Fachbereich AIHI
Universität Salzburg
Univ.-Prof. Dr. Wolfgang Trutschnig

Salzburg, April 2024
SS 23/24

Exercise Sheet 06 @'Applied AI Using R'

Main objective of the current sheet is again to combine the tools learned so far with some first regression parts. As mentioned last time, exercises like these do prepare you for participating in projects and for internships in companies.

Aufgabe 24.
Consider the following data set:

University	Mid-Career Salary	Yearly Tuition
Princeton	\$137 000	\$28 540
Harvey Mudd	\$135 000	\$40 133
CalTech	\$127 000	\$39 900
US Naval Academy	\$122 000	\$0
West Point	\$120 000	\$0
MIT	\$118 888	\$42 050
Babson College	\$117 000	\$40 400
Stanford	\$114 000	\$54 506

Calculate the linear regression coefficients by hand in order to test if higher tuition fees (on average) translate into higher-paying jobs.

Aufgabe 25.
Download the life expectancy dataset from the world bank (`https://data.worldbank.org/indicator/SP.DYN.LE00.IN`) and parse it into `R`:

1. We are interested in the life expectancy in the US. Use the `{dplyr}` verbs to arrange the downloaded data into the following form:

```
# A tibble: 5 x 3
  country_name   year life_expectancy
  <chr>         <dbl>           <dbl>
1 United States  1960            69.8
2 United States  1961            70.3
3 United States  1962            70.1
4 United States  1963            69.9
5 United States  1964            70.2
```

2. Produce a nice ggplot to visualize the timeseries.

3. Decide on which variable is the outcome and which is the predictor. Fit a linear model, either using `{tidymodels}` without a training/test-split and using the `extract_fit_engine()` function or work directly with the `lm()` function.

4. Use the regression line to estimate (extrapolate) life expectancy for a person born in 1950.

5. Repeat the previous steps with the date for Austria.

Aufgabe 26.
Download the "120 years of Olympic history" dataset from Kaggle (`https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results`) or here (`https://ufile.io/p460qizm`) and parse it into `R`.

1. Plot the number of participating athletes per year for the Olympic summer games. Can you explain the drop in 1956?

2. Consider the years 1936, 1976, 1996, 2016. For each of those years and each country (i.e. NOC) calculate the number of male and female participants.

3. For each of those years train a linear model to predict the number of female participants from the number of male participants.

4. How do the model coefficients change? How can this be interpreted?

Aufgabe 27.
Does `lm` really detect the *true* coefficients?

1. Pick your two favorite numbers (or sample them randomly), call them a and b, and choose some $\sigma^2 > 0$ (which will be the variance of the random errors in the model). We consider the regression function $f(x) := a + bx$.

2. Write a function that takes an integer n and returns a dataset with two columns `x` and `y` and n rows, whereby the entries in `x` are drawn from the uniform distribution on $[-5, 5]$ (use `runif(n,-5,5)`) and `y` is given by `y(x) = f(x) + eps`, where `eps` is normally distributed with variance σ^2 and mean 0 (use `rnorm(n,0,sigma2)`).

3. Write a function that takes an integer n, generates 100 of those datasets, trains a linear model of each of them and returns a dataframe with two columns `a_hat` and `b_hat` and 100 rows, where `a_hat` is the estimated intercept and `b_hat` is the estimated slope for each of the 100 models.

4. Run this function for different sample sizes n and different error variance σ^2, produce boxplots summarizing the obtained estimates and their deviation from the true coefficients. What can be observed - does the precision increase with increasing sample size n? Which effect does increasing the σ^2 have on the precision of the estimates?