Regression

Master Data Science
Summer 2021

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

Exercises 4 (incl. hints to solve)

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

Use the CPS1985 data (`require(AER); data(CPS1985)`) again. This time we want to consider subsamples for males and females:

```
males <- CPS1985[CPS1985$gender=="male",]
females <- CPS1985[CPS1985$gender=="female",]
```

Estimate for both subsamples a multiple regression model for $\log(\text{wage})$ on years of education, years of professional experience and squared experience. Consider the output from the respective `summary` for each of the subsamples:

(a) How could you determine the sample sizes of the two subsamples?

(b) Which of the coefficients are sigficantly different from 0, if we assume a level of significance of 5%? (Does this change if we would use 1%?)

(c) How could you determine/calculate the values of RSS for both models? Would it be useful to compare them?

(d) Predict $\log(\text{wage})$ for both models for a person with 12 years of education and 10 year of professional experience. What do you observe? (Is there a difference between females and males?)

(e) Generate graphs for the marginal effects of `experience` for both models, i.e. display the estimated quadratic functions while setting `education` equal to 12 for example. (Note that 12 is the median of `education` in the full sample.)

Exercise 3

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: $\mathbf{P}_2 \cdot \mathbf{P}_1 = \mathbf{P}_1\,,\quad \mathbf{P}_3 \cdot \mathbf{P}_1 = \mathbf{P}_1$ 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: $\mathbf{P}_1 \cdot \mathbf{P}_2 = \mathbf{P}_1\,,\quad \mathbf{P}_1 \cdot \mathbf{P}_3 = \mathbf{P}_1$ and $\mathbf{P}_2 \cdot \mathbf{P}_3 = \mathbf{P}_2$

due to symmetry:

$$(\mathbf{P}_1 \cdot \mathbf{P}_2)^\top = \mathbf{P}_1^\top \iff \mathbf{P}_2^\top \cdot \mathbf{P}_1^\top = \mathbf{P}_1^\top \iff \mathbf{P}_2 \cdot \mathbf{P}_1 = \mathbf{P}_1$$

Exercise 4

Consider linear model for a dataset 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.
depending on the reference categories:

reference $= a$ reference $= b$ reference $= c$

$$\mathcal{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{or} \quad \mathcal{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{or} \quad \mathcal{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

(b) Remember how we interpreted the estimated coefficients. Could you calculate the estimated coefficients using a pocket calculator, i.e. without using R? What are the predicted values $\widehat{y}_i$?
again depending on the reference categories:

reference $= a$

$$\widehat{\beta}_0^{(a)} = \frac{5+7+5}{3}$$
$$\widehat{\beta}_1^{(a)} = \frac{4+5+5+6}{4} - \widehat{\beta}_0^{(a)}$$
$$\widehat{\beta}_2^{(a)} = \frac{4+4+3}{3} - \widehat{\beta}_0^{(a)}$$

reference $= b$

$$\widehat{\beta}_0^{(b)} = \frac{4+5+5+6}{4}$$
$$\widehat{\beta}_1^{(b)} = \frac{5+7+5}{3} - \widehat{\beta}_0^{(b)}$$
$$\widehat{\beta}_2^{(b)} = \frac{4+4+3}{3} - \widehat{\beta}_0^{(b)}$$

reference $= c$

$$\widehat{\beta}_0^{(c)} = \frac{4+4+3}{3}$$
$$\widehat{\beta}_1^{(c)} = \frac{5+7+5}{3} - \widehat{\beta}_0^{(c)}$$
$$\widehat{\beta}_2^{(c)} = \frac{4+5+5+6}{4} - \widehat{\beta}_0^{(c)}$$

Exercise 5

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.