Exercise Sheet 2 – Data Mining Wirtschaftsinformatik, HTW Berlin

Martin Spott

last revision 21 April 2020

The exercises are about multiple linear regression. The data in `miete03.asc` has kindly been provided by the Institute for Statistics at the University of Munich. Import the data set into R using

```
mietspiegel <- read.table("miete03.asc", header = TRUE)
```

The data description:

- `nm`: Net rent in EUR
- `nmqm`: Net rent per m^2 in EUR
- `wfl`: Floor space in m^2
- `rooms`: Number of rooms in household
- `bj`: Year of construction
- `bez`: Urban district
- `wohngut`: Good residential area? (Y=1,N=0)
- `wohnbest`: Very good residential area? (Y=1,N=0)
- `ww0`: Hot water supply? (Y=0,N=1)
- `zh0`: Central heating? (Y=0,N=1)
- `badkach0`: Tiled bathroom? (Y=0,N=1)
- `badextra`: Supplementary equipment in bathroom? (Y=1,N=0)
- `kueche`: Well equipped kitchen? (Y=1,N=0)

The aim of building a linear regression model is to estimate the rent of an accommodation (house/flat) based on data.

Exercise 2.1

a) Explore the data set using `summary()` and `View()`. Note that `View()` does not work when you compile (knit) a RMarkdown document, but only in RStudio.

b) Plot the data using `pairs()`, i.e. a matrix of pairwise scatter plots of the variables. Interpret the results.

c) Compute the correlation matrix and show the lower half only for better readability. What do the numbers tell us?

```
miet_cor <- cor(mietspiegel)
miet_cor[upper.tri(miet_cor, diag = FALSE)] <- NA
miet_cor
```

```
##                 nm        nmqm         wfl        rooms          bj
## nm      1.00000000          NA          NA           NA          NA
## nmqm    0.47479666  1.00000000          NA           NA          NA
## wfl     0.70746267 -0.22683036  1.00000000           NA          NA
## rooms   0.54424729 -0.27290570  0.84064544  1.000000000          NA
## bj      0.04705929  0.28647882 -0.19907406 -0.152752907  1.00000000
## bez      -0.06675952 -0.07442558 -0.05216295  0.029348888  0.31158077
## wohngut   0.16056842  0.15038159  0.09125829  0.002111619 -0.10945956
## wohnbest  0.14749539  0.11045750  0.06284242  0.027458352  0.06055297
## ww0      -0.15863217 -0.28221689  0.07085171  0.083504068 -0.21570813
## zh0      -0.19011481 -0.29815097  0.02259344  0.029148598 -0.32158749
## badkach0 -0.13248538 -0.17268288 -0.02548533  0.001134735 -0.10244503
## badextra  0.29406790  0.06455749  0.27684649  0.211483094  0.04244939
## kueche    0.23200635  0.18824004  0.08621106  0.048319591  0.14628417
##                  bez     wohngut    wohnbest         ww0         zh0
## nm                NA          NA          NA          NA          NA
## nmqm              NA          NA          NA          NA          NA
## wfl               NA          NA          NA          NA          NA
## rooms             NA          NA          NA          NA          NA
## bj                NA          NA          NA          NA          NA
## bez       1.00000000          NA          NA          NA          NA
## wohngut  -0.31056198  1.00000000          NA          NA          NA
## wohnbest  0.06191412 -0.11998506  1.00000000          NA          NA
## ww0      -0.06935555  0.03710350 -0.02853963  1.00000000          NA
## zh0      -0.12715745  0.01626871 -0.04569784  0.52022396  1.00000000
## badkach0 -0.03953819 -0.01190246 -0.02851745  0.07958463  0.14645748
## badextra  0.05787463  0.06285792  0.04367286 -0.03371308 -0.04972802
## kueche    0.07630049  0.05112502  0.06023556 -0.05352421 -0.07229766
##             badkach0  badextra kueche
## nm                NA        NA     NA
## nmqm              NA        NA     NA
## wfl               NA        NA     NA
## rooms             NA        NA     NA
## bj                NA        NA     NA
## bez               NA        NA     NA
## wohngut           NA        NA     NA
## wohnbest          NA        NA     NA
## ww0               NA        NA     NA
## zh0               NA        NA     NA
## badkach0  1.00000000        NA     NA
## badextra -0.03606555 1.0000000     NA
## kueche   -0.05187616 0.1098271      1
```

Be aware that `cor()` uses Pearson correlation as a default, which is only sensible for metric features. In particular, it does not make sense for nominal features like `bez` (district). It can be used on binary features like `wohngut` that have an order (Y=1 is better than N=0). We will discuss these issues in the lectures.

d) Build a multiple linear regression model `miet_lm` with `nm` (*Nettomiete*) as target (output) variable and all the others as predictors (input variables).

e) Explore the model using `summary()`. Try to make sense of

   i) the residuals
   ii) the value of the coefficients
   iii) the p-value $Pr(>|t|)$ of the coefficients
   iv) R-squared
   v) the F-statistic and its p-value

f) Why are some of the coefficients in the model negative? Does that make sense? Refer to the data description above for help.

Exercise 2.2

Think about the following scenario. You would like to buy a flat. You gather data about it through a viewing and information from the estate agent. You want to use the data to estimate the price of the flat. For that purpose you build a multiple linear regression model.

a) Revisit the predictor (input) variables in your model from Exercise 2.1. Should we use all of them? Do we know all of them in our scenario?
b) Rebuilt the linear model without the one predictor variable we do not know in our scenario. Compare the model with the one including all predictors.
c) Interpret the coefficient of the variable `bez` describing the district of Munich. Does this make sense? Check out the data description.
d) Use the function `factor()` to change the variable `bez`. What does this do? Re-build the model. Compare the model with the previous one and interpret the differences. Look at the values of the coefficients for `bez`.

Exercise 2.3

What does the following code fragment do? Explain the value of R-squared and make the link to Exercise 2.2a.

```
mietspiegel_lm <- lm(nm ~ nmqm:wfl, mietspiegel)
summary(mietspiegel_lm)
```

```
##
## Call:
## lm(formula = nm ~ nmqm:wfl, data = mietspiegel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79255 -0.15308  0.00053  0.15102  0.77635
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.502e-02  1.186e-02     1.266    0.206
## nmqm:wfl    1.000e+00  1.911e-05 52321.109   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2125 on 2051 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.737e+09 on 1 and 2051 DF,  p-value: < 2.2e-16
```