STAT 504: Applied Regression
Problem Set 4

Winter 2022

Due date: Monday, February 28th, 2022.

Instructions: Submit your answers in a *single pdf file*. Your submission should be readable and well formatted. **Handwritten answers will not be accepted**. **All code should be in either R or Python.** You can discuss the homework with your peers, but *you should write your own answers and code.* ***No late submissions will be accepted.***

1 Monte Carlo Simulation (40 points)

In this exercise you will perform a short Monte Carlo simulation to compare traditional *versus* robust standard errors in the presence of heteroscedasticity. Consider the following data generating process:

$$X_i \sim \mathcal{N}(0, 1) \tag{1}$$

$$Y_i \mid X_i \sim \mathcal{N}(\mu_{y.x} = \alpha + \beta X_i,\ \sigma^2_{y.x} = \sigma^2 + \lambda X_i^2) \tag{2}$$

Answer the following questions:

(a) What is $\mathbb{E}[Y_i \mid X_i]$? Is it linear on X_i?

(b) What is $\text{Var}[Y_i \mid X_i]$? Is it, in general, constant on X_i?

(c) **Monte Carlo simulation.** For λ ranging from 0 to 10, do the following: generate $m = 1,000$ data sets of sample size $n = 100$ fixing the remaining parameters at $\alpha = 0.5$, $\beta = 1$, and $\sigma^2 = 1$. Construct two 95% confidence intervals for β in each of these data sets, one using traditional standard errors and the other using robust standard errors. Compute the percentage of datasets in which each type of confidence interval covers the true value, $\beta = 1$.

(d) Explain the results of the Monte Carlo simulation.

2 Discrimination (20 points)

For this exercise we will consider the dataset `bm.dta` from Bertrand and Mullainathan [2004]. The authors conducted a randomized experiment and sent 4,870 fictious resumes to employers in response

to job adverts. The resumes were randomly allocated to job openings, and some names were chosen to be distinctively "white sounding" (eg., Emily and Greg) while others distinctively "black sounding" (eg., Lakisha and Jamal). The main research question is to learn whether there are differences in callbacks for interviews due to different names. **Note:** `bm.dta` is in a STATA format, thus you may need to use a special package to read the data in `R` or `Python`.

(a) Run a linear regression of `call` (whether the candidate was called back) on `black` (whether the resume had a "black sounding" name). Interpret the estimated regression coefficients. What was the callback rate for white sounding names? What was the callback rate for black sounding names?

(b) Note that the outcome, `call`, is a binary variable. Is it a problem to run a linear regression in this case? Why, or why not?

(c) Construct a 95% confidence interval for the coefficient of `black` using *robust* standard errors.

(d) Run a regression of `call` on `black` further adjusting for `female` and `yearsexp` (years of experience). Does the regression coefficient of `black` change much? If yes (or no), why do you expect this to be the case?

(e) Can the previous estimates be interpreted causally? Why, or why not?

3 Elections (20 points)

For this exercise we will consider the dataset `hibbs.dat`, as appeared in Gelman et al. [2020].

(a) Make a scatter plot of `vote` against `growth`. Fit a linear regression model to predict `vote` using `growth`. Draw the regression line in the scatter plot.

(b) Interpret the estimated regression coefficients.

(c) Construct a 95% confidence interval for the population regression line, using both the nonparametric bootstrap ($B = 10,000$), robust standard errors, and traditional standard errors. Plot all of them in the previous scatter plot for comparison.

(d) Make a point prediction for the expected vote share of the incumbent party when the average growth is 2%. Construct a 95% *confidence* interval for the expected vote share when the average growth is 2% using the nonparametric bootstrap ($B = 10,000$), robust standard errors, and traditional standard errors.

4 House Prices (20 points)

For this exercise we will consider the dataset `SaratogaHouses.csv` from Corvetti [2007] and Scott [2020]. The dataset contains the sale price of 1,728 houses in Saratoga County, NY, in 2006. Beyond

price, it also contains other 15 variables, such as lot size, age, number of bedrooms, number of fireplaces, number of bathrooms, heating system etc.

(a) Regress `price` on `fireplaces`. Construct a 95% confidence interval for the coefficient of `fireplaces` using robust standard errors. How do you interpret these results?

(b) Regress `price` on `bedrooms`. Construct a 95% confidence interval for the coefficient of `bedrooms` using robust standard errors. How do you interpret these results?

(c) Regress `price` on *all* 15 variables. Give the point estimate and (robust) 95% confidence intervals for the coefficients of `fireplaces` and `bedrooms` for this regression. Did the coefficients of `fireplaces` and `bedrooms` change as compared to (a) and (b)? How do you interpret these results? In particular, how do you make sense of the sign and magnitude of the coefficient related to `bedrooms`?

5 Extra-Credit (10 points)

Consider the following structural causal model:

$$D_i = \lambda_{xd} X_i + U_d \tag{3}$$

$$Y_i = \lambda_{dy} D_i + \lambda_{xd} X_i + \lambda_{x^2 d} X_i^2 + U_y \tag{4}$$

Where X_i, U_d and U_y are independent, standard Gaussian random variables. Now suppose a researcher regress Y_i on D_i and X_i (without the quadratic term), as below:

$$OLS(Y_i|D_i, X_i) = \beta_0 + \beta_1 D_i + \beta_2 X_i$$

Answer the following: (i) is $OLS(Y_i|D_i, X_i) = \mathbb{E}[Y_i \mid D_i, X_i]$? (ii) is $\beta_1 = \lambda_{dy}$? Explain (and prove) your answer.

References

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94 (4):991–1013, 2004.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2020.

Candice Corvetti. *House Price Capitalization of Education by Part Year Residents*. PhD thesis, Citeseer, 2007.

James Scott. *Data Science: A Gentle Introduction*. Unpublished, 2020.