Machine Learning

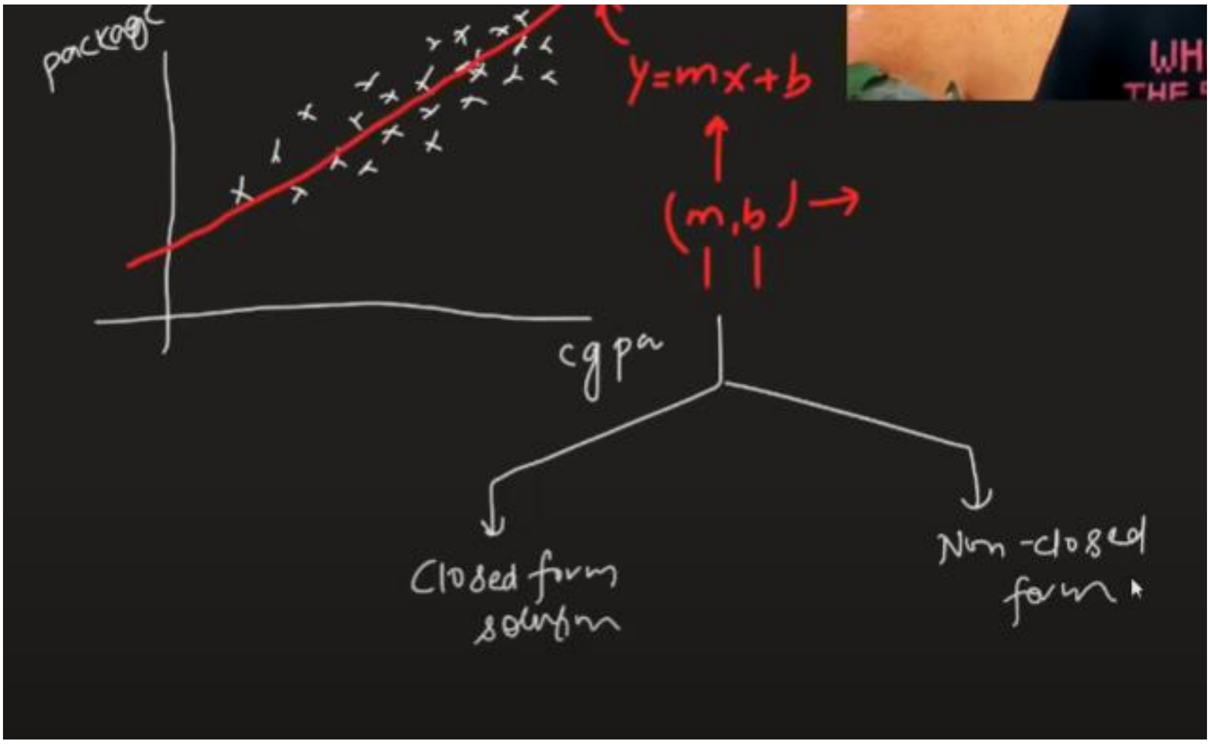
Video 51:

Simple Linear Regression | Mathematical Formulation

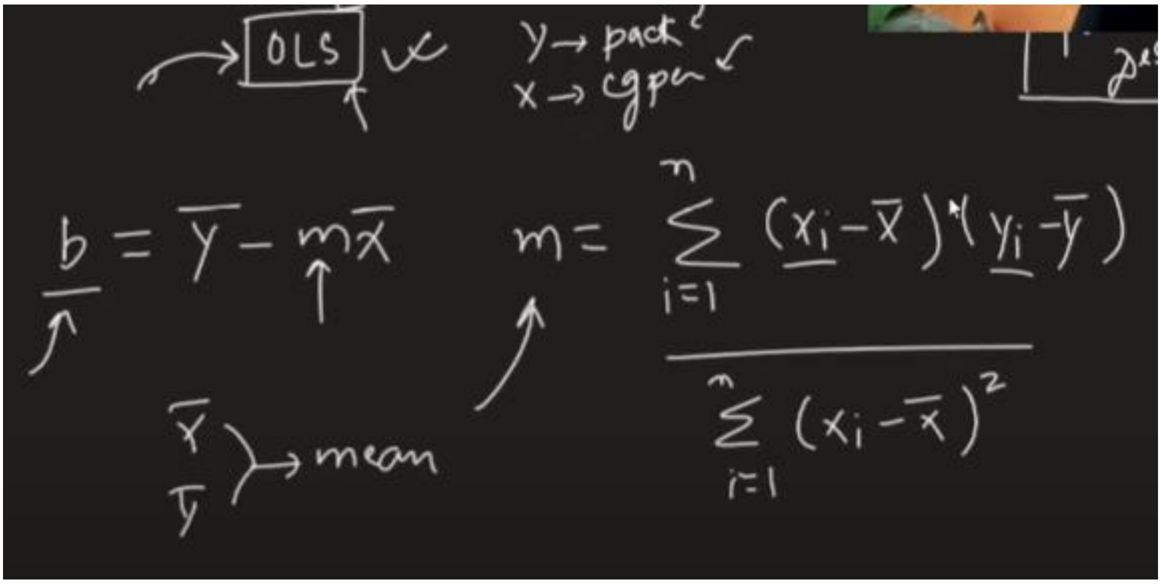
What we do in LR?

We find the value of m and b for y = mx + b

How to do so?

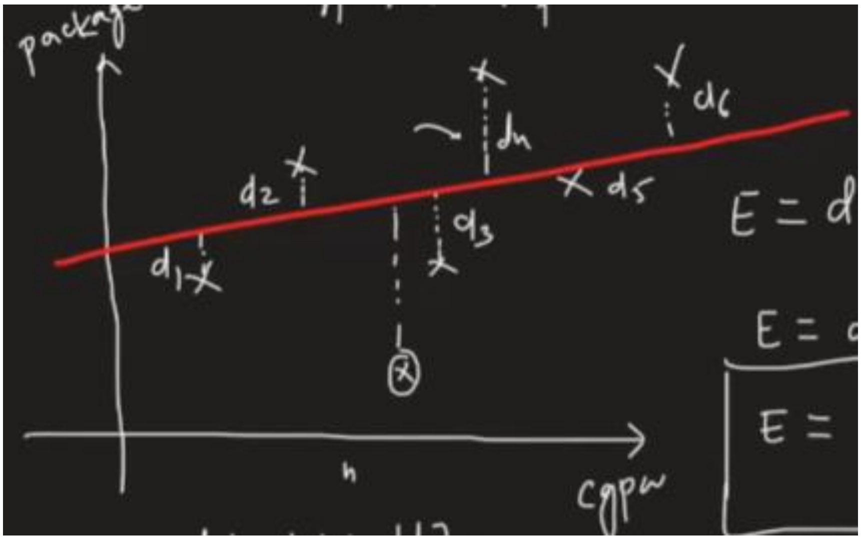


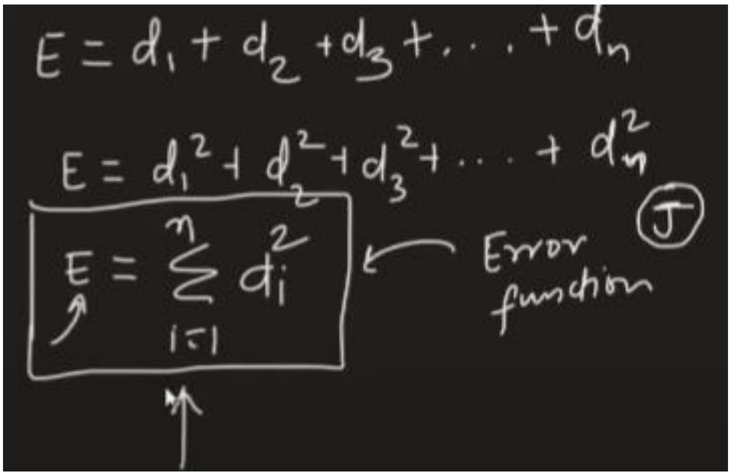
We have method to solve by closed-form: OLS



$$b = \bar{y} - m\bar{x} \qquad m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Derivation:





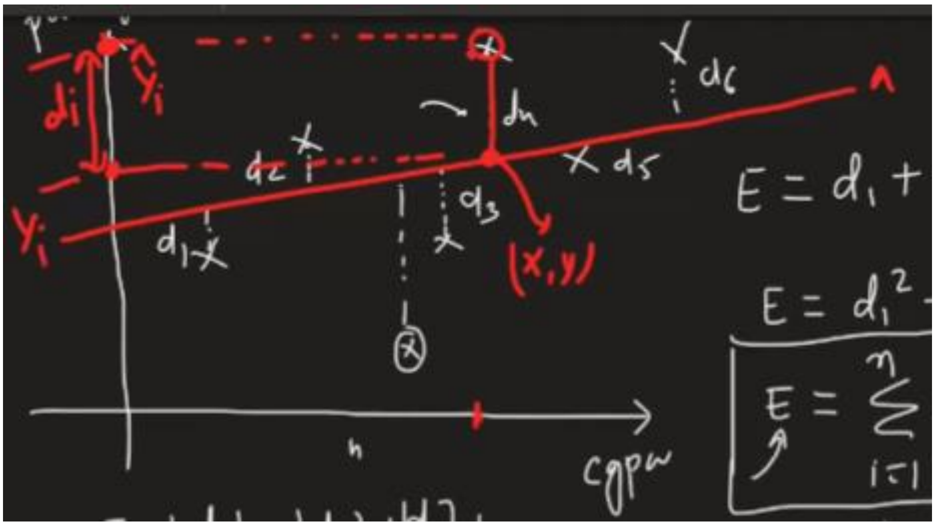
$$E = d_1 + d_2 + d_3 + \ldots + d_n$$

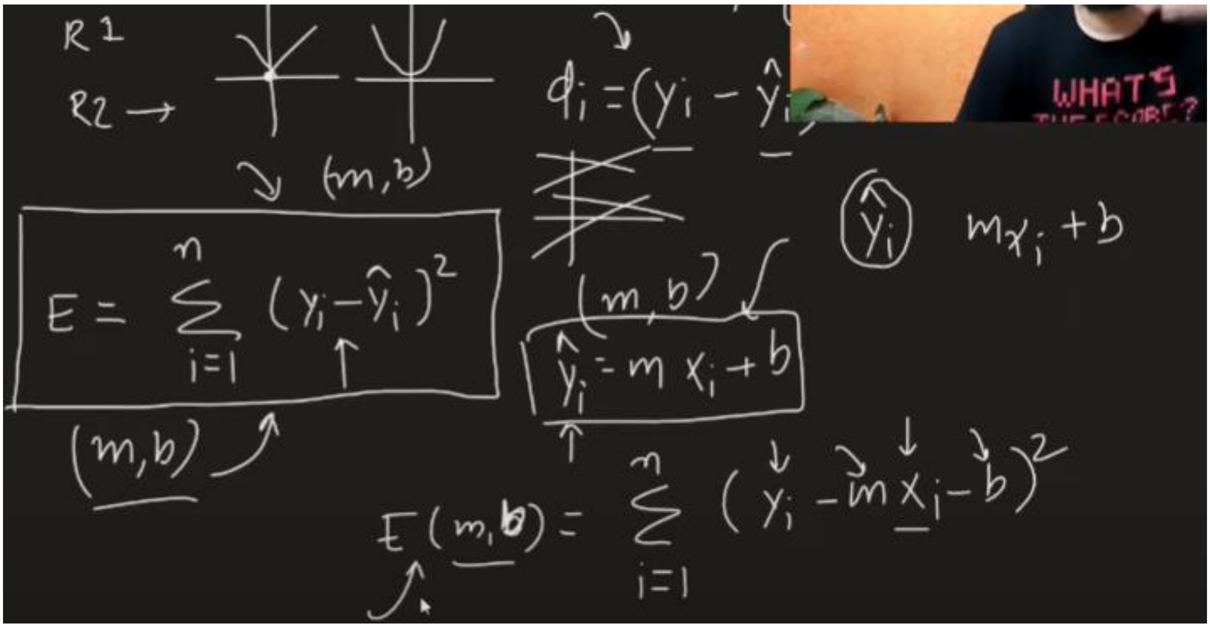
$$E = d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2$$

$$\boxed{E = \sum_{i=1}^{n} d_i^2}$$

Error function

J





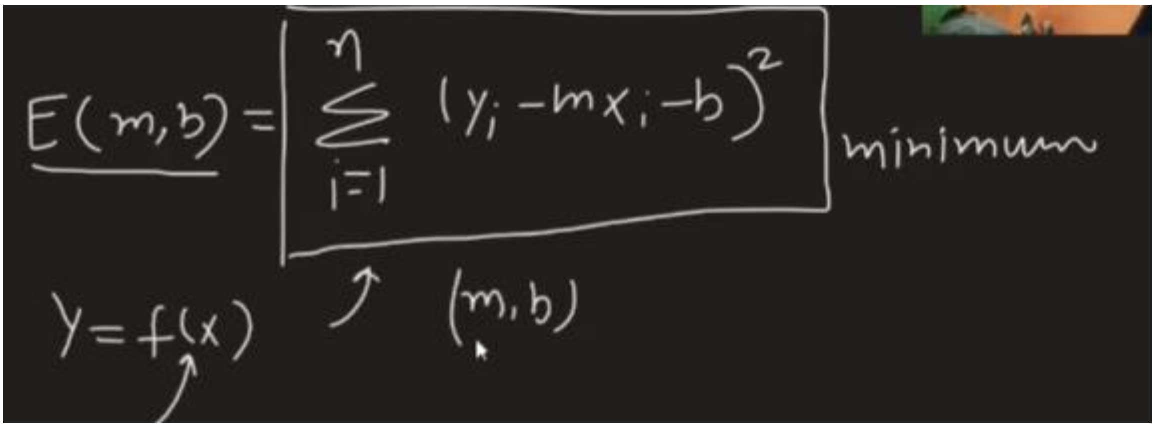
$$\boxed{E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$d_i = (y_i - \hat{y}_i$$

$$\boxed{\hat{y}_i = m x_i + b}$$

$$\hat{y}_i \quad m x_i + b$$

$$E(m,b) = \sum_{i=1}^{n} (y_i - m x_i - b)^2$$



$$E(m,b) = \boxed{\sum_{i=1}^{n} (y_i - m x_i - b)^2}$$

$$y = f(x)$$

Example:

Code link: https://colab.research.google.com/drive/1dwfdtMale-0wb8ygBOJftm-ZbiVRyQKo?usp=sharing

Data link:

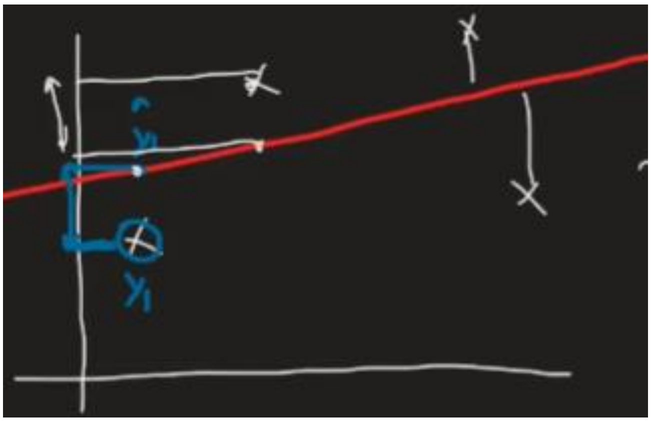
https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day48-simple-linear-regression

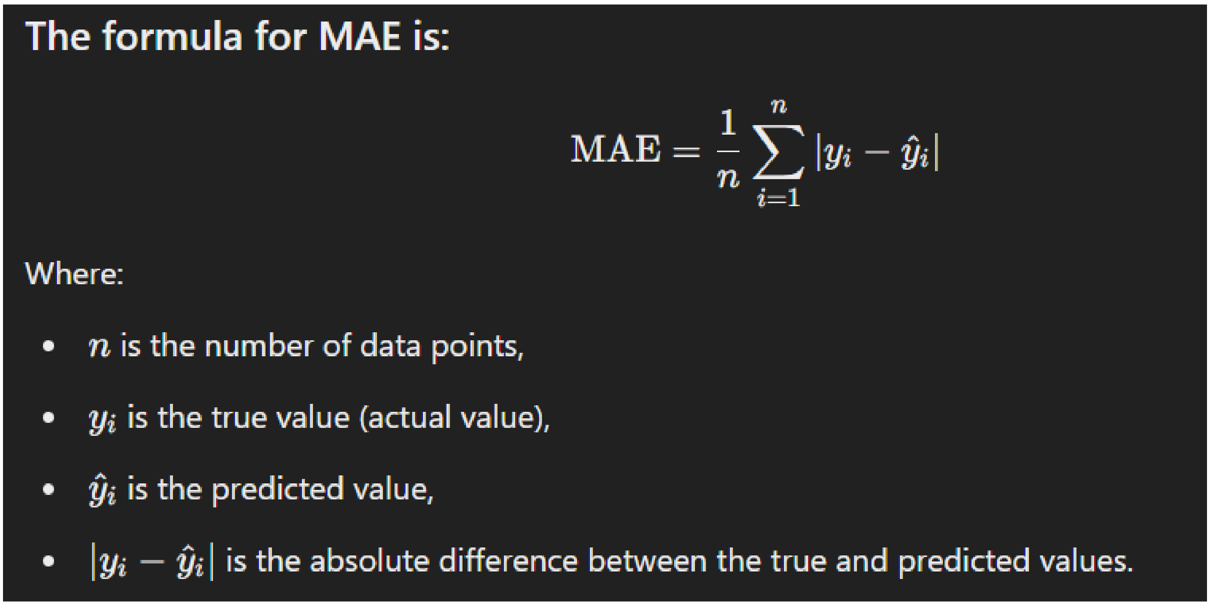
Video 52:

Regression Metrics | MSE, MAE & RMSE | R2 Score & Adjusted R2 Score

What is MAE?

In machine learning, **MAE** stands for **Mean Absolute Error**. It is a metric used to evaluate the performance of regression models. MAE calculates the average of the absolute differences between the predicted values and the actual values. It measures how far off the predictions are from the true values, but unlike other metrics like MSE (Mean Squared Error), it doesn't square the errors, which makes it less sensitive to outliers.





The formula for MAE is:

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

Where:

- n is the number of data points,
- y_i is the true value (actual value),
- $\hat{y}_i$ is the predicted value,
- $|y_i - \hat{y}_i|$ is the absolute difference between the true and predicted values.

Advantage: The advantage of MAE is its simplicity and interpretability, as it provides an easily understandable measure of prediction accuracy in the same units as the data.

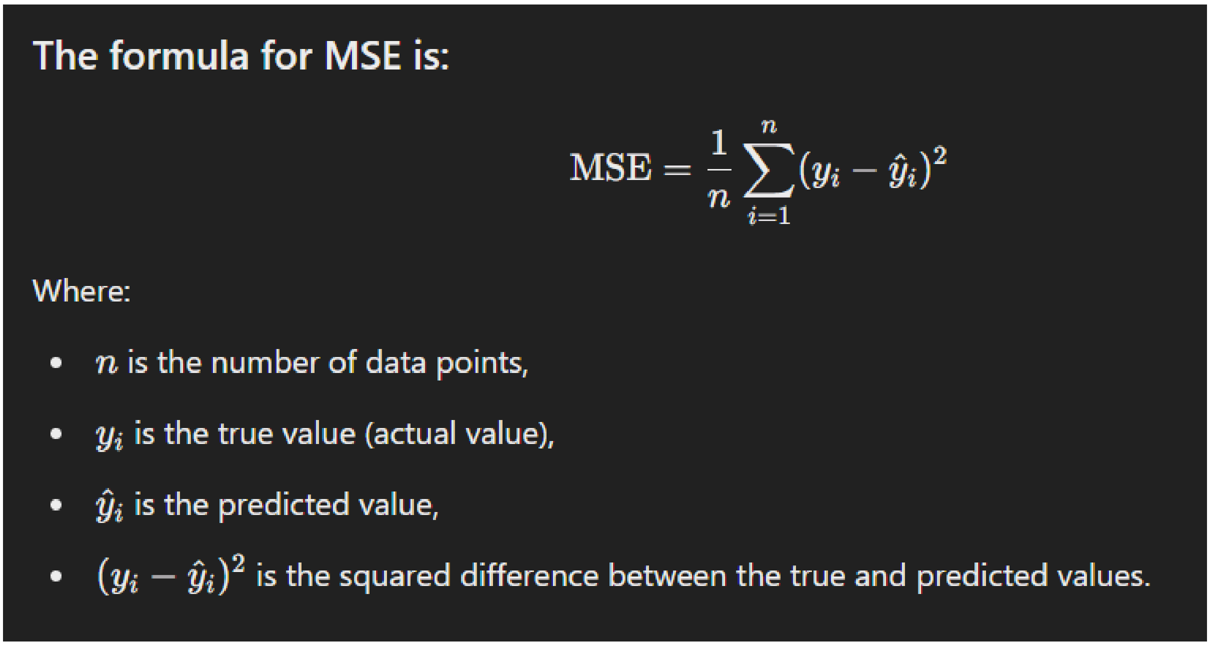
It's less sensitive to outliers compared to MSE, making it more robust in situations where large deviations may not be significant.

Disadvantage:

A disadvantage of MAE is that it doesn't penalize larger errors as heavily as MSE, potentially allowing significant mistakes to go unaddressed. It can also be less sensitive for models where large errors are more critical, and its optimization can be more challenging due to its non-differentiability at zero.

What is MSE?

MSE stands for **Mean Squared Error**, which is another common metric used to evaluate the performance of regression models. It measures the average of the squared differences between the predicted values and the actual values. Unlike MAE, MSE gives more weight to larger errors because it squares the differences, which can help identify models that make larger mistakes.



The formula for MSE is:

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

Where:

- n is the number of data points,
- y_i is the true value (actual value),
- $\hat{y}_i$ is the predicted value,
- $(y_i - \hat{y}_i)^2$ is the squared difference between the true and predicted values.

Advantage:

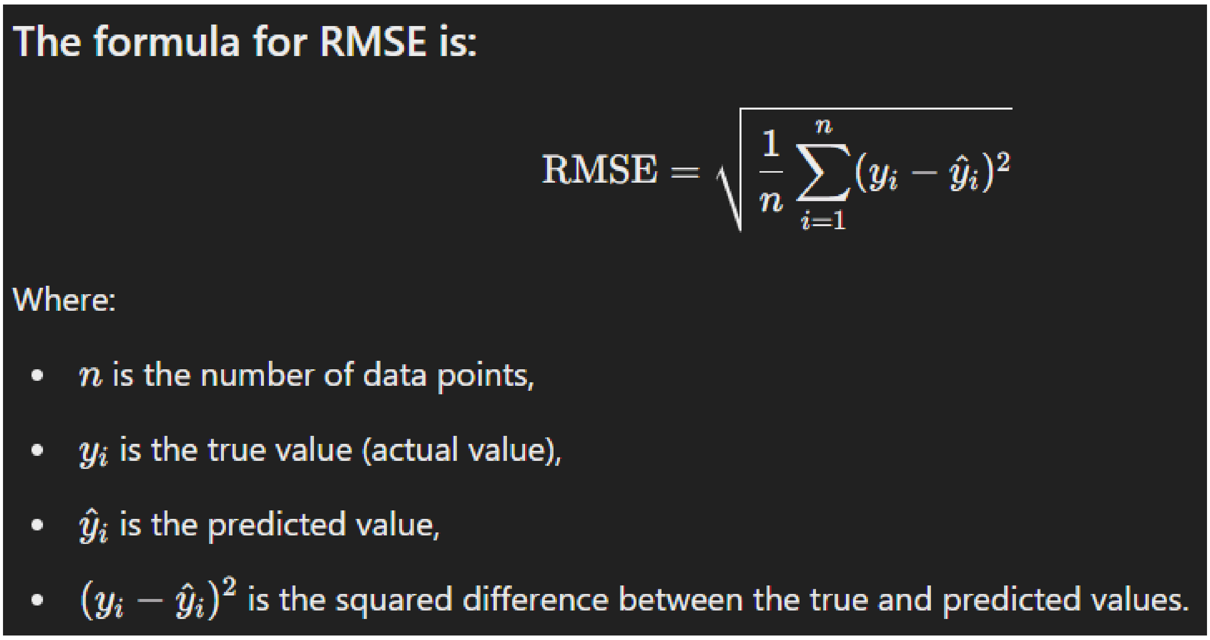
1. **Sensitive to Large Errors:** Since MSE squares the errors, it gives higher penalty to large deviations, which is useful when large errors are undesirable.
2. **Mathematical Convenience:** MSE is differentiable, making it easy to use with optimization techniques like gradient descent in machine learning algorithms.
3. **Better for Model Tuning:** It provides a smooth and continuous measure, helping fine-tune models more effectively during training.

Disadvantage:

1. **Sensitive to Outliers:** MSE can be overly influenced by outliers because it squares the errors, which can distort model performance when large errors are present.
2. **Not Easily Interpretable:** The result is in squared units of the original data, making it harder to interpret in a meaningful, real-world context.
3. **Ignores Direction of Error:** Like MAE, MSE doesn't distinguish between overestimations and underestimations, only focusing on the magnitude of errors.

What is RMSE?

RMSE stands for **Root Mean Squared Error**, which is a commonly used metric for evaluating the accuracy of regression models. RMSE is essentially the square root of the Mean Squared Error (MSE). It measures the average magnitude of the error in the same units as the original data, making it more interpretable compared to MSE.



The formula for RMSE is:

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

Where:

- n is the number of data points,
- y_i is the true value (actual value),
- $\hat{y}_i$ is the predicted value,
- $(y_i - \hat{y}_i)^2$ is the squared difference between the true and predicted values.

Advantages of RMSE:

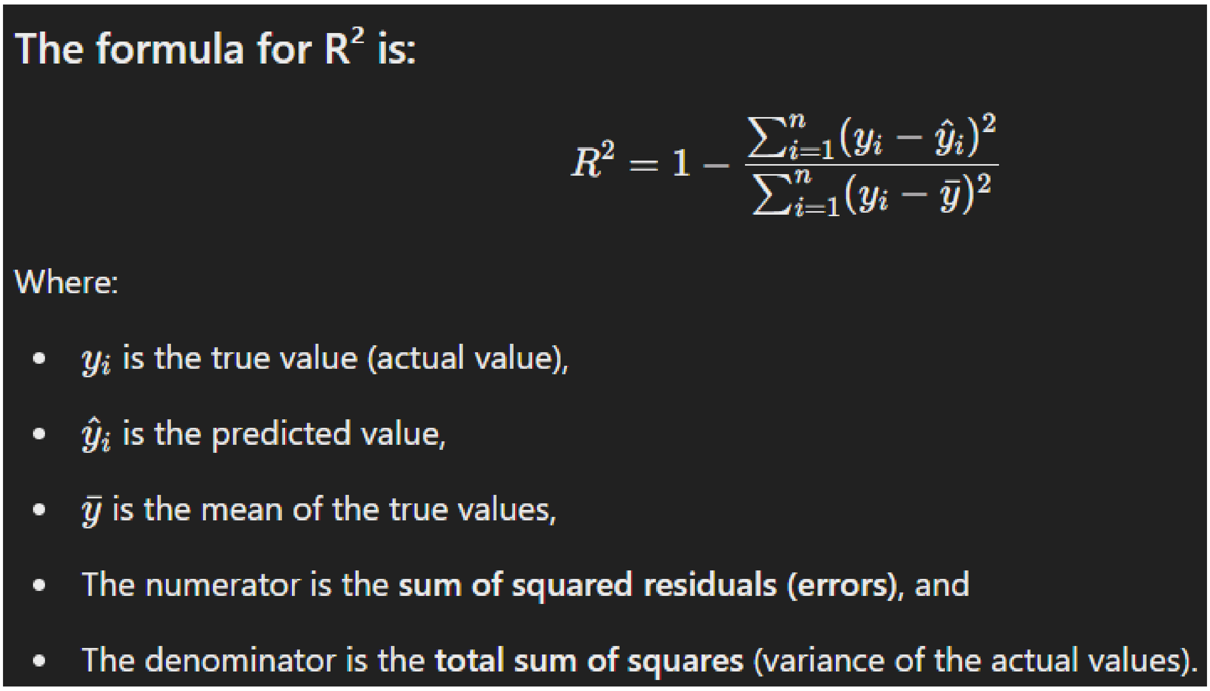
1. **Interpretability:** RMSE is in the same units as the original data, making it easier to interpret compared to MSE, which is in squared units.
2. **Sensitive to Large Errors:** Like MSE, RMSE penalizes larger errors more heavily, which can be useful when large deviations are particularly undesirable.
3. **Widely Used:** RMSE is a commonly accepted metric in many fields, making it easy to compare performance across different models and domains.
4. **Continuous Metric:** It provides a smooth and continuous performance measure, aiding in the optimization of models, especially in gradient-based training.

Disadvantages of RMSE:

1. **Sensitive to Outliers:** Since RMSE squares the differences between predicted and actual values, it can be heavily influenced by outliers, distorting the model evaluation if large errors are present.
2. **Not Robust for All Data Types:** In cases where the data contains many outliers, RMSE may give misleading results compared to other metrics like MAE, which is less sensitive to outliers.
3. **Doesn't Reflect Direction of Errors:** Like MSE, RMSE doesn't distinguish between overestimates and underestimates, only indicating the magnitude of errors.
4. **Less Robust in Certain Scenarios:** In some cases, RMSE might not be the best metric, especially when the cost of large errors is not proportional to their size.

What is R2 Score?

The **R^2 score** (also called **coefficient of determination**) is a statistical metric used to evaluate the goodness of fit of a regression model. It indicates how well the model's predictions match the actual data. R^2 measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variable(s) (features).



The formula for R² is:

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

Where:

- y_i is the true value (actual value),
- $\hat{y}_i$ is the predicted value,
- $\bar{y}$ is the mean of the true values,
- The numerator is the **sum of squared residuals (errors)**, and
- The denominator is the **total sum of squares** (variance of the actual values).

Advantages of R² Score:

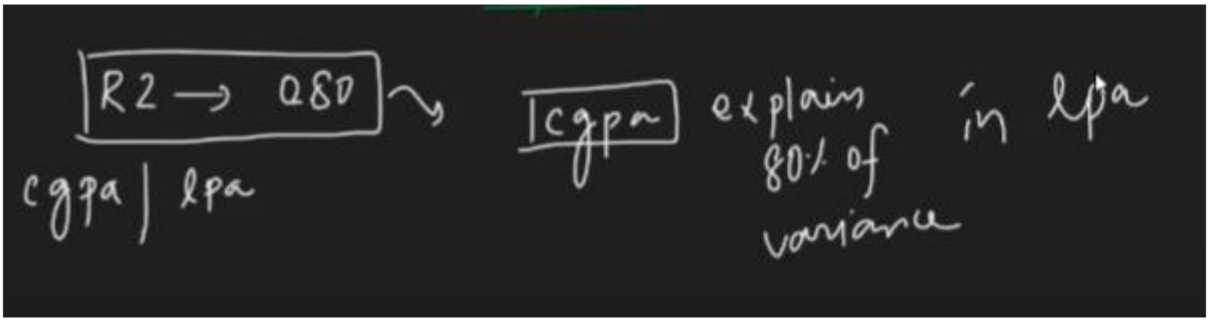
1. **Easy to Interpret:** R² provides a straightforward measure of how well the model fits the data, with values between 0 and 1 making it easy to gauge model performance.
2. **Widely Used:** It's a standard metric for evaluating regression models, making it easy to compare model performance across different datasets or approaches.
3. **Explains Variance:** R² quantifies the proportion of variance explained by the model, which is useful for understanding how much of the data is captured by the features.
4. **Good for Comparing Models:** When comparing models with similar complexity, a higher R² score indicates a better fit.

Disadvantages of R² Score:

1. **Doesn't Handle Non-Linearity Well:** R² might not provide meaningful insight for models that are not linear or when there is complex non-linearity between predictors and the target.
2. **Insensitive to Overfitting:** A model can have a high R² value but still be overfitting the data (especially with many predictors), which may lead to poor performance on new, unseen data.
3. **Doesn't Reflect Model Accuracy:** A high R² doesn't necessarily mean accurate predictions—it only measures how well the model fits the data's variance, not the actual prediction error.
4. **May Be Misleading for Small Datasets:** R² might be less reliable for small datasets, as its value can fluctuate significantly based on the size and distribution of the data.
5. **Negative R² Values:** In some cases (like poorly fitting models), the R² score can be negative, indicating the model fits worse than a baseline model (predicting the mean). This can be misleading for evaluation.

In conclusion, while R² is a useful metric, it should be used in conjunction with other evaluation metrics (such as MAE, MSE, or RMSE) to get a more complete picture of model performance. Make sure R2 goes to 1 rather than 0.

Also,

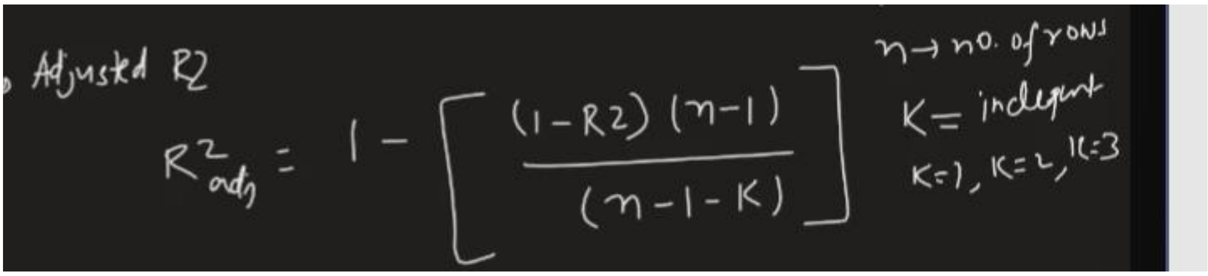


$\boxed{R2 \rightarrow 0.80}$

cgpa | lpa

$\boxed{cgpa}$ explain 80% of variance in lpa

Also, Adjusted R2 score: The **Adjusted R²** score is a modified version of the R² score that adjusts for the number of predictors in the model. It accounts for the fact that adding more variables to a model will always increase the R², even if those variables don't actually improve the model. Adjusted R² provides a more accurate measure of model fit, especially when comparing models with different numbers of predictors.



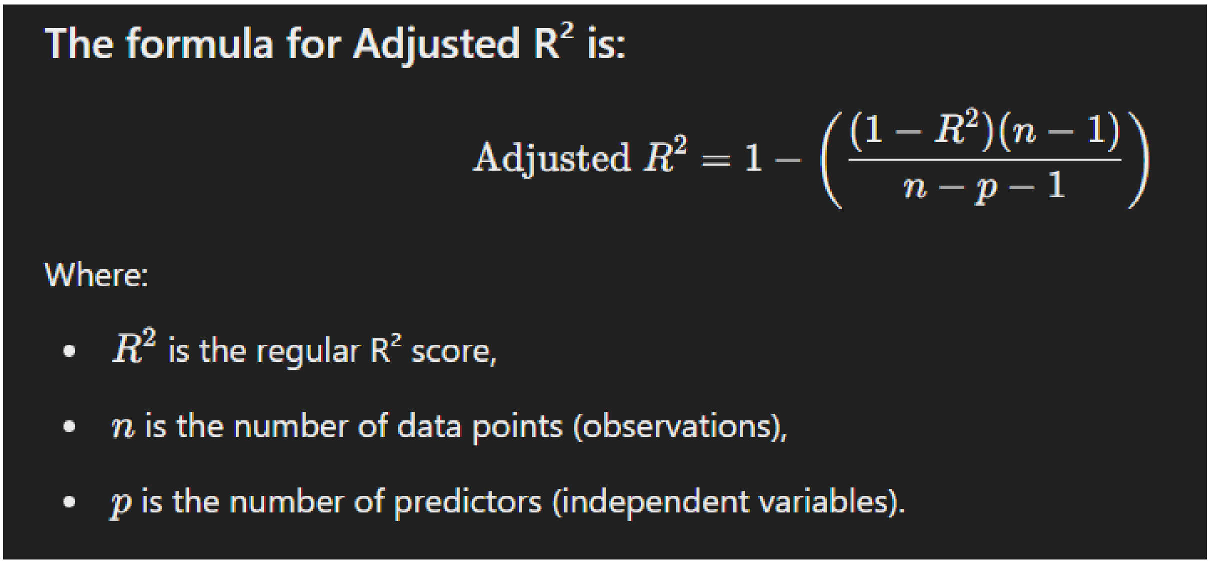
Adjusted R2

$$R^2_{adj} = 1 - \left[\frac{(1-R2)(n-1)}{(n-1-K)} \right]$$

$n \rightarrow$ no. of rows

K = indepnt

K=1, K=2, K=3



The formula for Adjusted R² is:

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

Where:

- R^2 is the regular R² score,
- n is the number of data points (observations),
- p is the number of predictors (independent variables).

Key Points:

- **Penalty for Adding Predictors:** Adjusted R² decreases if the added predictors do not improve the model sufficiently. It only increases when a new variable improves the model's predictive power more than would be expected by chance.
- **Range:** Adjusted R² can be negative if the model is worse than a simple mean prediction. For a good model, it will be closer to 1.

Example:

Code link:

https://colab.research.google.com/drive/1FNMvM6NVk4XjOyjG0TqRIq-9aLpLwKIY?usp=sharing

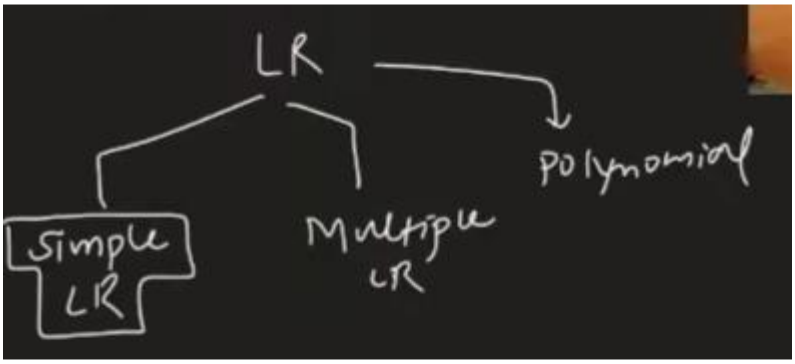
Data Link:

https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day49-regression-metrics

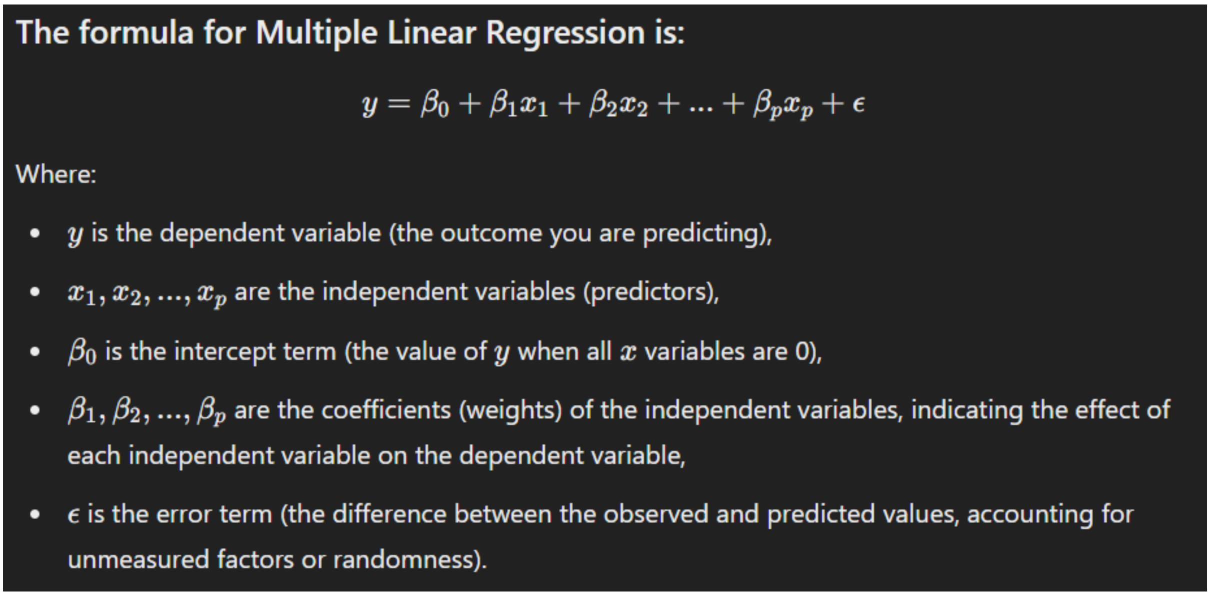
Video 53:

Multiple Linear Regression | Geometric Intuition & Code

We know LR is of following types:



Multiple Linear Regression (Multiple LR) is a statistical technique used to model the relationship between two or more independent variables (predictors) and a dependent variable (outcome). The goal is to understand how the dependent variable changes when the independent variables change, while keeping other variables constant.



The formula for Multiple Linear Regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Where:

- y is the dependent variable (the outcome you are predicting),
- $x_1, x_2, \ldots, x_p$ are the independent variables (predictors),
- β_0 is the intercept term (the value of y when all x variables are 0),
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (weights) of the independent variables, indicating the effect of each independent variable on the dependent variable,
- ϵ is the error term (the difference between the observed and predicted values, accounting for unmeasured factors or randomness).

Example:

Code link:

https://colab.research.google.com/drive/1FNMvM6NVk4XjOyjG0TqRIq-9aLpLwKIY?usp=sharing

Data link:

https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day50-multiple-linear-regression

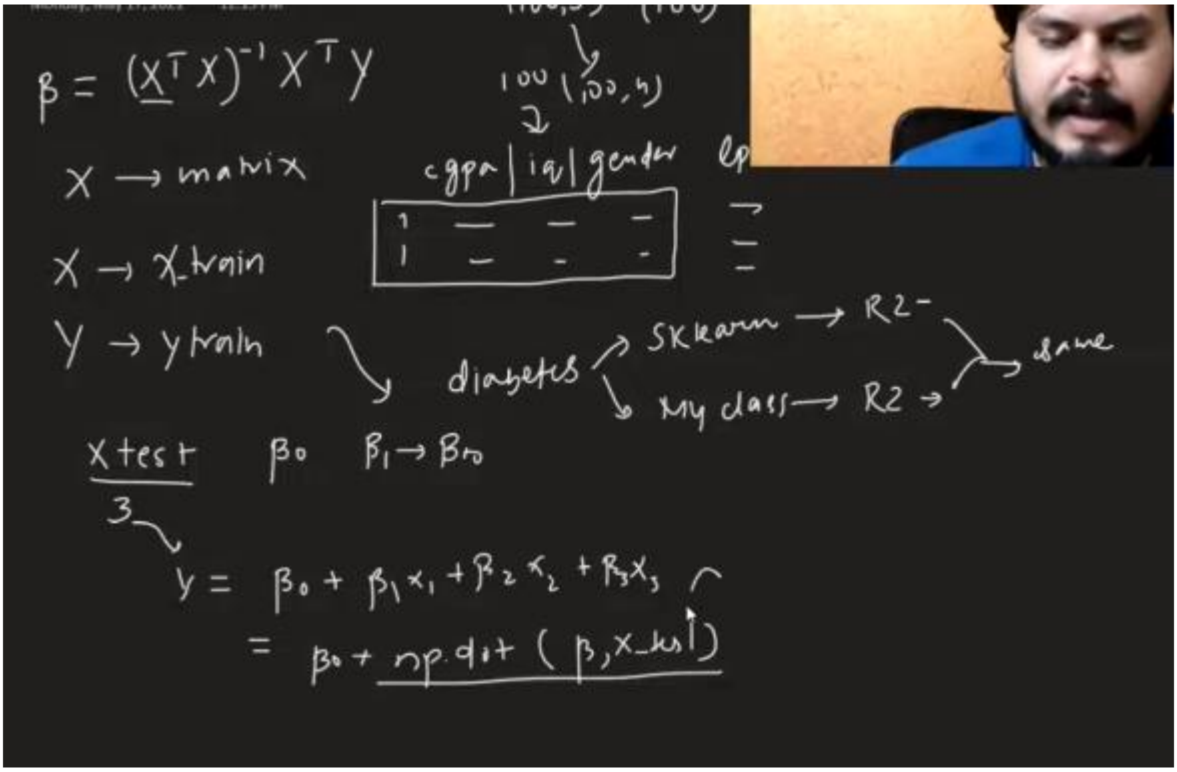
Video 54:

Multiple Linear Regression | Part 2 | Mathematical Formulation From Scratch

#Skipped the derivation

Video 55:

Multiple Linear Regression | Part 3 | Code From Scratch



Example:

Code link:

https://colab.research.google.com/drive/1CTjgtHLWpmGihopQajl0UPWVN01L74bD?usp=sharing