EMPIRICAL ANALYSIS AND RESEARCH METHODOLOGY EXAMINATION
Yale University

Department of Political Science

August 2011

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 8.5x11 handwritten (not photocopied) sheet of notes.

PART I.

Professor Smedley wants to know how economic interests affect people's views of the estate tax. He plans to examine two dimensions of the question. First, Smedley wants to know the influence of income on estate tax policy opinions. He interviews a large random sample of adults, collecting information about estate-tax attitudes, income, and other variables that may be related to estate-tax attitudes. He proposes to use matching to estimate the effect on attitudes of earning more than $100,000 per year.

Second, Smedley wants to examine how winning large sums of money in the state-sponsored lottery affects people's view about the estate tax. He interviews a random sample of adults and compares the attitudes of those who report winning more than $10,000 in the lottery to those who claim to have won little or nothing. He reasons that the lottery chooses winners at random, and therefore the amount that people report having won is random.

Smedley comes to you for suggestions about statistical analysis. Consider both his matching proposal for studying the influence of income and his lottery project. In each case, what are the important threats to unbiased inference? What alternative research design would you suggest to Smedley? Suggest one or two specific models to evaluate the effect of income and the effect of winning the lottery and discuss their potential advantages and disadvantages.

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).

Michael S. Lewis-Beck, Richard Nadeau and Angelo Elias. 2008. "Economics, Party, and the Vote: Causality Issues and Panel Data" *American Journal of Political Science*. 52(1): 84-95.

PART III. Statistical Reasoning

1. Suppose you estimate the model $Y_i = \beta X_i + \gamma Z_i + \epsilon_i$, with the usual OLS assumptions. Here X_i and Z_i are mean-zero scalar random variables, and ϵ_i is the disturbance term. One of your reviewers expresses the following concerns. Comment on the validity of the reviewer's observations.

   (a) The estimates $\hat{\beta}$ and $\hat{\gamma}$ are biased because your model has no intercept, so you are forcing the regression plane to go through zero.

   (b) X and Z are highly correlated, causing problems of multicollinearity. This problem, according to the reviewer means that the estimates and standard errors are consistent but biased in small samples.

2. Consider the model: $y = X\beta + Z\gamma + \epsilon$

   Let $M = I - X(X'X)^{-1}X'$. Now consider the following regressions:

   (a) $My = Z\gamma + \epsilon_1$

   (b) $y = Z\gamma + \epsilon_2$

   (c) $y = MZ\gamma + \epsilon_3$

   Which of the regressions could be used to provide unbiased estimates of γ?

3. Suppose that the true model is $Y_i = \beta_1 X_i + \beta_2 Z_i + \epsilon_i$ but the econometrician mistakenly postulates: $Y_i = \beta_1 X_i + \epsilon_i$. What are the implications, if any, of leaving Z_i out of the model?

4. Is the following statement true, false or uncertain? "Suppose you have panel data that follows people for five years. The use of a fixed effects estimator will solve any endogeneity by removing all unobserved differences among people".

5. 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?

6. A researcher argues that the introduction of a new work law should affect men more than women. She tests her theory by gathering data on income for both men and women in both the period before and the period after the law was introduced. She then performs the following difference-in-difference analysis: E[((Income for men in after period)-(Income for men in before period))-((Income for women in after period)-(Income for women in before period))]=12.

   Based on this information, which of the coefficients in the following regression could you identify:

$$\text{Income} = \alpha + \beta_1 \text{Men} + \beta_2 \text{After Period} + \beta_3 \text{Men*After Period} + \epsilon \tag{1}$$

   Justify your answer.

7. The probability distribution function for the poisson distribution can be written as:

$$\mathrm{f}(y_i|\lambda_i) = \begin{cases} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{for } \lambda_i > 0, y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Suppose that the rate of event occurrence λ is modeled as a function of an independent variable x with the following form: $\lambda_i = exp(\beta_0 + \beta_1 x_i)$. Further, assume that you have a data set with n observations of the variables y and x. Derive the log-likelihood for the poisson model. Explain each step of your derivation. Explain how you could use the resulting log-likelihood function to estimate the parameters β_0 and β_1.

8. Suppose $x_1 \ldots x_n$ and $y_1 \ldots y_n$ are real numbers with $s_x > 0$ and $s_y > 0$. The standardized versions of these variables are x^* and y^*; they have mean 0 and variance 1. Prove that $\mathrm{cor}(x, y) = \mathrm{cor}(x^*, y^*)$.

9. Consider the model $y_t = \beta_0 + \beta_1 y_{t-1} + e_t$, where $E\left[e_t \mid y_{t-1}, y_{t-2}, \ldots\right] = 0$. You estimate this model with OLS knowing that once y lagged one period has been controlled for, no further lags of y affect the expected value of y_t.

   (a) Are your estimates biased? Consistent? Do the answers to these questions depend on the values of β_0 or β_1?

   (b) Are the errors in the model serially correlated? If you are using OLS to estimate the model, does it matter whether the errors are serially correlated?

10. You have panel data that cover two periods. There is a treatment, $D_{i,t}$, that equals 0 for all subjects in the first period and 1 for some subjects in the second period. You use OLS to estimate two models:

    (a) $Y_{i,t} - Y_{i,t-1} = \beta_1 D_{i,t} + e_{i,t}$, and

    (b) $Y_{i,t} = \beta_2 Y_{i,t-1} + \beta_3 D_{i,t} + e_{i,t}$,

    where $D_{i,t}$ is a dummy variable for a treatment that i may receive at time t.

    A reviewer says that you should estimate a third model to get a new estimate of the treatment effect:

    (c) $Y_{i,t} - Y_{i,t-1} = \beta_4 Y_{i,t-1} + \beta_5 D_{i,t} + e_{i,t}$.

    Because you have already estimated (a) and (b), you tell the reviewer that it would be superfluous to estimate (c). What is your reasoning?