Classical Linear Regression
Unit 1

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute
New Delhi

Contents

1 Overview and Motivation

The classical linear regression model (CLRM) is the cornerstone of empirical economic analysis. It allows us to quantify and test the relationship between a dependent variable and one or more explanatory variables, providing a rigorous framework for estimation, inference, and prediction. This unit covers the two-variable regression model, properties of OLS estimators, and the extension to multiple regression.

1.1 Why Study Regression?

Regression analysis is indispensable in economics for:

- Quantifying causal and associative relationships (e.g., effect of education on wages).
- Testing economic hypotheses (e.g., is marginal propensity to consume less than one?).
- Forecasting (e.g., predicting crop yields from input use).
- Controlling for confounding influences in observational data.

1.2 Example: Agricultural Yield

Suppose we wish to study how fertilizer use (X) affects wheat yield (Y). The regression model provides a systematic way to estimate the average effect of fertilizer, accounting for random variation.

2 The Two-Variable Regression Model

2.1 Model Specification

The simple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where:

- Y_i = dependent variable (e.g., wheat yield for farm i)
- X_i = independent variable (e.g., fertilizer applied on farm i)
- β_0 = intercept (expected value of Y when $X = 0$)
- β_1 = slope (change in Y for a one-unit change in X)
- u_i = error term (captures all other factors affecting Y)

2.2 Conditional Mean Function

The regression function specifies the expected value of Y given X:

$$E[Y|X] = \beta_0 + \beta_1 X$$

On average, Y changes by β_1 units for each unit increase in X.

2.3 Population vs. Sample Regression

- **Population regression**: The true but unknown relationship in the population.
- **Sample regression**: The estimated relationship using observed data:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- **Residual**: $\hat{u}_i = Y_i - \hat{Y}_i$

2.4 Interpreting Coefficients: Example

Suppose we estimate:

$$\hat{Y}_i = 2.5 + 0.8X_i$$

Interpretation:

- The intercept (2.5) is the estimated average yield when no fertilizer is applied.
- The slope (0.8) means that, on average, each additional unit of fertilizer increases yield by 0.8 units.

3 Estimation by Ordinary Least Squares (OLS)

3.1 OLS Principle

The OLS method estimates β_0 and β_1 by minimizing the sum of squared residuals:

$$S = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

where b_0 and b_1 are estimators of β_0 and β_1.

3.2 OLS Estimator Formulas

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where $\bar{X}$ and $\bar{Y}$ are sample means.

3.3 Worked Example

Suppose for five farms, the data on fertilizer (X) and yield (Y) are as follows:

Farm (i)	Fertilizer (X_i)	Yield (Y_i)
1	2	5
2	4	7
3	6	9
4	8	10
5	10	13

Calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ using the above formulas.

4 Properties of OLS Estimators (Gauss–Markov Theorem)

4.1 Desirable Properties

Under the classical assumptions, OLS estimators possess several important properties:

- **Linearity**: OLS estimators are linear functions of the observed Y_i.
- **Unbiasedness**: $E[\hat{\beta}_j] = \beta_j$ for $j = 0, 1$; on average, OLS estimates hit the true parameter values.
- **Minimum Variance (Efficiency)**: Among all linear unbiased estimators, OLS estimators have the smallest variance (they are BLUE: Best Linear Unbiased Estimators).
- **Consistency**: As sample size increases, OLS estimators converge in probability to the true parameter values.

4.2 The Gauss–Markov Theorem

If the classical linear regression assumptions hold, the OLS estimators are the Best Linear Unbiased Estimators (BLUE) of the regression coefficients.

4.3 Implications and Example

- **Unbiasedness**: If we repeatedly sample data and estimate the regression, the average of $\hat{\beta}_1$ will equal the true β_1.
- **Efficiency**: No other linear, unbiased estimator has a smaller variance than OLS.
- **Example**: In estimating the effect of fertilizer on yield, OLS ensures that, on average, our estimate is correct and as precise as possible given the data and assumptions.

4.4 Variance of OLS Estimators

The variance of $\hat{\beta}_1$ is:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where σ^2 is the variance of the error term.

5 Extension to Multiple Regression

5.1 Motivation

Economic phenomena are rarely explained by a single factor. Multiple regression allows us to estimate the effect of one variable while controlling for others.

5.2 The Multiple Linear Regression Model

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

where:

- Y_i = dependent variable
- X_{ji} = jth explanatory variable for observation i
- β_j = coefficient for X_j
- u_i = error term

5.3 Interpretation of Coefficients

Each β_j measures the effect of a one-unit increase in X_j on Y, holding all other variables constant.

5.4 Example: Crop Yield

Suppose we model wheat yield as a function of fertilizer (X_1), rainfall (X_2), and seed variety (X_3):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Here, β_1 is the effect of fertilizer on yield, controlling for rainfall and seed variety.

5.5 OLS Estimation in Multiple Regression

The OLS estimators are obtained by minimizing the sum of squared residuals:

$$S = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

The solution yields estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

5.6 Partial Regression and Partial Effects

Multiple regression isolates the *partial effect* of each variable, allowing for more accurate estimation in the presence of confounding factors.

5.7 Goodness-of-Fit: R^2 and Adjusted R^2

- R^2 measures the proportion of variance in Y explained by all X variables.
- Adjusted R^2 corrects for the number of variables, penalizing overfitting.

5.8 Practical Example

In a wage regression, including education, experience, and gender as regressors allows us to estimate the effect of education on wages, net of experience and gender effects.

6 Assumptions of the Classical Linear Regression Model

The reliability of OLS estimates depends on several critical assumptions. Each is explained below, with practical examples.

6.1 1. Linearity in Parameters

Assumption: The model is linear in the parameters (β_j).

Example: $Y = \beta_0 + \beta_1 X + u$ is linear. $Y = \beta_0 + \beta_1 X^2 + u$ is also linear in parameters.

6.2 2. Random Sampling

Assumption: The data are a random sample from the population.

Example: If a survey of household income is conducted only in urban areas, the sample is not random and may not represent the entire population.

6.3 3. No Perfect Multicollinearity

Assumption: No explanatory variable is an exact linear function of others.

Example: In a multiple regression, including both "total expenditure" and "income" as regressors when they are always equal leads to perfect multicollinearity.

6.4 4. Zero Conditional Mean

Assumption: $E[u|X] = 0$. The error term has zero mean given any values of the explanatory variables.

Example: If unobserved ability affects both education and wages, omitting ability violates this assumption.

6.5 5. Homoscedasticity

Assumption: The variance of the error term is constant for all values of X.

Example: In income regressions, error variance may increase with income, violating homoscedasticity.

6.6 6. No Autocorrelation

Assumption: Error terms are uncorrelated across observations.

Example: In time series data, if weather shocks persist over years, errors in crop yield regressions may be autocorrelated.

6.7 7. Normality of Errors

Assumption: The error term is normally distributed.

Example: In small samples, normality is important for valid hypothesis testing. In large samples, the Central Limit Theorem mitigates this requirement.

7 Goodness-of-Fit and Interpretation

7.1 The Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

R^2 measures the proportion of the variance in Y explained by the model.

7.2 Adjusted R^2

Adjusted R^2 accounts for the number of regressors and sample size, providing a more accurate measure for model comparison.

7.3 Standard Error of Regression

The standard error of the regression (SER) measures the average distance that the observed values fall from the regression line.

7.4 Statistical Inference

- **t-tests:** Test whether coefficients are significantly different from zero.
- **Confidence intervals:** Provide a range of plausible values for coefficients.

8 Common Pitfalls and Practical Issues

8.1 Omitted Variable Bias

Leaving out a relevant variable that affects both Y and X biases the estimated effect of X.

8.2 Reverse Causality

If causality runs from Y to X, OLS estimates may be misleading.

8.3 Nonlinearity

If the true relationship is nonlinear but a linear model is estimated, the model may fit poorly.

8.4 Outliers and Influential Observations

Extreme values can disproportionately affect estimates. Always visualize data and check for outliers.

Glossary

- **Regression Analysis**: A statistical technique for estimating relationships among variables.
- **Ordinary Least Squares (OLS)**: Estimation method that minimizes the sum of squared residuals.
- **Residual**: The difference between the observed and predicted value of the dependent variable.
- **Homoscedasticity**: The property that the error variance is constant across all observations.
- **Autocorrelation**: Correlation of error terms across observations, often a problem in time series data.
- **Zero Conditional Mean**: The assumption that the error term has zero mean given any value of the independent variable(s).
- **Coefficient of Determination (R^2)**: The proportion of the variance in the dependent variable explained by the independent variable(s).
- **Omitted Variable Bias**: Bias in coefficient estimates resulting from leaving out a relevant variable.
- **BLUE**: Best Linear Unbiased Estimator.
- **Partial Regression Coefficient**: The effect of one explanatory variable on the dependent variable, holding other variables constant.

Practice Questions

1. Define the simple linear regression model. Provide an economic example.
2. Explain the OLS estimation procedure and derive the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$.
3. What does the slope coefficient represent in a two-variable regression? Illustrate with an example.
4. List and explain the properties of OLS estimators under the Gauss–Markov theorem.
5. What is the consequence of violating the zero conditional mean assumption?
6. How is the R^2 statistic interpreted in regression analysis?
7. Discuss the difference between population and sample regression functions.
8. What is omitted variable bias? Provide an example.
9. Why is it important to check for outliers in regression analysis?
10. How does the multiple regression model improve upon the two-variable model?

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.