LINEAR REGRESSION

INTRODUCTION:

Regression analysis is one of the most widely used methods for prediction. Linear regression is probably the most fundamental machine learning method out there and a starting point for the advanced analytical learning path of every aspiring data scientist.

A linear regression is a linear approximation of a causal relationship between two or more variables. Regression models are highly valuable, as they are one of the most common ways to make inferences and predictions. Apart from this, regression analysis is also employed to determine and assess factors that affect a certain outcome in a meaningful way. As many other statistical techniques, regression models help us make predictions about the population based on sample data.

KEY CONCEPTS:

Dependent Variable (Y): The outcome we want to predict.

Independent Variable (X): The predictor variable we use to make predictions.

THE LINEAR REGRESSION MODEL:

The simple linear regression model is expressed as:

$$Y = \beta 0 \ + \beta 1 \ X + \epsilon$$

Where ,

Y= dependent variable or Estimated (predicted) value;

X= independent variable or Sample data for independent variable;

β 0 ,β 1 = constants or Coefficients.

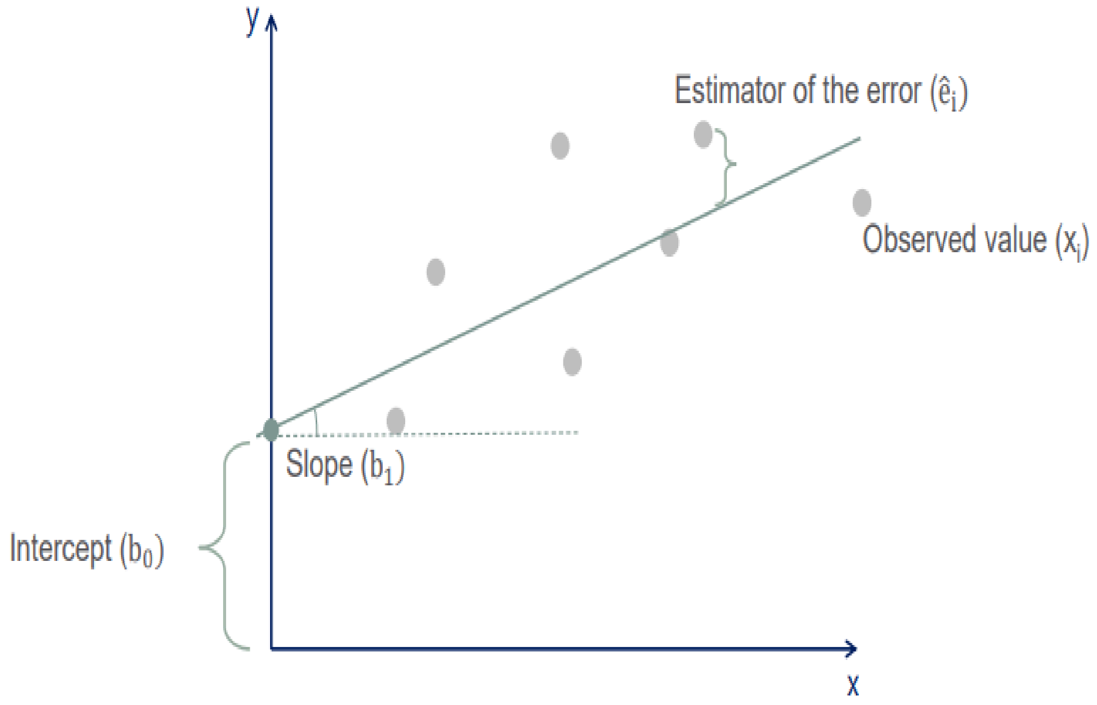
ϵ = Error

ASSUMPTIONS:

- The relationship between X and Y is linear.
- Observations are independent.
- The variance of error terms is constant.
- Error terms are normally distributed.

GEOMETRICAL REPRESENTATION OF LINEAR REGRESSION:

$yi = b0 + b1xi$



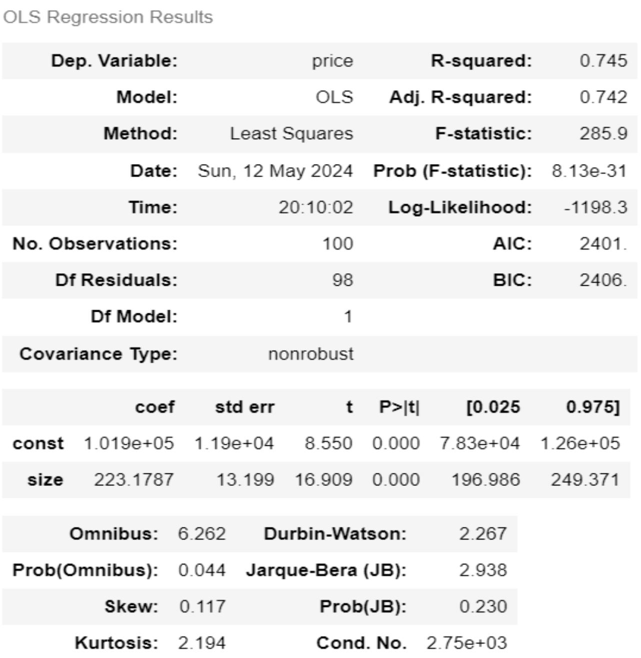
FITTING THE MODEL:

The least squares method finds the line that minimizes the sum of squared residuals.

EVALUATING THE MODEL:

R-squared measures the proportion of variability in Y explained by X.

SUMMARY TABLE AND IMPORTANT REGRESSION METRICS:



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.745
Model:	OLS	**Adj. R-squared:**	0.742
Method:	Least Squares	**F-statistic:**	285.9
Date:	Sun, 12 May 2024	**Prob (F-statistic):**	8.13e-31
Time:	20:10:02	**Log-Likelihood:**	-1198.3
No. Observations:	100	**AIC:**	2401.
Df Residuals:	98	**BIC:**	2406.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1.019e+05	1.19e+04	8.550	0.000	7.83e+04	1.26e+05
size	223.1787	13.199	16.909	0.000	196.986	249.371

Omnibus:	6.262	**Durbin-Watson:**	2.267
Prob(Omnibus):	0.044	**Jarque-Bera (JB):**	2.938
Skew:	0.117	**Prob(JB):**	0.230
Kurtosis:	2.194	**Cond. No.**	2.75e+03

R-squared: Variability of the data, explained by the regression model Range: [0;1].

Adj. R-squared: Variability of the data, explained by the regression model, considering the number of independent variables Range: <1; could be negative, but a negative number is interpreted as 0.

Prob (F-statistic): P-value for F-statistic; F-statistic evaluates the overall significance of the model (if at least 1 predictor is significant, F-statistic is also significant).

P>|t|: P-value of t-statistic; The t-statistic of a coefficient shows if the corresponding independent variable is significant or not.

EXAMPLES OF LINEAR REGRESSION:

https://drive.google.com/drive/folders/1iOdMW514BAwcbSf1q-dGA5LsJKMC6J2B?usp=sharing