Monte Carlo Study of Multicollinearity in Linear Regression

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

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

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}. In fact, we can always normalize regressors to mean 0 and standard deviation 1. For simplicity, we set $\beta_0 = \beta_1 = 1$, $\beta_2 = 2$.

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

$$\rho \in \{0, 0.5, 0.8, 0.9, 0.95, 0.99, 0.995, 0.999\}.$$

For each value of ρ, I will generate $k = 500$ Monte Carlo samples of size $n = 100$, fit the linear regression model to each sample, and collect the resulting coefficient estimates. For every ρ, I will compute the Monte Carlo mean, variance and significance rate (i.e., the proportion of simulations in which the coefficient is deemed statistically significant) of the estimated coefficients, and examine how these quantities changes as ρ increases in magnitude. In addition, I will compute the empirical coverage rate (i.e., the proportion of intervals that contain the

true parameter) as well as the average length of the nominal 95% confidence intervals for β_1 and β_2.

Next, to assess whether a larger sample size can alleviate the effect of multicollinearity, I will fix $\rho = 0.9$ and consider

$$n \in \{20, 30, 50, 80, 100, 150, 200, 250, 300, 500\}$$

For each sample size, I will generate $k = 500$ Monte Carlo samples, fit the model, and compute the Monte Carlo variances of the coefficient estimates. Then I will plot the relationship between sample size and the Monte Carlo variances. For comparison, I also include the corresponding curves for the case $\rho = 0.95$ and $\rho = 0.99$.

3. Result

The table below reports the values of various statistics as I vary ρ. The Monte Carlo means of β_1 and β_2 (column "beta1" and "beta2") remain close to 1 and 2, respectively, regardless of the value of ρ . This indicates that multilinearity does not affect the unbiasedness of the linear regression estimators.

However, the Monte Carlo variances of the two estimated coefficients (column "var1" and "var2") increase steadily as ρ grows. Both variances rise rapidly once ρ exceeds 0.9. From column "sr1" and "sr2", we also see as ρ approaches 1, the significance rates drop quickly. Moreover, the smaller coefficient β_1 is affected earlier and more severely than β_2 as ρ increases.

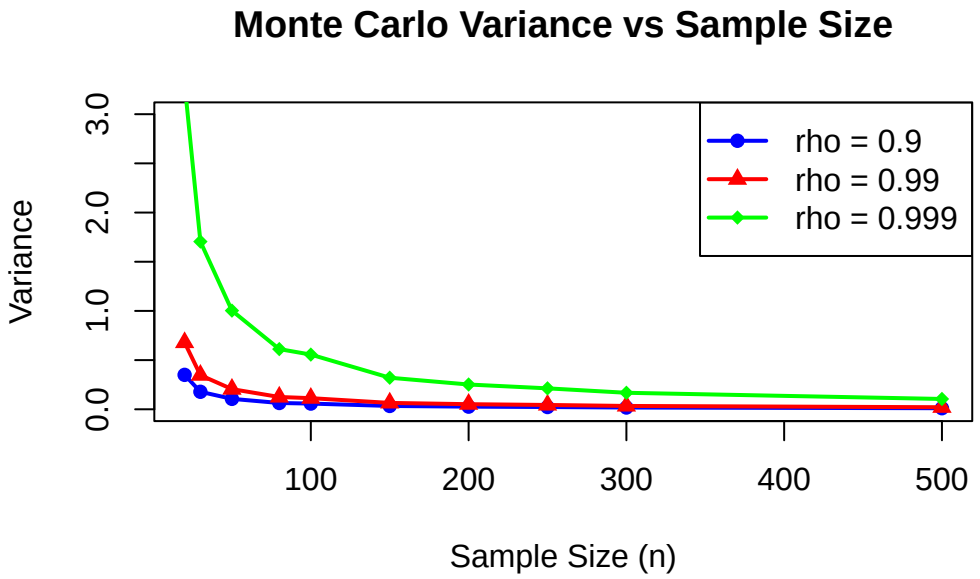
Finally, although the empirical coverage rates of the 95% confidence intervals (column "cr1" and "cr2") remain close to 95%, the average interval lengths (column "len1" and "len2") become extremely large under high multicollinearity, making the resulting inference uninformative.

	rho	beta1	beta2	var1	var2	sr1	sr2	cr1	cr2	len1	len2
1	0.000	0.997	2.006	0.011	0.009	1.000	1.000	0.954	0.956	0.403	0.403
2	0.500	0.990	1.998	0.013	0.015	1.000	1.000	0.958	0.940	0.466	0.467
3	0.800	0.998	2.006	0.031	0.030	1.000	1.000	0.942	0.948	0.675	0.675
4	0.900	0.990	2.010	0.054	0.055	0.990	1.000	0.956	0.948	0.917	0.917
5	0.950	1.006	1.991	0.101	0.102	0.866	1.000	0.954	0.960	1.292	1.292
6	0.990	1.032	1.970	0.550	0.555	0.308	0.756	0.936	0.936	2.847	2.842
7	0.995	1.059	1.939	1.050	1.047	0.202	0.454	0.958	0.954	4.040	4.038
8	0.999	0.944	2.061	5.515	5.503	0.078	0.150	0.942	0.946	8.970	8.970

The graph below shows the relationship between sample size and the Monte Carlo variances of estimated β_1 (the pattern for β_2 is nearly identical, so it is omitted). When sample size is below

100, the variance differs substantially across the three correlation levels $\rho = 0.9, 0.95, 0.99$. In particular, the variance under $\rho = 0.99$ is much larger than that of the other two values.

However, as the sample size increases, the variance gap across different values of ρ narrows rapidly. Once the sample size exceeds 300, the differences become less than 0.1. In addition, all three variances converge toward zero, approaching the variance in the ideal case with no multicollinearity ($\rho = 0$).



4. Conclusion

Inference for linear regression coefficients can be substantially affected by multicollinearity. Although the estimators remain unbiased, their variances can become much larger than in the ideal case with no multicollinearity. When the correlation between regressors is modest, this inflation may be mild; however, as ρ approaches 1, the variance increases sharply, causing coefficient estimates, especially the smaller ones, to lose statistical significance. In extreme cases, even though the 95% confidence intervals still achieve coverage rate close to 95%, the intervals become excessively wide and hence make no sense.

When sample size is small (e.g., below 100), regression estimates can suffer seriously from high multicollinearity. However, as the sample size grows, the adverse effects diminish rapidly. Once the sample size exceeds approximately 500, the coefficient variances can approach 0 even when $\rho = 0.99$, suggesting that high multicollinearity poses far less concern in large-sample settings. As a result, one must be caution when handling strongly correlated regressors in small samples, and one straightforward way to solve this issue is to increase the sample size.

Github link: https://github.com/SONG-Yunqi/STATS-506-Final-Project