Understanding Linear Regression

Key Metrics and Methods. — Chinmay Rozekar

- Linear regression is one of the most fundamental algorithms in ML & statistics.
- It is used to model the relationship b/w dependent variable (target) and one or more independent variables (features)
- While, the concept of linear regression is simple, evaluating its performance requires understanding of several key metrics.

What is Linear Regression?

Linear Regression aims to find the "best-fit-line" that minimizes the difference b/w the predicted and the actual values. such that the dependent variable can be expressed in terms of independent variable.

$$y = mx + b$$

- y: dependent variable (target)
- m: slope of the line
- x: Independent variable
- b: y intercept

"The goal of Linear Regression is to find values of m and b that minimizes the error b/w predicted and actual values"

Evaluating Linear Regression Models :

Once a linear regression model is trained, its essential to evaluate its performance.

(1) Mean Absolute Error (MAE) :

- MAE measures the absolute difference b/w the predicted and actual values.
- It provides a straight forward interpretation on of how far off the predictions are on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where, y_i = actual value

$\hat{y}_i$ = predicted value

n = number of observations.

Interpretation :

- MAE is easy to understand because it has the same units as the target variable.
- It is robust to outliers since it doesn't square the error.

- A lower MAE indicates that the performance of the model is better.
- It tells us how much the model deviates from the actual observations.

(2). Root Mean Square Error (RMSE):

- RMSE, measures the ~~actual~~ square root of the average squared differences b/w predicted & actual values.
- It penalizes larger errors heavily than MAE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Interpretations:

- RMSE is also in the same units as the target variable.
- It is sensitive to outliers because Errors are squared before averaging.

> • A lower RMSE indicates better model performance. !

- RMSE is a more robust measure of model performance if your data set has outliers.

MAE vs. RMSE

	Aspect	MAE	RMSE
1.	Sensitivity	Less sensitive to Outliers	More Sensitive to Outliers
2.	Interpretation	Avg. absolute Error	Sq. root of the avg. squared error.
3	Penalty	Equal penalty for all Errors	Higher penalty for larger errors.
4	Use case	When Outliers are NOT Critical	when large errors are undesirable

3. R^2

- R^2 is also known as the coefficient of determination.
- It measures the proportion of the variance in the dependent variable that is predicted from the independent variables.
- It ranges from 0 to 1.

$R^2 = 0$ means model explains none of the variability in the dependent variable.

$R^2 = 1$ means that the model explains all the variability in the dependent variable.

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

- SS_{res} is the sum of squared residuals
- SS_{tot} is the total sum of squares.

or,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Numerator → measure of the Error
Denominator → variability in y

Interpretation :

- A higher R^2 indicates a better fit of the model to the data.
- However, R^2 alone DOES NOT tell you whether the model is appropriate
  OR
  whether the independent variables are meaningful.

Example :

- If $R^2 = 0.85$

It means that 85% of the variability in the dependent variable is explained by the model.

Limitations of R^2 :

(1) Overfitting :

R^2 always increases as you add more independent variables or x's in your linear equation, even if they are irrelevant. This leads to overfitting.

(2) No indication of Causality.

A high R^2 doesn't imply that the independent variables cause the changes in the dependent variables.

(3) Sensitive to Outliers :

R^2 can be misleading if the data contains outliers.

(4) Adjusted R^2

- Adjusted R^2 is a modified version of R^2 that adjusts the independent variables in the model.
- It penalizes the addition of unnecessary variables that don't improve the models explanatory power.

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1} \right)$$

n = number of Observations
k = number of Independent variables

Interpretation:

- Adjusted R^2 can be < 0 if the model performs worse than the mean of the dependent variables.
- It provides a more accurate measure of the model's goodness of fit when multiple variables are included.

Example:
If Adjusted $R^2 = 0.80$, it means that 80% of the variability in the dependent variable is explained by the model, after accounting for the number of predictors.

Difference b/w R^2 and Adjusted R^2.

	Aspect	R^2	Adjusted R^2.
(1)	Purpose	Measures Explained Variance	Adjusts for Unnecessary Variables.
(2)	Behavior	Always increases with more variables	Increases only if new variables improve the model.
(3)	Range	0 to 1	Can be less than 0
(4)	Use case	Simple Models	Models with multiple predictors.