Final Exam, Regression Models in Economics and Business, February 2023

Directions

1. Provide an answer to each of the following problems. Include in your answer the Stata or R output from the estimation of the models and the commands you submitted to Stata or R. Do not submit output containing results that are not relevant to your answers.

2. Your answers to all three problems should be combined and submitted as a single Microsoft Word or Adobe Acrobat file. This file should be sent electronically as an email attachment to *miroslav.verbic@ef.uni-lj.si* by 23:59, Saturday, 18 February 2023.

3. You may consult your lecture notes, textbooks, and Stata or R manuals as you work on this examination. However, you should not seek help from other colleagues or artificial intelligence. The exam is individual; one participant per exam paper.

4. Please, read each question carefully. Be sure that you do everything as asked.

Exam Questions

1. In the Stata data file `carmarket.dta` we have quarterly data on the U.S. car market for the period 1976Q1–1990Q4. The variables are the following:

   - *avto*: number of cars sold (in millions);
   - *cena*: price index of new cars (1982Q1=100);
   - *rrdp*: real disposable income per capita (in 1,000 USD);
   - *om*: reference commercial bank interest rate (in %).

   a) Define the quarterly time dimension appropriately, i.e. create a time variable. Then estimate the following classical linear regression model:

   $$avto_t = \beta_1 + \beta_2 cena_t + \beta_3 rrdp_t + \beta_4 om_t + \varepsilon_t .$$

   Interpret all the regression coefficients. Do the estimates seem reasonable?

   b) Interpret the value of R^2. Prove algebraically that $R^2 = r^2_{y\hat{y}}$ by proceeding from the definition of $r_{y\hat{y}}$ (the correlation coefficient between *y* and $\hat{y}$).

   c) Test formally for normality of the disturbances, multicollinearity, homoscedasticity and autocorrelation. Write down the procedures and the results (hypotheses, test statistics, their distributions etc.). What do you find?

   d) Create a time series plot of *rrdp*. Does it appear to be stationary? Explain. Check for stationarity using an appropriate test; write down the null and alternative hypothesis, the test statistic and the results. How many lags should you use?

   e) If necessary, make the series *rrdp* stationary by using first differences. Does it appear to be stationary now; conditionally and unconditionally? Test for stationarity again.

2. In the Stata data file `ngrades.dta` we have cross-section data on the following variables:

   - *gpa*: grade point average;
   - *tuce*: test score on teaching college level economics;
   - *psi*: program participation variable (by school district);
   - *grade*: response 1 if the individual passes, 0 otherwise.

   Estimate the effect that *gpa*, *tuce* and *psi* have on the probability that an individual passes an end-of-term examination in economics (*grade* = 1 if passes, 0 otherwise). To estimate the effect of *gpa*, *tuce* and *psi* on *grade,* use the probit model, because the dependent variable is a binary variable.

   a) Are the explanatory variables in the model jointly significant at the 5% level? If not, what is their level of joint significance?
   b) What is the interpretation of the coefficient on *tuce*? Calculate and interpret its marginal effect as well.
   c) Produce a cross-tabulation of actual with predicted values for *grade*. How many of the observations in which *grade* = 1 are correctly predicted by your model? How many of the observations in which *grade* = 0 are correctly predicted?
   d) You just predicted that *grade* = 1 whenever the predicted $P(grade = 1)$ is greater than 0.5; otherwise *grade* = 0. Starting from your predicted values, improve this procedure (find a new threshold) to obtain a more accurate measure of the predictive power of this model. Explain your procedure and produce a new cross-tabulation. How many of the observations in which *grade* = 0 are correctly predicted by your new method?
   e) Based on the general specification of a probit model and its likelihood function, show theoretically that the value of the log likelihood of a probit model is always negative.

3. Stata data file `crime.dta` contains data for estimating the economic model of crime taken from Cornwall and Trumball (1994). The variables are the following:

   - *county*: county identifier;
   - *year*: years 1981 to 1987;
   - *crmrte*: crimes committed per person;
   - *prbarr*: "probability" of arrest;
   - *prbconv*: "probability" of conviction;
   - *prbpris*: "probability" of prison sentence;
   - *avgsen*: average sentence, days;
   - *polpc*: police per capita;
   - *density*: people per sq. mile;
   - *taxpc*: tax revenue per capita;
   - *west*: 1 if in western NC;
   - *central*: 1 if in central NC;
   - *urban*: 1 if in SMSA;
   - *pctmin80*: percentage minority, 1980;
   - *wcon*: weekly wage; construction;
   - wtuc: weekly wage; transport, utilities and communications;
   - *wtrd*: weekly wage; wholesale, retail trade;
   - *wfir*: weekly wage; financial institutions, real estate;
   - *wser*: weekly wage; service industry;
   - *wmfg*: weekly wage; manufacturing;

- *wfed*: weekly wage; federal employees;
- *wsta*: weekly wage; state employees;
- *wloc*: weekly wage; local government employees;
- *mix*: offense mix: face-to-face/other;
- *pctymle*: percent young male.

Additionally, the data file already contains the year dummy variables (*d82*, ..., *d87*), the logs of most variables (*lcrmrte*, ..., *lmix*), and first differences in logs of some variables (*clcrmrte*, *clprbarr*, *clprbcon*, *clprbpri*, *clavgsen*, *clpolpc*, *cltaxpc* and *clmix*).

a) Estimate a fixed effects model relating *lcrmrte* to *lprbarr*, *lprbconv*, *lprbpris*, *lavgsen* and *lpolpc*. Include also the year dummy variables.
b) Add the wage variables in logarithmic form, *lwcon* to *lwloc*, and test for joint significance of these added variables after estimation by fixed effects. Write down the procedure and the results.
c) Estimate the equation in part b) by first differences and comment on any notable changes. Do the standard errors change much between fixed effects estimates and first differences estimates?
d) Finally, estimate the model from part b) without the year dummy variables with the random effects estimator. Should you deal with the local effects in this case as fixed effects or random effects? Explain with reference to particular test statistics.