Regression

Master Data Science
Summer 2018

Prof. Dr. Marlene Müller
marlene.mueller@beuth-hochschule.de

Exercises 5

Exercise 1

Simulate artificial regression data using the follwing R code:

```
n <- 30
set.seed(19062017)      ## to always obtain the same numbers
eps <- rnorm(n, sd=0.5)
x1 <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n)
y <- 1 + x1 + eps
```

Estimate the following linear models:

```
lm1 <- lm( y ~ x1 )

lm2 <- lm( y ~ x1 + x2 )

lm3 <- lm( y ~ x1 + x3 )

lm4 <- lm( y ~ x1 + x2 + x3)
```

(a) Which of the coefficients in which of the models are significant at level 5%?

(b) Which of the models could be compared by F tests? Do these tests in R and comment the results.

(c) Draw a scatterplot of `x1` and `y` and plot the regression line corresponding to model `lm1`. Do also calculate the R^2 values for all models (how could you obtain the value from the model in R?).

If you look at the R code above that generated the data: Which of the parameters should be changed such that the R^2 values for all four models do increase? (How does this change the plot?)

Exercise 2

Consider a data set with an explanatory variable X and a dependent variable Y having the following values:

x_i	a	a	a	b	b	b	b	c	c	c
y_i	5	7	5	4	5	5	6	4	4	3

(a) Use R to fit a linear model to these data. Which possibilities do you have to code the variable X? Try to write the possible design matrices first on paper, then check with R.

(b) Could you calculate the estimated coefficients using a pocket calculator, i.e. without using R? What are the predicted values $\widehat{y}_i$?

Exercise 3

The cdf of the standard logistic distribution is given by $F(x) = \dfrac{1}{1 + e^{-x}}$.

(a) What are the properties of a cdf? Explain why F fulfills these.

(b) Calculate the pdf $f(x)$.

(c) Use the R function `rlogis` to generate pseudo-random numbers for the logistic distribution. (The standard logistic has `location=0` and `scale=1`.) Simulate samples from the standard logistic distribution and illustrate that its expectation is 0 and the variance equals $\dfrac{\pi^2}{3}$ ist.

Exercise 4

We consider again the standard logistic distribution (see previous exercise).

(a) Plot the curves for the pdf and the cdf.

(b) Add the corresponding curves for the Gaussian (standard normal) to your plots. How would you describe the differences between both distributions?

(c) Which parameters of the normal distribution could you choose in order to have a distribution that resembles the standard logistic? Do also compare the pdf and cdf curves.