POL SCI 231b: Final Exam. Spring 2017

Prof. Thad Dunning/GSI Natalia Garbiras Díaz

University of California, Berkeley

Distributed at 9:00 AM on May 4, May 5, or May 6

Due by 9:00 AM, 48 hours after you receive it

You should mail your completed midterm to Natalia Garbiraz Díaz at nataliagarbirasdiaz+231b@berkeley.edu—put "Final" somewhere in the subject line.

You can write answers to the theoretical/conceptual problems by hand and scan the hard copy before emailing it to Natalia; alternately, use LaTeX, R markdown, and/or a word processor. For coding problems, use R markdown or sweave and show all your code. Be concise in your answers, but please show all of your work for all questions.

To facilitate blind grading, please **do not include your name in your exam or in the names of any files you send us**. Instead, use your Berkeley student ID. **Please add your student ID to the header of every page of your exam.** To do this in TeX (or Sweave for those compiling TeX through `R`), add the following to your preamble:

```
\usepackage{fancyhdr}
\pagestyle{fancy}
\rhead{SID: [INCLUDE YOUR STUDENT ID HERE]}
```

The exam is open note and open book. Thus, you may consult any of the materials assigned this semester (readings, lecture notes, problem set solutions, etc.). You may also use the Internet, for example, for help on some R routine; cite any materials you find if that is necessary (use your judgement). You should **not** search for answers to the exam questions (which you are unlikely to find online in any case); should you find answers, it would be cheating to use them (and would be plagiarism to do so without reporting the source). **You may not work with others or consult anyone else about the exam answers.**

There are 235 points possible. You might want to use the number of points per question as a rough guide to allocating your time.

1. **(10 points)**

   (a) **(5 points) A scholar writes, "A t-test for a difference of means compares two groups of observations and tests the hypothesis that the average of the two groups is identical." Is this correctly stated? Say why or why not. If not, how would you rewrite the sentence?**

   The sentence is not correctly stated: you do not need a hypothesis to test whether the average of two groups of observations is the same! Simply compare the two numbers $\cdots$

   The null hypothesis being tested is instead about unobserved values in the population from which the groups of observations were drawn. You could rewrite the sentence as follows:

   "A t-test for a difference of means compares two groups of observations and tests the hypothesis that they are drawn from populations with the same average value."

   Note that the null hypothesis is also not that they are drawn from the same population. Two populations could have different second moments (i.e. different variances) but the same average. The right null hypothesis for the difference-of-means test is that the population averages are the same.

   (b) **(5 points) Suppose the data are drawn from an experiment, and the observations are the treatment and control groups. Now is the statement in (a) correct, and why or why not? If not, how would you rewrite the sentence?**

   The statement still is not correct, for the reasons given in (a). The sentence could be rewritten as follows (assuming we are comparing a treatment group to a control group):

   "A t-test for a difference of means compares the treatment sample to the control sample and tests the hypothesis that in the experimental population (study group), the average potential outcomes under treatment equal average potential outcomes under control."

2. **(50 points) During the 2017 presidential election, pollsters such as 538.com and The New York Times' The Update published running estimates of the probability that Hillary Clinton would win the election.**

   (a) **(40 points) How is the probability that Clinton would win the election like the probability of an earthquake in California? How, if at all, are the concepts different? What are the difficulties involved in defining and estimating the probabilities, in each case?**

For your answer, draw in detail on the discussion in David Freedman and Philip B. Stark. 2010. "What is the Chance of an Earthquake?" In David Collier, Jasjeet S. Sekhon, and Philip B. Stark, eds., *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. New York: Cambridge, Chapter 8.

(b) **(10 points) In light of your discussion in (a), do you think a numerical estimate of the probability that Clinton would win the election is helpful and defensible, or unhelpful and misleading? And how would we know which is the better answer? Overall, do you support the practice of pollsters such as 538.com and The Update?**

Answer omitted.

3. **(40 points) A scholar hypothesizes that past vote share for a party in a given constituency (Z) influences current vote share (Y) through the channel of local party organization (X). Thus, where a party received a high past vote share, it will have a stronger local party organization, which will lead to increases in current vote share. To test this hypothesis, and distinguish the "direct" and "indirect" effects of Z on Y, the scholar regresses Y on Z and X.**

(a) **(6 points) What is the "indirect effect" of Z on Y working through X? What is the "direct effect?" Write out the relevant regression models to answer this question.**

The scholar might assume a system of regression equations such as the following:

$$Y_i = aX_i + bZ_i + \epsilon_i \quad (1)$$
$$X_i = cZ_i + u_i \quad (2)$$

(Here, for simplicity, we are representing a system with mean-deviated variables, so there is no intercept in the equations).

Thus, substituting the second equation into the first,

$$Y_i = a(cZ_i + u_i) + bZ_i + \epsilon_i \quad (3)$$
$$= (ac + b)Z_i + nu_i, \quad (4)$$

where $nu_i = au_i + \epsilon_i$. Here, the "direct effect" of Z_i on Y_i is b, while the "indirect effect" working through X_i is ac.

(b) **(6 points) Now, use potential outcomes notation to define the direct and indirect effects. Since Z_i is continuous, define these effects in terms of a marginal change in Z_i, i.e., from z_i to z_i'. Is the effect of a marginal change in Z_i the same for every z_i? Does the answer differ for the models you wrote in (a), and why or why not?**

We can write potential current vote share as a function of both past vote share Z_i and local party organization X_i. Since X_i is a potential mediator, it also depends on Z_i, thus we can write $X_i(Z_i)$. (Here, formal notation differs a little from the example in lecture, because X_i and Z_i are both potentially multi-valued, e.g., past vote share Z_i may be a continuous variable. Party organization could potentially be dichotomous, e.g. $X_i = 1$ is "strong" local organization and $X_i = 0$ is "weak" organization).

Then, potential current vote share is $Y_i(X_i(\cdot), Z_i)$. The observed outcome Y_i is the value of $Y_i(X_i(z_i), z_i)$ at the particular value of $Z_i = z_i$, given also the resulting value of mediator $X_i(z_i)$.

In this framework, the "indirect effect" involves changing the mediator $X_i(\cdot)$ while holding constant $Z_i = z_i$. For example,

$$Y_i(X_i(z_i'), z_i) - Y_i(X_i(z_i), z_i) \tag{5}$$

is the indirect effect of $X_i(\cdot)$ on Y_i holding Z_i constant at z_i. (If $X_i(\cdot)$ is dichotomous, this is not a marginal change in Y_i induced by a marginal change in X_i; that interpretation would hold if X_i is continuous and we are using linear regression).

The "direct effect" of Z_i holding X_i constant at $X_i(z_i)$ is

$$Y_i(X_i(z_i), z_i') - Y_i(X_i(z_i), z_i). \tag{6}$$

If Z_i has a constant marginal effect as in a standard linear regression, then this effect is presumed the same whatever the initial value z_i. Here, however, there is no assumption that the marginal effect is the same for every z_i.

(c) **(8 points) Show how to write the total effect of a marginal change in Z_i as the sum of the direct and indirect effects. (You will need to "hold constant" Z_i at z_i' for purposes of defining potential changes to X_i). Where do "complex potential outcomes" come into play here?**

Since Z_i is continuous, one could think of the "total effect" of Z_i as the change in Y_i induced by a marginal change in Z_i: for example, $Y_i(X_i(z_i'), z_i') - Y_i(X_i(z_i), z_i)$, where $z_i' - z_i$ is a marginal change in Z_i starting from z_i.

Conceptualizing direct and indirect effects involves imagining impossibilities ("complex potential outcomes"). For example, in the definition of the "direct effect" in (6), of course, the two inputs $X_i(z_i)$ and z_i' in the first term cannot occur simultaneously: when vote share moves from z_i to z_i', the result is $X_i(z_i')$, the value of the mediator (local party organization) under assignment to vote share z_i'.

The same logic applies to the "indirect effect" defined in (5). Again, this definition involves imagining an impossibility, because if past vote share remains fixed at z_i, the value of local party organization remains fixed as well.

Notice that for the indirect effect, we can imagine a change in the mediator from $X_i(z_i)$ to $X_i(z_i')$ holding constant $Z_i = z_i$, as in (5); or, we can imagine a change in the mediator from $X_i(z_i)$ to $X_i(z_i')$ holding constant Z_i at z_i'.

$$Y_i(X_i(z_i'), z_i') - Y_i(X_i(z_i), z_i') \tag{7}$$

Under this second definition, the total effect of a change in Z_i from z_i to z_i' is just the sum of the direct and indirect effects:

$$[Y_i(X_i(z_i), z_i') - Y_i(X_i(z_i), z_i)] + [Y_i(X_i(z_i'), z_i') - Y_i(X_i(z_i), z_i')] = Y_i(X_i(z_i'), z_i') - Y_i(X_i(z_i), z_i). \tag{8}$$

Here, the term in the first brackets on the left-hand side is (6), the term in the second brackets is (7), and the right-hand side is the total effect.

(d) **(10 points) In her regression, the scholar finds that the coefficient on X is positive and statistically significant, while the coefficient on Z is negative and statistically significant. She concludes that the direct effect of past vote share on current vote share is actually negative, once the indirect effect on local party organization is taken into account. Comment on this analysis. Is this a valid conclusion? Why or why not?**

This appears to be a classic case of bias in mediation analysis. Suppose the scholar estimates the system of equations in part (a). The difficulty is that ϵ_i and u_i in (1) are not likely to be independent. Indeed, positive shocks to party organization X_i (such as the appearance of some charismatic candidate who induces local militants to organize for the party) are likely to positively influence the party's current vote share Y_i as well (and thus be "captured" in u_i). Thus, X_i is endogenous in the first line of (1). Moreover, by (1), X_i is correlated with Z_i (as long as $c \neq 0$). Thus, bias "propagates." When the unobserved error terms covary positively, this will lead to an inflation of the estimate of a and depress the estimate of b (see e.g. discussion in Chapter 10 of Gerber and Green). Thus, a finding that "the direct effect of past vote share on current vote share is actually negative, once the indirect effect on local party organization is taken into account" can actually just result from this classic bias in mediation analysis, where the apparent indirect effect is inflated and the direct effect is deflated.

Note: in the answer in part (a), we defined direct and indirect effects using a system of two equations, rather than just the regression of Y on X and Z. It is common (though misguided), however, for scholars to extend e.g. a regression of Y on Z with a regression of Y on X and Z, and to interpret a non-zero significant estimate on X as evidence for an indirect effect. Morale of the story: don't ever do this! :)

(e) **(10 points) Another scholar, drawing on these results, proposes to use Z as an instrumental variable for X in a different analysis of the effect of X on Y. Is this a good idea? Why or why not? Do the results mentioned in part (d) prove your point, or not?**

Not a good idea. After all, the exclusion restriction (which requires no direct effect of Z on Y beyond its indirect effect through X) is explicitly violated by

the first scholar's hypothesis. Whether the exclusion restriction actually holds is essentially untestable: for example, as we saw in part (d), a regression of Y on X and Z is uninformative, so the results mentioned in (d) do not prove the point. However, *a priori* reasoning and some evidence can be brought to bear to assess whether the exclusion restriction is at least plausible. Here, it is not: there must be many channels through which past vote share "affects" current vote share, besides the strengthening of local party organization (e.g., past success attracts donors, creates a "scare-off" effect for challengers, etc. etc. etc. . . .).

4. **(20 points) "Experimental data should be analyzed according to the Neyman potential outcomes model, because the model is assumption free." Evaluate this statement. Is the model assumption free? Should experimental data be analyzed according to the Neyman model? How about non-experimental data?**

   Answer omitted.

5. **(30 points) An analyst is interested in the determinants of civil war in Africa. Let** Conflict_{it} **be a dummy variable equal to 1 if there is an armed conflict in country** i **in year** t **and 0 otherwise;** AgGrowth_{it} **be the growth rate in the agricultural sectors of the economy in country** i **in year** t**; and** IndGrowth_{it} **be the growth rate in the industrial sectors of the economy in country** i **in year** t**. The design matrix** X **has** AgGrowth_{it} **and** IndGrowth_{it} **as columns. The analyst assumes the following model:**

$$\text{Prob}(\text{Conflict}_{it} = 1|X, \epsilon_{it}) = \beta_1 \text{AgGrowth}_{it} + \beta_2 \text{IndGrowth}_{it} + \epsilon_{it}, \tag{9}$$

   where

$$E(\epsilon_{it}) = 0, \tag{10}$$

$$\text{Var}(\epsilon_{it}) = \sigma^2, \tag{11}$$

$$\text{AgGrowth}_{it} \perp\!\!\!\perp \epsilon_{it}, \tag{12}$$

   and

$$\text{IndGrowth}_{it} \perp\!\!\!\perp \epsilon_{it} \tag{13}$$

   for all i **and all** t**. Here, the** ϵ_{it} **are i.i.d. random variables. The analyst has subtracted the mean of** AgGrowth **from each observation** AgGrowth_{it} **and the mean of** IndGrowth **from each observation** IndGrowth_{it}**, so the independent variables are mean-deviated. The analyst estimates the model by OLS.**

   (a) **(2 points) What is the name for a linear model like (9), which has a probability on the left-hand side? Which terms in equation (9) are observable and which are unobservable? What are the parameters of the model?**

   This is a linear probability model. In equation (9), β_1 and β_2 are unobservable, as is the random error term ϵ_{it}. The variables Conflict_{it}, AgGrowth_{it}, and IndGrowth_{it} are all observable.

The parameters are β_1, β_2, and σ^2 (the variance of the error term). Naming these three parameters is sufficient for full credit on the question. Lurking in the background, there is also one additional parameter—the intercept that disappears because equation (9) is written in mean-deviated form.

(b) **(3 points) True or false, and explain: if the assumption in (12) is false, the OLS estimator is unbiased, but the estimated standard errors may be seriously wrong.**

False. If the assumption in equation (12) is false, (12) is endogenous and the OLS estimator is biased. Whether the standard errors are given by the usual OLS formulas depends on other assumptions—such as, whether the ϵ_{it} are in fact i.i.d. with $\text{var}(\epsilon_{it}) = \sigma^2$.

(c) **(5 points) If the assumption in (13) is true, under what conditions, if any, is OLS a biased estimator for β_2, the coefficient on IndGrowth_{it}?**

OLS a biased estimator for β_2 if AgGrowth_{it} and IndGrowth_{it} are correlated, and ϵ_{it} is not independent of AgGrowth_{it}.

(d) **(4 points) Homoscedasticity implies that the variance of Y given X is the same for all X. Does this hold for equation (9)? Why or why not?**

The expected value of the dependent variable—that is, the mean of the random variable—is $\beta_1\text{AgGrowth}_{it} + \beta_2\text{IndGrowth}_{it}$. Denote the mean of a 0-1 random variable as p; then its variance is $(p)(1-p)$. Thus,

$$\begin{aligned}\text{var}(Y_{it}|\text{AgGrowth}_{it}, \text{IndGrowth}_{it}) = &[\beta_1\text{AgGrowth}_{it} + \beta_2\text{IndGrowth}_{it}] \\ &[1 - (\beta_1\text{AgGrowth}_{it} + \beta_2\text{IndGrowth}_{it})].\end{aligned}$$

Clearly, $\text{var}(Y_{it})$ depends on i and t—in particular, the values of IndGrowth_{it} and AgGrowth_{it} for unit i at time t. Thus the variance of Y given IndGrowth_{it} and AgGrowth_{it} cannot be the same for all values of the independent variables.

(e) **(2 points) A critic says that because assumption (12) does not hold, the analyst should use instrumental variables least squares (IVLS) regression. The critic suggests instrumenting for agricultural growth with rainfall growth, RainGrowth_{it}, which is the proportionate change in rainfall in country i in year t over the previous year. The analyst assumes that**

$$\text{RainGrowth}_{it} \perp\!\!\!\perp \epsilon_{it}. \tag{14}$$

What is a typical row of the matrix of instruments, Z_{it}?

A typical row Z_{it} is given by $[\text{Rain}_{it}\ \ \text{IndGrowth}_{it}]$.

(f) (4 points) **Now, let the IVLS estimator for equation (9) be given by**

$$\hat{\beta}_{IVLS} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y, \tag{15}$$

where X and Z are as defined above and Y is a column vector of observations of Conflict_{it}. True or false, and explain: if equations (9), (10), (11),

(13), and (14) hold, but equation (12) is violated, $\hat{\beta}_{IVLS}$ is an unbiased estimator for $\beta = (\beta_1 \ \beta_2)'$.

False. Under these assumptions, $\hat{\beta}_{IVLS}$ is a consistent estimator for $\beta = (\beta_1 \ \beta_2)'$, but it is not unbiased.

(g) **(5 points) Suppose that rainfall diminishes the probability of civil war by making it harder for the government and insurgents to deploy fighters to the field. If so, how many of the assumptions in (9)–(14) would be violated? (Zero through six are possible answers). If no assumptions are violated, say why; if one or more assumptions are violated, say which ones are violated and why. Which of the assumptions are testable? How could you (at least partially) test those assumptions?**

Equation (9) is violated: Rain_{it} should be in the model. That is, the exclusion restriction is not satisfied. Also, if Rain_{it} is left out of the model, we would then also have violations of (12) and (13), to the extent that rainfall is correlated with either agricultural and industrial growth.

(h) **(5 points) Suppose that assumption (12) is true but (13) is false. What are the implications for the IVLS estimator in (15)? In your opinion, how likely is it that (12) is true while (13) is false? Discuss more generally the implications of your answers for IVLS estimation of multiple regression models.**

The IVLS estimator is then biased and inconsistent. Indeed, if IndGrowth_{it} is endogenous, the assumption that $Z \perp\!\!\!\perp \epsilon$ is violated.

For the likelihood that that (12) is true while (13) is false, you should make your own reasoned argument. But, it is generally hard to see why AgGrowth_{it} would be exogenous and IndGrowth_{it} endogenous, or vice versa. For instance, omitted political institutions that shape agricultural growth and the probability of conflict would probably shape industrial growth as well. If AgGrowth_{it} and IndGrowth_{it} are both endogenous, we'll need two new instrumental variables: neither AgGrowth_{it} nor IndGrowth_{it} can be included in the matrix of instruments Z.

The lesson for IVLS estimation more generally is that it is not enough to instrument for a single endogenous right-hand side variable, in a multiple regression set-up: the other right-hand side variables will need instruments as well, unless we can make strong arguments for their exogeneity.

6. **(85 points)** Professor Smith is interested in studying the effect of election observers on the incidence of electoral fraud and violence on the day of the elections. She studies this question in the context of Mexican legislative elections. In Mexico, each electoral district (300 in total) elects one deputy for the Lower House of the Federal Congress. Suppose that each electoral district has 100 polling stations.[1]

She theorizes that election observers should increase the incumbent party's costs of engaging on illegal behavior, aimed at slanting the elections on their favor, and thus,

[1] This number is actually hypothetical but assume this for the purpose of the question.

should reduce electoral fraud and violence on the day of the elections. The main vehicle for fraud in this context is ballot-box stuffing (i.e., inflated reports of voter turnout).

In addition, she argues that political parties should respond to the presence of observers by shifting electoral fraud and violence to stations without observers. In other words, the presence of election observers in a polling station should have spillover effects on the polling stations with no observers that belong to the same electoral district.[2] However, according to the researcher, these spillover effects should be more likely to take place as the number of polling stations with election observers within the same electoral district increases.

Following this discussion, she formulates the following two hypotheses:

- ***H1 (direct effect):*** *The presence of an election observer reduces electoral fraud and violence in this polling station.*
- ***H2 (spillover effect):****Political party activists engaged in illegal behavior are, therefore, likely to respond to the presence of observers by shifting these activities to stations without observers. That is, observers may create spillover effects, and these effects should be more likely the higher the number of treated polling stations within the electoral district.*

The researcher conducted a field experiment to test these hypotheses. She first randomly sampled 60 electoral districts. Then she randomly assigned each one of these electoral districts to one of the following levels of saturation: *control*, *low*, *high*. In electoral districts assigned to a *high* level of saturation, 80% of the polling stations in the district were randomly assigned to the election observer treatment. In electoral districts assigned to a *low* level of saturation, 30% of the polling stations were assigned to the election observer treatment. Finally, in electoral districts assigned to *control* none of the polling stations was assigned to an election observer. The researcher therefore hypothesizes that spillovers from treated to control polling stations will only occur in *high* saturation districts, but not in the *low* saturation condition.

Within each electoral district, polling stations were randomly assigned to have an election observer or not, based on the saturation level randomly assigned to the electoral district.

The two outcomes of interest are electoral fraud and violence. For the former, the researcher uses turnout as a proxy for fraud. Electoral violence is measured as reported instances of violence on the election day.

The researcher has the following variables in her database:

- `cluster.id`: Electoral district ID
- `polling.id`: Polling station ID
- `cluster_random`: Saturation level (0=control, 1=low, 2=high)
- `polling_treat`: Election observer (0= control, 1=treatment)

[2]The researcher assumes that spillovers will only take place *within* the same electoral district.

- `turnout`: Turnout in the election being studied
- `violence_pre`: Pre-treatment measure of electoral violence
- `violence`: Reported instances of electoral violence in the election being studied
- `rural`: Rurality index (0-100, with lower number meaning higher rurality) in the year before the elections.
- `ethnic_fract`: Ethnic fractionalization in the year before the elections.
- `polcomp`: Political competition (measured as %winner - %runner up in past legislative elections)
- `poverty`: Percentage of the population below the poverty line by the end of the year before the elections.
- `gini`: Gini coefficient (measure for inequality, ranges from 0-1, with 1=perfect equality) in the year before the elections.
- `homicides`: Average homicide rate (per 100,000 pop.) in the year before the elections.

The dataset `Final.RData` contains these variables and is available in the final exam compressed folder.

(a) **(5 points) Write the potential outcome for a polling station in terms of the indicators T_{ij} and S_j, where T_{ij} indicates assignment to an electoral observer to polling station i in constituency j, and S_j indicates the saturation level of constituency j. How many potential outcomes does each polling station have? Are all of these observable, in principle? Explain your answers.**

Based on the structure of the randomization, we can define a polling station's potential outcomes as follows.

$$Y_{ij}(T_{ij}, S_j)$$

where Y_{ij} is the level of electoral fraud or violence for polling station i located in constituency j. T_{ij} indicates whether the polling station has an election observer ($T_{ij} = 1$) or not ($T_{ij} = 0$). S_j corresponds to one of the three possible saturation levels (i.e., pure control, *low* or *high*).

For instance, $Y_{ij}(1, 1)$ corresponds to the potential outcome (i.e., either level of electoral violence or fraud) of polling station i when it is assignted to an election observer, and that belongs to electoral district j that is assigned to a *low* saturation level.

Each polling station has 6 potential outcomes:

- $Y_{ij}(0, 0)$
- $Y_{ij}(0, 1)$
- $Y_{ij}(0, 2)$

- $Y_{ij}(1,0)$
- $Y_{ij}(1,1)$
- $Y_{ij}(1,2)$

All are in principle observable except for $Y_{ij}(1,0)$, because in the pure control no polling stations are assigned to treatement T.

(b) **(10 points) Now, write out the estimands that the researcher will seek to estimate to test H1 and H2.**

Direct effects (H1)

Let's first think about what is the N in this research design. We may want to define estimands in terms of the population of $30,000$ polling stations, particularly because (as we discuss below), this parameter is estimable. Alternatively, one could define the estimands for the $6,000$ polling stations (60 sampled electoral districts times the 100 polling stations in each district). This distinction corresponds to a contrast between Population Average Treatment Effects (for the $30,000$ polling stations) and Sample Average Treatment Effects (for the $6,000$ polling stations in our sample). (In much inference under the Neyman model, estimands are defined for the study group of units, and formal inferences are confined to this group; in this richer setting, with a randomly sampled study group, it may make sense to think of PATEs).

According to the hypothesis of no spillover effects in low saturation districts, we can define the estimand for the average direct treatment effect of election observers on the polling station's level of electoral fraud and violence as:

$$\tau_{direct} = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}(1,1) - \sum_{i=1}^{N} Y_{ij}(0,1) \tag{16}$$

This is equivalent to:

$$\frac{1}{N}\sum_{i=1}^{N} Y_{ij}(1,0) - \sum_{i=1}^{N} Y_{ij}(0,0) \tag{17}$$

Of course, as noted above, $Y_{ij}(1,0)$ is not observable, but we could still use this to define the estimand, under the assumption of no spillover effects for low saturation or pure control districts.

Spillover effects (H2)

We need to compare two parameters to define H2:

$$\frac{1}{N}\sum_{i=1}^{N} Y_{ij}(1,2) - \sum_{i=1}^{N} Y_{ij}(0,2) \tag{18}$$

and

$$\frac{1}{N}\sum_{i=1}^{N} Y_{ij}(1,1) - \sum_{i=1}^{N} Y_{ij}(0,1) \tag{19}$$

Expression (18) is the difference in average potential otucomes under treatment and control in the high saturation condition, while (19) is the direct effect defined above.[3]

Under the null hypothesis of no spillover effects, the difference between the estimand in (18) and the estimand in (19) should be zero. Under H2, the alternate hypothesis holds: this difference is non-zero.

(c) **(20 points) Consider the following balance test tables:**

	Mean 1	Mean 0	Difference	SE Diff	t-stat	N	df	p-value
Rurality index	43.4725	43.1946	0.2779	0.1697	1.6375	4000	3996.4818	0.1016
Ethnic Fract.	0.5104	0.5125	-0.0021	0.0194	-0.1099	4000	3985.6491	0.9125
Political comp.	35.4722	35.3162	0.1560	0.2806	0.5559	4000	3997.9896	0.5783
Poverty	37.2879	40.9089	-3.6210	0.4976	-7.2776	4000	3853.2804	0.0000
Gini	0.5143	0.5202	-0.0059	0.0064	-0.9204	4000	3989.5534	0.3574
Homicides Rate	36.1944	36.2254	-0.0309	0.1006	-0.3075	4000	3997.9803	0.7585

Table 1: Balance Tests for low saturation versus pure control conditions

	Mean 1	Mean 0	Difference	SE Diff	t-stat	N	df	p-value
Rurality index	43.3320	43.1946	0.1374	0.1714	0.8018	4000	3997.9999	0.4227
Ethnic Fract.	0.4858	0.5125	-0.0266	0.0185	-1.4399	4000	3990.7262	0.1500
Political comp.	35.2700	35.3162	-0.0463	0.2785	-0.1661	4000	3997.3120	0.8681
Poverty	42.8505	40.9089	1.9416	0.4154	4.6736	4000	3902.5479	0.0000
Gini	0.5321	0.5202	0.0119	0.0067	1.7919	4000	3990.6399	0.0732
Homicides Rate	36.2431	36.2254	0.0177	0.0999	0.1771	4000	3997.3734	0.8595

Table 2: Balance Tests for high saturation versus pure control conditions

Are these balance tests conducted correctly, or is there a flaw in the analysis? Explain your answer. If there is a flaw, write code to correct the analysis. Do these results suggest that the randomization failed?

From the tables it is not clear whether the researcher is testing random assignment to saturation levels (which is done at the cluster level) or to election observers (which is done at the polling station level).

Thus, we could run two types of balance tests. We could assess whether random assignment to election observers at the polling station level was successful or not. Alternatively, we could test whether random assignment to saturation level

[3]Notice that, based on the above discussion, the latter estimand is equivalent to (17).

conditions was successful or not. For the former, the analysis conducted by the researcher is correct. Randomization to election observers is done within cluster and, hence, we could run simple t-tests to assess if something went wrong with the randomization. However, we may have higher power if we were to use the entire sample (as opposed to running separate tests for low and high saturation levels). This should be good practice since, when running balance tests, the researcher may want to maximize her power to reject the null.

For this purpose we will need our function for the `t-test`:

```
# t-test
source("~/Dropbox/231B_Spring2017/final 231b_Spring 2017/Applied_question/t_tes
```

```
X <- cbind(data$rural, data$ethnic_fract,
    data$polcomp, data$poverty, data$gini,
    data$homicides)

# Balance low vs control:

ttest_covs <- NA
for (i in 1:ncol(X)) {
    table <- ttest(X[, i], data$polling_treat)
    ttest_covs <- rbind(ttest_covs,
        table)
}

ttest_covs <- ttest_covs[-1, ]
rownames(ttest_covs) <- c("Rurality index",
    "Ethnic Fract.", "Political comp.",
    "Poverty", "Gini", "Homicides Rate")
```

	Mean 1	Mean 0	Difference	SE Diff	t-stat	N	df	p-value
Rurality index	43.4307	43.2765	0.1541	0.1440	1.0703	6000	4616.4239	0.2845
Ethnic Fract.	0.5026	0.5031	-0.0005	0.0160	-0.0287	6000	4665.8902	0.9771
Political comp.	35.3597	35.3488	0.0110	0.2358	0.0465	6000	4638.1090	0.9629
Poverty	41.2685	39.8168	1.4517	0.3856	3.7651	6000	4967.8542	0.0002
Gini	0.5294	0.5180	0.0114	0.0056	2.0446	6000	4485.9604	0.0410
Homicides Rate	36.2434	36.2079	0.0355	0.0847	0.4194	6000	4618.8047	0.6749

Table 3: Balance Tests for Treatment Assignment to Election Observers

Did the randomization fail? This is hard to tell. In an experiment unlucky things could happen (e.g., it could have been the unusual luck of a draw). Yet, a third of the tests suggest imbalance.

In terms of testing whether random assignment to saturation levels was successful, we would need to take into account the clustered design.

We can reassess balance getting the correct standard errors. First, we create our own function to get unbiased standard errors for a cluster randomized experiment:

```
ttest_cl <- function(Y, Z, clust_id) {

    stopifnot(require(sandwich))

    # Get cluster-level means

    cl_treat <- unique(clust_id[Z ==
        1])
    cl_control <- unique(clust_id[Z ==
        0])

    cl_mean_treat <- unlist(lapply(cl_treat,
        FUN = function(x) mean(Y[clust_id ==
            x])))
    cl_mean_control <- unlist(lapply(cl_control,
        FUN = function(x) mean(Y[clust_id ==
            x])))

    Y_cl <- c(cl_mean_treat, cl_mean_control)
    Z_cl <- c(rep(1, length(cl_treat)),
        rep(0, length(cl_control)))

    model <- lm(Y_cl ~ Z_cl)

    se <- sqrt(vcovHC(model, type = "HC1")[2,
        2])
    p.value <- pt(abs(model$coefficients[2]/se),
        df = summary(model)$df[2], lower.tail = F)

    N1 <- length(na.omit(Y_cl[Z == 1]))
    N0 <- length(na.omit(Y_cl[Z == 0]))

    # Preparing output
    res <- c(model$coefficients[2],
        model$coef[1], (N1 + N0), se,
        p.value)
    names(res) <- c("Diff", "Y0", "N",
        "SE diff", "p-value")

    return(c(res))
```

```
}
```

Now, we use it to run the correct balance tests for treatment assignment to saturation levels.

```
# Subset samples treat:

X_low <- X[(data$cluster_random == 0 |
    data$cluster_random == 1), ]
low_data <- data[(data$cluster_random ==
    0 | data$cluster_random == 1), ]
low_data$treat <- ifelse(low_data$cluster_random ==
    1, 1, 0)
X_high <- X[(data$cluster_random ==
    0 | data$cluster_random == 2), ]
high_data <- data[(data$cluster_random ==
    0 | data$cluster_random == 2), ]
high_data$treat <- ifelse(high_data$cluster_random ==
    2, 1, 0)

# Balance low vs control
pval_cluster <- NA
N <- NA
for (i in 1:ncol(X)) {
    pval_cluster[i] <- ttest_cl(X_low[,
        i], low_data$treat, low_data$cluster.id)[5]
    N <- ttest_cl(X_low[, i], low_data$treat,
        low_data$cluster.id)[3]
}

## Loading required package:  sandwich

# Check we have the right N:
N

##  N
## 40

pval_cluster[pval_cluster < 0.05]

## numeric(0)

which(pval_cluster < 0.05)

## integer(0)
```

```r
# Balance high vs control
pval_cluster <- NA
N <- NA
for (i in 1:ncol(X)) {
    pval_cluster[i] <- ttest_cl(X_high[,
        i], high_data$treat, high_data$cluster.id)[5]
    N <- ttest_cl(X_low[, i], low_data$treat,
        low_data$cluster.id)[3]
}
# Check we have the right N:
N

##  N
## 40

pval_cluster[pval_cluster < 0.05]

## numeric(0)

which(pval_cluster < 0.05)

## integer(0)
```

When assessing success of random assignment to the saturation level, we see that our N goes down to 40. This is the case because treatment assignment is done at the electoral district level. Here, we see none of our pre-treatment covariates is imbalanced.

(d) **(25 points) Test both H1 and H2. What can you conclude in each case?**
First, let's take a look at the matrix of treatments:

```r
# Saturation level
table(data$cluster_random)

##
##    0    1    2
## 2000 2000 2000

# Election observer
table(data$cluster_random, data$polling_treat)

##
##         0    1
##   0 2000    0
##   1 1400  600
##   2  400 1600
```

A. Test H1.

Now, we can test the researcher's first hypothesis by estimating the estimand defined in (16) using the districts assigned to the low saturation condition. Here, given random assignment, we can use the sample equivalents to get an estimate of (16).

```
# Subset data for the analysis

low.treat <- data[data$cluster_random ==
    1, ]

# Fraud
round(ttest(low.treat$turnout, low.treat$polling_treat),
    4)

##     Mean 1     Mean 0 Difference
##    55.6685    59.7038    -4.0353
##    SE Diff     t-stat          N
##     0.6558    -6.1528  2000.0000
##         df    p-value
##  1123.8119     0.0000

# Violence
round(ttest(low.treat$violence, low.treat$polling_treat),
    4)

##     Mean 1     Mean 0 Difference
##     7.5866    17.6319   -10.0452
##    SE Diff     t-stat          N
##     0.3635   -27.6324  2000.0000
##         df    p-value
##  1131.6207     0.0000
```

The t-tests show that the presence of election observers reduces the prevalence of fraud (≈ -4) and violence (≈ -10) These effects are statistically significant.

B. Test H2.

Following our discussion in part (b), in order to test H2, we can use our sample equivalents to estimate (18) and (19).

To be sure, following the researcher's theory, under the null hypothesis of no spillover effects, the difference between (18) and (19) should be zero. Again, given random assignment, we can use the difference-in-differences estimator to test this hypothesis.

Fraud

```
# Subset data for analysis
low_data <- data[data$cluster_random ==
    1, ]
high_data <- data[data$cluster_random ==
    2, ]

# Low saturation
round(ttest(low_data$turnout, low_data$polling_treat),
    4)
```

```
##     Mean 1     Mean 0 Difference
##    55.6685    59.7038    -4.0353
##    SE Diff     t-stat          N
##     0.6558    -6.1528  2000.0000
##         df    p-value
##  1123.8119     0.0000
```

```
# High saturation
round(ttest(high_data$turnout, high_data$polling_treat),
    4)
```

```
##     Mean 1     Mean 0 Difference
##    55.7654    67.7219   -11.9565
##    SE Diff     t-stat          N
##     0.6518   -18.3448  2000.0000
##         df    p-value
##   608.0130     0.0000
```

```
# Difference in differences
# estimator:

estimate <- ttest(high_data$turnout,
    high_data$polling_treat)[3] - ttest(low_data$turnout,
    low_data$polling_treat)[3]
se_diff <- sqrt(ttest(high_data$turnout,
    high_data$polling_treat)[4]^2 +
    ttest(low_data$turnout, low_data$polling_treat)[4]^2)

estimate
```

```
## Difference
##  -7.921158
```

```
se_diff
```

```
##   SE Diff
## 0.9246271
```

The difference in differences estimator is -7.9211581. This estimate is consistent with the presence of spillover effects. Let's think about this in detail. Within saturation level conditions, the effect of election observers should be negative (i.e., reduce fraud). Yet, under our model of spillover effects, this negative effect should be higher in high saturation districts than in low saturation districts, as the average of control units in the former is pulled up due to spillover effects. This should lead to a larger difference and, then, a negative difference in differences estimator. Looking at the standard error of the difference in differences estimator (0.9246271), we reject the null of no spillover effects, concluding that the difference between (18) and (19) may not be equal to zero.

Violence:

We see the same pattern when using electoral violence as the outcome of interest.

```
# Low saturation
round(ttest(low_data$violence, low_data$polling_treat),
    4)

##     Mean 1     Mean 0 Difference
##     7.5866    17.6319   -10.0452
##    SE Diff     t-stat          N
##     0.3635   -27.6324  2000.0000
##         df    p-value
##  1131.6207     0.0000

# High saturation
round(ttest(high_data$violence, high_data$polling_treat),
    4)

##     Mean 1     Mean 0 Difference
##     8.6543    22.5993   -13.9450
##    SE Diff     t-stat          N
##     0.4663   -29.9046  2000.0000
##         df    p-value
##   614.2653     0.0000

# Difference in differences
# estimator:

estimate <- ttest(high_data$violence,
    high_data$polling_treat)[3] - ttest(low_data$violence,
    low_data$polling_treat)[3]
```

```
se_diff <- sqrt(ttest(high_data$violence,
    high_data$polling_treat)[4]^2 +
    ttest(low_data$violence, low_data$polling_treat)[4]^2)

estimate

## Difference
##  -3.899778

se_diff

##   SE Diff
## 0.5912744
```

Here, the difference in difference estimator suggests that the difference between the effect of election observers in high saturation election districts is greater than this treatment effect among low saturation districts, and this difference is statistically significant.

In sum, for both fraud and violence, evidence supports H2.

(e) **(15 points) Can you think of any ways to increase the researcher's power to test H1 and H2, for example, by collecting additional data? What other types of analysis could the researcher run to test H2 further?**

At least for the fraud outcome, the researcher could use official turnout data on the 240 control districts that were not sampled into the study group. (These provide a random sample of control potential outcomes under the pure control saturation condition.)

(f) **(10 points) Suppose the researcher's hypothesis that spillovers from treated to control polling stations will only occur in high saturation districts were false; i.e., suppose there are spillovers in the low saturation condition. Can we test H1?**

No, at least not by the methods the researcher has outlined. Following the discussion in part (d), if spillover effects were also present in low saturation level districts, we would not be able test H1. Specifically, we no longer can use the outcomes of treated and control polling stations in low saturation level districts to estimate the effects of election observers without spillover.