Other Issues in Regression
Unit 6

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute
New Delhi

Contents

1 Other Issues in Regression Analysis **2**
- 1.1 Outliers, Leverage, and Influential Observations . . . 2
- 1.2 Measurement Error . . . 2
- 1.3 Endogeneity and Simultaneity . . . 2
- 1.4 Structural Breaks and Parameter Instability . . . 3
- 1.5 Omitted Variable Bias and Irrelevant Variables . . . 3
- 1.6 Nonlinearity and Functional Form Misspecification . . . 3
- 1.7 Data Issues: Missing Data, Sample Selection, and Non-random Sampling 3
- 1.8 Non-normality of Errors . . . 4
- 1.9 Multilevel and Hierarchical Data . . . 4

2 Comprehensive Comparison Table of Regression Issues **5**

1 Other Issues in Regression Analysis

While classical regression analysis focuses on violations such as multicollinearity, heteroscedasticity, autocorrelation, and model misspecification, several other important issues can affect the reliability, interpretation, and application of regression results. This unit introduces and discusses such issues, drawing from advanced chapters in both *Basic Econometrics* (Gujarati & Porter) and *Introductory Econometrics* (Wooldridge).

1.1 Outliers, Leverage, and Influential Observations

- **Outliers** are data points with unusual values for the dependent or independent variables. They can disproportionately affect regression estimates.
- **Leverage points** are observations with extreme values for the explanatory variables. High-leverage points can pull the regression line toward themselves.
- **Influential observations** are those that, if omitted, would substantially change the regression estimates. Influence combines leverage and outlier status.
- **Detection:** Use standardized residuals, leverage statistics (hat values), Cook's distance, and DFBETAS.
- **Remedies:** Investigate data quality, consider robust regression methods, or re-estimate models excluding problematic points.

1.2 Measurement Error

- **Measurement error** occurs when the variables used in regression are measured with error, either in the dependent or independent variables.
- **Consequences:** Measurement error in the dependent variable increases variance but does not bias OLS coefficients. Measurement error in independent variables leads to attenuation bias (coefficients biased toward zero) and inconsistency.
- **Detection:** Compare with alternative data sources, use validation samples, or check for implausible results.
- **Remedies:** Use instrumental variables, better data, or error-in-variables models.

1.3 Endogeneity and Simultaneity

- **Endogeneity** arises when an explanatory variable is correlated with the error term, often due to omitted variables, measurement error, or simultaneity.
- **Simultaneity** occurs when causality runs in both directions (e.g., supply and demand).
- **Consequences:** OLS estimators become biased and inconsistent.
- **Detection:** Hausman test, theory-based reasoning, or using lagged variables.
- **Remedies:** Instrumental variables (IV), two-stage least squares (2SLS), or structural modeling.

1.4 Structural Breaks and Parameter Instability

- **Structural breaks** occur when the relationship between variables changes at certain points (e.g., due to policy changes, crises).
- **Parameter instability** refers to regression coefficients changing over time or across subgroups.
- **Detection:** Chow test, recursive estimation, CUSUM plots, or rolling regressions.
- **Remedies:** Model sub-periods separately, include interaction/dummy variables, or use time-varying parameter models.

1.5 Omitted Variable Bias and Irrelevant Variables

- **Omitted variable bias** occurs when a relevant explanatory variable is left out, causing bias in estimated coefficients.
- **Inclusion of irrelevant variables** increases variance but does not bias estimates.
- **Detection:** Theory-based reasoning, specification tests (e.g., Ramsey RESET), and changes in coefficient estimates when variables are added or removed.
- **Remedies:** Careful model selection, use of theory, and specification tests.

1.6 Nonlinearity and Functional Form Misspecification

- **Functional form misspecification** arises when the true relationship is nonlinear but a linear model is used.
- **Detection:** Residual plots, RESET test, comparison of alternative models (e.g., log-linear, quadratic).
- **Remedies:** Transform variables, use nonlinear models, or flexible functional forms.

1.7 Data Issues: Missing Data, Sample Selection, and Non-random Sampling

- **Missing data** can lead to loss of efficiency or bias if not missing at random.
- **Sample selection bias** occurs when the sample is not representative of the population.
- **Detection:** Compare sample and population characteristics, analyze missingness patterns.
- **Remedies:** Imputation, Heckman correction, or use of sample weights.

1.8 Non-normality of Errors

- **Non-normality** of errors affects the validity of small-sample inference (t, F tests).
- **Detection:** Histogram and Q-Q plots of residuals, formal tests (Jarque-Bera, Shapiro-Wilk).
- **Remedies:** Transformations, robust inference, or bootstrap methods.

1.9 Multilevel and Hierarchical Data

- **Hierarchical data** occur when observations are nested (e.g., students within schools).
- **Consequences:** Ignoring clustering can underestimate standard errors.
- **Detection:** Recognize data structure; intraclass correlation.
- **Remedies:** Use clustered standard errors or multilevel (mixed-effects) models.

2 Comprehensive Comparison Table of Regression Issues

Issue	**What is it?**	**Consequences**	**Detection**	**Remedies**
Multicollinearity	High linear correlation among explanatory variables	Large standard errors, imprecise estimates, unstable coefficients, insignificant t-ratios despite high R^2	Correlation matrix, VIF (> 10), auxiliary regressions, condition index (> 30)	Drop or combine variables, collect more data, principal component or ridge regression, centering, do nothing if prediction is main goal
Heteroscedasticity	Non-constant variance of error terms across observations	OLS still unbiased, but inefficient; standard errors incorrect, inference invalid	Residual plots (fan shape), Breusch-Pagan, White, Goldfeld-Quandt tests	Robust standard errors, transform variables (logs), weighted least squares, model specification
Autocorrelation	Error terms are correlated across time (or space)	OLS unbiased (if regressors exogenous), but inefficient; standard errors wrong, inference invalid	Residual/ACF plots, Durbin-Watson, Breusch-Godfrey, runs test	GLS, Newey-West standard errors, add lags, model dynamics, correct functional form
Model Misspecification	Model does not represent the true relationship (wrong variables, form, or error structure)	Bias, inconsistency, unreliable inference, poor prediction	Residual analysis, RESET test, omitted/irrelevant variable tests, Hausman test, information criteria	Specification tests, theory-driven modeling, add relevant variables, correct functional form, use IV/2SLS if endogeneity
Outliers/Leverage/Influential Points	Unusual or extreme observations in Y or X	Can distort OLS estimates, inflate standard errors, affect inference	Standardized residuals, leverage (hat) values, Cook's distance, DFBETAS	Check data quality, robust regression, re-estimate without influential points
Measurement Error	Errors in measuring variables (especially regressors)	In X: bias and inconsistency (attenuation); in Y: increased variance	Compare with alternative sources, validation subsamples, implausible results	Use better data, instrumental variables, error-in-variables models
Endogeneity/Simultaneity	Explanatory variable correlated with error term (due to omitted variables, simultaneity, or measurement error)	OLS is biased and inconsistent	Hausman test, theoretical reasoning, lagged variables	Instrumental variables, 2SLS, structural modeling

Issue	What is it?	Consequences	Detection	Remedies
Structural Breaks/Parameter Instability	Relationships change over time or across groups (e.g., policy changes, crises)	Coefficients change, model fit deteriorates, inference invalid	Chow test, recursive estimation, CUSUM, rolling regressions	Model sub-periods, use dummies/interactions, time-varying parameter models
Omitted/Irrelevant Variables	Leaving out relevant variables (omitted) or including unnecessary ones (irrelevant)	Omitted: bias; Irrelevant: higher variance	Specification tests, theory, changes in coefficients when variables added/removed	Use theory, specification tests, careful model selection
Functional Form Misspecification	Using an incorrect (often linear) relationship	Bias, poor fit, invalid inference	Residual plots, RESET test, compare alternative models	Transform variables, nonlinear/flexible models
Missing Data/Sample Selection	Some data are missing or sample is not random	Loss of efficiency, bias if not missing at random	Compare sample to population, analyze missingness	Imputation, Heckman correction, sample weights
Non-normality of Errors	Error terms are not normally distributed	Affects small-sample inference (t, F tests)	Histogram, Q-Q plots, Jarque-Bera/Shapiro-Wilk tests	Transformations, robust/bootstrapped inference
Hierarchical/Clustered Data	Data are nested (e.g., students in schools)	Underestimated standard errors if clustering ignored	Recognize structure, intraclass correlation	Clustered standard errors, multilevel models

Glossary

- **Outlier**: An observation with an unusual value for the dependent or independent variable.
- **Leverage**: The potential of an observation to influence the fit of a regression model due to its extreme value in the predictor space.
- **Influential Observation**: An observation whose removal significantly changes the regression estimates.
- **Measurement Error**: The difference between the true value and the measured value of a variable.
- **Endogeneity**: When an explanatory variable is correlated with the error term.
- **Simultaneity**: When causality runs in both directions between variables.
- **Structural Break**: A change in the relationship between variables at a certain point in time.
- **Parameter Instability**: Variation in regression coefficients over time or groups.
- **Omitted Variable Bias**: Bias in coefficient estimates due to leaving out relevant variables.
- **Functional Form Misspecification**: Using the wrong mathematical relationship between variables.
- **Sample Selection Bias**: Bias arising when the sample is not representative of the population.
- **Non-normality**: When the error terms do not follow a normal distribution.
- **Hierarchical Data**: Data with a nested structure (e.g., students within schools).

Practice Questions

1. Define outliers, leverage, and influential observations. How do they affect regression results?
2. What is measurement error? Discuss its consequences and remedies in regression analysis.
3. Explain the concept of endogeneity and simultaneity. How can they be detected and addressed?
4. What are structural breaks? How can they be detected and what are the remedies?
5. Discuss the difference between omitted variable bias and inclusion of irrelevant variables.
6. How can functional form misspecification be detected and corrected?

7. What are the potential issues with missing data and sample selection? How can they be addressed?

8. Why is non-normality of errors a concern in regression? How can it be detected and remedied?

9. Compare and contrast the issues of multicollinearity, heteroscedasticity, autocorrelation, and model misspecification.

10. Using the comparison table, summarize the main detection methods and remedies for each major regression issue.

References

- Gujarati, D. N., & Porter, D. C. (2010). *Basic Econometrics* (5th ed.). McGraw-Hill.

- Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning.