EMPIRICAL ANALYSIS AND RESEARCH METHODOLOGY EXAMINATION
Yale University

Department of Political Science

January 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.

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 effect of turnout on budgetary transfers to municipalities in Japan. He argues that turnout might reflect past budgetary transfers, which also influence current budgetary transfers; or that omitted variables might influence both turnout and transfers. Thus, he suggests that a regression of current transfers on turnout would lead to biased inferences about the impact of turnout. He therefore considers two alternate research designs for studying this question:

1. Smedley proposes to use election-day rainfall as an instrumental variable, in a regression of budgetary transfers on turnout. (Here, municipalities are the units of analysis.)

   Comment on this approach. What are some potential concerns? What tools could Smedley use to assess the validity of the approach? What are its overall strengths and limitations?

2. Smedley knows that campaign consultants have some tools at their disposal—such as in-person get-out-the-vote contacting—that influence turnout. He therefore proposes a randomized controlled experiment, in which some municipalities will be selected at random for get-out-the-vote efforts and subsequent budgetary transfers to municipalities will be studied.

   How should the data from this experiment be analyzed? What are the strengths and limitations of different ways of analyzing the experimental data? What are the potential costs and benefits of this second research design, relative to the first?

PART II. Read the essay attached to your exam. Offer a critical evaluation of its methodology. 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).

Besley, Timothy, and Marta Reynal-Querol. 2011. “Do Democracies Select More Educated Leaders?” *American Political Science Review.* 105 (3): 552-566.

PART III. Statistical Reasoning

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

2. Suppose that one seeks to estimate the parameter β in the equation $Y_i = \alpha + \beta X_i + u_i$. However, one observes only values of $Y_i > 0$. Missing values of Y (when $Y_i < 0$) are dropped from the analysis. Does this cause biased estimates of β? Why or why not?

3. Let Y be an $n \times 1$ vector, X be an $n \times p$ matrix with full rank, and define $e = Y - X(X'X)^{-1}X'Y$, an $n \times 1$ vector with typical element e_i indexed by $i = 1, ..., n$. Denote the average of the elements of e by $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$. Show that e is orthogonal to X. Under what condition does $\bar{e} = 0$?

4. An analyst assumes that $Y_i = a + bX_i + cZ_i + dX_iZ_i + \epsilon_i$ for all $i = 1, ..., n$. Here, X_iZ_i is the product of X_i and Z_i, and ϵ_i is a random variable.

   (a) Suppose this researcher runs a regression of Y_i on a constant, X_i, Z_i, and X_iZ_i, obtaining the fitted values
   $$\hat{Y}_i = \hat{a} + \hat{b}X_i + \hat{c}Z_i + \hat{d}X_iZ_i \tag{1}$$
   for each i. What is the estimated effect of X_i, given Z_i? What assumptions are needed for this quantity to be a valid estimator of the marginal effect of intervening to change X_i with Z_i held fixed?

   (b) Now, express the variance of this estimated marginal effect in terms of the coefficient estimators in equation (1). (For convenience, treat X_i and Z_i as fixed, rather than as random variables).

5. Let $y_1, ..., y_n$ be a list of real numbers with a positive and finite standard deviation. For each value, subtract the mean of the list and divide by the standard deviation to create a new list of numbers $y_1^*, ..., y_n^*$. Prove that the mean of the new list is zero and its standard deviation is 1.

6. A researcher assumes the following regression model:
   $$Y = X\beta + \epsilon, \tag{2}$$
   where Y is an $n \times 1$ vector of observable random variables. Here, X is a fixed $n \times p$ matrix with a vector of 1's as the first column, and ϵ is mean-zero vector of i.i.d. random variables with $\text{var}(\epsilon_i) = \sigma^2$. The analyst fits this model to a data set with n observations. The OLS estimator is $\hat{\beta} = (X'X)^{-1}X'Y$. The residuals from the OLS fit are $e = Y - X\hat{\beta}$.

   Now suppose the analyst bootstraps the regression model. In particular, for the kth bootstrap replicate, she uses a computer to sample at random with replacement from the vector e to produce an $n \times 1$ vector of bootstrap errors, $\epsilon_{(k)} = \epsilon_{(k)1}, ..., \epsilon_{(k)n}{}'$. For each bootstrap replicate, she then constructs $Y_{(k)} = X\hat{\beta} + \epsilon_{(k)}$ and fits the OLS estimator, $\hat{\beta}_{(k)} = (X'X)^{-1}X'Y_{(k)}$. There are 100 bootstrap replicates. Finally, let

$$\begin{aligned}
\hat{\epsilon}_{(k)} &= Y_{(k)} - X\hat{\beta}_{(k)} \\
s_k^2 &= \frac{\hat{\epsilon}'_{(k)}\hat{\epsilon}_{(k)}}{n-p} \\
\hat{\beta}_{\text{ave}} &= \frac{1}{100}\sum_{k=1}^{100}\hat{\beta}_{(k)} \\
V &= \frac{1}{100}\sum_{k=1}^{100}[\hat{\beta}_{(k)} - \hat{\beta}_{\text{ave}}][\hat{\beta}_{(k)} - \hat{\beta}_{\text{ave}}]'.
\end{aligned}$$

Say whether the following statements are true or false, and explain your answers (you will only get full credit for a correct answer accompanied by a correct explanation):

(a) $E(\hat{\beta}_{(k)}) = \beta$.

(b) $E(s_k^2) = \sigma^2$.

(c) $E(s_k^2) = \frac{1}{n-p} e'e$.

(d) The square roots of the diagonal elements of V are the bootstrap standard errors.

(e) The sample SD of the $\hat{\beta}_{(k)}$s is a good approximation to the SE of $\hat{\beta}$.

(f) If $\hat{\epsilon}_{(k)} \perp X$ for all k, this suggests that $\epsilon \perp\!\!\!\perp X$. (Here, $\perp$ means "orthogonal" and $\perp\!\!\!\perp$ means "independent").

7. Under what conditions is the least square estimator undefined?

8. A researcher argues that the effect of a college degree on ideological orientation (measured with a left-to-right scale) is positive. To test this argument, the researcher gathers data on ideology, and college degree status, as well as other variables including whether a person has a white or blue collar job. She runs an OLS regression where the dependent variable is a person's ideology and the independent variable is a dummy that takes the value of 1 when a person finished college, and 0 otherwise.

   Now, she is thinking about whether to include occupation in the regression. A college degree clearly opens the door to higher-paying white collar jobs. Should occupation therefore be seen as an omitted variable in a regression of ideology on college degree status? Would you recommend that she include or exclude occupation from her equation? Would your recommendation be different if college degree completion is randomly assigned?

9. Suppose that (y, x, z) all have a zero mean. Suppose you wish to estimate the regression $y = \beta x + \epsilon$ but you are concerned that $Cov(x, \epsilon) \neq 0$, but you know that $Cov(z, \epsilon) = 0$. (These covariances are defined over random variables). Now, given the following empirical covariances, construct the OLS and IV estimates.

	y	x	z
y	3.2	1.8	1.5
x	1.8	1.2	0.5
z	1.5	0.5	1.0

10. Suppose that $Y_i = \alpha + \beta X_i + \epsilon_i$ for all $i = 1, ..., n$, where X_i is a scalar. There is another scalar variable Z_i with $Z_i \perp\!\!\!\perp Y_i$. Now, a researcher regresses the $n \times 1$ vector Y on the $n \times 1$ vectors X and Z and uses the coefficient of X to estimate β. What are the consequences, if any, of including Z in this regression? Does this depend on the dependence between X and Z?