Barcelona School of Economics

Assignment 5

Foundations of Econometrics

Group 1

Corneel Jill Samuel

October 27, 2025

Question 1

Data file `medical.dta` includes data from the United States on medical expenditures for a group of people 65 years and older, who qualify for the federal health insurance program (Medicare). Medicare does not cover all medical expenses, and some citizens purchase supplementary private insurance coverage for out of pocket expenses. The goal of this exercise is to use this dataset to model the relationship between medical expenses and supplementary insurance. Regression model proposed to analyze this data is:

Model(1):

$$\text{lnmedexp} = \beta_1 + \beta_2 \text{privateins} + \beta_3 \text{phylim} + \beta_4 \text{actlim} + \beta_5 \text{totchr} + \beta_6 \text{age} + \beta_7 \text{female} + \epsilon$$

where lnmedexp = natural log of total medical expenditure, privateins = 1 if has supplementary private insurance, phylim = 1 if has functional limitation, totchr = # of chronic problems, age = age, female = 1 if person is a female. All software related questions in this exercise should be answered using Stata.

Part (a)

Estimate by OLS the model above, using default standard errors. Present the output as your answer.

Answer:

Part (b)

In as much detail as possible, provide the expression that has been used to calculate se($\hat{\beta}_2$) in the output in 1a. Substitute in the expression all the values, if any, available from the output itself.

Answer:

Part (c)

What meaning could we give to the estimate of parameter β_2? Rigorously justify.

Answer:

Part (d)

Which graph would you use to get information as to whether you are dealing with heteroskedastic disturbances? Include the graph in your answer and comment on what it seems to indicate regarding the role of the unobservables.

Answer:

Part (e)

Test for the presence of heteroskedastic disturbances using Breusch-Pagan (BP) test. In the simplest form, this test considers two steps:

Step (1): setting the following artificial regression:

$$\text{(AR)} \quad \hat{\epsilon}_i^2 = \alpha_1 + \alpha_2 \widehat{\ln\text{medexp}}_i + v_i,$$

with $\hat{\epsilon}_i$ and $\widehat{\ln\text{medexp}}_i$ being the OLS residual and fitted value for observation i, respectively.

Step (2): testing $H_0 : \alpha_2 = 0$ versus $H_1 : \alpha_2 \neq 0$ using the following test statistic:

$$n \cdot R_e^2 \underset{H_0}{\sim} \chi^2(1),$$

where n is the number of observations and R_e^2 is the coefficient of determination of the artificial regression (AR) above.

(i) Calculate the value of BP the test-statistic and associated p-value, by following these two steps.

Answer:

(ii) Verify the value of the BP test statistic and p-value you got by running command for this test: `estat hettest, iid`.

Answer:

(iii) Was the result of the test expected given your answer to 1d? Comment.

Answer:

Part (f)

Given the answer to the two previous questions, do you see the need to adjust the calculation of your standard errors. Rigorously argue.

Answer:

Part (g)

With the help of Stata estimate by OLS Model(1) again, but using heteroskedasticity robust standard errors. Present the output as your answer.

Answer:

Part (h)

Did you expect for R^2 to change when moving from OLS estimation to robust OLS? Rigorously argue.

Answer:

Part (i)

Perform a significance test of regressor privateins. Comment.

Answer:

Part (j)

(i) Estimate the model again but using bootstrapped standard errors, resampling 1000 times. Do not forget to include a seed for reproducibility. Include the output as your answer.

Answer:

(ii) Explain how the standard error has been calculated (just provide the expression, indicating how many bootstrapped samples you have used).

Answer:

(iii) Now, reproduce the calculation of the bootstrapped standard error for $\hat{\beta}_2$ by running do-file `Boot.do`. Please, notice you need to add the same seed you used in answering (i)!

Answer:

Question 2

The goal is to consistently estimate parameter β_2 in the following regression:

$$y = \beta_1 + \beta_2 x + \epsilon.$$

Part (a)

Explain why if β_2 is defined as the solution of the population least squares problem, β_2 can be consistently estimated using OLS.

Answer:

Part (b)

Now, consider that β_2 is defined as a causal parameter. Departing from the expression of the OLS:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

prove that $\hat{\beta}_2$ is not a consistent estimator of β_2 if x is an endogenous regressor (Hint: Apply plim$(\cdot)$ to both sides of the expression).

Answer:

Part (c)

Provide the expression of the instrumental variable estimator of parameter β_2, if only one instrument for x is available.

Answer:

Part (d)

Consider the instrument available is a binary instrument. That is, an instrument taking only two values: $z = 1$ and $z = 0$. Label n_1 as the number of observations where $z = 1$, and n_0 as the number of observations where $z = 0$, with $n_0 + n_1 = n$. Prove that in this case, $\hat{\beta}^{IV}$ can be written as:

$$\hat{\beta}^{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

where $\bar{y}_1$ and $\bar{x}_1$ indicate the sample average of variables y and x when $z = 1$ and, similarly, $\bar{y}_0$ and $\bar{x}_0$ are the sample averages when $z = 0$.

Hint: Use fact that $\sum_i (z_i - \bar{z})(y_i - \bar{y}) = n_1(\bar{y}_1 - \bar{y})$.

Answer:

Part (e)

Provide an intuition for the expression of β^{IV} provided in question 2d.

Answer:

Part (f)

Any examples of use of binary instruments that you recall?

Answer: