

PHS2000B Exam 2 (50 points)

Econometrics, Missing Data & Meta-Analysis

Spring 2021

Welcome to our 2nd exam! As a reminder, you have 24 hours to complete and upload your exam to the Assignments tab on Canvas. You may refer to your notes and any books you would like, but please do not discuss the exam with any other students from class or your tutor. By submitting your exam on Canvas, you affirm that the work you are submitting is yours alone and that you have not discussed the questions or responses with anyone other than the instructional team.

We wish to emphasize that you are more than welcome to handwrite your responses on paper, paste them into a word document, or submit your write up as a scanned PDF file (please submit as a single file if possible). If you choose to handwrite, please circle or otherwise indicate your final answer. Alternatively, you are welcome to type your responses in your program of choice (e.g. Word, TeX, markdown, etc.) Please note that any math or drawings you want to include can be typeset into the document or can be pasted in as an image. You will NOT be graded based on the format you choose to submit your exam.

If you have clarifying questions for the instructional team, please email the whole team and we will get back to you as soon as possible. During our regular 11:30 am – 1:00 pm lecture period, you may also log on to the [Zoom meeting](#) and get an immediate response from the instructional team.

There is a [Google doc](#) where we will be posting any clarifications that arise from questions we receive during the exam. If we post a clarification, we will also post an announcement referring you to the Google doc.

Good luck!

1 Instrumental Variables/RCTs (12 points)

Imagine it is 1995 and you are part of a research team that has just won a grant to study the effect of hospital length of stay (LOS) following delivery on risk of hospital readmission for newborns in the three months after delivery. Lengths of postpartum stay in the United States had decreased from an average over 5 days pre-1970 to a low of 1.1 days on average in 1995 and the medical community is concerned about potential negative effects of this trend.

Your team decides to perform a small randomized controlled trial to answer this question by recruiting a cohort of pregnant individuals who are randomized to either 1 day ($Z = 1$) or 3 day ($Z = 0$) hospital lengths of stay after a healthy vaginal delivery. (In 1996, Congress passed the Newborns' and Mothers' Health Protection Act, which required health insurance providers to extend hospital stay benefits to at least 48 hours after vaginal delivery.)

Questions 1-3 are based on this randomized controlled trial design and the following results.

Randomized to 1 day				Randomized to 3 days			
		Postpartum LOS				Postpartum LOS	
		1 day	3 days			1 day	3 days
Rehospitalization	Yes	8	0	Rehospitalization	Yes	0	2
	No	22	0		No	0	28

1. Draw a DAG depicting this study design including the randomized treatment assignment. Remember to define any variables you use. (1 point)
2. Based on the data in Table 1, calculate and interpret the intention-to-treat (ITT) effect. (2 points)
3. What are two requirements that must be true for the ITT you calculated in question 2 to be equal to the average treatment effect (ATE) for the effect of postpartum hospital length of stay on risk of rehospitalization? (1 point)

Your team is concerned that the randomized controlled trial consists of a very small sample selected from a single hospital and will fail to generalize to a wider population. To address this issue, you decide to perform an observational study for the effect of postpartum hospital length of stay on the risk of hospital readmission for newborns in three months post-delivery using medical records data. Your team is worried about a host of unmeasured confounders of the association that are not captured in administrative data. To allay these concerns, you suggest an instrumental variable analysis using day of week of delivery as an instrument for postpartum hospital length of stay with the idea being that deliveries at the end of the week (e.g., Friday/Saturday) will have longer length of stay compared to earlier in the week (e.g., Monday/Tuesday) because hospital billing offices close on Fridays at 5 PM and do not reopen until Monday mornings.

Questions 4-8 are based on this instrumental variable design and the following results.

Monday/Tuesday				Friday/Saturday			
		Postpartum LOS				Postpartum LOS	
		≤ 2 days	> 2 days			≤ 2 days	> 2 days
Rehospitalization	Yes	1,575	1,565	Rehospitalization	Yes	1,129	945
	No	14,354	26,949		No	9,956	30,124

4. Your colleague runs a first stage linear model (identity link) in which postpartum LOS is regressed against day of delivery and obtains the following output. Is the relevance assumption satisfied? Is day of delivery a strong instrument? What is the concern with weak instruments in IV estimation? (Low_LOS = 1 for ≤ 2 days, Low_LOS = 0 for > 2 days; MT = 1 for Monday/Tuesday delivery, MT = 0 for Friday/Saturday delivery) (2 points)

Call:

```
lm(formula = Low_LOS ~ MT, data = df)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-0.3584 -0.3584 -0.2630  0.6416  0.7370
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.262964    0.002245  117.16  <2e-16 ***
MT           0.095450    0.003133   30.46  <2e-16 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.4608 on 86595 degrees of freedom

Multiple R-squared: 0.0106, Adjusted R-squared: 0.01059

F-statistic: 928.1 on 1 and 86595 DF, p-value: < 2.2e-16

5. Explain the exclusion assumption in the context of this study and provide one plausible example of how the exclusion assumption could be violated in this study. (1 point)
6. Explain the independence assumption in the context of this study and provide one plausible example of how the independence assumption could be violated in this study. (1 point)
7. Based on the data in Table 2, calculate an IV estimate of the effect of hospital LOS on rehospitalization and provide a causal interpretation. (2 points)
8. What assumption is needed to interpret your effect estimate from Question 7 as a LATE? Describe an individual who would violate this assumption in this study. (2 points)

2 Regression Discontinuity (12 points)

In the United States, minimum legal drinking ages varied widely before 1975, when several states began to raise their drinking ages in response to drunk driving fatalities. In 1984, the National Minimum Drinking Age Act required states to raise their age of possession to 21, or else lose 10% of federal highway funds. Most states responded by raising both the age of possession and the minimum legal drinking age to 21.

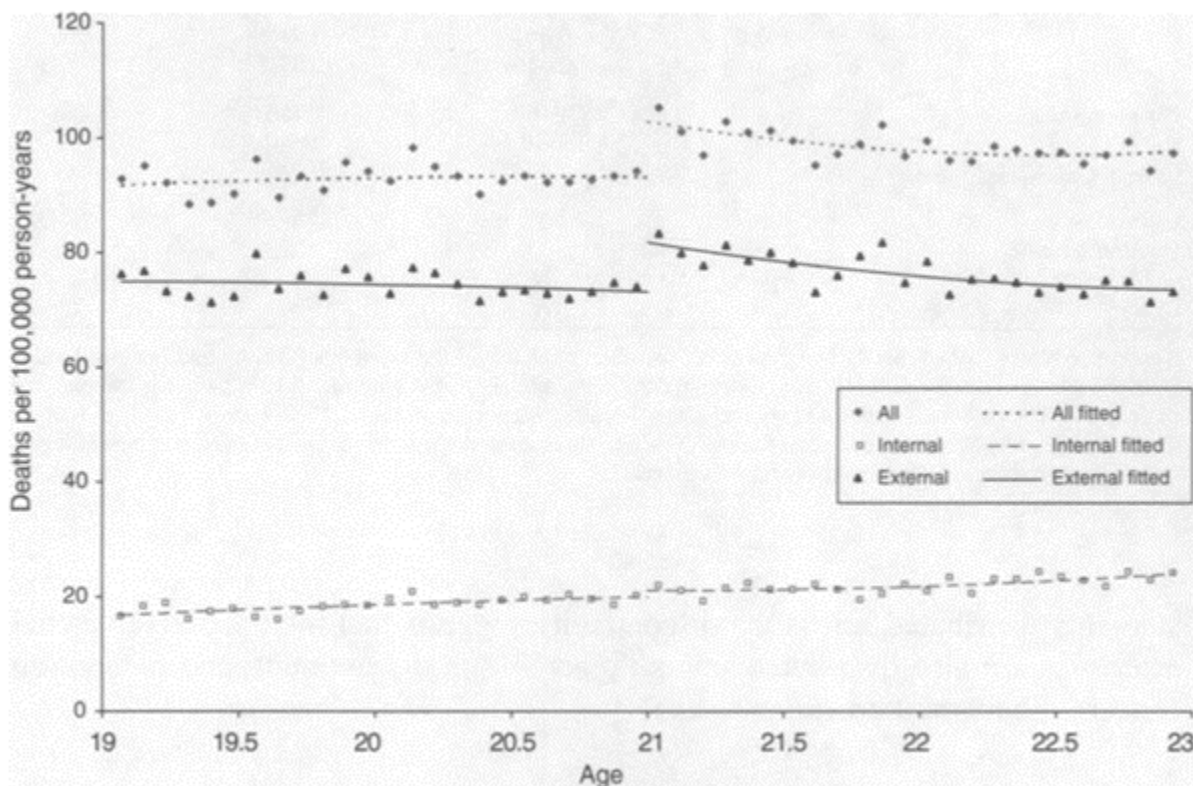
While the 21-year minimum legal drinking age remains **highly popular**, it is not without critics. The “age of majority” (i.e. the threshold of adulthood as recognized or declared in law) in most states is 18, while the minimum legal drinking age is 21, **causing some critics to observe** that young adults can vote, drive, and enlist in the military before they are viewed by the law as responsible enough to drink a beer. Others argue that the law is widely disregarded, with many young adults drinking before they are old enough to do so legally. As a **former college president wrote in the New York Times**, “the overwhelming response of young people has been, not compliance, but contempt for the law.” The effects of minimum legal drinking age laws are therefore of significant interest, both for public health researchers, economists, and policy-makers and for the public at large. So far, the evidence suggests that the minimum legal drinking age of 21, despite the criticisms mentioned above, **saves lives and protects health**. Motor vehicle crashes saw significant declines when minimum ages were raised to 21, reducing alcohol-related mortality in young drivers. And while young adults do drink before turning 21, the overall effect of minimum legal drinking age laws has been a reduction in drinking, even among individuals *above* the minimum age. Because the minimum drinking age of 21 is a widespread policy of significant interest, econometric studies have exploited a discontinuity in drinking habits just above and just below the 21-year-old threshold. The following questions are based on work by **Carpenter and Bobkin (2009)** which uses a regression discontinuity design to explore the causal effect of alcohol consumption on mortality at the threshold of the minimum drinking age.

1. To overcome endogeneity, researchers can randomize exposure in experimental studies or attempt to adjust for confounding variables in an observational study. Briefly explain why these approaches are not appropriate in this context, or why they would not necessary yield an unbiased causal estimate. **(2 points)**
2. As one of the critics of the minimum drinking age law who is quoted in the prompt argues, many people under the age of 21 drink, despite it being illegal to do. If the critic is correct in saying that disregard for the minimum legal drinking age is the “overwhelming” response of individuals under 21, 1) what would this imply about the discontinuity of drinking behaviors at the threshold and 2) how might this cause a challenge to fuzzy RD estimation in this study? In particular, what are the implications of this for the relevance assumption? **(4 points)**
3. Recall that in a fuzzy RD, the estimand will yield the complier average causal effect (or the CACE) under the assumption of monotonicity, or no defiers at the threshold. In one or two sentences, briefly describe who a defier would be in the context of this study. **(2 points)**
4. Econometricians using regression discontinuity designs may be concerned with manipulation of the forcing variable at the threshold, as well as measurement error if people’s exposure is self-reported. Consider the following:
 - The minimum legal drinking age is well publicized
 - There can be significant disciplinary, social, and legal consequences for individuals who are caught drinking under the age of 21.

- Data for this study were obtained from the National Health Interview Survey, which is a general longitudinal survey of health, and from the National Center for Health Statistics' mortality data files, which contains a registry of all deaths in the U.S. from 1997 to 2004. The mortality data include the decedent's date of birth and date of death.

In two to three sentences, briefly discuss what manipulation of the forcing variable would mean in the context of this study and whether the facts above suggest it is likely to be present in this study (Note: be careful in distinguishing manipulation of the forcing variable from other problems for RD studies.) (2 points)

The authors find that individuals just over age 21 report drinking on ~21 percent more days than individuals just under 21 years of age. They find a similar discontinuity in heavy drinking days. Satisfied that the alcohol consumption is discontinuous at the threshold, the researchers use the discontinuity to explore three outcomes of interest: all-cause mortality, mortality due to external causes (homicides, suicides, motor vehicle accidents, etc.), and mortality due to internal causes. They present the following graph, which shows lines fitted above and below the threshold for these three outcomes.

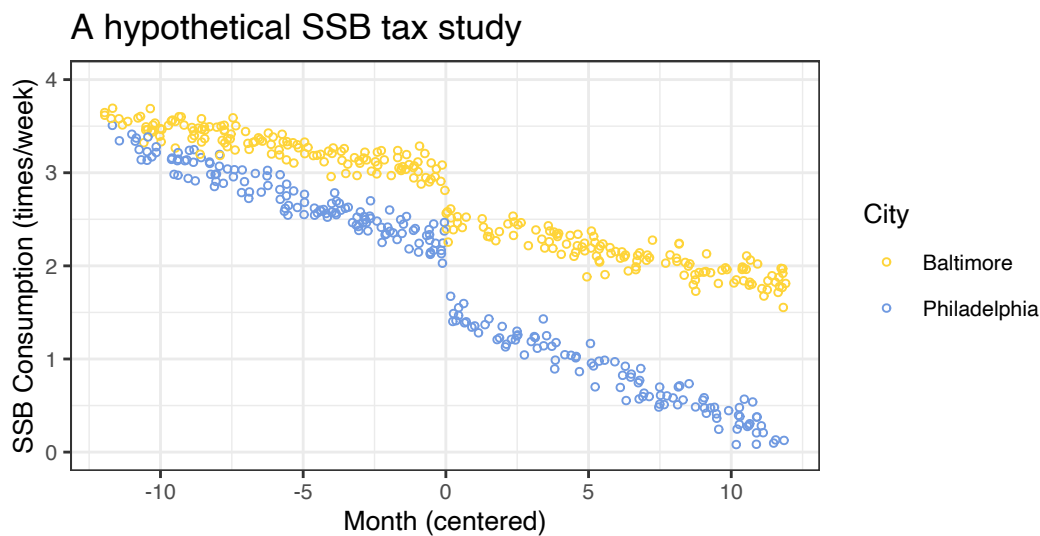


5. Based on the information in the prompt and in the graph, what conclusions do you draw from this graph and what might they imply regarding the impacts of alcohol on mortality? (2 points)

3 Interrupted Time Series/Difference in Differences (12 points)

Sugar-sweetened beverages (SSBs) are the main source of added sugars in the American diet: in 2013-2014, 60.7% of children and 50.0% of adults consumed SSBs on a given day according to dietary recall data from the National Health and Nutrition Examination Survey (NHANES).¹ To reduce consumption, some cities in the US have enacted taxes on sugar-sweetened beverages to raise prices and discourage purchasing. You are a researcher at the Philadelphia Department of Public Health charged with **evaluating the effect of an SSB tax that was implemented on January 1, 2017**.

As part of your work, you've collected weekly consumption data (number of SSBs consumed per week) on 500 individuals - 250 from Philadelphia, PA (*tax, i.e. treated*) and 250 from Baltimore, MD (*no tax, i.e. control*) - for 12 months before and after the tax was implemented. **These data are plotted below (with a centered version of your month variable):**



1. You plan on implementing an **controlled interrupted time series (CITS)** approach to estimate the effect of the Philadelphia SSB tax on SSB consumption frequency. You have variables for week of measurement (**week**), an indicator for whether this measurement was before or after 01/01/2017 (**post**), an indicator for whether an individual lived in Philadelphia (**trt**), and their weekly consumption frequency (**ssb**). Write out the CITS model you would fit to this data that includes terms for a tax-attributable level change and a tax-attributable slope change in SSB consumption frequency. (2 points)
2. Write down the regression parameters from your CITS model in question 1 that correspond to the tax-attributable level change in SSB consumption frequency, and the tax-attributable slope change in SSB consumption frequency. (1 point) What two key assumptions are needed to identify these as causal effect? Describe them within the context of the study. (2 points)
3. A data analyst runs your model in R, and they email you an **incomplete** snippet of the output (shown below). What are your estimate(s) of the causal effect you wrote down in question 2? Interpret within the context of the study. (3 points)

¹See Bleich et al. (2017)

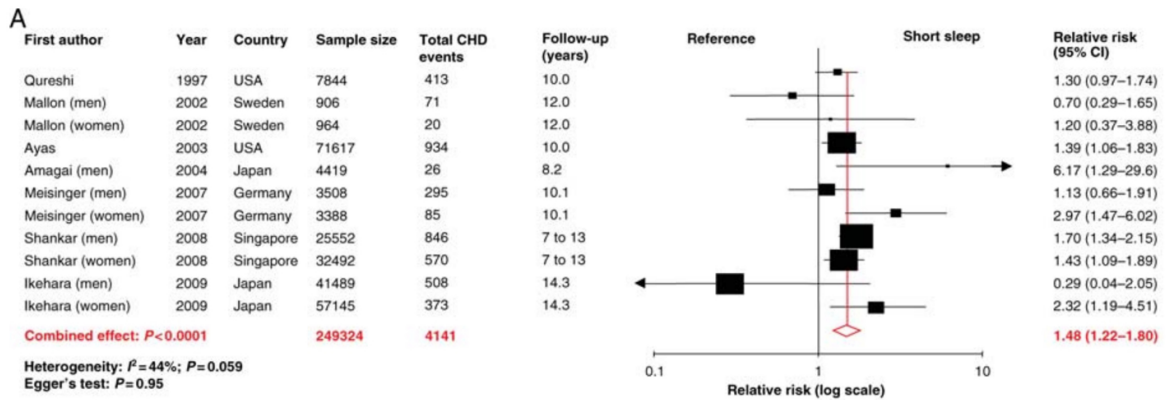
```
## Coefficients (**abbreviated!**) :
##           Estimate Std. Error t value Pr(>|t|)
## ...           ...           ...           ...
## week:post    -0.006095   0.003686  -1.654   0.0989 .
## week:trt     -0.047896   0.003756 -12.750 < 2e-16 ***
## post:trt     -0.194403   0.037083  -5.242 2.36e-07 ***
## week:post:trt 0.000000   0.005446  -1.623   0.8999
## ...           ...           ...           ...
```

4. A colleague suggests that people may consume SSBs more frequently during the summer months when it's warmer, and less frequently during the winter months. What challenge does this present to the analysis, and how could you address this problem? (2 points)
5. Yet another colleague suggests you implement a **difference-in-differences** approach, since this is what has been done in some research studies of SSB taxes previously. Describe what the **common shocks assumption** and **parallel pre-trends** mean in the context of the study. (2 points)

4 Missing Data/Meta-Analysis (14 points)

You want to assess the association between ambient noise pollution and sleep quality in Boston, Massachusetts. Luckily, you had unlimited funding, so you provided personal noise dosimeters (a device used to measure noise exposure) to all study participants. Participants were requested to carry their dosimeters with them for 24 hours prior to filling out a questionnaire. The questionnaire collected data about sleep quality, personal characteristics such as age, lifestyle factors, and other medical conditions.

1. Suppose that age is a predictor of noise exposure and sleep quality. You also realize that older study participants were more likely to forget to carry around their dosimeter compared to younger study participants, and thus data on ambient noise exposure is more likely to be missing for older study participants. Assume that there are no other predictors of missing dosimeter data and that there are no other confounders of the exposure-outcome association.
 - a. Identify the missingness mechanism for ambient noise exposure. (1 point)
 - b. Based on the description, draw a DAG to represent a Complete Case Analysis (CCA) in this scenario. Use RA to represent the missingness indicator for exposure. (1 point)
 - c. Under what condition would a CCA in this scenario yield an unbiased estimate of the effect of ambient noise on sleep quality? (1 point)
2. Upon further examination, you see that about 20% of the data is missing. Your colleague wants to use CCA to address the missingness scenario presented in question 2. What is a reason that you might argue in favor of using multiple imputation instead? (1 point)
3. You decide to use multiple imputation to address the missingness scenario in question 2. You include a set of variables in the imputation model that you were not planning on including in your main analysis. Your colleague is nervous that failing to include this additional set of variables in your main analysis will lead to biased estimates. Is this the case? Why or why not? (2 points)
4. Suppose that occupation and number of hours worked are confounders of the association between noise exposure and sleep quality. You then discover that several study participants who work long hours (students, medical professionals, etc.) were less likely to respond to questions about occupation and number of hours worked. Identify the missingness mechanism at play in this scenario. (1 point)
5. In the missingness scenario described in question 5, would using inverse probability weighting to address the missingness mechanism yield an unbiased estimate? Why or why not? (2 points)
6. Your friends ask you why it is important to study factors that may influence sleep quality. In response, you show them a meta-analysis from 2011 that assessed the association between short duration of sleep and risk of coronary heart disease (CHD). Below is the forest plot of the risk of developing or dying of CHD associated with short duration of sleep compared with the reference group.



- Interpret the pooled relative risk value of 1.48. (1 point)
- In this meta-analysis, authors reported an I^2 statistic. What does the I^2 represent? What kind of meta-analysis is this? (2 points)
- The I^2 statistic is 44% with a p-value of 0.059. Interpret the magnitude and significance of the I^2 in the context of this meta-analysis. (2 points)