Causal Inference Midterm Exams

Instructions:

You will have one class period to complete this exam. Please show all work (do not skip steps during proofs) and cite relevant assumptions and theorems to justify your analytical proofs. Good luck.

Problem 1

(24 points)
The Weak Foundation for Political Research (WFPR) decides to fund a study to assess the effect of watching the Fox News channel on presidential approval ratings. With that purpose, the researchers of the WFPR select 80 voters that currently do not have access to Fox News (assume they reside in a remote area) and randomly provide 40 of them with satellite access to the Fox News cable channel (i.e. the treated) while excluding the other 40 from accessing Fox News (i.e. the controls). One year after the intervention, the presidential approval rating of all the voters in the study is recorded. Each voter is asked: "Do you approve or disapprove of the way Barack Obama is handling his job as President?" The results are displayed in the table below.

	Approve	Disapprove
Treated	20	20
Control	24	16

a)

The investigators from the WFPR estimate the effect of the treatment on approval ratings using a difference in means between treated and controls. What do the WFPR results suggest about the effect of Fox News on approval ratings? (Note: We are not asking you to account for statistical significance here. Simply compute the point estimate.)

Solution:

$$
\begin{aligned}
\bar{Y_T} &= 20/(20+20) = .5 \\
\bar{Y_C} &= 24/(24+16) = .6 \\
\bar{Y_T} - \bar{Y_C} &= -.1 \\
t &= \frac{-.1}{\sqrt{Var(Y_T)/N_T + Var(Y_C)/N_C}} \\
&= \frac{-.1}{\sqrt{((1/N_T-1)*\sum_{i=1}^{N_T}(Y_{Ti}-\bar{Y_T})^2/N_T + ((1/N_C-1)*\sum_{i=1}^{N_C}(Y_{Ci}-\bar{Y_C})^2)/N_C}} \\
&= \frac{-.1}{\sqrt{((1/39)*\sum_{i=1}^{40}(Y_{Ti}-.5)^2)/40 + ((1/39)*\sum_{i=1}^{40}(Y_{Ci}-.6)^2)/40}} \\
&= \frac{-.1}{\sqrt{((1/39)*[20*(-.5)^2+20(.5)^2]/40 + ((1/39)*[16*(-.6)^2+24(.4)^2]/40}} \\
&\approx -.89
\end{aligned}
$$

Since the t statistic is less than 1.96 in absolute value, we cannot reject the null that Fox News had no effect at the 5% significance level.

b)

A group of researchers from Confusion State University (CSU) are not convinced by the results of the study so they ask the WFPR for the data. When they disaggregate the data by party identification (i.e. whether voters identify themselves as Republicans or Democrats prior to the study) they obtain the following table:

	Approve	Disapprove
Republicans:		
Treated	12	18
Control	3	7
Democrats:		
Treated	8	2
Control	21	9

The researchers from CSU similarly estimate the effect, but separately for Democrats and Republicans using the data in the table. Reproduce their analysis. What do the CSU results suggest about the effect of Fox News on approval ratings?

Solution:

$$\begin{aligned}
\bar{Y}_{TR} &= 12/(12+18) = .4 \\
\bar{Y}_{CR} &= 3/(3+7) = .3 \\
\bar{Y}_{TR} - \bar{Y}_{CR} &= .1 \\
t_R &= \frac{.1}{\sqrt{Var(Y_{TR})/N_{TR} + Var(Y_{CR})/N_{CR}}} \\
&= \frac{.1}{\sqrt{((1/N_{TR}-1) * \sum_{i=1}^{N_{TR}}(Y_{TRi} - \bar{Y_{TR}})^2/N_{TR} + ((1/N_{CR}-1) * \sum_{i=1}^{N_{CR}}(Y_{CRi} - \bar{Y_{CR}})^2)/N_{CR}}} \\
&= \frac{.1}{\sqrt{((1/29) * \sum_{i=1}^{30}(Y_{Ti} - .4)^2)/30 + ((1/9) * \sum_{i=1}^{10}(Y_{Ci} - .3)^2)/10}} \\
&= \frac{.1}{\sqrt{((1/29) * [18 * (-.4)^2 + 12(.6)^2]/30 + ((1/9) * [7 * (-.3)^2 + 3(.7)^2]/10}} \\
&\approx .56
\end{aligned}$$

We fail to reject the null.

$$
\begin{aligned}
\bar{Y}_{TD} &= 8/(8+2) = .8 \\
\bar{Y}_{CD} &= 21/(21+9) = .7
\end{aligned}
$$

$$
\begin{aligned}
\bar{Y}_{TD} - \bar{Y}_{CD} &= .1 \\
t_D &= \frac{.1}{\sqrt{Var(Y_{TD})/N_{TD} + Var(Y_{CD})/N_{CD}}} \\
&= \frac{.1}{\sqrt{((1/N_{TD}-1) * \sum_{i=1}^{N_{TD}} (Y_{TDi} - \bar{Y_{TD}})^2/N_{TD} + ((1/N_{CD}-1) * \sum_{i=1}^{N_{CD}} (Y_{CDi} - \bar{Y_{CD}})^2)/N_{CD}}} \\
&= \frac{.1}{\sqrt{((1/9) * \sum_{i=1}^{10} (Y_{Ti} - .8)^2)/10 + ((1/29) * \sum_{i=1}^{30} (Y_{Ci} - .7)^2)/30}} \\
&= \frac{.1}{\sqrt{((1/9) * [18 * (-.8)^2 + 12(.2)^2]/30 + ((1/9) * [7 * (-.7)^2 + 3(.3)^2]/10}} \\
&\approx .63
\end{aligned}
$$

We fail to reject the null.

c)

The researchers from CSU send their results to the WFPR asking for an explanation. They receive an answer from the WFPR which explains that the discrepancy in the results between the WFPR and CSU is due to a small sample problem. People at CSU do not know what to think. Why are the results different? Comment.

Solution:

The estimated effect in the full sample was negative, but within each party, the estimated effect is positive. In neither case do the tests return statistically significant results. The divergence in results is due to the fact that the sample was not large enough to ensure equal assignment probabilities within each party. Among Democrats, 1/3 of the individuals were treated, and among Republicans, 2/3 were treated. The relatively lower approval rating among Republicans combined with the higher observed probability (by chance due to the small sample) of Republicans being treated tilts the approval rating in the overall treated group downwards, resulting in a negative estimate for the treatment effect in the whole sample, despite these positive sub-group estimates. In large samples, or in a blocked design, assignment probabilities would be roughly equal within subgroups which would avoid this problem.

Problem 2

(24 points) Are the following statements true or false? Justify your answers for credit.

1. To test the null hypothesis of no difference in two sample means, one needs a sufficiently large sample in order to use the standard normal table for determining the critical region of the test. However, if we assume that the random variables are each exactly normally distributed, and if the assumption is correct, then we can use the standard normal table for the test even if we only have a small sample size.

   False. The first sentence is correct. The second sentence is wrong because we need to use the t table instead of the z table in that case. (3pt)

2. A key identification assumption in a randomized experiment is that there is no interference between units (the SUTVA assumptions). To make one's inference robust to the violation of that assumption within clusters of units, one must correct standard errors using methods such as cluster-robust standard errors and block bootstrap.

   False. Cluster robust SEs cannot correct for violations of SUTVA. They can only address the possible correlation between potential outcomes of the units belonging to the same clusters, which is a different from interference (i.e. treatment on unit A directly affecting the outcome of unit B belonging to the same cluster). (3pt)

3. Matching is often preferred to regression adjustment for causal inference in observational data because the latter relies on conditional ignorability while the former does not.

   False. Matching requires conditional ignorability. Its advantage over typical regression adjustment is less dependence on functional form assumptions and extrapolation. (3pt)

4. A researcher is analyzing the effect of a treatment in a randomized experiment and uses a two-sample t-test with unequal variances to reject the null hypothesis of zero average treatment effect. The researcher could have tested the same null hypothesis with a randomization test like Fisher's exact test to avoid any large sample approximations.

   False. A randomization test will test the sharp null of no effect instead of the zero average treatment effect. (3pt)

5. Assume we have a non-randomized binary treatment variable and we use the naive difference in means as an estimator of the ATT. If we are able to make the assumption that there is no baseline bias (i.e. $E[Y_0|D=0] = E[Y_0|D=1]$) our estimate of the ATT is still potentially biased due to the possibility of differential treatment effect bias.

   False. False. By focusing on the ATT as our estimand of interest, we need only worry about baseline bias. Differential treatment effect bias is irrelevant to estimating the ATT. (3pt)

6. Matching with replacement is less biased than matching without replacement, but is also less efficient.

   True. Re-using matches is inefficient because we are using less information to derive estimates, but this approach helps with bias because the matches we obtain are better (closer) matches to the treated units. The converse is true for matching without replacement. (3pt)

Problem 3

(24 points)
A recent study examined the income gains from migration using data on New Zealand (NZ) and Tonga. Tongans need a NZ visa to migrate to NZ. A lottery is used to issue visas to applicants, and only lottery winners get a visa. However, some lottery winners do not migrate after all. Here is a summary of the data:

	Group size	Mean weekly income (in NZ$)
Lottery winners		
Migrants	65	424.5
Non-migrants	55	81.1
Total	120	194.7
Lottery losers		
Migrants	0	
Non-migrants	78	104.1
Total	78	104.1

With these data, a research group interested in the effects of migration on earnings presents a variety of estimators.

1. First, they present the difference between the mean weekly income of lottery winners that migrate and the mean income of lottery losers. One critic in the audience claims that since Tongans self-select into migration, this comparison is misleading. The researchers argue, however, that the lottery generated random assignment, and thus the criticism is invalid. Who is correct and why? What is it that the researchers are estimating?

   The critic is right. The lottery is randomly assigned, but the migration is determined by unobservables (when doing the simple mean difference). The estimated quantity is of no substantive interest.

2. The researchers then present the difference between the mean income of all lottery winners and all lottery losers, and argue that this consistently estimates the average treatment effect of migration on earnings. The critic again claims that this is incorrect, now because not everyone who won the lottery migrated, and it is thus incorrect to attribute this difference to migration. Who is correct and why? What is it that the researchers are estimating?

   The critic is right. While the described difference in means is consistent for the average treatment effect of winning the lottery on earnings, it is not the ATE of migration on earnings.

3. Finally, the researchers show the difference in migration rates between the lottery winners and the lottery losers. They argue that this is a consistent estimate for the average treatment effect of the lottery on migration. The critic, once again, complains that this is incorrect because the lottery losers do not even have the option of migrating. Who is correct and why? What is it that the researchers are estimating?

   The researchers are right. The inability of Tongans to migrate in the event of losing in the lottery is irrelevant; the random nature of the lottery enables the estimation of the average treatment effect of the lottery on migration.

Problem 4

(28 points)

A researcher wants to figure out once and for all if giving students high school vouchers for private schools would significantly improve educational outcomes, in this case measured by students' standardized test scores. The researcher receives a grant to conduct a randomized controlled trial. Assume that the researcher's experiment does not suffer from any of the standard problems we typically encounter in implementing randomized trials, namely there are no compliance issues. This means that everyone who received a private school voucher goes to private school and everyone who did not receive a voucher (those people in the control group) does not go to private school. After running the experiment and collecting all of her data, the researcher estimates the following model to determine the treatment effect:

$$Y_i = \alpha + \beta_1 \text{voucher} + \beta_2 \text{race} + \beta_3 \text{parent's education}_i + \beta_4 \text{parent's income}_i + \epsilon_i$$

where Y is the student's test score at the end of the experiment, "voucher" is a dummy variable for whether a student received a voucher or not, and the other variables are "controls" which predict test scores in particular and other educational outcomes in general. The researcher finds that her estimate for β_1 is positive and significant and sends in her work for publication.

Many months later she receives her reviews. One reviewer really doesn't like the researcher's model. The reviewer argues, since the experiment has no compliance problems, and because it is a randomized controlled trial, the researcher should only estimate the following model:

$$Y_i = \alpha + \beta_1 \text{voucher} + \epsilon_i$$

The researcher estimates the recommended model, but finds that while her point estimate for β_1 has not changed, her standard errors have grown and her estimate is no longer significant.

a)

What might have been the reviewer's rationale for recommending the reduced model?

Solution:

The reviewer is probably concerned that the researcher may have been mining for a statistically significant treatment effect by trying out different sets of control variables.

b)

Prove analytically why using the reduced model does not cause the estimated treatment effect to change (in expectation) but could inflate the standard errors. You can assume homoskedasticity, and to simplify things, you can assume demeaned variables in order to suppress the intercept.

Solution:

The estimated treatment effects will not differ in expectation whether other pre-treatment covariates are included or not. To see this, suppose the true model of Y (using demeaned variables) is:

$$Y_i = \beta_1 \text{voucher} + \mathbf{Z}\delta + \epsilon_i$$

where $\mathbf{Z}$ is a matrix of control variables. Now imagine we estimate the following reduced model:

$$Y_i = \tilde{\beta}_1 \text{voucher} + v_i$$

Let X represent an indicator of whether a voucher was given. The estimated treatment effect will be:

$$\begin{aligned}
\hat{\tilde{\beta}} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta_1 + \mathbf{Z}\delta + \epsilon_i) \\
&= (X'X)^{-1}X'X\beta_1 + (X'X)^{-1}X'\mathbf{Z}\delta + (X'X)^{-1}X'\epsilon_i \\
E[\hat{\tilde{\beta}}|X] &= E[\beta_1|X] + E[(X'X)^{-1}X'\mathbf{Z}\delta|X] + E[(X'X)^{-1}X'\epsilon_i|X] \\
&= \beta_1 + \delta(X'X)^{-1}E[X'\mathbf{Z}|X] + (X'X)^{-1}E[X'\epsilon_i] \\
&= \beta_1
\end{aligned}$$

where the last step follows from the fact that the covariance between the randomly assigned treatment X and the matrix $\mathbf{Z}$ is zero by design; and the covariance between X and the error term is zero by the zero conditional mean assumption. So the reduced model will estimate the true average treatment effect in expectation. Excluding the controls will make no difference, on average. The estimated variance of the estimated treatment effect under homoskedasticity in the reduced model is:

$$\frac{\hat{\sigma}_v^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

where $\hat{\sigma}_v^2$ is estimated as:

$$\frac{\hat{v}'\hat{v}}{n-k-1} = \frac{\hat{v}'\hat{v}}{n-2}$$

where $\hat{v} = y - X\hat{\beta}$ is an nx1 vector of residuals and k is the number of independent variables in the model.

The estimated variance of the estimated treatment effect under homoskedasticity in the full model is:

$$\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2 (1 - R_Z^2)}$$

where R_Z^2 is the R^2 from the regression of X on all the covariates in $\mathbf{Z}$. Since X is randomly assigned and does not covary with the variables in Z, R_Z^2 will be approximately zero, leaving the denominator of the estimated variance unchanged. With the addition of four covariates in the model, the numerator of the variance will now be:

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{n-k-1} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-6}$$

If n is reasonably large, the denominator of this expression will basically remain the same, but since the additional covariates explain variation in y, $\hat{\epsilon}'\hat{\epsilon} < \hat{v}'\hat{v}$ (i.e. the regression surface "fits" the data better and the residuals are smaller). Therefore, the standard error on the treatment effect in the full model will likely be smaller and we will be more likely to reject the null hypothesis that the treatment effect is zero.

c)

Which model should we prefer? Why? Is this generally true, or are there exceptions?

Solution:

We should prefer the full model because the estimates will be more precise. One exception to this general approach would be if one of the covariates is causally posterior to the treatment. In that case, the inclusion of the post-treatment variable would bias the estimated treatment effect.

d)

In the paper, the researcher also estimated:

$Y_i = \alpha + \delta_1 \text{voucher} + \delta_2 \text{race} + \delta_3 \text{voucher} \times \text{race} + \delta_4 \text{parent's education}_i + \delta_5 \text{parent's income}_i + \epsilon_i$

Using this model, how can the researcher estimate the treatment effect and its standard error for each race sub-group? (Assume only two race groups, i.e., race$\in \{0, 1\}$)

Solution:

The treatment effect can be recovered through the marginal effect of receiving a voucher:

$$\frac{\partial Y_i}{\partial voucher} = \delta_1 + \delta_3 \text{race}$$

This quantity can be estimated with $\hat{\theta} = \hat{\delta}_1 + \hat{\delta}_3 \text{race}$. The researcher can estimate its standard error by: $\hat{Var}(\hat{\theta}) = \sqrt{\hat{Var}(\hat{\delta}_1) + \hat{Var}(\hat{\delta}_3) \times race^2 + 2 \times race \times \hat{Cov}(\hat{\delta}_1, \hat{\delta}_3)}$

s

Problem 5 - EXTRA CREDIT (If you have time . . .)

(5 points)

1. Imagine you run a randomized experiment. Let D_i be the treatment indicator for individual i, Y_{di} be the potential outcome given treatment assignment, and Y_i be the observed outcome. You feel good about yourself and estimate the ATE using a regression of $Y_i = \alpha + \tau D_i + \varepsilon_i$.

   Now your good old friend Mr Critic comes by and starts pestering you: "Your estimate is not valid because the regression assumes a constant treatment effect τ. In reality, individuals react differently to the treatment and so this assumption is violated." Do you agree? Why or why not? If you agree, can you think of a valid alternative?

   Solution:

   This criticism is invalid. The ATE is identified by the difference in means estimator while allowing for full treatment effect heterogeneity τ_i. The regression is just a convenient way to estimate this.

2. Mr Critic goes on and with a triumphant smile proclaims: "In fact, you can test for heterogeneous treatment effects. If you reject the null hypothesis that $Var[Y_{1i}] = Var[Y_{0i}]$ this implies rejection of the null hypothesis that the treatment effects are constant." Do you agree or not? Explain why or why not (preferably using a bit of algebra).

   Solution:

   Note that by definition $\tau_i = Y_{1i} - Y_{0i}$ so we have that

   $$Var[Y_{1i}] = Var[Y_{0i}] + Var[\tau_i] + 2Cov[Y_{0i}, \tau_i]$$

   so the equality $Var[Y_{1i}] = Var[Y_{0i}]$ holds if

   $$Var[\tau_i] = -2Cov[Y_{0i}, \tau_i]$$

   Under the null that τ_i is constant, both sides of this equation are zero since the covariance between a variance and a constant is zero. Thus rejecting the null hypothesis of equal variances means rejecting the null hypothesis of constant treatment effects. Note that there is also the remote possibility that the covariance is negative and offsets the variance in τ_i which would also mean the equality of variances holds.