EXAMINATION: EMPIRICAL ANALYSIS AND RESEARCH METHODOLOGY

Yale University

Department of Political Science

January 2013

This exam consists of two parts. Provide complete answers to BOTH sections.

Your answers should be succinct and to the point.

Use algebra to back up your assertions wherever appropriate, and remember to show your work.

Do not answer questions that have not been asked.

Do not leave sub-parts of questions unanswered.

You can use the points assigned to each question as a (rough!) guide to allocation of your time.

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.

For Part I of the exam, please turn in a hard copy of your Stata or R code to Colleen, then later email a digital copy of the code to the exam committee (thad.dunning@yale.edu, allan.dafoe@yale.edu, daniel.butler@yale.edu).

1 Design and Data Analysis (3.5 hours, 100 points)

This question will ask you to analyze a dataset for the purposes of generating a causal inference. We generated this data in a manner that is much simpler than reality, but we tried to add a certain degree of realism. You may use any approach to analyze the data; you should use the approach that is best suited to your goals. In your answer, be sure to explain what tools you used and why, and also why you chose to not pursue other approaches that might seem appropriate to others.

Resources: Data for this question is available on your computers (all three files are equivalent, just in different formats). For this question you may use the reference materials that are included as part of the R or Stata software. We have also provided you with a pdf version of "An Introduction to R" by Venables and Smith. Other than your one page of notes and calculator, no other resources are permitted.

Advice: For this question you will need to be efficient with your time. You might want to spend the first hour just thinking through the question, and writing out in words what you intend to do and why. This might help guide your analysis. And it will provide some clear evidence of the reasoning behind your analysis. Also, since this question will involve coding you should feel free to write your answers as comments in your code.

Some approaches will take substantially more time than others for a small gain in the quality of your estimates. You might want to work through this question first with those tools that are easiest for you to use. Time permitting you can go back to refine your analyses with more appropriate tools and go into greater detail.

Submitting your Answer: For your answer you should submit your code as a computer file as well as in hard copy. You may include your code along with your answer, you may write up your answer as comments in the code, or any other approach that allows the graders to follow your answer. We recommend you put brief comments to guide the reader through your code; even if your code has errors in it, a comment explaining what you intended to do will help us award you partial points.

***

Professor Smedley is studying the effects of losing one's job on willingness to vote. Smedley theorizes that losing one's job will make an individual less likely to vote because they feel disempowered and less welcome in their community. Smedley has a large dataset potentially relevant to this question with information from two government run surveys implemented in years t and $t + 4$. Participation in these surveys is required by federal law and the survey teams are well-funded; there is no missing data. Surveys are administered through in-person interviews; unless stated or hinted at otherwise, you may assume that measures are fairly reliable and valid. Variables are described below. The following scenario description will consist of more details than you need; we added these so that it's not obvious which details are relevant. That is for you to figure out.

Smedley will now describe to you some details of the region from which he drew his sample, and of other covariates he collected. Your job is to propose and implement strategies of

analyzing this data to estimate the causal effect of losing one's job on willingness to vote. You may analyze the data in whatever manner is appropriate, and point out any relevant inferences to Smedley's question.

1.1 Research Context and Dataset

The dataset consists of a random sample of individuals aged 14-70 from a city in an imaginary country that is in some ways similar to the U.S., though the federal government is much more interventionist, as you will see. People younger than 16 are by law not allowed to work. People above age 16 may work, and they work until 70. People get their high school degree when they're 18, and their college degrees when they're 23. Not everyone completes high school or college.

People vary on a number of observed and unobserved dimensions. For example, people observably differ on their enjoyment of exercise, on the number of pets they own, on their conscientiousness, and other characteristics.

Many people in this population suffer from depression. Depression is thought to be heavily determined by genetic and other background determinants. This is evident from studies of depression that have found there to be very strong correlation over time in levels of depression, even after controlling for a large set of other factors. Other factors that are thought to have some effect on depression include exercise, age, and job loss. It is known that for this population depression is associated with, and probably causes, reduced: levels of education, chances of getting and keeping a job, willingness to respond to surveys, desire to exercise, etc... The government survey only asked about depression starting in year $t+4$.

Most of the population is employed. People work in firms. There are 99 firms in the economy, with names "0.01", "0.02"... "0.99". Higher numbered firms are generally considered more white-collar firms and typically employ workers with more education.

Some firms use oil as a primary input, others do not. The profit level for firms that use oil as a primary input are highly negatively correlated with the price of oil. The profit level for firms that do not use oil as a primary input is largely uncorrelated with the price of oil. A major conflict in the Middle East in year t led to a sharp and unexpected increase in the price of oil that has persisted to year $t+4$. This increase in the price of oil was unexpected by most experts and by futures markets.

In year $t+1$ a natural disaster struck the area. A river ran over the city's levees, flooding low lying areas. Specifically, all buildings lower than 45 ft (from sea level) were flooded. Scientists who had studied the possibility of such a flood believed that it could happen about once every thirty years, and that when it did occur it would flood all areas below H, where simulations suggested that H was drawn from a normal distribution with mean $\mu_H = 48$ feet and standard deviation $\sigma_H = 11$.

A number of firms went bankrupt between years t and $t+4$. Bankrupt firms are required by law to fire just under 80% of their employees (technically they are required to fire as many employees as required to get as close to 80% as possible, without going over). After the firings, management of the firm is taken over by another group, and the business life for the firm continues as before. The employees who were not fired continue working. The

unexpected increase in the price of oil made those firms that depend on oil as a primary input more likely to go bankrupt. Firms that had their buildings flooded were also more likely to go bankrupt (each firm has only one building).

The federal government believes that it is economically beneficial to require firms to have turnover in their employees. As such, all firms that did not go bankrupt were required by federal law to fire just under 20% of their employees between the years t and $t+4$ (as above, the firms fire employees until the firing of an additional employee would put them over the 20% threshold). No employees are fired above and beyond these two federal rules.

Attempting to alleviate public concern about the many job losses that are caused by federal policy, the federal government implemented an employment security program in year $t-1$ that granted immunity to certain individuals from losing their job for the next five years: firms are required by law to keep these immunity winners employed until year $t+5$. Immunity was granted to a simple random sample of the population, making up 30% of the population. Smedley does not have a record of those who won immunity, though in year $t+5$ he tried to contact all of the participants who had not lost their job to ask them whether they had won immunity. He called each individual at least twice, and sent a letter by mail.

The federal government provides unemployment insurance. All individuals without a job receive some money from the federal government. Individuals who recently lost their job receive substantially more than those who lost their job earlier. The level of unemployment insurance granted depends on an individual's application for it, as well as other factors.

1.2 Variables

- obs: Observation ID

- Y.t4: Smedley's primary dependent variable measured in year $t+4$. This is a measure of willingness to vote in year $t+4$, with more positive values implying a greater willingness to vote. The measure is constructed using a survey instrument, which for the sake of this question we will assume provides a valid and reliable measure of willingness to vote. Smedley then demeaned the measure so that the metric is centered at 0, but he otherwise does not transform the variable

- jl: An indicator for whether the individual lost their job between years t and $t+4$. Smedley coded all individuals who were not employed at year t as jl$=0$.

- age.t: Individual's age at year t.

- em.t: Whether the individual was employed at any time between t and $t+4$. Those with em.t$=0$ could not lose their job.

- oil.t: An indicator for whether the individual's firm relies on oil as a primary input. Those firms that do rely on oil experienced a negative profit shock during years t through $t+4$. Firms experiencing such a shock were more likely to go bankrupt. Smedley codes a 0 for individuals who were unemployed.

- alt: The altitude of the building of the firm in which the individual works, measured in feet.
- flood: An indicator for whether the building of the firm in which the individual works was flooded.
- bkr.t: Indicator for whether the individual's firm went bankrupt between years t and $t+4$.
- con.t: A measure from year t of the conscientiousness of the individual.
- resp.t: A measure from year t of the difficulty that the administrators of the government survey had in contacting these individuals; larger values correspond to individuals who were easier to contact.
- ed.t: The individual's education at year t. 0 means did not finish high school. 1 means completed high school. 2 means completed college.
- ex.t: A measure of the extent to which the individual enjoys exercise in year t.
- f.t: A numerical code representing the firm that the individual works in. There are 99 firms, named "0.01", "0.02"... "0.99". Smedley coded individuals who were unemployed in year t as having f.t$= 0$.
- pet.t: The number of pets the individuals own at year t.
- sen.t: A measure of the individual's seniority in their firm in year t. The variable provides their seniority in percentiles. Smedley codes individuals who are unemployed at year t as having 0 seniority, a percentile which is otherwise not in the dataset.
- dep.t4: A measure of the extent to which an individual is depressed in year $t+4$. Larger values correspond to more severe depression.
- unem.t4: The dollar value of the unemployment insurance that the individual received in year $t+4$.
- rep.t5: An indicator for whether Smedley was able to contact this individual for a follow-up question about their immunity status. Smedley tried to contact in year $t+5$ all individuals in the survey who were employed and did not lose their job. Smedley does this because he wants to code whether individuals have immunity status, and he knows that if someone did lose their job they must not have had immunity status. rep.t5$= 1$ if Smedley could contact them, rep.t5$= 0$ if he could not *or* if he did not try to contact them because they had lost their job or were not employed.
- imun.t5: An indicator coded by Smedley for whether the individual won immunity status. Amongst those who Smedley contacted (rep.t5$= 1$), imun.t5 is coded as 1 if they had won immunity status, and 0 if they had not. Smedley codes imun.t5 as

0 for those who had lost their job since they clearly did not have immunity status. Smedley also codes all those who were unemployed as having imun.t5= 0. Assume that individuals report truthfully.

1.3 Smedley's Natural Experiments (30 points; app. 45 min)

Smedley has two ideas for natural experiments inducing job loss, one using the oil price shock, the other using the flood. First, Smedley argues that the sudden and unexpected increase in the price of oil is a natural experiment, since it is causally external to the economy under study and was not anticipated. Second, Smedley argues that the flood is a natural experiment since it was caused by a rare unexpected natural disaster.

Smedley asks you for advice. Can either or both of these shocks be used to help estimate an ATE of job loss? If so, describe what you recommend doing, explain why it is appropriate, and implement the analysis if you can. If not, explain and justify why not.

1.4 Analyze the Data (60 points; app. 2 hours)

Help Smedley with his problem. **Analyze this data in whatever manner is best to evaluate the causal effect of job loss.** Justify and explain what you do. If there are approaches that you choose not to follow, though they might seem appropriate to others, you should explain your reasoning for not pursuing them.

There are many ways to approach this question. An excellent answer will demonstrate understanding of the strengths and weaknesses of different approaches to analyzing the data, will be able to articulate and will examine the assumptions underlying one's estimates, and will otherwise demonstrate sophistication in data analysis and causal inference.

1.5 Summary (10 points)

State your best estimate of the causal effect of job loss. A good estimate should include a statement of the conditions under which the estimate is reliable, and a statement of your uncertainty.

2 Statistical Reasoning (3.5 hours, 100 points)

1. (5 points) "Measurement error in regressors biases coefficient estimators towards zero." Under what conditions, if any, is this statement true? Explain.

2. (20 points, 5 for each subquestion). Let Y be an $n \times 1$ column vector and let X be an $n \times p$ matrix. Define $e \equiv Y - X(X'X)^{-1}X'Y$, which has typical element e_i for $i = 1, ..., n$. Let $\bar{e} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i$.

   Also, $\text{Var}(Z) \equiv \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

   Say whether the following statements are true or false, or whether you need more information to answer. For each statement, provide algebra or an informal proof to support your answer. (Giving the correct answer without a correct explanation will not earn full credit).

   (a) $e \perp X$. (Here, $\perp$ means "orthogonal to").

   (b) $\bar{e} = 0$.

   (c) The correlation between e and any column of X is zero.

   (d) $\text{Var}(Y) = \text{Var}(X(X'X)^{-1}X'Y) + \text{Var}(e)$.

3. (15 points, 5 for each subquestion) A scholar conducts a voter mobilization experiment in which subjects are randomly assigned to one of three conditions: a control group (no contact is attempted), a treatment group (canvassers attempt to deliver a face-to-face encouragement to vote), and a placebo group (canvassers attempt to deliver a face-to-face encouragement to recycle).[1] Suppose there are two types of subjects in the experimental study group: those who answer the door to canvassers (Compliers) and those who do not (Never Treats). Use the results in Table 1 to answer the following questions.

   (a) Estimate the number of Never Treats in the control group. Attach a standard error to the estimate. Justify both your estimate and its estimated standard error, i.e., say why your proposed estimators are best.

   (b) Estimate the turnout rates among Never Treats and Compliers assigned to the control group.

   (c) Estimate the effect of treatment among Compliers (i.e., the effect of assignment to the treatment group relative to assignment to control, for Compliers). Is your estimate unbiased? Why or why not?

4. (10 points) "Experimental data should be analyzed according to the Neyman potential outcomes model, because the model is non-parametric and assumption-free." Evaluate this statement. Is the model non-parametric? Assumption-free? Should experimental data be analyzed according to the Neyman model? How about non-experimental data?

[1] This example is loosely based on Nickerson (2005, 2008; see Gerber and Green 2012).

Table 1: **Results from a Get-Out-the-Vote Experiment**

Treatment Assignment	Treated?	N	Voting %
Control	No	2572	31.22%
Treatment	Yes	486	39.09%
	No	2086	32.74%
Placebo	Yes	182	29.79%
	No	818	32.15%

5. (10 points) What are the assumptions necessary for an instrumental variable to be valid? In answering this question, discuss both instrumental-variables analysis based on the potential outcomes model, and instrumental-variables least-squares (IVLS) regression. Identify any differences in the assumptions invoked in these two cases.

6. (10 points) A researcher estimates a regression with n = 100 iid observations. The researcher comes to you with a methodological question: if she were to go out and collect 10 times as much data, how much smaller would the standard errors of her regression be?

7. (10 points) A researcher is planning on studying whether changes in public opinion over time are related to changes in public policy. To measure the independent variable (changes in public opinion) the researcher has gathered together all results from surveys where the same binary question (i.e., the responses to the question can always be coded as 0 or 1 - e.g., approve or disapprove) is asked at two different times. The researcher wants to limit the analysis to those instances where the change in opinion is statistically significant at the 0.05 level. Assuming that 1,000 people are included in each survey and that there is a fairly even division of opinion on the issue, how much difference (in terms of percentage points) does there need to be between the two surveys in order for it to be included in the sample.

8. (20 points, 5 for each subquestion) An analyst assumes the following model:

$$Y_i = a + bX_i + cZ_i + dX_iZ_i + \epsilon_i \tag{1}$$

for all $i = 1, ..., n$. The researcher posits the usual OLS assumptions. Here, X_iZ_i is the product of X_i and Z_i.

   (a) What are the "usual OLS assumptions" here?

   (b) According to the model, what is the marginal effect of intervening to change X_i with Z_i held fixed? And what is the marginal effect of intervening to change Z_i with X_i held fixed?

(c) Suppose that after fitting equation (1) to data, an analyst finds that

$$\widehat{Y}_i = 1.2 + 2.3X_i + 0.5Z_i - 2.1X_iZ_i. \tag{2}$$

Moreover, the estimated variance-covariance matrix of $\widehat{\beta} = (\hat{a}\ \hat{b}\ \hat{c}\ \hat{d})'$ is given by

$$\widehat{\text{cov}}(\widehat{\beta}|X, Z, XZ) = \begin{pmatrix} 0.5 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0.9 & 0.5 & -0.3 \\ 0.4 & 0.5 & 0.4 & 0.3 \\ 0.7 & -0.3 & 0.3 & 0.7 \end{pmatrix} \tag{3}$$

Conduct a t-test of the null hypothesis that the marginal effect of intervening to change X_i, with Z_i held fixed at $Z_i = 1$, is zero. (That is, find the relevant t-statistic, and state whether or not you reject the null based on this t-statistic). Does the size of the n matter for the validity of this test, and if so, why?

(d) Suppose this model is fit to data from a natural experiment, in which Pakistani applicants are assigned at random to receive visas to attend the Hajj pilgrimage ($Z_i = 1$) or not ($Z_i = 0$). Y_i measures attitudes towards the West. An analyst is interested in the moderating effect of the age of pilgrims (measured by X_i) on the effect of the pilgrimage. In this context, can you interpret the parameters of the model as in (b)? What are the implications for the test of the causal hypothesis discussed in (c)?