ASSIGNMENT 2

Instructions

1. The assignment is due on April 9 (23:59).
2. Hand in your assignment in Canvas.
3. We highly encourage you to use R Markdown throughout the exercise. You can write both text and equations, and a lot of help on how to use it can be found online.[1] From the Markdown file, you can generate a PDF file by clicking on the "Knit" button, which then is nicely formatted and shows clear answers. Furthermore, we provide a template solution file, which you can easily fill in accordingly.
4. When answering the empirical exercise, the code used and the results (i.e. values of statistics) from the code need to be clearly stated and linked to your answers. In particular, first state the R code you use, then provide the values of the statistics you calculate below the code, and beneath that provide your answer that relates to the code. We highly encourage you to use R Markdown, since it combines code with text in a readable format, as we show in the answer template we provide.
5. Using built-in R functions is not allowed. Build up everything from basic matrix algebra. For simple operations such as calculating a mean, and for operations related to distributions (e.g. taking random draws, getting the quantile), you can use the built-in R functions.
6. Work in groups of three.
7. State your full name and your student number on the front page of the assignment.

[1] E.g. http://www.stat.cmu.edu/~cshalizi/rmarkdown/ provides a great summary of all necessary commands.

Theoretical Exercises

1. Consider the following regression model:

$$Y_i = X_i^\top \beta + \varepsilon_i, \qquad X_i \in \mathbb{R}^k,$$

where the error term is heteroskedastic, i.e., $\mathsf{E}[\varepsilon_i^2 \mid X_i] = v(X_i)$ for some function $v : \mathbb{R}^k \to \mathbb{R}_{>0}$. Assume that the rest of the standard assumptions hold. Using the steps below, verify that it is possible to construct an estimator of β that is asymptotically more efficient than the OLS estimator when the function v is known.

(1) Let $w : \mathbb{R}^k \to \mathbb{R}_{>0}$ be such that $\mathsf{E}[X_1 X_1^\top w(X_1)]$ and $\mathrm{Var}(X_1 w(X_1)\varepsilon_1)$ are finite and positive definite. Find the asymptotic distribution of the *weighted least squares* estimator

$$\hat\beta_n^{\mathrm{WLS}} = \underset{b}{\mathrm{argmin}} (\boldsymbol{y} - \boldsymbol{X} b)^\top W (\boldsymbol{y} - \boldsymbol{X} b),$$

where $W = \mathrm{diag}(w_1, \ldots, w_n)$ and $w_i \equiv w(X_i)$. Specifically, verify that the asymptotic variance of $\hat\beta_n^{\mathrm{WLS}}$ is given by

$$V = \left(\mathsf{E}[X_1 X_1^\top w_1]\right)^{-1} \mathsf{E}[X_1 X_1^\top w_1^2 \varepsilon_1^2] \left(\mathsf{E}[X_1 X_1^\top w_1]\right)^{-1}.$$

(2) Set $w(X_i) = (v(X_i))^{-1}$ and compute the asymptotic variance, V^*, in this case, assuming that v is such that V^* is finite and positive definite.

(3) Find the matrix H such that

$$V^{*-1} - V^{-1} = H^\top \mathsf{E}[Z_1 Z_1^\top] H,$$

where

$$Z_i = \begin{bmatrix} X_i / \sqrt{v(X_i)} \\ X_i \sqrt{v(X_i)} w_i \end{bmatrix}.$$

(4) Conclude $V - V^*$ is positive semidefinite for any suitable weighting function w. In particular, this holds for $w \equiv 1$. [**Hint:** use the fact that if A and B are positive definite matrices of the same size, then $B - A$ is positive semidefinite iff $A^{-1} - B^{-1}$ is positive semidefinite.]

2. Consider a model with a single endogenous regressor:

$$\begin{aligned} Y_{1i} &= X_i^\top \beta + \gamma W_i + \varepsilon_i, \\ W_i &= Z_i^\top \pi + v_i. \end{aligned}$$

That is, W_i is endogenous, and Z_i is the vector of all exogenous variables such that $\dim(Z_i) \geq \dim(X_i)+2$. We estimate $\theta = (\beta^\top, \gamma)^\top$ using the following 2-step procedure:

(1) Estimate the second equation by OLS and compute the residuals $\{\hat v_i\}$;

(2) Estimate θ using the following regression:

$$Y_i = X_i^\top \beta + \gamma W_i + \rho \hat v_i + u_i.$$

Show that the OLS estimator of θ in this case is identical to the 2SLS estimator:

$$\hat{\theta}_n^{\text{2SLS}} = (\boldsymbol{X}'^\top P_{\boldsymbol{Z}} \boldsymbol{X}')^{-1} \boldsymbol{X}'^\top P_{\boldsymbol{Z}} \boldsymbol{y},$$

where $\boldsymbol{X}' = [\boldsymbol{X} \quad \boldsymbol{W}]$. [**Hint:** use Equation (1) in Lecture 3 and the fact that $\boldsymbol{X}^\top \hat{\boldsymbol{v}} = 0$.]

Empirical Exercise

In this exercise we continue exploring the Blackburn & Neumark's dataset (use the dataset attached to this assignment as it contains more variables compared to the one of Assignment 1; the description of the variables is given in Figure 1.) You want to estimate the following wage equation:

$$lwage_i = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + \beta_3 black_i + \beta_4 educ_i + \varepsilon_i,$$

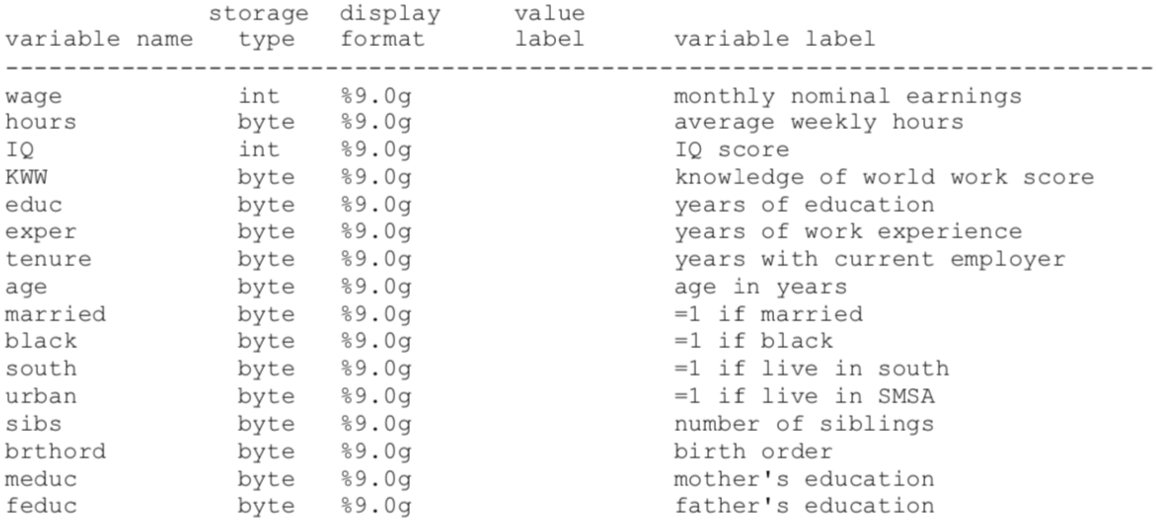
and you are particularly concerned that $educ_i$ is potentially endogenous.

(a) Estimate the model using OLS. Interpret the coefficient estimate of education. (Note that the dependent variable is in logs and this has an implication on how you interpret the coefficient.)

(b) Perform the IV regression using $meduc_i$ as an instrument for $educ_i$. Report the estimated coefficients and the corresponding standard errors. Compare the estimated return to education with one you obtained in part (a).

(c) Perform the 2SLS regression using $meduc_i$, $feduc_i$, and $sibs_i$ as instruments for $educ_i$ and compare the results with part (b). Discuss the validity of the suggested instrumental variables.

(d) As we argued in class, the 2SLS estimator of the vector of coefficients is less efficient than the OLS estimator when $educ_i$ is exogenous. Carry out the Hausman test to check whether IV estimation is required and discuss the results.



```
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
wage            int    %9.0g                  monthly nominal earnings
hours           byte   %9.0g                  average weekly hours
IQ              int    %9.0g                  IQ score
KWW             byte   %9.0g                  knowledge of world work score
educ            byte   %9.0g                  years of education
exper           byte   %9.0g                  years of work experience
tenure          byte   %9.0g                  years with current employer
age             byte   %9.0g                  age in years
married         byte   %9.0g                  =1 if married
black           byte   %9.0g                  =1 if black
south           byte   %9.0g                  =1 if live in south
urban           byte   %9.0g                  =1 if live in SMSA
sibs            byte   %9.0g                  number of siblings
brthord         byte   %9.0g                  birth order
meduc           byte   %9.0g                  mother's education
feduc           byte   %9.0g                  father's education
```

Figure 1: Description of the variables