Introductory Econometrics

Multicollinearity

Yaohan Chen

School of Big Data and Statistics, Anhui University

Spring, 2025

Basic Concepts

- Multivariate linear regression model expressed as follows

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i.$$

- An implicit assumption is explanatory variables $X_{i1}, \ldots, X_{ik}$ are mutually independent.
- If this assumption is violated, then we say there exists multicollinearity.

Basic Concepts

- Perfect multicollinearity

$$c_1 X_{i1} + c_2 X_{i2} + \ldots + c_k X_{ik} = 0$$

  where there exists $1 \leqslant j \leqslant k$ such that $c_j \neq 0$.

- Approximate multicollinearity

$$c_1 X_{i1} + c_2 X_{i2} + \ldots + c_k X_{ik} + \nu_i = 0$$

  where there exists $1 \leqslant j \leqslant k$ such that $c_j \neq 0$ and ν_i refers to a random variable.

Basic Concepts

- For multivariate regression model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

  multicollinearity implies $\operatorname{rank}(\boldsymbol{X}) < k + 1$.
- There exists at least one column of $\boldsymbol{X}$ can be perfectly (or approximately) linearly represented by other columns of $\boldsymbol{X}$. For instance, $X_2 = (\text{or } \approx)\lambda X_1$.

Problems of Multicollinearity

- Problems associated with perfect multicollinearity. We cannot solve OLS as (no unique solution)

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

  since $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ does not exist as $\boldsymbol{X}'\boldsymbol{X}$ is a singular matrix.
- Identification problem arises if there exists perfect multicollinearity. For instance, for multivariate regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

  it would degenerate to simple linear regression model if $X_2 = \lambda X_1$

$$Y = \beta_0 + \left(\beta_1 + \lambda\beta_2\right) X_1 + u$$

Problems of Multicollinearity

Example (Identification problem)

Consider following regression model for consumption function

$$C = \beta_0 + \beta_1 N + \beta_2 S + \beta_3 T + u$$

where C is the consumption, N is nonlabor income, S is salary, T is total income, and u refers to the stochastic error term. Since $N + S = T$, we cannot pin down the β_1, β_2 and β_3 uniquely.

Problems of Multicollinearity

- Problems associated with approximate multicollinearity. $\boldsymbol{X}'\boldsymbol{X}$ is approximately a singular matrix, $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$, which leads to the problem that the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ are large. Therefore,

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Problems of Multicollinearity

- For multivariate regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \boldsymbol{X}_{i2}\boldsymbol{\beta}_2 + u_i$$

where $\boldsymbol{X}_{i2} = (1, X_{i2}, \ldots, X_{ik})$, $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_k)'$. Then

$$\text{Var}\left(\hat{\beta}_1\right) = \sigma^2 \left(\sum x_{i1}^2\right)^{-1} \left[1 - R^2_{X_1,\boldsymbol{X}_2}\right]^{-1}$$

where $x_{i1} = X_{i1} - \bar{X}_1$ and $R^2_{X_1,\boldsymbol{X}_2}$ refers to the R^2 when running OLS of X_1 on $\boldsymbol{X}_2$.

Problems of Multicollineartiy

- Specifically when $\boldsymbol{X}_{i2}$ degenerate to a scalar X_{i2}, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$$

and

$$\operatorname{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum x_{i1}^2} \cdot \frac{1}{1 - r^2}$$

and $r^2 = R^2_{X_1, X_2}$.

- More specifically,

$$r^2 = \frac{\left(\sum x_{i1} x_{i2}\right)^2}{\sum x_{i1}^2 \sum x_{i2}^2}.$$

Problems of Multicollinearity

- When there exists multicollinearity, large variance associated with OLS estimator implies that either hypothesis testing or prediction may fail.
- The signs of coefficients estimates may not be as expected due to the identification problem or the high variance of the estimator.
- Unless there exists perfect multicollinearity, multicollinearity **DOES NOT** suggest violation of the classical assumptions.
- The OLS estimator, when faced with multicollinearity, is not “perfect” due to the potentially induced high variance.

Multicollinearity Detection

Detecting through the symptoms of multicollinearity:

- The simplest way to detect multicollinearity is to calculate the correlation matrix for the regressor (explanatory variable).
- Regressing X_1 on X_2 and using corresponding r^2 to determine whether there exists strong multicollinearity between X_1 and X_2.
- Insignificant t-statistic on all or many coefficients, but large F-statistic for testing whether all of the coefficients are zero.
- The coefficient estimates may be sensitive to the deletion of a statistically insignificant variable.

Multicollinearity Detection

- The coefficient estimates may be very sensitive to the addition of one or a small number of observations.
- One may get very odd coefficient estimates possibly with wrong signs due to the high variance of the estimator.
- Using F-statistic,

$$F_j = \frac{R_j^2/(k-1)}{(1-R_j^2)/(n-k)} \sim F(k-1, n-k)$$

where R_j^2 refers to the R^2 when regressing the jth variable on other $(k-1)$ variables included the regression model.

Remedies for Multicollinearity

- Drop variables suspected of causing the multicollinearity problem.
- To get more data. If you really want to know the separate effects of X_1 and $\boldsymbol{X}_2$ for example, you need to get as much data as possible. This may not be always possible because the data you need may be macro time series or it my be very costly to get additional data.
- Try to impose any prior linear restrictions, if any, provided by economic theory.
- Conduct ridge regression.
- **Adding regressor step-by-step**.

Adding Regressor Step-by-Step: An Example

- Grain output as the explained variable Y. We collect 6 explanatory variables, denoted by X_1 to X_6.
- Establishing the regression model by including all the collected explanatory variables,

$$\begin{aligned}\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \\ \beta_4 \ln X_4 + \beta_5 \ln X_5 + \beta_6 \ln X_6 + u\end{aligned}$$

Adding Regressor Step-by-Step: An Example

- OLS estimation yields

$$\begin{aligned}\ln\hat{Y} = \; & \underset{(0.367)}{-0.767} + \underset{(0.092)}{0.757\ln X_1} + \underset{(0.097)}{0.246\ln X_2} + 0.0002\ln X_3 \; \boxed{(0.108)} \\ & +\underset{\boxed{(0.032)}}{0.030\ln X_4} - \underset{\boxed{(0.034)}}{0.032\ln X_5} + \underset{\boxed{(0.042)}}{0.051\ln X_6}\end{aligned}$$

$$R^2 = 0.9850 \qquad \bar{R}^2 = 0.9812 \qquad F = 262.32$$

- Since $F = 262.32 > F_{0.05}(6, 24) = 2.51$, we can claim that β_1, β_2, β_3, β_4, β_5 and β_6 are not equal to zero simultaneously.
- But insignificant t-statistic for β_3, β_4, β_5 and β_6.

Adding Regressor Step-by-Step: An Example

- Regressing Y on $\ln X_1$. (Model (1))

$$\begin{array}{rcccc} \ln\hat{Y} & = & -0.684 & + & 1.004\ln X_1 \\ & & (-3.88) & & (35.14) \end{array}$$

$$R^2 = 0.9771 \qquad \bar{R}^2 = \boxed{0.9763}$$

- Adding X_2 and regressing Y on $\ln X_1$ and $\ln X_2$. (Model (2))

$$\begin{array}{rcccccc} \ln\hat{Y} & = & -0.915 & + & 0.812\ln X_1 & + & 0.238\ln X_2 \\ & & \boxed{(0.215)} & & \boxed{(0.072)} & & \boxed{(0.083)} \end{array}$$

$$\bar{R}^2 = \boxed{0.9810}$$

	(1)	(2)	(3)	(4)	(5)	(6)
$\ln X_1$	1.004	0.812	0.769	0.813	0.820	0.761
	(0.029)	(0.072)	(0.089)	(0.074)	(0.071)	(0.075)
$\ln X_2$		0.238	0.209	0.241	0.281	0.231
		(0.083)	(0.091)	(0.086)	(0.087)	(0.080)
$\ln X_3$			0.071			
			(0.088)			
$\ln X_4$				-0.005		
				(0.028)		
$\ln X_5$					-0.041	
					(0.029)	
$\ln X_6$						0.050
						(0.029)
Cons	-0.684	-0.915	-0.722	-0.930	-1.072	-0.734
	(0.222)	(0.215)	(0.321)	(0.234)	(0.238)	(0.231)
$\bar{R}^2$	0.9763	0.9810	0.9808	0.9803	0.9817	0.9823

Ridge Regression

- Multicollinearity arises when $\boldsymbol{X}'\boldsymbol{X}$ is singular or approximately singular. How can we fix it ?

Ridge Regression

- Multicollinearity arises when $\boldsymbol{X}'\boldsymbol{X}$ is singular or approximately singular. How can we fix it ?
- Ridge regression is defined a special constrained regression,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \text{s. t. } ||\boldsymbol{\beta}||^2 &\leq s\end{aligned}$$

  where $||\boldsymbol{\beta}||^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$.
- By establishing the Lagrangian we can equivalent claim that

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||^2.$$

Ridge Regression

- By taking the first order derivative with respect to $\boldsymbol{\beta}$, we obtain

$$-2\boldsymbol{X}'\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ridge}}\right)+2\lambda\hat{\boldsymbol{\beta}}_{\text{ridge}}=\boldsymbol{0}$$

  which finally yields

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}=\left(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

- $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ always exists regardless of the behavior of $\boldsymbol{X}$ since $\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I}$ is positive definite matrix.
- Ridge estimator is essentially about shrinkage, which introduces bias but reduces the variance.