PMDS504L: Regression Analysis and Predictive Models

Introduction to Simple Linear Regression

Dr. Jisha Francis

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

Linear Models and Regression Analysis

Suppose the outcome of any process is denoted by a random variable Y , called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by $X_1, X_2, \dots, X_k$.

- In statistical modeling, the **dependent variable** (Y) is the variable to be predicted or explained.
- The **independent variables** $(X_1, X_2, \dots, X_k)$ influence Y.

Linear Models and Regression Analysis

- Suppose the behaviour of Y can be explained by a relationship given:

$$Y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \epsilon$$

- f: A well-defined function that describes how $X_1, X_2, \ldots, X_k$ interact with parameters $\beta_1, \beta_2, \ldots, \beta_k$.
- ϵ: Stochastic (random) term that accounts for variability not explained by the model.

Mathematical vs Statistical Models

- If $\epsilon = 0$: The relationship is **exact**, and the model is called a **mathematical model**.
- If $\epsilon \neq 0$: The relationship is **approximate**, and the model is called a **statistical model**.

The term "model" is broadly used to represent any phenomenon in a mathematical framework.

Linearity in Parameters

- A model is **linear** if it is linear in the **parameters** $(\beta_1, \beta_2, \ldots, \beta_k)$.
- **Condition for Linearity:**

$$\frac{\partial Y}{\partial \beta_i} \text{ is independent of } \beta_i, \quad \forall i$$

- If any partial derivative depends on β_i, the model is **nonlinear**.

Linearity vs Nonlinearity of Models

Key Distinction:

- Linearity or nonlinearity depends on the **parameters**, not the **explanatory variables**.

Examples:

- **Linear Model:**

$$Y = \beta_1 X_1 + \beta_2 X_2^2$$

  Linear in parameters (β_1, β_2) despite X_2^2 being nonlinear.

- **Nonlinear Model:**

$$Y = \beta_1 e^{\beta_2 X_2}$$

  Nonlinear because $\frac{\partial y}{\partial \beta_2}$ depends on β_2.

Exercise: Identifying Linear and Nonlinear Models

Question: Consider the following models. Determine whether they are linear or nonlinear. Justify your answers.

1. $$Y = \beta_1 X^2 + \beta_2 X + \beta_3 \log X + \epsilon$$

2. $$Y = \beta_1^2 X + \beta_2 X + \beta_3 \log X + \epsilon$$

Linear and Nonlinear Models

Linear Model:

- If the function f is linear in parameters, then the model

$$Y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \epsilon$$

  is called a *linear model*.

Nonlinear Model:

- If the function f is nonlinear in parameters, then the model is called a *nonlinear model*.

Key Points: Linear and Nonlinear Models

In general, the function f for a linear model is chosen as:

$$f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Key Points:

- $X_1, X_2, \ldots, X_k$ are pre-determined variables.
- Y is the observed outcome, hence known.
- The knowledge of the model depends on the parameters $\beta_1, \beta_2, \ldots, \beta_k$.

Regression Analysis

Regression Analysis:

- Regression analysis is a technique that helps in determining the statistical model by using the data on study (dependent) and explanatory (independent) variables.
- The classification of regression analysis into:
  - **Linear Regression:** Based on the determination of linear models.
  - **Nonlinear Regression:** Based on the determination of nonlinear models.

Understanding Regression: An Example

Example: Regression Analysis

- Suppose the yield of a crop (Y) depends linearly on two explanatory variables:
    - Quality of fertilizer (X_1).
    - Level of irrigation (X_2).
- The relationship is given by:

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

- The true values of β_1 and β_2 exist but are unknown to the experimenter.
- Data on Y, X_1, and X_2 is collected through experiments.

Regression Analysis Approach

- The collected data reflects the systematic relationship between Y, X_1, and X_2.
- The experimenter uses the data to estimate the parameters β_1 and β_2.
- This process of moving from observed data to model estimation is termed **regression analysis**.
- The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

Steps in Regression Analysis

Regression analysis involves the following steps:

1. **Statement of the problem:** Clearly define the problem under consideration.
2. **Choice of relevant variables:** Identify variables that are significant to the problem.
3. **Collection of data:** Gather data on the relevant variables.
4. **Specification of model:** Define the mathematical or statistical model to represent the relationship.
5. **Choice of fitting method:** Select an appropriate method to fit the data (e.g., least squares).
6. **Fitting the model:** Apply the chosen method to estimate the model parameters.
7. **Model validation and criticism:** Assess the model's accuracy and suitability.
8. **Using the model:** Apply the model to solve the stated problem.

Types of Regression Analysis (Part 1)

Type of Regression	Conditions
Univariate	Only one quantitative response variable.
Multivariate	Two or more quantitative response variables.
Simple	Only one explanatory variable.
Multiple	Two or more explanatory variables.
Linear	All parameters enter the equation linearly, possibly after transformation of the data.

Table: Regression Methodologies (Part 1)

Types of Regression Analysis (Part 2)

Type of Regression	**Conditions**
Nonlinear	The relationship between the response and some explanatory variables is nonlinear or some parameters appear nonlinearly, and no transformation is possible to make the parameters appear linearly.
Analysis of Variance	All explanatory variables are qualitative variables.
Analysis of Covariance	Some explanatory variables are quantitative, while others are qualitative.
Logistic	The response variable is qualitative.

Objectives of Regression Analysis

- **Determine the Explicit Form:** The primary goal is to identify the explicit form of the regression equation, representing a valid relationship between the study variable and explanatory variables.
- **Policy Formulation:** Use the regression equation to determine the role of explanatory variables in joint relationships to aid in policy-making.
- **Forecasting:** Predict the values of the response variable for a given set of explanatory variable values.
- **Understanding Interrelationships:** Analyze and interpret the interrelationships among variables.

Simple Linear Regression Model

Model Definition:

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

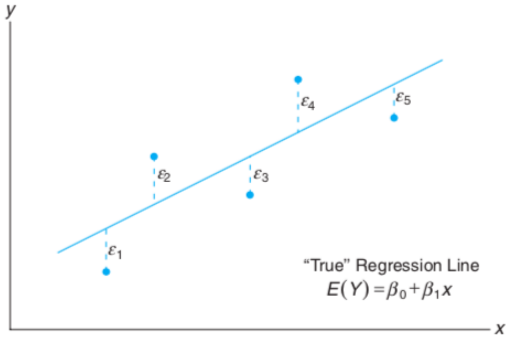
where:

- Y: Dependent (study) variable
- X: Independent (explanatory) variable
- β_0: Intercept term (regression coefficient)
- β_1: Slope parameter (regression coefficient)
- ϵ: Error term

Simple Linear Regression Model

- Y: Dependent (study) variable
  - Response Variable
  - Outcome Variable
  - Predicted Variable
  - Target Variable
  - Criterion Variable
- X: Independent (explanatory) variable
  - Predictor Variable
  - Explanatory Variable
  - Input Variable
  - Feature (commonly used in machine learning)
  - Regressor

Simple Linear Regression Model



Simple Linear Regression Model

Error Term (ϵ): The unobservable error component (ϵ) accounts for the failure of data to lie on a straight line and represents the difference between the true and observed realization of Y.

- The quantity ϵ, often called a random error or random disturbance,
- Represents differences between true and observed values of Y.
- Reasons for differences:
  - the effect of all deleted variables in the model
  - variables may be qualitative
  - Inherent randomness in observations

Assumptions of the Error Term and Model Properties

Assumptions of Error Term (ϵ):

- ϵ is an **independent and identically distributed (i.i.d.)** random variable.
- Mean of ϵ: $E(\epsilon) = 0$.
- Variance of ϵ (error variance): $\text{Var}(\epsilon) = \sigma^2$ (constant).
- Later assumption: $\epsilon \sim N(0, \sigma^2)$ (normal distribution).

Model Properties:

- Independent variable (X) is **non-stochastic**, controlled by the experimenter.
- Dependent variable (Y) is a random variable since ϵ is random.

$$E(Y) = \beta_0 + \beta_1 X, \quad \text{Var}(Y) = \sigma^2$$

Conditional Mean When X is Random

Scenario:

- In some cases, the independent variable (X) is a **random variable**.
- This changes the focus of analysis from overall properties to **conditional properties**.

Key Adjustment:

- Instead of using the **sample mean and variance** of Y, we consider:

$$E(Y|x) = \beta_0 + \beta_1 x$$

- This represents the **conditional mean** of Y given $X = x$.

Implication:

- The regression equation now describes the relationship of Y given specific values of X.
- Helps analyze the expected value of Y based on observed x.

Conditional Variance and Estimation in Regression

Conditional Variance:

- Given $X = x$, the **conditional variance** of Y is:

$$\text{Var}(Y|x) = \sigma^2$$

Complete Description of the Model:

- The model $Y = \beta_0 + \beta_1 X + \epsilon$ is fully defined when:

$$\beta_0, \beta_1, \text{ and } \sigma^2$$

  are known.
- In practice, these parameters are **unknown** and must be **estimated**.

Estimation of Parameters:

Least Squares Estimation

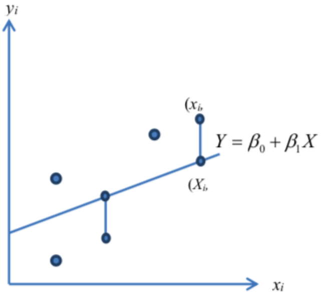
- Suppose a sample of n sets of paired observations (x_i, y_i) $(i = 1, 2, \ldots, n)$ is available.
- These observations satisfy the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

- The principle of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of the difference between the observations and the line in the scatter diagram.

Least Squares Estimation

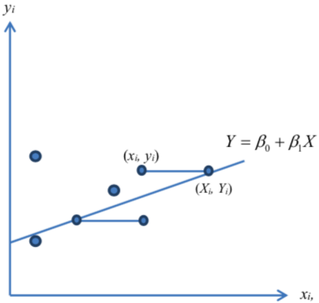
When the vertical difference between the observations and the line in the scatter diagram is considered, and its sum of squares is minimized, this method is known as *direct regression*.



Direct regression

Reverse Regression Method

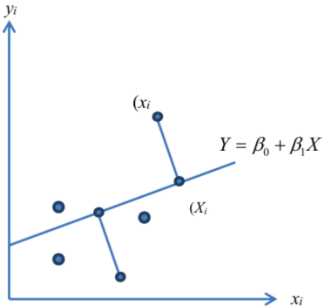
- Alternatively, the sum of squares of the differences between the observations and the line in the **horizontal direction** in the scatter diagram can be minimized.
- This approach yields estimates for β_0 and β_1 and is termed as the *reverse (or inverse) regression method*.



Reverse regression method

Orthogonal Regression Method

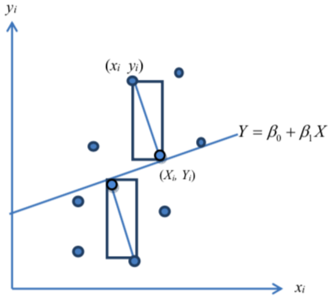
- Instead of horizontal or vertical errors, the **sum of squares of perpendicular distances** between the observations and the line in the scatter diagram is minimized.
- This approach provides estimates for β_0 and β_1 and is referred to as the *orthogonal regression* or *major axis regression* method.



Major axis regression method

Reduced Major Axis Regression Method

- Instead of minimizing distances, the **sum of the areas of rectangles** defined between the observed data points and the nearest point on the line is minimized.
- This method provides estimates of regression coefficients by minimizing these areas.



Reduced major axis method

Least Absolute Deviation Regression

- The method of **Least Absolute Deviation Regression** minimizes the **sum of the absolute deviations** of the observations from the regression line in the vertical direction.
- This approach is similar to direct regression but uses absolute deviations instead of squared deviations.
- Estimates of β_0 and β_1 are obtained by solving:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

Assumptions of Least Squares Estimates (1/3)

- The least squares estimates do not require any assumption about the probability distribution of ϵ_i.
- For statistical inferences, the following assumptions are made:

$$
\begin{aligned}
E(\epsilon_i) &= 0, \\
\mathrm{Var}(\epsilon_i) &= \sigma^2, \\
\mathrm{Cov}(\epsilon_i, \epsilon_j) &= 0 \quad \text{for all } i \neq j \ (i, j = 1, 2, \ldots, n).
\end{aligned}
$$

Properties of Least Squares Estimates (2/3)

- These assumptions are necessary to derive:
  - The mean and variance of the least squares estimates.
  - Other statistical properties of the estimates.
- The assumption of normality of ϵ_i is utilized for:
  - Constructing tests of hypotheses.
  - Building confidence intervals for the parameters.

Ordinary Least Squares Method (3/3)

- Different approaches yield different estimates for β_0 and β_1 with varying statistical properties.
- Among these, the **direct regression approach** is the most commonly used.
- The estimates obtained through this approach are called:
  - **Least Squares Estimates** or
  - **Ordinary Least Squares (OLS)** Estimates.

Direct Regression Method (1/3)

- Also known as the **ordinary least squares (OLS) estimation**.
- Assume a set of n paired observations $(x_i, y_i), i = 1, 2, \ldots, n$ satisfy the linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- For each observation, the model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (i = 1, 2, \ldots, n).$$

- The direct regression method minimizes the sum of squares:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Direct Regression Method (2/3)

- To minimize $S(\beta_0, \beta_1)$, take the partial derivatives:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i.$$

- Setting these derivatives to zero gives:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

Direct Regression Method (3/3)

- Solving these equations gives the OLS estimators of β_0 and β_1:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- Where:

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i (x_i - \bar{x}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Fitted Linear Regression Model and Residuals

- The fitted line or fitted linear regression model is:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- The predicted values of y are:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \text{for } i = 1, 2, \ldots, n.$$

- The difference between the observed value y_i and the fitted (or predicted) value $\hat{y}_i$ is called a residual . The i-th residual is defined as:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

Numerical Problem

SCUBA divers have specific maximum dive times that they must not exceed at various depths to ensure safety. The table below shows the relationship between the depth (in feet) and the corresponding maximum dive time (in minutes):

Depth (feet), X	50	60	70	80	90	100
Maximum Dive Time (minutes), Y	80	55	45	35	25	22

1. **Fit the Least Squares Regression Line:** Use the data to calculate the least squares regression equation of Y (maximum dive time) on X (depth).
2. **Predict:** Based on the regression equation, predict the maximum dive time for a depth of 110 feet.
3. **Plot the Data:** Create a scatter plot of the data points, draw the fitted regression line, and sketch the relationship between depth and maximum dive time.

Solution

To fit a simple linear regression line using the method of least squares, we calculate the coefficients β_0 and β_1 using the following formulas:

$$\hat{\beta}_1 = \frac{n\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{n\sum_i^n x_i^2 - (\sum_i^n x_i)^2}$$

$$\hat{\beta}_0 = \frac{\sum_i^n y_i - \hat{\beta}_1 \sum_i^n x_i}{n}$$

Data and Calculations for Regression Line

Given data:

x	y	xy	x^2
50	80	4000	2500
60	55	3300	3600
70	45	3150	4900
80	35	2800	6400
90	25	2250	8100
100	22	2200	35500
450	262	17700	408

Summary of Values and Regression Coefficients

$$S_{xx} = 1750$$
$$S_{xy} = -1950$$
$$\beta_1 = -1.114285714$$
$$\beta_0 = 127.2380952$$

The estimated regression line is:

$$Y \approx 127.2380952 - 1.114285714X$$

Prediction for $X = 110$

Substitute $X = 110$: The maximum dive time for 110 feet is

$$127.2380952 - 1.114285714 * 110 = 4.667s.$$

The predicted maximum dive time at 110 feet is 4.667 seconds.

Residuals Calculation for the Regression Model

The regression equation is: $\hat{Y} = 127.2380952 - 1.114285714X$

The residuals ($e = Y - \hat{Y}$) are calculated as follows:

X (Depth)	Y (Observed)	$\hat{Y}$ (Predicted)	Residual ($e = Y - \hat{Y}$)
50	80	$127.2381 - 1.1143 \times 50 = 71.5238$	$80 - 71.5238 = 8.4762$
60	55	$127.2381 - 1.1143 \times 60 = 60.3809$	$55 - 60.3809 = -5.3809$
70	45	$127.2381 - 1.1143 \times 70 = 49.2381$	$45 - 49.2381 = -4.2381$
80	35	$127.2381 - 1.1143 \times 80 = 38.0952$	$35 - 38.0952 = -3.0952$
90	25	$127.2381 - 1.1143 \times 90 = 26.9524$	$25 - 26.9524 = -1.9524$
100	22	$127.2381 - 1.1143 \times 100 = 15.8095$	$22 - 15.8095 = 6.1905$

Table: Residuals for Each Data Point

Summary of Residuals

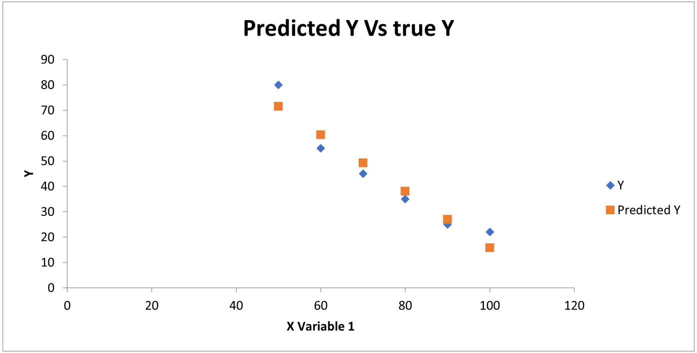
- Positive residuals indicate observed values are greater than predicted values.
- Negative residuals indicate observed values are less than predicted values.

Plot of Data and Regression Line

Steps to Plot:

- Scatter the observed data points.
- Plot the regression line $Y \approx 127.2380952 - 1.114285714X$.

Plot of Data and Regression Line



Reference

- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.
- https://home.iitk.ac.in/ shalab/course5.htm

Thank You!

Thank you for your attention!