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

FS 2024
2024-05-15
Assignment 2

Microeconometrics and Statistical Learning
Homework Assignment 2

Rules: This assignment is due within 14 days, i.e., by 2024-05-29, 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.
- Handwritten solutions for theoretical questions are fine, provided they are legible.
- It is recommended – though not mandatory – to work with `.Rmd`, `.Rnw` or `.qmd` files, where it is easy to incorporate code and text (that's what they are for). Use of `.docx` is possible, but discouraged. **In any case, we need your results *and* your interpretation. For this reason, do not hand in solely (R) codes.**
- Submissions: Files should be named `X+Y-ass2.postfix`, where X and Y are the names of the authors, and `postfix` is some of `pdf, Rnw, rmd, qmd, txt, R, ...`. Please upload your solutions to OLAT.

Problem 1: (geometric distribution, again)
Consider the geometric regression model

$$P(Y = y_i \mid x_i) = \frac{\mu_i^{y_i}}{(1+\mu_i)^{(y_i+1)}}, \quad \text{for } y_i = 0, 1, 2, \ldots, \tag{1}$$

with $\mathsf{E}(y_i \mid x_i) = \mu_i(x_i^\top \beta) =: \mu_i$.

(a) Show that the geometric regression model is a generalized linear model and identify all of its components.

(b) In the GLM framework, there is a connection (mentioned in passing in class) between moments and the 'b function' in the exponential family representation. Specifically, $\mathrm{Var}(y_i \mid x_i) = b''(\theta)$ in the absence of a dispersion parameter. Compute the variance of the geometric distribution using this approach. (You can check your result against the expression for the variance of the NB II distribution that is available from the lecture notes.)

Problem 2: (doctor visits, again)
Consider the same data set as in Assignment 1, i.e. `dv.rda`. In Assignment 1, the task was to model a binary factor indicating doctor visits. Now the task is to fit a count data model to the number of doctor visits.

(a) Fit a Poisson regression model. Briefly explore the fit. In which ways is the Poisson model satisfactory, in which ways not?

(b) Is there overdispersion in the data? Explain.

(c) Try to find a better model than the Poisson model. Defend your choice.

NB. Among other tools, I advertised rootograms in class for exploring model fit. Given that currently the packages **countreg** and **topmodels** are moving targets you do not need rootograms for this problem, although they are available via these packages. There are quite a few alternative options for exploring and modelling the data.

Problem: 3 (regularization)
The dataset `gr.rda` (in R binary format) contains data on the average economic growth of several countries for the years 1960–1992. There are 34 potential explanatory variables available. The purpose of the problem is to compare OLS, ridge and LASSO estimates. (The data are originally from the data archive of the *Journal of Applied Econometrics*, but here we are mainly interested in playing with our new tools.)

(a) Run an ordinary least squares regression for `y` using all explanatory variables.

(b) Run the corresponding ridge regression for `y` using all explanatory variables. Determine the tuning parameter λ using 10-fold cross validation. Obtain a plot of the MSE as a function of $\log(\lambda)$.

(c) Repeat (b) for LASSO.

(d) Compare the coefficients of all three models (e.g., by using a table or a suitable plot). Do the results match your expectations?