Project II

BUSN 5000

Zhengdong, Peng

Academic honesty statement: I have been academically honest in all of my work and will not tolerate academic dishonesty of others, consistent with UGA's Academic Honesty Policy (https://honesty.uga.edu/Academic-Honesty-Policy/).

Signature:Zhengdong Peng

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You *must* submit your completed assignment as both an `.Rmd` file and a single pdf. If you knit to html, save the knitted file as a pdf. Your submission will not be considered complete without both files. To ensure you earn credit for all of your work, be sure to review your pdf before submitting.

Upload both files to eLC by **1159p on Dec 9** using the filenames `sectiontime_lastname_firstname.pdf` (e.g. `935_cornwell_chris.pdf`) and `sectiontime_lastname_firstname.Rmd` (e.g. `935_cornwell_chris.Rmd`).

View the instructions posted to eLC for Dr. Schmutte's project if you need additional guidance on how to save a document knitted to html in a pdf format.

Signing the above honesty statement is required.

Notes:

- Include your first and last names as the `author` in the `yaml`.
- Do not alter the formatting code in this template.
- For questions requiring analytical solutions, you can type them in using markdown math code or you can submit handwritten solutions embedded in the knitted document as clearly readable images.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio.

Problem 1 (15 points)

Consider the cross-section regression model,

$$y_i = \beta_0 + \delta D_i + \beta_1 x_{i1} + u_i, \quad i = 1, \ldots N,$$

where y_i is some outcome of interest, D_i is a binary treatment indicator for unit i and x_{i1} is some unit-specific characteristic that is being controlled in the model.

a. What is the key identifying assumption for interpreting the OLS estimator of δ as a causal effect?
b. Write down the formula for the OLS estimator of δ. Your answer should reflect the application of the FWL theorem.

c. Write down the formula for the test statistic to test $H_0 : \delta = 0$. How is this test statistic justified? By default, does `lm` produce the correct version of the test statistic? Why or why not?

Answers

a. For Casual Inference, we need OLS to have two properties: Consistency and Asymptotic normality.

b. $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, with $E(u_i | x_{i1}, x_{i2}) = 0$. The FWL says the effect of x_1 on y by regressing y on $r_1 hat$ is the same.

c. $t = \beta_k - 0/SEof\beta_k$ For any null hypothesis H, if the p-value of H and the observed test statistic is below the significance level α, then we should reject H_0. By default, the the classical assumption adopted by `lm` us wrong because it assumes homoscedasticity.

Problem 2 (20 points)

Figure 1 reproduces the Project STAR class-size effects estimates from assignment 2. Recall that the dependent variable is a student test score. Use the table (in the “figure”) to answer the questions below.

a. Write a sentence that interprets the estimated effects of class size and teacher aides on test scores in Column (1). Are the estimated effects statistically significant at the 5% level? Are they causal?

b. Write a formal expression of the regression estimated in Column (2). Does adding teacher experience affect your answer in part (a)? Is the estimated effect of teacher experience statistically significant at the 5% level? Write a sentence comparing the value of a move to a small class with having a teacher with 10 years experience.

c. What is the difference between the specifications in Columns (2) and (3)? How does the specification in Column (3) affect the estimated class-size and teacher-experience effects?

d. Column (4) adds controls for student gender, race and eligibility for free lunch. How much do they add to the overall explanatory power of the regression? Are the estimated class-size effects robust to their inclusion?

Answers

a. Controlling all other variable, the estimated effects of class size and teacher aides are equivalent to the difference in means, 918.043 for small and 918.357 for regular+aids. The estimated effects are statistically significant at the 1% level and may have a casual relationship with the outcome.

b. $E(y_i | x_i, D_i) = \beta_0 + \delta D_i + \beta_1 x_{i1}$, where $\delta = E(y_i | x_i, D_i = 1)$ - $E(y_i | x_i, D_i = 0)$ Adding teacher experience to the model has essentially no impact on the simple differences in means estimate given in column(1). The estimated effects are statistically significant at the 1% level. Since the average teacher experience is between 9 and 10, having a teacher with 10 years experience is no difference.

c. The difference between Column(2) and Column(3) is that column(3) added the school effects. Adding school effects in column(3) increased the class-size coefficient estimate slightly from 14 to 15.9, decreased the experiencek from 1.469 to 0.743, and it remains highly statistically significant.

d. I check the R^2 value to determine how it affects the explanatory power, the greater the R^2, the greater the power. The R^2 value is increasing every time when adding new controls.

Problem 3 (15 points)

a. Briefly describe the intuition behind regression discontinuity (RD) research design for estimating the effect of drinking on mortality in assignment 3. What is the key identifying assumption for the results from such a design to have a causal interpretation?

b. Figure 3 reproduces the regression discontinuity (RD) plots for `mva` , `suicide` and `homicide` from assignment 3. What is the main message of the plots?

c. Figure 3 reproduces the estimated effects of MLDA laws on `mva` , `suicide` and `homicide` from assignment 3. Write a sentence that interprets the RD estimates for motor-vehicle accidents. Repeat this exercise for suicides and homicides. Write a sentence indicating which results are statistically significant at the 5% level. Based on the quadratic specification, we find that the MVA deaths is increased by 4.66 after 21. Suicides is increased by 1.81 after 21. Homicide is increased by 0.2 after 21. The estimated effect on suicides is statistically significant at the 5% level.

Answers

a. We used RD design to address the drinking on mortality question by assessing the MLDA law to distinguish external(potential alcohol-related) from external causes. The key identifying assumption of RD design is the average potential outcomes are continuous through the cutoff. An ATE measured at cutoff would be identified under the assumption.

b. The figure suggests that, after 21, MVA deaths and suicides rises, but the homicide rate stays the same.

c. Based on the quadratic specification, we find that the MVA deaths is increased by 4.66 after 21. Suicides is increased by 1.81 after 21. Homicide is increased by 0.2 after 21. The estimated effect on suicides is statistically significant at the 5% level.

Problem 4 (15 points)

a. Briefly describe the intuition behind difference-in-differences (DD) research design for estimating the effect of worker's compensation on injury duration in assignment 4. What is the key identifying assumption for the results from such a design to have a causal interpretation?

b. Write down the population regression model corresponding to Column (1), explicitly defining each variable.

c. Figure 3 reproduces the difference-in-differences (DD) estimates of the effect of worker's compensation on injury duration from assignment 4. Write a sentence that interprets the simple DD estimate of the effect of the WBA increase in KY. How does adding covariates to the model affect the DD estimate? How does the simple DD evidence from MI compare with the results from KY?

Answers

a. In assignment 4, instead of estimating the effect of worker's comp on injury duration, we estimated the log injury duration. We compared the difference in mean log duration for high and low earners before and after the WBA increase. It turns out the ldurant DD is 0.20. The key identifying assumption is that the treatment and control-group outcomes would follow parallel trends in the absence of the treatment, and the DD will identify an ATT.

b. $yhat = \mu + \gamma treat + \eta after \ \delta treat * after + u$

Below are the variables used in the regression:

afhigh: time out of work changes in ky after increasing the WBA afchnge: after the change in benefits

highearn: average time out of work changes for high earns

c. The effect of the WBA increase is that the time out of work rose 19.1% in ky. Adding covariates reduced bias and increased R square. Controlling other variables increases the overall fit of the regression by 2 points. The time out of work rose 19.2% in mi because if the WBA increase. The increased percentage is basically the same.

Problem 5 (15 points)

a. What is the rationale for “regularizing” OLS regression for prediction purposes? What does lasso stand for and what does the lasso penalty do?

b. Figure 5 reproduces the lasso $CV(M)$ plot from assignment 5. What do the red dots represent? Give the steps in the cross-validation process behind their calculation.

c. Explain how to use the results depicted in Figure 5 to choose the best model for out-of-sample prediction.

Answers

a. Regularization penalizes model complexity by restricting the regression coefficients, the purpose is to reduce the variance. LASSO stands for least absolute shrinkage and selection operator. It is a shrinkage estimator. The lasso penalty shrinks the OLS coefficient estimates towards zero and forces some coefficients of the least relevant variables to be exactly zero when λ is sufficiently large.

b. The red dots are MSPE values that varies with penalty strength log(λ). The cross-validation split data into train and test sets for example, 70% of the data going to train and 30% of the data going to test. Then divide the train sets into 10 folds and train each fold individually. Measure the model performance for each folds and find the average model performance, then we can get the CV accuracy.

c. In order to choose the best model for out-of-sample prediction, we could look at its improvement from OLS MSPE. The best-performing estimator should have the least MSPE.

Problem 6 (20 points)

In this final problem you will replicate some of the analysis in B.3 of assignment 1 using a different sample from the NLSYM. The data come from Blackburn and Neumark (1992) (https://academic.oup.com/qje/article-abstract/107/4/1421/1846978?redirectedFrom=fulltext) (hereafter, BN) and are available in the `wage2` dataset of the `wooldridge` package. BN’s sample is based on the 1980 survey year, but it is otherwise similar to the Card (1995) sample you used in assignment 1. You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio. You will also find a description in their paper.

Unlike in the homework assignments, you will be a little more “off the chain” here. First, you will have to fill out the code chunk on your own. Don’t fret though, because you have everything you need in assignment 1. Plus, there is TAL if you get really stuck. Second, the analysis write-up is less scripted by the instructions, so you will have to string the relevant sentences together on your own.

a. Begin by constructing a table of summary statistics for the main model variables (`wage` , `educ` , `exper` , `black` , `south` , `urban`) that reports the mean, standard deviation, min and max. Write a short paragraph describing the sample based on the table you constructed.

b. Estimate the return to schooling controlling for `exper` and its square, `black` , `south` , and `urban` . Then, as in assignment 1, address the concern that the estimated education coefficient is biased because the model does not control for unobserved ability by adding `IQ` as a proxy. The sample also provides each young man's Knowledge of the World of Work (`KWW`) score. In a third and final regression, add `KWW` as an additional proxy for unobserved ability.

   Present your results in a proper table using `modelsummary` report standard errors that are robust to heteroscedasticity. Write a short paragraph interpreting the returns-to-schooling estimates from this analysis, being sure to indicated whether they are statistically significant. An obvious approach to this paragraph would be to start with a sentence about the finding in Column (1) and then proceed to columns (2) and (3) highlighting how the results change.

Answers

a. The summary statistics table reported that all variables has 935 samples constructed, the average wage among these samples is 957.95 with minimum wage of 115 and maximum wage of 3078. The `black` , `south` , and `urban` are dummy variables that only has values 0 or 1. By viewing the mean, min and max, the unit of wage should be monthly dollar amount, and the unit of educ&exper should be years.

b. For estimating the log wage return, firstly, use the simple regression model on `exper` and its square, `black` , `south` , and `urban` . I have noticed that the wage2 dataset has no expersq column, so I mutated a new column into the dataset. Secondly, create a second model for adding `IQ` as a proxy. Then create the third model for adding the knowledge of world work score. Combine 3 models and create a coefficient map which contains all added variables and then create a good-of-fit map. Construct a table of results for the combined model by using the modelsummary function, the vcov option is included to report standard errors that are robust to heteroscedasticity.I have also include p-values into the table. The p-values shows that all results are statistically significant at 1% level except for experience and its square. Also, in Column (1) and then proceed to columns (2) and (3), the result is decreasing by adding IQ and KWW.

```
# Construct table of summary statistics of (`wage`, `educ`, `exper`, `black`, `south
`, `urban`) that reports the mean, standard deviation, min and max.
data <- wage2
datasummary(wage + educ + exper + black + south + urban ~
            N + Mean + SD + Min + Max, data = data,
            title="Summary statistics, Blackburn and Neumark (1992)")
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

Summary statistics, Blackburn and Neumark (1992)

	N	Mean	SD	Min	Max
wage	935	957.95	404.36	115	3078
educ	935	13.47	2.20	9	18
exper	935	11.56	4.37	1	23
black	935	0.13	0.33	0	1

	N	Mean	SD	Min	Max
south	935	0.34	0.47	0	1
urban	935	0.72	0.45	0	1

```
# Estimate regression models
#head(card), this dataset has expersq column
#head(wage2), this dataset has no expersq, in order to include expersq to the model,
we need to create one.
data <- dplyr::mutate(wage2, expersq = exper^2)
model_1 <- lm(lwage ~ educ + exper + expersq + black + south + urban, data)

#adding IQ
data_iq <- filter(data, !is.na(IQ))
model_2 <- lm(lwage ~ educ + exper + expersq + black + south + urban + IQ, data_iq)

#adding World work score
model_3 <- lm(lwage ~ educ + exper + expersq + black + south + urban
               + IQ + KWW, data_iq)

#Combine models
models <- list(
  "(1)" = model_1,
  "(2)" = model_2,
  "(3)" = model_3)

# Create coefficient map with variable labels.
cm <- c('educ'        = 'Education',
        'exper'       = 'Experience',
        'expersq'     = 'Experience$^2$',
        'black'       = 'Black',
        'south'       = 'South',
        'urban'        = 'Urban',
        'IQ'          = 'IQ',
        'KWW'         = 'World work score',
        '(Intercept)' = 'Constant')

# Create good-of-fit map.
gm <-  tibble::tribble(
  ~raw, ~clean, ~fmt,
  "nobs", "$N$", 0,
  "r.squared", "$R^2$", 2)
# Estimate the models and construct a table of results.
modelsummary(models,
             coef_map = cm,
             gof_map = gm,
             vcov = c("robust","robust","robust"),
             stars = FALSE,
             title = "Table 2. Estimated returns to schooling, Blackburn and Neumark
(1992)",
             notes = ('Columns (1)-(3) also include Urban dummies,
                          IQ test, and KWW.'),

             statistic = c("p = {p.value}")

             )
```

Table 2. Estimated returns to schooling, Blackburn and Neumark(1992)

	1.	**2.**	**3.**
Education	0.067	0.056	0.049
	p = <0.001	p = <0.001	p = <0.001
Experience	0.014	0.014	0.016
	p = 0.303	p = 0.298	p = 0.238
Experience^2	0.0002	0.0002	0.00005
	p = 0.678	p = 0.693	p = 0.929
Black	–0.213	–0.165	–0.145
	p = <0.001	p = <0.001	p = <0.001
South	–0.093	–0.081	–0.084
	p = <0.001	p = 0.004	p = 0.003
Urban	0.176	0.174	0.166
	p = <0.001	p = <0.001	p = <0.001
IQ		0.004	0.003
		p = <0.001	p = 0.002
World work score			0.006
			p = 0.008
Constant	5.609	5.378	5.350
	p = <0.001	p = <0.001	p = <0.001
N	935	935	935
R^2	0.21	0.22	0.23

Columns (1)-(3) also include Urban dummies, IQ test, and KWW.