EMPIRICAL ANALYSIS AND RESEARCH METHODOLOGY EXAMINATION
Yale University

Department of Political Science

January 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 eight hours to complete the exam. You may use a calculator and one 8.5x11 handwritten (not photocopied) sheet of notes.

PART I.

Professor Smedley is interested in reexamining the argument that left governments are more likely to default on their sovereign debt. Smedley has a dataset with all countries in the world which have outstanding sovereign debt and he also has a measure for each of whether or not they defaulted on their debt. Smedley comes to you for suggestions about statistical analysis. Suggest one or two models to evaluate the relationship between left governments and debt default and discuss their potential advantages and disadvantages. What are the important threats to unbiased inference? What alternative research design would you suggest to Smedley?

PART II. Read the essay attached to your exam.

http://sppq.press.illinois.edu/view.php?vol=8&iss=3&f=malhotra.pdf

Malhotra, Neil. 2008. “The Impact of Public Financing on Electoral Competition: Evidence from Arizona and Maine.” *State Politics and Policy Quarterly* 8(3): 263-281.

(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). Offer a critical evaluation of its methodology. Justify each of your claims and suggest ways in which this line of research might be improved.

PART III. Statistical Reasoning

1. Consider the regresion model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 W_i + \epsilon_i$, where $W_i = X_{i1} + X_{i2}$.

   (a) What is *really* being asserted by the null hypothesis $H_0 : \beta_1 = 0$?

   (b) Suggest another way to test the assertion of H_0? Provide some brief detail as to how you would implement this test.

2. Under what conditions is the least squares estimator undefined?

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. A researcher runs a simple bivariate regression using data from congressional House races where the dependent variable is incumbent vote share and the independent variable is the amount of money spent by the incumbent given in $10K increments (i.e. 1=$10,000, 2=$20,000, etc.). The coefficient on money spent by the incumbent is 0.2. If the independent variable were instead coded in $100K increments (i.e. 1=$100,000, 2=$200,000, etc.) how would the coefficient in the regression change? Show work to justify answer.

5. In November 1993, the state of Pennsylvania conducted elections for its state legislature. The result in the Senate election in the 2nd district (based in Philadelphia) was challenged in court. The facts of the matter are that the Democratic candidate won 19,127 of the votes cast by voting machines on election day, while the Republican won 19,691 votes cast by voting machines, giving the Republican a lead of 564 votes. However, the Democrat won 1,396 absentee ballots, while the Republican won just 371 absentee ballots, more than offsetting the Republican lead based on the votes recorded by machines on election day. The Republican candidate sued, claiming that many of the absentee ballots were fraudulent. The judge in the case solicited expert analysis from Orley Ashenfelter, an economist at Princeton University. Ashenfelter examined the relationship between absentee vote margins and machine vote margins in 21 previous Pennsylvania Senate elections in seven districts in the Philadelphia area over the preceding decade, yielding a data set with the following summary statistics:

   We now use linear regression to examine the relationship between y and x (as defined in the previous table), i.e., $E(y_i|x_i) = \alpha + \beta x_i$, in order to assess the possibility that the disputed election (not included

Table 1: default

Variable	Mean	Min	Max
x, Democratic Margin in Machine Balloting (percentage point lead among two-party votes)	40.68	-13.16	89.32
y, Democratic Margin in Absentee Ballots (percentage point lead among two-party votes)	29.06	-34.67	72.97

in the data set) is suspect. For this regression, we have y

$$\begin{aligned}
X^{'}X &= \begin{pmatrix} 21 & 854.21 \\ 854.21 & 53530.38 \end{pmatrix} \\
(X^{'}X)^{-1} &= \begin{pmatrix} .14 & -.0022 \\ -.0022 & .000053 \end{pmatrix} \\
X^{'}y &= \begin{pmatrix} 610.36 \\ 40597.39 \end{pmatrix} \\
y^{'}y &= 35141.94
\end{aligned}$$

(a) What is the ordinary least squares estimate of β_0 and β_1?

(b) What are the estimated standard errors of the least squares estimates?

(c) Provide an interpretation of the estimated intercept in the regression you just estimated.

6. Is the following statement true, false or uncertain? "For a single regression coefficient, the F-test for restricting the coefficient to be zero is the square of the t-statistics testing the hypothesis that the coefficient is zero" Justify your answer.

7. Suppose that (y, x, z) all have a zero mean. Suppose you wish to estimate the regression $y = \beta x + \epsilon$ but you are concern that $Cov(x, \epsilon) \neq 0$, but you know that $Cov(z, \epsilon) = 0$. 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

8. Consider the following RDD estimates. The running variable, x, has been normalized so that the jump point is zero and D is the indicator variable for treatment:

$y = 5 + 1.8x - 2.4D + 1.2D * x + \epsilon$

(a) What is the impact of treatment? Why?

(b) Is this parameter identifying the average impact of treatment on the treated (ATT)?

(c) In one or two paragraph briefly mention the disadvantages of RDD estimation strategies

9. You want to estimate the true model $y = \gamma + \beta x^* + \epsilon$, but you don't observe x^*. Instead you have x which you know $x = x^* + v$. What are the implications, if any, of using x in your model? (Assume that the $Cov(x, x^* = 0)$)

10. Are the empirical residuals of OLS uncorrelated with the independent variables?