Universität Basel
Faculty of Business and Economics
Econometrics and Statistics
Prof. Christian Kleiber

FS 2024
2024-03-27
Assignment 1

Microeconometrics and Statistical Learning
Homework Assignment 1

Rules: This assignment is due within 14 days, i.e., by 2023-04-10, 10:15am (i.e., *before class*). No late homework assignments will be accepted.

You may work in groups of two. Solutions in English or German will be accepted. Please note that **you are expected to show results *and* to comment on them. For this reason, do not hand in solely (R) codes.** For example, some people use – this is not mandatory! – **rmarkdown**, `Sweave()` or **knitr**, which is OK provided you include results and your interpretation.
Handwritten solutions for theoretical questions are fine, provided they are legible.

Problem 1: (maximum likelihood)
Consider a simple random sample of size n from a (discrete) distribution with probability mass function ('density')

$$f(y;\theta) = \theta\ (1-\theta)^{y-1}, \quad \theta \in [0,1],\ y \in \mathbb{N}, \tag{1}$$

which has $\mathsf{E}(y) = 1/\theta$.

(a) Give the likelihood and the log-likelihood.

(b) Obtain the score $s(\theta)$ and the Hessian $H(\theta)$.

(c) Obtain the MLE of θ. (Don't forget to check the second-order condition!)

(d) Obtain the observed information and the expected (i.e., Fisher) information. Are they identical?

(e) Obtain estimates of the variance of the MLE using the observed and the expected information. Are the estimates identical?

(f) The distribution used here has an alternative version, with $f(y;\theta) = \theta\ (1-\theta)^y$, that is supported on $\mathbb{N}_0$ (not on $\mathbb{N}$ as the one from above). For this alternative form the first moment is $(1-\theta)/\theta$. What is the MLE of this quantity, using the MLE of the previously studied version?

(g) Obtain a standard error for the estimate from the preceding question.

Remarks: You may use that the ML regularity conditions are satisfied here. The MLE is approximately normally distributed with the "usual" variance.

Problem 2: (binary response)
The data set `dv.rda` (in R binary format, alternatively `dv.csv` in .csv format) contains cross-sectional data originating from an Australian health survey.
We are interested in modelling whether an individual visited a doctor as a function of certain explanatory variables. The data set contains the following variables:

`visits`	Number of doctor visits in past 2 weeks.
`gender`	Factor indicating gender.
`illness`	Number of illnesses in past 2 weeks.
`reduced`	Number of days of reduced activity in past 2 weeks due to illness or injury.
`freepoor`	Factor. Does the individual have free government health insurance due to low income?

(a) Estimate a logit model with all explanatory variables using an indicator of doctor visits as the dependent variable. Note that you will first have to construct this binary indicator.

(b) Briefly interpret the coefficient on `gender`.

(c) Predict the probability that a male person without illnesses but 4 days of reduced activity in the past two weeks and free health insurance visits a doctor.

(d) Obtain McFadden's R^2 and the ROC curve. Comment on the fit.

(e) Find the cutoff c^* which provides the highest accuracy. Obtain the confusion matrices for both $c = 0.5$ and $c = c^*$. Proceed by computing the corresponding hit rates and briefly comment on your results.