STAT 504: Applied Regression
Problem Set 5

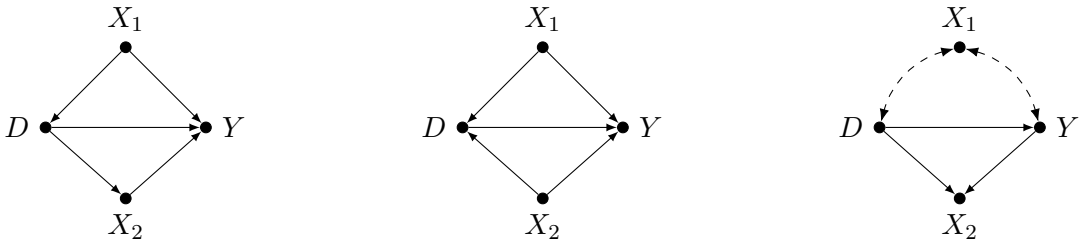
Winter 2022

Due date: Friday, March 11th, 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 Causal Effects via Regression Adjustment

Consider the three models given by the DAGs of Figure 1. All variables are binary, and the dataset for each model can be found on canvas.



(a) Model 1 (b) Model 2 (c) Model 3

Figure 1: Models 1, 2, and 3

For each model, estimate the average treatment effect (ATE) of D on Y, namely, $\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, using a saturated regression model (i.e, a linear regression with fully interacted dummy variables) with the appropriate control variables:

(a) Estimate the ATE for Model 1 (use the data in `model1.csv`).

(b) Estimate the ATE for Model 2 (use the data in `model2.csv`).

(c) Estimate the ATE for Model 3 (use the data in `model3.csv`).

2 Bias Amplification

Consider the model below, where U denotes a latent variable.

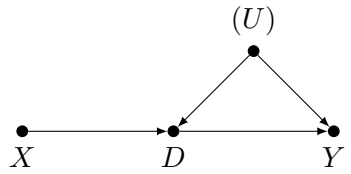


Figure 2: Bias amplification.

Further assume we have a linear *structural* model, of the form:

$$X_i = U_{xi} \tag{1}$$

$$D_i = \lambda_{xd} X_i + U_i \tag{2}$$

$$Y_i = \lambda_{dy} D_i + U_i \tag{3}$$

Where U_i and U_{xi} are independent, standard Gaussian random variables. Our goal is to identify the average treatment effect (ATE) of D on Y, namely, $\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \lambda_{dy}$ using regression adjustment.

(a) Show that it is actually not possible to identify the ATE via simple regression adjustment.

(b) Show that the bias of the regression coefficient adjusting for X is actually larger in magnitude than the bias of the regression coefficient not adjusting for X.

(c) **(Extra-credit).** Show that, although λ_{dy} cannot be identified by a simple regression coefficient, it can nevertheless be identified by the *ratio* of two regression coefficients:

$$\lambda_{dy} = \frac{\beta_{yx}}{\beta_{dx}}$$

where β_{yx} is the regression coefficient of X in the regression $Y \sim X$ and β_{dx} is the regression coefficient of X in the regression $D \sim X$.

3 Cross Validation for Polynomial Regression

Consider the following data generating processes:

- DGP 1: $Y = -2 \times \mathbb{I}(X < -3) + 2.55 \times \mathbb{I}(X > -2) - 2 \times \mathbb{I}(X > 0) + 4 \times \mathbb{I}(X > 2) - 1 \times \mathbb{I}(X > 3) + \epsilon$
- DGP 2: $Y = 6 + 0.4X - 0.36X^2 + 0.005X^3 + \epsilon$
- DGP 3: $Y = 2.83 \times \sin\left(\frac{\pi}{2} \times X\right) + \epsilon$

- DGP 4: $Y = 4 \times \sin(3\pi \times X) \times \mathbb{I}(X > 0) + \epsilon$

where X is drawn from the uniform distribution in $[-4, 4]$ and ϵ is standard Gaussian.

(a) For each DGP, do the following: generate a dataset of size $n = 100$. Calculate the in-sample MSE and the 10-fold cross-validated MSE for polynomial regression (using OLS) of orders $d = 1, 2, \ldots, 10$. Plot the in-sample mean squared error and the cross-validated MSE as a function of d. Report the order of the polynomial regression selected by cross-validation. Fit the best polynomial in the full data, and plot the estimated function, together with the data, as well as the true function.

(b) Now repeat the previous exercise with $n = 1,000$. What polynomial order is selected by cross-validation? Explain why they are the same or different from (a).

4 OLS, Ridge and Lasso

This question is based on Exercise 6.1 of Kuhn et al. [2013]. Read the passage below [Kuhn et al., 2013, p.137]:

> Infrared (IR) spectroscopy technology is used to determine the chemical makeup of a substance. The theory of IR spectroscopy holds that unique molecular structures absorb IR frequencies differently. In practice a spectrometer fires a series of IR frequencies into a sample material, and the device measures the absorbance of the sample at each individual frequency. This series of measurements creates a spectrum profile which can then be used to determine the chemical makeup of the sample material. A Tecator Infratec Food and Feed Analyzer instrument was used to analyze 215 samples of meat across 100 frequencies (...) in addition to an IR profile, analytical chemistry determined the percent content of water, fat, and protein for each sample. If we can establish a predictive relationship between IR spectrum and fat content, then food scientists could predict a sample's fat content with IR instead of using analytical chemistry. This would provide costs savings, since analytical chemistry is a more expensive, time-consuming process:

Consider the dataset `tecator.csv`. The dataset has 215 observation, and 101 columns. The colum `fat`, denotes the percent content of fat for each sample, and the columns `X1` to `X100` denote the 100 IR frequencies of the IR profile. Our goal is to predict `fat` using the IR frequencies.

(a) Use all the data and estimate a simple OLS regression of `fat` on all features. What is the in-sample mean squared error (MSE) of this model? Is this a good estimate of out-of-sample performance?

(b) Estimate the MSE of the simple OLS model using 5-fold cross validation. Is it different from the in-sample MSE?

(c) Using 5-fold cross validation, report the mean squared error of the best Lasso model with the penalty parameter varying from 0 to 1 (grid of 100 values). What is the best penalty selected by cross-validation?

(d) Using 5-fold cross validation, report the mean squared error of the best Ridge model with the penalty parameter varying from 0 to 1 (grid of 100 values). What is the best penalty selected by cross-validation?

5 Lasso, "covariate selection," and causal effects

Consider the following linear structural model:

$$X_k = U_k \text{ (for k = 1 to 98)} \tag{4}$$

$$D = U_d \tag{5}$$

$$Y = U_y \tag{6}$$

$$Z = D + Y + U_z \tag{7}$$

Where all the disturbances U are independent standard Gaussian random variables. A simulated dataset of this DGP is given in `lasso.csv`. Suppose you are interested in estimating the average treatment effect of D on Y (ATE).

(a) What is the ATE of D on Y according to this model? Which variables should you adjust for to identify the ATE via regression?

(b) Now run a Lasso regression of Y on D, Z and $X_1 \dots X_{98}$. Use 10-fold cross validation to pick the best penalization parameter. What are the variables selected by Lasso?

(c) Does the previous Lasso regression provide a good estimate of the causal effect of D on Y? Why, or why not? Is this reason specific to Lasso regression?

Acknowledgements: Q3 based on Hazlett, STAT 201B, UCLA.

References

Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.