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?

```
rm(list=ls()) # clean environment

# install our package sozoekds

# install.packages("remotes") # hashtagged because already installed
library(remotes)
remotes::install_gitlab("BAQ6370/sozoekds", host="gitlab.rrz.uni-hamburg.de")
library(sozoekds)

# load other libraries
library(tidyverse)
library(stargazer)
library(Hmisc)
library(dplyr)

# load the data & store it as "airbnb_data"
airbnb_data <- airbnbsmall

# get an overview of the data
```

```
N=nrow(airbnb_data)
N # shows nr of obs: 20634

colnames(airbnb_data) # variables
str(airbnb_data) # variable types
summary(airbnb_data) # summary of all variables in the data set
summary(airbnb_data$price) # summary of only the price variable
describe(airbnb_data$price) # provides further descriptive information (e.g. nr of missings)
```

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.

```
{r}
# equation of model: price = b0 + b1*d_breakfast + e

attach(airbnb_data) # attach data set

# calculate by hand

b1 <- cov(d_breakfast, price)/var(d_breakfast) # the covariance of the two variables divided b
print(b1) # -24.14064
b0 <- mean(price)-b1*mean(d_breakfast)  # the mean endogenous variable - the estimated coeffic
print(b0) # 107.3325

# using lm

mod1 <- lm(price~d_breakfast,data=airbnb_data) # create the object "mod1" that contains the li
summary(mod1) # prints the regression output

# results are the same
```

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.

```
# equation of model: price = b0 + b1*d_breakfast + b2*n_accommodates + b3*n_number_of_reviews
```

```
mod2<-lm(price~d_breakfast+n_accommodates+n_number_of_reviews,data=airbnb_data) # create the o
summary(mod2) # prints the regression output
```

```
Call:
lm(formula = price ~ d_breakfast + n_accommodates + n_number_of_reviews,
    data = airbnb_data)

Residuals:
    Min      1Q  Median      3Q     Max
-278.65  -30.15   -9.13   20.41  777.48

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          18.50938    0.90212  20.518   <2e-16 ***
d_breakfast         -11.19510    1.24789  -8.971   <2e-16 ***
n_accommodates       25.70602    0.20870 123.171   <2e-16 ***
n_number_of_reviews  -0.11277    0.01371  -8.223   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.44 on 20630 degrees of freedom
Multiple R-squared:  0.4303,    Adjusted R-squared:  0.4303
F-statistic:  5195 on 3 and 20630 DF,  p-value: < 2.2e-16
```

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.

```
in_predictions <- predict(mod2, airbnb_data) # calculate predictions given your model and the 
in_error <- (in_predictions-airbnb_data$price) # calculate the error (predicted values - obser
sqrt(mean(in_error^2)) # calculate the RMSE
```

```
[1] 59.43094
```

```
#prints 59.43
```