Final Project Proposal

1 Introduction

When fitting a linear regression model, we typically seek to avoid high multicollinearity among the regressors, because it can inflate the variance of the estimated coefficients and weaken statistical inference. In this final project, I will use a Monte Carlo study to examine how the correlation between two regressors affects inference on their coefficients. In addition, I will investigate whether increasing the sample size can mitigate this problem.

2 Method

Consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and is independent of (x_{1i}, x_{2i}). To induce multicollinearity, we let the regressors follow a bivariate normal distribution,

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

where $\rho \in (-1, 1)$ denotes the correlation between x_{1i} and x_{2i}. For simplicity, we set $\beta_0 = \beta_1 = \beta_2 = 1$.

To study how the correlation between the regressors affects coefficient inference, I will vary

$$\rho \in \{-0.9, -0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8, 0.9\}.$$

For each value of ρ, I will generate $k = 100$ Monte Carlo samples of size $n = 500$, fit the linear regression model to each sample, and collect the resulting coefficient estimates. For every ρ, I will compute the Monte Carlo mean and variance of the estimated coefficients, and examine how these quantities change as ρ increases in magnitude. In addition, I will compute the empirical coverage rate of the nominal 95% confidence intervals for β_1 and β_2, i.e., the proportion of intervals that contain the true parameter.

Next, to assess whether a larger sample size can alleviate the effect of multicollinearity, I will fix $\rho = 0.5$ and consider

$$n \in \{500, 1000, 1500, 2000\}.$$

For each sample size, I will generate $k = 1000$ Monte Carlo samples, fit the model, and again compute the Monte Carlo mean and variance of the coefficient estimates, as well as the empirical 95% coverage rates. Comparing these results across sample sizes will show to what extent increasing n mitigates the loss of precision caused by multicollinearity.