Problem Set 3
MGMT 737

Spring 2025

You should have gotten this homework assignment from the Github classroom environment (https://classroom.github.com/classrooms/192971645-yale-mgmt-737-spring-2025-classroom). In submitting your problem set, you should have two files in your Github repository:

1. `homework3-code.R`, which contains your code,
2. `homework3-writeup.pdf`, which contains your writeup

You may have other files / folders if necessary when there are images or auxiliary files. My very strong preference is for you to write this up in R. If this is not possible, you can use Python. Please let me know if you're planning on this. (This is not a coding preference, but mainly a grading issue.)

1. **Standard Errors I** For this problem, use the dataset `networth_delta_elas.csv`, where `county_fips` is the county FIPS code, `statename` is the state FIPS code, `elasticity` is the Saiz elasticity measure, `total` is the number of households in each county, and `netwp_h` is the change in net worth within a county from 2006 to 2009.

   (a) Write a function to estimate the linear regression of networth change against a constant and the Saiz elasticiy. Report the coefficient on the elasticity.
   → Label the calculated value as `lin_reg_coef` in your code.

   (b) Next, estimate the homoskedastic SE, heteroskedasticity-robust SE, HC2, and HC3 standard errors for the elasticity estimate. [See the formulas in the slides]
   → Label the calculated values as `homoskedastic_se1`, `hetero_se1`, `hc2_se1`, `hc3_se1` in your code.

   (c) Now, we will estimate the three standard errors from Abadie et al. (2020) [see section 4 for details]. I will walk you through the estimation. Let

$$\begin{aligned} V^{causal} &= n^{-1}\Gamma^{-1}(\rho\Delta^{cond} + (1-\rho)\Delta^{ehw})\Gamma^{-1} \\ V^{causal,sample} &= n^{-1}\Gamma^{-1}\Delta^{cond}\Gamma^{-1} \\ V^{descr} &= n^{-1}(1-\rho)\Gamma^{-1}\Delta^{ehw}\Gamma^{-1} \\ V^{ehw} &= n^{-1}\Gamma^{-1}\Delta^{ehw}\Gamma^{-1}. \end{aligned}$$

   Our elasticity measure is X_i, and the outcome is Y_i. Let Z denote our constant (and potentially an additional control).

   i. Estimate $\hat{\epsilon}_i$ as the standard residual from the linear regression of Y on X and Z
   ii. Estimate the short regression of X on Z (X as the outcome, Z as the right hand side) to calculate $\hat{\gamma}$, the projection of X on Z (note that when Z is a constant, this is just the mean of X).
   iii. Estimate $\hat{\Gamma} = n^{-1}\sum_i (X_i - \hat{\gamma}Z_i)^2$
   iv. Estimate $\hat{\Delta}^{\text{ehw}} = n^{-1}\sum_i (X_i - \hat{\gamma}Z_i)\hat{\epsilon}_i^2(X_i - \hat{\gamma}Z_i)$.
   v. Now, estimate $V^{EHW} = (1/n) * \hat{\Gamma}^{-1}\hat{\Delta}^{\text{ehw}}\hat{\Gamma}^{-1}$. Check that this coincides with your previous EHW estimates (it should differ slightly b/c of no degree of freedom corrections, but quite close).
   vi. Estimate $\rho = n/N$ using the following fact: the data is observed at the county level, and in the United States, there are 3,006 counties. Recall that (in my notation) n is the number of observations in the sample, and N is the "population." Using this measure, estimate $V^{descr} = (1-\rho)V^{EHW}$ and report the standard error. Note the relative size of each.
   → Label the calculated value as `abadie_se_desc`.

vii. Next, calculate:

$$\hat{G} = \left(n^{-1} \sum_i (X_i - \hat{\gamma} Z_i) \hat{\epsilon}_i Z_i' \right) \left(n^{-1} \sum_i Z_i Z_i' \right)^{-1} \tag{1}$$

Note that this is exactly zero in our current case.

viii. Now, let $\hat{\Delta}^Z = n^{-1} \sum_i ((X_i - \hat{\gamma} Z_i)\hat{\epsilon}_i - \hat{G} Z_i)^2$. Note that this should be equal to $\hat{\Delta}^{EHW}$.

ix. Finally, calculate V^{causal} and $V^{\text{causal,sample}}$ using $\hat{\Delta}^Z$ in the place of Δ^{cond} and and report the standard errors. (We cannot estimate Δ^{cond} feasibly, so we use Δ^Z in its place.) Note that in this setting, we have identical estimates for the causal estimates b/c we cannot do better than the EHW estimate.

→ Label the calculated value as `abadie_se_causal` and `abadie_se_causal_sample` in your code.

x. Now reimplement this approach, but include state fixed effects as controls in Z. Report your estimates for the standard errors using V^{EHW}, V^{descr}, V^{causal} and $V^{causal,sample}$ in this setting.

→ Label the calculated value as `abadie_se_desc2`, `abadie_se_causal2` and `abadie_se_causal_sample2` in your code.

2. **Standard Errors II - Clustering:** This problem will teach you how to generate data and then test your model on the data. First, you will construct code to randomly sample data. Then you will implement different estimators on the randomly generated data, and study the properties of these estimators over many random simulations. *N.B. Make sure to set your seed* `set.seed(123)`.

(a) Generate a sample of $n_c = 100$ clusters, each of size $n = 100$ (hence $N = 10,000$). For each observation, you will randomly assign a treatment D_i with probability $p = 0.5$. Specify the outcomes as:

$$Y_i = D_i \tau_i + \varepsilon_i \tag{2}$$

where

$$\varepsilon_i = \varepsilon_{c(i)} + \varepsilon_{indiv,i} \tag{3}$$
$$\varepsilon_{c(i)} \sim \mathcal{N}(0, \rho) \tag{4}$$
$$\varepsilon_{indiv,i} \sim \mathcal{N}(0, 1-\rho) \tag{5}$$
$$\tau_i = \tau_{c(i)} + \tau_{indiv,i} \tag{6}$$
$$\tau_{indiv,i} \sim \mathcal{N}(0.1, 0.5^2) \tag{7}$$
$$\tau_{c(i)} \sim \mathcal{N}(0.1, 0.5^2) \tag{8}$$

where $c(i)$ denotes the cluster that person i is a part of, and $\rho = 0.5$.

For a sample of 1000 simulations, what is the overall average estimate of $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$? This should be very close to `0.2`. What is the standard deviation of the average effect across simulations?

→ Label the calculated value as `est_tau1` and `sd_tau1` in your code.

(b) Now for each simulation, estimate the standard error of the treatment effects using (1) HC2 robust standard errors, and (2) clustered standard errors, clustering at the cluster level. What is the average coverage of the true value, using a 95% confidence interval, for each of these estimators? What is the average ratio of the clustered standard errors to the robust standard errors?

→ Label the calculated value as `coverage_hc2`, `coverage_cluster` and `ratio1` in your code.

(c) Now, set $\tau_{c(i)} = 0.1$ for all i. What is the average estimate of the treatment effect in this setting? What is the average coverage of the true value, using a 95% confidence interval, for each of these estimators? What is the average ratio of the clustered standard errors to the robust standard errors?

→ Label the calculated value as `coverage_hc2b`, `coverage_clusterb` and `ratio1b` in your code.

(d) Now remove your change from part (c). Instead, set $\rho = 0.05$. What is the average coverage of the true value, using a 95% confidence interval, for each of these estimators? What is the average ratio of the clustered standard errors to the robust standard errors?

→ Label the calculated value as `coverage_hc2c`, `coverage_clusterc` and `ratio1c` in your code.

(e) Consider what the implications are for the use of clustered standard errors in the presence of heterogeneous TE. If there is no correlation within a cluster for treatment effects, do you need to cluster your standard errors?