Econometrics - Problem Set #2

Ryan Safner

Fall 2017

Concepts and Critical Thinking

Please answer the following questions briefly (1-2 sentences).

1. (5 points) In your own words, describe what R^2 means. How do we calculate it, what does it tell us, and how do we interpret it?

2. (5 points) In your own words, describe what the standard error of the regression (SER) means. How do we calculate it, what does it tell us, and how do we interpret it?

3. (5 points) In your own words, describe what exogeneity and endogeneity mean, and how they are related to bias. What can we learn about the bias if we know X is endogenous?

4. (5 points) In your own words, describe what homoskedasticity and heteroskedasticity mean: both in ordinary English, and in terms of the graph of the OLS regression line.

5. (10 points) A researcher is interested in examining the impact of illegal music downloads on commercial music sales. The author collects data on commercial sales of the top 500 singles from 2017 (Y) and the number of downloads from a web site that allows 'file sharing' (X). The author estimates the following model

$$\text{music sales}_i = \beta_0 + \beta_1 \text{illegal downloads}_i + \epsilon_i$$

The author finds a large, positive, and statistically significant estimate of $\hat{\beta}_1$. The author concludes these results demonstrate that illegal downloads actually *boost* music sales. Is this an unbiased estimate of the impact of illegal music on sales? Why or why not? Do you expect the estimate to overstate or understate the true relationship between illegal downloads and sales?

6. (10 points) A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned the treatment group or into the control group. Half of the people are given the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress the following model:

$$\text{cholesterol level}_i = \beta_0 + \beta_1 \text{dosage level}_i + \epsilon_i$$

For people in the control group, $\text{dosage level}_i = 0$ and for people in the treatment group, dosage level_i measures milligrams of the active ingredient. In this case, the authors find a large, negative, statistically significant estimate of $\hat{\beta}_1$. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not? Do you expect the estimate to overstate or understate the true relationship between dosage and cholesterol level?

Problems

Please perform the requested calculations and provide interpretations where asked. Unless otherwise specified, round to 2 decimal places.

7. (10 points) Suppose a researcher, using data on class size and average test score from 100 classes, estimates the following OLS regression:

$$\widehat{\text{Test score}} = 520.4 - 5.82\text{Class size}, R^2 = 0.08, SER = 11.5$$

   (a) Interpret what $\hat{\beta}_0$ means in this context.

   (b) Interpret what $\hat{\beta}_1$ means in this context.

   (c) A class has 22 students. What is the regression's prediction for that classroom's average test score?

   (d) Last year a classroom has 19 students, and this year it has 23 students. What is the regression's prediction in the change in class average test score?

   (e) It turns out the class with 22 students had an actual average test score of 401. What is the residual for this class?

8. (15 points) A researcher wants to estimate the relationship between average weekly earnings (AWE, measured in dollars) and age (measured in years) using a simple OLS model. Using a random sample of college-educated full-time workers aged 25-65 yields the following:

$$\widehat{AWE} = 696.7 + 9.6 \times Age, R^2 = 0.023, SER = 624.1$$

   (a) Interpret what the coefficients 696.7 and 9.6 mean.

   (b) What are the units of the SER, and what does it mean? Is the SER large in the context of this regression?

   (c) The R^2 for the regression is 0.023. What are the units of the R^2, and what does it mean?

   (d) What does the regression predict will be the earnings of a 25 year-old worker? How about a 45 year-old worker?

   (e) Will the regression give reliable predictions for a 99 year-old worker? Why or why not?

   (f) What does the error term (ϵ_i) represent in this case, and why might individuals have different values of ϵ_i?

   (g) Do you think it's likely that age is exogenous? Why or why not? Would we expect $\hat{\beta}_1$ to be too large or too small?

9. (15 points) Suppose a researcher is interested in estimating the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and in a sample of 48 observations, generates the following descriptive statistics:

- $\bar{X} = 30$
- $\bar{Y} = 63$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 6900$
- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 29000$
- $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 13800$
- $\sum_{i=1}^{n}\hat{\epsilon}^2 = 1656$

(a) What is the OLS estimate of $\hat{\beta}_1$?

(b) What is the OLS estimate of $\hat{\beta}_0$?

(c) Suppose the OLS estimate of $\hat{\beta}_1$ has a standard error of 0.072. Without running a t-test, could we probably reject a null hypothesis of $H_0 : \beta_1 = 0$ at the 95% level?

(d) What is the R^2 for this model? Does this model explain a lot of variation in Y_i?

(e) How large is the average residual?

Stata Exercises

Please perform the requested computations on Stata. Write the answers to questions below in the same document as the rest of your answers, and also include a `.log` file when you turn in your problem set.

10. (20 points) Download the `MLBattend` dataset from Blackboard. This data contains data on attendance at major league baseball games for all 32 teams from the 1970s-2000.
    (a) Make a table of the summary statistics (count, mean, sd, min, max) for *home_attend* and *runs_scored*
    (b) Create a boxplot for `home_attend` over time (that is, over the `seasons`). How does attendance seem to change over time?
    (c) Create two histograms (each in percents), one for `home_attend` and one for `runs_scored`. Describe the skew of each distribution, and why this makes sense.
    (d) Create a scatterplot between `home_attend` (as dependent variable) and `runs_scored` (as independent variable).
    (e) Estimate the following model:

$$\text{Home attendance rate}_i = \beta_0 + \beta_1 \text{runs scored}_i + \epsilon_i$$

    Write the equation of the regression, placing standard errors in parentheses beneath the coefficients. Round to the nearest whole number. Assume the errors are homoskedastic. Then interpret each coefficient. Finally, can we reject the null hypothesis at the 5% level that there is no relationship between runs and attendance?
    (f) Use Stata to predict the attendance for a team that scores 500 runs in a year. Also predict the residual(s) for having 500 runs in a year. Is the residual larger or small than the average?
    (g) Plot the regression line on the scatterplot.
    (h) Make a residual plot.
    (i) Let's look at some other variables that might affect attendance, and present them in a nice table. Run separate regressions of attendance on runs allowed, wins, losses, and games behind. Present them in a nice table using `outreg2`.
    (j) Let's only look at the 2000 season. Make a scatterplot, run a regression, and then plot the regression on the scatterplot between attendance and runs scored, but only for the year 2000 (hint: for each command, set a condition with '`if`' your condition at the end)