Econ 31703: Assignment 1

Due date: April 21, 2021

Exercise 1

Consider the following data generating process:

$$
\begin{aligned}
(X_1, \cdots, X_p, \varepsilon) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{p+1}\right), \\
Y &= X_1 - X_2 + \varepsilon.
\end{aligned}
$$

An i.i.d. sample from $(Y, X_1, \cdots, X_p)$ of size N can be expressed in the following matrix form:

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1{}^\intercal \\ \vdots \\ \mathbf{X}_N{}^\intercal \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{Np} \end{pmatrix}.
$$

(a) Set $p = 90$ and simulate 10,000 samples of size $N = 100$. These samples will be used for **(b)**-**(d)**.

(b) Suppose an econometrician does not have knowledge on the DGP and considers the following linear regression model:

$$
Y_i = X_{i1}\beta_1 + u_i.
$$

Let $\hat{\beta}_1$ be the OLS estimator of β_1 from this regression model. Compute $\hat{\beta}_1$ for each sample and report the average and the variance of $\hat{\beta}_1$ across the simulated samples.

(c) The econometrician also considers alternative linear models with varying $\tilde{p}$:

$$
Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{i\tilde{p}}\beta_{\tilde{p}} + u_i.
$$

Let $\hat{\beta}_1(\tilde{p})$ denote the OLS estimator of β_1 in a linear model that takes $X_{i1}, \cdots, X_{i\tilde{p}}$ as regressors. Compute $\hat{\beta}_1(\tilde{p})$ for each sample while varying $\tilde{p} = 5, 10, 50, 85, 90$. Report the average and the variance of $\hat{\beta}_1(\tilde{p})$ across the simulated samples, for each $\tilde{p}$. Discuss the results.

(d) Let

$$\mathbf{X}_{-1}(\tilde{p}) = \begin{pmatrix} X_{12} & \cdots & X_{1\tilde{p}} \\ \vdots & \ddots & \vdots \\ X_{N2} & \cdots & X_{N\tilde{p}} \end{pmatrix}$$

and $\hat{X}_{i1}(\tilde{p})$ be the fitted value from a linear model:

$$X_{i1} = X_{i2}\gamma_2 + \cdots + X_{i\tilde{p}}\gamma_{\tilde{p}} + u.$$

Compute $\hat{X}_{11}(\tilde{p}), \cdots, \hat{X}_{N1}(\tilde{p})$ and the lowest eigenvalue of $\mathbf{X}_{-1}(\tilde{p})^\intercal \mathbf{X}_{-1}(\tilde{p})$ for each sample while varying $\tilde{p} = 5, 10, 50, 85, 90$. Report the average of $1/N \sum_i \hat{X}_{i1}(\tilde{p})^2$ and the average of the lowest eigenvalue across the simulated samples, for each $\tilde{p}$. Interpret the results.

(e) Now, let us introduce the dependence structure within the covariate. Assume that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}\right).$$

This time, we will vary N: $N = 100, 200, 500, 1000$. Simulate 1,000 samples for each N. Compute $\hat{\beta}_1(90)$ and the lowest eigenvalue of $\mathbf{X}^\intercal\mathbf{X}$. Report the average and the variance of $\hat{\beta}_1(90)$ and the average of the lowest eigenvalue across the simulated samples, for each N. Repeat this for $\rho = 0, 0.5, 0.9$. Interpret the results.

(f) Using the simulated samples you generated in **(e)** for $N = 1000$, plot the averages of eigenvalues across the simulated samples for each ρ, from the second biggest eigenvalue to the smallest eigenvalue. Also, report the average of the biggest eigenvalue for each ρ separately. Discuss the results.

(g) Lastly, we will let p grow with N. Repeat **(e)** with $p_N = 0.9N$ and again with p_N being the biggest integer smaller than $20 \cdot \log N$. Discuss the results.