Applied Econometrics - Pset 1
Introduction to Stata
and Review of Linear Models

Hosein Joshaghani

Due date: **Monday 1396/11/23 - 9:30 AM**

1 Stratified Sampling vs Simple Random Sampling

Imagine that your research funds are limited to survey input and output information of only $N^{sample} = 100$ manufacturing establishments while there are $N^{pop} = 100,000$ of those establishments in the economy, in other words you are restricted to draw a 0.1% sample from the population. Using last year's census data you have the identification code for each of N^{pop} establishments.

1.1 Simple Random Sampling

Explain how you would draw a simple random sample of size N^{sample} for the population and what would be the likelihood of each establishment to be chosen in your sample.

1.2 Stratified Sampling

There are two kinds of establishments in the economy: small-medium establishments with less than 1,000 workers and large establishments with more than 1,000 workers. In total, there are $N^{pop}_{s-m} = 99,000$ small-medium establishments and only $N^{pop}_{large} = 1,000$ large establishments in the economy.

1.2.1 Employment Share

If average employment in large establishments is 2,500 workers, and small-medium establishments have 50 workers on average, then compute share of manufacturing employment, who are

employed in large establishments.

1.2.2 Some Unpleasant Simple Sampling Arithmetics

Briefly explain the problem with simple random sampling and why you may want stratified sampling. [hint: how many large firms do you expect to be found in any simple random sample?]

1.2.3 Stratified Sampling and Probability of being Sampled

You are convinced that in order to have reliable inference about large establishments you need at least 40 observations. Explain how you draw stratified sample with $N_{large}^{strata} = 40$ large firms and $N_{s-m}^{strata} = N^{sample} - N_{large}^{strata}$ small and medium establishments. What is the probability for each large firm to be chosen in your stratified sample? What is the probability for each small-medium establishment to be chosen in your stratified sample?

1.2.4 Stratified Sampling and Sampling Weights

Define weight of each observation i with w_i such that

$$\sum_{i=1}^{N^{sample}} w_i = N^{pop}$$

Derive w_i for the case of simple sampling. Then derive w_i for observations in each strata for the stratified sampling case. Compare sampling weights with probabilities you derived in previous section.

2 Best Linear Prediction

Suppose the following linear conditional mean

$$y = \alpha + \mathbf{x}'\gamma + u$$

Show that for best linear prediction, first order conditions are

$$E[u] = 0$$
$$Cov[\mathbf{x}, u] = 0$$

then show that FOCs imply

$$\gamma = (V[\mathbf{x}])^{-1} Cov[\mathbf{x}, y]$$
$$\alpha = E[y] - E[\mathbf{x}']\gamma$$

3 Data Description

Medicare doesn't cover all medical expenses. About half of eligible individuals therefore purchase supplementary insurance in the private market that provide insurance coverage against various out-of-pocket expense. In this exercise, we consider the impact of this supplementary insurance on total annual medical expenditures of an individual, measured in dollars.

1. List various feature of the variable used in the regression.
2. Give a summary statistics of the variables.
3. Check how many observations in income are negative?
4. Give a more detailed summary statistics of total expenditure (*totexp*). Interpret it. How it makes a problem?
5. Give a summary statistics showing the relation of gender and having insurance. Do women have more chronic problems?
6. Is the average of the medical expenditure equal for those with and without supplementary health insurance? Test it, and plot the distribution of medical expenditure for both groups in one graph.
7. Plot the histogram of medical expenditure? How do you fix the problem of highly skewed in plotting?
8. Distribute the data to ten income groups. Report the mean of health status (*hvgg*), income, years of education and total medical expenditure in each group. Plot the bar chart of health status in ten income groups.

4 Basic Regression Analysis

1. Investigate the pair vise correlation of variables.
2. Regress the total annual medical expenditure on the supplementary insurance. Is the coefficient of independent variable reliable? What are the problems?
3. Add some control variables one by one and report the results in one table. Compare the results with the first regression.
4. Create a variable for the fitted values. Report the summary statistics of this new variable and *totexp*.

5. Fit the last regression model on levels, except use all observations rather than those with just positive expenditures, and report robust standard errors. Predict medical expenditures. Use *correlate* to obtain the correlation coefficient between the actual and fitted value and show that, upon squaring, thjs equals R^2. Show that the linear model *mfx* without options reproduces the OLS coefficients. Now use *mfx* with an appropriate option to obtain the income elasticity of medical expenditures evaluated at sample means.

6. Add the variables *phylim* and *actlim* to the regression (if you don't use them in the previous step) and test the jointly significance of the coefficients of them.

4.1 Specification Analysis

1. Plot the residual of your model against the fitted value. An extreme outlier would lead to the big residual; Is there any extreme outlier? List the three observation with bigger residual.

2. Some observations may have unusual influence in determining parameter estimates and resulting model predictions. A commonly used measure is *dfits* which can be shown to equal the scaled difference between predication of with and without the i^{th} observation in the OLS regression. A rule of thumb is that observations with $|dfits| > 2\sqrt{k/N}$ may be worthy of further investigation, where k is the number of independent variables and N is the number of observations . List these observations. Are they serious problems?

4.2 Specification Tests

1. Test the functional form of the last regression model.

2. Test the heteroscedasticity. Use the Breusch-Pagan Lagrange multiplier test. Use the *estat hettest* command.