PHP 2610 Problem Set 2

Due: October 14 by 11:59pm

Instructions: Please upload your answer to the Canvas course page as a pdf file. You can submit your answers in a separate pdf file (please be sure to properly mark the question number to your responses), or you can work on this pdf file, scan it, and upload it to the Canvas course page.

Late or missed assignments: Problem sets and the final report must be turned in online at or before the posted due date. Every one day (24 hours) of delay will result in a ten point (out of 100) downgrade.

Question 1 (15 points)

Experiments with cells suggest that chromium and nickel can damage DNA. Werfel et al. (1998) used 1:1 matched pairs for a welder exposed to chromium and nickle and an exposed control, matching for age and smoking. The full description about this study is described in the following paper, but it is not necessary to read the paper to answer the questions below.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics, 63*(2), 456-464.

The data (`erpcp`) can be found in the `sensitivitymv` package in `R`.

```
library(sensitivitymv)
data(erpcp)
```

Q1.1. (5 points) The following `R` code calculates the p-value of testing the sharp null hypothesis of no treatment effect in a matched observational study with a sensitivity parameter $\Gamma = 1$.

```
senmv(erpcp, gamma = 1)
```

Based on the results after running the code above, which of the following statements are correct?

a. Without no confounding by age and smoking, the p-value of rejecting the null of no treatment effect is less than 0.01.

b. Without unmeasured confounding but with confounding by age and smoking allowed, the p-value of rejecting the null of no treatment effect is less than 0.01.

c. With unmeasured confounding, the p-value of rejecting the null of no treatment effect is larger than 0.01.

d. Without unmeasured confounding, the p-value of rejecting the null of no treatment effect is larger than 0.01.

e. Without unmeasured confounding, the p-value of rejecting the null of no treatment effect is less than 0.01.

Answer:

Q1.2. (5 points) The following `R` code calculates the p-value of testing the sharp null hypothesis of no treatment effect in a matched observational study with a sensitivity parameter $\Gamma = 3$.

```
senmv(erpcp, gamma = 3)
```

Based on the results after running the code above, which of the following statements are correct?

a. A p-value of 0.018 is the lower bound of the one-sided p-value.
b. Given a specified Γ, one unit might be three times as likely as another to be exposed to chromium and nickel due to unmeasured confounding.
c. Given a specified Γ, one unit might be three times as likely as another to have managed DNA due to unmeasured confounding.
d. Given a specified Γ, one unit might be three times as likely as another to be a smoker due to unmeasured confounding.
e. In this case, we can reject the null of no treatment effect at type-I error $\alpha = 0.01$.

Answer:

Q1.3. (5 points) When $\Gamma = 4$, we cannot reject the sharp null of no treatment effect at the Type-I error $\alpha = 0.05$ level.

a. TRUE
b. FALSE

Answer:

Question 2 (20 points)

Suppose that one study found the association between obesity and cardiovascular diseases as RR = 1.25 (95% CI: [1.04, 1.46]). Let RR_{AU} denote the maximum risk ratio for any specific level of the unmeasured confounders comparing individuals with and without obesity, conditional on observed covariates. Let RR_{UY} denote the maximum risk ratio for cardiovascular disease incidence comparing any two categories of the unmeasured confounders within each treatment group, conditional on observed covariates.

Q2.1. (7 points) In that case, what is the E-value for the RR estimate (round to 2 decimal places)?

Answer: 1.81

Q2.2. (7 points) In the above case, what is the E-value for the lower confidence interval limit (round to 2 decimal places)?

Answer: 1.24

Q2.3. (6 points) What is the correct interpretation of the above results?

a. If both of RR_{UY} and RR_{AU} are greater than than 1.81, then an unmeasured confounder cannot explain way the observed RR.
b. If both of RR_{UY} and RR_{AU} are smaller than than 1.81, then an unmeasured confounder cannot explain way the observed RR.
c. If either of RR_{UY} and RR_{AU} are greater than than 1.81, then an unmeasured confounder cannot explain way the observed RR.
d. If either of RR_{UY} and RR_{AU} are greater than than 1.81, then observed confounders can explain way the observed RR.
e. If either of RR_{UY} and RR_{AU} are greater than than 1.81, then both of observed and unobserved confounders cannot explain way the observed RR.

Answer:

Question 3 (15 points)

Nyugen et al. (2016) showed the protective effect of education against risk of dementia in older adulthood. In this study, the level of education is measured by educational attainment operationalized as self-reported years of schooling. (https://doi.org/10.1016/j.annepidem.2015.10.006; reading this article is not necessary to answer the homework questions)

Q3.1. (5 points) If you were able to design a study to investigate the effects of education on the risk of dementia, what would be the ideal way to conduct such a study, independent of ethical or feasibility concerns?

a. Randomize subjects to receive encouragement to have each levels of years of schooling
b. Randomize subjects to have each level of years of schooling.
c. Randomize subjects to receive treatments for dementia versus no treatments
d. Randomize subjects to receive encouragement to take treatments for dementia versus no encouragement.
e. None of the above
f. All of the above

Answer:

Nyugen et al. (2016) used several different variables as an instrumental variable. Among those, three independent single-nucleotide polymorphisms (SNPs) that have been previously identified as genome-wide significant predictors of education attainment in a large genome-wide association study.

Q3.2. (5 points) What does the exclusion restriction assumption imply here?

a. Each of the three SNPs should have a non-zero association with education attainment
b. Each of the three SNPs should have a non-zero association with the risk of dementia
c. Each of the three SNPs should be randomized
d. The three SNPs do not have a direct effect on the risk of dementia
e. The three SNPs should be independent of any confounders between education attainment and dementia

Answer:

Q.3.3. (5 points) Suppose that A denotes a continuous scale showing the risk dementia; B denotes one of the three independent SNPs; and C denotes the measured educational attainment. Please choose the equations that can be modelling together to make the two stage least squares model:

a. $A_i \sim \beta_0 + \beta_1 \times C_i$ and $C_i \sim \alpha_0 + \alpha_1 \times B_i$
b. $A_i \sim \beta_0 + \beta_1 \times B_i$ and $C_i \sim \alpha_0 + \alpha_1 \times B_i$
c. $A_i \sim \beta_0 + \beta_1 \times B_i$ and $B_i \sim \alpha_0 + \alpha_1 \times C_i$
d. $A_i \sim \beta_0 + \beta_1 \times C_i$ and $B_i \sim \alpha_0 + \alpha_1 \times C_i$
e. Models in (a) and (b)

Answer:

Question 4 (15 points)

Let Y denote a continous outcome, Z denote a binary instrumental variable, A denote a binary treatment variable, and X denote the observed confounders. Y^z denotes the potential outcome variable when assigned to $Z = z$ and A^z denotes the treatment variable when assigned to $Z = z$. Suppose that $Y^0, Y^1 \perp\!\!\!\perp Z \mid X$ but $Y^0, Y^1 \not\perp\!\!\!\perp Z$, and that $A \perp\!\!\!\perp Z \mid X$ but $A \not\perp\!\!\!\perp Z$. Demonstrate why the following equation holds using the causal assumptions we learned (please use consistency, ignorability, exclusion restriction, and monotonicity).

$$\frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(A \mid Z = 1) - E(A \mid Z = 0)} \quad = \quad E(Y^1 - Y^0 \mid \text{complier})$$

Answer:

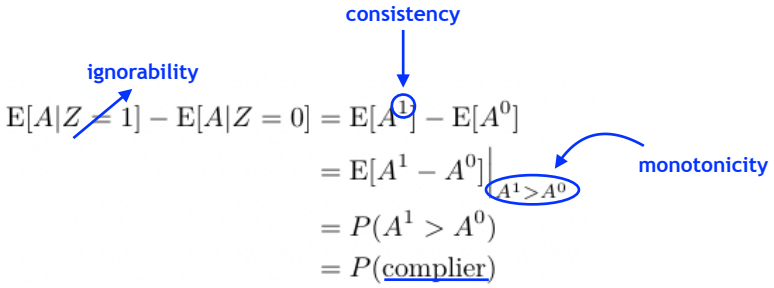
(See next 2 pages)

Numerator:

$$
\begin{aligned}
\mathrm{E}[Y|Z=1] &= \mathrm{E}[Y|Z=1,\ \text{defier}]P(\text{defier}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{complier}]P(\text{complier}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{never-taker}]P(\text{never-taker}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{always-taker}]P(\text{always-taker}|Z=1) \\
&= \mathrm{E}[Y|Z=1,\ \text{defier}]\cdot 0 \quad \longleftarrow \text{no defiers} \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{complier}]P(\text{complier}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{never-taker}]P(\text{never-taker}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{always-taker}]P(\text{always-taker}|Z=1) \\
\mathrm{E}[Y|Z=0] &= \mathrm{E}[Y|Z=0,\ \text{defier}]P(\text{defier}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{complier}]P(\text{complier}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{never-taker}]P(\text{never-taker}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{always-taker}]P(\text{always-taker}|Z=0) \\
&= \mathrm{E}[Y|Z=0,\ \text{defier}]\cdot 0 \quad \longleftarrow \text{no defiers} \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{complier}]P(\text{complier}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{never-taker}]P(\text{never-taker}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{always-taker}]P(\text{always-taker}|Z=0)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0] &= \Big[\mathrm{E}[Y|Z=1,\ \text{complier}]P(\text{complier}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{never-taker}]P(\text{never-taker}|Z=1) \\
&\quad + \mathrm{E}[Y|Z=1,\ \text{always-taker}]P(\text{always-taker}|Z=1)\Big] \\
&\quad - \Big[\mathrm{E}[Y|Z=0,\ \text{complier}]P(\text{complier}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{never-taker}]P(\text{never-taker}|Z=0) \\
&\quad + \mathrm{E}[Y|Z=0,\ \text{always-taker}]P(\text{always-taker}|Z=0)\Big] \qquad \text{equal} \\
&= \mathrm{E}[Y|Z=1,\ \text{complier}]P(\text{complier}|Z=1) \qquad \text{ignorability} \\
&\quad - \mathrm{E}[Y|Z=0,\ \text{complier}]P(\text{complier}|Z=0) \\
&= \Big[\mathrm{E}[Y^{1,A^Z}|\ \text{complier}] - \mathrm{E}[Y^{0,A^Z}|\ \text{complier}]\Big]P(\text{complier}) \qquad \text{consistency} \\
&= \mathrm{E}[Y^1 - Y^0|\ \text{complier}]P(\underline{\text{complier}}) \qquad \text{exclusion restriction}
\end{aligned}
$$

Denominator:



Result: (*LATE / CASE*)

$$\therefore \frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[A|Z=1] - \mathrm{E}[A|Z=0]} = \frac{\mathrm{E}[Y^1 - Y^0|\ \text{complier}]P(\text{complier})}{P(\text{complier})}$$
$$= \mathrm{E}[Y^1 - Y^0|\ \underline{\text{complier}}]$$

Question 5 (3 points)

Suppose that Y is a response variable subject to missingness and X is a vector of always observable covariates. Y is observable if and only if $A = 1$. Then what does the missingness at random (MAR) assumption imply?

a. $Y \perp\!\!\!\perp X$
b. $A \perp\!\!\!\perp Y$
c. $Y \perp\!\!\!\perp X|A$
d. All of the above
e. None of the above

Answer:

Question 6 (2 points)

Suppose that we have prior knowledge about the outcome distribution given the covariates set but are not sure about the missingness mechanisms. Then the IPW estimator is more appropriate than the multiple imputations.

a. TRUE
b. FALSE

Answer:

Question 7 (30 points)

We are going to use the data set `CigarettesSW` which comes with the package `AER` to examine the relation between the demand for and the price of cigarettes. It is a panel data set that contains observations on cigarette consumption and several economic indicators for all 48 continental federal states of the U.S. from 1985 to 1995. We will consider data for the cross section of states in 1995 only.

```
library(AER)
data("CigarettesSW")
```

Please refer to https://www.rdocumentation.org/packages/AER/versions/1.2-9/topics/CigarettesSW for more information about the variables. We will transform the data to obtain deflated cross section data for the year 1995.

```
# compute real per capita prices
CigarettesSW$rprice <- with(CigarettesSW, price / cpi)

#  compute the sales tax
CigarettesSW$salestax <- with(CigarettesSW, (taxs - tax) / cpi)

# generate a subset for the year 1995
c1995 <- subset(CigarettesSW, year == "1995")
```

Let Q_i (`"packs"`) denote the number of cigarette packs per capital sold and P_i (`"rprice"`) denote the after-tax average real price per pack of cigarettes in state i. The instrumental variable we are going to use for $\log(P_i)$ is sales tax (`"salestax"`) measured in dollars per pack.

Q7.1. (5 points) Suppose that we are interested in estimating β_1 in:

$$\log(Q_i) = \beta_0 + \beta_1 \log(P_i) + \epsilon_i$$

Use `lm()` and obtain $\hat{\beta}_1$ (round to 2 decimal places):

Answer: **-1.21**

Q7.2. (5 points) Using the following two linear functions, derive the estimate for β_1 (round to 2 decimal places):

```
fit1 <- lm(log(rprice) ~ salestax, data = c1995)
fit2 <- lm(log(packs) ~ salestax, data = c1995)
```

Answer: **-1.08**

Q7.3. (5 points) Using the following two linear functions, derive the two-stage least squares estimate for β_1 (round to 2 decimal places):

```
fit1 <- lm(log(rprice) ~ salestax, data = c1995)
c1995$pred <- fit1$fitted.values
fit2 <- lm(log(packs) ~ pred, data = c1995)
```

Answer: **-1.08**

Q7.4. (5 points) Please use `ivreg()` from and obtain the two-stage least square estimate for β_1 (round to 2 decimal places):

Answer: **-1.08**

Q7.5. (10 points) Please describe why the estimate from Q.7.1. can produce biased causal effect and why the estimate from Q.7.4. can provide an unbiased causal effect related to the instrumental variable assumptions.

Answer:

The first estimate is biased is due to the indirect effect of "salestax" (instrumental variable) on "packs" (outcome) through "rprice" (treatment). Not accounting for the instrumental variable in the model, in turn, confounds the causal effect of treatment on the outcome. Specifically, this violates the ignobility assumption in that the indicator variable is not independent of the treatment and outcome. On the other hand, the remaining estimates satisfy this assumption in that the indicator variable's effect on treatment ("rprice") is now accounted for, thus removing bias due to the instrument and making "packs" and "rprice" mutually independent of "salestax".