## 1. What is Linear Regression?

Linear regression is a statistical method used to model and analyze the relationships between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship (straight line) that predicts the dependent variable based on the values of the independent variables.

**Key Components:**

1. **Dependent Variable (Y)**: The outcome or the variable we are trying to predict or explain.
2. **Independent Variable (X)**: The variable(s) used to predict the dependent variable.
3. **Equation**: The relationship is typically expressed as:

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

   - $\beta_o$ is the intercept (the value of Y when X is 0).
   - $\beta_1$ is the slope (the change in Y for a one-unit change in X).
   - ϵ is the error term (the difference between the observed and predicted values of Y).

**Example of Simple Linear Regression**

Let's say you want to predict a person's weight (Y) based on their height (X).

**Step-by-Step Example:**

1. **Collect Data**: Suppose you have data for 5 individuals:

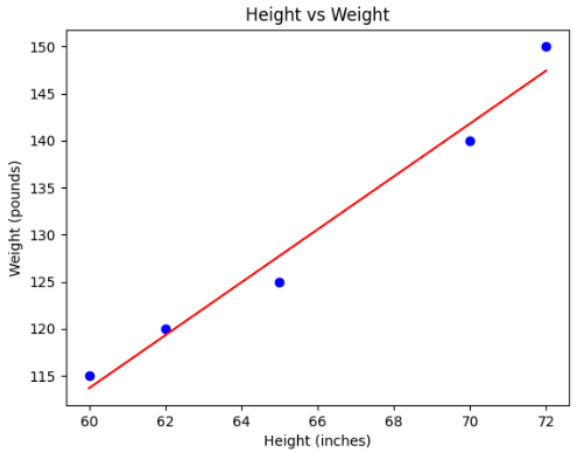
| Height (inches) | Weight (pounds) |
|---|---|
| 60 | 115 |
| 62 | 120 |
| 65 | 125 |
| 70 | 140 |
| 72 | 150 |

2. **Visualize the Data**: You can plot this data on a scatter plot with height on the X-axis and weight on the Y-axis.
3. **Fit a Linear Model**: Use the method of least squares to find the best-fitting line:

$$Weight = \beta_o + \beta_1 \times Height$$

4. **Calculate the Coefficients**: Using statistical software or manual calculations, determine the values of $\beta_o$ (intercept) and $\beta_1$(slope).

5. **Example in Python**:




6. **Interpret the Results**:

- **Intercept ($\beta_o$)**: The estimated weight when the height is 0 (not always meaningful in a real-world context).
- **Slope ($\beta_1$)**: The estimated change in weight for each inch increase in height.

**Example Output and Interpretation**

- Suppose the output is:

$$Weight = 50 + 1.5 \times Height$$

  - **Intercept**: 50 (not interpretable in this context since a height of 0 inches is unrealistic).
  - **Slope**: 1.5, meaning for every additional inch in height, the weight increases by 1.5 pounds.

## 2. WHEN SHOULD WE USE LINEAR REGRESSION?

Linear regression is a versatile statistical tool used in various fields for different purposes. Here are some scenarios and conditions under which linear regression is appropriate:

**When to Use Linear Regression**

1. **Understanding Relationships**:
   - To explore and quantify the relationship between two or more variables. For example, understanding how weight changes with height.
2. **Prediction**:
   - To make predictions based on the relationship between variables. For example, predicting future sales based on advertising spend.
3. **Trend Analysis**:
   - To analyze trends over time or other continuous variables. For example, studying the trend of temperature changes over the years.

**Conditions for Using Linear Regression**

1. **Linearity**:
    - The relationship between the independent and dependent variables should be linear. This means a straight-line relationship should reasonably approximate the data.
2. **Independence**:
    - The observations should be independent of each other. For example, the weight of one person does not affect the weight of another in a sample.
3. **Homoscedasticity**:
    - The variance of the errors should be constant across all levels of the independent variable. This means the spread of residuals (errors) should be consistent across the range of the independent variable.
4. **Normality of Errors**:
    - The residuals (errors) should be approximately normally distributed. This is particularly important for hypothesis testing and constructing confidence intervals.
5. **Minimal Multicollinearity**:
    - In multiple linear regression, the independent variables should not be too highly correlated with each other. High multicollinearity can make it difficult to interpret the coefficients.
6. **Adequate Sample Size**:
    - A larger sample size provides more reliable estimates and more power for statistical tests.

**Practical Examples**

1. **Economics**:
    - Analyzing how changes in interest rates affect consumer spending.
2. **Healthcare**:
    - Predicting patient outcomes based on treatment plans and patient characteristics.
3. **Marketing**:
    - Estimating the impact of different advertising channels on sales.
4. **Environmental Science**:
    - Modeling the relationship between pollution levels and health outcomes.

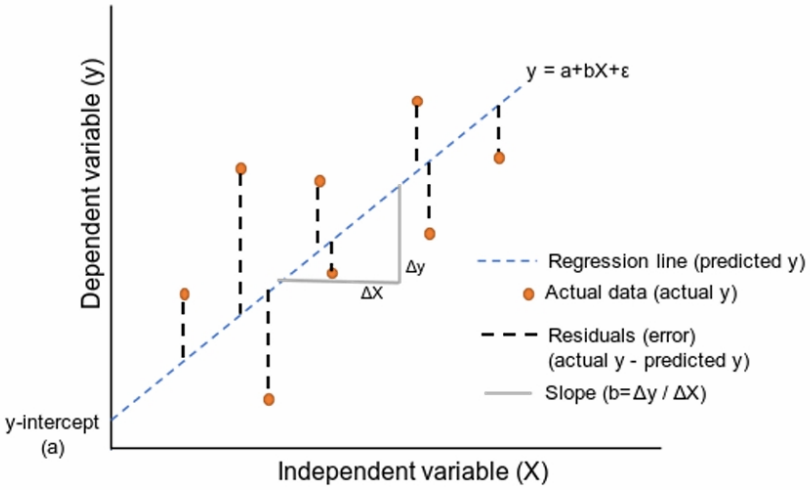
### 3. When Linear Regression May Not Be Appropriate?

1. **Non-linear Relationships**:
    - If the relationship between the variables is not linear, other methods like polynomial regression, logistic regression, or non-linear models may be more appropriate.

2. **Autocorrelation**:
    - When there is a correlation between residuals (errors), especially in time-series data. In such cases, time-series models like ARIMA may be more suitable.
3. **Significant Outliers**:
    - When the data contains significant outliers that can distort the results. Robust regression methods may be more appropriate in such cases.
4. **Categorical Dependent Variables**:
    - When the dependent variable is categorical (e.g., binary outcomes), logistic regression or other classification methods should be used instead.

*Linear regression is a powerful tool for modeling relationships between variables, making predictions, and analyzing trends. However, it's essential to ensure the data meets the assumptions of linear regression to achieve reliable and interpretable results. If the assumptions are violated, considering alternative methods or transforming the data may be necessary.*

F(x) = mx + c = y = predicted output




$$f(\mathbf{x}) = \sum_{j=1}^{D} w_j x_j + \epsilon = \mathbf{y}$$

The **model parameters** are:

$$\boldsymbol{\theta} = \{w_0, \ldots, w_n, \sigma\} = \{\mathbf{w}, \sigma\}$$

## 4. What are Residuals?

Residuals are the differences between the observed values and the values predicted by a regression model. They represent the error or deviation of the predicted values from the actual values. In the context of linear regression, residuals are used to assess the fit of the model.

**Formula**

For a given data point iii, the residual $e_i$ is calculated as $e_i = y_i - \hat{y}_i$ where:

- $y_i$ is the observed value of the dependent variable.
- $\hat{y}_i$ is the predicted value of the dependent variable from the regression model.

**Example**

Consider a simple linear regression model that predicts the weight of individuals based on their height. Suppose you have the following data:

| Height (inches) | Actual Weight (pounds) | Predicted Weight (pounds) |
|---|---|---|
| **60** | 115 | 113 |
| **62** | 120 | 118 |
| **65** | 125 | 123 |
| **70** | 140 | 138 |
| **72** | 150 | 145 |

The residuals for each data point would be:

$e_1 = 115 - 113 = 2$ , $e_2 = 120 - 118 = 2$ , $e_3 = 125 - 123 = 2$ , $e_4 = 140 - 138 = 2$ etc.

**Interpretation**

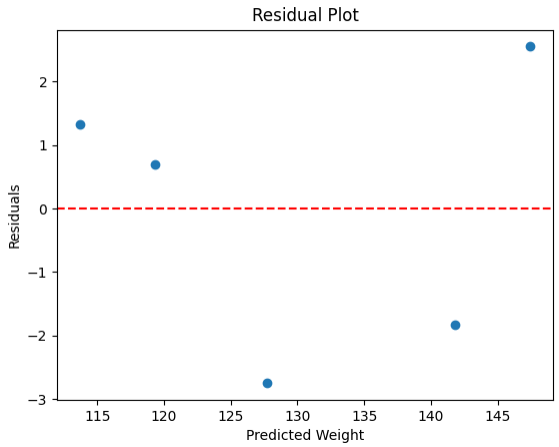
- **Positive Residual**: The observed value is greater than the predicted value (underestimation by the model).
- **Negative Residual**: The observed value is less than the predicted value (overestimation by the model).
- **Zero Residual**: The observed value is exactly equal to the predicted value.

**Use of Residuals**

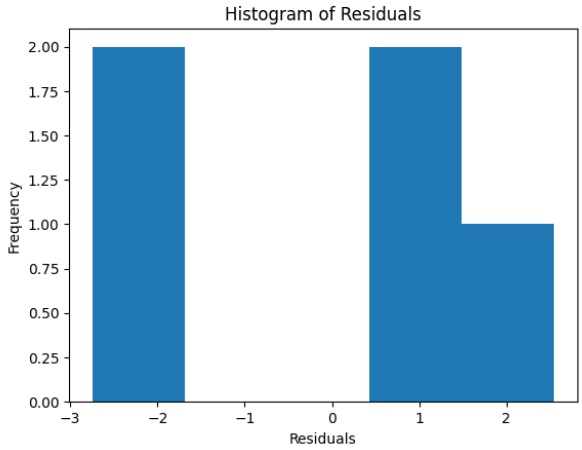
1. **Assessing Model Fit**:
    - Residuals help to determine how well the model fits the data. Ideally, residuals should be randomly scattered around zero, indicating no systematic errors in the predictions.
2. **Detecting Outliers**:
    - Large residuals indicate data points that are not well explained by the model and may be outliers.
3. **Checking Assumptions**:
    - **Linearity**: Residuals should not show patterns when plotted against predicted values or independent variables.
    - **Independence**: Residuals should not be correlated with each other.
    - **Homoscedasticity**: The spread of residuals should be constant across all levels of the independent variable.
    - **Normality**: Residuals should be approximately normally distributed for inference purposes.

Visualizing Residuals

1. Residual Plot




2. Histogram Diagram




Equation: $y = mx + c$ , we need to find the value of m and c. List Square Method

$$\boldsymbol{c = y - mx}$$

Formula 1: $\boldsymbol{m = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}}$ where:

- N is the number of observations
- x and y are the individual single points of the independent variable and dependent variable respectively
- $\sum xy = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + \cdots + x_ny_n$
- $\sum x^2 = (x_1)^2 + ([\![x_2)]\!]^2 + (x_3)^2 + ([\![x_4)]\!]^2 + \cdots + (x_n)^2$

Formula 2: $\boldsymbol{m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}}$ where:

- $\bar{x} = \sum x / N$ and $\bar{y} = \sum y / N$

## 5. What is Data Splitting?

Data splitting is the process of dividing a dataset into two or more subsets to validate and test the performance of machine learning models. The main subsets typically include:

1. **Training Set:** The portion of the data used to train the model.
2. **Validation Set:** A separate portion used to tune model parameters and prevent overfitting.
3. **Test Set:** A final portion used to assess the model's performance on unseen data.

**Purpose of Data Splitting**

- Prevent Overfitting: By using separate data for training and testing, we ensure that the model generalizes well to new data rather than just memorizing the training data.
- Model Validation: Data splitting allows for the evaluation of model performance and comparison of different models or configurations.
- Parameter Tuning: The validation set helps in fine-tuning model parameters, such as the learning rate or the depth of a decision tree.

**Common Data Splitting Techniques** Data Splitting example

1. **Holdout Method:**
    - The dataset is split into two parts: training and test sets.
    - A typical split ratio is 70% for training and 30% for testing, or 80% for training and 20% for testing.
2. **Train/Validation/Test Split:**
    - The dataset is split into three parts: training, validation, and test sets.
    - A common split might be 60% training, 20% validation, and 20% testing.
3. **Cross-Validation:** Cross-validation Example
    - The dataset is divided into k subsets (folds).
    - The model is trained k times, each time using a different fold as the test set and the remaining folds as the training set.
    - Commonly used cross-validation techniques include k-fold cross-validation, stratified k-fold cross-validation, and leave-one-out cross-validation.

## 6. what is Cross Validation and Cross Function?

Cross-validation is a statistical method used to estimate the performance and generalizability of a machine learning model. It involves partitioning the data into subsets, training the model on some subsets, and validating it on the remaining subsets. This process is repeated multiple times to ensure the model's performance is consistent and reliable.

**Types of Cross-Validation**

1. **K-Fold Cross-Validation**:
    - The dataset is divided into k equally sized folds.

- The model is trained k times, each time using a different fold as the validation set and the remaining folds as the training set.
- The performance metrics are averaged over the k trials to provide a robust estimate.

2. **Stratified K-Fold Cross-Validation**:
   - Similar to k-fold cross-validation but ensures that each fold has a similar distribution of the target variable. This is especially useful for imbalanced datasets.
3. **Leave-One-Out Cross-Validation (LOOCV)**:
   - A special case of k-fold cross-validation where k is equal to the number of data points.
   - Each data point is used as a single test case while the remaining data points are used as the training set.
4. **Repeated K-Fold Cross-Validation**:
   - The k-fold cross-validation process is repeated multiple times with different random splits of the data.
   - This helps to further ensure the reliability of the model's performance estimates.

**Benefits of Cross-Validation**

1. **Better Performance Estimation**:
   - Provides a more accurate estimate of a model's performance on unseen data compared to a single train-test split.
2. **Robustness**:
   - Helps in identifying if the model's performance is consistent across different subsets of data.
3. **Model Selection**:
   - Assists in comparing different models or hyperparameters to choose the best-performing model.

*Cross-validation is a powerful technique used to evaluate the performance and generalizability of machine learning models. By using different subsets of data for training and validation, cross-validation ensures that the model is not overfitting and provides a reliable estimate of its performance on new, unseen data.*

**1. Loss (or Error) for a Single Sample:**

• When you calculate the difference between the actual value and the predicted value for a single data point, it's generally referred to as a "loss" or "error" for that specific data point.

• This term is used to describe the discrepancy between the prediction and the true value for a single instance.

**2. Cost (or Loss) for the Entire Dataset:**

• When you calculate the average or total of these losses/errors across the entire dataset, it's often referred to as the "cost" or "loss" for the dataset.

• The term "cost" or "loss" is used to describe the overall quality of the model's predictions for the entire dataset.

### 7. How to calculate Cost(or Loss)?

***Mean Absolute Square (MAE):*** Mean Absolute Error (MAE) is a metric used to evaluate the performance of a regression model. It measures the average magnitude of the errors between the predicted values and the actual values, without considering their direction. In other words, MAE is the average of the absolute differences between predicted and actual values.

$$MAE = \frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i| \quad \text{where:}$$

- n is the number of data points.
- $y_i$ is the actual value.
- $\hat{y}_i$ is the predicted value.

Example of MAE:

| Height(inches) $x$ | Actual Weight (pounds) $y$ | Predicted Weight(pounds) $\hat{y}$ |
|---|---|---|
| 60 | 115 | 110 |
| 62 | 120 | 112 |
| 65 | 125 | 115 |
| 70 | 140 | 120 |
| 72 | 150 | 122 |

Calculate Absolute Errors:

$$\sum_{i=0}^{n} |y_i - \hat{y}_i| = |y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + |y_3 - \hat{y}_3| + |y_4 - \hat{y}_4| + |y_5 - \hat{y}_5|$$

$$= |115 - 110| + |120 - 112| + |125 - 115| + |140 - 120| + |150 - 122|$$

$$= 5 + 8 + 10 + 20 + 28 = 71$$

$$\therefore MAE = \frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i| = \frac{71}{5} = 14.2$$

***Mean Squared Error (MSE):*** Mean Squared Error (MSE) is a common metric used to evaluate the performance of a regression model. It measures the average of the squares of the errors between the predicted values and the actual values. The squaring of the errors ensures that larger errors have a disproportionately larger effect on the metric, making MSE sensitive to outliers.

$$MAE = \frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i|^2 \quad \text{where:}$$

- n is the number of data points.
- $y_i$ is the actual value.

- $\hat{y}_i$ is the predicted value.

**Example of MSE:**

| Height(inches) $x$ | Actual Weight (pounds) $y$ | Predicted Weight(pounds) $\hat{y}$ |
|---|---|---|
| 60 | 115 | 110 |
| 62 | 120 | 112 |
| 65 | 125 | 115 |
| 70 | 140 | 120 |
| 72 | 150 | 122 |

$$\sum_{i=0}^{n} |y_i - \hat{y}_i|^2 = |y_1 - \hat{y}_1|^2 + |y_2 - \hat{y}_2|^2 + |y_3 - \hat{y}_3|^2 + |y_4 - \hat{y}_4|^2 + |y_5 - \hat{y}_5|^2$$

$$= |115 - 110|^2 + |120 - 112|^2 + |125 - 115|^2 + |140 - 120|^2 + |150 - 122|^2$$

$$= (5)^2 + (8)^2 + (10)^2 + (20)^2 + (28)^2 = 1373$$

$$MAE = \frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i|^2 = \frac{1373}{5} = 274.6$$

***Root Mean Squared Error (RMSE):*** Root Mean Squared Error (RMSE) is a widely used metric to evaluate the performance of regression models. It is the square root of the Mean Squared Error (MSE) and provides a measure of the average magnitude of the errors between the predicted values and the actual values. RMSE is in the same units as the target variable, which makes it more interpretable compared to MSE.

$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{\left\{\frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i|^2\right\}}$$

$$RMSE = \sqrt{274.6} = 16.57$$

**Comparison with Other Metrics**

- **Mean Absolute Error (MAE)**: Measures the average of the absolute differences between predicted and actual values. Unlike RMSE, it is less sensitive to outliers because it does not square the errors.
- **Mean Squared Error (MSE)**: Measures the average of the squared differences between predicted and actual values. It is more sensitive to outliers compared to MAE.
- **R-squared ($R^2$)**: Measures the proportion of variance in the dependent variable that is predictable from the independent variables. It does not directly measure error but gives a sense of model fit.

**Correlation**

Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It quantifies how much two variables change together, and it can help in identifying whether and how strongly pairs of variables are related.

**Types of Correlation**

1. **Positive Correlation**: When one variable increases, the other variable also increases. For example, height and weight often have a positive correlation.
2. **Negative Correlation**: When one variable increases, the other variable decreases. For example, the speed of a car and the time taken to travel a fixed distance have a negative correlation.
3. **No Correlation**: There is no apparent relationship between the variables. For example, the number of books in a library and the amount of rainfall.

**Correlation Coefficient**

The correlation coefficient (often denoted as r) is a numerical measure of the direction and strength of the linear relationship between two variables. The value of r ranges from -1 to 1.

- r = 1: Perfect positive linear relationship.
- r = −1: Perfect negative linear relationship.
- r = 0: No linear relationship.

**Pearson Correlation Coefficient**

The Pearson correlation coefficient is the most common measure of correlation. It is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum((y_i - \bar{y})^2}}$$