

# PLSC 503: Final Exam

Thad Dunning

Department of Political Science

Yale University

Spring 2010

Due Friday, April 23, 2010 at 10:00 AM

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. Remember to read and answer each question carefully and completely. The exam is open-book and open-note but please work independently.

### Exam Questions

1. An analyst is interested in the determinants of civil war in Africa. Let  $\text{Conflict}_{it}$  be a dummy variable equal to 1 if there is an armed conflict in country  $i$  in year  $t$  and 0 otherwise;  $\text{AgGrowth}_{it}$  be the growth rate in the agricultural sectors of the economy in country  $i$  in year  $t$ ; and  $\text{IndGrowth}_{it}$  be the growth rate in the industrial sectors of the economy in country  $i$  in year  $t$ .

The analyst assumes the following model:

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

where

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

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

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

and

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

for all  $i$  and all  $t$ . Here, the  $\epsilon_{it}$  are i.i.d. random variables. The analyst has subtracted the mean of  $\text{AgGrowth}$  from each observation  $\text{AgGrowth}_{it}$  and the mean of  $\text{IndGrowth}$  from

each observation  $\text{IndGrowth}_{it}$ , so the independent variables are mean-deviated. The analyst estimates the model by OLS.

- (a) What is the name for a linear model like (1), which has a probability on the left-hand side?
  - (b) Which terms in equation (1) are observable and which are unobservable?
  - (c) What are the parameters of the model in equations (1)–(5)?
  - (d) A critic suggests that the analyst’s model specification is naive. In particular, she is concerned about the assumption in equation (4). In a few sentences, tell a substantive story that would produce a violation of the assumption in equation (4).
  - (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.
  - (f) If equation (5) is true, under what conditions, if any, is OLS a biased estimator for  $\beta_2$ , the coefficient on  $\text{IndGrowth}_{it}$ ?
  - (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  $\beta_1$ . But her critic says that this is silly if OLS is a biased estimator. Comment on the critic’s logic.
  - (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).
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,  $\text{Rain}_{it}$ , which is

the proportionate change in rainfall in country  $i$  in year  $t$  over the previous year. The analyst assumes that

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

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,  $Z_{it}$ ?
- (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  $\text{Conflict}_{it}$ . 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  $\beta = (\beta_1 \ \beta_2)'$ .

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

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

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,  $\beta_3$  is negative, while  $\beta_2$  is positive.
- (b) The analyst can test that  $\beta_2 = \beta_3 = 0$  by conducting an  $F$ -test.
- (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).
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.
- (b) However, if the assumptions of the classical linear model are violated, we should not use the  $F$ -statistic unless  $n$  is large.
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.
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?
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?
- (b) How can we formulate the causal hypotheses under investigation in terms of a response schedule?

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?

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

#### 10. Computer Exercise

- (a) Simulate a single data with the following characteristics: (i)  $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$  for all  $i = 1, \dots, 50$ , where  $\beta_0 = 0.4$ ,  $\beta_1 = 0.3$ ,  $\beta_2 = 0.2$ ,  $X_1$  is distributed  $N(2, 1)$ , and  $X_2$  is distributed  $N(E[X_1 + Z], \text{var}(X_1) + \text{var}(Z))$ . Here,  $Z$  is a standard normal variable.  
(ii) The error term  $\epsilon_i$  is normally distributed with  $E(\epsilon_i) = 0$  and  $\text{var}(\epsilon_i|X_i) = \sigma^2 X_1^2$ .
- (b) For this single data set,

- (i) estimate the parameters  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  using OLS.
  - (ii) estimate the parameters using fGLS.
- (c) Generate 1000 data sets as outlined in (a) and save the OLS and fGLS estimates for each of these 1000 replicates. Use your simulated data sets to create a histogram showing the empirical distribution of the OLS and fGLS estimators. Turn in all of your code and output.
- (d) Is the variance of the fGLS estimator of the parameters  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  smaller than the variance of the OLS estimator?
- (e) Comment on your answer in (d).