Regression

Master Data Science
Summer 2021

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

Exercises 5 (incl. hints to solve)

Exercise 1

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.
monotonously increasing, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$, right continuous (both-side continuous in case of a continuous distribution

(b) Calculate the pdf $f(x)$.
$f(x) = F'(x) = \dfrac{e^{-x}}{(1+e^{-x})^2} = \dfrac{e^x}{(1+e^x)^2}$

(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 2

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.

Exercise 3

Load the dataset `Affairs` from the R package `AER` (`library(AER); data(Affairs)`). Check the data documentation: `?Affairs`.

To estimate a logit model, we need a dependent variable Y with only to values 1 and 0. A useful approach is to generate Y from the variable `affairs` (say $Y = 1$ if the number of of affairs is positive and $Y = 0$ otherwise). Do the following analyses:

(a) Explore graphically the effect of single explanatory variables on Y (use for example: `spineplot`, `barplot`, or `mosaicplot`).

(b) Fit at least three different logit models (some of them should be nested) and do interpret the estimated coefficients.

(c) Use one of your models to predict. The canonical link here is logit, so the inverse link for prediction is $F(u) = \dfrac{1}{1+e^{-u}}$. Derive the formula for doing it using a pocket calculator in an exam situation.

(d) Compare your estimated models. Instead of the F test for linear models we do now use χ^2 tests. The syntax is similar: `anova(glm1,glm2,test="Chisq")`

Additionally, also compare AIC values and apply `stepAIC`.

Exercise 4

Load again the dataset `Affairs` from the R package `AER`. We aim to estimate a poisson regression now, so that we use $Y =$ `affairs`. Do the following analyses:

(a) Explore graphically the effect of single explanatory variables on Y (cf. Exercise 3(a)).

(b) Fit at least three different poisson models (some of them should be nested) and do interpret the estimated coefficients. Use the canonical link function.

(c) Compare your estimated models. Again, here we use χ^2 tests (remember to use `anova` with the option `test="Chisq")`). Additionally, also compare AIC values and apply `stepAIC`.

(d) Use one of your models to predict. The canonical link here is $\log$, so the inverse link for prediction is $\exp$. Derive the formula for doing it using a pocket calculator in an exam situation.