PMDS504L: Regression Analysis and Predictive Models

Model Adequacy Checking in Regression Analysis

Dr. Jisha Francis

Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology
Vellore Campus, Vellore - 632 014
India

Table of Contents

Introduction to Model Adequacy Checking

In regression analysis, we make several assumptions about the model. These assumptions need to be checked to ensure the adequacy of the model. Below are the key assumptions we consider:

1. The relationship between the response Y and the regressors is linear, at least approximately.
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

Assumptions in Detail

Key Points:

- Assumptions 4 and 5 imply that the errors are independent random variables.
- Assumption 5 is crucial for hypothesis testing and interval estimation.

Why Model Adequacy Checking?

- These assumptions are essential for model stability and the validity of conclusions drawn.
- It is important to verify if the assumptions hold, as gross violations can lead to misleading or unstable models.

Consequences of Violating Assumptions

What can happen if the assumptions are violated?

- Unstable models: A small change in the sample may lead to drastically different conclusions.
- Model performance and predictions could be unreliable.
- It can lead to incorrect estimations and inference.

Can We Detect Violations?

- Metrics like R^2, t-tests, and F-tests only provide global measures.
- They do not capture underlying violations of model assumptions.

Diagnostics for Model Assumption Violations

In this module, we discuss several methods to detect violations of basic regression assumptions. These diagnostic methods primarily rely on analyzing the **model residuals**.

Checking Linear Relationship

- **Scatter Diagram:** When there is only one explanatory variable X in the model, a scatter diagram is useful to visualize the linear relationship between the dependent variable Y and X.
- **Interpretation:**
  - If the scatter diagram shows a linear trend, it indicates that the relationship between Y and X is linear.
  - If the pattern appears curved or complex, it suggests that the relationship is nonlinear.

Linear vs Nonlinear Trends

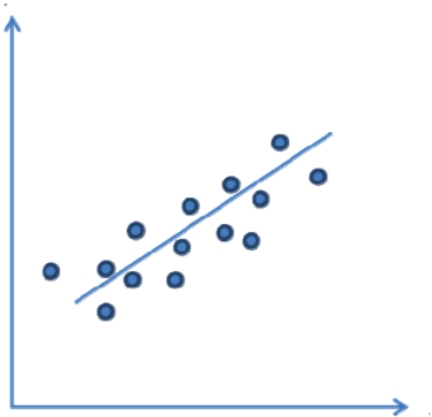


Figure: Example of a Linear Trend

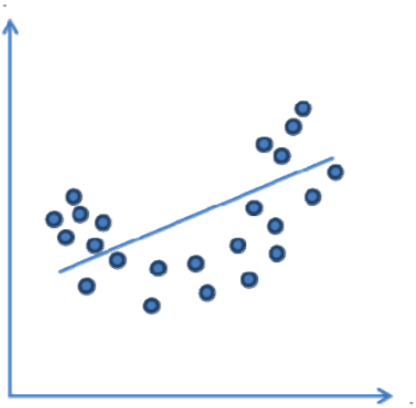


Figure: Scatter Plot Suggesting a Nonlinear Trend

Case of more than one explanatory variables

Scatterplot Matrix for Model Linearity

- A scatterplot matrix is a two-dimensional grid of plots where:
  - Each off-diagonal cell contains a scatterplot between two variables.
  - Diagonal cells may contain histograms or remain empty.
- This matrix provides information about the relationships between pairs of variables:
  - Helps visualize linear or nonlinear patterns.
  - Gives insight beyond simple correlation coefficients.

Displaying the Scatterplot Matrix

- Two common approaches for displaying the scatterplot matrix:
  - Upper triangular part: Scatterplots.
  - Lower triangular part: Corresponding correlation coefficients.
- Example for a model with two explanatory variables:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Displaying the Scatterplot Matrix

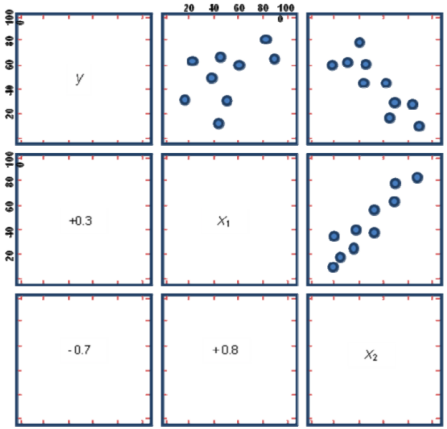


Figure: Scatterplot Matrix Example

Interpretation and Cautions

- Correlation coefficients measure only linear relationships and can be influenced by outliers.
- Pairwise scatterplots provide insight, but the assumption of linearity involves the joint relationship of Y with all explanatory variables.
- Interrelationships between explanatory variables may cause scatterplots to be misleading.
- Alternative methods are sometimes required to explore relationships between study and explanatory variables.

Introduction to Normal Probability Plots

Purpose of Normal Probability Plots:

- Verify the assumption of normal distribution of residuals.
- Assess the adequacy of regression models.
- Detect departures from normality.

Why Normality Matters:

- Ensures the validity of hypothesis testing, confidence intervals, and prediction intervals.
- Gross deviations from normality can severely affect model performance.

Plot Construction

Steps for Creating a Normal Probability Plot:

1. Order the residuals $e_{[1]} < e_{[2]} < \cdots < e_{[n]}$.
2. Compute cumulative probabilities $P_i = \frac{i-0.5}{n}$ for $i = 1, 2, \ldots, n$.
3. Plot the ordered residuals against the cumulative probabilities.

Interpretation: Points should approximately lie along a straight line if residuals are normally distributed.

Rationale Behind Plotting

Key Considerations:

- Divide the area under the normal curve into n equal sections.
- Assume the cumulative area up to each midpoint as $P_i = \frac{i-0.5}{n}$.
- Plot residuals against these probabilities to visualize deviations.

Interpreting Normal Probability Plots

- (a) **Ideal plot:** Points lie approximately on a straight line indicating normal distribution.
- (b) **Heavy-tailed distribution:** Sharp curves at extremes, thicker tails than normal.
- (c) **Light-tailed distribution:** Flattening at the extremes, thinner tails than normal.
- (d) **Positively skewed distribution:** Upward trend change from the midpoint.
- (e) **Negatively skewed distribution:** Downward trend change from the midpoint.

Example Plots

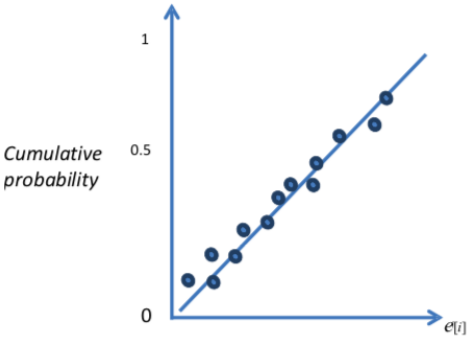


Figure: (a) Ideal normal probability plot: Points on a straight line.

Heavy-Tailed Distribution

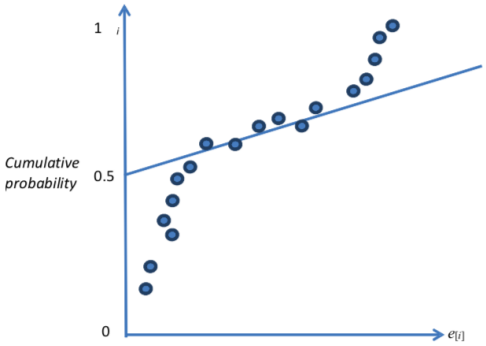


Figure: (b) This figure has sharp upward and downward curves at both extremes. This indicates that the underlying distribution is heavy-tailed, i.e., the tails of the underlying distribution are thicker than the tails of normal distribution.

Light-Tailed Distribution

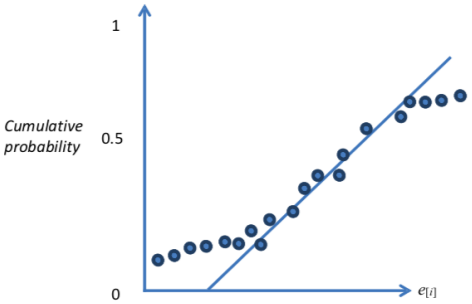


Figure: (c) This figure has flattening at the extremes for the curves. This indicates that the underlying distribution is light-tailed, i.e., the tails of the underlying distribution are thinner than the tails of normal distribution.

Positively Skewed Distribution

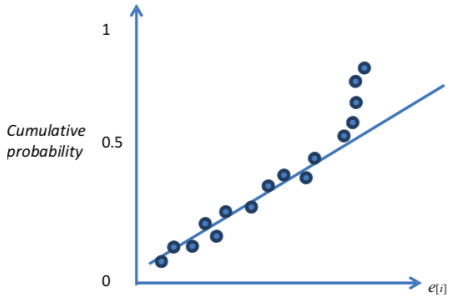


Figure: (d) This figure has a sharp change in the direction of the trend in an upward direction from the mid. This indicates that the underlying distribution is positively skewed.

Negatively Skewed Distribution

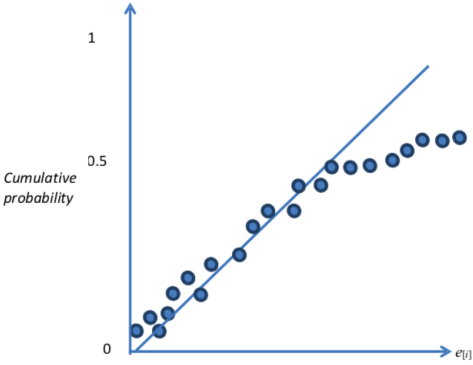


Figure: (e) This figure has a sharp change in the direction of the trend in the downward direction from the mid. This indicates that the underlying distribution is negatively skewed.

Important Notes on Interpretation

- Small sample sizes ($n \leq 16$) often deviate substantially from linearity.
- Larger sample sizes ($n \geq 32$) produce more stable and interpretable plots.
- Parameter fitting can destroy evidence of non-normality.
- Outliers may appear as large residuals, pulling the plot off the expected line.

Conclusion: Normal Plots

- Normal probability plots help validate model assumptions.
- Departure from straight-line behavior indicates potential issues.
- Understanding plot patterns aids in selecting appropriate estimation techniques.
- Expertise is required to interpret subtle deviations.

Residual Analysis

Definition of Residuals

- Residuals are the differences between observed and predicted values in a regression model:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

- Here, y_i is the observed response and $\hat{y}_i$ is the fitted or predicted value.
- Residuals measure the variability in the response variable not explained by the regression model.

Why Residual Analysis?

Purpose of Analyzing Residuals:

- Residuals are realized or observed values of model errors.
- Any deviations from assumptions about the model errors will be reflected in the residuals.
- Plotting residuals helps investigate:
  - Model fit.
  - Linearity assumptions.
  - Homoscedasticity (constant variance).
  - Normality of errors.

Key Properties of Residuals

Important Properties of Residuals:

- **Zero Mean:** The sum of residuals is approximately zero. $\sum_{i=1}^{n} e_i = 0$
- **Variance Estimate:** The approximate average variance of residuals is given by:

$$MS_{Res} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p}$$

  where $p = k + 1$.

- **Degrees of Freedom:** Residuals are not independent, as they have only $n - p$ degrees of freedom, where: n is the number of observations and p is the number of parameters in the model.
- Despite non-independence, residual analysis remains effective if n is large relative to p.

Types of Residuals

- **Original Residuals** (e_i):

$$e_i = y_i - \hat{y}_i$$

- **Scaled Residuals**: Residuals adjusted to maintain constant variance.
- **Studentized Residuals**: Residuals divided by an estimate of their standard deviation.

$$r_i = \frac{e_i}{\hat{\sigma}_i}$$

- Preferred for plotting due to their constant variance property.

Methods for Scaling Residuals

Why Scale Residuals?

- Scaled residuals help in identifying outliers or extreme values.
- Useful for detecting observations that are significantly different from the rest of the data.

Standardized Residuals

- The approximate average variance of a residual is estimated by:

$$MS_{\text{Res}} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p}$$

- A logical scaling for residuals is the **standardized residual**:

$$d_i = \frac{e_i}{\sqrt{MS_{\text{Res}}}}$$

where:

- $e_i = Y_i - \hat{Y}_i$ (residual)
- MS_{Res} is the mean square error of the residuals.

Properties of Standardized Residuals

- Standardized residuals have:

$$\text{Mean} = 0, \quad \text{Variance} \approx 1$$

- A large standardized residual ($|d_i| > 3$) potentially indicates an **outlier**.

Studentized Residuals

Why Improve Residual Scaling?

- Standardized residuals use an approximation for variance.
- **Studentized residuals** improve scaling by dividing e_i by its exact standard deviation.

Residuals and the Hat Matrix

- The residual vector is given by:

$$e = (I - H)y$$

  where $H = X(X'X)^{-1}X'$ is the **hat matrix**.

- The hat matrix properties:
  - Symmetric: $H' = H$
  - Idempotent: $HH = H$

Studentized Residuals

Variance of Residuals

- The covariance matrix of the residuals:

$$\text{Var}(e) = \sigma^2(I - H)$$

- The variance of the ith residual:

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

  where h_{ii} is the ith diagonal element of H.
- The covariance between residuals e_i and e_j:

$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Definition of Studentized Residuals

$$r_i = \frac{e_i}{\sqrt{MS_{res}(1 - h_{ii})}} \tag{1}$$

where:

- e_i is the raw residual.
- MS_{res} is the mean squared residual.
- h_{ii} is the leverage of the i^{th} observation.

Properties of Studentized Residuals

- Expected value: $E(r_i) = 0$
- Variance: $\text{Var}(r_i) = 1$
- This holds regardless of the location of x_i when the model is correct.

Influential Points and Leverage

- Any point with:
  1. A large residual.
  2. A large leverage value h_{ii}.
- Such points are potentially highly influential on the least-squares fit.
- If there is only one explanatory variable, r_i is generally recommended for examination.

Conclusion: Residuals

- Studentized residuals r_i are useful for assessing model fit.
- They are preferable in identifying influential data points.
- Their standardized variance ensures stability in interpretation.

Pure Leverage Points

- A data point is a pure leverage point when it is remote in the regressor space.
- The observed response is consistent with predictions from other data points.

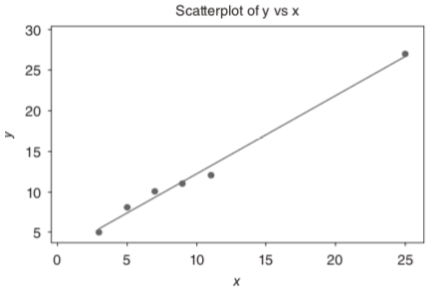


Figure 4.1 Example of a pure leverage point.

Figure: Scatter plot showing a pure leverage point.

Influential Points

- A data point is influential when it is both remote in regressor space and inconsistent with predicted values.
- Influential points "drag" the prediction equation toward themselves.

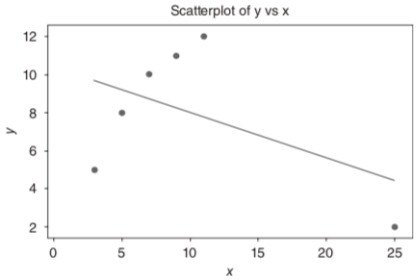


Figure 4.2 Example of an influential point.

Residual Plots in Regression Analysis

- Residual plots against fitted values are useful in detecting model inadequacies.
- They help in identifying patterns that indicate violations of regression assumptions.

Residual Plots

- Graphical analysis of residuals is a powerful way to:
  - Investigate the adequacy of the regression model fit.
  - Check the validity of underlying assumptions (e.g., linearity, constant variance).
- Types of plots generated by regression software:
  - Residuals vs. fitted values
  - Normal probability plots
  - Residuals over time/order

Interpreting Residual Plots

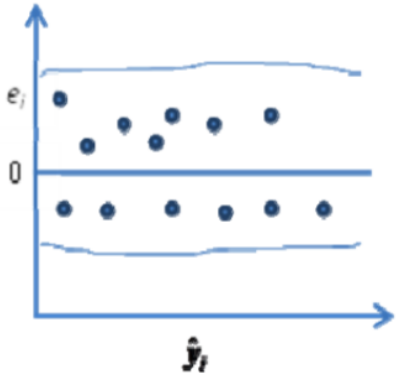
- **Random Pattern:** Indicates a good model fit.
- **Systematic Pattern:** Suggests model inadequacy (e.g., nonlinear relationships).
- **Fan-shaped Patterns:** Indicate non-constant variance.
- **Outliers:** Can be detected through studentized residuals.

Residual Plot Interpretations

- A plot of residuals e_i or scaled residuals (d_i, r_i) versus fitted values $\hat{y}_i$ is helpful in diagnosing model issues.
- Different patterns provide different insights into the regression model?s adequacy.

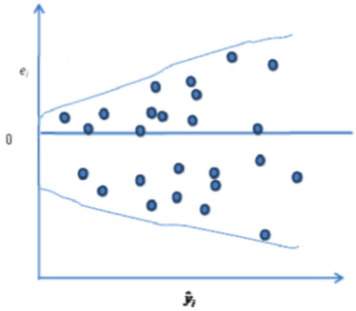
Horizontal Band Pattern

- If residuals fluctuate randomly within a horizontal band, there are no visible model defects.
- This indicates that the model assumptions are reasonably satisfied.



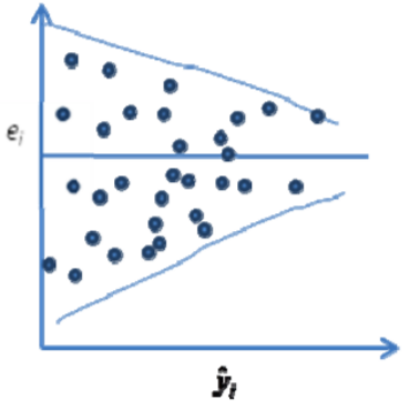
Outward Opening Funnel Pattern

- If residuals form an outward opening funnel, it suggests that error variance increases with $\hat{y}$.
- This indicates heteroscedasticity, meaning the variance of errors is not constant.



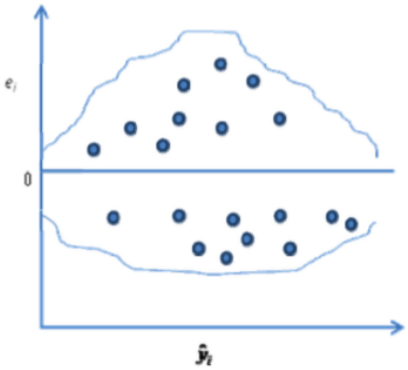
Inward Opening Funnel Pattern

- If residuals form an inward opening funnel, the variance of errors decreases with $\hat{y}$.
- This also indicates heteroscedasticity, but in the opposite manner of the outward funnel.



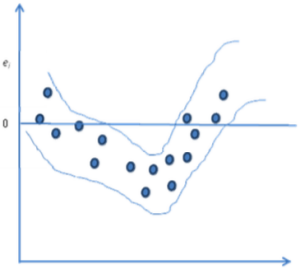
Double Bow Pattern

- If residuals form a double bow, y may be a proportion between 0 and 1.
- This suggests a nonlinear relationship between y and the explanatory variables.



Curved Pattern

- If residuals exhibit a curved pattern, it suggests that the assumed relationship between y and the predictors is nonlinear.
- Possible solutions:
  - Include additional explanatory variables.
  - Use polynomial terms (e.g., squared terms).
  - Apply transformations to explanatory or response variables.



Handling Variance Inequality

- Common methods to deal with non-constant variance:
  - Transformations (log, square root, Box-Cox transformation).
  - Weighted least squares regression.

Outliers and Influential Points

- Residual plots can reveal unusually large residuals.
- Points with large residuals at extreme $\hat{y}_i$ values might indicate:
  - Non-constant variance.
  - Nonlinear relationship between y and explanatory variables.
  - Potential outliers that should be investigated further.

Conclusion: Residual plots

- Residual plots provide essential insights into model adequacy.
- Different patterns in residual plots indicate different problems in the model.
- Addressing these issues improves the reliability of regression models.

Introduction to Homoskedasticity and Heteroskedasticity

- In regression analysis, one of the fundamental assumptions is that the variance of the error terms remains constant across all observations.
- When this assumption holds, the error terms are said to exhibit **homoskedasticity**.
- However, in many real-world scenarios, the variance of the error terms varies across observations, leading to **heteroskedasticity**.

Mathematical Representation of Homoskedasticity

Homoskedasticity: The covariance matrix of the disturbance terms is given by:

$$\text{Var}(\varepsilon) = \sigma^2 I_n$$

where:

- σ^2 is the constant variance of the errors.
- I_n is the identity matrix of size $n \times n$.
- The diagonal elements are equal to σ^2, and the off-diagonal elements are zero, indicating that errors are uncorrelated.

Mathematical Representation of Heteroskedasticity

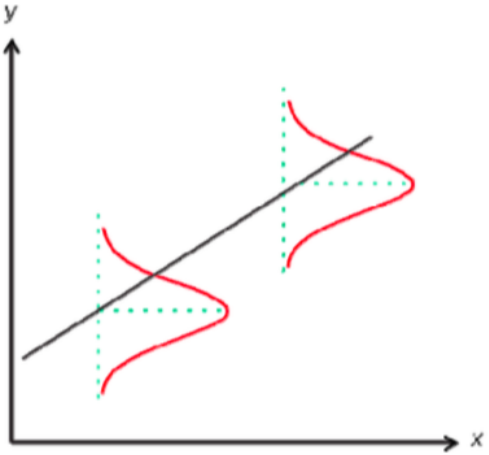
Heteroskedasticity: When the variance of the errors is not constant across observations, the covariance matrix takes the form:

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

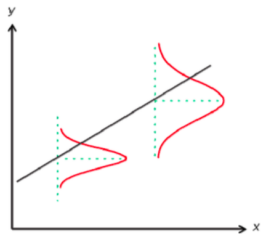
Here:

- The diagonal elements represent the variance of each observation, which varies across data points.
- The off-diagonal elements remain zero, implying that errors are uncorrelated.
- The presence of different variances indicates heteroskedasticity in the data.

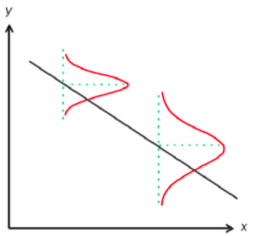
Graphical Representation of Homoskedasticity and Heteroskedasticity



Homoskedasticity



Heteroskedasticity ($Var(y)$ increases with x)



Heteroskedasticity ($Var(y)$ decreases with x)

Heteroskedasticity: Variance of residuals changes with independent variable values.

Real-World Examples of Heteroskedasticity

- Consider a regression model where x represents income and y represents expenditure on food. As income increases, food expenditure varies due to increased choices, leading to changing variance in y.
- Suppose x represents the number of hours practiced for typing and y represents the number of errors per page. More practice generally reduces errors, but the variance in errors is likely to change as well, violating the homoskedasticity assumption.

Why Heteroskedasticity Matters?

- When heteroskedasticity is present, standard errors of regression coefficients may be biased, leading to incorrect statistical inferences.
- It violates one of the key assumptions of the classical linear regression model, potentially affecting hypothesis testing and confidence intervals.
- Identifying and correcting heteroskedasticity is crucial for reliable regression analysis.

Autocorrelation in Regression Models

- In linear regression, one fundamental assumption is that the random error components (disturbances) are **identically and independently distributed**.
- The model is given by:

$$y = X\beta + u, \tag{2}$$

where $E(u_t, u_{t-s})$ is assumed to be:

$$E(u_t, u_{t-s}) = \begin{cases} \sigma_u^2, & \text{if } s = 0, \\ 0, & \text{if } s \neq 0. \end{cases} \tag{3}$$

Violation of Assumptions

- If $E(u_t, u_t) \neq \sigma_u^2$, for $s = 0$ then the variance of the disturbance term is not constant, leading to **heteroskedasticity**.
- If $E(u_t, u_{t-s}) \neq 0$ for $s \neq 0$, the disturbance terms are correlated, leading to **autocorrelation**.

Understanding Autocorrelation

- When autocorrelation is present, some or all off-diagonal elements in $E(uu')$ are nonzero.
- Sometimes the study and explanatory variables have a natural sequence order over time, i.e., the data is collected with respect to time. Such data is termed as time-series data. The disturbance terms in time series data are serially correlated.

Understanding Autocorrelation

- The **autocovariance** at lag s is defined as:

$$\gamma_s = E(u_t, u_{t-s}), \quad s = 0, \pm 1, \pm 2, ... \tag{4}$$

- At lag $s = 0$, this reduces to the variance:

$$\gamma_0 = E(u_t^2) = \sigma^2. \tag{5}$$

Autocorrelation Coefficient

- The **autocorrelation coefficient** at lag s is given by:

$$\rho_s = \frac{E(u_t u_{t-s})}{\sqrt{Var(u_t)Var(u_{t-s})}} = \frac{\gamma_s}{\gamma_0}, \quad s = 0, \pm 1, \pm 2, ... \tag{6}$$

- Assuming stationarity, ρ_s and γ_s depend only on the lag s, not on time t.

Orders of Autocorrelation

- The **first-order autocorrelation** measures correlation between successive terms:

$$\rho_1 = Corr(u_2, u_1), Corr(u_3, u_2), ..., Corr(u_n, u_{n-1}). \tag{7}$$

- The **second-order autocorrelation** measures correlation between every second term:

$$\rho_2 = Corr(u_3, u_1), Corr(u_4, u_2), ..., Corr(u_n, u_{n-2}). \tag{8}$$

Sources of Autocorrelation

1. **Carryover Effect:** The expenditure in one month may influence expenditure in the next month, causing autocorrelation in time-series data.
2. **Omitted Variables:** If an important variable is omitted from a regression model, the impact of that variable may be absorbed into the error term, introducing autocorrelation.
3. **Misspecification:** If a non-linear relationship is wrongly modeled as linear, it can lead to autocorrelation. For example, an exponential trend should not be modeled as a simple linear equation.
4. **Measurement Errors:** Errors in measuring the dependent variable can create a pattern in residuals, leading to autocorrelation.

Conclusion: Autocorrelation

- Autocorrelation in regression models affects the efficiency of estimators and can lead to misleading statistical inferences.
- Understanding autocorrelation is crucial for analyzing time-series data and applying appropriate corrections.

Introduction to Polynomial Regression Models

- A model is said to be linear when it is linear in parameters.
- Polynomial regression models are an extension of linear regression models, allowing for curvilinear relationships.
- These models are particularly useful in cases where the relationship between the response variable and the explanatory variables is nonlinear but can be approximated using polynomial terms.

Mathematical Representation of Polynomial Models

Example of a second-order polynomial model in one variable:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \tag{9}$$

where:

- y is the dependent variable,
- x is the independent variable,
- $\beta_0, \beta_1, \beta_2$ are the model parameters,
- ϵ represents the error term.

Polynomial Models in Multiple Variables

Example of a second-order polynomial model in two variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon. \tag{10}$$

Explanation:

- This model includes quadratic terms x_1^2 and x_2^2 to account for curvature in the relationship.
- The interaction term $x_1 x_2$ allows for the combined effect of both explanatory variables on the dependent variable.

General Form of a Polynomial Model in One Variable

The k-th order polynomial model in one variable is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \epsilon. \tag{11}$$

Interpretation:

- This model captures increasing levels of curvature by including higher-degree polynomial terms.
- The choice of the polynomial degree k depends on the complexity of the underlying relationship between y and x.

Polynomial Regression as a Special Case of Multiple Linear Regression

If $x_j = x^j$ for $j = 1, 2, ..., k$, then the polynomial model can be rewritten as a multiple linear regression model:

$$y = X\beta + \epsilon, \tag{12}$$

where:

- X is the design matrix containing polynomial terms of x,
- β is the vector of coefficients,
- ϵ is the error term.

This means standard techniques used for fitting linear regression models can also be applied to fit polynomial regression models.

Example: Quadratic Model in One Variable

Definition: A second-order polynomial regression model, also known as a quadratic model, is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon. \tag{13}$$

Explanation of Parameters:

- β_1 is called the **linear effect parameter**, which determines the rate of change of y with respect to x.
- β_2 is called the **quadratic effect parameter**, which introduces curvature to the model.

Applications:

- Used in economics, engineering, and sciences where relationships exhibit nonlinearity.
- Helps in capturing diminishing or increasing returns.

Considerations in Fitting a Polynomial in One Variable

Some important considerations when fitting a polynomial model:

1. **Order of the Model**
2. **Model Building Strategy**
3. **Extrapolation**
4. **Ill-conditioning**
5. **Hierarchy**

Order of the Model

- The order of the polynomial should be kept as low as possible.
- Transformations can be used to maintain a first-order model where possible.
- If necessary, a second-order polynomial is considered.
- Arbitrarily fitting higher-order polynomials can be a misuse of regression analysis.
- The model should align with the knowledge of data and its environment.

Model Building Strategy

Choosing the Order of a Polynomial Model

- A structured approach is essential for determining the order of the model.
- **Forward Selection:**
  - Start with the lowest order model.
  - Increase the order step by step.
  - Check the significance of regression coefficients using the t-test.
  - Stop when the highest order term is non-significant.
- **Backward Elimination:**
  - Start with a high-order polynomial.
  - Remove terms one at a time, starting with the highest order.
  - Continue until the highest remaining term has a significant t-statistic.
- Forward selection and backward elimination may result in different models.
- In practice, first and second-order polynomials are most commonly used.

Extrapolation

- Extrapolation using polynomial models requires caution.
- The curvature in the observed data region may differ from the extrapolated region.
- Example: A trend increasing in the data region may decrease in the extrapolated region.
- Polynomial models can exhibit unexpected turns, leading to incorrect inferences.
- This applies to both interpolation and extrapolation.

Ill-conditioning

- A key assumption in linear regression is that the X-matrix has full column rank.
- In polynomial regression, as the order increases, the $X'X$ matrix becomes ill-conditioned.
- This leads to numerical instability in estimating $(X'X)^{-1}$.
- High ill-conditioning results in large estimation errors.
- If x values lie within a narrow range, multicollinearity increases.
- Example: If x varies between 2 and 3, then x^2 varies between 4 and 9, introducing strong multicollinearity.

Hierarchy in Polynomial Models

- A model is hierarchical if it includes all lower-order terms.
- Example of a hierarchical model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$$

- Example of a non-hierarchical model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_4 x^4 + \epsilon$$

- The second model is not hierarchical as it lacks the x^3 term.
- Hierarchical models are preferable because they are invariant under linear transformations.

Special Cases in Model Hierarchy

- Some models require specific terms for interpretation.
- Example: Interaction models

$$y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \epsilon$$

- A strict hierarchical model would require the inclusion of x_1^2, which may not always be necessary.
- The choice of hierarchy depends on the statistical significance of terms and the research context.

Polynomial Regression Model

General Form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Matrix Representation:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y}$: Response vector $(n \times 1)$
- X: Design matrix $(n \times (k+1))$
- $\boldsymbol{\beta}$: Parameter vector $((k+1) \times 1)$
- $\boldsymbol{\epsilon}$: Error term $(n \times 1)$

Challenges with Traditional Polynomials

- Multicollinearity: Columns of X (i.e., $1, x, x^2, \ldots, x^k$) are not orthogonal.
- Adding a new term $\beta_{k+1}x^{k+1}$ affects all lower-order coefficients.
- Computationally expensive to recompute $(X'X)^{-1}$.

Orthogonal Polynomial Regression

New Model Representation:

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \cdots + \alpha_k P_k(x_i) + \epsilon_i$$

where $P_u(x_i)$ is the u^{th} order orthogonal polynomials defined as:

$$\sum_{i=1}^{n} P_r(x_i) P_s(x_i) = 0, \quad \text{for } r \neq s,\ s, r = 0, 1, 2, \ldots, k,\ P_0(x_i) = 1$$

Orthogonality Condition

New Design Matrix:

$$X = \begin{bmatrix} P_0(x_1) & P_1(x_1) & \dots & P_k(x_1) \\ P_0(x_2) & P_1(x_2) & \dots & P_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(x_n) & P_1(x_n) & \dots & P_k(x_n) \end{bmatrix}$$

Orthogonality ensures: (Since this X matrix has orthogonal columns, so $X'X$ matrix becomes)

$$X'X = \begin{bmatrix} \sum P_1^2(x_i) & 0 & \dots & 0 \\ 0 & \sum P_2^2(x_i) & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sum P_k^2(x_i) \end{bmatrix}.$$

Least Squares Estimator

OLS Estimator: The ordinary least squares estimator is $\hat{\alpha} = (X'X)^{-1}X'y$ which for α_j is

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{n} P_j(x_i)y_i}{\sum_{i=1}^{n} P_j^2(x_i)}, \quad j = 0, 1, 2, \ldots, k.$$

Variance of the Estimator: $V(\hat{\alpha}) = \sigma^2(X'X)^{-1}$ as

$$\text{Var}(\hat{\alpha}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} P_j^2(x_i)}.$$

If σ^2 is unknown, it is estimated using ANOVA.

Advantages of Orthogonal Polynomials

- Avoids multicollinearity in polynomial regression.
- Adding higher-degree terms does not affect lower-order coefficients.
- Efficient computation due to diagonal $X'X$ matrix.

Conclusion: Orthogonal Polynomials

- Orthogonal polynomials simplify polynomial regression models.
- They provide more stable parameter estimates.
- Useful in practical applications where multicollinearity is a concern.

Introduction to Transformations on Y or X

- Transformations are applied to response (Y) or predictor (X) variables.
- Used to stabilize variance, linearize relationships, and meet model assumptions.

Transformations on Y (Response Variable)

Common Transformations:

- **Log Transformation**: $Y^* = \log(Y)$ (stabilizes variance, normalizes data)
- **Square Root Transformation**: $Y^* = \sqrt{Y}$ (for count data)
- **Inverse Transformation**: $Y^* = 1/Y$ (reduces right skew)
- **Box-Cox Transformation**: $Y^* = \frac{Y^\lambda - 1}{\lambda}$ (optimal λ chosen)

Transformations on X (Predictor Variable)

Common Transformations:

- **Log Transformation**: $X^* = \log(X)$ (handles exponential growth)
- **Polynomial Transformation**: $X^* = X^2, X^3$ (captures non-linearity)
- **Reciprocal Transformation**: $X^* = 1/X$ (used when Y decreases rapidly with X)
- **Standardization**: $X^* = \frac{X - \bar{X}}{\sigma_X}$ (for better interpretability)
- **Orthogonal Polynomials**: Used to avoid correlation among polynomial terms

Example: Polynomial Transformation

Consider fitting a quadratic model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \tag{14}$$

If the columns of X are not orthogonal, then:

- The parameters $\beta_0, \beta_1, \beta_2$ change if a higher-order term is added.
- Using orthogonal polynomials ensures that X'X is diagonal.

Conclusion: Transformations on Y or X

- Transformations improve model assumptions and performance.
- Choose transformations based on residual analysis and model diagnostics.

Introduction to Inverse Regression

- Inverse regression models the predictor variable X as a function of the response variable Y.
- Useful when X is subject to measurement errors.
- Commonly applied in calibration problems and measurement error models.

Standard Regression vs. Inverse Regression

Classical Regression Model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{15}$$

Inverse Regression Model:

$$X = \alpha_0 + \alpha_1 Y + \eta \tag{16}$$

- Classical regression predicts Y given X.
- Inverse regression estimates X based on Y.

Estimation of Parameters

Given n observations (X_i, Y_i), we minimize:

$$S = \sum_{i=1}^{n} (X_i - \alpha_0 - \alpha_1 Y_i)^2 \tag{17}$$

Estimates:

$$\alpha_1 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (Y_i - \bar{Y})^2}, \quad \alpha_0 = \bar{X} - \alpha_1 \bar{Y} \tag{18}$$

Relation Between Regression Models

$$\alpha_1 = \frac{1}{\beta_1} \quad (\text{if } \beta_1 \neq 0) \tag{19}$$

- If standard regression is known, inverse regression coefficients can be derived.
- Assumes normally distributed data with constant variance.

Example: Chemical Calibration

Problem: Determine the concentration of a chemical solution X from its absorbance Y.

$$X = \alpha_0 + \alpha_1 Y + \eta \tag{20}$$

Solution: Given observed Y-values, use the estimated model to predict X.

Conclusion: Inverse Regression

- Inverse regression is useful when X is difficult to measure directly.
- Common in calibration studies and measurement error models.
- Provides an alternative perspective in regression analysis.

References

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis, Fifth Edition. Wiley.
- MTH 416 : Regression Analysis — Shalabh, IIT Kanpur

Thank You!

Thank you for your attention!