Problem Set 4

AAEC 5126 / Instructor: Klaus Moeltner

General Instructions

Complete the **R**-based assignments in a sweave file that shows your code, output, and discussion. Hand in the compiled (pdf) version. For the theory questions, you may want to work out your answer on paper first, then type up the key steps in LaTeX. You can insert your LaTeX code directly in in your sweave file, or - if you prefer, in a separate file or files in TeXnicCenter. For full credit, typed answers are required.

You can work with others, but please hand in your own version. Please report any glitches as soon as you discover them - thanks!

Q1: Instrumental Variables / TSLS

Consider the following regression model: $h_i = \beta_1 age_i + \beta_2 ex_i + \varepsilon_i$

where h_i is a (continuous) health index for professional worker *i*, age_i is the age of worker *i*, and ex_i is the hours of exercise per week for worker *i*. Assume all these (and subsequent variables) are expressed as deviations from their respective mean (So we don't have to worry about intercept terms, which will make the following a bit easier). The full model can thus be written as

$\mathbf{h} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X} = [\mathbf{age} \quad \mathbf{ex}]$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$.

(a) Why might you suspect that exercise could be correlated with the error term? (provide some reasoning / intuition).

(b) If this is the case (i.e $\text{plim}\left(\frac{1}{n}\mathbf{ex}'\boldsymbol{\varepsilon}\right) = \varphi \neq 0$) determine whether b_{OLS} is a consistent estimator for $\boldsymbol{\beta}$.

(Hint: note that $\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \begin{bmatrix} \frac{1}{n}\mathbf{age}'\boldsymbol{\varepsilon} \\ \frac{1}{n}\mathbf{ex}'\boldsymbol{\varepsilon} \end{bmatrix}$). Assume that plim($\mathbf{X}'\mathbf{X}$/n) = $\mathbf{Q_{XX}}$, a well-behaved finite matrix

(c) Suppose you have information on all workers in your sample for two additional variables: "distance from home to nearest health club" (dh_i), and "distance from work to nearest health club" (dw_i). Assume neither of these variables are correlated with $\boldsymbol{\varepsilon}$, i.e. $\text{plim}(\mathbf{dh}'\boldsymbol{\varepsilon}) = \text{plim}(\mathbf{dw}'\boldsymbol{\varepsilon}) = \mathbf{0}$. Why might these variables be good instruments for *exercise*?

(d) Show how these additional variables can be used to derive a consistent TSLS estimator for $\boldsymbol{\beta}$ (show all detailed steps). Proof that this estimator is indeed consistent. (Assume that plim($\mathbf{Z}'\mathbf{Z}$/n) = $\mathbf{Q_{zz}}$ and plim($\mathbf{Z}'\mathbf{X}$/n)= $\mathbf{Q_{zx}}$ are well-behaved finite matrices.)

(e) What would you use for a consistent estimator for σ^2? (show detailed expression)

(f) Outline in detail how a Hausman test and a Wu test could be performed to test H_0: $\varphi = 0$

Q2.) Instrumental Variables and Specification Tests in ***R***

Use Greene's quarterly macroeconomic data (data set "consumption" on our course web site).

Consider the model

$$y_t = \beta_0 + \beta_1 dpi_t + \beta_2 cpi_t + \beta_3 rate_t + \varepsilon_t \quad (1)$$

where *"t"* indexes the current time period, *y* = aggregate consumption (billion dollars, denoted as "*realcons*" in the variable list), *dpi* = aggregate disposable income ("realdpi" in the list), *cpi* = consumer price index, and *rate* = real interest rate ("realint" in the list).

You suspect that *dpi* is correlated with the error term for the same time period. You decide to instrument it with "dpi_{t-1}" and "y_{t-1}", i.e. lagged dpi and lagged consumption.

Use the procedure outlined in script *mod4s1b* to generate all needed lagged variables.

1. Run the simple OLS model given in (1).
    a. Comment on the significance levels of the estimated coefficients. Are the signs of the significant coefficients as expected? Explain.

2. Run the TSLS model with the instruments given above. Comment on any changes in coefficient estimates and significance levels compared to the OLS model.

3. Perform a Hausman test.
    a. State the null hypothesis (H_0) and alternative hypothesis (H_1) for this test.
    b. Using the p-value generated by ***R*** to draw a conclusion for your H_0.

4. Perform a Wu test.
    a. State the null hypothesis (H_0) and alternative hypothesis (H_1) for this test.
    b. Using the p-value generated by ***R*** to draw a conclusion for your H_0.

Q3: Heteroskedasticity – ***R***

Sample data for the analysis of home prices as a function of home and neighborhood features are notorious for heteroskedasticity problems. For example, as you can imagine, the value of certain home features and thus home prices are more likely to fluctuate more widely for larger homes.

Consider the data set "homeprice" (on our web site). It contains observation on home prices and features for a Seattle suburb for home sales during 1985-1989. There are 14 columns and 100 rows. The variables are as follows:

id	running id
price	sale price (in 1989 dollars)
ln_price	log of price
tsqft	total square footage
bedrms	number of bedrooms
bathrms	number of bathrooms
age	age of home
garage	existence of a garage (1=yes, 0=no)
view	existence of a mountain view (1=yes, 0=no)
firepl	number of fireplaces
porch	existence of a porch (1=yes, 0=no)
distance	distance to lake (in 100 feet)

sewer	hooked up to municipal sewer system (1=yes, 0=no)
year	year of sale

- load the data into ***R***
- define **price/1000** as your dependent variable (**y**).
- define **X** to include a column of ones, **tsqft/1000**, bedrms, bathrms, garage, view and distance. So **X** should be 100 by 7;

a. Run a generic OLS regression and show your output.

b. You suspect potential HSK, if present, to be related to total square footage (tsqft), number of bedrooms, and number of bathrooms. Derive a residual-vs.-predictor plot for each, using *mod4_2b* for guidance. Do the plots provide indication for HSK? Make sure the graphs are added to your output.

c. Perform a Breusch-Pagan score test using the same three explanatory variables as HSK-driving suspects. Show the test results and state your test decision.

d. Then perform a White test, capture the results and state your test decision. Make sure to include all *permissible* interactions in your augmented data matrix.

e. Estimate a robust OLS model with White-corrected standard errors. Show your output.

f. Using the same HSK suspects, estimate your model through FGLS, using a multiplicative *(don't forget the Harvey correction)* form to model HSK. Show your output.

g. Compare your original OLS estimates, the White corrected estimates, and the FGLS results and elaborate:
    a. Compare the s.e.'s and t-values between OLS and robust OLS. Are there any noteworthy changes in significance levels? In light of your finding, how does the naive OLS model mis-represent the significance of one or more coefficients?

    b. Compare the s.e.'s and t-values between the robust OLS and the FGLS model. Are there any noteworthy changes in significance levels?

    c. Assume the main focus of your research is on the effect of "view" and "distance" on home prices. Overall, which model would you choose? (think: Are the gains in significance via FGLS worth the risk of misspecification bias? What about the sample size?).