Econ 488, Fourth problem set

2025-02-01

- **Due Date:** 11:59pm on February 6. Submission should be via gradescope (accessible via canvas).
- Other instructions:
  - You may work on this problem set in groups of up to 3 students. One submission per group (via Gradescope), and please indicate your group's members.
  - You are welcome to typeset the solutions in LaTeX or Rmarkdown to receive bonus points.
  - Unless otherwise stated, please justify your answer. Also, try to answer concisely: please limit each response to at most two sentences. **Look out for typos!**

Set-up

In this problem set you will replicate a paper by Arik Levinson, titled "How Much Energy Do Building Energy Codes Save? Evidence from California Houses" (Levinson 2016). The author constructs a dataset of houses (and their occupants) in California, and uses regression to learn the effect of introducing the California Residential Building Code in 1978 on the energy use of households.

The author's empirical strategy is to assume that the conditional independence assumption holds in their setting. We focus on the regression presented in Table 3 Column (3) and Figure 4—somewhat unusually, the author has split the regression results into a table and a figure, since the treatment value is multi-valued. Column (1) of Table A3 in the "Online Appendix" presents the full results for this regression. The outcome variable is:

- $lnBTUgas_i$, indicating (the natural logarithm of) household i's usage of natural gas in British Thermal Units (BTUs) in a single year (2003 for most houses).

Since the author is interested in the effect of the energy efficiency regulations embedded in the California Residential Building Code (specifically the post 1978-building code), he defines the treatment variable as the house year of construction:

- $constYRs_i$, indicating during which years house i was constructed.

$constYRs_i$ is a categorical (factor) variable that takes one of twelve values, depending on the year of construction. We will focus on houses constructed just before and after the introduction of the energy efficiency building code:

- $post_i$, which equals 0 if house i is built between 1975 and 1977 ($constYRs_i = 6$) and equals 1 if house i is built between 1978 and 1982 ($constYRs_i = 7$).

The following code reads in the data file and selects the subset of data of houses built between 1975 and 1982:

```
library(haven)
df0 <- read_dta("Levinson 2016/Levinson 2016/table3-cols12.dta")
df <- df0[df0$constYRS==6 | df0$constYRS==7, ]
df <- df[!is.na(df$lnBTUgas), ]
df$BTUgas <- exp(df$lnBTUgas)
```

Notice in the final line we have created the variable $BTUgas$, which we will use as the outcome variable:

- $BTUgas_i$ indicates household i's usage of natural gas in British Thermal Units (BTUs) in a single year (2003 for most houses).

Given these simplifications, the CIA (no confounders) assumption is

$$(BTUgas_i(0), BTUgas_i(1)) \perp Post_i \mid X_i,$$

where X_i is the following long list of possibly confounding variables:

- "cddzip30_00", Cooling degree-days (100s)
- "hddzip30_00", Heating degree-days (100s)
- "lnsqftK", Log size of house (in sq ft)
- "numroom", number of bedrooms
- "elecboth", whether electric stove
- "remodeled", if house is remodeled
- "lnyrs_res", log yrs resident
- "lnrescnt", log number of residents
- "lnhhinc", log household income
- "nr0_5", number of household members between 0 and 5 years old
- "nr65_99", number of household members older than 65
- "college", indicator for whether household head graduated college
- "anydisabled", indicator for whether any disabled resident
- "hohblk1", indicator for any black residents
- "hohlat1", indicator for any latino residents
- "own", indicator for own home
- "year09", indicator for RASS 2009
- "cecfast", climate zones

On the final page of this problem set I provide code that generates the regression results in Table A3 Column (1) of Levinson (2016), which may be useful for responding to the questions in the problem set. However, for all questions that follow you should use $Post_i$ as the treatment variable and $gasBTU_i$ as the outcome variable.

QUESTIONS

Treatment effect

1. Describe in words the potential outcome. Describe in words and mathematically the unit-level treatment effect.

CIA assumption

2. Explain why the simple difference in means

$$E[BTUgas_i \mid Post_i = 1] - E[BTUgas_i \mid Post_i = 0]$$

   is unlikely to identify the ATE. In your answer, please refer to selection bias or omitted variable bias (i.e., if you claim there there is likely to be selection bias, explain why).

3. Describe in words the CIA assumption.

4. Explain why the conditional independence assumption is far more plausible than the (unconditional) independence assumption in this example. Can you think of a confounding variable not contained in X_i?

COC assumption

5. Express the COC assumption mathematically and in words.

6. In previous problem sets we considered one single control variable, whereas in this application there are over 15 control variables. Explain why, in this application, it is more difficult to verify whether the COC assumption holds.

Various approaches to estimating the ATE

Now you will estimate the ATE using each of the methods we discussed in class.

To answer each question, please provide the code you have used. You do not need to display the results until the final question.

Simple difference in means

As a benchmark, compute the simple difference in mean outcomes between treatment and control. Call this $\widehat{ATE}_{dim}$.

Regression

Estimate the linear model

$$BTUgas_i = \alpha + \beta Post_i + \gamma \cdot X_i + U_i,$$

and store your estimate of β as $\widehat{ATE}_{reg}$. ***Hint:*** *I suggest using the lm function in R. If some of the variables are categorical, be sure to tell R. For the following questions it will be useful to keep the $n \times dim(X)$ matrix of control variables: you can do so by adding the argument 'keepX=TRUE' to the lm function.*

Matching on covariates

For $K = 1, 5, 10$, compute a K-nearest neighbor matching estimator of the ATE via the following steps:

1. For each unit i in your dataset compute a K-nearest neighbor estimate of i's counterfactual outcome by the following steps:
    a. Standardize each column of the matrix of covariates X to have variance equal to 1, by dividing each column by its variance. Denote $Xstd$ as the standardized matrix of covariates. Confirm that each column has variance equal to 1.[1]
    b. Compute $dist(i,j) = \sqrt{\sum_{l=1}^{L}(Xstd_{i,l} - Xstd_{j,l})^2}$, a measure of distance of between i and j's covariates. Here l indexes the L columns of $Xstd$
    c. If $D_i = D_j$, set $dist(i,j)$ to a very large number.
    d. Find the K units with the smallest values of $dist(i,j)$. ***Hint:*** *You can use the 'order' function.*
    e. Set the estimated counterfactual outcome for unit i to the average outcome among the K units determined in the previous step.
2. Use the estimated counterfactual outcome from step 1 to estimate the unit-level treatment effect for each unit i, $Y_i(1) - Y_i(0)$. (Keep in mind that the counterfactual outcome for the treated units is a different potential outcome than the counterfactual outcome for untreated units.)[2]
3. Form the ATE estimator as the mean value of the estimated unit-level treatment effects. Denote this as $\widehat{ATE}_{Knn,x}$

Matching on the propensity score

For $K = 1, 5, 10$, estimate the ATE via matching on the propensity score via the following steps:

1. Compute the propensity score via logistic regression of the treatment variable on the set of control variables. ***Hint:*** *In R, you can implement logistic regression via the 'glm' function. Its usage is essentially the same as 'lm', except you must specify the 'link' function via the argument 'family="binomial"'. To get the estimated propensity scores for each unit, you can use the 'fitted.values' value of the 'glm' output,*

[1] The point of this step is to try to make sure that differences between different parts of the covariates are equally penalized. To illustrate, this step would ensure that it does not matter if income is denominated in dollars or 1,000s of dollars.

[2] Keep in mind that the ULTE are not identified, so it does not make sense to take the estimator in this step very seriously.

and transform them via the cdf of the logistic distribution, 'plogis'. Confirm your estimated p-scores are between 0 and 1.

2. Follow the same steps for constructing $\widehat{ATE}_{KNN,X}$ but use the distance function (now $dist_{ij} = |\hat{p}(x_i) - \hat{p}(x_j)|$). Denote your estmator as $\widehat{ATE}_{Knn,p}$

Inverse probability weighting

Given the estimated propensity scores, compute the ATE via the IPW estimator (given in equation (25) on slide 44 in Lecture 3). Denote your estimator as $\widehat{ATE}_{ipw}$.

Doubly robust estimation

One way to interpret the regression we computed above is that we are approximating the conditional mean function $E[BTUgas_i|Post_i = d, X_i = x]$ as

$$m_{app}(d, x) = \alpha + \beta d + \gamma \cdot x,$$

So then $\beta = m_{app}(1, x) - m_{app}(0, x)$. Then, for the propensity-score matching and IPW estimators, we approximated the propensity score $\Pr(Post_i = d \mid X_i = x)$ as

$$p_{app}(x) = \Lambda\left(\gamma_p \cdot x\right),$$

where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. In this question, we combine both approximating functions to form a 'doubly robust' estimator.

On Slide 49 in Lecture 3, we discussed the doubly robust identifying equations, which were in terms of population level expectations. The sample analogs would be:

$$\widehat{E}[Y(1)] = \sum_{i=1}^{n} \left(\widehat{m}_{app}(1, X_i) + \frac{D_i(Y_i - \widehat{m}_{app}(1, X_i))}{\widehat{p}_{app}(X_i)} \right)$$
$$\widehat{E}[Y(0)] = \sum_{i=1}^{n} \left(\widehat{m}_{app}(0, X_i) + \frac{(1 - D_i)(Y_i - \widehat{m}_{app}(0, X_i))}{1 - \widehat{p}_{app}(X_i)} \right)$$

Where $\widehat{m}(d, x)$ and $\widehat{p}(x)$ are the estimated approximating equations that you generated in the "regression" and "matching on propensity score" sections, respectively.

$$\widehat{m}_{app}(d, x) = \widehat{\alpha} + \widehat{\beta} d + \widehat{\gamma} \cdot x, \qquad \widehat{p}_{app}(x) = \Lambda\left(\hat{\gamma}_p \cdot x\right),$$

Compute each of these sample analogs, and then form the doubly robust estimator as the difference, denote this as $\widehat{ATE}_{dr}$. ***Hint:*** *You could confirm this for yourself, but you can get the fitted values $\widehat{m}(D_i, X_i)$ directly as the 'fitted.values' value of the 'lm' function used for the regression.*

Presenting your results and discussion

7. Present each of your 10 estimators for the ATE. Briefly discuss your results. For comparison, the paper estimates that $BTUgas$ decreases by 5% ($= -0.161 + 0.113$, see Table A3 Column (1)) due to the introduction of building code reform—given the average level of BTUgas before the reform is '`mean(df$BTUgas[df$post==0])` equals 49.809, a 5% decrease corresponds to an average effect of -2.49.

8. Briefly reflect on your experience implementing the different estimators. For example: which were harder/easier to implement, what would you try differently, any other relections?

Appendix: code to replicate Table A3, Column (1)

```
col1 <-  lm(lnBTUgas ~ 1 +
               cddzip30_00 + # Cooling degree-days (100s)
               hddzip30_00 + # Heating degree-days (100s)
                lnsqftK+ # Log size of house (in sq ft)
              numroom+ # number of bedrooms
              elecboth+ # whether electric stove
              remodeled+ # if house is remodeled
                lnyrs_res + # log yrs resident
                lnrescnt + # log number of residents
              lnhhinc+ # log household income
              nr0_5+ # number of household members between 0 and 5 years old
              nr65_99+ # number of household members older than 65
              college+ # indicator for whether household head graduated college
              anydisabled+ # indicator for whether any disabled resident
              hohblk1+ # indicator for any black residents
              hohlat1+ # indicator for any latino residents
               own + # indicator for own home
                year09 + # indicator for RASS 2009
                factor(constYRS) # treatment variable: year of construction
               + factor( cecfast) # climate zones
             , data=df0)
require(magrittr)
```

```
## Loading required package: magrittr
```

```
print(col1$coefficients %>% round(3))
```

```
##        (Intercept)        cddzip30_00        hddzip30_00            lnsqftK
##              3.103             -0.016              0.029              0.377
##            numroom          elecboth          remodeled         lnyrs_res
##              0.016             -0.037             -0.010              0.011
##           lnrescnt           lnhhinc              nr0_5            nr65_99
##              0.131              0.064              0.014              0.058
##            college        anydisabled            hohblk1            hohlat1
##             -0.034              0.080              0.116             -0.029
##                own             year09  factor(constYRS)2  factor(constYRS)3
##             -0.078             -0.161             -0.055             -0.037
##  factor(constYRS)4  factor(constYRS)5  factor(constYRS)6  factor(constYRS)7
##             -0.051             -0.094             -0.113             -0.161
##  factor(constYRS)8  factor(constYRS)9 factor(constYRS)10 factor(constYRS)11
##             -0.196             -0.255             -0.278             -0.313
## factor(constYRS)12   factor(cecfast)2   factor(cecfast)3   factor(cecfast)4
##             -0.350              0.081              0.124             -0.014
##   factor(cecfast)5   factor(cecfast)7   factor(cecfast)8   factor(cecfast)9
##             -0.013              0.119              0.014              0.083
##  factor(cecfast)10  factor(cecfast)11  factor(cecfast)12  factor(cecfast)13
##              0.028             -0.097             -0.050             -0.115
```