Assignment #1, 2 pages
Labor and Public Economics, CERGE-EI, Fall 2006, Daniel Münich

Using basic mincerian human capital model, the purpose of this assignment is to make you acquainted with artificial data generation mimicking empirical data and use of standard econometrics techniques on data of completely known properties.

Using Stata package, your written output should consist of:

a) **A *Stata* script** as a *.DO file we (I&Jan) could easily inspect. Sections with headlines separating individual tasks should be clearly labeled. Hints on useful *Stata* commands are provided in *italics bellow*. Variables names mentioned bellow should be given the same name (including letter size) in your script so that we could inspect your data easily.

b) **A feedback on each task 1a - 5, briefly & clearly** i) explaining how you created artificial data, ii) explaining economic or econometrics issue in question, iii) supplying necessary i.e. not redundant(!) computational outputs, iv) presenting comparison to the *underlying model*. Fit your output on 6-7 pages.

1. **Create artificial dataset** of EARN, AGE, EDU for 200 independent individuals according to the *Mincerian Human Capital Model* presented at the lecture as logY=a+b1*EDU+c1*EXP+c2*EXP^2+e, where EXP=AGE-EDU-6. In the following, the deterministic part of this equation is called underlying model.
   a. Generate right hand side variables. Variables should have plausible (in line with empirical evidence) range of values, distributions and should have reasonable V-C matrix (reflecting likely correlations of RHS variables).
   b. Set reasonable coefficients *a, b, c*, generate stochastic term *e* of plausible *iid*, and generate left-hand size variable.
   c. Present statistics sufficiently describing dataset you created and will use (*summarize, table, inspect, etc.*…).
2. **Using the dataset, do the following** (+ clarify your approach in your written output):
   a. Estimate *underlying* model by OLS using *underlying* functional form. Check how your estimates compare to actual parameters of the underlying model (*test*).
   b. One by one, exclude RHS variable from the specification and run OLS.
   c. Estimate *underlying* model using earnings levels instead of logs.
   d. Estimate experience of maximum earnings, EXP*, and test its statistical difference from

point value 35 years and from the value given by the *underlying model*. (*test, nltest*)

3. Case by case, introduce additional peculiarities into the original dataset:
   a. Introduce heteroskedasticity of plausible kind defining logYHET instead of logY. Re-estimate the model by simple OLS and by more suitable estimation method. Present a scatter plot showing presence of heteroskedasticity.
   b. Introduce measurement error of plausible structure in the EDU variable creating new variable EDUERR. Re-estimate *underlying* model in #1. Generate valid instrument EDUIV and re-estimate the underlying model by IV method. (*ivreg*)
   c. Incorporate plausible measurement error in the LHS variable, logYERR and re-estimate the *underlying* model.
   d. Re-estimate underlying model considering 3rd order polynomial of EXP.
   e. Use 2nd order polynomial of AGE instead of EXP and re-estimate the underlying model.
4. Assume that the *underlying* functional form of the earning-experience profile is not known to you and employ method of splines. Re-estimate underlying model. (*mkspline*)
5. Do following exercise: repeat (loop) task #1) 300times using different *seed* for each run and save estimated B1, of estimated coefficient b1 from each run. Show distribution of estimated B1 (*svmat, kdensity*).