Instructions: Please Follow Exactly

Please email your write-up to Professor Grimmer by 6pm on June 8th. Late work will **not** be accepted, so please **start early and plan to finish early**. Remember that this exam may take longer to finish than you might expect.

- Your exam must be typed. No handwritten sections please.
- The solution to each question should be presented with tables and figures included in the document and referenced within the text. I should not have to guess which figure/table you are talking about. Furthermore, your code file should be clearly commented, and I should be able to reproduce all of your plots, tables and results. You will not receive full credit for plots, tables and results that I cannot replicate from your code.
- Your write-up should explain very clearly exactly what you did. In your write-up, all results should be presented clearly in tables and figures, and **not** in code. You should assume in your write-up that I cannot understand code, so you must describe and interpret the results in words. If I do not understand what you did from your write-up, I will not go through your code to try to figure it out.
- Your solution to each of the questions should start on a **new page**, and **must be kept within the page limits indicated for each question**. However, it is not necessary to utilize the total allotment of pages in order to receive full points, and extraneous writing can hurt your score. In general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions I have asked, you will not receive points when you demonstrate knowledge about questions I have not asked, and you will **lose** points when you make inaccurate statements (whether or not they relate to the question asked). Any text beyond the page limit will be ignored.
- If specific code instructions have not been given for a particular problem, then you can use any available function in R. If you cannot successfully write code that works, describe in your write-up what you were attempting to do and explain how you would have incorporated these results into your solution.
- I will answer questions of clarification during the week. I will not answer statistical or computational questions relating to the exam or any of the course materials until after the exam is handed in. If you have a question **do not post it publicly on Piazza**. Instead please send an email to me. If I deem your question to be clarifying in nature, I will strip all identifying information from the email, and then send it to the course list.

- You may use your books and notes to answer the questions below. However, you are to work on the exam *alone*: **you are not allowed to forward, copy, discuss, or share any content of this exam with anybody (even after the exam is due), or receive help on it. If I receive evidence that students have violated the academic code, for example by forwarding the exam and or working on any portion of the exam in groups, all students in such groups will receive a failing grade.**

1. Treatment Effects (3-page limit)

   Suppose that we are interested in understanding the income of undocumented immigrants in California. Let Y denote income.

   (a) First, we are given the true distribution of the income of undocumented immigrants in California. A journalist asks you to provide your best guess of the income of an immigrant she interviewed (the interviewee did not state his income). Without any information about the interviewee, what is your best guess of his income? Show formally that your answer follows from minimizing the expectation of the squared difference between your guess and the truth.

   (b) Now, suppose that we also have information on the voter registration in each immigrant's county of residence. In particular, we know X, the proportion of registered Republican voters in each county. The journalist now asks you to guess the interviewee's income given the proportion of registered Republican voters in the county of residence of her interviewee (given $X = x$). What is your best guess? Show formally that your answer follows from minimizing the expectation of the squared difference between the truth and an arbitrary function of X, $h(X)$. Hint: you might want to think about the Law of Iterated Expectations.

   (c) California Proposition 187 was a 1994 ballot initiative that sought to exclude undocumented immigrants from public services (it passed and then was later found unconstitutional). Imagine that some counties implemented the initiative by aggressively excluding undocumented immigrants from schools and hospitals, and that other counties did not implement the initiative. Suppose that each county decided whether or not to implement, and assume that immigrants cannot move between counties.

   Define D as an indicator for whether or not a given county chose to implement Prop 187. Your journalist friend notes that immigrant incomes are lower in counties where $D = 1$ than in counties where $D = 0$ and asks whether she should interpret this as an effect of Prop 187. To answer her, write an expression for the difference in expected incomes conditional on D, and decompose this difference into the causal effect of implementing Prop 187 and a bias term. How do you interpret the overall difference in expected incomes in this context?

   (d) Suppose we now know that implementation of Prop 187 is as good as random once we condition on X (county partisanship). Or, more formally, assume that

   $$\left(Y_{0i}, Y_{1i}\right) \perp D_i \Big| X_i \tag{1}$$

   Assume also that there is no level of partisanship for which the probability of implementing Prop 187 is zero (or one). In other words, assume

   $$0 < Pr(D_i = 1|X_i) < 1 \tag{2}$$

   Write an expression for the expected average treatment effect of Prop 187 on income Y for counties with a given partisanship level ($X = x$). Given assumptions (1) and (2), is $\tau(x) = E(Y_i(1) - Y_i(0)|X = x)$ identified? If so, derive an estimable expression for $\tau(x)$. For credit, be very explicit about the assumptions that justify each step of the derivation.

   (e) Given the expression derived in (d), how would you estimate $\tau(X = x)$? Explain how assumption (2) affects the estimation.

(f) How would you estimate τ_{ATE} (that is, the treatment effect averaged across all values of X)?

(g) Now suppose that we relax the assumption that immigrants cannot move from county to county. How does this affect our ability to estimate τ?

2. **Panel** (4-page limit)

There is a vivid debate in public policy about the effect of housing allowance programs, which are sometimes used to support disadvantaged low-income households with their rent payments. A sticking point in these debates is the extent to which particular programs would actually increase the rent paid by low-income families – presumably a measure of the space and quality of housing – or whether participants simply spend the money on other things, such as clothing, food, cars, televisions, and the like.

To inform these debates, researchers in the mid-90s conducted a large scale experiment in two cities, Phoenix and Pittsburgh. Low-income families in both cities were first recruited and then randomly assigned to a treatment or control group. The experiment lasted two years. In the first year (i.e. the pre-experimental period) families in both groups received no allowance payments. In the second year (i.e. the post-period) families in the treatment group were given experimental monthly lump sum payments, which varied by family income and family size. These payments were simply extra income that the families could spend as they pleased. Families in the control group just received a $10 one-time participation reward. The outcome of interest is the average monthly rent paid by the families.

The dataset `house.dta` contains the data for both years in a standard long panel format. The list of variables is as follow:

- id: unique family identifier
- treated: indicator for being in the treatment group (1 = treated family, 0 = control family). This is constant for each family (i.e. for treated families, this variable is 1 in both the pre-period and post-period).
- post: indicator for second year, i.e. 1 = post-period, 0 = pre-period
- rent: average monthly rent paid in dollars
- calcpay: average monthly experimental payments to the families during the post period (includes $10 one-time participation reward to control families)
- inc: annual family income in dollars
- race indicator: 1 = nonwhite, 0 = white
- phoenix indicator: 1 = Phoenix, 0 = Pittsburgh
- famsiz: family size
- hed: head of family's level of education (in years)
- femh: 1 = female head of family, 0 = male head of family

(a) Social experiments are costly, compared to analysis of observational data. Thus it is important to evaluate the informational gains derived from the experimental approach. Using only the control families, fit the following regression to estimate the effect of additional family income on monthly rental payments (here and for all other estimates in this problem, make sure you pick an appropriate method to estimate standard errors and briefly justify your choice):

$$rent = \beta_0 + \beta_1 inc + \beta_2 famsiz + \beta_3 phoenix + \beta_4 post + \beta_5 (phoenix \cdot post) + \beta_6 hed + \beta_7 femh + \beta_8 race + \varepsilon$$

Report the results. What do the estimates imply about the effect of one additional dollar in annual income on average monthly rents paid?

(b) Again using only control-families, re-estimate the model in first-differences and report the results. What do the estimates imply about the effect of one additional dollar in annual income on average monthly rents paid? Do the results differ from the ones in model (a)? If so why? Which estimate do you prefer? Be precise about the specific identification assumptions.

(c) Again using only control-families, estimate the following model (where $t \in \{0, 1\}$ indexes the pre- and post-period year respectively):

$$rent_{t=1} = \beta_0 + \beta_1 rent_{t=0} + \beta_2 inc_{t=0} + \beta_3 famsiz_{t=0} + \beta_4 phoenix + \beta_5 hed_{t=0} + \beta_6 femh_{t=0} + \beta_7 race + \varepsilon$$

Report the results. What do the estimates imply about the effect of one additional dollar in annual income on average monthly rents paid? Do the results differ from the ones in models (a) and (b)?

(d) Use your three estimates about the marginal effect of income from (a), (b), and (c) and combine them with the information about the average monthly experimental payments received in the treatment group during the post period (*calcpay*) to predict the amount of additional monthly rental payment we would expect for the treatment group if the income of these families were increased by the experimental amount. In other words, from the regression results, how much would you expect the rent of the treatment group to increase on average, given the monthly rent subsidy payments? (Remember to accommodate the fact that *inc* is annual while *calcpay* is monthly). Report your three estimates with their standard errors and 95% confidence intervals.

(e) Now using all families (treated and control), consider the following model:

$$rent = \beta_0 + \beta_1 post + \beta_2 treated + \beta_3 (treated \cdot post) + \varepsilon$$

Before estimating the model, precisely describe the meaning of each coefficient, referencing any specific identification assumptions necessary for that interpretation. Given the study design, what should your expectations be regarding the general sign and magnitudes of each of the coefficients?

(f) Now, report the results. What do the estimates imply about the effect of the treatment? How do the results compare with those from (d) above? Which ones do you prefer and why? To economize on space you may want to present the estimates of the models from (f) and (h) (below) in a single table.

(g) Now consider the following model:

$$rent = \beta_0 + \beta_1 treated + \varepsilon$$

Imagine estimating this model twice. In the first estimation you use the entire set of data, producing the estimate $\hat{\beta}_1^A$. In the second estimation, you use only the data from the post period, producing the estimate $\hat{\beta}_1^B$. Would you expect $\hat{\beta}_1^A$ to be larger, smaller, or roughly the same as $\hat{\beta}_1^B$? Why?

(h) Now again using all families, estimate the following model:

$$\begin{aligned} rent &= \beta_0 + \beta_1 post + \beta_2 treated + \beta_3 (treated \cdot post) + \beta_4 phoenix \\ &+ \beta_5 (phoenix \cdot post) + \beta_6 (treated \cdot phoenix) + \beta_7 (treated \cdot post \cdot phoenix) + \varepsilon \end{aligned}$$

Reports the results, and precisely describe the meaning of each coefficient (where appropriate, you may combine up to two coefficients for interpretation). What do the estimates imply about the effect of the treatment?

(i) How much do you trust the results from the experiment? Use the data and appropriate checks from previous questions to inform your judgment.

3. **LATE** (4-page limit)

A famous recent study reports the results of a canvassing experiment conducted in Beijing on the eve of a local election. The researchers randomly assigned dorm rooms on the campus of Peking University to a treatment and control group. Canvassers attempted to contact the students in the dorm rooms assigned to the treatment group and encouraged them to vote using scripted mobilization information. No contact with the control group was attempted. The researchers then subsequently measured whether the students voted in the election or not. The data from the experiment is provided in the file `beijing.dta`.

The variables are as follows:

turnout	voted in 2003 elections
treat2	treatment group indicator
contact	contact made indicator (script delivered)
deptid	department ID
dormid	dorm ID
id	student ID

(a) Estimate the ITT effect and its standard error. How do you interpret this estimate? Very briefly, under what assumptions is this a valid estimate? Do these assumptions seem plausible?

(b) Estimate the LATE and its standard error. How do you interpret this estimate? Very briefly, under what assumptions is this a valid estimate? Do these assumptions seem plausible?

(c) What is the compliance ratio and pattern of compliance in this experiment? How does this affect your interpretation of the results?

(d) Very briefly, explain whether each of the following statements is true or false for the case of one-sided non-compliance (assuming that the experiment satisfies the other necessary assumptions).

i. If the ITT is negative, the LATE must be negative.

ii. The smaller the compliance ratio, the larger the LATE.

iii. One cannot identify the LATE if no one in the experiment receives the treatment.

iv. You will never know which of your subjects are Compliers and which of your subjects are Never-Takers.

(e) Imagine the situation where, at every dorm room they visited, even those where no one answered, canvassers left a leaflet encouraging students to vote. How would this affect the validity of the ITT and the LATE estimate from above?

(f) Assume that this leaflet raised the probability of voting by two percentage points among both Compliers and Never-Takers. In other words, suppose that the treatment group's turnout rate would have been two percentage points lower had the leaflets not been distributed. Write down a model of the expected turnout rates in the treatment and control groups, incorporating the average effect of the leaflet.

(g) Given this assumption about the ATE of leaflets, estimate the LATE of canvassing for compliers.

4. **Selection on Observables** (4-page limit)

   The data for this problem are `Exam2018sim.csv`. In this problem we will compare methods for estimating causal effects. In the csv file, you'll find a 5000 x 5 matrix with the following columns:

   (a) treatment: each observation's treatment status

   (b) response: the value of each unit's response variable

   (c) cov1, cov2, cov3: the values of three covariates for each unit

   The treatment was not randomly assigned, but the assumption of selection on observables is met. Furthermore, the data were generated such that the treatment effect for all units in these simulated data is 2.5.

   In the following, we will first define our quantities of interest, and then we will attempt to estimate the treatment effect using several methods.

   (a) Using the potential outcomes framework, define:
      i. The Average Treatment Effect (ATE)
      ii. The Average Treatment Effect on the Treated (ATT)
      iii. The Average Treatment Effect on the Control (ATC)
      iv. Express the ATE in terms of the ATT and ATC

   (b) Estimate each observation's propensity score using a probit regression of the treatment on the three covariates (just the covariates and intercept alone; no interactions, polynomials, etc.). We will call this new variable `est.prop`.

   (c) Using the estimated propensity scores, we will create a matched sample for estimating the ATC.
      i. Using only the propensity scores, perform nearest-neighbor matching *without* replacement in order to pair each control unit with one treated unit.
         - Note: You can use the `Matching` package to implement this. For extra credit, you can also code your own matching algorithm. If you choose the latter, be sure to verify that it works properly.
      ii. Compare the balance in the covariates before and after matching. Does the matching improve balance?
      iii. Compare the balance in propensity scores before and after matching. Does the matching improve this balance much? Why or Why Not?

   (d) We are now in a position to estimate the effect of the intervention. For each of the following specifications, run a linear regression in order to estimate the treatment effect and its standard error. Report your results. In addition, explicitly denote the implied treatment effect of interest for each specification.
      i. Using all units: **response** $\sim$ **treatment**
      ii. Using all units: **response** $\sim$ **treatment + cov1 + cov2 + cov3**
      iii. Using all units: **response** $\sim$ **treatment + est.prop**
      iv. Using the matched sample from (c): **response** $\sim$ **treatment**

   (e) Comparing the four specifications, comment on how well each specification yielded the causal effect of interest. For each specification, explain why it is or might be the case that it performed relatively well/poorly.