**Advanced Econometrics II Winter 2022/2023**

**Weak instruments assignment**

**Due date: Friday January 20–th, printed version before 5pm in the box in the common space on the 4-th floor of the second tower of the E-building. Max number of pages is four including figures! No code. Just plain answers to the questions, no need for a story.**

1. Size distortions: To illustrate the size distortions of the 2SLS $t$-statistic, we simulate data from the following model:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ X &= Z\Pi + V \end{aligned}$$

where $Y$ and $X$ are $N \times 1$ vectors which contain the endogenous variables and $Z$ is a $N \times k$ matrix of instruments. $\varepsilon$ and $v$ are $N \times 1$ vectors that contain the disturbances. The different rows of $(\varepsilon \vdots V)$, $\binom{\varepsilon_i}{V_i}'$, $i = 1, \ldots, N$, are independently normal distributed: $\binom{\varepsilon_i}{V_i} \sim N(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \vdots & \rho \\ \rho & \vdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \vdots & 0 \\ \rho & \vdots & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} 1 & \vdots & \rho \\ 0 & \vdots & \sqrt{1-\rho^2} \end{pmatrix}$. We use $N = 100$, $k = 10$, $\Pi = a \times e_{10}$ with $e_{10}$ a $10 \times 1$ vector whose top element is one and all remaining elements are equal to zero. All elements of $Z$ are independently standard normal distributed. We only simulate them once and keep them fixed throughout the simulation experiment.

   We use eight values of $a$ : (0.4 0.3 0.25 0.2 0.16 0.1 0.05 0) and ten different values of $\rho$ : (0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.9 0.95).

   For each value of $a$ construct a graph of the rejection frequency as a function of $\rho$ when testing $\mathrm{H}_0 : \beta = 0$ with 95% significance using the 2SLS t-statistic. Use 5000 simulations from the above model. You can combine these eight graphs into one figure. What do you conclude? (Start each simulation run with the same seed so you use the same random numbers for each of the graphs. The command: "np.random.seed(123)" sets the seed for the Python random number generator).

2. Compute and make a figure of the 95% critical value function of the LR statistic as a function of $r(\beta_0)$ for $k = 10$. Can you say anything specific about the critical value when $r(\beta_0) = 0$ or infinite? (Use the same seed for each simulation run!).

3. Repeat the excercise from 1 for the AR, score and LR statistics. What do you conclude? (Use the same seed for each simulation run!).

4. Compute and make a figure of the 95% critical value function of the LR statistic as a function of $r(\beta_0)$ for $k = 3$. (Use the same seed for each simulation run).

5. Card (1993)[1] analyzes the return on education. He uses different proximity to college variables as instruments. The file assigmentweakinstruments.csv contains that part of the Card data which we use for this assignment. The different variables in the file are: nearc2: if near a 2 year college, nearc4: if near a 4 year college, nearc4a: if near a 4 year community college, nearc4b: if near a 4 year private college, ed: years of education, wage: log-earnings, age: age in years, age2: squared age, exper: experience, exper2: experience squared, south: lives in the South, smsa: lives in a metropolitan area, race: racial indicator.

   The variables wage and ed constitute the endogenous variables ($y$ and $x$), nearc2, nearc4, nearc4a are instruments ($z$) and exper, exper2, south, smsa, race and the constant term are the included exogenous variables ($w$) (We do not use age and age2. Do not explicitly compute the $M_W$ matrix when you partial out $W$. This is like a $3000\times3000$ matrix which Python or your computer doesn't like).

   (a) Using only nearc2 as an instrument, construct the 95% confidence set for the return on education using the 2SLS $t$-statistic and the AR statistic.

   (b) Is there a difference between these confidence sets and if so can you explain why this difference occurs?

   (c) What is the value of the first stage F-statistic and what does the value of the AR statistic look like when the tested parameter is large. Can you also prove this hunch?

   (d) We did not use the LM and LR statistics in a or did we?

   (e) Using nearc4, nearc2 and nearc4a as instruments, construct the 95% confidence set for the return on education using the 2SLS $t$-statistic, AR, LM and LR statistics.

   (f) Is there a difference between these confidence sets and if so can you explain why this difference occurs?

   (g) Since the model is over-identified, we can test whether the instruments are exogenous. Do so using the over-identification test (which coincides with the minimal value of the AR-statistic over all values of the return on education parameter). Use the correct degrees of freedom for the $\chi^2$ limiting distribution which is ...?

[1] Card, D. Using geographic variation in college proximity to estimate the return to schooling. In L.N. Christofides, E.K. Grant and R. Swidinsky, editor, Aspects of Labour Market Behaviour: essays in honor of John Vanderkamp, pages 201—222. University of Toronto Press, Toronto, Canada, 1995. (NBER Working Paper 4483 (1993)).