

PLSC 503: Solutions to Final Exam

Thad Dunning

Department of Political Science

Yale University

Spring 2010

The exam had 100 points possible. The average score was 70 (SD of 13), and the median was 71.5. The top score was 87.5. We gave 2 “freebie” points to everybody to compensate for one poorly worded question (Question 3a). Here is the allocation of points for each question:

Question 1: 18 points overall; 1 for (a), 2 for (b), 2 for (c), 3 for (d), 3 for (e), 3 for (f), 4 for (g), 2 bonus points possible for (h).

Question 2: 18 points overall; 1 for (a), 3 for (b), 4 for (c), 3 for (d), 3 for (e), 4 for (f).

Question 3: 6 points overall; 2 for (a), 2 for (b), 2 for (c).

Question 4: 6 points overall; 3 for (a), 3 for (b).

Question 5: 5 points.

Question 6: 5 points.

Question 7: 6 points overall; 3 for (a), 3 for (b).

Question 8: 10 points overall.

Question 9: 10 points overall.

Question 10 (Computer Exercise): 16 points overall; 4 for (a), 4 for (b), 6 for (c), 2 for (d), and 4 for (e).

Exam Questions and Solutions

Note: The answers given below are in many cases just sketches; more complete answers are sometimes warranted.

Question 1: (OLS) An analyst is interested in the determinants of civil war in Africa. Let Conflict_{it} be a dummy variable equal to 1 if there is an armed conflict in country i in year t , AgGrowth_{it} be the growth rate in the agricultural sectors of the economy in country i in year t , and IndGrowth_{it} be the the growth rate in the industrial sectors of the economy in country i in year t .

The analyst assumes the following model:

$$\text{Prob}(\text{Conflict}_{it} = 1 | \text{AgGrowth}_{it}, \text{IndGrowth}_{it}, \epsilon_{it}) = \beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it} + \epsilon_{it}, \quad (1)$$

$$E(\epsilon_{it}) = 0, \quad (2)$$

$$\text{var}(\epsilon_{it}) = \sigma^2, \quad (3)$$

$$\text{AgGrowth}_{it} \perp \epsilon_{it}, \quad (4)$$

and

$$\text{IndGrowth}_{it} \perp \epsilon_{it} \quad (5)$$

for all i and all t . Here, the ϵ_{it} are i.i.d. random variables. The analyst has subtracted the mean of AgGrowth from each observation AgGrowth_{it} and the mean of IndGrowth from each observation IndGrowth_{it} , so the independent variables are mean-deviated. The analyst will estimate this model by OLS.

(a) What is the name for a linear model like (1), which has a probability on the left-hand side?

Solution: This is a linear probability model.

(b) Which terms in equation (1) are observable and which are unobservable?

Solution: The parameters β_1 and β_2 are unobservable, as is the random error term ϵ_{it} . The variables $\text{Conflict}_{it} = 1$, AgGrowth_{it} , and IndGrowth_{it} are all observable.

(c) What are the parameters of the model given by equations (1), (2), (2), (4), and (5)?

Solution: The parameters are β_1 , β_2 , and σ^2 (the variance of the error term). Naming these three parameters is sufficient for full credit on the question. Lurking in the background, there is also one additional parameter—the intercept that disappears because equation (1) is written in mean-deviated form.

(d) A critic suggests that the analyst's model specification in equations (1)–(5) is naive. In particular, she is concerned about equation (4). In a few sentences, tell a substantive story that would lead to a violation of the assumption in equation (4).

Solution: If civil conflict influences agricultural growth (reciprocal causation)—as seems likely if, say, soldiers burn crops—then we cannot have $\text{AgGrowth}_{it} \perp \epsilon_{it}$ for all i and all t . There might be omitted variables, too. For example, perhaps foreign aid to the agricultural sector boosts agricultural growth but also independently inhibits the probability of civil conflict. Then, the error term ϵ_{it} is not independent of AgGrowth_{it} .

(e) True or false, and explain: if the assumption in equation (4) is false, the OLS estimator is unbiased, but the estimated standard errors may be seriously wrong.

Solution: False. If the assumption in equation (4) is false, the OLS estimator is biased. Whether the standard errors are given by the usual OLS formulas depends on other assumptions—such as, whether the ϵ_{it} are in fact i.i.d. with $\text{var}(\epsilon_{it}) = \sigma^2$.

(f) If equation (5) is true, under what conditions, if any, is OLS a biased estimator for β_2 , the coefficient on IndGrowth_{it} ?

Solution: OLS a biased estimator for β_2 if AgGrowth_{it} and IndGrowth_{it} are correlated, and ϵ_{it} is not independent of AgGrowth_{it} .

(g) The analyst is willing to believe that equation (1) represents a response schedule; for example, if we intervene to change growth in the agricultural sector in country i at time t by 1 unit, the expected change in the probability of conflict is β_1 . But her critic says that this is silly if OLS is a biased estimator. Comment on the critic's logic.

Solution: The critic is not appropriately distinguishing the response schedule—the model—from the estimators of parameters of that model. The response schedule (1) could well be valid, and β_1 could well represent the true causal effect of intervening to change growth in the agricultural sector in country i at time t —even if the OLS estimator $\hat{\beta}_1$ is biased. So the critic is a little mixed up here.

(h) **(Bonus points)** *Homoscedasticity* implies that the variance of Y given X is the same for all X . Do you think that this holds for equation (1)? Why or why not? (Hint: the variance of a 0-1 random variable is $p(1 - p)$, where p is the mean of the random variable).

Solution: The expected value of the dependent variable—that is, the mean of the random variable—is $\beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it}$. Thus,

$$\text{var}(Y_{it} | \text{AgGrowth}_{it}, \text{IndGrowth}_{it}) = [\beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it}] [1 - (\beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it})].$$

Clearly, $\text{var}(Y_{it})$ depends on i and t —in particular, the values of IndGrowth_{it} and AgGrowth_{it} for unit i at time t . Thus the variance of Y given IndGrowth_{it} and AgGrowth_{it} cannot be the same for all values of the independent variables.

Question 2 This continues the previous question. A critic says that because assumption (4) does not hold, the analyst should use instrumental variables least squares (IVLS) regression. The critic

suggests instrumenting for agricultural growth with rainfall growth, Rain_{it} , which is the proportionate change in rainfall in country i in year t over the previous year. The analyst assumes that

$$\text{Rain}_{it} \perp \epsilon_{it}. \quad (6)$$

Let X be the design matrix in equation (1), with typical row $X_{it} = [\text{AgGrowth}_{it} \text{ IndGrowth}_{it}]$, and let Z be the matrix of instruments.

(a) What is a typical row of the matrix of instruments, Z_{it} ?

Solution: A typical row Z_{it} is given by $[\text{Rain}_{it} \text{ IndGrowth}_{it}]$.

(b) Let the IVLS estimator for equation (1) be given by

$$\hat{\beta}_{IVLS} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y, \quad (7)$$

where X and Z are as defined above and Y is a column vector of observations of Conflict_{it} . True or false, and explain: if equations (1), (2), (3), (5), and (6) hold, but equation (4) is violated, $\hat{\beta}_{IVLS}$ is an unbiased estimator for $\beta = (\beta_1 \beta_2)'$.

Solution: False. Under these assumptions, $\hat{\beta}_{IVLS}$ is a consistent estimator for $\beta = (\beta_1 \beta_2)'$, but it is not unbiased.

(c) Suppose that rainfall diminishes the probability of civil war by making it harder for the government and insurgents to deploy fighters to the field. If so, how many of the assumptions in equations (1)–(6) would be violated? (Zero through six are possible answers). If no assumptions are violated, say why; if one or more assumptions are violated, say which ones are violated and why.

Solution: Equation (1) is violated: Rain_{it} should be in the model. That is, the exclusion restriction is not satisfied. Also, if Rain_{it} is left out of the model, then equation (6) is violated, because rain is correlated with omitted determinants of conflict. Finally, we would then also have violations of (4) and (5), to the extent that rainfall is correlated with either agricultural and industrial growth.

(d) Suppose that equation (4) is true but (5) is false. What are the implications for the IVLS estimator in (7)?

Solution: The IVLS estimator is then biased and inconsistent. Indeed, if IndGrowth_{it} is endogenous, the assumption that $Z \perp \epsilon$ is violated.

(e) In your opinion, how likely is it that equation (4) is false but (5) is true, or that (4) is true while (5) is false? Defend your answer. You may want to refer to your answer to question 1(d). Discuss the implications of your answer for IVLS estimation more generally.

Solution: Here, you are intended to make your own reasoned argument. But for many of the arguments made in 1(d), it is generally hard to see why AgGrowth_{it} would be exogenous and IndGrowth_{it} endogenous, or vice versa. For instance, omitted political institutions that shape agricultural growth and the probability of conflict would probably shape industrial growth as well. If AgGrowth_{it} and IndGrowth_{it} are both endogenous, we'll need two instrumental variables. The lesson for IVLS estimation more generally is that it is not enough to instrument for a single endogenous right-hand side variable, in a multiple regression set-up: the other right-hand side variables will need instruments as well, unless we can make strong arguments for their exogeneity.

(f) Which of the assumptions in equations (1)–(6) are testable? How could you (at least partially) test those assumptions?

Solution: Equation (1), the specification of the model, is not fully testable. Neither are equations (2). Assumptions like equations (4), (5), and (6) could be probed though not proved; for example,

if rainfall growth is as-if randomly assigned, we should not be able to reject the null hypothesis of the independence of rainfall growth and various pre-existing covariates. If the assumptions in equations (1), (2), (4), (5), and (6) are correct, we can partially test the assumption in equation (3), for example by plotting residuals against X-values and looking for evidence of heteroskedasticity (or conducting more formal tests for heteroskedasticity).

Question 3: Suppose a critic tells an analyst that the effects of industrial growth on the probability of conflict in Africa are conditional on the growth rate of the agricultural sector of the economy. In particular, industrial decline only increases the probability of conflict if agriculture is in decline. The analyst posits the following model:

$$\begin{aligned} \text{Prob}(\text{Conflict}_{it} = 1 | \text{AgGrowth}_{it}, \text{IndGrowth}_{it}, \epsilon_{it}) &= \beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it} \\ &+ \beta_3 (\text{AgGrowth}_{it} \times \text{IndGrowth}_{it}) + \epsilon_{it}, \end{aligned}$$

where the variables are defined as above. True or false, and explain:

(a) If the critic's hypothesis is right, β_3 is negative, while β_2 is positive.

Solution: False. According to the problem, industrial decline ($\text{IndGrowth}_{it} < 0$) only increases the probability of conflict if agriculture is in decline—i.e., $\text{AgGrowth}_{it} < 0$; similarly, industrial growth must decrease the probability of conflict only if agriculture is growing. Otherwise, $\text{IndGrowth}_{it} < 0$ has no effect on the probability of conflict. Thus, we have $\beta_3 < 0$ and $\beta_2 = 0$.

NOTE: We decided this question was poorly worded, so while we graded it, we also gave two “freebie” points to all exams.

(b) The analyst can test that $\beta_2 = \beta_3 = 0$ by conducting an F -test.

Solution: This is true, as long as ϵ_{it} is normally distributed or if the number of observations is relatively large. This is the standard setting in which the F -test would be used.

(c) If the analyst rejects the null hypothesis that $\beta_3 = 0$ but fails to reject the null that $\beta_1 = \beta_2 = 0$, then she should assume equation (1).

Solution: False. The model is not validated by a hypothesis test about the size of a particular coefficient, so such a finding should not have much bearing on the specification of the model. But in any case, if the analyst rejects the null hypothesis that $\beta_3 = 0$, this if anything strengthens the justification for including the interaction term in the model.

Question 4: An analyst regresses a variable Y on an $n \times p$ matrix X that has rank $p < n$. The analyst wants to test the null hypothesis that the coefficients of the last p_0 columns of X are all equal to zero, so she also regresses Y on the first $p - p_0$ columns of X , where $1 < p_0 < p$. Let e be the vector of residuals from the first (full) regression and e^s be the vector of residuals from the second (smaller) regression. The analyst forms the F-statistic as

$$F = \frac{(\|e^s\|^2 - \|e\|^2)/\|e\|^2}{p_0/(n - p)}. \quad (8)$$

True or false, and explain your answer:

(a) Equation (8) involves only the residuals from two regressions. Thus, no statistical model is needed to fit these two regressions to any data set and use the F-statistic to test the hypothesis that some of the coefficients are jointly zero.

Solution: False. First, the coefficients are parameters of the model (if there is no model, there are no coefficients). Second, the model specifies a random error term, so that the data are observed values of random variables. To use the F-distribution, the error term must be normally distributed (or else the number of observations must be large, so that we can appeal to a central limit theorem).

(b) However, if the assumptions of the classical linear model are violated, we should not use the F-statistic unless n is large.

Solution: True, where by “assumptions of the *classical* linear model” we mean, the assumption that the error term in the model is normally distributed. We need normality of the error term or large n to use the usual F-test.

Of course, if other assumptions of the linear model—such as the assumption that data are generated according to a linear function of parameters, or that there is a random error term—are violated, then the statement is false: the usual F-test doesn’t apply, no matter how large is the n . Partial credit was also given for answers that took this latter view.

Question 5: An analyst uses data from an alleged natural experiment, in which a large number of subjects are as-if randomly assigned to treatment and control conditions. She gathers data on 20 “pre-treatment covariates” such as age, sex, and so on. She finds that the treatment group is substantially older than the control group, and the difference is highly significant. She notes that “as-if random assignment should balance the treatment and control groups on pre-treatment covariates. Moreover, there are a large number of subjects, lending substantial statistical power. Thus, treatment assignment must not have been as-if random.” Comment on the analyst’s reasoning.

Solution: Here, we’ve got 20 pre-treatment covariates. Suppose subjects are randomly assigned to treatment and control, and consider a single pre-treatment covariate. In expectation, in 5 out of 100 assignments, the differences between the treatment and control group on this covariate will be large enough that we would reject at the 0.05 level the null hypothesis that the groups are sampled at random from the same population. What we’ve got here is a “significant” difference for 5/100 covariates—that is, 1 out of 20—which is what we will see in expectation even if the null hypothesis is true. So this doesn’t provide strong evidence against as-if random assignment.

Question 6: Another team of political scientists conduct an experiment in which a relatively small number of subjects are randomly assigned to treatment and control conditions. They find a

statistically significant difference between mean outcomes in the treatment and control groups. These scholars write that “covariates are balanced across the treatment and control groups in expectation but not necessarily in their realization. To check that our results are not driven by omitted variables, we focus attention on two potential confounds.” Can the difference in mean outcomes across the treatment and control groups be due to confounding or omitted variable bias? Why or why not?

Solution: Randomization, if it is conducted correctly, ensures that covariates are balanced in expectation—just as the analysts write. This means that treatment assignment is independent of potential confounders. The analysts are confusing chance error with bias—two very different concepts.

Question 7: A social scientist is interested in the effects of religion on political participation in Africa. She analyzes data from surveys taken in all 48 sub-Saharan African countries. Controlling for many covariates in a regression model, she finds no relationship on average between strength of individual religious beliefs and political participation. However, suspecting that this relationship might differ in different countries, she runs separate regressions for each country. She writes “interestingly, we found a statistically-significant and positive relationship between individual attachment to religion and political participation in Nigeria, Sudan, and the Ivory Coast – all countries in which religion has been substantially politicized in recent years.”

(a) Does the evidence suggest that the effects of religious beliefs may be conditional on the country-level politicization of religion? What is a different interpretation of the evidence?

Solution: There are many different issues to which you could point—for instance, what is the evidence for a causal “effect” here? But the main point to make is that there are 48 countries, and “significant” relationships in 3 of them—which is well within the realm of chance error. This is similar to the logic of Question 5.

(b) How can we formulate the causal hypotheses under investigation in terms of a response schedule?

Solution: The key idea is that there are structural parameters that are invariant to intervention. Really, there are two response schedules involved. One would formalize the idea that if we intervene to strengthen individual religious beliefs, we will see a change of some amount (possibly zero) in political participation. (The data are measured at the aggregate level—which raises issues of *ecological inference*—but the hypothesis appears to be about individual-level beliefs and participation. The second response schedule would specify how an intervention to politicize religion (at the country level) shapes the relationship between beliefs and participation. Given the assumptions of the response schedules, regression might allow us to estimate the size of, e.g., the change in political participation that is due to an intervention to strengthen religious beliefs.

Question 8: In your view, which of these statements is closer to the truth? (i) Regression analysis can demonstrate causation; (ii) Regression analysis assumes causation but can be used to estimate the size of a causal effect—if the assumptions of the regression models are correct.

Pick one of these two statements and defend your choice in detail. Does your answer change, depending on whether we are analyzing experimental or observational data?

Solution: This is for you to argue, and our evaluation is based on how well you argue your position. In my view, (ii) is closer to the truth. For making causal inference from regression models, the idea of a response schedule plays a key role: the response schedule is a theory of how the data were generated, and it tells us how one variable would respond if we intervened to manipulate other variables. For instance, in multiple regression, the response schedule tells us the functional form of the relationship between the key independent variable (treatment), other measured and unmeasured independent variables, and the treatment. Given the model—the response schedule—we can use regression to estimate the size of the parameters of that model.

This doesn't really change, whether we are analyzing experimental or observational data. To make causal inferences, we need a response schedule that links treatments to outcomes. In many (but not all) experimental analyses, the response schedule is Neyman's potential outcomes model.

Question 9: What is the relationship, if any, between causal inference and statistical hypothesis testing?

Solution: Statistical inference depends on a statistical model, according to which data are observed values of random variables. Under the assumptions of the statistical model, we can conduct tests of null hypotheses such as, the value of a single parameter in the model is zero. Yet the assumed response schedule is what justifies attaching a causal interpretation to the parameters of the model, because it tells us the counterfactual response of the dependent variable to hypothetical manipulations of the independent variable.

Computer Exercise

Solutions to the computer exercises will be posted online.