STAT 504: Applied Regression
Midterm

Winter 2022

Due date: Sunday, February 20th, 2022.

Instructions: Submit your answers in a *single pdf file*. Your submission should be readable and well formatted. **All code should be in either R or Python.** ***No late submissions will be accepted.*** Read all instructions carefully and give complete answers (that is, explain your answer, do not merely show the output of the code).

1 Derivations multivariate regression (sample version) (20 points)

For all the questions below, suppose we have an independent and identically distributed sample of size n of $P(V_i)$, where $V_i = (Y_i, X_i)$, Y_i is the outcome, and X_i is a $p \times 1$ vector of covariates, including a constant, for observation i. Assume all relevant moments and inverses exist. Let $\mathbf{X}$ denote the $n \times p$ *matrix* of n observations and p covariates (including a constant), and Y the $n \times 1$ vector of n individuals of the outcome.

1.1 OLS estimator

Consider the empirical least squares problem:

$$\hat{\beta}_{ols} = \arg\min_{\beta} \frac{1}{n} \sum_i (Y_i - X_i^\top \beta)^2 \tag{1}$$

Show that:

$$\hat{\beta}_{ols} = \left[\frac{1}{n} \sum_i X_i X_i^\top \right]^{-1} \left[\frac{1}{n} \sum_i X_i Y_i \right] \tag{2}$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y \tag{3}$$

1.2 OLS estimator and the classical model

(a) Show that we can write the OLS estimator as:

$$\hat{\beta}_{ols} = \beta_{ols} + \left[\frac{1}{n} \sum_i X_i X_i^\top \right]^{-1} \left[\frac{1}{n} \sum_i X_i e_i \right] \tag{4}$$

Then argue that:

$$\mathbb{E}[\hat{\beta}_{ols}|\mathbf{X}] = \beta_{ols} + \left[\frac{1}{n}\sum_i X_i X_i^\top\right]^{-1}\left[\frac{1}{n}\sum_i X_i\,\mathbb{E}[e_i|X_i]\right]$$

Where β_{ols} is the population regression coefficient, and e_i is the population regression residual, of the regression of Y_i on X_i.

(b) Is the estimator $\hat{\beta}_{ols}$, in general, an unbiased estimator of β_{ols}? Why, or why not?

(c) Now suppose the CEF is linear, that is:

$$\mathbb{E}[Y_i \mid X_i] = X_i^\top \beta_{ols}$$

Show that in this case $\hat{\beta}_{ols}$ is both conditionally and unconditionally unbiased.

(d) Beyond linearity, suppose that homoskedasticity holds; that is, assume $\text{Var}(Y_i|X_i) = \sigma^2$ is constant. Show that, in this case:

$$\text{Var}(\hat{\beta}_{ols} \mid \mathbf{X}) = \frac{\sigma^2}{n}\left[\frac{1}{n}\sum_i X_i X_i^\top\right]^{-1} \tag{5}$$

$$= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \tag{6}$$

(e) Finally, argue that, if we additionally assume e_i is normally distributed, then:

$$\hat{\beta}_{ols} \mid \mathbf{X} \sim \mathcal{N}(\beta_{ols},\ \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

2 Applied Analysis: violence in Darfur (80 points)

In this exercise you will reproduce and extend some of the regression analyses found in Cinelli and Hazlett [2020], Hazlett [2020], and Cinelli et al. [2020]. We will consider a dataset on attitudes of Darfurian refugees in eastern Chad. The main "treatment" variable is `directlyharmed`, which indicates that the individual was physically injured during attacks on villages in Darfur, between 2003 and 2004. The main outcome of interest is `peacefactor`, a measure of pro-peace attitudes. Key covariates include `herder_dar` (whether they were a herder in Darfur), `farmer_dar` (whether they were a farmer in Darfur), `age`, `female` (indicator for female), and `past_voted` (whether they report having voted in an earlier election, prior to the conflict).

This dataset can be found both in the R and Python package `sensemakr`. If you are using R, install the package with:

```
install.packages("sensemakr")
```

And load the data with:

```
library("sensemakr")
data("darfur")
```

If you are using Python, install the package with:

```
pip3 install PySensemakr
```

And load the data with:

```
import sensemakr as smkr
darfur = smkr.load_darfur()
```

You can find further help and documentation in `https://carloscinelli.com/sensemakr/` (for R) and `https://pysensemakr.readthedocs.io/en/latest/` (for Python).

Throughout this exercise, let:

- $Y_i =$ `peacefactor` (outcome);
- $D_i =$ `directlyharmed` (binary indicator of "treatment");
- $F_i =$ `female` (binary indicator);
- $V_i =$ `village` (a matrix of binary village indicators);
- $X_i =$ a matrix with a constant and the covariates `herder_dar`, `farmer_dar`, `age`, and `past_voted`.

In what follows, many passages of the description of the problem draws directly from Cinelli and Hazlett [2020] and Cinelli et al. [2020].

2.1 Research Question

In 2003 and 2004, the Darfurian government orchestrated a horrific campaign of violence against civilians, killing an estimated two hundred thousand people. This application asks whether, on average, being directly injured or maimed in this episode (D_i) changed individuals attitudes towards peace (Y_i). Did exposure to violence make individuals more likely to feel "vengeful" and unwilling to make peace with those who perpetrated this violence, or, more likely to feel "weary," and motivated to see it end by making peace?

More specifically, suppose we are interested in the average treatment effect of D_i on Y_i (ATE).

(a) Write down mathematically our target query, the ATE, using either potential outcome or do notation.

2.2 Identification

The purpose of these attacks was to punish civilians from ethnic groups presumed to support the opposition, and to kill or drive these groups out so as to reduce support for the opposition. Violence against civilians included aerial bombardments by the government as well as ground assaults by the Janjaweed, a pro-government militia. Now suppose a researcher argues that, while some villages were more or less intensively attacked, within village violence was largely indiscriminate. The bombings could not be finely targeted owing to their crudeness, and there were not many reasons to target them. Similarly, the Janjaweed had no reason to target certain individuals rather than others, and no information with which to do so—with one major exception: women were targeted and often subjected to sexual violence.

Given these considerations, this researcher may argue that adjusting for village V_i and female F_i is sufficient for control of confounding.

(a) Express mathematically (in counterfactual terms) the claim "village and female are sufficient for control of confounding."

(b) Given the assumption above (and consistency), can we identify the ATE from the observed data? If so, write down the identifying expression, nonparametrically.

(c) According to the assumption above, is X_i necessary for identification of the ATE? Why or why not? And if not, why could we still consider adjusting for X_i in our regression?

2.3 Estimation: specification 1

Now suppose the researcher is willing to entertain that the CEF of Y_i can be reasonably approximated with a linear function of the covariates,

$$\mathbb{E}[Y_i \mid D_i, F_i, V_i, X_i] \approx \tau_1 D_i + \beta_{1f} F_i + V_i^\top \beta_{1v} + X_i^\top \beta_{1x}.$$

(a) Show that, under the specification given above, the ATE can be identified by the regression coefficient τ_1.

(b) Estimate the ATE using linear regression considering the specification given above.

(c) Construct an estimate of τ_1 using the FWL theorem, and show it is numerically identical to the previous estimate. Namely:

- Regress Y_i on F_i, V_i and X_i. Save the residuals of this regression. Call such residuals $\tilde{Y}_i$;
- Regress D_i on F_i, V_i and X_i. Save the residuals of this regression. Call such residuals $\tilde{D}_i$;
- Regress $\tilde{Y}_i$ on $\tilde{D}_i$.

(d) Construct a 95% confidence interval for the ATE using the nonparametric bootstrap (with 1,000 samples). Compare it with the classical 95% interval.

(e) Explain your findings.

2.4 Estimation: specification 2

Now suppose a fellow researcher argues that we should consider an interaction term between D_i and F_i, as in:

$$\mathbb{E}[Y_i \mid D_i, F_i, V_i, X_i] \approx \tau_2 D_i + \beta_{2f} F_i + \beta_{2fd} D_i \times F_i + V_i^\top \beta_{2v} + X_i^\top \beta_{2x}.$$

(a) Explain why the ATE is not equal a single regression coefficient anymore.

(b) Estimate the ATE using the plug-in principle and linear regression considering the specification given above.

(c) Construct a 95% confidence interval for the ATE using the nonparametric bootstrap (with 1,000 samples).

(d) Does this new specification change your previous findings?

2.5 Sensitivity analysis: traditional OVB

From here on, consider only specification 1, namely:

$$\mathbb{E}[Y_i \mid D_i, F_i, V_i, X_i] \approx \tau_1 D_i + \beta_{1f} F_i + V_i^\top \beta_{1v} + X_i^\top \beta_{1x}.$$

Note that the causal interpretation of the previous estimates requires the assumption of no unobserved confounders. While this may be supported by the claim that there was no targeting of violence within village-gender strata, not all investigators may agree with this account. For example, a reasonable argument could be made that, although the bombing was crude, bombs were still more likely to hit the center of the village, and those in the center would also likely hold different attitudes towards peace.

Let C_i denote the unobserved binary confounder `center`. Thus, one could argue that unconfoundedness would hold only conditional on F_i, V_i *and* C_i, and, to estimate (approximate) the ATE, we should have instead run the regression:

$$\mathbb{E}[Y_i \mid D_i, F_i, V_i, X_i, C_i] \approx \tau D_i + \beta_f F_i + V_i^\top \beta_v + X_i^\top \beta_x + \gamma C_i.$$

Now denote by $W_i = [F_i, V_i, X_i]$ and let $\delta := \mathrm{Cov}(D_i^{\perp W_i}, C_i^{\perp W_i}) / \mathrm{Var}(D_i^{\perp W_i})$.

(a) Express mathematically the new unconfoundedness assumption.

(b) How would including the hypothetical confounder C_i in our regression equations have changed our inferences? Write mathematically the expression for the difference between τ_1 and τ, in terms of γ and δ. What is the interpretation of each term?

(c) Suppose a researcher claims that the covariate Center could have a conditional impact of $\gamma = 0.2$ in attitudes towards peace, and conditional imbalance of $\delta = 0.2$, meaning that on average, those who were physically injured were also 20 percentage points more likely to live in the center of the village. Using the plug-in principle, compute a point estimate for τ under such assumptions, assuming that the direction of the bias reduces the magnitude of τ. Construct a 95% confidence interval for τ using the nonparametric bootstrap (1,000 samples). Is such strength of confounding sufficient to change the main conclusions of the study?

2.6 Sensitivity analysis: partial R^2 parameterization

Beyond `center`, one may also argue that the *Janjaweed* may have observed signals that indicate, for example, the wealth of individuals. Or, perhaps, that an individual's prior political attitudes could have led them to take actions that exposed them to greater risks. To complicate things, all these factors could interact with each other or otherwise have non-linear effects, all acting as confounders. Let Z_i denote a vector of all these covariates, including nonlinear transformations if necessary.

(a) Write down the formula for the omitted variable bias in terms of the partial R^2 of Z_i with Y_i and D_i. Which terms can be estimated from the data, and which terms need to be limited by hypothesis regarding the strength of confounding?

(b) Using the plug-in principle, compute the terms of the bias that can be estimated from the data. Now suppose that Z_i can explain, collectively, at most 12% of the residual variation of the outcome, and 1% of the residual variation of the treatment. Assuming that the direction of the bias reduces the magnitude of τ, compute a point estimate for τ using the plug-in principle. Compare this with the result of the package `sensemakr`, using the function "adjusted estimate" (they should be numerically identical).

(c) Now bootstrap the whole procedure above, and construct a 95% confidence interval for τ using the nonparametric bootstrap (1,000 samples). Is such strength of confounding sufficient to change the main conclusions of the study?

References

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

Chad Hazlett. Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, 64(5):844–870, 2020.

Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. sensemakr: Sensitivity analysis tools for ols in r and stata. *Available at SSRN 3588978*, 2020.