GAEE
Global Association of Economics Education
INDIA

# REGRESSION PLAYBOOK

*FOUNDATIONS OF APPLIED ECONOMETRICS*

*FIRST EDITION*

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# CHAIRPERSON'S NOTE

I am thrilled to present the first edition of the **Regression Playbook**.

As an economics student, my days revolve around data and numbers. Working with data can often seem intimidating at first; however, I have learned that with the right guidance and approach, it becomes one of the most powerful tools for understanding crucial economic questions around us. As Chairperson, my vision has always been to make economic education more practical. Our aim at GAEE is to bridge the gap between theory and application by making regression more intuitive and accessible. This Playbook is one such effort, bringing together practical examples, case studies, and econometric applications.

I am incredibly grateful to my entire team for their efforts. Putting this together has been a great learning experience for our whole team, and we hope it turns out to be one for you too.

Warm Regards,
Aanya Narula

# EDITOR'S NOTE

Numbers rarely speak for themselves. Regression analysis helps economists uncover relationships, test hypothesis, and draw meaningful insights from data, but interpreting results requires more than statistical technique. It demands an understanding of the assumptions, limitations, and economic reasoning behind the model.

This playbook is designed to make regression more approachable, combining intuitive explanations, practical examples, and structured workflows to bridge theory and application.

Whether you're learning regression for the first time or strengthening your understanding, we hope this playbook serves as a valuable guide in your research journey.

Warm Regards,
Mitakshra, Neerja, and Gurnoor

# ABOUT THE PLAYBOOK

## PURPOSE

This playbook bridges econometric theory and real-world practice through hands-on applications in R and Excel, supported by case studies and policy examples. It emphasises understanding, implementing, and interpreting regression models, helping learners connect theory with empirical analysis while developing practical skills for research and placement interviews.

## IMPORTANCE OF APPLIED ECONOMETRICS

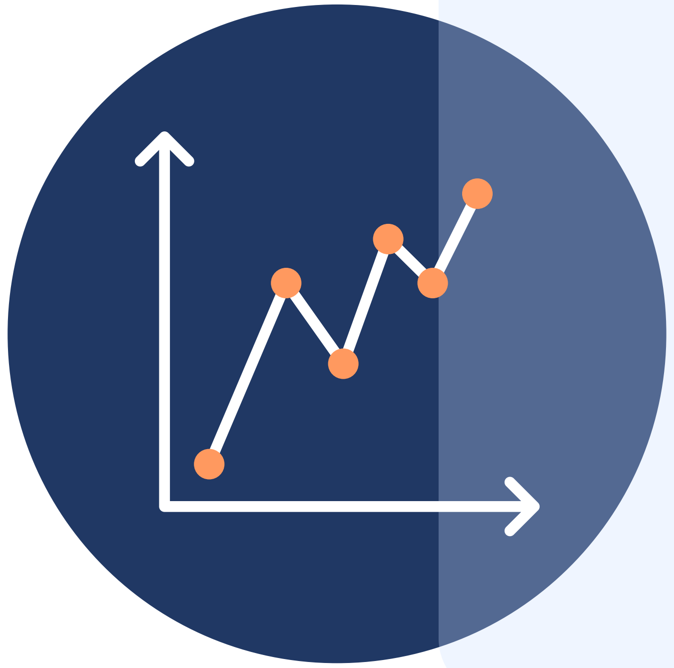
- Applies statistical methods to real-world economic data.
- Uses theory-based models to explain and predict outcomes.
- Relies on regression to estimate relationships between variables.
- Bridges economic theory and empirical evidence.
- Widely used in economics, business, and social sciences.
- Requires a strong framework for hypothesis testing.

## HOW IS IT DIFFERENT FROM STANDARD BOOKS?

Unlike standard textbooks that focus on theory, this playbook emphasises real-world applications of econometrics using R and Excel. It provides a structured approach to data analysis, helping students understand concepts such as multicollinearity and heteroskedasticity while guiding them through the process of building and interpreting econometric models. Ultimately, it makes econometric analysis more practical, consistent, and easier to understand.

# FOUNDATIONS OF REGRESSION

# WHAT IS REGRESSION?



Regression analysis examines how independent variables affect a dependent variable. Linear regression fits a straight line to the data.

It measures the relationship's direction, strength, and predictive power. Model performance is assessed using $R^2$ and error measures.

## WHAT IS THE PURPOSE OF REGRESSION ANALYSIS IN ECONOMICS?

- QUANTIFY RELATIONSHIPS BETWEEN VARIABLES
- DETERMINE THE IMPACT OF ONE VARIABLE, CONTROLLING FOR OTHER VARIABLES
- VERIFY ECONOMIC THEORY USING ACTUAL OBSERVATIONS
- FORECAST FUTURE DEVELOPMENTS
- GUIDE POLICYMAKERS BY PROVIDING NUMERICAL ESTIMATES
- HANDLE LARGE DATASETS WITH MANY VARIABLES SIMULTANEOUSLY
- ESTABLISH THE MAGNITUDE AND DIRECTION OF IMPACTS, RATHER THAN SIMPLY THEIR PRESENCE
- IDENTIFY PATTERNS THAT ARE NOT VISIBLE THROUGH SIMPLE OBSERVATION

## FORECASTING RETAIL SALES- A BUSINESS EXAMPLE

A retail chain used multiple regression to show that every ₹1 lakh increase in digital ad spend was linked to ₹4.2 lakh in added revenue, improving marketing budgets and forecasts.

## POLICY EXAMPLE 1 –CARD AND KRUEGER (1994)

A regression study comparing fast-food employment in New Jersey and Pennsylvania found that higher minimum wages did not necessarily increase unemployment, shaping U.S. minimum wage debates.

## POLICY EXAMPLE 2 –OREGON MEDICAID LOTTERY (2008)

Oregon's Medicaid lottery study found that coverage reduced financial hardship and improved mental health, but had limited short-term physical health benefits, informing debates on Medicaid expansion.

# CORRELATION VS CAUSATION

## CONCEPTUAL DIFFERENCE

| PARAMETER | CORRELATION | CAUSATION |
|---|---|---|
| DEFINITION | Exists when one variable changes and another tends to change in a specific direction (no cause-effect link implied. | Exists when a change in one variable directly triggers a change in another. |
| KEY CHARACTERISTIC | Indicates two variables move together, but does not confirm one causes the other. | Indicates a direct cause-and-effect relationship — X causes Y. Causation generally implies correlation, but correlation does not imply causation. |
| EXAMPLE | As height increases, weight tends to increase (positive correlation); as absences rise, grades tend to fall (negative correlation). | Increasing the dosage of a medicine causes symptoms to decrease. |

## MISCONCEPTIONS

1. Association alone doesn't prove causation - establishing it requires randomized experiments or other sound identification methods with proper data collection.
2. Omitting an important factor can distort results, wrongly attributing its effect to other variables
3. Causality can run in either direction - analysis must account for reverse causality or simultaneity, where variables influence each other in a feedback loop
4. Some correlations arise only because variables follow the same broader trend over time, not because they're logically connected
5. Statistical significance only shows a link is unlikely due to chance - it doesn't explain why the link exists or prove causation.

# STRUCTURE OF A REGRESSION MODEL

A regression model uses mathematical methods to determine how different variables interact with each other. The model enables us to predict one variable's value using information about other variables.

**Dependent Variable (Y)**
Also known as the outcome. This is the factor you are trying to predict or explain. The value of this variable depends on other factors.

**Independent Variable (X)**
Also known as the predictor. This is the factor you believe has an impact on the dependent variable.

**Coefficient (β)**
The estimated parameter that quantifies the relationship between each predictor and the outcome. $\beta_0$ is the intercept; each $\beta_i$ shows how Y changes with a one-unit increase in $X_i$, holding all other variables constant.

## EXAMPLE: STUDY HOURS AND EXAM SCORES

**Y (Dependent): Exam Score** → **X (Independent): Hours Studied** → **β (Coefficient): Points per hour**

# STRUCTURE OF REGRESSION MODEL

GAEE Global Association of Economics Education INDIA

## ROLE OF THE ERROR TERM

In real-world situations, independent variables cannot explain all variations in the dependent variable. The error term (or residual) captures this unexplained portion by accounting for omitted or unmeasured variables, random factors, and measurement errors in the data.

For example, a student may report studying for 4 hours when they actually studied for only 2.5 hours. The classical regression model assumes that error terms have a mean of zero and constant variance.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

**The Goal:** In a "good" model, we want the error term to be as small and random as possible, meaning the X variables explain a large portion of Y.

### WHAT MAKES UP THE STATISTICAL MODEL?

**Model prediction** $\beta_0 + \beta_1 X$ **+** **ERROR TERM: epsilon (ε)** Unexplained variation **=** **ACTUAL Y** Exam score

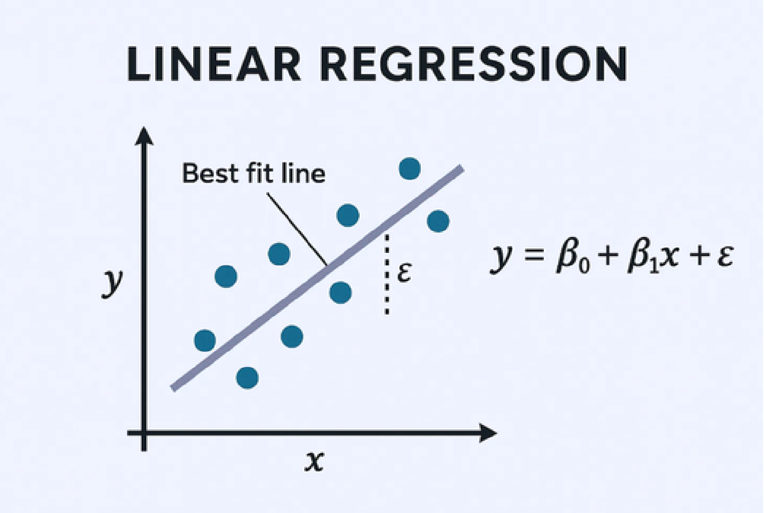
**ε captures these parts the model missed:**

- **OMITTED VARIABLES** Prior knowledge, exam difficulty
- **RANDOM VARIATION** A student just had a bad day
- **MEASUREMENT ERROR** Hours logged inaccurately

## THE "LINE OF BEST FIT"

The basic idea of regression involves locating a trend line that lies closest to all scatterplot data points.

- The Intercept ($\beta_0$): Where the line hits the vertical axis (the value of Y when X is zero).
- The Slope ($\beta_1$): The slope of the line indicates its steepness. It shows the expected Y increase for each unit X increase.
- The Prediction: By establishing this "best fit" line, we can move from simply looking at past data to forecasting future outcomes.




Source: statistical aid

The slope together with the intercept allows us to use any X value for predicting values of Y

### COMMON VARIANTS

- **Simple linear** One continuous predictor
- **Multiple linear** Several predictors
- **Logistic** Binary outcome
- **Polynomial** Non-linear relationships

# THE LINEAR REGRESSION MODEL

Linear Regression is a statistical tool that allows researchers to capture and ascertain the relationship between a Dependent Variable (also referred to as the target or outcome variable) and one or more independent variables (the predictors used to support the model).

## DEFINING THE RELATIONSHIP

Regression models use a mathematical formula to describe how changes in X affect Y.

$$Y = B_0 + B_1 X + E$$

Y = DEPENDENT VARIABLE

X = INDEPENDENT VARIABLE

$B_0$ (INTERCEPT) = CONSTANT TERM

$B_1$ (SLOPE COEFFICIENT) = EFFECT OF X ON Y

E (ERROR TERM) = UNEXPLAINED FACTORS AFFECTING Y

## MEANING OF THE PARAMETERS

- **Intercept ($\beta_0$):** Value of Y when X = 0 (baseline).
- **Slope ($\beta_1$):** Change in Y for a one-unit change in X, indicating the relationship's direction and strength.
- **Error Term ($\varepsilon$):** Unexplained variation in Y due to omitted factors, measurement errors, or randomness.

## ELASTICITY IN REGRESSION

Elasticity measures how responsive Y is to a 1% change in X, expressed as the percentage change in Y.

- **Elastic (>1):** Y changes more than proportionally.
- **Inelastic (<1):** Y changes less than proportionally.
- **Unitary (=1):** Y changes exactly proportionally.

In regression, elasticity helps predict how changes in one variable affect another. For example, a price elasticity of -2 means a 1% price increase leads to a 2% fall in demand, helping policymakers assess policy impacts.

# ORDINARY LEAST SQUARES (OLS)

## OLS FRAMEWORK

### MODEL REPRESENTATION

- **Y:** Dependent (outcome) variable
- **$X_1, X_2, \dots, X_n$:** Independent (predictor) variables
- **$\beta_0$:** Intercept (constant)
- **$\beta_1, \beta_2, \dots, \beta_n$:** Regression coefficients
- **ε:** Error term (unexplained variation)

### MINIMISING ERRORS

- Estimates the coefficients ($\beta_0, \beta_1, \beta_2, \dots, \beta_n$)
- Minimizes the sum of squared residuals (errors)
- Finds the best-fitting regression line for the data

### ESTIMATING COEFFICIENTS

- Estimates regression coefficients ($\beta_0, \beta_1, \dots, \beta_n$)
- Minimizes the sum of squared errors
- Produces the best-fitting regression equation

### ASSUMPTIONS

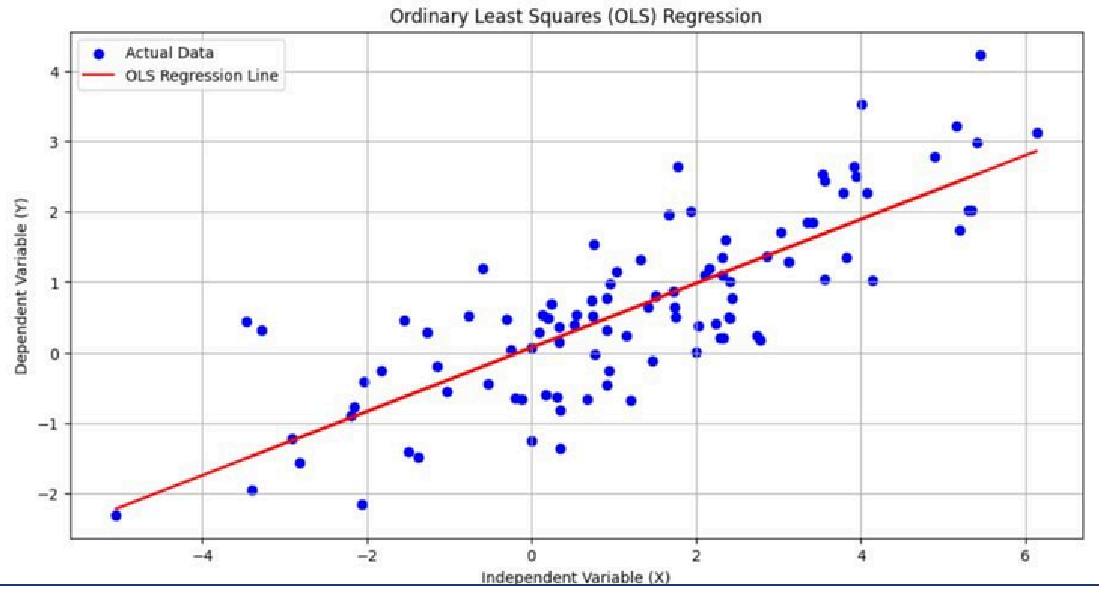
- Linear relationship between variables
- Zero-mean errors
- No perfect multicollinearity
- Homoscedastic, uncorrelated errors
- Normally distributed errors

### INTERPRETING RESULTS

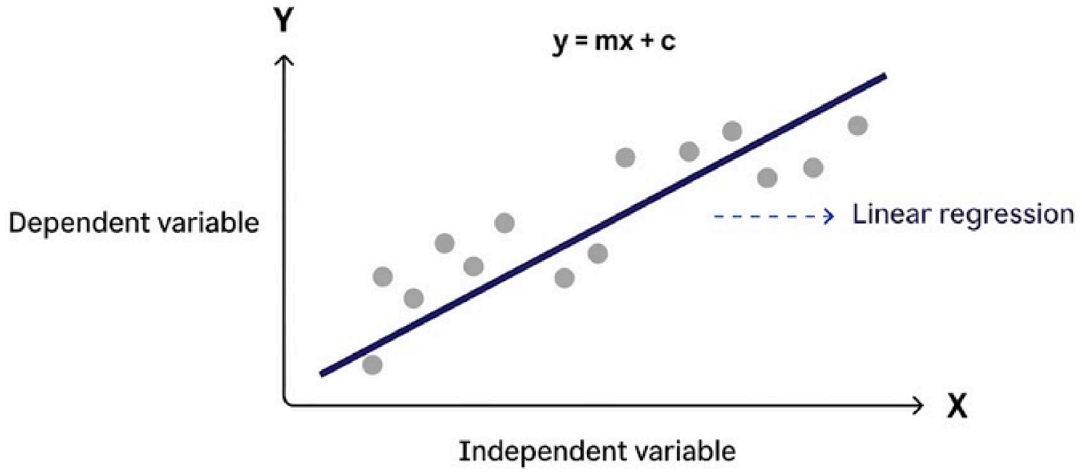
- Coefficients measure each predictor's effect on the outcome
- A one-unit increase in X changes Y by β (ceteris paribus)
- Useful for prediction and interpretation

## WHAT'S OLS

Ordinary Least Squares (OLS) estimates the parameters of linear regression by finding the best-fitting line, helping model and interpret relationships between variables.




## LINEAR MODEL




A linear regression model uses a straight line to describe the relationship between Y and one or more X variables. In OLS, the fitted line represents the expected value of Y for a given X.

## FITTING OF LINEAR MODEL

### Best Fit Criterion

OLS finds the regression line by minimising the sum of squared residuals (errors) between actual and predicted values.

### Squared Errors

Squaring makes all errors positive and gives greater weight to larger errors, preventing errors from canceling out.

### Basic Properties

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ putting $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X$

Taking the summation on each side,

$\Sigma \hat{Y}_i = n\bar{Y} + \hat{\beta}_1 \Sigma (X_i - \bar{X})$

$\Sigma (X_i - \bar{X}) = 0$

$\bar{\hat{Y}} = \bar{Y}$

### Outcome

OLS produces the best-fitting line and estimates the population parameters.

# CLRM ASSUMPTIONS

## The Classical Linear Regression Model (CLRM)

It is based on a set of assumptions that ensure the Ordinary Least Squares (OLS) estimators are reliable, unbiased, and efficient. These assumptions form the theoretical foundation of regression analysis and enable valid statistical inference.

**Linearity**
The dependent variable must be a linear function of the explanatory variables and an error term.

**Random Sampling**
Observations must be randomly drawn from the population, ensuring the sample is representative.

**Exogeneity**
Explanatory variables must be uncorrelated with the error term, ruling out omitted factors or reverse causality.

**Zero Conditional Mean of Errors**
The expected value of the error term given any value of X must equal zero: $E(u/X) = 0$

**No Perfect Multicollinearity**
No explanatory variable should be a perfect linear combination of another, as this makes isolating individual effects impossible.

**Homoscedasticity**
Error variance must remain constant across all observations: $Var(u/X) = \sigma^2$

**No Autocorrelation**
Error terms across observations must be uncorrelated: $Cov(u_i,u_j) = 0$ for $i \neq j$

**Correct Model Specification**
The model must include all relevant variables, exclude irrelevant ones, and use the correct functional form.

**Normality of Errors**
The error term should be normally distributed: $u \sim N(0,\sigma^2)$, supporting hypothesis testing and confidence intervals.

# RANDOM SAMPLING AND VIOLATIONS

## RANDOM SAMPLING

### What It Means

Observations must be drawn independently from the target population through a probability-based process, where every unit has a known chance of selection and no unit's inclusion affects another's.

### Enables Valid Inference

Each observation pair ($X_i$, $Y_i$) is independent of others, meaning data points are independently and identically distributed (i.i.d.) draws from the same underlying population.

### Ensures Unbiasedness

If the sample systematically differs from the population, estimated coefficients will reflect the sample's quirks rather than the true population relationship.

### Formal Implication

Standard errors, t-statistics, and confidence intervals all rely on observation independence. Non-random sampling distorts these quantities and invalidates hypothesis tests.

Within the Classical Linear Regression Model (CLRM) framework, random sampling is a key assumption that ensures OLS estimators possess desirable statistical properties.It helps ensure that:

### Role in OLS (Ordinary Least Squares)

**Ensures $u_i$ is not systematically related across observations.**

**Allows the Law of Large Numbers and Central Limit Theorem to apply, enabling reliable inference as sample size grows.**

**Guarantees estimates converge to true population values as sample size increases.**

## COMMON VIOLATIONS AND CONSEQUENCES

### Selection bias

Occurs when inclusion in the sample relates to the outcome or predictor, biasing results.

1 → **Biased, inconsistent coefficients**

### Survivorship bias

Only "survivors" of a process are observed, skewing conclusions toward successful outcomes.

2 → **Understated standard errors**

### Cluster dependence

Observations share unobserved characteristics, making standard errors too small and inflating significance.

3 → **Skewed coefficient estimates**

### Convenience Sampling

Using the easiest available data rarely produces a representative sample, yielding unreliable estimates.

4 → **Limited external validity**

# GOODNESS OF FIT

## R SQUARE

$R^2$ measures how much of the total variance in the dependent variable is explained by the model, on a scale of 0 to 1.

$$R^2 = 1 - (SSres / SStot)$$

SSres= *sum of squared residuals*

Stot = total sum of squares

Adjusted $R^2$ measures how well the model explains variation in the dependent variable while penalizing unnecessary predictors.

$$R^2adj = 1 - (1-R^2) \times (n-1)/(n-k-1)$$

## DIFFERENCES

| R - Squared | Adjusted R-Squared |
|---|---|
| Always increases or stays the same | Decreases if the new predictor adds no value |
| Creates a misleading impression of model effectiveness | More reliable indicator of true model quality |
| Models with the same number of predictors | Models with different numbers of predictors |

## GOODNESS OF FIT

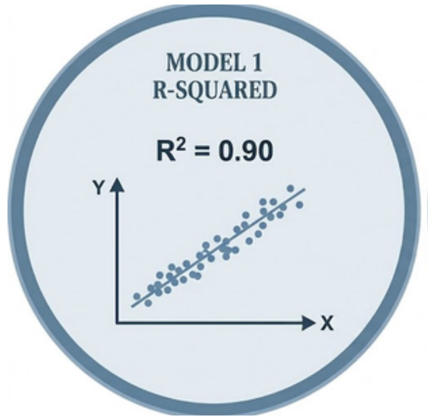
**High goodness of fit**

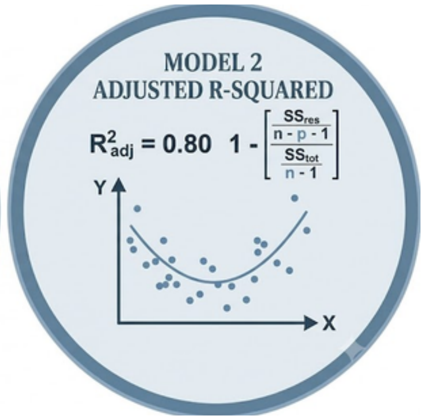
Observed values are close to expected values, so the model explains the current data well. However, it may not predict new data accurately if it is overfitting.

**Low goodness of fit**

The model explains the data less well, but this does not automatically mean its conclusions are misleading. Its usefulness depends on the context and purpose.

## INTERPRETATION




Model 1 has a higher $R^2$, suggesting it explains more of the overall variance in the data.

Model 2 has a higher adjusted $R^2$, suggesting a better balance between fit and model complexity.

Model 2 would be preferable for comparison purposes, though overfitting and model reliability depend on additional diagnostics beyond adjusted $R^2$ alone.

## LIMITATIONS

### R SQUARED

- Including irrelevant predictors can inflate $R^2$.
- A high $R^2$ may falsely indicate strong model performance.
- Outliers, multicollinearity, and non-linearity can distort $R^2$.

### ADJUSTED R SQUARED

- Adjusted $R^2$ does not detect multicollinearity; VIF does.
- It can be distorted by outliers in the data.
- A high Adjusted $R^2$ does not guarantee a good model.

## ANOVA TABLE

The ANOVA table tests the overall statistical significance of the regression model by comparing explained and unexplained variation.

| Source | df | Sum of Sq. | Mean Sq. | F - Stat | P - value |
|---|---|---|---|---|---|
| Between | $k-1$ | $SSB$ | $s_b^2 = \frac{SSB}{df_b}$ | $f = \frac{s_b^2}{s_w^2}$ | $P(F_{k-1,n-k} \geq f)$ |
| Within | $n-k$ | $SSW$ | $s_w^2 = \frac{SSW}{df_w}$ | | |
| Total | $n-1$ | $SST$ | | | |

High F-stat + low p-value = model is statistically significant.

# STATISTICAL INFERENCE

## STATISTICAL SIGNIFICANCE

Statistical significance indicates whether an observed relationship is likely real rather than due to random chance.

## KEY MEASURES

### p value

**p-value is the probability of obtaining a result as extreme as the observed one.**

**DECISION RULE:**

$p < \alpha \rightarrow$ Reject $H_0$

$p > \alpha \rightarrow$ Fail to Reject $H_0$

- Purpose: Determines whether a relationship is likely real or due to chance.
- Interpretation: Smaller p-values provide stronger evidence against $H_0$.
- Rule: If $p < 0.05$, reject $H_0$; the relationship is statistically significant.

### t-statistic

**Measures how far an estimated coefficient is from its hypothesized value.**

**FORMULA:**

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

- $\beta = 0$: Predictor has no effect on the dependent variable.
- Higher |t|-values: Stronger evidence against $H_0$ ($\beta = 0$).
- Measures coefficient precision by accounting for its variability.
- Used when population variance is unknown.

## MODEL SETUP

**OBJECTIVE:** Examine the impact of study hours on students' academic performance.

**REGRESSION MODEL:**

$$Score = \beta_0 + \beta_1(Study\ Hours) + u$$

**NULL HYPOTHESIS:** $\beta_1 = 0$ (Study hours have no effect on scores)

**ALTERNATE HYPOTHESIS:** $\beta_1 \neq 0$ (Study hours affect scores)

**REGRESSION OUTPUT**

| Variable | Coefficient | p-value |
|---|---|---|
| Study Hours | 2.5 | 0.02 |

**INTERPRETATION:**

- Coefficient (2.5): Each additional study hour increases marks by 2.5.
- p-value (0.02): Only a 2% chance of observing this result if study hours had no effect.
- Since $0.02 < 0.05$, reject $H_0$.
- Conclusion: Study hours have a statistically significant positive effect on marks.

## IMPORTANCE OF STATISTICAL SIGNIFICANCE IN ECONOMETRICS

**Reliability Check**: Helps determine whether estimated relationships are reliable and consistently supported by the data.

**Theory Testing**: Enables economists to validate theoretical predictions using real-world observations and empirical evidence.

**Policy Support**: Provides a strong statistical basis for policy formulation, business strategies, economic recommendations.

**Reduces Random Error**: Ensures that regression results reflect genuine relationships rather than sampling fluctuations.

# CORE ECONOMETRIC FRAMEWORK

# COEFFICIENTS AND INTERPRETATION IN REGRESSION OUTPUT

## COEFFICIENTS

The value of the coefficient shows the extent to which the average value of the dependent variable changes when there is a one-unit change in the value of the independent variable while keeping the rest of the variables constant. If the coefficient of X1 is 5, it indicates that whenever X1 rises by one unit, Y increases by 5 units.

**Unstandardize Coefficients ($B$)**

Expressed in the original units of the data.

**Standardized Coefficients ($\beta$)**

Expressed in standard deviations, enabling comparison across predictors with different scales.

## Interpreting Signs and Magnitudes in Regression Output

When interpreting regression output, the coefficients (β) are the primary focus. Each coefficient should be evaluated based on its sign and magnitude.

- Positive (+): X and Y move in the same direction—as X increases, Y increases.
- Negative (−): X and Y move in opposite directions—as X increases, Y decreases.

Coefficient signs should be checked first against economic theory. An unexpected sign may indicate issues such as multicollinearity, omitted variable bias, or model misspecification.

**The Magnitude**

The coefficient magnitude shows how much Y changes for a one-unit increase in X, holding other variables constant. Its interpretation depends on the units of both variables.

## PRACTICAL CHECKS WHEN INTERPRETING OUTPUT

**Statistical vs. Practical Significance:**
A significant coefficient may still have little practical impact—consider both p-value and magnitude.

**Standard Error (SE):**
Measures the uncertainty of a coefficient estimate.

**Confidence Interval:**
Shows the plausible range of the true coefficient; narrower intervals indicate greater precision.

**Interaction Terms:**
Coefficients must be interpreted jointly, not in isolation.

**Log Variables:**
Coefficients represent percentage changes rather than unit changes.

**t-statistic**:
Larger coefficients or smaller SEs produce larger t-values and lower p-values. t=Coefficient/Standard Error (SE)

**Model Statistics:**
$R^2$ = 0.72: Explains 72% of the variation in the dependent variable. Adjusted $R^2$ = 0.70: Adjusts for the number of predictors; useful for model comparison. F-statistic ($p < 0.001$): Indicates the overall model is statistically significant.

# MULTICOLLINEARITY

## WHAT IS MULTICOLLINEARITY?

Multicollinearity occurs when two or more independent variables are highly correlated, making their individual effects difficult to estimate accurately.

## HOW MULTICOLLINEARITY ARISES?

### INSUFFICIENT SAMPLE SIZE

A small sample size can exaggerate correlations between variables, making multicollinearity appear more severe.

### NATURALLY RELATED VARIABLES

Naturally related variables (e.g., income and wealth) often move together, making their individual effects difficult to estimate.

### SMALL VARIABLE RANGES

A limited sample can make variables appear more strongly related than they actually are.

### POORLY DEFINED MODELS

Using multiple representations of the same variable can increase correlation among predictors, leading to multicollinearity.

### EXCESS VARIABLES WITH TOO LITTLE DATA

Too many variables relative to observations make individual effects difficult to estimate.

## DETECTION AND INTERPRETATION

| | |
|---|---|
| Correlation Matrix | Examines correlations between independent variables. If r is between −0.8 and +0.8, multicollinearity is generally not a concern. |
| Tolerance Analysis | Tolerance represents the proportion of a predictor's variance that is not explained by other predictors. Higher values are generally desirable. A tolerance value above 0.2 is considered acceptable. A value below 0.1 suggests severe multicollinearity. |
| Variance Inflation Factor | VIF measures how much the variance of a regression coefficient is inflated due to correlations. A VIF of 1 indicates no multicollinearity; 1–5 suggests moderate multicollinearity; above 5 indicates high multicollinearity. |
| Eigenvalue Analysis | This method evaluates multicollinearity using eigenvalues of the predictor matrix. A condition index below 10 suggests no multicollinearity; 10–30 indicates moderate multicollinearity, and above 30 reflects severe multicollinearity that may distort coefficient estimates and statistical inference. |
| Condition Index | The Condition Index measures the severity of multicollinearity based on the eigenvalues of the predictor matrix. A CI below 10 indicates little multicollinearity; 10–30 suggests moderate multicollinearity; above 30 indicates severe multicollinearity |

## REMEDIES

### INCREASE SAMPLE SIZE

With an increase in the number of samples, there will be an increase in variance in the variables, thus making it easy to differentiate between the effects of various predictors.

### REMOVING REDUNDANT PREDICTORS

When two or more predictor variables are highly correlated and convey similar information, one of them should be removed to reduce multicollinearity and improve the reliability of the regression model.

### COMBINING VARIABLES

In case two or more predictors are highly correlated with each other, they should be combined together into one predictor. It can be done through Principal Component Analysis (PCA).

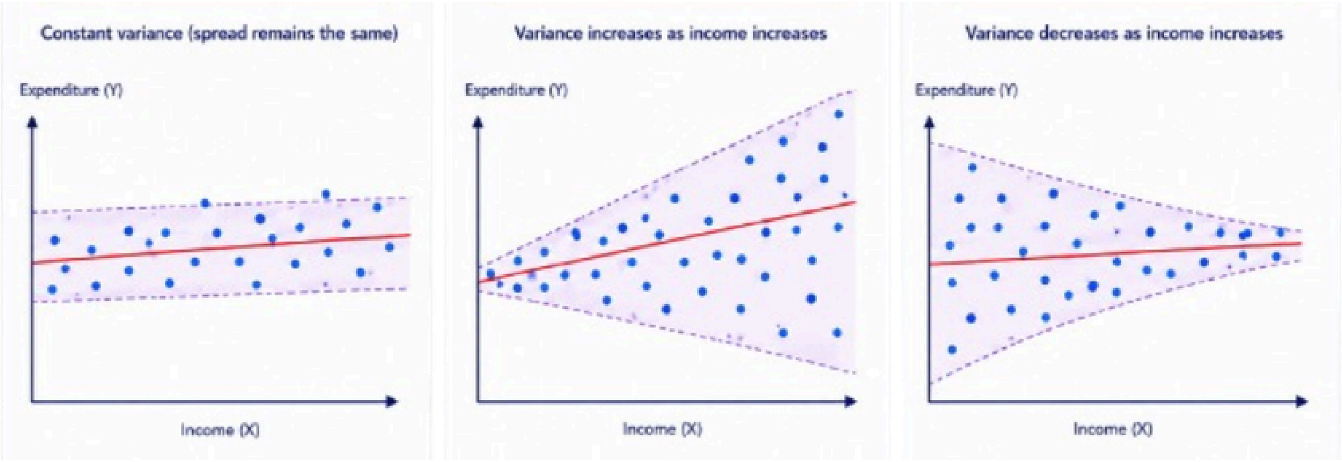
# HETEROSCEDASTICITY

## WHAT IS HETEROSCEDASTICITY?

Heteroscedasticity states that the variance of the disturbance term, conditional on the explanatory variables, is not the same for all combinations of outcomes of the explanatory variables. This means that the variance is not constant across observations.

### EXAMPLE

If an increase in income also increases the expenditure, and the individual expenditure values stay far away from the average expenditure (i.e., the spread/variability is not constant) at every income level, then we say there is heteroscedasticity.




## REMEDIAL MEASURES

1. **$\sigma^2$ IS KNOWN:** We transform the regression function in a way that the transformed function's error term becomes homoscedastic. If we apply the OLS in the transformed model, it will contain no heteroscedasticity.

2. **$\sigma^2$ IS UNKNOWN:** If  2 not known a priori, then heteroscedasticity is corrected by hypothesizing a relationship between the error variance and one of the explanatory variables. (Feasible Generalized Least Square Method ) There can be several versions of the hypothesized relationship.

## CONSEQUENCES

- OLS estimators are still linear and unbiased; however, they no longer have minimum variance. This means the OLS estimators are inefficient i.e. they are no longer BLUE, now they are LUE.
- Due to heteroscedasticity, the usual formula to estimate the variances of the OLS estimators becomes biased.
- The usual confidence interval and hypothesis tests based on t and F distribution are unreliable because the standard errors and variances become large.

## DETECTION

- **Graphical Method**: If there is an observed pattern between Xi & e2i or Yi & e2i. The pattern can be U-Shaped, Inverted U-Shaped or Linear Funnel.
- **Park Test:** A regression-based test which checks for heteroscedasticity by evaluating whether there is a correlation between log ( e2i) and log ( Xi )
- **Glejser Test:** A test for heteroscedasticity that involves regressing | ei | from the original regression on the explanatory variable.
- **White's General Test for Heteroscedasticity**: A general and widely-used test that detects heteroscedasticity using an auxiliary regression equation:  ei2  =A1 +A2 X2i+A3 X3i+A4X2i2 +A5X23i+A6X2iX3i+Vi
- **Goldfeld Quandt Test**: A test for heteroscedasticity that divides the dataset into two groups (e.g., high and low values) and compares the error variances between them using separate regressions.

# CHOOSING A FUNCTIONAL FORM

## PRACTICAL CONSIDERATIONS

**LINEAR REGRESSION**

- **OLS Model:** Estimates linear relationships using

$$\boxed{Y = \beta_0 + \beta_1 X + u;}$$

- **Log-Linear Models:** Use logarithms to capture percentage effects, where beta represents elasticity; (1% increase in X → β% change in Y).

**BINARY RESPONSE REGRESSION MODEL**

PRACTICAL CONSIDERATIONS

- Estimates the probability of an outcome occurring using nonlinear functions:
- Coefficients are interpreted through marginal effects, shows how a one-unit change in X affects the probability of event. $\boxed{P(Y = 1|X)}$
- Used when the dependent variable is binary (e.g., Export = 1, Not Export = 0), making OLS inappropriate.

**PANEL DATA MODEL**

- Analyze data across entities and time periods: where i = entity and t = time.

$$\boxed{Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}}$$

- Fixed effects models control for unobserved time-invariant factors reducing omitted variable bias.
- Useful for studying effect of changes within the same firm, country, or individual on outcomes over time.

**INSTRUMENTAL VARIABLES**

- A valid instrument Z must satisfy **Relevance** $(Cov(Z, X) \neq 0)$ and **Exogeneity** $(Cov(Z, u) = 0)$.
- Used when OLS is biased due to endogeneity caused by simultaneity, omitted variables, or measurement error.
- Two-Stage Least Squares (2SLS) uses the instrument to isolate exogenous variation and estimate a causal effect rather than a simple correlation.

# CHOOSING A FUNCTIONAL FORM

## ECONOMIC THEORY

| | |
|---|---|
| **IDENTIFYING VARIABLES** | Economic theory guides model selection by identifying relevant variables, expected signs, and causal relationships between variables. |
| **REFLECTS BEHAVIOR** | A well-specified econometric model should reflect the underlying economic behavior while remaining consistent with the available data. |
| **AVOID MISSPECIFICATION** | Theory helps avoid model misspecification and improves the interpretation of estimated coefficients. |

## GRAPHICAL INTUITION

- Graphical analysis provides an initial understanding of relationships through scatter plots, trend lines, and data visualization.
- It helps identify trends, outliers, nonlinear patterns, potential data issues before formal estimation.
- Visual inspection assists researchers in selecting an economically and statistically appropriate model specification.

## ADJUSTED $R^2$ AND MODEL FIT

- Adjusted $R^2$ measures the proportion of variation in the dependent variable explained by the model while accounting for the number of predictors.
- Unlike $R^2$, Adjusted $R^2$ penalises unnecessary variables, making it a more reliable measure of model fit.
- A higher Adjusted $R^2$ indicates better explanatory power, but model selection should also be guided by economic theory and statistical significance.

# ANOVA VS ANCOVA

| BASIS | ANOVA | ANCOVA |
|---|---|---|
| **Definition** | Compares group means to test for significant differences. | Compares group means while controlling for covariates. |
| **Purpose** | Its purpose is to determine whether groups differ significantly from one another. | Uses independent and dependent variables, with covariates to adjust group comparisons through regression. |
| **Variables Involved** | Uses one dependent variable and one or more categorical independent variables to compare group means. | Uses independent, dependent, and covariate variables to adjust group comparisons. |
| **Type of independent variable** | Independent variables are categorical and define the groups. | Independent variables are categorical, but *ANCOVA* also includes continuous covariates. |
| **Control of external factors** | Compares group means without controlling external factors. | Controls external continuous variables affecting the dependent variable. |
| **Error variance** | Divides total variation into between-group and within-group (error) variation. | Reduces error variance by adjusting for covariates; precision improves if they explain outcome variation. |
| **Statistical technique used** | ANOVA uses the F-test to test significance. | Combines ANOVA and regression techniques. |
| **Accuracy of results** | Results are reliable when assumptions are properly satisfied. | Results are more precise because covariates are controlled for. |
| **Assumptions** | ANOVA assumes normality, independence of observations, and equal variances. | Assumes ANOVA assumptions, a linear covariate–outcome relationship, and equal regression slopes. |
| **Example** | Comparing the average marks of students taught by different teaching methods. | Compares sales across strategies while controlling for advertising expenditure (covariate). |

# FUNCTIONAL FORMS I



## LINEAR - LINEAR MODEL

A linear-linear model uses both the dependent variable (Y) and independent variable (X) in their original form. It assumes a straight-line relationship between the variables and estimates the best-fit line to predict Y from X.

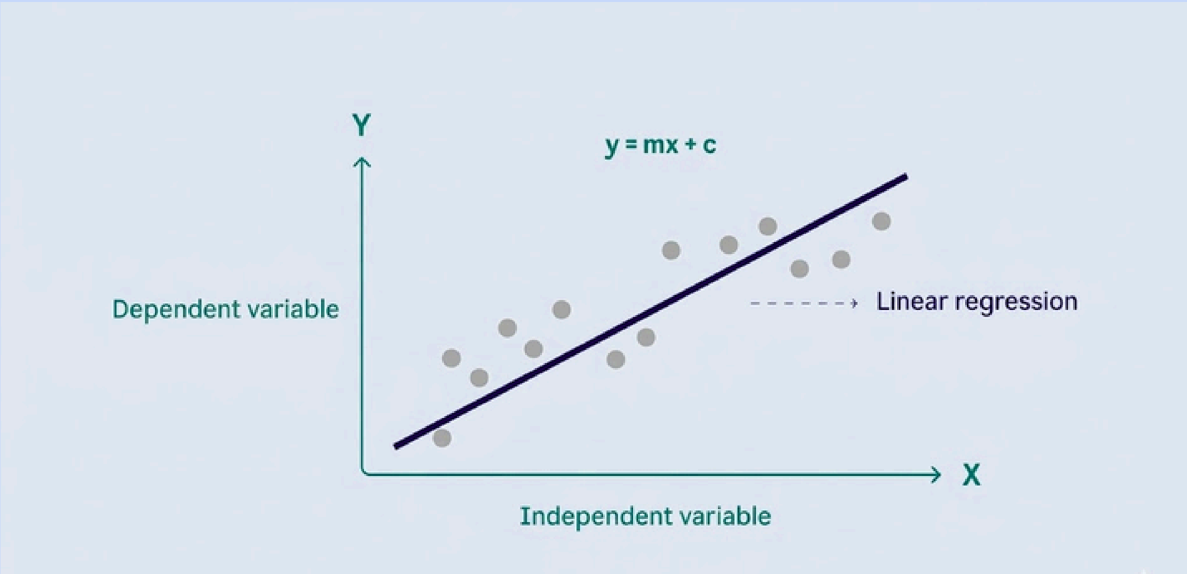
### LINEAR REGRESSION EQUATION

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \epsilon$$

### UNDERSTANDING BASIC TERMS

*β1,β2,...,βn* -COEFFICIENTS OF INDEPENDENT VARIABLE

*Y* is the dependent variable




*β0* is the intercept term

*ϵ* IS THE ERROR TERM

## LOG - LOG MODEL

A log-log regression applies natural logarithms to both variables. Its coefficient represents elasticity: a 1% change in the predictor is associated with a β% change in the outcome

### Y = B0 + B1X INTO THE LOG–LOG FORM

$$\ln(Y) = \beta 0 + \beta 1 \ln(X) \quad (X, Y > 0)$$

### DEMAND CURVE EXAMPLE

In the demand function Q=αP^β ,Q is quantity demanded, P is price, α shifts the demand curve, and β measures price responsiveness.

When β<0 the demand curve slopes downward: higher prices reduce demand, while lower prices increase demand.

As a power function, it has constant elasticity: a given % change in price causes the same % change in quantity demanded, equal to β.

If β=−1, demand is unitary elastic, a 1% rise in price leads to a 1% fall in quantity demanded, and elasticity is −1 .

*Taking logarithms converts the function into a linear form that makes it easier to estimate and analyze.*

### LOG-LOG TRANSFORMATION

1. $Yi = \alpha Xi\beta$
2. $\ln(Yi) = \ln(\alpha) + \beta \ln(Xi)$
3. $\ln(Yi) = \beta 0 + \beta 1 \ln(Xi)$

## ANALYSIS OF COEFFICIENT

In a log-log model, the coefficients represent elasticities. Each coefficient shows the percentage change in the dependent variable (Y) resulting from a 1% change in the independent variable (X).

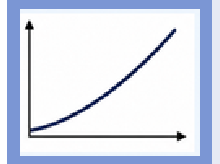
### DERIVING THE ELASTICITY

Differentiating a log–log model converts absolute values into relative changes, allowing elasticity to be measured directly by the regression coefficient:

$$\Delta Y/Y = B1 \cdot (\Delta X/X)$$

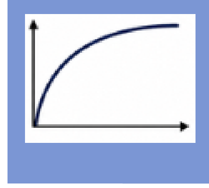
*Right side = %ΔX, left side = %ΔY → β1 measures the elasticity.*

### THREE OUTCOMES FOR β1



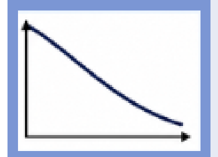
**β1 > 1**

Constantly magnified - Y always responds by more than 1% for every 1% rise in X.



**0 < β1 < 1**

Positive but dampened - Y always responds by less than 1% for every 1% rise in X.



**β1 < 0**

Negative - Y moves opposite to X for every 1% rise in the independent variable.

# FUNCTIONAL FORMS II

## LOG- LINEAR MODEL

A log-linear model is a regression specification in which the dependent variable is expressed in logarithmic form while the independent variable(s) remain in their original scale

### LINEAR REGRESSION EQUATION

$$LogY_t = \beta_1 + \beta_2 t + \mu_t$$

### DERIVATION:

$\beta_2$ *is the instantaneous growth rate*

$Y_0$ ———————————— $Y_t = Y_0(1+r)$

$$Y_t = Y_0(1+r)^t$$
$$LogY_t = LogY_0 + tLog(1+r)$$
$$\text{Let } LogY_0 = \beta_1$$
$$Log(1+r) = \beta_2 \qquad [r = e^{\beta_2} - 1]$$
$$LogY_t = \beta_1 + \beta_2 t$$

***R*** *is the compound growth rate*

***t*** *is the time period*

## LINEAR - LOG MODEL

A linear-log model is a regression model in which the dependent variable remains in its original form while the independent variable is expressed in logarithmic form

### LINEAR REGRESSION EQUATION

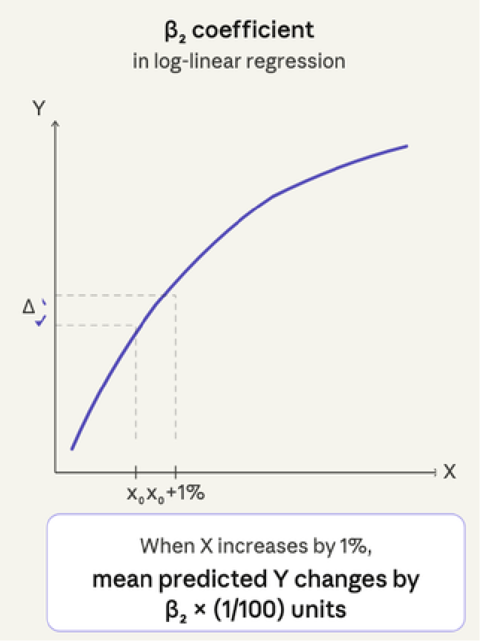
$$Y_i = \beta_1 + \beta_2 LogX_i + \mu_i$$

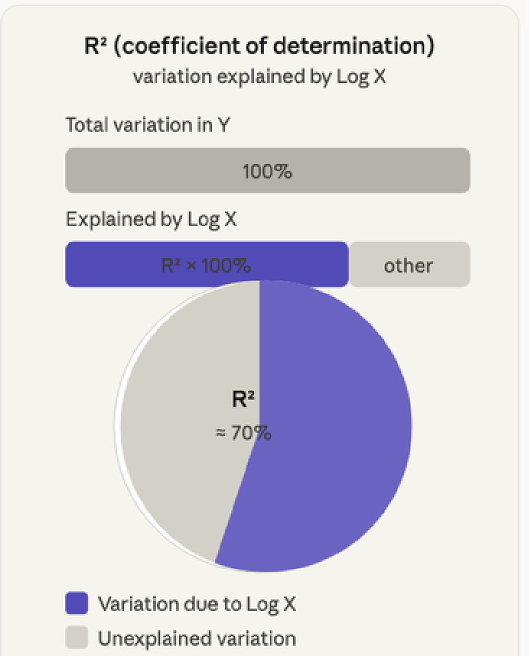
$$\text{Consider, } \widehat{Y_i} = \beta_1 + \beta_2 LogX_i + \mu_i$$

$$\frac{dY}{dLogX} = \beta_2$$

$$\Rightarrow \beta_2 = \frac{Absolute\ Change\ in\ Y}{Relative\ Change\ in\ X}$$

### INTERPRETATIONS




## POLYNOMIAL REGRESSION MODEL

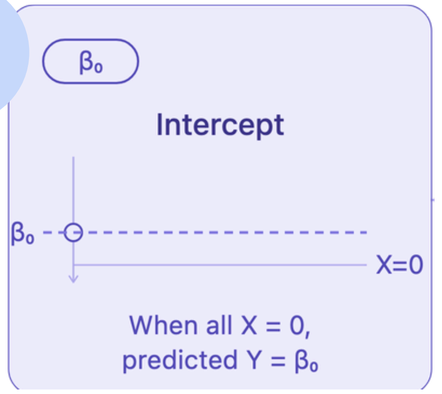
This model shows the relationship between Yi , the dependent variable and Xi , the explanatory variable, as an nth degree polynomial.

### DERIVING THE FORMULA

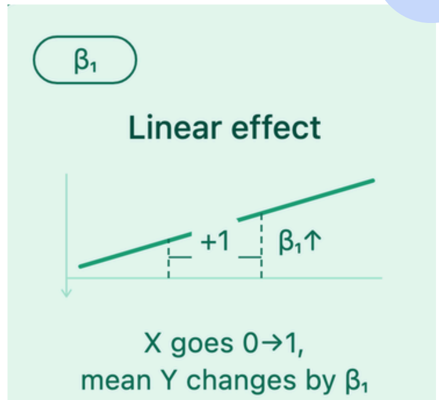
$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + \ldots\ldots + \beta_n X_i^{n-1} + \mu_i$$

### INTERPRETATIONS

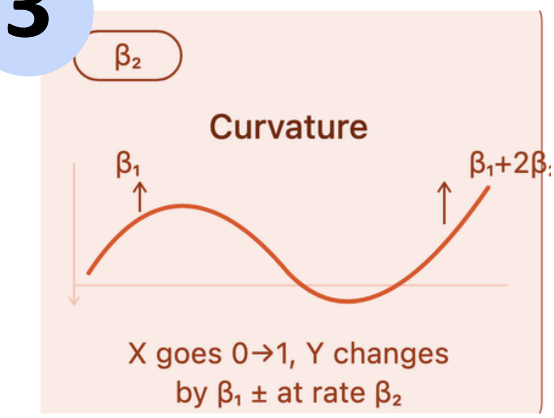
1




2




3



# REGRESSION IN PRACTICE

# HOW DO ECONOMISTS USE REGRESSION?

**Economists use regression analysis to study relationships. Regressions quantify the relationship between one variable and others thought to explain it, and identify its strength. In economics, correlations are common, but identifying causality is rarely easy. The aims include forecasting, explanation, causal inference, theory testing, and policy evaluation.**

## STEPS OF ANALYSIS

### 1 STATEMENT OF THEORY AND ASSUMPTIONS :

State of economic theory and assumptions, helps in specifying null and alternative hypotheses if theory is known a priori.

### 2 SPECIFYING THE MATHEMATICAL EQUATION

Specify the mathematical equation; depending on needs and variables, an appropriate model (simple or multiple linear regression) is developed, translating theory into an econometric model.

### 3 COLLECTION OF DATA:

Collect data to obtain parameter values, using time-series, cross-sectional, or pooled data depending on the research question and model.

### 4 ESTIMATION OF PARAMETERS:

Estimate population parameters indirectly through sample data.

### 5 HYPOTHESIS TESTING:

Conduct hypothesis testing using sample estimates to make inferences about population parameters.

### 6 PREDICTION/ FORECASTING OR POLICY EVALUATION:

Predict the dependent variable using sample estimates and explanatory variables. This is sometimes done for policy evaluation and theory testing, not merely forecasting.

# DUMMY VARIABLES AND THE DUMMY VARIABLE TRAP

GAEE Global Association of Economics Education INDIA

## PURPOSE OF DUMMY VARIABLE

Dummy variables (coded 0 or 1) represent categorical data in regression. A category with k levels needs k-1 dummies, with the remaining category serving as the reference.

For example, marital status (married, single, divorced) uses two dummies; divorced becomes the reference when both equal 0.

Examples of Categorical Variables:
- Eye color (blue, green, brown)
- Gender (male, female)
- Marital status (married, single, divorced)

Each dummy is compared against the reference group. A positive coefficient means higher income than the reference; negative means lower.

## THE REGRESSION EQUATION

$INCOME = b_0 + b_1X_1 + b_2X_2$

Where, $b_0$, $b_1$, and $b_2$ are regression coefficients.

### DUMMY VARIABLE DEFINITIONS

$X_1 = 1$, if Republican; $X_1 = 0$, otherwise.

$X_2 = 1$, if Democrat; $X_2 = 0$, otherwise.

### EXAMPLES & DUMMY VARIABLES

$X_1$ & $X_2$ are dummy variables representing political affiliation categories.

Reference group: Independent

### INTERPRETATION OF BINARY VARIABLES

Statistical significance confirms whether the gap is real.

Independents serve as the reference group.

### THE DUMMY VARIABLE TRAP

In a model with Republican, Democrat, and Independent voters, Independents serve as the reference group. Including all k dummies for k categories causes perfect multicollinearity, known as the dummy variable trap. Always use k-1 dummies.

# SPECIFICATION BIAS AND MISSPECIFICATION OF THE MODEL

## INCORRECT FUNCTIONAL FORM:

A functional form defines how the dependent variable responds to independent variables. An incorrect form can produce biased estimates and misleading conclusions.

## EXCLUSION OF RELEVANT VARIABLES:

Occurs when an important independent variable is omitted from the regression model.

For Example,

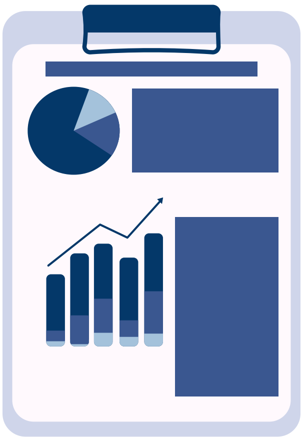
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ → $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, where $X_2$ is omitted.

True Regression Model → Estimated Model

## BIAS IN COEFFICIENT ESTIMATES

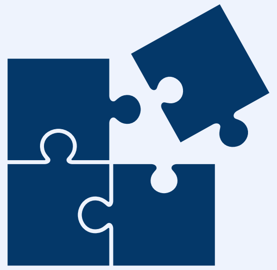
- Produces biased coefficient estimates
- Leads to incorrect conclusions about variable effects
- Can create false statistical significance
- Bias depends on correlations with omitted variables

## MODEL MISSPECIFICATION



- Omitting relevant variables causes model misspecification.
- The model fails to capture the true relationship.
- Reduces prediction accuracy and model fit.
- Can lead to flawed business or policy decisions.

## INCLUDING IRRELEVANT VARIABLES:



- Including irrelevant variables that do not meaningfully explain changes in the dependent variable leads to model misspecification

**FOR EXAMPLE :**

| True Regression Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. | → | Estimated Model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, where $X_2$ is omitted. |
|---|---|---|

# SPECIFICATION BIAS AND MISSPECIFICATION OF THE MODEL

## INCLUDING IRRELEVANT VARIABLES:

- Reduces degrees of freedom, making inference less reliable.
- $R^2$ mechanically increases with each added variable regardless of relevance, making it a misleading fit measure.
- Does not bias OLS estimates but reduces their efficiency, shown as larger standard errors, making it harder to detect significant effects.
- If irrelevant variables are uncorrelated with others, efficiency drops further.
- Adjusted residual sum of squares still provides an unbiased estimate of unexplained variation.

## SELECTION BIAS AND MODEL MISSPECIFICATION

Selection bias occurs when the sample is unrepresentative, while misspecification occurs when the regression model is incorrectly specified. Both can produce biased and unreliable results.

### DEFINITION

Selection bias occurs when non-random sampling produces an unrepresentative sample, distorting parameter estimates, undermining inference, and leading to flawed policy conclusions. Mathematically, for true parameter θ, bias is:

$$B = E(\hat{\theta}) - \theta$$

### EXAMPLES

- **Self-selection bias:** Including only program completers can overestimate its effect.
- **Sample-selection bias:** Studying only treated patients may underestimate a treatment's true effectiveness.

### SELECTION BIAS MATTERS

- **Validity:** Bias leads to unreliable inference and policy decisions.
- **Comparability:** Uncontrolled bias makes study comparisons misleading.
- **Efficiency:** Bias reduces estimator precision and reliability.

# SPECIFICATION BIAS AND MISSPECIFICATION OF THE MODEL

## ECONOMETRIC STUDIES ARE SUSCEPTIBLE TO SELECTION BIAS FROM VARIOUS AVENUES, INCLUDING:

Non-random Sampling: Systematic deviation from random selection produces an unrepresentative sample.

Attrition: Participants dropping out over time skews results away from the original sample.

Self-Selection: Individuals choosing to participate based on characteristics that also affect the outcome introduce bias.

## MISSPECIFICATION

Misspecification is a structural model error stemming from wrong variables, functional form, or assumed relationships. Unlike estimation error, which shrinks with larger samples, misspecification bias persists regardless of sample size, producing inconsistent estimates that never converge to true values. A misspecified model with a million observations is still misspecified.

A classic example is fitting a linear model to a quadratic relationship. Omitting the quadratic term causes the model to overpredict in some ranges and underpredict in others, leading to incorrect conclusions about the size and direction of effects.

## ESTIMATOR ERROR VS MISSPECIFICATION BIAS

VS

**Random sampling variation that decreases with larger sample sizes.**

**Systematic model error that persists regardless of sample size.**

## THE IMPLICATIONS OF MODEL MISSPECIFICATION ARE SIGNIFICANT AND CAN LEAD TO:

### BIASED ESTIMATES

When a model is misspecified, the estimated coefficients may not reflect the true relationships, leading to incorrect conclusions about the effects of variables.

### INVALID HYPOTHESIS TESTING

Misspecification can affect the reliability of statistical tests, potentially leading to false positives or negatives in hypothesis testing.

# ENDOGENEITY: CAUSES & IMPACT

**DEFINITION :** Endogeneity occurs when an explanatory variable (X) is correlated with error , violating a key OLS assumption and leading to biased , inconsistent estimates.

## WHY DOES IT HAPPEN ?

### 1. OMITTED VARIABLE BIAS

- Important variable omitted from model
- Omitted factor affects both X and Y
- Effect gets absorbed into the error term
- Creates correlation between X and error term

**EXAMPLE :** Ability omitted in education-based studies

### 2. REVERSE CAUSALITY

- X and Y influence each other simultaneously
- Direction of causation becomes unclear
- Error term indirectly effects X
- Leads to correlation between X and Error term

**EXAMPLE :** Unexpected increase in crime leads to more police deployment

### 3. MEASUREMENT ERROR

- X is measured inaccurately
- True variation in X mixed with noise
- Leads to correlation between X and error term
- Estimated coefficient become biased

**EXAMPLE :** Survey reporting errors

## SOLUTIONS

| Include relevant control variable | Instrumental Variable(IV) | Fixed Effects Model | Natural Experiments |
|---|---|---|---|

## CONSEQUENCES

| Biased Coefficient Estimates | Inconsistent Regression Results | Over/Under-estimation of effects | Poor policy and business decision |
|---|---|---|---|

# ENDOGENEITY AND SOURCES

## MEASUREMENT ERROR

Measurement error occurs when observed data deviates from the true underlying construct, creating "errors-in-variables."

Classical measurement error in explanatory variables (regressors) causes endogeneity, leading to biased and inconsistent ordinary least squares (OLS) estimates, specifically toward zero, known as attenuation bias.

The difference between the observed value (X) and the true, unobserved value X*, is usually modeled as:

$$\boxed{X = X^* + \epsilon_x}$$

Strength and direction of the bias are determined by ρXu, the correlation between the error term and the regressor.

## OMITTED VARIABLE BIAS

Omitted variable bias is the bias in the OLS estimator that arises when the regressor, X. is correlated with an omitted variable.

### CONDITIONS FOR OVB TO OCCUR

1. X is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable Y.

### VIOLATION OF OLS ASSUMPTION

Together they result in a violation of the OLS assumption

E(ui /Xi ) = 0

### VIOLATIONS OF OLS ASSUMPTION

Formally, the resulting bias can be expressed as

$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{xu} (\sigma_u / \sigma_x)$

### WHY IS OVB A PROBLEM ?

OVB is a problem that cannot be solved by increasing the number of observations used to estimate β1, as 1  is inconsistent: OVB prevents the estimator from converging in probability to the true parameter value. An omitted variable leads to biased and inconsistent coefficient estimates.

# LIMITATIONS OF REGRESSION

## WHAT REGRESSION CANNOT SOLVE

**Regression shows association, not causation**
A significant regression coefficient indicates that two variables are related, but it does not prove that one variable causes the other.

**Omitted variables can bias results.**
If important explanatory variables are excluded or if the dependent and independent variables influence each other (e.g., income ↔ education).

**Poor data quality & incorrect specification**
Missing data, measurement errors, or using the wrong functional form lead to unreliable estimates regardless of how advanced the statistical model is.

**Statistical significance is not economic significance**
A coefficient may have a very low p-value but represent only a negligible effect in practical or policy terms.

**Regression cannot resolve endogeneity**
When explanatory variables are correlated with the error term due to reverse causality, simultaneity, or omitted factors.

## IMPORTANCE OF JUDGMENT

- **Regression models should be guided by economic theory**
- **Results should be evaluated for economic plausibility**
- **Context matters when interpreting findings**
- **Understand data-generating process improves model quality.**

## CORRELATION VS. STRUCTURAL CAUSALITY

**Correlation means variables move together, whereas causality explains why they move together.**

- Measures statistical association between variables, showing that they move together without explaining the underlying reason.
- Does not prove cause-and-effect, as the relationship may arise from confounding variables, coincidence, or reverse causality.

**Sources of Spurious Correlation**

- Confounding variables may influence both variables, creating a misleading relationship.
- Reverse causality occurs when the outcome actually causes the explanatory variable instead of the reverse.

**Structural Causality Methods**

- Instrumental Variables (IV) address endogeneity by using an external variable related to the treatment but not the outcome directly.
- Difference-in-Differences (DiD) estimates causal impacts by comparing changes between treatment and control groups over time.

**Goal of Structural Causality**

- Structural methods estimate the effect of changing one variable while holding other influences constant.
- Causal inference enables researchers to explain why outcomes occur, rather than simply describing patterns in the data.

## KEY TAKEAWAYS

1. Good econometrics combines statistical methods with economic intuition.
2. Regression outputs should be interpreted beyond p-values and $R^2$.
3. Ultimate goal is to identify the true drivers of economic outcomes.

# CASE STUDIES

# CASE STUDY 1: OKUN'S LAW

**Evaluating pre & post 2008 policy impacts using Okun's law.**

Okun's Law is a macroeconomic rule of thumb that describes the empirical relationship between GDP growth and unemployment.

## DATA AND METHODOLOGY

1. Collected annual time-series dataset for India for the periods 1998 - 2008 and 2008 - 2018.
   **Find the data <u>here</u>**
2. Defined – **Dependent Variable:** Change in Unemployment Rate
   **Explanatory Variable (Baseline Model):** Output Gap
   Note: Output Gap calculated as the difference between Actual GDP and Potential GDP.
3. Autocorrelation Check: Durbin–Watson Test
4. Compared Model 1 (1998–2008) and Model 2 (2008–2018) using:
   - $R^2$
   - Adjusted $R^2$
   - F-statistic
   - p-value
   - Coefficient (β) changes

**Note:** The analysis is based on annual time-series data and estimated using a simple linear regression framework for the two study periods.

## DESCRIPTIVE SUMMARY

### MODEL 1

```
      Year          Actual GDP       Potential GDP       Output Gap          Actual Unemployment
 Min.   :1998   Min.   : 708.3   Min.   : 718.0   Min.   :-0.0420000   Min.   :5.410
 1st Qu.:2000   1st Qu.: 819.8   1st Qu.: 832.5   1st Qu.:-0.0175000   1st Qu.:5.565
 Median :2003   Median : 939.5   Median : 963.0   Median :-0.0060000   Median :5.595
 Mean   :2003   Mean   : 982.4   Mean   : 997.8   Mean   :-0.0002727   Mean   :5.623
 3rd Qu.:2006   3rd Qu.:1138.7   3rd Qu.:1114.0   3rd Qu.: 0.0125000   3rd Qu.:5.738
 Max.   :2008   Max.   :1312.4   Max.   :1295.0   Max.   : 0.0450000   Max.   :5.740

 Change in Unemployment        β
 Min.   :-0.09000         Mode:logical   Min.   :-0.1444
 1st Qu.: 0.06500         NA's:11        1st Qu.: 6.6200
 Median : 0.10000                        Median :15.0000
 Mean   : 0.09455                        Mean   :29.2096
 3rd Qu.: 0.13500                        3rd Qu.:28.1250
 Max.   : 0.24000                        Max.   :140.0000
```

### MODEL 2

```
> summary(data_1)
      YEAR      Unemployment_Rate Actual_Growth_Rate_of_Output Potential_Growth_Rate_Of_Output Change_in_Unemployment_Rate
 Min.   :2008   Min.   :7.609     Min.   :3.090                Min.   :6.500                   Min.   :-0.35266
 1st Qu.:2010   1st Qu.:7.628     1st Qu.:5.925                1st Qu.:6.700                   1st Qu.:-0.26124
 Median :2013   Median :7.631     Median :6.800                Median :6.800                   Median : 0.09176
 Mean   :2013   Mean   :7.637     Mean   :6.678                Mean   :6.982                   Mean   : 0.11076
 3rd Qu.:2016   3rd Qu.:7.649     3rd Qu.:7.930                3rd Qu.:7.200                   3rd Qu.: 0.38630
 Max.   :2018   Max.   :7.678     Max.   :8.500                Max.   :7.800                   Max.   : 0.91270
```

*The selected data provide sufficient variation to evaluate the relationship between economic activity and changes in unemployment across both periods.*

$$\Delta u = \alpha + \beta \cdot \Delta Y + \varepsilon$$

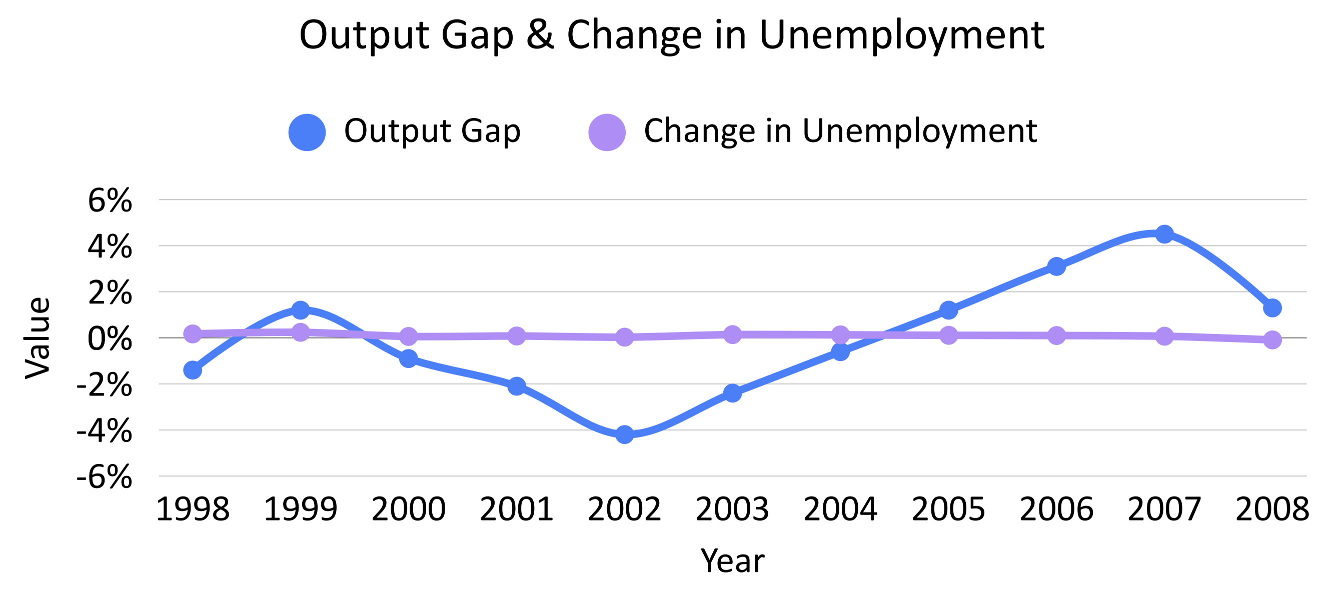
Where: Δu is the change in unemployment, ΔY is output gap, β is the Okun coefficient (typically -0.4 to -0.5 in absolute value)

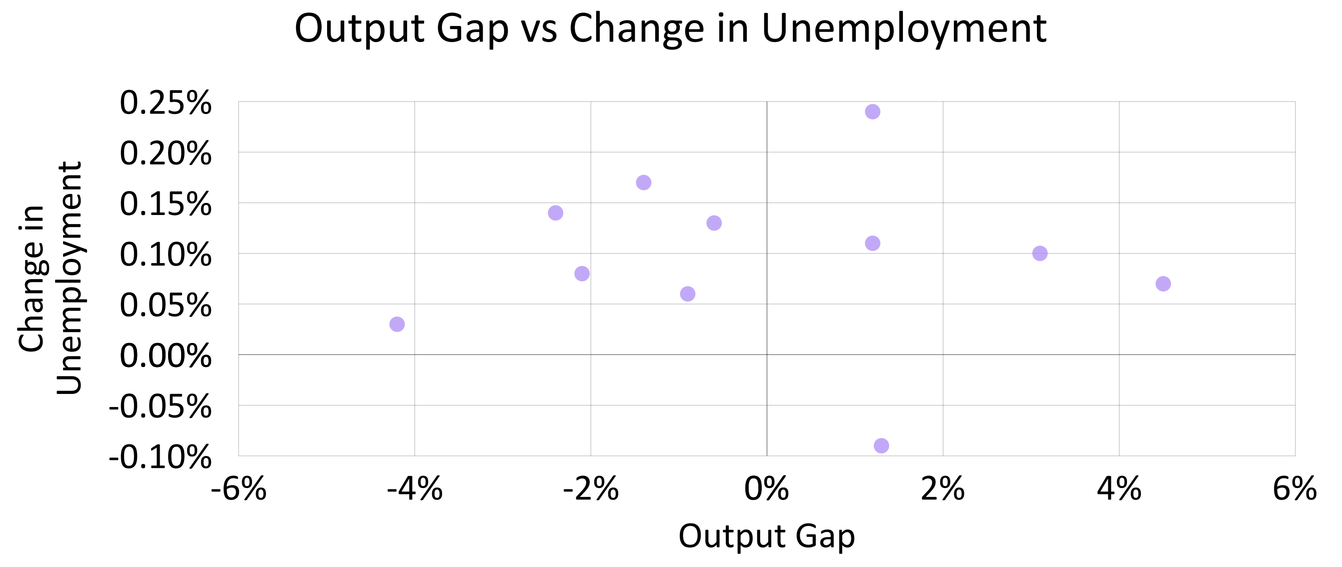
| EQUATION OF MODEL 1: | EQUATION OF MODEL 2: |
|---|---|
| $Y_t = 0.09453 - 0.06676\ X_t$ | $Y_i = 0.06219 - 0.15995\ X_i$ |

# CASE STUDY 1: R WORKFLOW

## PRE 2008 CODES

```
install.packages("readxl"); library(readxl)
data <- read_excel("C:/Users/Kalyani/Downloads/Okuns law case study.xlsx")
model_1 <- lm(`Change in Unemployment` ~ `Output Gap`, data=data)
summary(model_1)
```



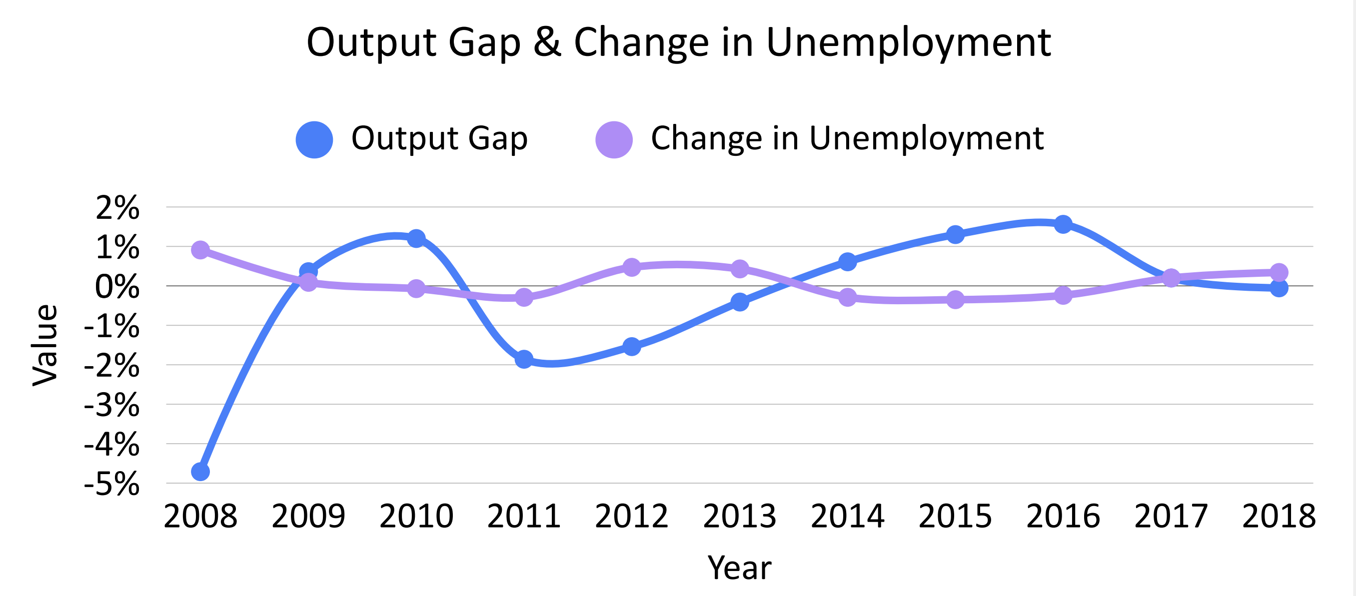

## INTERPRETATION

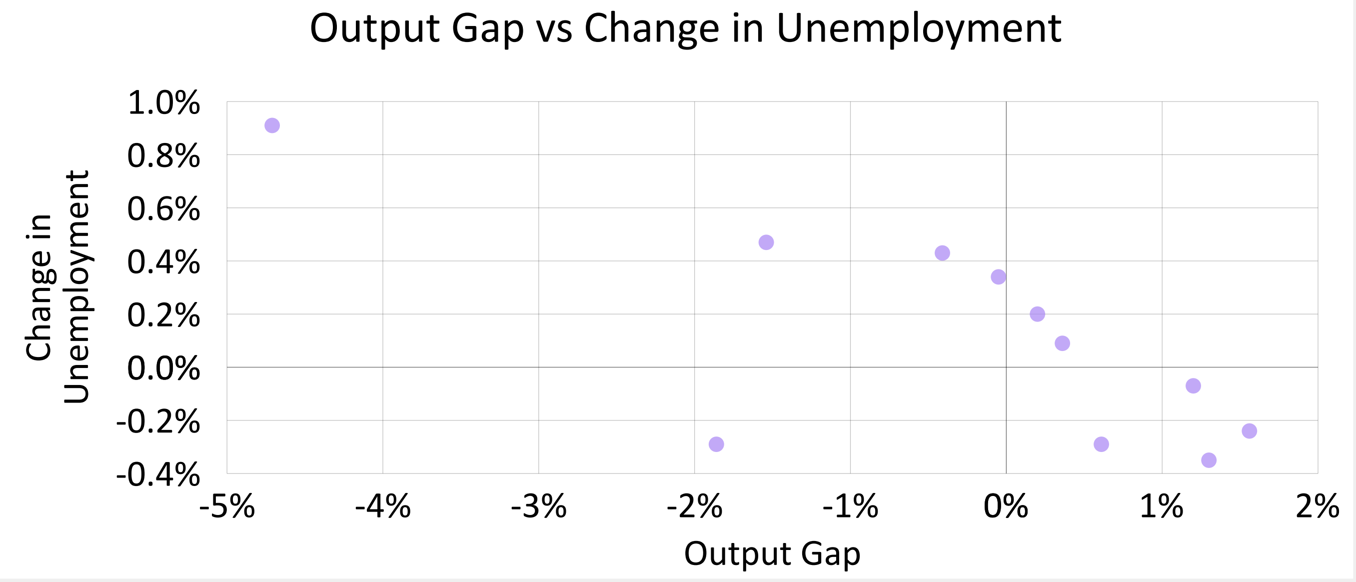
| PARTICULARS | 1998 - 2003 | 2003 - 2008 |
|---|---|---|
| LACK OF JOBS | Rose in 1999, eased in 2000 - 01 with telecom and IT hiring, then climbed in 2002 - 03 as tech slowed. | Improved from 2003 to 2004 (infrastructure + IT BPO hiring), then rose again from 2005 to 2008 despite strong output growth |
| OUTPUT GAP | Peaked at + 0.24 in 1999, declined to + 0.03 by 2002 - 03. | Rose from + 0.03 in 2003 to a peak of + 0.14 in 2004, remained elevated (+ 0.10 – 0.13) through 2007, then turned negative in 2008. |
| OKUN'S FIT | Clear inverse relationship in 2000 - 02, falling output matched rising unemployment. | Strong in 2003 to 2004. Breaks down from 2005 to 2007, the clearest case of jobless growth in the dataset |
| DRIVERS | Positives: telecom liberalization, software exports, sanctions easing; Negatives: dot-com crash, 9/11, Gujarat quake. | Infrastructure spending, IT-BPO outsourcing. |
| TAKEAWAY | IT/telecom growth lifted skilled jobs; low-skilled stagnation highlighted need for vocational training. | Growth quality mattered → support textiles, food, small manufacturing for jobs. |

# CASE STUDY 1: 2008 - 2018

## POST 2008 CODES

```
library(readxl)
data_1<-
read_excel(path="C:/Users/USER/Desktop/Okun's_Law.xlsx")
output_gap<-(data_1$Actual_Growth_Rate_of_Output-
data_1$Potential_Growth_Rate_Of_Output)
Okuns_model<-lm(data_1$Change_in_Unemployment_Rate~output_gap)
summary(Okuns_model)
```




## INTERPRETATION

### 2008 - 10: Crisis then Cushion

- 2008 shock, NREGA cushions recovery
- Output gap -4.71 to +1.2, unemployment eases
- Okun holds in 2008, then breaks

### 2010 - 12: Tightening & Paralysis

- Rate hikes, 2G/Coalgate scandals
- Output gap +1.2 to -1.54, unemployment lags
- Okun breaks in 2011, re-aligns by 2012

### 2012 - 14: Lagged Pain, New Growth

- NREGA fades, Make in India begins
- Output Gap -1.54 to +0.61, unemployment falls
- Okun delayed 2011-12 pain surfaces late

### 2014 - 16: Jobless Growth

- Make in India, Jan Dhan, cheap oil
- Output gap peaks +1.56, unemployment falling
- Weak Okun fit - classic jobless growth

### 2016 - 18: GST & IL&FS shock

- Demonetization, rushed GST hit SMEs
- Output gap to -0.05, unemployment +0.20, +0.34
- Okun fit sharpens as gap nears zero

| PARTICULARS | 2008 - 2013 | 2013 - 2018 |
|---|---|---|
| LACK OF JOBS | Spiked in 2008, eased by 2011, rebounded in 2012-13. | Fell steadily despite strong GDP growth. |
| OUTPUT GAP | Output gap rose to 2010, then fell sharply by 2011. | Output gap rose to 2016, then collapsed by 2018. |
| OKUN'S FIT | Held in 2008, weakened, then realigned by 2012-13. | Jobless growth persisted until Okun's Law reappeared. |
| DRIVERS | Crisis, stimulus, NREGA, then policy shocks slowed growth. | Make in India, Jan Dhan, and low oil supported growth. |
| TAKEAWAY | Relief delayed job losses; slowdown restored the link. | Capital-intensive growth; SME shocks restored the output-employment |

# CASE STUDY 1: AUTOCORRELATION

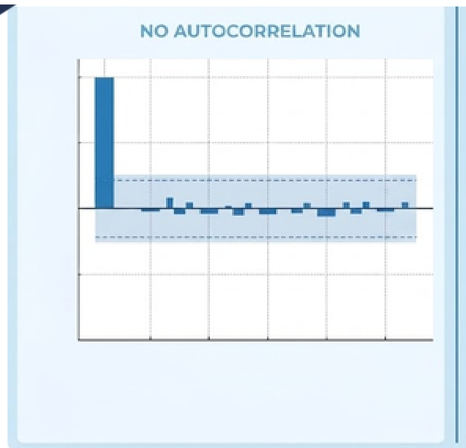
## AUTOCORRELATION FOR 1998-2008

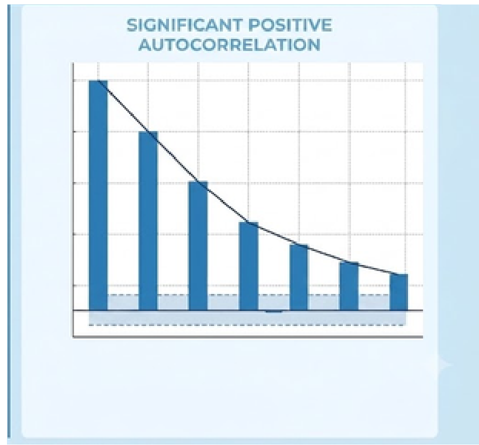
```
        Durbin-Watson test

data:  model_1
DW = 1.1476, p-value = 0.03296
alternative hypothesis: true autocorrelation is greater than 0
```

⇓

Since the Durbin–Watson test is significant & p-value < 0.05, we reject the null hypothesis and conclude that the model's residuals exhibit significant autocorrelation.




## AUTOCORRELATION FOR 2008-2018

```
        Durbin-Watson test

data:  Okuns_model
DW = 2.0522, p-value = 0.952
alternative hypothesis: true autocorrelation is not 0
```

⇓

Since the Durbin–Watson test is not significant & p-Value > 0.05, we fail to reject the null hypothesis and conclude that the model's residuals do not exhibit significant autocorrelation.

## COMPARISON

### 1998-2008

| Coefficients | R-square | Adjusted R-square | F- Stat | P-Value |
|---|---|---|---|---|
| Intercept (estimate): 0.09453<br>Output Gap: -0.06676 | 0.0004122 | -0.1107 | 0.003711 | 0.9528 |

### 2008-2018

| Coefficients | R-square | Adjusted R-square | F stats | P value |
|---|---|---|---|---|
| Intercept(estimate): 0.06219<br>Output gap (estimate): -0.15995 | 0.2925 | 0.4724 | 9.952 | 0.01165 |

# CASE STUDY 2: WAGE AND EDUCATION

GAEE Global Association of Economics Education INDIA

***RESEARCH QUESTION** - What are the determinants of wages, and how significant is education in explaining wage variation?*

## DATASET

**SOURCE -** *Click here to download dataset*

**MAJOR DETERMINANTS**

- Education
- Work Experience
- Gender

These variables significantly influence wage rates

**MINOR DETERMINANTS**

- Education Squared
- Model Specification
- Gender Interaction

These variables do not significantly affect wage prices

- +\$0.54/hr per additional year (simple regression)
- +\$0.61/hour per additional year.
- Earn \$0.99/hour more than unmarried workers
- Earn \$2.27/hour less than comparable males.
- Education explains: 16.5% of wage variation alone ($R^2 = 0.1648$)
- Full Model explains: 24.0% of wage variation ($R^2 = 0.2401$).

## METHODOLOGY

**Research Design**

Quantitative study using Multiple Linear Regression (OLS).

**Data Collection**

- Used the Wooldridge Wage1 dataset containing 526 observations from the 1976 U.S. Current Population Survey (CPS)
- The dataset was cleaned and prepared for analysis, with variables coded appropriately for regression models (including dummy, logarithmic, interaction, and quadratic terms).

**Variables Used**

Dependent Variable: Hourly Wage
Independent Variables: Education, Work Experience, Job Tenure, Gender (Female), Marital Status, Education² (Quadratic Term), and Education × Gender (Interaction Term).

**Statistical Analysis**

- Conducted Simple Linear Regression, Multiple Linear Regression, Log-Linear Regression, Quadratic Regression, and Interaction Regression using OLS.

**Inference**
Coefficient of education is 0.541, implying that an additional year of education increases hourly wage by approximately \$0.54, holding all constant.

- Education has the largest positive impact on wages.
- Experience, tenure, and marriage also increase earnings.
- A gender wage gap exists, but returns to education are similar for men and women.

## SIMPLE LINEAR REGRESSION

| Variable | Coefficient | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | -0.905 | 0.685 | -1.321 | 0.187 |
| Education | 0.541 | 0.053 | 10.167 | <0.001 |

# CASE STUDY 3: HOUSING PRICE DETERMINANTS

***RESEARCH QUESTION** - Which housing characteristics significantly influence residential property prices in India?*

## DATASET

**SOURCE** - *Click here to download dataset*

- 358 residential property listings analyzed
- 7 Indian cities represented
- Mean Property Price: ₹1.07 Cr
- Median Property Price: ₹71 Lakh
- Mean Property Area: 1,382 sq.ft.
- 94.2% of properties are 2–3 BHK
- Price Range: ₹69,900 – ₹7.07 Cr

### MAJOR DETERMINANTS

- Number of Bedrooms
- Number of Bathrooms
- Property Age

### MINOR DETERMINANTS

- Built-up Area
- Parking Availability
- Furnishing Status

## METHODOLOGY

**Research Design**

- Quantitative study using Multiple Linear Regression (OLS).
- Objective: To identify the housing characteristics that influence residential property prices in India

**Data Collection & Preparation**

- Collected data from online real-estate platforms.
- Cleaned dataset by removing inconsistencies and standardizing variables

**Variables Used**

- Dependent Variable: Property Price
- Independent Variables: Bedrooms, Bathrooms, Property Age, Built Area, Parking,Furnishing Status

## REGRESSION MODEL

$$\text{Price} = \beta_0 + \beta_1(\text{Bedrooms}) + \beta_2(\text{Bathrooms}) + \beta_3(\text{Age}) + \beta_4(\text{Area}) + \beta_5(\text{Parking}) + \beta_6(\text{Furnishing}) + \varepsilon$$

Where:

Price = Residential Property Price

$\beta_0$ = Intercept

$\beta_1$–$\beta_6$ = Coefficients measuring the impact of each housing characteristic

$\varepsilon$ = Random error term

# CASE STUDY 3: HOUSING PRICE DETERMINANTS

***RESEARCH QUESTION** - Which housing characteristics significantly influence residential property prices in India?*

| *Regression Statistics* | | ANOVA | | | | | |
|---|---|---|---|---|---|---|---|
| Multiple R | 0.035432136 | | *df* | *SS* | *MS* | *F* | *Significance F* |
| R Square | 0.001255436 | Regression | 1 | 61234043278128 | 61234043278128 | 0.4474971 | 0.503959492 |
| Adjusted R Square | -0.001550026 | Residual | 356 | 48713876486592200 | 136836731703911 | | |
| Standard Error | 11697723.36 | Total | 357 | 48775110529870300 | | | |
| Observations | 358 | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | |
| Intercept | 10685997.34 | 619280.4051 | 17.25551 | 0 | 9468089.531 | 11903905 | |
| X Variable 1 | -1.451347372 | 2.169582898 | -0.66895 | 0.503959492 | -5.718157568 | 2.8154628 | |

| Aspect | Interpretation |
|---|---|
| **Model Fit** | The model has very poor explanatory power ($R^2$ = 0.13%). The negative Adjusted $R^2$ indicates that the predictor does not improve the model beyond the mean property price. |
| **Hypothesis Testing** | Since $p = 0.504 > 0.05$, the null hypothesis cannot be rejected, indicating insufficient evidence that the predictor affects property prices. |
| **Coefficient Analysis** | The coefficient (−1.45) suggests a negative relationship, but the effect is not statistically significant, making it unreliable for prediction. |
| **Confidence Interval** | The 95% confidence interval (−5.72 to 2.82) includes zero, indicating that the true effect of the predictor may be insignificant. |

Since housing prices depend on several characteristics simultaneously, a simple linear regression may not capture these combined effects. A multiple regression model offers a more comprehensive analysis by evaluating the impact of multiple variables together.

## KEY TAKEAWAYS

- Multiple R = 0.035 indicates an extremely weak linear relationship.
- $R^2$ = 0.0013 means the predictor explains only 0.13% of the variation in property prices.
- The overall regression model is not statistically significant($F = 0.447$, $p = 0.504$).
- The predictor (X Variable 1) is not a significant determinant of residential property prices
- The model fails to reject the null hypothesis.
- 95% confidence interval includes zero, indicating no significant effect.
- The negative coefficient is too weak to establish a meaningful relationship.

# CASE STUDY 4: DIGITAL CURRENCY ADOPTION

GAEE Global Association of Economics Education INDIA

**Research Objective -** To evaluate factors influencing digital currency ownership across countries, using Gretl

## DATA & METHODOLOGY

1. Collected cross-sectional dataset of 29 countries
   Find the raw data here
2. Defined - **Dependent Variable:**
   Digital Currency Adoption Rate (%), and
   **Explanatory Variable/Regressors (baseline model):**
   - Internet Penetration (%)
   - Remittances Received (% of GDP)
   - Adults with an Account (%)
   - Inflation (%)
   - Income Dummy for country

   ***Note:*** *Lagged Xs taken*
3. Multicollinearity check: Correlation Matrix + VIF
4. Endogeneity test via OVB test: Add HDI as a proxy f or omitted development factors
5. Compare Model 1 and Model 2 using $R^2$, F-statistic and coefficient changes

## BASELINE MODEL OR MODEL 1

$$Y = \beta_0 + \beta_1 Internet + \beta_2 Remittances + \beta_3 Account + \beta_4 Inflation + \beta_5 Developed$$

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-29 (n = 24)
Missing or incomplete observations dropped: 5
Dependent variable: AdoptionRate_2024

                    coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------------
  const             11.4975       11.0414       1.041    0.3115
  InternetPenetrat~  0.168194      0.121288     1.387    0.1825
  RemittancesGDP_2~ −0.599068      0.430215    −1.392    0.1807
  AdultswithanAcco~ −0.143578      0.0971494   −1.478    0.1567
  Inflation_2023     0.0270217     0.0337419    0.8008   0.4337
  Developed_Dummy_~ −1.41438       2.55762     −0.5530   0.5871

Mean dependent var   12.87083   S.D. dependent var   4.483931
Sum squared resid    299.2020   S.E. of regression   4.077050
R-squared            0.352978   Adjusted R-squared   0.173250
F(5, 18)             1.963954   P-value(F)           0.133278
Log-likelihood      -64.33131   Akaike criterion     140.6626
Schwarz criterion    147.7309   Hannan-Quinn         142.5378

Excluding the constant, p-value was highest for variable 8 (Developed_Dummy_2023)
```

**Regression Equation:** $Y = 11.50 + 0.168X_1 - 0.599X_2 - 0.144X_3 + 0.027X_4 - 1.414X_5$

These weak results suggest that an important explanatory variable may be missing.

| | | | |
|---|---|---|---|
| $\sum$ n = 24 | $R^2$ = 0.353 | $R^2$ (Adjusted) = 0.173 | No individual coefficient is significant at 5% p-values > 0.18 |

# CASE STUDY 4: DIGITAL CURRENCY ADOPTION

## Multicollinearity & Omitted Variable Bias (OVB) Investigation

### MULTICOLLINEARITY CHECK

| Variables | r | p-value | Interpretation |
|---|---|---|---|
| **Adults with an account & Internet Penetration** | 0.628 | 0.0006 | Strong |
| **Internet Penetration & Income Dummy for country** | 0.63 | 0.0002 | Strong |
| **Adults with an Account & Income Dummy for country** | 0.737 | 0 | Strongest |
| **Internet Penetration & Inflation** | -0.071 | 0.7182 | No |
| **Adults with an Account & Remittances GDP** | -0.428 | 0.0327 | Moderate |

**Variance Inflation Factor (VIF)**

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

InternetPenetration2023     1.975
         RemittancesGDPX2     1.473
AdultswithanAccount2021     2.751
          Inflation2023X4     1.215
               DDeveloped     2.345

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables
```

**Inference:**

Multicollinearity exists but is not severe enough to explain the weak model performance (highest VIF = 2.751 < 10)

| **Internet Penetration** | **Remittances % GDP** | **Adults with Account** | **Inflation** | **Income Dummy** |
|---|---|---|---|---|
| 1.975 | 1.473 | 2.751 | 1.215 | 2.345 |

### OMITTED VARIABLE BIAS (OVB)

**Suspected Omitted Variable:** Human Development Index

**Why HDI?**

- Influences internet access
- Influences financial inclusion
- Likely influences digital currency adoption

### DEMONSTRATING ENDOGENEITY VIA OVB

**Model 2: Add HDI**

```
gretl: model 2
File  Edit  Tests  Save  Graphs  Analysis  LaTeX
Model 2: OLS, using observations 1-29 (n = 24)
Missing or incomplete observations dropped: 5
Dependent variable: AdoptionRate_2024

                      coefficient   std. error   t-ratio   p-value
  -----------------------------------------------------------------
  const                49.1288      20.6900       2.375    0.0296   **
  InternetPenetrat~     0.549177     0.213878     2.568    0.0200   **
  RemittancesGDP_2~    -0.974025     0.433955    -2.245    0.0384   **
  AdultswithanAcco~    -0.125820     0.0896011   -1.404    0.1783
  Inflation_2023        0.0645325    0.0358184    1.802    0.0894   *
  Developed_Dummy_~     7.12379      4.71813      1.510    0.1494
  HDI_2023            -88.8730      42.5959      -2.086    0.0523   *

Mean dependent var   12.87083   S.D. dependent var   4.483931
Sum squared resid    238.2053   S.E. of regression   3.743271
R-squared            0.484883   Adjusted R-squared   0.303077
F(6, 17)             2.667036   P-value(F)           0.051996
Log-likelihood      -61.59547   Akaike criterion     137.1909
Schwarz criterion    145.4373   Hannan-Quinn         139.3787

Excluding the constant, p-value was highest for variable 4 (AdultswithanAccount_2021)
```

If coefficient estimates change substantially after adding HDI, it suggests potential omitted variable bias.

$$Y = \beta_0 + \beta_1 Internet + \beta_2 Remittances + \beta_3 Account + \beta_4 Inflation + \beta_5 Developed + \beta_6 HDI$$

# CASE STUDY 4: DIGITAL CURRENCY ADOPTION

## RESULTS COMPARISION

| Metric | Model 1 | Model 2 |
|---|---|---|
| $R^2$ | 0.353 | 0.485 |
| Adj. $R^2$ | 0.173 | 0.303 |
| F-test p-value | 0.133 | 0.052 |

**Coefficient Changes** (Key Variables):

| Variable | Model 1 | Model 2 |
|---|---|---|
| Internet | 0.168 | 0.549 |
| Remittances | -0.599 | -0.974 |
| Income Dummy | -1.414 | 7.124 |

**Takeaway:** HDI was correlated with both digital currency adoption and several regressors. Omitting it caused biased coefficients in Model 1. After including HDI, model fit improved substantially and coefficient estimates changed significantly, illustrating endogeneity arising from omitted variable bias.

## STEP-BY-STEP RUNDOWN IN GRETL

1. **Import Data** — File → Open Data → Import → Excel/CSV
2. **Descriptive Statistics** — Variable → Summary Statistics
3. **Estimate Model 1** — Model → Ordinary Least Squares (OLS) → Select Variables → OK
4. **Correlation Matrix** — View → Correlation Matrix
5. **Multicollinearity Test (VIF)** — Analysis → Collinearity
6. **Estimate Model 2** — Model → Ordinary Least Squares (OLS) → Add HDI → OK
7. **Compare Models** — Compare $R^2$, Adjusted $R^2$, F-statistic, coefficients and p-values
8. **Interpret Endogeneity** — Compare Model 1 and Model 2 coefficient estimates
9. **Report Limitations** — Document methodological limitations

## KEY TAKEAWAYS

- Theoretical reasoning should guide variable selection before model estimation.
- Always diagnose multicollinearity before interpreting coefficients.
- Significant changes in coefficients after adding a relevant control variable may indicate omitted variable bias.
- Better model fit does not automatically imply causality.
- Regression is as much about specification and interpretation as it is about estimation.

## LIMITATIONS

Sample restricted to high-adoption countries due to data availability. Results may not be fully representative of all countries, creating potential sample selection bias and limiting generalizability.

# ECONOMIC AND POLICY FRAMEWORKS

# LINKING REGRESSION TO ECONOMIC THEORY

Regression models rely on economic theory to guide variable selection, expected coefficient signs, and functional form. Without theory, results may be statistically valid but economically meaningless.

## 01 INCENTIVES AND BEHAVIOURAL CONSIDERATIONS

- Economic agents respond to incentives, making some independent variables endogenous and causing OLS estimates to be biased.

- Example: If a central bank raises interest rates when credit demand is already high, the observed relationship reflects both policy and demand, making the causal effect difficult to isolate.

## 02 DEMAND AND SUPPLY RELATIONSHIPS

Economic theory guides regression by specifying relevant variables and expected coefficient signs.

- **$\beta_1 < 0$:** Higher price lowers demand.
- **$\beta_2 > 0$:** Higher income increases demand for normal goods.
- **$\beta_3$:** Positive for substitutes, negative for complements.

Theory-driven models produce more meaningful and interpretable results.

## 03 TRANSLATING THEORY INTO EMPIRICAL ANALYSIS

- Moving from theory to estimation requires choosing appropriate variables, proxies, and functional forms. Poor proxies or incorrect functional forms can weaken model validity.
- Control variables help satisfy the ceteris paribus condition, making regression a more reliable test of economic theory.

# POLICY EVALUATION USING REGRESSION



## IMPACT EVALUATION BASICS

Impact evaluation measures causal effects, not just correlations. The counterfactual what would have happened without the policy is estimated using methods like:

- **RCTs**: random assignment to treatment/control
- **DiD**: compare changes over time across groups
- **RDD**: use eligibility cut-offs

Regression estimates the Average Treatment Effect (ATE) with a treatment dummy. Controls for confounders (age, education, income, etc.) ensure unbiased results.

**Example**

A skill program's effect on youth employment is measured by comparing participants and non-participants, controlling for education and experience.

Evidence-based policymaking follows:
***Implement → Evaluate → Learn → Refine***

This cycle improves effectiveness, transparency, and accountability.

## CONCLUSION

Regression enables impact evaluation and evidence-based policymaking. By estimating causal effects, controlling confounders, and measuring outcomes, it helps governments allocate resources wisely, improve welfare, and refine policies continuously.

## REGRESSION IN GOVERNMENT DECISION-MAKING

Governments use regression to evaluate subsidies, taxes, welfare, healthcare, and education. Coefficients show policy impact:

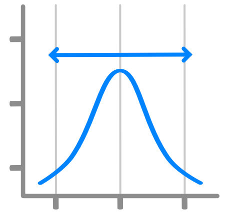
- **+ = beneficial**
- **− = adverse**
- **> = stronger effect**

Regression supports cost-benefit analysis, guiding efficient resource use. ***Example***: Minimum wage effects on employment are estimated while controlling for growth and industry factors.

## EVIDENCE-BASED POLICYMAKING

Regression provides quantitative evidence for policy goals.

- **p-values**: small (<0.05) indicate real effects



- **Confidence intervals:** show range and uncertainty
- Reveals unintended consequences across groups

**Example:** A scholarship program raises attendance by 12%, with stronger effects in rural areas, guiding better targeting.

# INTERPRETING RESULTS FOR POLICY AND BUSINESS

## REGRESSION MODEL INTERPRETATION

**Coefficients:** Understanding the direction and magnitude of the relationship between input variables and the outcome variable

**p-values:** Determining the statistical significance of the coefficients

**R-squared:** Evaluating the goodness of fit of the model

**Residual plots:** Checking for assumptions of linearity, homoscedasticity, and normality

## WHAT REGRESSION OUTPUT INCLUDES

**Coefficients:** Estimated values of the coefficients

**Standard Errors:** Measure the variability of the coefficients

**T-statistics:** Measure the statistical significance of the coefficients

**P-values:** Probabilities of observing the t-statistics under the null hypothesis

**R-squared:** Proportion of the variance explained by the model

**F-statistic:** Measures the overall significance of the model

## PRACTICAL APPLICATIONS & TRADE-OFFS

**Policy Example**
Unemployment regressed on output gap:
Coefficient -0.6, p = 0.01,
$R^2 = 0.45$, CI [-0.9, -0.3]
- Coefficient: substantial effect
- p-value: statistically significant
- $R^2$: explains 45% variation
- CI: consistently negative

Decision: policymakers justify counter-cyclical spending; businesses anticipate labor shortages.

**Business Applications**
Regression helps:
- Predict outcomes under scenarios
- Identify key drivers
- Optimize resources
- Evaluate strategies

Example: Sales predicted from marketing spend and seasonality → forecast sales, identify effective channels, optimize expenditure.

**Trade-offs & Uncertainty**
- p-values: test if coefficients differ from zero
- Confidence intervals: range of likely values
- Confidence bands: uncertainty around prediction line, narrowest at mean, widest at extremes

Practical trade-off: central predictions more reliable; edge predictions need caution. Narrow intervals support confident action; wide bands call for more data.

## RISKS OF MISINTERPRETATION

**Correlation vs Causation**
Hidden drivers can mislead (e.g., ice cream sales vs drowning rates). A significant coefficient does not prove causation.

**Extrapolation Beyond Data**
Results may not hold outside the observed range. Models only describe relationships within the data collected.

**Statistical vs Practical Significance**
Small but statistically significant effects may be irrelevant in real-world policy or business decisions.

**Overconfidence & p-hacking**
Small samples or multiple tests mislead. Running many tests inflates false positive rates without correction.

**Biases**
Omitted variables, measurement error, and selection bias can all distort coefficient estimates and limit generalizability.

# INTERVIEW AND EXAM TOOLKIT

# RUNNING AND READING A REGRESSION

## RUNNING A BASIC REGRESSION IN R

INSTALL AND LOAD PACKAGES: INSTALL.PACKAGES("PACKAGE_NAME") LIBRARY(PACKAGE_NAME) → IMPORT DATASET AND ASSIGN TO A VARIABLE. → EXPLORE DATA USING SUMMARY(DATASET). → ACCESS COLUMNS WITH DATASET$COLUMN_NAME. → RUN REGRESSION: MODEL < -LM(Y ~ X, DATA = DATASET) SUMMARY(MODEL)

## READING OUTPUT

- **Coefficient:** Shows direction and size of effect.
- **p-value:** Indicates statistical significance.
- **Standard Error:** Precision of estimate.
- **Constant:** Baseline value.
- **$R^2$:** Goodness of fit.
- **Sample size:** Number of observations.

## KEY INTERPRETATIONS

- **Coefficients:** Measure impact of X on Y (e.g., slope = 2 means Y rises 2 units per 1 unit of X).
- **Standard Errors:** Smaller = more precise; larger = less precise.
- **Significance:** $p < 0.05$ or test statistic > critical value → significant.

## APPLIED CASES

**Case 1:** Regression line equation: $y = 40 + 6x$, where sales (units sold each day, y) depends on advertising cost (in hundreds of dollars, x). If x = 5, the predicted value from the equation is 70 units, but the actual value observed is 60 units.. What is the residual value of this observation? (Answer: Residual value equals −10, meaning that the prediction overestimates the number of units sold.)

**Case 2:** Regression equation predicts productivity score (y) from training hours (x). Slope $\beta_1 = 0.5$, with 95% confidence interval (0.1, 0.9). What is the most applied conclusion for managers? (Answer: One hour of training increases the productivity score by 0.5.)

# COMMON MISTAKES IN ECONOMETRICS INTERVIEWS AND EXAMS

**Definitions Without Interpretation:** Giving textbook definitions without practical implications. Example: defining heteroskedasticity but not its effect on standard errors and hypothesis testing.

**Ignoring Economic Intuition:** Treating econometrics as pure math, failing to link results to theory or real-world behavior.

**Jumping to Conclusions:** Declaring significance without discussing null hypothesis, test statistic, p-value, and reasoning.

**Statistical vs Economic Significance:** Assuming statistical significance equals importance, ignoring coefficient magnitude and practical impact.

**Forgetting Model Assumptions:** Neglecting OLS assumptions and issues like multicollinearity, heteroskedasticity, autocorrelation, endogeneity. Skipping diagnostic tests.

**Overloading with Equations:** Writing formulas without explanation or interpretation.

**Poor Answer Structure:** Mixing assumptions, methods, and conclusions, leading to unclear flow.

## FRAMEWORK FOR STRUCTURING ANSWERS

**Define the Concept** concise definition, key terms.

→ **Explain Economic Intuition** why it matters, economic problem addressed.

→ **Present Statistical Framework** model, hypotheses, estimation.

→ **Interpret Results** coefficient sign, magnitude, statistical and economic significance.

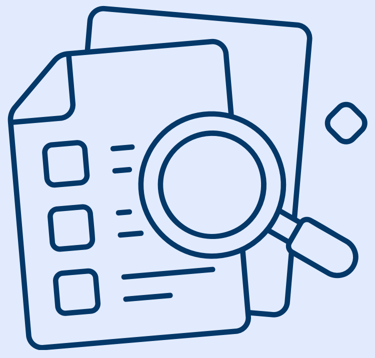
→ **Discuss Assumptions & Limitations** possible violations, data constraints, omitted variables, measurement errors.

→ **Provide Practical Example** link to business, policy, or economic application.

# IMPLEMENTATION AND CONCLUSION

# BEYOND OLS

## INTRODUCTION TO ADVANCED ECONOMETRIC METHODS



While OLS is a useful starting point, complex economic questions often require advanced methods. These techniques address issues like **limited dependent variables, unobserved heterogeneity, and time-series data**, providing more reliable estimates when OLS assumptions are violated.

## INTRODUCTION TO ADVANCED ECONOMETRIC METHODS

### PANEL DATA COMBINES: CROSS SECTIONAL DATA + TIME SERIES DATA

**Cross sectional data:** (different individuals, firms, or countries)
**Time series data:** (observations over multiple years)

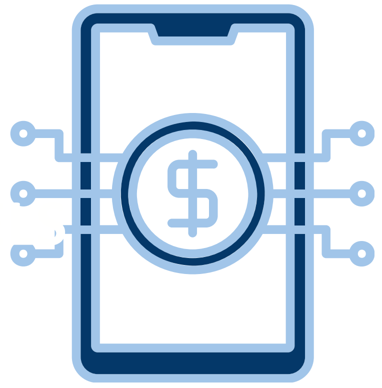
A **Fixed Effects Model** controls for characteristics that remain constant over time but differ across entities.

### BENEFITS

- **Reduces omitted variable bias**
- **Captures changes within entities over time**
- **Produces reliable estimates to analyze longitudinal data.**

## PANEL DATA AND FIXED EFFECTS

**Logit and Probit models** are used when the **dependent variable** is **binary**. For instance, it takes only two values:



- **YES OR NO**
- **1 OR 0**

**Examples**

- *Whether a person adopts digital currency* ***(Yes/No)***
- *Whether a household owns a vehicle* ***(Yes/No)***

| ADVANTAGES | LIMITATIONS |
|---|---|
| Produces probabilities between 0 and 1. | Less intuitive interpretation than OLS. |
| Better suited for decision - making outcomes than OLS. | Requires larger sample sizes for reliable estimates. |
| Widely applied in economics, finance, marketing & public policy. | Assumes correct specification of link function (normal or logistic) |

# REFERENCES AND FURTHER READINGS

## RECOMMENDED TEXTBOOKS


- Gujarati, D. N., & Porter, D. C. *Basic Econometrics*.
- Wooldridge, J. M. *Introductory Econometrics: A Modern Approach*.
- Stock, J. H., & Watson, M. W. *Introduction to Econometrics*.
- Kennedy, P. *A Guide to Econometrics*.


## RESEARCH PAPERS


- White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator."
- Breusch, T. S., & Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation."
- Durbin, J., & Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression."
- Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis."


## DATA SOURCES AND DATABASES

- World Bank Open Data
- International Monetary Fund (IMF) Data
- Reserve Bank of India (RBI) Database
- Ministry of Statistics and Programme Implementation (MOSPI)
- OECD Data

## SOFTWARE AND LEARNING SOURCES

- Gretl (GNU Regression, Econometrics and Time-series Library)
- R and RStudio
- Stata
- Python (Statsmodels, Scikit-learn)

# APPENDIX

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.58743351 |
| R Square | 0.34507813 |
| Adjusted R Square | 0.33388288 |
| Standard Error | 9539823.06 |
| Observations | 358 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regressior | 6 | 1.68E+16 | 2.81E+15 | 30.8236318 | 1.0504E-29 |
| Residual | 351 | 3.19E+16 | 9.1E+13 | | |
| Total | 357 | 4.88E+16 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -13595668 | 2131904.21 | -6.37724 | 5.7E-10 | -1.8E+07 | -9402755 | -17788582 | -9402755.3 |
| X Variable 1 | 0.20784601 | 1.78867642 | 0.1162 | 0.90756 | -3.31003 | 3.725717 | -3.3100254 | 3.72571742 |
| X Variable 2 | 4391015.14 | 914132.408 | 4.80348 | 2.3E-06 | 2593149 | 6188881 | 2593149.28 | 6188880.99 |
| X Variable 3 | 4405914.34 | 856485.424 | 5.14418 | 4.5E-07 | 2721425 | 6090403 | 2721425.43 | 6090403.24 |
| X Variable 4 | 188728.615 | 78330.7256 | 2.40938 | 0.01649 | 34672.01 | 342785.2 | 34672.0084 | 342785.222 |
| X Variable 5 | 1491063.35 | 1101362.5 | 1.35384 | 0.17666 | -675036 | 3657163 | -675036.46 | 3657163.17 |
| X Variable 6 | 316015.453 | 1055386.47 | 0.29943 | 0.76479 | -1759661 | 2391692 | -1759661.2 | 2391692.11 |

# R CODES

```r
#Regression through R
#Setting working directory
getwd()
setwd("C:/Users/hp/Desktop/Ecotrix")

#Installing packages
install.packages("wooldridge")
install.packages("car")
install.packages("lmtest")
install.packages("sandwich")
install.packages("stargazer")
#Setting up packages
library(wooldridge)
library(car)
library(lmtest)
library(sandwich)
library(stargazer)

#Using dataset Wage1
data("wage1")
#Viewing data
str(data)
head(data)

#Simple linear regression of wage on education
model_educ <- lm(wage ~ educ, data = wage1)
summary(model_educ)

#Multiple linear regression
model_multi <- lm(wage ~ educ + exper + married, data = wage1)
summary(model_multi)
```

```r
#Log-linear model
model_loglin <- lm(log(wage) ~ educ + exper + tenure, data = wage1)
summary(model_loglin)

#Scatter Plot
plot(wage1$educ, wage1$wage,
     xlab="Education",
     ylab="Wage",
     main="Wage vs Education")

#Confidence Intervals
confint(model_educ)
confint(model_educ, level = 0.90)

#Dummy Variables
model_dummy <- lm(wage ~ educ + female, data = wage1)
summary(model_dummy)

#Quadratic Model
model_quad <- lm(wage ~ educ + I(educ^2),
                 data=wage1)
summary(model_quad)

#Interaction Model
model_interaction <- lm(wage ~ educ*female,
                        data=wage1)
```

```r
summary(model_interaction)

#Scatter plot using ggplot
library(ggplot2)
ggplot(wage1, aes(x=educ, y=wage)) +
  geom_point()

#Fitted Values
fitted(model_multi)

#Residuals (u-hat)
residuals(model_multi)

#Store fitted values and residuals
wage1$y_hat <- fitted(model_multi)
wage1$u_hat <- residuals(model_multi)

#Verify residual
wage1$wage[1] - wage1$y_hat[1]
wage1$u_hat[1]

# Sum of residuals
sum(wage1$u_hat)

#VIF
library(car)
vif(model_multi)

#Ramsey Reset Test
library(lmtest)
resettest(model_multi,
          power=2:3,
          type="fitted")
```

# OKUN'S MODEL

```r
library(readxl)
data_1<-read_excel(path="C:/Users/USER/Desktop/Okun's_Law.xlsx")

output_gap<-(data_1$Actual_Growth_Rate_of_Output-
data_1$Potential_Growth_Rate_Of_Output)

Okuns_model<-lm(data_1$Change_in_Unemployment_Rate~output_gap)
summary(Okuns_model)


anova(Okuns_model)

plot(data_1$Change_in_Unemployment_Rate, output_gap, xlab="Change In
Unemployment",
     ylab="OutputGap",pch=16,col="blue",las=1,main="Okun's Law For
India(2008-18)")
abline(a=0.06219,b=-0.15995, col="red",lwd=4)


library(lmtest)
dwtest(Okuns_model,alternative="two.sided")

plot(data_1$YEAR, output_gap,
     type = "b",
     ylim = range(c(output_gap,
                    data_1$`Change_In_Unemployment_Rate`)),
     xlab = "Year",
     ylab = "Value",
     main = "Output Gap and Change in Unemployment")

lines(data_1$YEAR,
      data_1$Change_in_Unemployment_Rate,
      type = "b",
      lty = 2)

legend("bottomright",
       legend = c("Output Gap", "Change in Unemployment"),
       lty = c(1, 2))
```

```r
install.packages("readxl")
library(readxl)
data <- read_excel("C:/Users/Kalyani/Downloads/Okuns law case
study.xlsx")
model_1 <- lm(`Change in Unemployment` ~ `Output Gap`, data=data)
summary(model_1)
plot(data$`Change in Unemployment`, data$`Output Gap`,
main = "Scatter Plot: unemployment vs output",
xlab= "output gap",
ylab="unemployment change",
pch=19,
col="blue")
install.packages("lmtest")
library(lmtest)
dwtest(model_1)
install.packages("ggplot2")
library(ggplot2)

ggplot(data,
       aes(x = `Output Gap`,
           y = `Change in Unemployment`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Okun's Law",
       x = "Output Gap",
       y = "Change in Unemployment")
plot(data$Year,
     data$`Output Gap`,
     type = "b",
     ylim = range(c(data$`Output Gap`,
                    data$`Change in Unemployment`)),
     xlab = "Year",
     ylab = "Value",
     main = "Output Gap and Change in Unemployment")

lines(data$Year,
      data$`Change in Unemployment`,
      type = "b",
      lty = 2)

legend("topright",
       legend = c("Output Gap", "Change in Unemployment"),
       lty = c(1, 2))
```

# CASE STUDY 4: GRETL COMMAND LOG

```
# This script has been generated from the Gretl command log
# and lightly edited for readability.
# Import dataset
open "Digital Currency Adoption_Raw Data.csv"
# Create dummy variable
dummify Developed_DummyX5
# Rename dummy variable with developed = 1
rename 8 Developed_Dummy_2023

# Model 1: Baseline OLS Regression
ols AdoptionRate_2024 0 \
InternetPenetration_2023 \
RemittancesGDP_2023 \
AdultswithanAccount_2021 \
Inflation_2023 \
Developed_Dummy_2023

# Pairwise Correlation Analysis
corr AdultswithanAccount_2021 InternetPenetration_2023
corr InternetPenetration_2023 Developed_Dummy_2023
corr AdultswithanAccount_2021 Developed_Dummy_2023
corr InternetPenetration_2023 Inflation_2023
corr AdultswithanAccount_2021 RemittancesGDP_2023

# Multicollinearity Diagnostics
vif
# Model 2: OLS Regression with HDI
# (Testing for Potential Omitted Variable Bias)
ols AdoptionRate_2024 0 \
InternetPenetration_2023 \
RemittancesGDP_2023 \
AdultswithanAccount_2021 \
Inflation_2023 \
Developed_Dummy_2023 \
HDI_2023

# End of Script
```

# CASE STUDY 4: GRETL OUTPUT TABS

```
Model 1: OLS, using observations 1-29 (n = 24)
Missing or incomplete observations dropped: 5
Dependent variable: AdoptionRate_2024

                      coefficient   std. error   t-ratio   p-value
  -----------------------------------------------------------------
  const                11.4975      11.0414       1.041    0.3115
  InternetPenetrat~     0.168194     0.121288     1.387    0.1825
  RemittancesGDP_2~    -0.599068     0.430215    -1.392    0.1807
  AdultswithanAcco~    -0.143578     0.0971494   -1.478    0.1567
  Inflation_2023        0.0270217    0.0337419    0.8008   0.4337
  Developed_Dummy_~    -1.41438      2.55762     -0.5530   0.5871

Mean dependent var   12.87083   S.D. dependent var   4.483931
Sum squared resid    299.2020   S.E. of regression   4.077050
R-squared            0.352978   Adjusted R-squared   0.173250
F(5, 18)             1.963954   P-value(F)           0.133278
Log-likelihood      -64.33131   Akaike criterion     140.6626
Schwarz criterion    147.7309   Hannan-Quinn         142.5378

Excluding the constant, p-value was highest for variable 8 (Developed_Dummy_2023)
```

```
Model 2: OLS, using observations 1-29 (n = 24)
Missing or incomplete observations dropped: 5
Dependent variable: AdoptionRate_2024

                      coefficient   std. error   t-ratio   p-value
  -----------------------------------------------------------------
  const                49.1288      20.6900       2.375    0.0296   **
  InternetPenetrat~     0.549177     0.213878     2.568    0.0200   **
  RemittancesGDP_2~    -0.974025     0.433955    -2.245    0.0384   **
  AdultswithanAcco~    -0.125820     0.0896011   -1.404    0.1783
  Inflation_2023        0.0645325    0.0358184    1.802    0.0894   *
  Developed_Dummy_~     7.12379      4.71813      1.510    0.1494
  HDI_2023            -88.8730      42.5959      -2.086    0.0523   *

Mean dependent var   12.87083   S.D. dependent var   4.483931
Sum squared resid    238.2053   S.E. of regression   3.743271
R-squared            0.484883   Adjusted R-squared   0.303077
F(6, 17)             2.667036   P-value(F)           0.051996
Log-likelihood      -61.59547   Akaike criterion     137.1909
Schwarz criterion    145.4373   Hannan-Quinn         139.3787

Excluding the constant, p-value was highest for variable 4 (AdultswithanAccount_202
```

# CASE STUDY 4: GRETL OUTPUT TABS

```
corr(InternetPenetration_2023, Developed_Dummy_2023) = 0.63021498
Under the null hypothesis of no correlation:
 t(27) = 4.21768, with two-tailed p-value 0.0002
```

```
corr(AdultswithanAccount_2021, InternetPenetration_2023) = 0.62837256
Under the null hypothesis of no correlation:
 t(24) = 3.95724, with two-tailed p-value 0.0006
```

```
corr(AdultswithanAccount_2021, Developed_Dummy_2023) = 0.73718686
Under the null hypothesis of no correlation:
 t(24) = 5.34486, with two-tailed p-value 0.0000
```

```
corr(InternetPenetration_2023, Inflation_2023) = -0.07137399
Under the null hypothesis of no correlation:
 t(26) = -0.364868, with two-tailed p-value 0.7182
```

```
corr(AdultswithanAccount_2021, RemittancesGDP_2023) = -0.42830031
Under the null hypothesis of no correlation:
 t(23) = -2.2731, with two-tailed p-value 0.0327
```

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

InternetPenetration_2023    1.975
     RemittancesGDP_2023    1.473
AdultswithanAccount_2021    2.751
          Inflation_2023    1.215
    Developed_Dummy_2023    2.345

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables
```

# THANK YOU!

GAEE

Global Association of Economics Education

INDIA

/gaeeindia northindia@gaee.org