PS C236A / Stat C239A
Midterm Exam - Solutions

I. True or False

Answer *True* or *False*. Explain your answer in a sentence or two.

1. TRUE

In the study, the scientist wants to estimate the average treatment effect (ATE), where the potential outcomes are *assumed* to be a function only of treatment T and gender M. For the $4n$ subjects, of which exactly $2n$ are men ($M_i = 1$) and $2n$ are women ($M_i = 0$), the scientist assigns treatment to exactly half of each group.

Thus, in this design, $M \perp\!\!\!\perp T$, so ATE can be estimated from the observed quantities from four groups:

$$\begin{aligned}\widehat{ATE} &= \frac{1}{4n}\left\{\sum_{i\in\{1,1\}} Y_i(1,1) - \sum_{i\in\{0,1\}} Y_i(0,1) + \sum_{i\in\{1,0\}} Y_i(1,0) - \sum_{i\in\{0,0\}} Y_i(0,0)\right\} \\ &= \frac{1}{4n}\left\{\sum_{i\in\{1,M\}} Y_i(1,M) - \sum_{i\in\{0,M\}} Y_i(0,M)\right\}\end{aligned}$$

which is unbiased, so long as only treatment and gender are relevant in predicting the potential outcomes *and* treatment assignment is balanced across gender groups.

The first part of this is also assumed in the following linear model:

$$Y_i = \alpha + \beta_1 T_i + \beta_2 M_i + \epsilon_i$$

This model assumes that Y_i is a systematic and additive function only of T and M, with the other variance in Y_i to be random noise in ϵ_i according to the OLS assumptions.

And since $T \perp\!\!\!\perp M$ and thus $Cov(T, M) = 0$, it can be shown in OLS that the estimator for β_1 is $\hat{\beta}_1 = Cov(Y, T)/Var(T) = Cov(Y, T)/p(1 - p) = 4 \times Cov(Y, T)$. By the definition of covariance with a binary variable T, this estimator is identical to:

$$\hat{\beta}_1 = \frac{1}{4n}\left\{\sum_{i=1}^{2n} Y_i(1,M) - \sum_{i=1}^{2n} Y_i(0,M)\right\}$$

which is the difference in means estimator for observed outcomes under the assigned treatment and control for men and women *balanced* across treatment assignment.

Note, that for this OLS estimator to be an unbiased estimate of ATE, the potential outcomes cannot be a function of any factor other than treatment and gender, since treatment is not assigned randomly.

2. FALSE

The condition being tested is that balancing the mean of x for treated and the mean of x for control reduces bias in an estimate of the effect of T on Y, so long as other moments of x across treated and control do not become more dissimilar acording to a Kolmogorov-Smirnov (KS) test. Although this is possible (and may often occur), there is no guarantee that bias is reduced in this case. For example, x may confound the estimate of a treatment in its second-, third-, or higher-order moments, but not due to its mean moment. In such a case, let's assume that the means and variances get better, but higher-order moments get worse, which contribute to bias if left imbalanced. Since the KS test is sensitive to improvements in the balance of many moments, it is possible that under such a condition the *results* of the KS test do not change, but certain important moments get worse increasing bias.

3. FALSE

   Since the design is implemented at the state level, the researcher is breaking the design by ignoring this assignment mechanism and only focusing on the study of individual-level outcomes. In this case, treatments are correlated by geographical location and are not independent. A consequence of this is that the probability of rejecting the null hypothesis of no effect is likely to be biased. For instance, in a permutation approach, this is equivalent to permitting permutations of a treatment that have zero probabiity of occuring, but that contribute probability to a test of the null. Typically, this failure to adhere to the clustered nature of the assignment will understate uncertainty and overstate power in such a test since the number of n of independently assigned units is much greater when assuming individual-level rather than state-level assignment, which assumes the test has much more independent information about the treatment effect than it should given the assignment.

II. Regression Discontinuity Design

Consider a large medical trial for a new weight loss drug. Before the trial, each patient has their weight, height, and body fat percentage measured. A goodness-of-health score is calculated for each patient based on these measurements (higher scores are a proxy for worse health). Assume that patients do not have time to manipulate their weight or body fat once selected to participate in the trial. Historically, a histogram of patient goodness-of-health scores closely follows a normal distribution with mean c. It is thought that the effect of the drug varies with the value of this score.

a) Suppose that the trial is conducted so that people with a goodness-of-health score of c or above are given the drug and that people with a score below c are not given the drug. Under this setup, what meaningful inferences can be made about the effect of the drug on weight loss? Discuss the parameter of interest and the methods used to estimate this parameter. What assumptions are required to estimate this parameter?

b) Consider the following mechanism for treatment assignment: Before being assigned to treatment or control, each patient rolls a 6-sided die. If the die comes up as 1, 2, 3, or 4 and the patient has a score of c or above, that patient takes the weight loss drug, otherwise they receive a placebo. If the die comes up as a 5 or 6 and the patient has a score below c, that patient takes the weight loss drug, otherwise they receive a placebo. Suppose that the die roll is unknown to the experimenter.

   What does the estimate in a) measure now? Can the parameter of interest in part a) still be estimated? If so, how? If not, why not?

c) Suppose that the effect of the drug is thought to be the same for all patients with goodness-of-health scores within the interval $(c-5, c+5)$. Suppose that patients with scores below c are ineligible to receive the weight loss drug. Patients with scores of c or above are given an appointment to receive the drug. The drug is administered only once during the trial, and only at this appointment. Some patients fail to arrive at their appointment. Under this setup, discuss at least two types of inference possible for measuring the effect of the drug on weight loss? Discuss the parameters of interest and the methods used to estimate these parameters. What assumptions are required to estimate these parameters? Which estimate will be larger (in absolute value)?

III. Sensitivity Analysis

Suppose that there is a study with a total of $2n$ subjects. Exactly n of these people are smokers. A height and weight are measured for each subject. Suppose that there are enough people so that the joint distribution of the heights and weights is extremely close to a multivariate normal distribution. The researcher wants to test whether smoking affects 40-yard dash times.

a) A statistician notices some imbalance in the average weight and height between the smokers and the non-smokers. To fix the imbalance, the statistician matches smokers to non-smokers by matching on the Mahalanobis distance with height and weight covariates (with replacement, nearest neighbor). Will the differences in average height and average weight between the smokers and matched non-smokers be as small or smaller as they were before matching? Why or why not?

b) Suppose instead that all subjects in the study are twins. For each set of twins, one twin is a smoker and one twin is a non-smoker, and both twins in each set have the same height and weight. In his analysis, the statistician believes that the smoking sibling in a twin pair is essentially random, though he concedes that some unobserved trait may help explain a twin's propensity for being a smoker.

   For each set of twins s, let $(1, s)$ denote the twin that smokes, and let $(2, s)$ denote the non-smoking twin. Let T_{is} denote random smoking indicators; $T_{is} = 1$ if the ith unit in the sth twin pair smokes, $i = 1, 2$. For this study, for each pair s, the smoking indicators are observed to be $T_{1s} = 1$ and $T_{2s} = 0$. The statistician models the probability that a subject smokes in the following way:

$$\log\left(\frac{P(T_{is}=1)}{1-P(T_{is}=1)}\right) = \alpha + \kappa_1 h_{is} + \kappa_2 w_{is} + \gamma u_{is} \tag{1}$$

   where h_{is} and w_{is} are the height and weight of twin (i, s), and u_{is} is the value of an unobserved covariate for that twin. The statistician also assumes that any subject cannot influence any other subject to smoke or not smoke (smoking is independent across all subjects).

   Show that, under this model, the probability that subject $(1, s)$ is a smoker is:

$$P(T_{1s} = 1 | T_{1s} + T_{2s} = 1) = \frac{e^{\gamma u_{1s}}}{e^{\gamma u_{1s}} + e^{\gamma u_{2s}}} \tag{2}$$

   Hint: Use $P(A|B) = P(A \cap B)/P(B)$, and find an expression for

$$\frac{P(T_{1s} = 1 \cap T_{2s} = 0)}{P(T_{1s} = 0 \cap T_{2s} = 1)} = \left(\frac{P(T_{1s}=1)}{1-P(T_{1s}=1)}\right)\left(\frac{P(T_{2s}=0)}{1-P(T_{2s}=0)}\right)$$

c) Suppose that $0 \leq u_{is} \leq 1$ and that $\gamma > 0$. Find an upper and lower bound (sharper than just 1 and 0) for the probability $P(T_{1s} = 1 | T_{1s} + T_{2s} = 1)$. Denote these bounds by p_s^+ and p_s^- respectively. Do the same for $P(T_{1s} = 0 | T_{1s} + T_{2s} = 1)$. Comment, in one sentence, on how these bounds change if $\gamma < 0$.

d) Let y_{is} denote the 40-yard dash time of subject (i, s) in milliseconds. Let Z_s denote an indicator variable for the smoker having the faster 40-yard dash time: $Z_s = 1$ if and only if the smoking twin has a faster 40-yard dash time than the non-smoking twin. Let d_s denote the rank of $|y_{1s} - y_{2s}|$; higher ranks denote larger absolute values. Assume there are no ties between y_{1s} and y_{2s} within any twin pair s, and that $|y_{1s} - y_{2s}| \neq |y_{1t} - y_{2t}|$ for all distinct twin pairs s, t.

   The Wilcoxon signed rank statistic is:

$$W = \sum_{s=1}^{n} d_s Z_s.$$

   Let Z_s^+ and Z_s^- be independent and identically distributed bernoulli random variables (or indicator variables) with $P(Z_s^+ = 1) = p_s^+$ and $P(Z_s^- = 1) = p_s^-$. Consider the following random variables:

$$\begin{aligned} W^+ &= \sum_{s=1}^{n} d_s Z_s^+ \\ W^- &= \sum_{s=1}^{n} d_s Z_s^- \end{aligned}$$

Show that, under the null hypothesis that smoking does not effect 40-yard dash times, the following property holds:

$$\mathbb{E}(W^-) \leq \mathbb{E}(W|T_{1s} + T_{2s} = 1) \leq \mathbb{E}(W^+)$$

e) In fact, it can be shown that under this null hypothesis, for any a:

$$P(W^- \geq a) \leq P(W \geq a|T_{1s} + T_{2s} = 1) \leq P(W^+ \geq a) \tag{3}$$

Discuss, in about 3 -5 sentences or so, how property (3) can be exploited to test the exact null of no treatment effect.

f) [BONUS QUESTION] Prove property (3).

IV. Media Bias

For this problem, you will compare the research design from three papers studying the effects of media bias on political attitudes and choices. The first paper is "The Fox News Effect", by S. DellaVigna and E. Kaplan (DVK), and can be found here `http://sekhon.berkeley.edu/causalinf/papers/DellaVignaFoxNews.pdf`. The second paper is "Exploiting a Rare Shift in Communication Flows to Document News Media Persuasion", by J. Ladd and G. Lenz (LL), and can be found here `http://sekhon.berkeley.edu/causalinf/papers/LaddLenzBritish.pdf`. And the third paper is "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions", by A. Gerber, D. Karlan, and D. Bergan (GKB), and can be found here `http://sekhon.berkeley.edu/causalinf/papers/GerberNewspapers.pdf`.

Please write a page or two addressing the following questions:

a. Compare the identification strategies of the three papers. Which strategy do you find the most convincing? The least? Why?

b. Given the different types of interventions being studied (e.g., biased media exposure v. change in media bias, television v. newspaper media, etc), in what sense are the findings across these three studies 'comparable'? Do these studies give us useful information to test the same theoretical claim or different theoretical claims?

c. Imagine at a future point in time, Fox News expanded to every major cable and media market in the US. Would we expect to see a similar media effect as measured by DVK as a result of this national expansion? Why or why not? What would be an analogous type of issue in the studies conducted by LL and GKB? Do any of the three studies seem more robust to this type of issue than the others?

d. Which paper do you find the most interesting, weighting both the scope and significance of the effect being estimated, as well as the *external* and *internal* validity of the respective estimates? Generally speaking, which study is *most* informative about the substantive impact of media bias on public opinion or vote choice?

V. Data and Matching

For this problem, you will perform several matching exercises using Ladd and Lenz's "Exploiting a Rare Shift" data. The unit of observation is the individual respondent in a UK election survey, and the treatment under study is whether an individual is a reader of a newspaper that switched its party endorsement from Tory to Labour in the run-up to the 1997 election. The main outcome is change in Labour party vote support between 1992 and 1997. To control for confounding, the authors condition on a number of covariates (listed in Table 3 and Table 1A of their paper) that may predict both readership and party voting behavior.

The Ladd and Lenz data is available here: `http://sekhon.berkeley.edu/causalinf/data/midterm.dta`. The variables are described in the following file: `http://sekhon.berkeley.edu/causalinf/data/midterm_codebook.xsls`

For parts (a) - (e) below, be sure to explicitly set seeds to ensure that GenMatch recovers reproducible results, i.e. `set.seed` in general, and in GenMatch `unif.seed`, `int.seed`.

a. Estimate the causal effect of being a typical reader of a newspaper that switched party endorsement (from Tory to Labour) on the *change* in Labour party vote support between 1992 and 1997. In doing so, select a set of relevant covariates to condition on. In matching, first use a custom loss function and then use GenMatch's default loss function. Provide some justification for your custom loss function. Choose the matched dataset with the best balance on the relevant covariates. Are the media effects on voting you estimate significantly different from zero? What are the mean differences in change in party vote suport you recover after matching? What are the three *worst* balanced covariates in this best-matched dataset? What are the standardized mean differences across matched treated and control on these three covariates?

b. Using the best-matched dataset from part (a), restrict your analysis of the matched-pairs to include only those treated individuals who are habitual readers of a newspaper that switched its party endorsement. Check balance on this 'restricted' dataset using `MatchBalance`. Does balance change considerably in this dataset, compared with that recovered in (a)? Now, use GenMatch to match on the same covariates used in (a), utilizing habitual readers as the treatment indicator. Does balance in this matched dataset improve compared to that recoverd in the 'restricted' matched dataset? What media effects do you recover in these two matched datasets? Are these different from that found in (a)?

c. Choose one matched dataset from (a) or (b) that you think is the most convincing in recovering conditional exchangeability (for either habitual or typical readers), and conduct two robustness tests of the conditional exchangeability assumption. The first robustness check should be a Rosenbaum sensitivity test using the `rbounds` package in R. The second robustness check either should be a post-matching parametric bias adjustment on the matched data (e.g., a probit regression including covariates and treatment to model the outcome on the matched data), or a placebo test of the effect of treatment on a prior party vote outcome before and after matching. What is the Γ magnitude of confounding due to an unobserved covariate in the Rosenbaum sensitivity test at which the estimated treatment effect is indistinguishable from zero? How does this Γ compare to the imbalance recovered on the three worst-balanced covariates in the best-balanced dataset in (a)? Are these robustness tests convincing that conditional exchangeability holds?

d. Repeat the analysis in part (a), this time using the *level* of Labour party vote support in 1997 (rather than change in vote support). Is this estimate consistent with the one recovered in (a)? Is this causal estimate more or less persuasive than the difference-in-difference estimate you recovered in (a)? Overall, what do we learn about the effects of media from this analysis?

e. [BONUS QUESTION] Fully replicate the regression analysis in Table 1A (excluding the 1992 instrument column), on both the *level* and *change* in party vote support. That is, do the bivariate analysis, the exact matching on the same covariates used by Ladd and Lenz, and the GenMatch analysis on the same coveriates, and also perform linear adjustment on each matched data set. Can you replicate the table exactly? If not, which parts can you replicate exactly? How confident are you that this analysis is recovering an unbiased estimate of the persuasive effect of media on vote choice behavior? Does replicating the analysis change your assessment?