Empirical Analysis and Research Methodology Examination
Yale University
Department of Political Science
January 2008

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

Your answers should be succinct and to the point. Do not answer questions that have not been asked.

Do not leave sub-parts of questions unanswered.

You have 6 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 the effects the digital divide (i.e., the fact that some people have computer and internet access, while others do not) on political inequality. Smedley's hypothesis is that access to computer resources exacerbates problems of political inequality, because it gives those who can afford computers easier access to public officials and information necessary to learn about public hearings, legislative votes, and other aspects of politics that potentially lead to public involvement. Smedley argues that computers widen the already large political participation gap between rich and poor.

In order to test this claim, Smedley makes use of a large and well-executed national survey of American adults. The survey has a very high response rate and asks an extensive array of questions about political participation, computer use, and respondents' background attributes. More than 20,000 interviews were conducted so as to measure accurately even rare activities, such as contacting elected officials. Using regression analysis, Smedley finds statistically significant effects of computer use on various forms of participation, even controlling for background attributes such as education, income, interest in politics, parental interest in politics, and past involvement in campaigns. A two-stage least squares analysis shows even larger effects of computer use on participation; in this analysis, the same control variables are used, and the excluded instrumental variable is the respondents' interest in technology while growing up.

Smedley comes to you for suggestions about statistical analysis, wondering whether a regression analysis would to be informative. Is regression helpful here? If so, what regression model 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.

The Power of TV: Cable Television and Women's Status in India

Robert Jensen, Emily Oster

NBER Working Paper No. 13305
Issued in August 2007

(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 serial correlation and explain its implications for regression analysis.

B. Consider the model: $Y_i = a + bX_i + u_i$. What are the practical implications, if any, of the u_i being distributed non-normally?

C. Suppose that one seeks to estimate the parameter b in the equation $Y_i = a + bX_i + u_i$, where X_i are distributed normally with mean 0 and variance of 1. However, one observes only values of $X_i > 0$. Missing values of X_i (when $X_i < 0$) are dropped from the analysis. Does this cause biased estimates of b? Why or why not?

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 "weak instrumental variables"? What consequences do weak instruments have for instrumental variables estimation?

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?

G. Sometimes it is suggested that researchers devote special attention to large positive and negative residuals. The claim is that by reflecting on the possible reasons why some observations are severely over- or underpredicted, the researcher may discover new variables that will better predict the dependent variable. Comment on the soundness of this recommendation.

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 unbiased regressions to estimate the effect of X on Y: a regression of Y on X and a regression of Y on X and Z. Under what conditions is the second regression superior when judged by the criterion of mean-squared error? Use algebra to back up your answer.

J. Recent years have seen a surge of the use of “matching” to estimate causal effects. What is matching and how is it used? Under what conditions does it provide estimates of causal effects that are more reliable than those generated by regression?