Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

1 Introduction to Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable (often called the outcome or target variable) and one or more independent variables (often called predictors or features). The goal is to find the best-fitting straight line (regression line) that describes the relationship between these variables.

2 Key Components of Linear Regression

- **Dependent Variable (Y):** The outcome we are trying to predict or explain.
- **Independent Variable (X):** The predictor or feature we use to predict the dependent variable.
- **Regression Line:** A straight line that best fits the data points on a scatter plot. In simple linear regression, this line is represented by the equation:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

  where:

  - Y is the predicted value of the dependent variable.
  - β_0 is the y-intercept (the value of Y when $X = 0$).
  - β_1 is the slope of the line (how much Y changes for a one-unit change in X).
  - ϵ is the error term, representing the difference between the predicted and actual values.

3 Understanding the Line of Best Fit

The "line of best fit" minimizes the differences (residuals) between the observed values and the values predicted by the line. The method commonly used to find this line is called **Ordinary Least Squares (OLS)**, which minimizes the sum of the squared differences between the observed values and the predicted values.

4 Assumptions of Linear Regression

- **Linearity:** The relationship between the independent and dependent variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** The residuals (differences between observed and predicted values or Ypred - Y) have constant variance.
- **Normality:** The residuals are normally distributed.

5 Assumptions of Linear Regression Explanation

Linear regression relies on several key assumptions to ensure that the model provides valid and reliable results. These assumptions are as follows:

5.1 Linearity

Explanation: The assumption of linearity means that there is a straight-line relationship between the independent variable(s) and the dependent variable. This implies that the effect of the predictor(s) on the outcome is constant across all values of the predictor(s).

Why It's Important: If the relationship between the variables is not linear, the predictions and inferences made by the model will be incorrect. The linear regression model will not fit the data well if this assumption is violated.

5.2 Independence

Explanation: The independence assumption states that the residuals (errors) are independent of each other. This means that the error for one observation should not predict or influence the error for another observation.

Why It's Important: When observations are not independent (e.g., in time series data where past values influence future values), the model's estimations of the coefficients and their standard errors can be biased, leading to incorrect conclusions.

5.3 Homoscedasticity

Explanation: Homoscedasticity means that the residuals have constant variance across all levels of the independent variable(s). In other words, the spread (or "scatter") of the residuals is consistent for all values of the independent variable(s).

Why It's Important: If the residuals have non-constant variance (a condition known as heteroscedasticity), the model's estimates of the coefficients may still be unbiased, but the standard errors could be incorrect, leading to unreliable hypothesis tests and confidence intervals.

5.4 Normality

Explanation: The normality assumption states that the residuals (errors) are normally distributed. This does not mean that the independent and dependent variables themselves need to be normally distributed, but rather that the errors of the model's predictions follow a normal distribution.

Why It's Important: This assumption is particularly important for constructing confidence intervals and performing hypothesis tests. If the residuals are not normally distributed, the results of these tests may not be valid, especially in small sample sizes.

5.5 Why These Assumptions Matter

These assumptions are fundamental for linear regression because they ensure that the model is the right tool for analyzing the data and making predictions. If any of these assumptions are violated, the model's predictions and statistical inferences may be invalid.

When teaching linear regression, it is important to communicate to students the importance of checking these assumptions when applying linear regression to real-world data. Emphasize that these assumptions are the foundation for why and how the model works. Without them, the linear regression model might not give accurate or meaningful results, which is why checking for assumption violations is a critical step in the modeling process.

6 Evaluating the Model

- **Coefficient of Determination (R^2):** Measures the proportion of variation in the dependent variable that can be explained by the independent variable(s).
- **Residual Analysis:** Examining the residuals to check for patterns that might indicate a violation of the regression assumptions.

7 Applications of Linear Regression

Linear regression is widely used in fields such as economics, biology, engineering, and social sciences to model relationships and make predictions. Examples include predicting a student's test score based on study hours, estimating a company's revenue based on advertising spending, and forecasting housing prices.

Problem Statement

A company wants to understand the relationship between the number of hours their sales team works and the total sales they generate in thousands of dollars. The data collected for 5 employees is as follows:

Hours Worked (X)	Sales ($1000) (Y)
2	4
4	5
6	7
8	10
10	15

Using this data, perform a linear regression analysis to find the relationship between the hours worked and the sales generated. Predict the sales if an employee works for 7 hours.

Solution

Step 1: Calculate the Means of X and Y

$$\bar{X} = \frac{\sum X}{n} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{4+5+7+10+15}{5} = \frac{41}{5} = 8.2$$

Step 2: Calculate the Slope (β_1) and Intercept (β_0)

The formulas for the slope (β_1) and intercept (β_0) of the regression line $Y = \beta_0 + \beta_1 X$ are:

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Calculations:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (2-6)(4-8.2) + (4-6)(5-8.2) + (6-6)(7-8.2) + (8-6)(10-8.2) + (10-6)(15-8.2)$$

$$= (-4)(-4.2) + (-2)(-3.2) + (0)(-1.2) + (2)(1.8) + (4)(6.8)$$

$$= 16.8 + 6.4 + 0 + 3.6 + 27.2 = 54.0$$

$$\sum (X_i - \bar{X})^2 = (2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2$$

$$= (-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2$$

$$= 16 + 4 + 0 + 4 + 16 = 40$$

Now, calculate β_1:

$$\beta_1 = \frac{54.0}{40} = 1.35$$

Calculate β_0:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 8.2 - (1.35)(6) = 8.2 - 8.1 = 0.1$$

Step 3: Form the Regression Equation

The regression line equation is:

$$Y = 0.1 + 1.35X$$

Step 4: Predict Sales for 7 Hours of Work

To predict the sales when an employee works 7 hours:

$$Y = 0.1 + 1.35(7)$$

$$Y = 0.1 + 9.45 = 9.55$$

So, if an employee works for 7 hours, the predicted sales are \$9,550 (since sales are in thousands of dollars).

Problem 1

A study is conducted to understand the relationship between the number of hours students study and their final exam scores. The data collected for 6 students is as follows:

Hours Studied (X)	**Exam Score (Y)**
1	50
2	55
3	65
4	70
5	75
6	85

Perform a linear regression analysis to find the relationship between hours studied and exam scores. Predict the exam score if a student studies for 7 hours.

Solution:

Means: $\bar{X} = 3.5$, $\bar{Y} = 66.7$
Slope: $\beta_1 = 7.5$
Intercept: $\beta_0 = 40.8$
Regression Equation: $Y = 40.8 + 7.5X$
Prediction for 7 hours: $Y = 40.8 + 7.5(7) = 93.3$

Problem 2

A company wants to examine the effect of advertising expenditure on their product sales. The data for 5 different months is provided below:

Advertising Expenditure (\$1000) (X)	**Sales (\$1000) (Y)**
10	25
15	35
20	45
25	55
30	65

Find the regression line that predicts sales based on advertising expenditure. Predict the sales when the advertising expenditure is $35,000.

Solution:

Means: $\bar{X} = 20$, $\bar{Y} = 45$

Slope: $\beta_1 = 2$

Intercept: $\beta_0 = 5$

Regression Equation: $Y = 5 + 2X$

Prediction for $35,000: $Y = 5 + 2(35) = 75$

Problem 3

Researchers are studying the relationship between the number of years of education and annual income. The data for 5 individuals is shown below:

Years of Education (X)	Annual Income ($1000) (Y)
10	30
12	35
14	40
16	50
18	60

Perform a linear regression to find the relationship between years of education and income. Estimate the income for someone with 20 years of education.

Solution:

Means: $\bar{X} = 14$, $\bar{Y} = 43$

Slope: $\beta_1 = 3.75$

Intercept: $\beta_0 = -7.5$

Regression Equation: $Y = -7.5 + 3.75X$

Prediction for 20 years: $Y = -7.5 + 3.75(20) = 67.5$

Problem 4

A business analyst is examining the relationship between the number of products sold and the profit made by the company. The data for 5 months is given below:

Products Sold (X)	Profit ($1000) (Y)
50	200
60	220
70	250
80	270
90	300

Determine the regression equation to predict profit based on the number of products sold. What would be the expected profit if 100 products are sold?

Solution:

Means: $\bar{X} = 70$, $\bar{Y} = 248$

Slope: $\beta_1 = 2$

Intercept: $\beta_0 = 108$

Regression Equation: $Y = 108 + 2X$

Prediction for 100 products: $Y = 108 + 2(100) = 308$

Problem 5

A researcher is investigating the relationship between temperature (in degrees Celsius) and the energy consumption (in kilowatt-hours) of a household. The data for 5 days is given below:

Temperature (X)	Energy Consumption (Y)
15	400
20	350
25	300
30	250
35	200

Find the linear regression equation for predicting energy consumption based on temperature. Predict the energy consumption when the temperature is 40 degrees Celsius.

Solution:

Means: $\bar{X} = 25$, $\bar{Y} = 300$

Slope: $\beta_1 = -10$

Intercept: $\beta_0 = 550$

Regression Equation: $Y = 550 - 10X$

Prediction for 40 degrees: $Y = 550 - 10(40) = 150$