EMPIRICAL ANALYSIS AND RESEARCH METHODOLOGY EXAMINATION
Yale University

Department of Political Science

August 2012

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, and remember to show your work wherever appropriate.

Do not answer questions that have not been asked.

Do not leave sub-parts of questions unanswered.

You have seven 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 of unionization on the political attitudes and behaviors of workers. He hypothesizes that working in a unionized firm will lead workers to embrace left-wing ideology and to vote for left-leaning parties. However, he also believes that unionization will decrease support for unemployment subsidies, that is, direct payments by the government to laid-off workers.

To test these hypotheses, Smedley makes use of a large and well-executed cross-national survey of workers in both unionized and non-unionized firms. The survey has a very high response rate and asks an extensive array of questions about ideological placement on left-right scales, support for left-wing parties, attitudes towards unemployment subsidies, and respondents' background attributes. More than 30,000 interviews were conducted.

Smedley is considering the merits of two alternative research designs and comes to you asking for advice:

1. Using OLS regression analysis, Smedley finds statistically significant positive effects of unionization on support for left-wing parties and self-placement to the left on an ideology scale and negative and significant effects on support for unemployment subsidies, even controlling for background attributes such as education, income, interest in politics, parents' interest in politics, and sector of the respondent's firm. A two-stage least squares analysis shows even larger effects of unionization; in this analysis, the same control variables are used, and the excluded instrumental variable is a dichotomous variable for whether the respondent's father belonged to a union.

2. Smedley knows that in a subset of the countries in his sample, firms become unionized when a majority of workers vote for unionization. For respondents in these countries, he therefore gathers data on the most recent vote in each respondent's firm. He then drops respondents from firms in which the margin of approval or rejection of unionization in the previous vote was greater or lesser than 2.5%. For this sub-sample, he then conducts a regression of each dependent variable (left-wing ideology, support for left-wing parties, and support for government subsidies) on a constant, an indicator variable for unionization, and the host of control variables mentioned above. Here, he finds no significant effect of unionization on any of the three dependent variables.

What are the strengths and weaknesses of these two designs? What are the important threats to unbiased inference in each case, and are there any empirical analyses that Smedley could perform to test for the presence of these threats? What might account for the different results Smedley finds using the two designs? Finally, given practical limitations, what alternative research design and/or statistical analyses would you suggest to Smedley?

PART II. Read the essay attached to your exam and cited below. Offer a critical evaluation of its methodology and research design. Justify each of your claims and suggest ways in which this line of research might be improved. (We do not expect you to have special expertise in the topic area, but we do expect you to bring to bear your general analytical skills as a political scientist. Also, the second section with the brief formal model is not critical to evaluate and may be skimmed).

In answering this question, you should consider the following topics in particular:

1. What are the primary threats to internal validity with this research design? What threats are controlled by the design?

2. List the key assumptions involved in equation (8) on p. 607. How plausible are those assumptions?

3. The measured average real price of land per hectare fell by $106,000 in Chile's Central Urban and North Central Valley provinces after the reform of 1958—from $266,000 to $160,000 (Table 1). However, reporting results from Table 3, the authors note that "farm prices should have fallen by about $70,000 per hectare in [these] provinces" (p. 611). Which number provides the best estimator of the causal effect of the reform—or would a different number be better? Explain your answer.

Baland, Jean-Marie, and James A. Robinson. 2012. "The Political Value of Land: Political Reform and Land Prices in Chile." *American Journal of Political Science* 56 (3): 601-19.

PART III. Statistical Reasoning

1. 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.

2. Let X be an $n \times 2$ matrix, where the first column is all 1's and the second column is $(x_1, x_2, ..., x_{n-1}, x_n)'$. Let Y be an $n \times 1$ column vector consisting of $(y_1, y_2, ..., y_{n-1}, y_n)'$.

   (a) Find $(X'X)^{-1}X'Y$. (That is, write out the elements of the matrix $(X'X)^{-1}X'Y$, using notation such as n, $\sum_{i=1}^{n}(x_i)^2$, and so forth). Don't forget to show your work!

   (b) Show that the $(2,1)$ element of $(X'X)^{-1}X'Y = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$.

   (c) Show that the $(1,1)$ element of $(X'X)^{-1}X'Y = \bar{y} - b\bar{x}$, where $b = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$.

   N.B.: It may be helpful to remember the following alternate definitions of covariance and variance:

$$\begin{aligned} \text{Cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} y_i) \\ &= \overline{xy} - (\bar{x})(\bar{y}) \end{aligned} \tag{1}$$

   and

$$\begin{aligned} \text{Var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} x_i)] \\ &= \overline{x^2} - (\bar{x})^2. \end{aligned} \tag{2}$$

3. Political scientists use statistical models to provide quantitative characterizations of different electoral systems. Electoral systems take vote shares for a party in election t , $V_t \in [0,1]$ and generate seat shares, $S_t[0,1]$. In particular, interest centers on two features of an electoral system:

   - *unbiasedness* or the property that $E(S_t|V_t = .5) = .5$
   - *responsiveness*: how much does S_t change in result to change in V_t?

   The following statistical model is often used to estimate these features of electoral systems:

$$ln\frac{S_t}{1-S_t} = \alpha + \beta ln\frac{V_t}{1-V_t} + \epsilon_t$$

   (a) Interpret α.

   (b) In two-party systems it was conjectured that "if the votes are divided in the ratio $A : B$ then the seats will be divided in the ratio $A^3 : B^3$, and the winning majority of the votes will be magnified in the proportion of seats won" (M.G. Kendall and A. Stuart, The Law of Cubic Proportion in Election Results, *British Journal of Sociology* 1 (1950), 183–97). An earlier statement of this so-called cube-law was made in 1898 by one of the founders of modern statistics, F. Y. Edgeworth. The cube law thus implies $\alpha = 0$ and $\beta = 3$. How would you test the cube law given data on S and V ?

4. Consider the bivariate regression model with no constant term given by $y_i = \beta x_i + \epsilon_i$. Here, y_i, x_i, and ϵ_i are all random variables. Define the mean ratio estimator as: $\widehat{\beta} = \frac{\overline{y}}{\overline{x}}$, where $\overline{y}$ denotes the sample mean of y and $\overline{x}$ denotes the sample mean of x. Assume that all the Gauss Markov assumptions hold. Is the mean ratio estimator unbiased? Back up your answer with algebra.

5. (a) Suppose that data are generated according to the following regression equation: $Y_i = \alpha + \beta X_i + \epsilon_i$ for each unit i. However, a researcher measures $X_i^* = X_i + \nu_i$. Here, ϵ_i and ν_i are both i.i.d. random variables and are independent of X_i. Now, suppose the researcher regresses Y_i on a constant and X_i^*. What are the consequences for the unbiasedness and variance of the OLS estimator of β?
(b) Does your answer change if the true data-generating process is $Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i$ and the researcher regresses Y_i on a constant, X_i^*, and Z_i?

6. In 2000 Arizona implemented a campaign finance system that provided participating candidates with matching public funds if they agree to certain fund raising limits. A researcher who believes that this should lead incumbents to spend more time on constituency service (instead of fund raising), collected data on how much time incumbents in Arizona and surrounding states in the election year before the law was in effect and the election year that the law was in effect. The researcher uses this data to estimate the regression given in Table 1. A colleague asks what result he would have gotten had he used a difference-in-difference estimator (i.e., looking at the difference in Arizona before and after compared to neighboring states before and after). Provide some formal justification for your answer.

Table 1: Determinants of Weekly Hours Devoted to Constituency Service.

Variable	DV=Weekly Hours
Arizona	-1.0
	(2.0)
post-Law period	-1.0
	(3.0)
Arizona*post-Law period	5.0
	(2.0)
Constant	12.0
	(4.0)

7. A researcher is interested in finding the effect of X on Y and plans to estimate the model $Y_i = \alpha + \beta X_i + \epsilon_i$. She is concerned that X and ϵ_i may not be independent. He thinks he has an instrument Z but he is not sure if Z is independent of ϵ_i. Therefore, he proposes the following specification test: regress Y on X and Z, perform a t-test to determine whether Z significantly predicts Y, and use IV regression only if the t-statistic proves to be insignificant. Evaluate this procedure.

8. 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?

9. Let Y_i^T denote the potential outcome if observation i is treated, and let Y_i^C denote the potential outcome for the same observation if it is not treated. The unit causal effect 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.

   (a) Show that the potential outcomes model

   $$Y_i = Y_i^C(1 - X_i) + Y_i^T X_i \tag{3}$$

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

   $$Y_i = a + bX_i + u_i, \tag{4}$$

where the disturbance term $u_i = Y_i^C - \bar{Y}^C + ((Y_i^T - \bar{Y}^T) - (Y_i^C - \bar{Y}^C))X_i$.

(b) Is it possible for the disturbance term u_i to be statistically independent of the independent variable? Explain your answer.

(c) Is it possible for the disturbance term u_i to be homoskedastic? Explain your answer.

(d) In light of your answers to (b) and (c), is the OLS estimator of the model in equation (4) unbiased? What are the implications for the nominal OLS standard errors (e.g., those reported by statistical software after a standard OLS regression)?

10. In your view, which of these statements is closer to the truth?

(a) Regression analysis can demonstrate causation;

(b) 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?