Regression

Master Data Science
Summer 2023

Prof. Dr. Marlene Müller
marlene.mueller@bht-berlin.de

Exercises 4

Exercise 1

(a) Plot the pdf of the χ^2-distribution for different degrees of freedom. See `?dchisq`. How would you describe the effect of an increasing `df` parameter?

(b) Plot the pdf of the t-distribution for different degrees of freedom. See `?dt`. What happens to the distribution when `df` increases? (You may also ask Google to answer. ☺)

(c) Generate artificial data for different χ^2- and t-distributions. You should use sufficiently large sample sizes and calculate means and variances. What is your guess about the expectations for both χ^2- and t as well as for the variance of χ^2?

Exercise 2

We generate artificial regression data:

```
x <- runif(10)
y <- 2 - 2*x + 0.5*x^2 + rnorm(length(x), sd=0.2)
lm1 <- lm( y ~ x )
lm2 <- lm( y ~ x + I(x^2))
lm3 <- lm( y ~ x + I(x^2) + I(x^3) )
```

(a) Do a scatterplot of the data and graphically display the 3 estimated regression functions.

(b) The R function `model.matrix` allows to extract the design matrix ($\mathcal{X}$ matrix) from an estimated regression model. Use this to calculate the hat matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$. Verify with R that all 3 matrices are projection matrices (which properties have to be checked?) and that their traces equal $p + 1$.

(c) Do also verify with R that: $\mathbf{P}_2 \cdot \mathbf{P}_1 = \mathbf{P}_1$, $\mathbf{P}_3 \cdot \mathbf{P}_1 = \mathbf{P}_1$ and $\mathbf{P}_3 \cdot \mathbf{P}_2 = \mathbf{P}_2$

(Remark: For our models we have `lm1` $\subseteq$ `lm2` $\subseteq$ `lm3`. So, if we already projected into the space spanned by the column vectors of a smaller design matrix, then the projection on to a larger space does not change the result anymore.)

(d) Prove that from (c) follows: $\mathbf{P}_1 \cdot \mathbf{P}_2 = \mathbf{P}_1$, $\mathbf{P}_1 \cdot \mathbf{P}_3 = \mathbf{P}_1$ and $\mathbf{P}_2 \cdot \mathbf{P}_3 = \mathbf{P}_2$

Exercise 3

Load the dataset `AnscombeQuartet.csv` (see Moodle, source: Wikipedia). The dataset contains columns for 4 different regressions, i.e. to model `y1` in dependence of `x1` until `y4` in dependence of `x4`.

(a) Estimate the 4 simple linear regressions first. Compute and compare the R^2 and RSS values. What do you observe? (Any big differences?)

(b) Now, do a graphical exploration: Plot the data as point clouds and add the respective regression lines. Describe the differences.

Exercise 4

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

(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 $\dfrac{\pi^2}{3}$.

Exercise 5

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.