

PLSC 503: Midterm

Thad Dunning

Department of Political Science

Yale University

Spring 2010

Due March 4, 2010 at the start of class

Please turn in two copies of your exam. If you write the exam by hand, photocopies are fine, but make sure both copies are everywhere legible (and keep the original for yourself!).

You may want to read the whole exam before beginning. Don't be alarmed by the length in pages; some of the questions should go relatively quickly. Answer each question carefully and completely. The exam is open-book and open-note but please work independently.

Exam Questions

1. Draws are being made at random with replacement from a box. The number of draws is getting larger and larger. Say whether each of the following statements is true or false, and explain. (Remember that "converges" means "gets closer and closer.")
 - (a) The probability histogram for the sum of the draws (when put in standard units) converges to the standard normal curve.
 - (b) The histogram for the numbers in the box (when put in standard units) converges to the standard normal curve.
 - (c) The histogram for the numbers drawn (when put in standard units) converges to the standard normal curve.
 - (d) The probability histogram for the product of the draws (when put in standard units) converges to the standard normal curve.
 - (e) The histogram for the numbers drawn converges to the histogram for the numbers in the box.
 - (f) The variance of the numbers drawn converges to zero.
 - (g) The variance of the histogram for the numbers drawn converges to zero.
 - (h) The variance of the average of the draws converges to zero.

2. An analyst fits a regression to a large data set. True or false, and explain:
- (a) If the usual OLS assumptions are violated, the analyst can't calculate the R^2 of the regression.
 - (b) If the usual OLS assumptions are violated, the R^2 statistic does not have any ready interpretation.
3. A large college course has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues this way:
- “If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That’s what happened to me—chance variation.”
- Is this a good defense? Answer yes or no, and explain briefly.
4. A newspaper article says that on average, college freshmen spend 7.5 hours a week going to parties. One administrator thinks that these figures do not apply at her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours. Is the difference between 6.6 and 7.5 real, or can it be easily explained by chance?
- (a) Formulate the null and alternative hypotheses in terms of a box model.
 - (b) Fill in the blanks. The null says that the average of the box is _____. The alternative says that the average of the box is _____.
 - (c) Now answer the question: is the difference real?
5. In a graduate class, 30 students take a midterm. 10 are left-handed and the other 20 are right-handed. The 10 left-handers score 83 (out of 100) on the exam on average, with an SD

of 7, while the right-handers score 89 on average, with an SD of 9. Is the difference between 89 and 83 statistically significant? Explain.

6. A geography test was given to a simple random sample of 250 high school students in a certain large school district. One question involved an outline map of Europe, with the countries identified only by number. The students were asked to pick out Great Britain and France. As it turned out, 65.8% could find France, compared to 70.2% for Great Britain. Is the difference statistically significant? Or can this be determined from the information given? Explain.
7. The great French kings of history had mediocre chief ministers, while the great ministers served under kings of lesser talent. Is this a fact of French history? Or of statistics? Explain briefly.
8. There is a study group of 10 subjects in a randomized controlled experiment. 7 of the subjects are assigned at random to treatment and 3 are assigned to the control group.

Observed data on the response variable look as follows:

Assigned to Treatment	Assigned to Control
3	—
2	—
5	—
6	—
3	—
4	—
5	—
—	2
—	4
—	3

In answering the questions below, you may use a calculator where appropriate, but discuss your work. Work out the matrices by hand (e.g., don't use statistical software).

- (a) Construct a box model for this experiment, drawing on our discussion of the Neyman model. What is in the box?
- (b) Define the intention-to-treat parameter in terms of the model you constructed in (a).
- (c) Estimate the intention-to-treat parameter, using the data in the table.
- (d) Attach a standard error to your estimate in (c). To do this, use the formula for the variance of a difference-of-means of two independent samples. (To estimate the variance of tickets in the box, you should divide by the number of tickets in the sample, minus one.)
- (e) Now, suppose an investigator assumes the OLS model:

$$Y_i = \alpha + \beta T_i + \epsilon_i, \tag{1}$$

where T_i is a 0-1 variable, with 1 indicating that a subject was assigned to treatment. Make a list of the “usual OLS assumptions”?

- (f) Under the OLS model, what is $E(Y_i|T_i = 0)$? How about $E(Y_i|T_i = 1)$?
- (g) Denote the design matrix as X . What is a typical row of this matrix? What size is X ? Denote the response variable as Y . What size is Y ?
- (h) Calculate $X'X$, $(X'X)^{-1}$, $X'Y$, and $(X'X)^{-1}X'Y$. Use $(X'X)^{-1}X'Y$ to estimate α and β .
- (i) Express $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0)$ in terms of your estimates $\hat{\alpha}$ and/or $\hat{\beta}$. How does this difference compare to your answer in (c)? Comment briefly.
- (j) Calculate the OLS residual for each subject, and calculate the sum of squared residuals. (Show and/or describe your work).
- (k) Now use the usual OLS formula to attach estimated standard errors to $\hat{\alpha}$ and $\hat{\beta}$.
- (l) Attach a standard error to the difference $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0)$ you found in (i). How does this compare to your estimated standard error in (d)?

- (m) Do you think the usual OLS assumptions are satisfied here? Why or why not? Which assumptions are the most plausible? What assumptions might be less plausible? Explain your answers carefully.
9. Suppose you assume the model $Y_i = \beta X_i + \gamma Z_i + \epsilon_i$, with the usual OLS assumptions. Here, X_i and Z_i are mean-zero scalar random variables, and ϵ_i is the disturbance term. A critic expresses the following concerns. Comment on the validity of the critic's observations.
- Omitting the intercept in the model will lead to bias, because we are forcing the regression plane to go through zero.
 - The disturbance term in the model may be correlated with the independent variables (X, Z) . Therefore, you should plot the residuals from your regression against X and Z to see if either of them show signs of being correlated with the disturbance term.
 - X and Y are highly correlated, causing problems of multicollinearity. This problem, according to the critic, means that the estimates and standard errors are consistent but biased in small samples.
10. Suppose the true model is $Y_i = \beta X_i + \gamma Z_i + \epsilon_i$, with the usual OLS assumptions. Here, X_i and Z_i are mean-zero scalar random variables, and ϵ_i is the disturbance term. We run a regression of Y on X , omitting Z . Show that:
- The fitted coefficient on X is a conditionally unbiased estimator for β if X and Z are statistically independent.
 - The fitted coefficient on X is a conditionally biased estimator for β if X and Z are not statistically independent. Derive an expression for the bias, in terms of the covariance of the random variables X and Z .

Computer exercise

11. Work Lab 9 in Freedman (2009: 303). Turn in your code for each part of the question.