Violation of Assumptions of Classical Regression Model: Model Misspecification

Unit 5

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute

New Delhi

Contents

1 What is Model Misspecification?

1.1 Definition and Concept

Model misspecification occurs when the regression model used in empirical analysis does not accurately represent the true underlying economic relationship. This can result from incorrect functional form, omitted relevant variables, inclusion of irrelevant variables, measurement errors, or incorrect assumptions about the stochastic error term. Model misspecification undermines the reliability of estimation, inference, and prediction.

1.2 Why is it a Violation?

The classical linear regression model assumes that the specified model is the correct one. Misspecification violates this assumption, potentially leading to biased, inconsistent, or inefficient estimates, and invalid statistical inference.

1.3 Types of Misspecification

- **Omitted variable bias**: Leaving out a relevant explanatory variable.
- **Inclusion of irrelevant variables**: Adding variables that do not belong.
- **Incorrect functional form**: Using a linear model when the true relationship is nonlinear.
- **Measurement error**: Errors in the measurement of variables.
- **Incorrect specification of the error term**: Assuming homoscedasticity or no autocorrelation when these are violated.

2 Reasons for Model Misspecification

2.1 Omitted Variables

- Important explanatory variables are left out due to lack of data, oversight, or theoretical misunderstanding.
- Example: Modeling crop yield as a function of fertilizer and rainfall, but omitting soil quality.

2.2 Inclusion of Irrelevant Variables

- Including variables that do not belong in the model can increase variance and reduce precision.
- Example: Adding a variable for advertising in a model where advertising has no effect on agricultural yield.

2.3 Incorrect Functional Form

- Using a linear model when the true relationship is nonlinear (e.g., quadratic, log-linear).
- Example: Modeling consumption as a linear function of income when the relationship is actually log-linear.

2.4 Measurement Error

- Errors in measuring the dependent or independent variables can bias estimates and lead to incorrect conclusions.
- Example: Using self-reported income, which may be systematically under- or over-reported.

2.5 Incorrect Specification of the Error Term

- Assuming errors are homoscedastic or uncorrelated when they are not.
- Example: Ignoring autocorrelation in time series data.

3 Practical Situations Where Misspecification Occurs

3.1 Economic Data Examples

- **Omitted variable bias**: Excluding education in a wage regression.
- **Irrelevant variable inclusion**: Adding weather variables in a model of stock returns with no theoretical justification.
- **Functional form**: Modeling Engel curves (expenditure vs. income) as linear when they are typically nonlinear.
- **Measurement error**: Using proxy variables or survey data prone to recall bias.

3.2 Panel and Time Series Data

Misspecification is common in panel and time series data when dynamics, fixed effects, or trends are ignored.

3.3 Example

In a demand model for electricity, omitting temperature as a regressor leads to omitted variable bias, while including a variable for advertising (when there is none) introduces irrelevant variable bias.

4 Detection and Identification of Model Misspecification

Detecting model misspecification is essential for ensuring the validity of regression results. Both informal and formal methods are used.

4.1 Symptoms in Regression Output

- Systematic patterns in residuals.
- Poor out-of-sample prediction.
- Unexpected signs or magnitudes of coefficients.
- High R^2 but insignificant t-statistics.

4.2 Formal Detection Methods

1. Ramsey RESET Test

- **Procedure:** Augment the regression with powers of fitted values (e.g., $\hat{Y}^2$, $\hat{Y}^3$) and test their joint significance.
- **Interpretation:** A significant F-statistic suggests functional form misspecification or omitted variables.

2. Tests for Omitted Variables

- **Procedure:** Add suspected omitted variables to the model and test their joint significance.
- **Interpretation:** If added variables are significant, the original model is misspecified.

3. Tests for Irrelevant Variables

- **Procedure:** Remove variables suspected to be irrelevant and test whether their exclusion significantly reduces model fit.
- **Interpretation:** If the model fit does not decrease, the variable is likely irrelevant.

4. Tests for Functional Form

- **Procedure:** Compare linear and nonlinear (e.g., log-linear, quadratic) models using information criteria (AIC, BIC) or likelihood ratio tests.
- **Interpretation:** Lower AIC/BIC or significant likelihood ratio test favors the correct functional form.

5. Residual Analysis

- **Procedure:** Plot residuals against fitted values or regressors.
- **Interpretation:** Systematic patterns (e.g., curvature, trends) indicate misspecification.

6. Hausman Specification Test

- **Procedure:** Compare OLS and instrumental variable (IV) estimates.
- **Interpretation:** Significant difference between OLS and IV estimates suggests endogeneity or omitted variable bias.

4.3 Summary Table: Interpreting Detection Methods

Method	Procedure	Evidence of Misspecification
Ramsey RESET	Add powers of fitted values	Significant F-statistic for added terms
Omitted Variable Test	Add suspected omitted variables	Significant coefficients for added variables
Irrelevant Variable Test	Remove suspected irrelevant variables	No significant loss in fit or explanatory power
Functional Form Test	Compare different model forms (AIC/BIC/LR)	Lower AIC/BIC or significant LR test for alternative form
Residual Analysis	Plot residuals vs. fitted values or regressors	Systematic patterns (curvature, trend, heteroscedasticity)
Hausman Test	Compare OLS and IV estimates	Significant difference indicates endogeneity/omitted variable bias

4.4 Practical Guidance

- Always begin with residual analysis and theory-based model building.
- Use formal tests to confirm suspicions raised by informal analysis.
- Consider out-of-sample prediction and cross-validation to assess model adequacy.

5 Consequences of Model Misspecification

5.1 Theoretical Consequences

- OLS estimators may become biased and inconsistent (especially with omitted variables or endogeneity).

- Variances of estimators may be inflated or deflated.
- Standard errors and test statistics become unreliable.
- Model predictions may be systematically wrong.

5.2 Practical Consequences

- **Policy errors:** Policy recommendations based on a misspecified model can be misleading or harmful.
- **Misleading inference:** Hypothesis tests and confidence intervals are invalid.
- **Poor forecasting:** Out-of-sample predictions are unreliable.

5.3 Example

Omitting education from a wage regression biases the estimated return to experience. Using a linear model for a relationship that is actually quadratic can result in systematic under- or over-prediction at the extremes.

6 Remedies for Model Misspecification

6.1 Theory-Driven Model Building

- Use economic theory to guide variable selection and functional form.
- Avoid purely data-driven specification.

6.2 Specification Tests and Model Selection

- Apply specification tests (RESET, omitted variable, Hausman, etc.) to check model adequacy.
- Use information criteria (AIC, BIC) to compare alternative models.

6.3 Data Improvement and Measurement

- Improve measurement accuracy of variables.
- Use instrumental variables or proxy variables when direct measurement is not possible.

6.4 Functional Form Flexibility

- Consider nonlinear, log-linear, or interaction terms if theory or data suggest.
- Use flexible modeling approaches (e.g., splines, polynomials) if justified.

6.5 Iterative Model Refinement

- Use diagnostic checking and iterative respecification to improve model fit.
- Validate model using out-of-sample or cross-validation techniques.

6.6 Interpretation and Cautions

If the model is used for policy or inference, careful specification and testing are crucial. For prediction, a model's predictive accuracy may be more important than perfect specification.

Glossary

- **Model Misspecification**: The use of an incorrect or incomplete regression model.
- **Omitted Variable Bias**: Bias resulting from leaving out relevant explanatory variables.
- **Irrelevant Variable**: A variable included in the model that does not belong.
- **Functional Form**: The mathematical relationship between dependent and independent variables.
- **Ramsey RESET Test**: A diagnostic test for general model misspecification.
- **Hausman Test**: A test for endogeneity or omitted variable bias by comparing OLS and IV estimates.
- **Specification Error**: Any error resulting from incorrect model specification.

Practice Questions

1. Define model misspecification. What are its main types?
2. Explain omitted variable bias and provide an economic example.
3. How does including irrelevant variables affect regression results?
4. Describe and interpret the Ramsey RESET test.
5. What are the consequences of model misspecification for OLS estimation and inference?
6. Discuss at least three remedies for model misspecification.
7. How can residual analysis help detect misspecification?
8. What is the Hausman test and when is it appropriate to use?
9. Why is it important to use theory in model specification?
10. How can out-of-sample validation aid in model selection?

References

- Gujarati, D. N., & Porter, D. C. (2010). *Basic Econometrics* (5th ed.). McGraw-Hill. [See especially Chapter 13: Econometric Modeling—Model Specification and Diagnostic Testing]
- Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning. [See discussion in chapters on model specification and diagnostic testing]