Final Assignment

This final assignment accounts for 40% of your final grade. It is **due on Dec 24th, 2024**. Please submit your report, **together with your code** used to analyze these questions, to the following email: zhenghuan0789@163.com. (结果输出，图表呈现等可通过截图方式；建议使用 Stata，但也可使用任何你所熟悉的统计软件进行分析处理)。

Part 1. Basic regression analysis.

Q1. OLS regression

Use the data in HPRICE1.RAW to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has *sqrft* = 2,438 and *bdrms* = 4. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was \$300,000 (so *price* = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

Q2. OLS partial regression coefficients

Using the 526 observations on workers in WAGE1.RAW, we include *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer) in an equation explaining log(*wage*). The estimated equation is

$$\widehat{\log(wage)} = .284 + .092\, educ + .0041\, exper + .022\, tenure \qquad [3.19]$$
$$n = 526.$$

Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out for Example 3.2. This first requires regressing *educ* on *exper* and *tenure* and saving the residuals, $\hat{r}_1$. Then, regress log(*wage*) on $\hat{r}_1$. Compare the coefficient on $\hat{r}_1$ with the coefficient on *educ* in the regression of log(*wage*) on *educ*, *exper*, and *tenure*.

Q3. IV regression

The data in FERTIL2.RAW include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

(i) Estimate the model

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

by OLS, and interpret the estimates. In particular, holding *age* fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

(ii) The variable *frsthalf* is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that *frsthalf* is uncorrelated with the error term from part (i), show that *frsthalf* is a reasonable IV candidate for *educ*. (*Hint:* You need to do a regression.)

(iii) Estimate the model from part (i) by using *frsthalf* as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from part (i).

(iv) Add the binary variables *electric*, *tv*, and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

Q4. GLS regression

Use the data set GPA1.RAW for this exercise.

(i) Use OLS to estimate a model relating *colGPA* to *hsGPA*, *ACT*, *skipped*, and *PC*. Obtain the OLS residuals.

(ii) Compute the special case of the White test for heteroskedasticity. In the regression of $\hat{u}_i^2$ on $\widehat{colGPA}_i$, $\widehat{colGPA}_i^2$, obtain the fitted values, say $\hat{h}_i$.

(iii) Verify that the fitted values from part (ii) are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?

Part 2. Simulation analysis.

Q5. Bias in OLS estimator caused by endogeneity.

Start by generating two samples of 10,000 random draws from the standard normal distributions. Name the two variables as x_i (endogenous variable) and u_i. Generating the disturbance term $\varepsilon_i = x_i + u_i$ and $y_i = 1 + 2x_i + \varepsilon_i$. We take this to be the population, and then do the following:

We drew **1,000** random samples of **500** observations on (y_i, x_i) from the above population, and for each of the 1,000 samples, run the simple regression of y_i on x_i and obtain the estimated coefficient on x_i. So we will get 1,000 such estimated coefficients.

a. Output the distributional statistics (including the mean, median, standard deviation, maximum, and minimum) of these 1,000 estimated coefficients;

b. Plot both the density of the OLS estimator and of the normal distribution in the graph for comparison. What can you conclude? (e.g., Stata code: kdensity b, xline(2) normal normopts(lp(dash)), where b is the variable containing the 1000 estimated coefficients on x_i from part a).

Q6. Exploring the consequences of weak instruments.

Start by generating three samples of 10,000 random draws from the standard normal distributions. Name the three variables as ε_i, u_i, and z_i (instrumental variable). Generating an endogenous variable $x_i = \mathbf{0.01}z_i + 0.2\varepsilon_i + 0.1u_i$ and $y_i = 1 + 2x_i + \varepsilon_i$. We take this to be the population, and then do the following:

We drew **1,000** random samples of **500** observations on (y_i, x_i, z_i) from the above population, and for each of the 1,000 samples, run the two-stage least squares regression (Stata command "ivregress 2sls") of y_i on x_i and obtain the estimated coefficient on x_i. So we will get 1,000 such estimated coefficients.

a. Output the distributional statistics (including the mean, median, standard deviation, maximum, and minimum) of these 1,000 estimated coefficients;
b. Plot both the density of the 2sls estimator and of the normal distribution in the graph for comparison. What can you conclude?

Part 3. Analyzing stock risk based on data sets from commercial platforms.

***Q7*. Estimating the beta coefficients of stocks**

Beta coefficient (贝塔系数) is a well-known measure for systematic risks (系统性风险) of stocks in financial markets. Empirically it is usually computed as the coefficient estimate of β in the following regression:

$$Return_{i,t} = \alpha + \beta \cdot Return_{m,t} + \varepsilon_{i,t}$$

Where $Return_{i,t}$(个股收益率) is the stock return for firm *i* during period *t* and $Return_{m,t}$ is the market return (市场收益率) during period *t*. *t* can be daily, weekly, or annual. Now we have 19 csv. files containing weekly returns for all listed companies (A 股上市公司) as well as for the total market through the period from 2010 to 2021 (these data are all downloaded from the RESSET 金融研究数据库 - RESSET/DB). Do the following tasks:

(i) Append all these 19 files into a single one to get the complete data set. You can use any software to append the data.

(ii) **For each firm in each year**, run the above regression to obtain the estimated beta ($\hat{\beta}$). Firms are identified by the variable '*股票代码_stkcd'*. To run the regression, use the variable '*周收益率_wkret*' as the dependent variable and '*总市值加权平均市场周收益率_wretmc*' as the independent variable. Output the distributional statistics (including the mean, median, standard deviation, maximum, and minimum) of these estimated coefficients and plot the density. (Note: To mitigate the impacts of outliers (极端值), winsorizing (缩尾处理, Stata command: winsor) the above estimated coefficients at the 1% and 99% percentiles of the distribution before output outputting summary statistics and plotting the density).

(iii) From all the companies, select 5 companies you are most familiar with or interested in, estimate the beta for each company in each year and plot a graph showing the time-series of the estimated coefficients (时间趋势图).