240A Winter 2022: Computing Assignment II

DUE: 6pm on 1/27. Upload to Canvas an RMD file and either a PDF or html file.

Your compiled file should contain slides that you will present during discussion (if your group is chosen).

One selected group will present problem 1 and the other will present problem 2

1. Simulation of Misspecified Linear Model

Investigate the linear projection of a quadratic function. You will simulate a DGP of the form: $y = log(x) + e, \quad e \sim N(0,1)$, where the scalar x is lognormal with 3 different moments: $log(x) \sim N(0,1)$, $log(x) \sim N(3,1)$, and $log(x) \sim N(0,3)$.

You will estimate the linear projection: $y = \beta_0 + \beta_1 x + e$, which is a misspecified model for the CEF.

1. What are the population values of β_1 for each of the three x distributions? You may derive the answer analytically as $cov(x,y)/var(x)$, or you may approximate it by generating a very large sample from the DGP and computing the OLS estimate $\hat{\beta}_1$. *Hint:* if $ln(x) \sim N(\mu, \sigma^2)$, then $cov(x, ln(x)) = \sigma^2 e^{\mu + 0.5\sigma^2}$
2. Simulate the distribution of $\hat{\beta}_1$ under the three different distributions for x, and for three different sample sizes: $n = 10, 100, 1000$. Summarise the results from the simulations in a data-frame. Is the estimator unbiased? What is the effect of the sample size? Try to explain what you observe using intuition, theory and formulas seen in class.

To complete part 2, you need to generate a large number of samples from the DGP (say, 1000), and then you will estimate β for each sample. Do this separately for each x distribution and each n. To present your results, you could make tables of means and standard deviations.

2. Heteroskedasticity Consistent Covariance Estimation

This problem is motivated by Section 4.16 in Hansen (2019) and Section 8.1 in Mostly Harmless Econometrics, by Angrist and Pischke. Consider the following data generating process:

$$y_i = \beta_0 + \beta_1 D_i + e_i$$

where $\beta_0 = \beta_1 = 0$ and $e_i \sim N(0, \sigma_i^2)$. Use a sample size $n = 100$. The variable D_i is a dummy variable with n_1 values equal to one and the remaining values equal to zero. The error variance is $\sigma_i^2 = \omega^2$ for the $D_i = 1$ observations and $\sigma_i^2 = 1$ for the $D_i = 0$ observations.

You will simulate the properties of 5 estimators for $V_{\hat{\beta}_1} = var(\hat{\beta}_1 | D)$.

- Consider the following settings: $\omega \in \{1, 2\}$ and $n_1 \in \{3, 10, 25\}$.
- Evaluate each of the following variance estimators defined on pages 113-115 of Hansen's book: $\{\hat{V}^0_{\hat{\beta}_1}, \hat{V}^{ideal}_{\hat{\beta}_1}, \hat{V}^{HC0}_{\hat{\beta}_1}, \hat{V}^{HC1}_{\hat{\beta}_1}, \hat{V}^{HC2}_{\hat{\beta}_1}\}$

For each of the 6 combinations of $\{\omega, n_1\}$ and each of the 5 estimators, simulate the following objects:

- True standard error (i.e, standard deviation of $\hat{\beta}_1$ across Monte Carlo draws)
- Average estimated standard error (i.e., average of the square root of $\hat{V}_{\hat{\beta}_1}$ across Monte Carlo draws)

- Probability of rejecting $H_0 : \beta_1 = 0$ when the estimated standard error is used to compute the t-statistic (i.e., proportion of t-statistics greater than 1.96 in absolute value)

(These objects are reported for different parameter settings in Table 8.1.1 of Mostly Harmless Econometrics.)

R hints:

- install the sandwich package. Then use the **vcovHC** function to compute the variance estimates.

```
install.packages("sandwich")
library(sandwich)
```

- For $\hat{V}^{0}_{\hat{\beta}_1}$, use **type="const"** in the **vcovHC** function. For $\hat{V}^{HC0}_{\hat{\beta}_1}$, use **type="HC0"**, etc.
- For the *ideal* estimator, we need to feed the omega argument in the `vcovHC()` function. Here omega refers to the diagonal elements of the Ω matrix. For example, if *omega* denotes ω, $n1$ denotes n_1, and reg is your regression:

```
sigmas <- c(rep(omega, n1), rep(1, n-n1))
vcovHC(reg, omega= sigmas^2)
```