Violation of Assumptions of Classical Regression Model: Multicollinearity
Unit 2

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute
New Delhi

Contents

1 What is Multicollinearity?

1.1 Definition and Concept

Multicollinearity is a phenomenon in multiple regression analysis where two or more explanatory variables are highly linearly related. In the case of *perfect multicollinearity*, one regressor is an exact linear function of others. More commonly, *imperfect* or *high multicollinearity* exists when regressors are highly, but not perfectly, correlated.

The classical linear regression model assumes that no explanatory variable is an exact linear combination of the others. Multicollinearity violates this assumption, making it difficult to estimate regression coefficients uniquely and reliably.

1.2 Mathematical Illustration

Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

If $X_2 = a + bX_1$, perfect multicollinearity exists, and OLS cannot estimate β_1 and β_2 uniquely. Even high but imperfect correlation can cause serious estimation problems.

2 Reasons for Multicollinearity

2.1 Data Collection Issues

- **Limited variation in data:** Data collected from a narrow range of observations (e.g., only large farms) can induce high correlation among explanatory variables.
- **Small sample size:** With few observations, random correlations among regressors can be high.

2.2 Model Specification Issues

- **Including variables that are functions of each other:** For example, including both total expenditure and its components.
- **Using highly correlated variables:** Such as including both age and years of experience in a wage regression.
- **Dummy variable trap:** Including all categories of a qualitative variable along with the intercept.
- **Polynomial and interaction terms:** Including X and X^2 or X and $\log(X)$, especially if X has little variation.

2.3 Economic or Physical Relationships

- **Genuine economic links:** In macroeconomic models, variables like income, consumption, and investment are often interrelated.
- **Trends in time series:** Variables with strong upward or downward trends (e.g., GDP and population) tend to be highly correlated.

2.4 Illustrative Example

In agricultural economics, fertilizer use and irrigation may be highly correlated if both increase together in modernized farms. In macroeconomic models, government expenditure and tax revenue may move closely due to fiscal policy.

3 Practical Situations Where Multicollinearity Occurs

3.1 Economic Data Examples

- **Agricultural input studies:** Fertilizer and irrigation are often used together, leading to high correlation.
- **Education and labor studies:** Years of education and years of experience are often correlated.
- **Macroeconomic models:** Consumption, income, and investment all tend to move together over time.

3.2 Dummy Variable Trap

Including k dummy variables for k categories and an intercept causes perfect multicollinearity. For example, in a model with three regions (North, South, East), including all three dummies and an intercept causes perfect collinearity.

3.3 Polynomial and Interaction Terms

Including X and X^2 or X and $\log(X)$ can induce high correlation, especially when X has little variation. This is common in models that attempt to capture nonlinear effects.

3.4 Micronumerosity

When the number of explanatory variables approaches the number of observations, even small correlations among regressors can lead to multicollinearity.

4 Detection and Identification of Multicollinearity

Detecting multicollinearity is essential for diagnosing instability in regression estimates. Both informal and formal methods are used, and interpretation is crucial.

4.1 Symptoms in Regression Output

- **High R^2 but few significant t-ratios:** The model fits well overall, but individual variables are not statistically significant.
- **Large changes in coefficients:** Coefficient estimates change dramatically with small changes in data or model specification.

- **High standard errors:** Coefficient estimates become imprecise.
- **Unstable signs and magnitudes:** Coefficient signs may flip or have implausible values.

4.2 Formal Detection Methods

1. Correlation Matrix

- **Procedure:** Calculate the pairwise correlation coefficients among all explanatory variables.
- **Interpretation:** Correlations above 0.8 or below -0.8 suggest potential multicollinearity.
- **Limitation:** High pairwise correlation is a sufficient but not necessary condition for multicollinearity; variables may be highly collinear even with low pairwise correlations if multivariate relationships exist.

2. Variance Inflation Factor (VIF)

- **Procedure:** For each regressor X_j, regress it on all other regressors and compute R_j^2. Then,
$$VIF_j = \frac{1}{1 - R_j^2}$$
- **Interpretation:** A VIF_j greater than 10 is a common rule-of-thumb for serious multicollinearity. If VIF_j is close to 1, there is little multicollinearity.
- **Example:** If $R_j^2 = 0.9$, then $VIF_j = 10$; this indicates that the variance of the OLS estimator for X_j is inflated tenfold due to collinearity.

3. Auxiliary Regressions

- **Procedure:** Regress each regressor on all others and examine the R^2.
- **Interpretation:** A high R^2 (e.g., above 0.9) indicates that the regressor is highly predictable from the others, signaling multicollinearity.

4. Condition Number / Condition Index

- **Procedure:** Compute the condition number or index from the eigenvalues of the scaled $X'X$ matrix.
- **Interpretation:** A condition index above 30 is considered evidence of strong multicollinearity.
- **Practical guidance:** Some econometric software provides this diagnostic directly.

5. Farrar-Glauber Test

- **Procedure:** This is a formal statistical test involving a sequence of hypothesis tests for the presence of multicollinearity, based on the determinant of the correlation matrix and other statistics.
- **Interpretation:** A significant test result (low p-value) indicates multicollinearity.
- **Note:** This test is less commonly used in practice than VIF or condition index.

4.3 Summary Table: Interpreting Detection Methods

Method	Procedure	Evidence of Multicollinearity
Correlation Matrix	Compute pairwise correlations among regressors	Correlation $\|r\| > 0.8$ (rule of thumb)
Variance Inflation Factor (VIF)	Regress X_j on all other X's, compute VIF_j	$VIF_j > 10$ signals serious multicollinearity
Auxiliary Regression	Regress each X_j on others, examine R^2	$R^2 > 0.9$ indicates high multicollinearity
Condition Index	Compute from eigenvalues of $X'X$	Index > 30 signals strong multicollinearity
Farrar-Glauber Test	Formal test based on correlation matrix	Significant test statistic (low p-value)

4.4 Practical Guidance

- Use more than one method for confirmation, as each test has different sensitivity.
- High VIFs or condition indices indicate the need for further investigation.
- Multicollinearity is problematic mainly for inference (coefficient interpretation), less so for prediction.

5 Consequences of Multicollinearity

5.1 Theoretical Consequences

- OLS estimators remain unbiased and consistent.
- Variances and covariances of OLS estimators increase, making estimates less precise.
- Confidence intervals for coefficients become wider.

5.2 Practical Consequences

- **Large standard errors:** Coefficient estimates become imprecise.
- **Insignificant t-ratios:** Even if variables are truly related to Y, their estimated coefficients may not be statistically significant.
- **Sensitivity:** Estimates and their standard errors can change dramatically with small changes in data or model specification.
- **High R^2 but few significant variables:** The model appears to fit well overall, but individual predictors seem unimportant.
- **Unreliable signs and magnitudes:** Coefficient signs may flip or have implausible values.

5.3 Example

In a wage regression including education and IQ, if these are highly correlated, the standard errors on both coefficients may be large, making it difficult to establish their separate effects. In macroeconomic time series, consumption and income may both trend upward, leading to high correlation and unstable coefficient estimates.

5.4 Micronumerosity and Overfitting

When the number of regressors is close to the number of observations, multicollinearity is almost inevitable, and OLS estimates become highly unstable.

6 Remedies for Multicollinearity

6.1 Data Remedies

- **Collect more data:** Increasing sample size can sometimes reduce multicollinearity, especially if the problem is due to small sample peculiarities.
- **Redesign the study:** Use a broader range of observations or experimental design if feasible.

6.2 Model Specification Remedies

- **Drop one of the correlated variables:** If two variables are nearly identical, consider omitting one.
- **Combine variables:** Aggregate highly correlated variables into an index or principal component.
- **Centering:** For polynomial terms, subtract the mean from X before squaring to reduce correlation between X and X^2.
- **Avoid the dummy variable trap:** Exclude one category when using dummy variables.

6.3 Analytical Approaches

- **Ridge Regression:** Introduce a small bias to reduce variance in the presence of severe multicollinearity.
- **Principal Component Regression:** Use principal components of regressors as predictors.

6.4 Interpretation and Cautions

If the goal is **prediction** rather than inference, multicollinearity may not be problematic. However, for interpreting individual coefficients, addressing multicollinearity is crucial. Sometimes, if the correlated variables are theoretically important, it may be better to retain them and interpret results with caution.

6.5 Do Nothing?

If multicollinearity is inherent in the data and cannot be resolved, and if the model's predictive power is satisfactory, it may be acceptable to proceed, acknowledging the limitations for inference.

Glossary

- **Multicollinearity**: The presence of high linear relationships among explanatory variables in a regression model.
- **Perfect Multicollinearity**: When one regressor is an exact linear function of others.
- **Variance Inflation Factor (VIF)**: A measure of how much the variance of an estimated regression coefficient increases due to multicollinearity.
- **Dummy Variable Trap**: Perfect multicollinearity caused by including all categories of a qualitative variable.
- **Condition Index**: A diagnostic based on the eigenvalues of the regressor correlation matrix, used to detect multicollinearity.
- **Ridge Regression**: A biased estimation technique that can reduce the impact of multicollinearity.
- **Micronumerosity**: A situation where the number of regressors approaches the number of observations, leading to near-perfect multicollinearity.

Practice Questions

1. Define multicollinearity. How does it violate the classical regression assumptions?
2. List and explain the main causes of multicollinearity in applied econometric work.
3. Give practical examples where multicollinearity is likely to occur in economic data.

4. Describe and interpret at least three methods for detecting multicollinearity in a regression model.
5. What are the consequences of multicollinearity for OLS estimation and inference?
6. Discuss at least three remedies for multicollinearity. When should each be used?
7. Is multicollinearity always a problem? Explain with reference to prediction versus inference.
8. How can the dummy variable trap be avoided in regression analysis?
9. Explain the use and interpretation of the variance inflation factor (VIF) in diagnosing multicollinearity.
10. What is ridge regression and when is it appropriate to use?

References

- Gujarati, D. N., & Porter, D. C. (2010). *Basic Econometrics* (5th ed.). McGraw-Hill. [See especially Chapter 10: Multicollinearity]
- Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning. [See discussion in the chapter on multiple regression]