Tutorial 7 - Linear Regression

AUTHOR
Victoria Hünewaldt

Tutorial 7

This tutorial will cover linear regression. You will learn:

- how to familiarise with a new data set
- how to derive regression coefficient estimates by hand
- how to run a linear regression model
- how to interpret it
- how to evaluate its performance

Exercises

Create a .qmd-file and solve the tasks there. Store it in the JupyterHub folder "Session 7".

1 Data Descriptives

Please use the airbnbsmall data set. Use our package "sozoekds" to load the data set. Make sure to have installed the package first.

Make yourself familiar with the dataset: How many observations does it comprise? How many variables?

2 Linear Regression

1 Simple linear regression by hand

Think about a simple linear regression model with one exogenous variable explaining "price" (= endogenous variable). Write down the equation of your model.

Calculate the intercept and regression coefficient estimate by hand. Remember the formula from the lecture (p.11).

Then, run the regression using a suitable R command. Compare your results to the ones you calculated by hand and interpret the coefficient and intercept.

2 Multiple linear regression

Create a multiple linear regression model explaining "price" (= endogenous variable). Choose the exogenous variables that you think explain "price" most. Write down the equation of your model. After

having decided upon your model, analyse the association between your exogenous variable(s) and endogenous variable by running the regression. Interpret your results.

3 Performance evaluation

Now evaluate the results of model 1 and model 2 using the R^2 metric. What does the R^2 value of your model tell you? How would you evaluate your models given their R^2 values?

Use model 2 and calculate the in-sample root mean squared error (RMSE). It provides an estimation of how well the model is able to predict the target value. You can calculate in three steps:

1. Calculate the predicted values of your model.
2. Calculate the error, i.e. the difference between predicted and observed values of the target variable.
3. Take the root mean square of the error.