ECOM-G314 Econometrics 1
Homework Assignment 3

This homework assignment will be discussed in the exercise session on **Tuesday** 5 December (groups at 12.15pm and 2.15pm) in seminar room 3–4 at Economicum. Please submit your solution **by 11.45 a.m. on Tuesday 5 December**. Note that the deadline and the exercise session are exceptionally on Tuesday! (Wednesday is the Finnish Independence Day, a bank holiday.)

Peer review and self-assessment should be done by **Monday 11 December at 6 p.m.** at the latest. Please note that the peer review and self-assessment are compulsory, and a prerequisite for gaining points from your submission.

Tutorials will be held on Mondays 27 November and 4 December at 2.15pm in the Economicum lecture hall. You can ask the TA for help with the the homework assignments and discuss the assignments with other students. You may also ask questions regarding assignment 4. There is also an **extra tutorial session** on Friday 8 December at 2.15pm in seminar room 3–4 at Economicum. If you have any questions, please contact the TA via email at heikki.korpela@helsinki.fi.

The share of each exercise of the maximum number of points from the assignment is given in brackets.

Please return your submission in Moodle as one PDF file. It is not strictly necessary to return the code used, but if there are errors in your results, the code may be helpful in deciding whether you've made a fundamental or a minor mistake.

The peer review is anonymous. For this reason, please do **not** include your name or student ID in your submission, in the filename or the file description.

1. Consider the wage equation in the empirical example in the video lecture on IV estimator. The data used in the example is in the file `schooling.txt`. [35%]

   The baseline model is also implied by the code on the lecture slides on instrumental variables (p. 8–9); it was

   $$\log(wage) = schooling + exper + exper^2 + black + smsa + south + iqscore,$$

   where *wage* stands for wage (`WAGE76` in data), *schooling* for education (`ED76` in data), *exper* for work experience (`EXP76` in data), *black* for being black (`BLACK` in data), *smsa* for living in a metropolitan SMSA ((`SMSA76` in data)) and *south* for living in the south (`SOUTH76` in data). Further details on the model and the data appear on the textbook, chapter 5.4.

   (a) It was argued that *schooling*, and thus also *exper* and its square that depend on it, are endogenous because of a missing general ability variable in the model. The file `schooling.txt` contains the variable *iqscore* (an intelligence test score) that can be used as a measure of general ability. Add *iqscore* to the model for the log wage discussed in the video, and estimate the augmented model by OLS. Interpret the coefficient estimates of *schooling* and *iqscore*. What might explain

the difference in the the estimate of the coefficient of *schooling* compared to the model excluding *iqscore*?

References/tips: the empirical example in the lectures is covered on slides "Instrumental variables estimator", p. 8–9. In the book, see chapter 5.4. As in the example code `schooling.R` (week 5), you probably want to read `schooling.txt` with `read.table` (the first row of the data has some extra whitespace which might confuse some import functions). Here, you might think about important unobserved variables that are correlated with both schooling and IQ scores; see also slides 3–5 on "Inconsistent OLS estimator".

The full dataset and its description are available from David Card's homepage; the original dataset includes rows where log wage is missing, has more variables, and some of the variables have slightly different names. The working paper on this topic is freely available.

The dataset contains missing values for the variable *iqscore*. One way to restrict the dataset to rows with no missing values in R is

```
schooling <- read.table('schooling.txt', header = T)
schooling <- subset(schooling, !is.na(iqscore))
```

(b) Test for heteroskedasticity of the error term using the Breusch-Pagan test in the model in (a). Compute the heteroskedasticity consistent standard errors, and test the significance of the coefficient of `schooling`. Is the conclusion different compared to the significance test based on assuming homoskedasticity?

References/tips: See the previous week's homework. Note that functions like `coeftest` also report standard errors when given a covariance matrix; you do not need to report the matrix itself separately.

(c) As pointed out by Verbeek, omitting the ability variable is not the only potential reason for the endogeneity of *schooling*. Estimate the augmented model using *age* (`AGE76` in data) and its square and *nearcollege* (`NEARC4` in data) as instruments for *exper*, its square and *schooling*. Interpret the coefficient estimates of *schooling* and *iqscore*, and compare them to those obtained by OLS in (a).

References/tips: as with example code `schooling.R` (week 5), you may find `ivreg` (from the package with the same name) helpful. In particular, consider this bit from `?ivreg`:

> "Regressors and instruments for ivreg are most easily specified in a formula with two parts on the right-hand side, e.g., `y ~ x1 + x2 | z1 + z2 + z3`, where x1 and x2 are the explanatory variables and z1, z2, and z3 are the instrumental variables. Note that exogenous regressors have to be included as instruments for themselves.
> For example, if there is one exogenous regressor ex and one endogenous regressor en with instrument in, the appropriate formula would be `y ~ en + ex | in + ex`. Alternatively, a formula with three parts on the right-hand side can also be used: `y ~ ex | en | in`. The latter is typically more convenient, if there is a large number of exogenous regressors."

See also chapters 5.3.1, 5.3.4, 5.4 in the book, and the slides on instrumental variables estimator.

(d) Interpret the Durbin-Wu-Hausman test result in (c). Is there any evidence of a weak instruments problem?

References/tips: You may find the `summary` on an object returned by `ivreg` helpful. For weak instruments and the DWH test, see slide 7 on "Instrumental variables estimator" (week 4) and slides 10–11 on "Generalised instrumental variables estimator" (week 5). In the book, see chapter 5.6.4 and pp. 153–154 in chapter 5.3.1.

(e) Suppose also *iqscore* is endogenous, and estimate the augmented model using as instruments *momseducation* (mother's education, `MOMED` in data), *dadseducation* (father's education, `DADED` in data) and *kww* (the score of a general knowledge test, `KWW` in data) in addition to *age* and its square and *nearcollege*. Interpret the coefficient estimates of *schooling* and *iqscore*, and compare them to those obtained by OLS in (a).

References/tips: see (c).

(f) Interpret the results of the over-identifying restrictions test and Durbin-Wu-Hausman test in (e).

References/tips: again, you may find the `summary` on an object returned by `ivreg` helpful. Regarding the over-identifying restrictions test, see slides 8–9 on "Generalised instrumental variables estimator" (week 5), and chapter 5.6.3 in the book.

2. Consider, again, the model for log wage discussed in the video lecture. Estimate the IV model in two stages involving OLS regressions. Verify that the coefficient estimates are identical to those obtained in one stage, but the standard errors are different. [15%]

References/tips: see how a regression for *schooling* on the left hand side is done in the example code `schooling.R`. Could you do the same for the other endogenous variables in the model? Note also that if `results_modelX` holds your OLS results, you can extract fitted values with `fitted(results_modelX)` in R.

The idea of the procedure is outlined on slide 7 (compare to slide 6 as needed) of "Generalized instrumental variables estimator".

3. The file `money.xlsx` contains the easonally adjusted quarterly time series of the logarithm of the real money supply, m_t, real GDP, y_t, and the 3-month Treasury Bill rate, r_t, for Canada for the period 1967:3 to 1998:4. In the file, the variables are named as `m`, `y`, and `r`, respectively, while `m1` and `m2` are m_{t-1} and m_{t-2}, respectively, and `r1` and `r2` are r_{t-1} and r_{t-2}, respectively. A conventional money demand function is [25%]

$$m_t = \beta_1 + \beta_2 r_t + \beta_3 y_t + \beta_4 m_{t-1} + \beta_5 m_{t-2} + \varepsilon_t.$$

(a) Estimate the model by OLS. Interpret the estimation result.

(b) Test for conditional homoskedasticity of the errors by the Breusch-Pagan test and for first-order error autocorrelation by the Breusch-Godfrey test. Compute the heteroskedaticity consistent covariance matrix estimator, and test for the significance of the coefficients using heteroskedasticity robust standard errors.

References/tips: see previous week's exercises.

(c) Estimate the model by 2SLS, treating r_t as an endogenous variable and using r_{t-1} and r_{t-2} as additional instruments. Are the estimates much different from the OLS estimates?

References/tips: see 1c).

(d) Show that the model estimated in (c) is over-identified. Test the over-identifying restrictions. Interpret the results of the Durbin-Wu-Hausman test. Is there evidence of a weak instruments problem?

References/tips: see 1d) and 1f).

4. Consider the simple linear regression model [25%]

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Assume that x_i is independently normally distributed with mean zero and variance $\sigma_X^2 = 1$, and the error term $\varepsilon_i = \rho x_i + \eta_i$, where η_i is independently normally distributed with mean zero and variance σ_η^2=1.

(a) Write down the moment conditions for OLS estimation of β_1 and β_2. Show that if $\rho \neq 0$, the error term is correlated with the regressor x_i, and one of the moment conditions is violated.

References/tips: see slide 3 on "Instrumental variables estimator" for moment conditions. For derivation of the correlation, recall that if random variables z and q are independent, then $\mathbb{E}(zq) = \mathbb{E}q\mathbb{E}z$ (or check ''Some useful results" from problem set 1). Also note that to derive a non-zero correlation, you only need a non-zero covariance, and that covariance can be written as $\text{Cov}(x_i, \varepsilon_i) = \mathbb{E}x_i\varepsilon_i - \mathbb{E}x_i\mathbb{E}\varepsilon_i$ (see, for example, appendix B.4 in the book).

(b) Let $\beta_1 = 0.0$ and $\beta_2 = 1.0$. Consider three values of ρ, $\rho = 0$, $\rho = 0.25$, and $\rho = 0.50$. For each ρ, generate $S = 5000$ samples of size $N = 25$ from the regression model, and for each generated sample, compute the OLS estimate of β_2 and the t-test statistic for $H_0 : \beta_2 = 1.0$ against $H_1 : \beta_2 \neq 1.0$.

Because 1.0 is the true value of β_2, H_0 should be rejected in 5% of the replications in the t-test conducted at the 5% level of significance (the nominal size of the test).

Compute the rejection rate of the test, i.e., find the proportion of the replications where the absolute value of the t-test statistic exceeds the critical value (1.96).

Plot also the histogram of the OLS estimator of β_2. How does the rejection rate and the distribution of the estimator vary with ρ? [Hint: If the model is estimated using the `lm()` function in R and the result is stored in `ols1`, the OLS estimate of β_2 is obtained as `coef(ols1)['x']` and the covariance matrix estimator of the OLS estimator as `vcov(ols1)`.]

References/tips: see the previous week's suggested solutions on how to perform t-tests on each iteration of repeatedly simulating data. Standard errors for the regressor x can be extracted by, for example, `sqrt(vcov(model)['x','x'])` or `coef(summary(model))['x','Std. Error']`. If the null hypothesis specifies the coefficient as β_k^0, the coefficient estimate is b_k and the standard error is $se(b_k)$, the t-test statistic is defined as $\frac{b_k - \beta_0^k}{se(b_k)}$; see chapter 2.5.1 in the book or slides 2–3 in week 1's slides Testing hypotheses under the normality assumption.

A histogram can be plotted using the function `hist` (or `geom_histogram` for fancier plots with `ggplot2`).

(c) Repeat (b) for sample size $N = 100$, and compare.