Empirical Analysis and Research Methodology Examination
Yale University
Department of Political Science
August 2009

This exam consists of three parts. Provide answers to ALL THREE sections.

Your answers should be succinct and to the point.

Use algebra to back up your assertions.

Do not answer questions that have not been asked.

Do not leave sub-parts of questions unanswered.

You have eight hours to complete the exam. You may use a calculator and one 8.5" x 11" handwritten (not photocopied) sheet of notes.

Part I.

Professor Smedley is interested in whether developing countries serve their populations better under democratic rule as opposed to autocratic or oligarchic rule. Smedley looks at infant mortality rates in 2003 for a set of 68 countries that meet the following criteria: (i) they had less than $1000 of GDP per capita in 1960 and (ii) experienced at least one change in regime such that they went from democracy to non-democracy or vice versa between 1975 and 2000.

Smedley is considering regressing 2003 infant mortality for each country on its 1975 infant mortality rate and a dummy variable scored 1 if the country went from dictatorship to democracy (0 otherwise).

Smedley comes to you for suggestions about statistical analysis. What do you think of Smedley's regression model? What alterations would you recommend? What are the important threats to unbiased inference? Given practical limitations, what alternative research design and/or statistical analysis would you suggest to Smedley?

Part II. Read the essay that is attached to your exam.

Terry M. Moe
Collective Bargaining and The Performance of the Public Schools
American Journal of Political Science
Volume 53, Issue 1, Date: January 2009, Pages: 156-174

We do not expect you to have special expertise in the topic area, but we do expect you to bring to bear your general analytic skills as a political scientist.

Offer a critical evaluation of its methodology. Are the estimates and standard errors unbiased? Why or why not? Suggest ways in which this line of research might be improved.

III. Statistical Reasoning

Provide short answers to the following questions. Where possible, back up your answers with algebra.

A. Define the term "maximum likelihood estimator." When is ordinary least squares a maximum likelihood estimator?

B. Let Y_i^T denote the outcome if observation *i* is treated, and let Y_i^C denote the outcome for the same observation if it is not treated. The causal effect for observation *i* is the difference between Y_i^T and Y_i^C. This causal effect may vary from one observation to the next. The random assignment of subjects to treatment ($X_i = 1$) and control ($X_i = 0$) is the only random component in the modeling framework.

(i) Show that the potential outcomes model

$$Y_i = Y_i^C(1 - X_i) + Y_i^T X_i$$

may be expressed in the form of a regression model such that *b* represents the average causal effect of the treatment:

$$Y_i = a + bX_i + u_i,$$

where $a = \overline{Y}^C$ and $u_i = Y_i^C - \overline{Y}^C + \left(\left(Y_i^T - \overline{Y}^T\right) - \left(Y_i^C - \overline{Y}^C\right)\right)X_i$.

(ii) Given this disturbance term, is it possible for the disturbance term to be statistically independent of the independent variable?

(iii) Given this disturbance term, is it possible for the disturbance term to be homoskedastic?

(iv) In light of your answers to (ii) and (iii), what are the statistical implications of applying regression to the model in (i) in terms of unbiased estimates and standard errors?

C. Suppose that one seeks to estimate the parameter β in the time-series equation

$$Y_t = \alpha + \beta Y_{t-1} + u_t,$$

where the u_t are independent and identically distributed (IID). Does this regression generate unbiased estimates of β? Explain why or why not in terms of Gauss-Markov assumptions.

D. Scholars sometimes express concern about publication bias, contending that the sampling distribution of published articles is not centered at the true parameter being estimated. Propose one or two ways of detecting publication bias in research literatures.

E. What are "robust cluster" standard errors? Under what conditions is it appropriate to use them instead of conventional standard errors?

F. Sometimes scholars seek to examine whether X's effect on Y is mediated through some variable M. Is a regression of Y on both X and M helpful here? What about a regression of Y on X or a regression of Y on M? Back up your answers with algebra.

G. What is a "local average treatment effect"? How does it differ from the average effect of the treatment on the treated?

H. Professor Smedley is interested in the causal effect of rural economic conditions and ethnic conflict in Africa. Smedley gathers annual data on 7 African countries over a 30 year period. Smedley believes that rainfall can be used as an instrumental variable for economic conditions, on the grounds that variation in weather patterns is nearly random. Using OLS, Smedley shows that rainfall is a significant predictor of rural economic conditions. Smedley also shows, using OLS, that when ethnic conflict is regressed on both economic conditions and rainfall, rainfall's estimated effect is zero. Smedley therefore concludes that rainfall is both a theoretically and empirically justified instrumental variable. Is this reasoning persuasive?

I. Suppose that a researcher is weighing two alternative regressions to estimate the effect of X on Y: a regression of Y on X and a regression of Y on X and Z. It turns out that X and Z have a correlation of exactly zero. The only thing the researcher cares about is the effect of X on Y. In terms of efficiency, does it matter which regression the researcher runs?

J. Suppose you were interested in the effect of electing Republican governors on the rate of growth in state government spending. You gather data on elections held since 1970 in which the governor was elected by fewer than 1000 votes. Suppose there are 50 such elections. You gather data on spending growth following each of these elections. Explain the steps by which you would conduct a regression discontinuity analysis using these data. In particular, describe your regression model and any analyses you would use to check the model's validity or robustness.