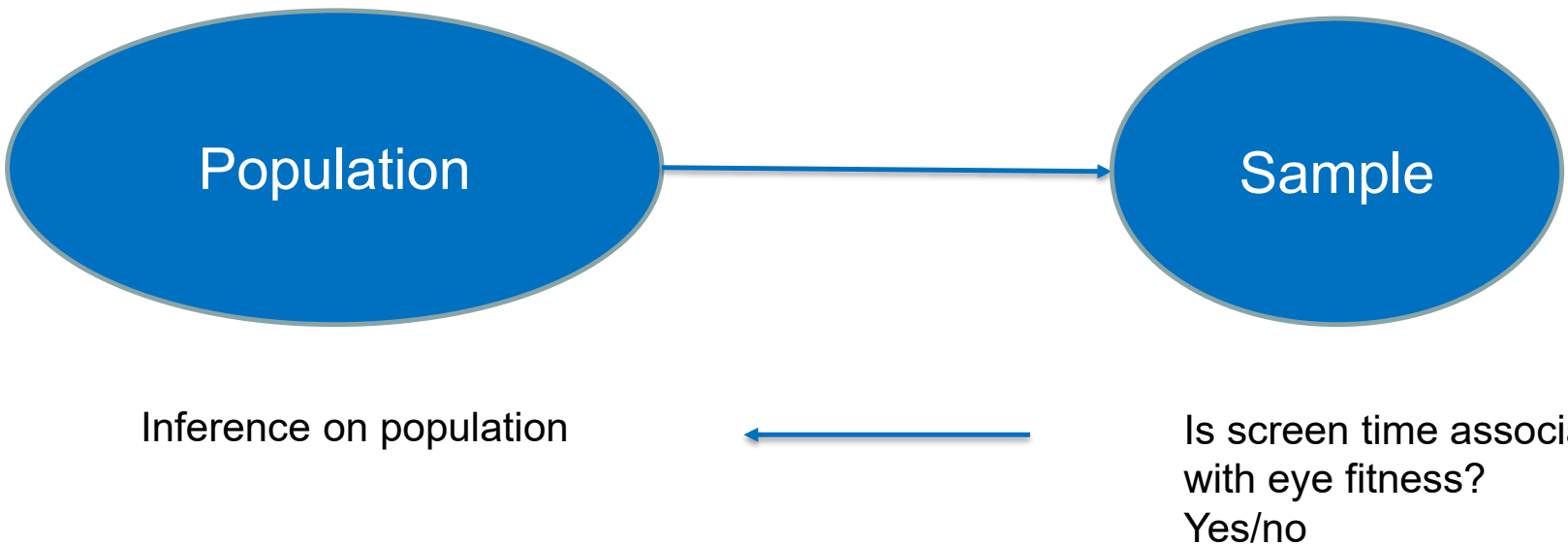
Goal

- The goal of a hypothesis test is to make a decision between a null hypothesis and a (complementary) alternative hypothesis.
- Example: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.

Review: Hypothesis testing

Goal

- The goal of a hypothesis test is to make a decision between a null hypothesis and a (complementary) alternative hypothesis.
- Example: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.



Review: Hypothesis testing

1. Assume: The null hypothesis is correct.

Review: Hypothesis testing

1. Assume: The null hypothesis is correct.
2. Calculate the probability (= p-value), that you obtain such (or more extreme) observations as you have in your sample, given that the null hypothesis is true.

Review: Hypothesis testing

1. Assume: The null hypothesis is correct.
2. Calculate the probability (= p-value), that you obtain such (or more extreme) observations as you have in your sample, given that the null hypothesis is true.
3. If this probability ...
   - is small (e.g. $< 5\% = \alpha$), then the empirical observations are hardly compatible with the assumption.
     $\longrightarrow$ Assumption must be wrong
     $\longrightarrow$ Reject null hypothesis
     $\longrightarrow$ Accept alternative hypothesis
   - is not small (e.g. larger than α), then there is not a strong evidence against the null hypothesis, therefore don't reject the null hypothesis.

Review: Hypothesis testing

- Approach: evaluate the evidence against the null hypothesis (=no association).
- Make your decision by (i) comparing the p-value to the pre-specified α level, or (ii) comparing the empirically calculated value of the test statistic to the theoretical "critical" value of the test statistic.[1]

[1] (i) and (ii) are identical.

Learning objectives of today

- Introduction to linear regression.
- Next lecture (but already here in the slides): Assumptions of linear regression.
- Next lecture: Get to know ways how to evaluate and select linear regression models, and how to do this in R, and do linear regression with multiple imputation.

Leading questions

- What is linear regression and how can you apply it in R?
- Next lecture (but already here in the slides): Which assumptions are made in linear regression, how can you check them, how bad is it if they are not satisfied, and what can you do if they are not satisfied?
- Next lecture: If I have 10 variables X_j to predict Y, how do I choose the relevant variables?
- Next lecture: How can you tell how good a linear regression model is?

Correlation

Covariance and correlation

Correlation

- Correlation = measure of association between two ordinal or metric variables X and Y.
- Has values between -1 (perfect negative association) and $+1$ (perfect positive correlation). A correlation of 0 means there is no association between X and Y.
- There exist multiple correlation coefficients, most popular are Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and Kendall's τ.

Covariance and correlation

Pearson's correlation coefficient

- ... is well-suited for normally-distributed (i.e. metric) variables.
- ... is based on the covariance *cov* between X and Y. This is defined, for a sample of n observations of X and Y, as:

$$cov(X, Y) = s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

- Correlation r between X and Y: $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$
- Hypothesis test: $H_0 : r = 0$ vs. $H_1 : r \neq 0$

Covariance and correlation

Pearson's correlation coefficient

- ... is well-suited for normally-distributed (i.e. metric) variables.
- ... is based on the covariance *cov* between X and Y. This is defined, for a sample of n observations of X and Y, as:

$$cov(X, Y) = s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

- Correlation r between X and Y: $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$
- Hypothesis test: $H_0 : r = 0$ vs. $H_1 : r \neq 0$

Spearman's correlation and Kendall's τ

- ... are based on ranks, i.e. suited for ordinal and non-normally distributed variables.

Correlation

In R

- Use the `cor` and `cor.test` functions.
- Also always look at the association in scatter plots, with the `plot` function.
- You can use `cor` and `plot` also for more than 2 variables!

Exercise 1

- In the KiGGS dataset, choose 3 or more different variables and compute pairwise different correlation coefficients.
- Look at scatter plots of the variables.
- Also try to use the `cor` and `plot` for all 3 variables at once.
- Look at the results (correlation coefficients, confidence intervals, hypothesis tests), compare and interpret them.
- See `R_8a_exercise_1_correlation.Rmd`

Simple linear regression

Overview

Aim

- Aim: Predict a variable Y by a variable X under the assumption of a linear relationship.

Overview

Aim

- Aim: Predict a variable Y by a variable X under the assumption of a linear relationship.

- Y = dependent variable, outcome, target variable, response
- X = independent variable, predictor, feature, covariate

Model equation

Set-up

n observations of two variables X and Y.

Model equation

Set-up

n observations of two variables X and Y.

Model equation

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, \ldots, n$, where

- β_0 is the intercept
- β_1 is the slope
- and more generally, β_0, β_1 are the regression coefficients (regression weights)
- ϵ_i are the residuals $=$ error terms
- and with the assumption that $\varepsilon_i \sim N(0, \sigma^2)$.

Model equations

- Model equation: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Alternative model equation 1: $Y = \beta_0 + \beta_1 x + \varepsilon$
- Alternative model equation 2: $E(Y) = \beta_0 + \beta_1 x$

Model equations

- Model equation: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Alternative model equation 1: $Y = \beta_0 + \beta_1 x + \varepsilon$
- Alternative model equation 2: $E(Y) = \beta_0 + \beta_1 x$

Interpretations of the model equation and of the regression coefficients?

Estimating regression coefficients

- Least squares estimate of the regression coefficients β_0 and β_1 (minimizing $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\text{cov}(X,Y)}{\text{Var}(X)}$$

- Interpretation?

Predicting the outcome

- Once you have obtained estimates of β_0 and β_1, you can plug them into the model equation to obtain estimates of Y based on the model equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

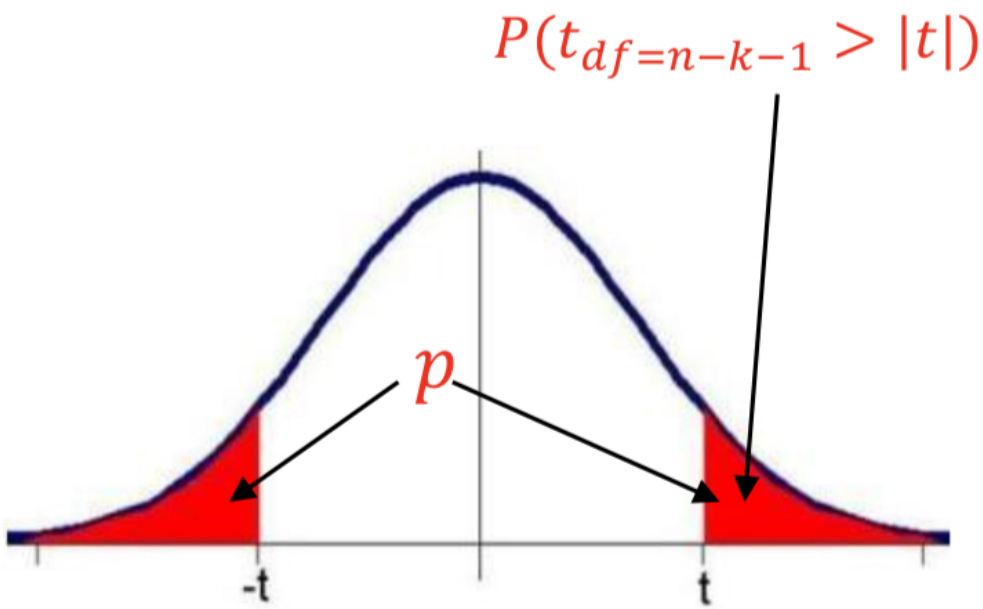
- This can be used, e.g., to predict the Y values of people that were not observed in this sample.

Testing regression coefficients

- Hypothesis tests of regression coefficients
  - Test of $H_0 : \beta_1 = 0$ vs. $H_0 : \beta_1 \neq 0$
  - Test statistic: $t = \widehat{\beta}_1 / \widehat{SE}(\widehat{\beta}_1)$, $\widehat{SE}(\widehat{\beta}_1) = \sqrt{\hat{\sigma}^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - has a t-distribution with $n - k - 1$ degrees of freedom, where n is the sample size and k is the number of independent variables in the regression (here: $k = 1$).

Testing regression coefficients

- Hypothesis tests of regression coefficients
  - Test of $H_0 : \beta_1 = 0$ vs. $H_0 : \beta_1 \neq 0$
  - Test statistic: $t = \widehat{\beta}_1 / \widehat{SE}(\widehat{\beta}_1)$, $\widehat{SE}(\widehat{\beta}_1) = \sqrt{\hat{\sigma}^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - has a t-distribution with $n - k - 1$ degrees of freedom, where n is the sample size and k is the number of independent variables in the regression (here: $k = 1$).



Confidence intervals of regression coefficients

- $(1-\alpha)\%$ confidence interval of regression coefficient β_1:

$$\beta_1 \pm t_{df=n-k-1,\alpha/2} \cdot SE(\beta_1)$$

Overall model fit

	Sum of squares	Degrees of freedom	Mean sum of squares	F-value	p-value
Regression	SS_{Regr}	k	$MS_{Regr} = SS_{Regr}/k$	$F = MS_{Regr}/MS_{Resid}$	$p = P(F_{df_1=k, df_2=n-k-1} > F)$
Residuals	SS_{Resid}	$n-k-1$	$MS_{Resid} = SS_{Resid}/(n-k-1)$		
Total	SS_{Total}	$n-1$			

where k is the number of predictors, and

$$SS_{Total} = SS_{Regr} + SS_{Resid},$$

$SS_{Total} = \sum_{i=1}^{n} (y_i - \overline{y})^2$, $SS_{Regr} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, $SS_{Resid} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Overall model fit

	Sum of squares	Degrees of freedom	Mean sum of squares	F-value	p-value
Regression	SS_{Regr}	k	$MS_{Regr} = SS_{Regr}/k$	$F = MS_{Regr}/MS_{Resid}$	$p = P(F_{df_1=k, df_2=n-k-1} > F)$
Residuals	SS_{Resid}	$n-k-1$	$MS_{Resid} = SS_{Resid}/(n-k-1)$		
Total	SS_{Total}	$n-1$			

where k is the number of predictors, and

$$SS_{Total} = SS_{Regr} + SS_{Resid},$$

$SS_{Total} = \sum_{i=1}^{n} (y_i - \overline{y})^2$, $SS_{Regr} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, $SS_{Resid} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Determination coefficient: $R^2 = \frac{SS_{Regr}}{SS_{Total}}$

Interpretation?

Adjusted (corrected) R^2: $R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$

Linear regression in R

- Use the `lm` function to fit linear regression models:
  `fit <- lm(Y ~ X1, data = dat)`
- Use the `summary` function to obtain a reader friendly output:
  `summary(fit)`
- Use the `predict` function to obtain predictions of Y:
  `predict(fit)`
- Use the `confint()` or `jtools::summ()` function to obtain confidence interval estimates of the regression coefficients.

How do you incorporate predictors in `lm`?

- ... as numeric or factor variables?
- And what are dummy variables?
- See exercise 2!

Exercise 2

In the KiGGS dataset:

1. Compute a linear regression by predicting systolic blood pressure (variable sys12) by BMI (variable bmiB) of children, and interpret the results
2. Compute a linear regression by predicting systolic blood pressure (variable sys12) by BMI categories (variable bmiKH) of children, and interpret the results
3. See `R_8b_exercise_2_linear_regression.Rmd`.

Multiple linear regression

Overview of multiple linear regression

- Aim: Predict a variable Y by multiple variables $X_1, \ldots, X_k$ under the assumption of a linear relationship.

Overview of multiple linear regression

- Aim: Predict a variable Y by multiple variables $X_1, \ldots, X_k$ under the assumption of a linear relationship.
- Model equation:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- In matrix formulation: $Y = X^T \beta + \varepsilon$
- Least squares estimate of the regression coefficient vector β:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Standard error of $\widehat{\beta}$:

$$SE(\widehat{\beta}) = \sqrt{\sigma^2 (X^T X)^{-1}}$$

Interpretation of the regression coefficients

- The regression coefficient β_j describes the effect of X_j on Y (increase of Y by β_j when X_j is increased by 1 unit), while holding all other X variables in the model constant.
- I.e. the effect of the other X variables in the model has been removed (=adjusted, corrected) from the regression coefficient β_j.
- Calculation in model with 2 independent variables X_1, X_2:

$$\hat{\beta}_1 = \frac{r_{X_1 Y} - r_{X_1 X_2} r_{X_2 Y}}{1 - r^2_{X_1 X_2}}$$

(cf. partial correlation coefficient)

- What can you conclude if all independent variables are uncorrelated?

Hypothesis tests

- Hypothesis tests and confidence intervals of the regression coefficients, and tests of the full model ($H_0 : \beta_1 = \cdots = \beta_k = 0$ vs. H_0 : at least β_j is $\neq 0$) are analogous to simple linear regression.

Exercise 3

In the KiGGS dataset:

1. Use linear regression to investigate the question if the amount and frequency of eating pancakes is associated with the BMI of children.
2. Adjust the analysis for age, sex, and further covariates, and interpret the new results.
3. See `R_8c_exercise_3_linear_regression.Rmd`.

Linear regression - Model assumptions

Overview

- So far we have neglected the assumptions that are made in linear regression
- What are they, how can we check them, and how bad is it if they are violated?

Assumptions of linear regression

- Y is continuous
- The relationsships between Y and all X_j are linear
- All relevant variables (covariates, confounders) are in the model
- All observations are independent
- There is no multicollinearity ($\approx$ not a "super strong" correlation between the X_j)
- Homoscedasticity (equal variance) of the residuals
- Normal distribution of the residuals

Assumptions of linear regression

- Y is continuous
- The relationsships between Y and all X_j are linear
- All relevant variables (covariates, confounders) are in the model
- All observations are independent
- There is no multicollinearity ($\approx$ not a "super strong" correlation between the X_j)
- Homoscedasticity (equal variance) of the residuals
- Normal distribution of the residuals

How do you check and deal with these assumptions?

Assumption: Y is continuous

Check

- yes/no (or approx. yes/no?)

Assumption: Y is continuous

Check

- yes/no (or approx. yes/no?)

What if assumption is not satisfied?

- Other regression (multinomial, ordinal?)

Assumption: Linearity

Check

- Visually in scatter plot
- in R: `plot(X, Y)`

Assumption: Linearity

Check

- Visually in scatter plot
- in R: `plot(X, Y)`

What if assumption is not satisfied?

- Add quadratic/polynomial terms or splines
- Adding a quadratic term in R: `lm(Y ~ X + X^2)`
- Transform predictor to factor?

Assumption: All relevant variables are in the model

Check

- Think, draw directed acyclic graphs, look at literature
- Compare the estimates of the regression coefficients in different models

Assumption: All relevant variables are in the model

Check

- Think, draw directed acyclic graphs, look at literature
- Compare the estimates of the regression coefficients in different models

What if assumption is not satisfied?

- Results can be strongly biased, in any direction
- Throw results into garbage and use other statistical model, or consider in interpretation

Assumption: Independent observations

Check

- Theoretical: Is there a structure in the data (time, hierarchy/cluster)?
- Compute ICC (intraclass correlation coefficient), Durbin-Watson statistic of autocorrelation

Assumption: Independent observations

Check

- Theoretical: Is there a structure in the data (time, hierarchy/cluster)?
- Compute ICC (intraclass correlation coefficient), Durbin-Watson statistic of autocorrelation

What if assumption is not satisfied?

- Use other statistical model (linear mixed models, time series)

Assumption: No multicollinearity

Definition

- 2 or more variables are collinear, if one variable can be written as a linear combination of the other variables.
- Multicollinearity here: $\approx$ no "super strong" correlation between the predictors X_j

Assumption: No multicollinearity

Definition

- 2 or more variables are collinear, if one variable can be written as a linear combination of the other variables.
- Multicollinearity here: $\approx$ no "super strong" correlation between the predictors X_j

Check

- Compute the correlation (and maybe VIF, variance inflation factor) between predictors
- in R: `cor(Xmatrix)`
- Check the predictive power of the model (R^2) eg using cross-validation

Assumption: No multicollinearity

What if assumption is not satisfied?

- If predictors are highly correlated (eg $r = 0.99$), then the estimates of the regression coefficients are still unbiased, and the standard error estimates of the regression coefficient estimates as well as the respective hypothesis tests are still valid. But: the standard error of the regression coefficient estimates is larger ($\longrightarrow$ smaller power) and R^2 is larger (SS_{Resid} and MSE are not affected, $Var(Y)$ and SS_{Regr} larger).
- But if X does not affect Y and instead both have a joint distribution, then also the regression coefficients can be affected and be biased (standard error estimates are ok).
- See `R_8d_assumpt_multicoll.R`
- Consider in interpretation (with respect to validity of estimates/tests, effect of X_1 or X_2?), remove effect of one variable, factor analysis/principal component analysis.

Assumption: Homoscedasticity of residuals

Check

- Graphically: scatter plot of residuals vs. predicted Y values, e.g. with `plot(lm(...))` function
- Extract residuals and predicted Y values in R: `residuals(lm(...))`, `predict(lm(...))`

Assumption: Homoscedasticity of residuals

Check

- Graphically: scatter plot of residuals vs. predicted Y values, e.g. with `plot(lm(...))` function
- Extract residuals and predicted Y values in R: `residuals(lm(...))`, `predict(lm(...))`

What if assumption is not satisfied?

- With heteroscedastic residuals, the estimates of the regression coefficients are still unbiased, but the standard errors are underestimated, i.e. the p-values are too small (inflation of type I error)
- See `R_8d_assumpt_homosced.R`
- Solutions: Consider in regression (eg weighted least-squares regression), use robust standard error estimates

Assumption: Normal distribution of residuals

Check

- Graphically: histograms and Q-Q-plots of the residuals
- For example, using the `plot(lm(...))` function

Assumption: Normal distribution of residuals

Check

- Graphically: histograms and Q-Q-plots of the residuals
- For example, using the `plot(lm(...))` function

What if assumption is not satisfied?

- Approximatively normally-distributed residuals are sufficient so that estimates, standard error estimates, and hypothesis tests of the regression coefficients are still valid - but the power decreases.
- Only for extremely non-normal residuals, estimates are biased.
- See `R_8d_assumpt_normality.R`
- Solutions: transform Y (log, Box-Cox), in R with `log()` and eg `MASS::boxcox()`

Further regression diagnostics

- The `plot(lm(...))` function yields further diagnostic plots, e.g. of outliers and leverage points.

Exercise 4

- Examine the assumptions in the linear regression of exercise 3.
- See `R_8d_exercise_4_assumpt.Rmd`.

Questions?

References

- Knight K (1999). Mathematical statistics. CRC Press.
- Harrell (2015). Regression modeling strategies. Springer.

Homework

Homework

No homework today. In the tutorial, the exercises in `R_8_more_exercises.Rmd` will be discussed.