STAT 504: Applied Regression
Problem Set 3

Winter 2022

Due date: Friday, February 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 Derivations multivariate regression

In this exercise you will reproduce some of the proofs we did in class, and also derive new results (extra-credit).

1.1 Best linear predictor (multivariate version)

Let β_{ols} be defined as,

$$\beta_{\text{ols}} := \mathbb{E}[X_i X_i^\top]^{-1} E[X_i Y_i].$$

Show that:

$$\beta_{\text{ols}} = \arg\min_{\beta} \mathbb{E}[(Y_i - X_i^\top \beta)^2].$$

1.2 If CEF is linear, it equals the linear regression

Show that, if the CEF is linear on X_i, namely,

$$\mathbb{E}[Y_i \mid X_i] = X_i^\top \beta^*,$$

then, $\beta^* = \beta_{\text{ols}}$

1.3 Linear regression is the best linear approximation to the CEF

Show that β_{ols} is also the minimizer of the following problem:

$$\beta_{\text{ols}} = \arg\min_{\beta} \mathbb{E}[(\mathbb{E}[Y_i \mid X_i] - X_i^\top \beta)^2].$$

1.4 Saturated regression

Let Y_i be the response random variable, and let X_{1i} and X_{2i} be binary random variables. Show that the CEF $\mathbb{E}[Y_i \mid X_{1i}, X_{2i}]$ can be written as a linear function of $X_i = [1, X_{1i}, X_{2i}, X_{1i} \times X_{2i}]$, as in,

$$\mathbb{E}[Y_i \mid X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \times X_{2i}$$

What is the meaning of each parameter above?

1.5 Quadratic regression

Suppose that the CEF of Y_i given X_{1i} is quadratic, as in,

$$\mathbb{E}[Y_i \mid X_{1i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2$$

How can we estimate this CEF using linear regression?

1.6 FWL theorem

Let the linear regression decomposition of Y_i on D_i and X_i be denoted by,

$$Y_i = \tau_r D_i + X_i^\top \beta_r + e_{r,i}$$

Define the "partialling out" operations $Y_i^{\perp X_i}$, $D_i^{\perp X_i}$ as:

$$\begin{aligned} Y_i^{\perp X_i} &:= Y_i - X_i^\top \, \mathbb{E}[X_i X_i^\top]^{-1} E[X_i Y_i] \\ D_i^{\perp X_i} &:= D_i - X_i^\top \, \mathbb{E}[X_i X_i^\top]^{-1} E[X_i D_i] \end{aligned}$$

Show that:

$$\tau_r = \frac{\mathrm{Cov}(Y_i^{\perp X_i}, D_i^{\perp X_i})}{\mathrm{Var}(D_i^{\perp X_i})}$$

1.7 OVB theorem

Let the linear regression decomposition of Y_i on D_i, X_i and Z_i be,

$$Y_i = \tau D_i + X_i^\top \beta + \gamma Z_i + e_i$$

And let the linear regression decomposition of Y_i on D_i and X_i be denoted by,

$$Y_i = \tau_r D_i + X_i^\top \beta_r + e_{r,i}$$

Show that:

$$\tau_r = \tau + \gamma \delta$$

Where:

$$\delta = \frac{\text{Cov}(Z_i^{\perp X_i}, D_i^{\perp X_i})}{\text{Var}(D_i^{\perp X_i})}$$

Explain in English what the theorem says. What is the meaning of γ and δ?

1.8 Partialling-out properties

This question is for extra-credit. Let V, X_1 and X_2 be a scalar random variables, and $Z = [Z_1, Z_2, \ldots, Z_p]^\top$ be a random column vector. Let $V = Z^\top\beta + e$ be the OLS decomposition of V on Z.

Show that:

(a) If $V = X_1 + X_2$, then $V^{\perp Z} = X_1^{\perp Z} + X_2^{\perp Z}$.

(b) $e^{\perp Z} = e$.

(c) $Z^{\perp Z} = \mathbf{0}$, where:

$$Z^{\perp Z} = [Z_1^{\perp Z}, Z_2^{\perp Z}, \ldots, Z_p^{\perp Z}]^\top$$

2 The association of access to clean water and infant mortality

In this exercise, you will use regression to investigate the association between access to clean water and infant mortality. We will use the *Quality of Governance* dataset (`qog_jan16.csv`) from January 2016 [Teorell et al., 2016, Aronow and Miller, 2019]. This exercise is based on the example found in Aronow and Miller [2019, Sec. 4.4].

The dataset contains 184 rows (countries), and four columns:

- `cname`: country name;
- `wdi_morinftot`: infant mortality rate, as measured by the number of infants died before reaching one year of age per 1,000 live births, in a given year;
- `epi_watsup`: access to clean water, as measured by the percentage of the population with access to a source of clean driking water.
- `wdi_accelectr`: access to electricity, as measured by the percentage of the population with acceess to electricity.

Here we will assume that the data is an i.i.d sample from the joint distribution of these three variables. For all questions below let Y_i denote `wdi_morinftot`, X_i denote `epi_watsup`, and Z_i denote `wdi_accelectr`. For all bootstrap exercises, you should write your own bootstrap code (do not use a bootstrap package), and use $10,000$ bootstrap samples. In what follows, we will use the notation $\beta_{ab.1cde}$ to denote the regression coefficient of A on B after adjusting for a constant, C, D and E, and $e_{a.1bcde,i}$ to denote the residuals of this same regression. This notation is more "verbose," but it helps clearly differentiating each regression coefficient and residual.

2.1 Simple linear model

We are going to start our investigation using a simple linear model.

(a) Construct a scatter plot between infant mortality and access to clean water. Does there seem to exist an association? In which direction?

(b) Fit a simple linear regression model to predict infant mortality using access to clean water,

$$Y_i = \widehat{\beta}_{y1.x} + \widehat{\beta}_{yx.1} X_i + \hat{e}_{y.1x,i}$$

Construct a scatter plot between the two variables and include the regression line. What are the estimates for the coefficients of this regression? How can you interpret these estimates?

(c) Construct a 95% confidence interval for the regression coefficients using the nonparametric bootstrap.

(d) Now fit a regression model to predict infant mortality using *both* access to clean water and access to electricity,

$$Y_i = \widehat{\beta}_{y1.xz} + \widehat{\beta}_{yx.1z} X_i + \widehat{\beta}_{yz.1x} Z_i + \hat{e}_{y.1xz,i}$$

Construct a 3d scatter plot between the three variables and include the regression plane. Is the new estimate for the regression coefficient related to X_i the same as before? If not, is its magnitude lower or higher, as compared to $\widehat{\beta}_{yx.1}$? How can you interpret $\widehat{\beta}_{yx.1z}$?

(e) **FWL theorem**. Here you will verify the FWL theorem numerically.

- Fit a linear regression of X_i on Z_i (and a constant). Save the residuals on a new variable called `x_z`.
- Fit a linear regression of Y_i on Z_i (and a constant). Save the residuals on a new variable called `y_z`.
- Construct a scatter plot of `y_z` with `x_z`. Now regress `y_z` on `x_z`. Draw the regression line in the previous scatter plot. The coefficient related to `x_z` on this regression should be numerically identical to the previous regression coefficient $\widehat{\beta}_{yx.1z}$.

(f) **OVB theorem**. Here you will verify the OVB theorem numerically.

- Fit a linear regression to predict access to electricity (Z_i) using access to clean water (X_i):

$$Z_i = \widehat{\beta}_{z.1x} + \widehat{\beta}_{zx.1} X_i + \widehat{e}_{z.1x,i}$$

What is the estimated value of $\widehat{\beta}_{zx.1}$, and how do you intepret it?

- Show numerically that:

$$\widehat{\beta}_{yx.1z} = \widehat{\beta}_{yx.1} - \widehat{\beta}_{yz.1x} \times \widehat{\beta}_{zx.1}$$

2.2 Quadratic model

We are now going to make our model a bit more flexible, and use a quadratic specification. You will note that, while the quadratic model provides a better fit to the data, for this particular example, most of the key insights of the association between infant mortality and access to clean water were already given in the simple linear model.

(a) Fit a quadratic model to the data using OLS:

$$Y_i = \widehat{\beta}_{y1.xx^2} + \widehat{\beta}_{yx.1x^2} X_i + \widehat{\beta}_{yx^2.1x} X_i^2 + \hat{e}_{y.1xx^2,i}$$

Construct a scatter plot between infant mortality and access to clean water and include the quadratic regression line. Does this seem to fit the data better? What are the estimates for the regression coefficients of this regression? In this quadratic model, can we interpret the impact of a change in X_i on our predictions of Y_i by simply reading the regression coefficient $\widehat{\beta}_{yx.1x^2}$? Why, or why not?

(b) Define the *average partial derivative* of X_i on Y_i as,

$$\text{APD}_{yx} = \mathbb{E}\left[\frac{\partial\, \mathbb{E}[Y_i \mid X_i]}{\partial X_i}\right]$$

Now let us use the previous quadratic regression model as an approximation to the CEF,

$$\mathbb{E}[Y_i \mid X_i] = \beta_{y1.xx^2} + \beta_{yx.1x^2} X_i + \beta_{yx^2.1x} X_i^2.$$

Show that:

$$\text{APD}_{yx} = \beta_{yx.1x^2} + 2\beta_{yx^2.1x}\, \mathbb{E}[X_i].$$

(c) Estimate the previous average partial derivative using the plug-in principle:

$$\widehat{\text{APD}}_{yx} := \widehat{\beta}_{yx.1x^2} + 2\widehat{\beta}_{yx^2.1x}\, \mathbb{E}_n[X_i]$$

Where $\mathbb{E}_n[X_i] = n^{-1} \sum_i X_i$ is the empirical mean. Construct a 95% confidence interval using the nonparametric bootstrap. How does the $\widehat{\text{APD}}_{yx}$ compare with the previous regression coefficient of the simple linear model $\widehat{\beta}_{yx.1}$?

(d) Recall the derivative can be defined as :

$$\left.\frac{\partial\, \mathbb{E}[Y_i \mid X_i]}{\partial X_i}\right|_{X_i = x} = \lim_{h \to 0} \frac{\mathbb{E}[Y_i \mid X_i = x + h] - \mathbb{E}[Y_i \mid X_i = x - h]}{2h}$$

This formula suggests an alternative to approximate the derivative numerically, by computing the above difference with a small enough h. Here we will estimate the APD using the plug-in

principle and this numerical approximation. Use $h = 0.0001$ and compute,

$$\widehat{\text{APD}'}_{yx} := \mathbb{E}_n \left[\frac{\widehat{\mathbb{E}}[Y_i \mid x_i + h] - \widehat{\mathbb{E}}[Y_i \mid x - h]}{2h} \right] = \frac{1}{n} \sum_i \left[\frac{\hat{\mathbb{E}}[Y_i \mid x_i + h] - \hat{\mathbb{E}}[Y_i \mid x - h]}{2h} \right]$$

Where $\widehat{\mathbb{E}}[Y_i \mid x]$ denotes the predictions of the quadratic model you fitted, with $X_i = x$. How does this new estimate compare to the prior estimate?

(e) Let us now consider a quadratic model to predict infant mortality, using *both* access to clean water and access to electricity. Using OLS, fit a quadratic model of the following form:

$$Y_i = \widehat{\beta}_{y1.xx^2zz^2} + \widehat{\beta}_{yx.1x^2zz^2} X_i + \widehat{\beta}_{yx^2.1xzz^2} X_i^2 + \widehat{\beta}_{yz.1xx^2z^2} Z_i + \widehat{\beta}_{yz^2.1xx^2z} Z_i^2 + \hat{e}_{y.1xx^2zz^2,i}$$

Construct a 3d scatter plot between the three variables and include the regression plane. Report the estimated coefficient values.

(f) Define the *average partial derivative* of X_i on Y_i, adjusting for Z_i, as,

$$\text{APD}_{yx.z} = \mathbb{E} \left[\frac{\partial \, \mathbb{E}[Y_i \mid X_i, Z_i]}{\partial X_i} \right]$$

Use a numerical approximation and the plug-in principle to compute the APD:

$$\widehat{\text{APD}}_{yx.z} := \mathbb{E}_n \left[\frac{\widehat{\mathbb{E}}[Y_i \mid x_i + h, z_i] - \widehat{\mathbb{E}}[Y_i \mid x_i - h, z_i]}{2h} \right] = \frac{1}{n} \sum_i \left[\frac{\hat{\mathbb{E}}[Y_i \mid x_i + h, z_i] - \hat{\mathbb{E}}[Y_i \mid x_i - h, z_i]}{2h} \right]$$

How does the estimate of the APD compare to the previous regression coefficient $\widehat{\beta}_{yx.1z}$?

(g) Construct a 95% confidence interval for $\text{APD}_{yx.z}$ using the nonparametric bootstrap.

2.3 Causal interpretation (a warm-up)

Can we interpret the previous estimates as the causal effect of access to clean water in infant mortality? Why, or why not? Explain your answer?

References

Jan Teorell, Staffan Kumlin, Stefan Dahlberg, Sören Holmberg, Bo Rothstein, Anna Khomenko, and Richard Svensson. The quality of government oecd dataset, version jan16. *University of Gothenburg: The Quality of Government Institute, http://www. qog. pol. gu. se doi*, 10, 2016.

Peter M Aronow and Benjamin T Miller. *Foundations of agnostic statistics*. Cambridge University Press, 2019.