Regression

Regression is a statistical method used to **find relationships** between variables and **make predictions.** It helps us understand how one or more **independent variables (inputs)** affect a **dependent variable (output).**

▼ Types of Regression

1. **Linear Regression** (Simple & Multiple) – The relationship is a straight line.
2. **Polynomial Regression** – The relationship is curved (quadratic, cubic, etc.).
3. **Logistic Regression** – Used for classification (yes/no, spam/not spam).

1. Simple Linear Regression

Used when there's **one independent variable (X)** affecting the dependent variable (Y).

📌 **Equation:**

$$Y = mX + c$$

Where:

- Y = Dependent variable (output)
- X = Independent variable (input)
- m= Slope (rate of change)
- c = Intercept (value of Y when X = 0)

🔷 **Example:**

Predicting **salary** based on **years of experience**.

2. Multiple Linear Regression

Used when there are **multiple independent variables**.

📌 **Equation:**

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Where:

- $X_1, X_2, ..., X_n$ = Different independent variables
- b_0 = Intercept

- $b_1, b_2, ..., b_n$ = Coefficients

🔷 **Example:**

Predicting **house prices** using **size, location, and number of rooms**.

3. Polynomial Regression

Used when the relationship is **non-linear (curved)**.

📌 **Equation:**

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \dots$$

🔷 **Example:**

Predicting the **growth of bacteria over time**, where the curve follows an **exponential** pattern.

Regression Metrics

Mean Squared Error (MSE)

Mean Squared Error (MSE) is a commonly used metric to measure the accuracy of a regression model. It quantifies the **average squared difference** between the **actual values** and **predicted values**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where:

- n = Number of data points
- Y_i = Actual value of the dependent variable
- $\hat{Y}_i$ = Predicted value from the model
- $(Y_i - \hat{Y}_i)$ = **Error (Residual)** (difference between actual and predicted values)
- **Squaring the error** ensures all errors are positive and penalizes larger errors more.

Why Use MSE?

✅ Measures Model Accuracy – **Lower** MSE means **better** predictions.

✅ Prevents Negative Errors from Canceling Out – Squaring ensures all errors contribute positively.

✅ Penalizes Large Errors More – Bigger mistakes have a greater impact.

Properties of MSE

✔️ **Always Positive** → Because errors are squared.

✔️ **Lower is Better** → Smaller MSE means more accurate predictions.

✔️ **Sensitive to Outliers** → Large errors get squared, increasing MSE significantly.

📈 Understanding the Derivative in Linear Regression

Root Mean Squared Error (RMSE)

The **Root Mean Squared Error (RMSE)** is a popular metric for measuring the accuracy of a regression model. It calculates the **square root of the average squared differences** between actual and predicted values.

$$RMSE = \sqrt{MSE}$$

Key Properties of RMSE:

1. **Non-Negative** – RMSE is always **≥ 0**.
2. **Same Units as Target Variable** – Unlike MSE, RMSE has the **same units** as the dependent variable, making it easier to interpret.
3. **Penalizes Large Errors More** – Since errors are **squared before averaging**, RMSE gives **more weight to large errors**, making it sensitive to outliers.
4. **Smooth and Differentiable** – This makes it useful for optimization in machine learning algorithms.

Why Use RMSE?

✅ **More sensitive to large errors** – Good for cases where big deviations matter.

✅ **Same units as the target variable** – Easier to interpret compared to MSE.

✅ **Works well for normally distributed errors** – When errors follow a normal distribution, RMSE is a great choice.

Mean Absolute Error (MAE)

The **Mean Absolute Error (MAE)** is a metric used to measure the accuracy of a regression model by calculating the average absolute difference between the actual and predicted values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{Y}_i|$$

where:

- N = Number of data points
- Y_i = Actual value of the target variable
- $\hat{Y}_i$ = Predicted value
- $|Y_i - \hat{Y}_i|$ = Absolute error for each data point

Explanation:

- **Absolute Error** ensures that negative and positive errors don't cancel out.
- MAE gives an average magnitude of errors **without considering direction** (i.e., overestimations and underestimations are treated equally).
- Lower MAE means better model performance.

Why Should We Use MAE?

✅ **Easy to Interpret** – Represents average error in the same unit as the target variable.

✅ **Less Sensitive to Outliers** – Unlike MSE, it does not heavily penalize large errors.

✅ **Balanced Error Contribution** – Treats all errors equally without squaring them.

✅ **Robust for Real-World Applications** – Useful in finance, sales, and forecasting where understanding absolute error is important.

Properties of Mean Absolute Error (MAE)

1. **Non-Negativity**:
    - MAE is always **≥ 0**. A perfect model would have **MAE = 0** (i.e., no error at all).
2. **Absolute Differences**:

- It considers the absolute differences between actual and predicted values, so it does not cancel out positive and negative errors.

3. **Linear Error Measurement**:
   - All errors contribute **equally** to the final metric since they are not squared (unlike MSE or RMSE).
4. **Robust to Small Variations**:
   - Small variations in predictions do not cause a large change in MAE, making it **less sensitive** than MSE to extreme errors.
5. **Scale Dependence**:
   - MAE has the same unit as the dependent variable, making it easy to interpret.

Difference Between MAE, MSE, and RMSE

Metric	Formula	Key Characteristics	Sensitivity to Outliers	Units
Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert Y_i - \hat{Y}_i\rvert$	Measures average absolute error	Less sensitive (treats all errors equally)	Same as target variable
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	Squares errors, penalizes large errors more	Highly sensitive (large errors impact more)	Squared units of target variable
Root Mean Squared Error (RMSE)	$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$	Square root of MSE, penalizes large errors but keeps unit consistency	Highly sensitive (like MSE)	Same as target variable

Key Takeaways:

- **MAE** is best when you need an intuitive measure of average error.
- **MSE** is useful when you want to penalize large errors more.
- **RMSE** is preferred when you want to balance penalizing large errors while maintaining interpretability.

R-Squared (R^2) – Coefficient of Determination

R-Squared (R^2) is a statistical measure that explains how well the independent variable(s) predict the dependent variable in a regression model. It represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).

Formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where:

- SS_{res} (Residual Sum of Squares) = $\sum(Y_i - \hat{Y}_i)^2$ → Measures unexplained variance (errors).
- SS_{tot} (Total Sum of Squares) = $\sum(Y_i - \bar{Y})^2$ → Measures total variance in Y.
- Y_i = Actual values, $\hat{Y}i$ = Predicted values, $\bar{Y}$ = Mean of actual values.

Key Properties of R^2:

1. **Ranges from 0 to 1:**
   - $R^2 = 1$ → Perfect model (100% variance explained).
   - $R^2 = 0$ → Model does not explain variance.
   - Can be **negative** if the model performs worse than just predicting the mean $\bar{Y}$
2. **Higher R^2 Means Better Fit:**
   - A higher value indicates that the model explains more variability in the data.
3. **Does Not Detect Overfitting:**
   - A high R^2 does **not** guarantee a good model, as it does not consider the number of predictors.

Why Use R^2?

✅ **Easy Interpretation** → It shows how much of the variation in Y is explained by X.

✅ **Compares Different Models** → Helps compare the goodness-of-fit for different regression models.

✅ **Works Well for Simple Linear Regression** → Provides a clear measure of how well the independent variable explains the dependent variable.

When NOT to Use R^2?

❌ **For Non-Linear Models** → R^2 assumes a linear relationship.

❌ **When You Need Adjusted R^2 for Multiple Predictors** → Adjusted R^2 accounts for extra predictors.

Adjusted R-Squared (R^2_{adj})

Adjusted R-Squared is an improved version of **R-Squared (R^2)** that accounts for the number of predictors in a regression model. It adjusts for overfitting by penalizing the addition of unnecessary independent variables.

Formula:

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(N - 1)}{N - k - 1} \right)$$

where:

- N = Total number of observations (data points)
- k = Number of independent variables (predictors)
- R^2 = Regular R-Squared value

Key Differences Between R^2 and Adjusted R^2

Feature	R-Squared (R^2)	Adjusted R-Squared (R^2_{adj})
Formula Accounts for Predictors?	❌ No, increases with more variables	✅ Yes, penalizes unnecessary variables
Overfitting Risk?	✅ Higher risk	❌ Lower risk
Value Can Decrease?	❌ No, always increases with more predictors	✅ Yes, if an added variable does not improve the model

Why Use Adjusted R^2?

✅ **Prevents Overfitting** → Avoids misleading high R^2 values when unnecessary variables are added.

✅ **Better for Multiple Regression** → More reliable when working with multiple predictors.

✅ **Compares Models Fairly** → Helps choose the best model without bias toward complexity.

When NOT to Use Adjusted R^2?

❌ **For Simple Linear Regression** → Adjusted R^2 is unnecessary when there's only one predictor.

❌ **If Model Selection Uses Other Criteria (AIC/BIC)** → Some methods use different evaluation metrics for model selection.
