Problem Set 02
review
casual infrance

TAs: Aida Hatami, Sam Fathinejad, Fatemeh Salehi, Reza Sahour

February 21, 2025

Question1

Consider the following regression model for a random sample of two variables. Assume that the following form correctly specifies the relationship between the two variables and also shows:

$$\sum (x_i - \bar{x})^2 > 0, \tag{1.1}$$

$$y_i = \alpha + \beta x_i + e_i, \quad e_i \sim (0, \sigma_e^2), \tag{1.2}$$

The variable x is observed m times, but instead of this variable, a noisy version is available. Mathematically, this is expressed as:

$$z_h = x + u_h, \quad (h = 1, ..., m), \tag{1.3}$$

where the errors u_h are independent of x and the errors themselves are two independent variables with a mean of zero and the same variance. Additionally, they satisfy:

$$\text{cov}(e, u_h) = 0, \tag{1.4}$$

for all h.

(a) Assume that one of the z_h's is used as the observed variable, and the regression obtained using the ordinary least squares (OLS) method is estimated. Discuss the probability limit of $\hat{\beta}$ (probability limit) and whether you can provide an intuitive explanation for this result.

(b) Assume that W represents the average of the observed variables:

$$W = \frac{1}{m} \sum z_h, \tag{1.5}$$

Now, instead of the unobservable variable x, the variable W is used in the regression. Discuss how this affects the consistency of $\hat{\beta}$ and analyze whether this approach improves the consistency issue previously mentioned.

Question2

A researcher is considering two regression specifications to estimate the relationship between a variable X and a variable Y:

$$\log Y = \beta_1 + \beta_2 \log X + U, \tag{2.1}$$

$$\log\left(\frac{Y}{X}\right) = \alpha_1 + \alpha_2 \log X + V. \tag{2.2}$$

(a) Determine whether (2.2) can be expressed as a restricted version of (2.1). Rewrite (2.2) as:

$$\log Y = \alpha_1 + (\alpha_2 + 1) \log X + v. \tag{2.3}$$

(b) Using the same n observations of the variables Y and X, the researcher fits the two specifications using OLS. The fits are:

$$\hat{\log} Y = \hat{\beta}_1 + \hat{\beta}_2 \log X + U, \tag{2.4}$$

$$\log\left(\frac{\hat{Y}}{X}\right) = \hat{\alpha}_1 + \hat{\alpha}_2 \log X + V. \tag{2.5}$$

Using the expressions for the estimates, write $\hat{\beta}_2$ in terms of $\hat{\alpha}_2$.

(c) Write $\hat{\beta}_1$ in terms of $\hat{\alpha}_1$.

(d) Demonstrate that:

$$\log \hat{Y} - \log X = \log\left(\frac{\hat{Y}}{X}\right). \tag{2.6}$$

(e) Demonstrate that the residuals of (2.4) are identical to those of (2.5).

(f) Demonstrate that the standard errors of $\hat{\beta}_2$ and $\hat{\alpha}_2$ are identical.

(g) Explain with detailed arguments whether R^2 would be the same in the two regressions.

Question3

Consider two unbiased estimators W_1 and W_2 for the parameter θ in a population. Assume that the variances of these estimators are σ_1^2 and σ_2^2, respectively, and their covariance is σ_{12}.

(a) Find the conditions on the parameters a and b such that the estimator $W = aW_1 + bW_2$ remains an unbiased estimator for θ.

(b) Given the condition obtained in part (a), determine the values of a and b that minimize the variance of the estimator W.

(c) What is the optimal variance? Can this value be less than the variances of both estimators W_1 and W_2? Explain this concept.

Question4

Data file: The data sets files along with the exercise has been uploaded. Combine data sets 1 and 2, and then answer the questions based on the new data set.(In this exercise, you should use Python for the coding section.)

Variable Definitions:

- $wage_i$ = average hourly earnings of worker i in 1976, in dollars per hour.
- $educ_i$ = years of formal education completed by worker i, in years.
- $female_i$ = an indicator variable equal to 1 if worker i is female, and 0 if worker i is male.

1. Compile a table of descriptive summary statistics for the sample data, including the sample mean, standard deviation, minimum, and maximum for each variable. Determine the number of females and males in the sample.
2. Compute and present OLS estimates of the following regression equation for the full sample of 436 paid workers:

$$wage_i = \beta_0 + \beta_1 educ_i + u_i. \tag{4.1}$$

   (a) Report the OLS coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
   (b) Interpret the value of the slope coefficient estimate $\hat{\beta}_1$.
   (c) Interpret the value of the intercept coefficient estimate $\hat{\beta}_0$.
   (d) Draw the estimated sample regression function, computing and labeling points for $educ_i = 12$ and $educ_i = 16$.

Question5

Using the data in `ATTEND.RAW`, answer the following questions: (In this exercise, you should use Python for the coding section.)

(a) Provide a summary statistics report for all the variables

(b) create histogram for all the variables

(c) To determine the effects of attending lectures on final exam performance, estimate a model relating `stndfnl` (the standardized final exam score) to `atndrte` (the percent of lectures attended). Include the binary variables `frosh` and `soph` as explanatory variables. Interpret the coefficient on `atndrte`, and discuss its significance.

(d) create a plot of the error term for result.

(e) How confident are you that the OLS estimates from part (c) are estimating the causal effect of attendance? Explain.

(f) As proxy variables for student ability, add to the regression `priGPA` (prior cumulative GPA) and `ACT` (achievement test score). Now what is the effect of `atndrte`? Discuss how the effect differs from that in part (c).

(g) What happens to the significance of the dummy variables in part (f) as compared with part (c)? Explain.

(h) Add the squares of `priGPA` and `ACT` to the equation. What happens to the coefficient on `atndrte`? Are the quadratic terms jointly significant?

(i) To test for a nonlinear effect of `atndrte`, add its square to the equation from part (h). What do you conclude?