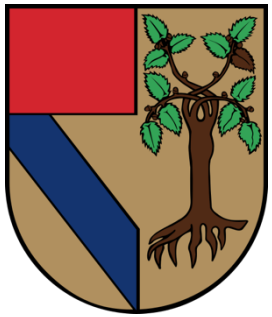


UNIVERSIDAD PANAMERICANA

APRENDIZAJE DE MÁQUINA

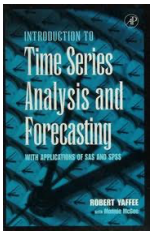
COM194

Enrique González N.

Profesor

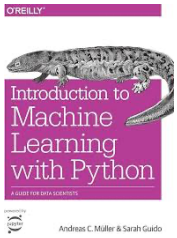
COM194-AM-SM1-24

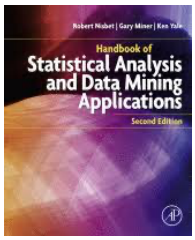
MATERIAL DE CONSULTA



Introduction to Time Series Analysis and Forecasting with applications and SAS and SPSS
Robert Yaffee
Academic Press

Introduction to Machine Learning with Pyhon
Andreas C. Müller & Sarah Guido
O'Reilly





Handbook of Statistical Analysis and Data Mining Applications
Robert Nisbet, Gary Miner & Ken Yale
Academic Press

PEARSON CORRELATION COEFFICIENT (PCC)

- AKA: Pearson coefficient, Pearson's r, or Pearson product-moment correlation coefficient.
- Defined by Karl Pearson (1911) – University College London (Department of Applied Statistics).
- It measures the linear correlation between two variables (sets of data).
- Is one of the most used to measure the grade or relationship between parametric variables:
  - Ex. The relationship between two stocks
- Is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale.
- Is a measure of the strength and direction of relationships between variables.
- Pearson correlation ranges from -1 to 1:
  - 1 indicates a perfect positive linear relationship.
  - -1 indicates a perfect negative linear relationship.
  - 0 indicates no linear relationship between the variables.

FORMULA

- It is the ratio between the covariance of two variables and the product of their standard deviations:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where

cov is the covariance
σ_X is the standard deviation of X
σ_Y is the standard deviation of Y

RESIDUAL SUM OF SQUARES (RSS)

- AKA: Sum of Squared Errors (SSE).
- It is a measure used in regression analysis to quantify the amount of variance that is not explained by the regression model.
- It essentially measures the discrepancy between the observed data and the values predicted by the regression model.

FORMULA

- Is computed by taking the squared differences between the observed values and the predicted values, then summing up these squared differences across all data points.

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

Where

y_i is the actual value for observation *i*.
$\hat{y}_i$ is the predicted value for observation *i*.
n is the total number of observations.

REGRESSION SUM OF SQUARES (SSR)

- AKA: Explained Sum of Squares (ESS).
- It measures the amount of variance in the dependent variable that is explained by the regression model.

- It represents the amount of variability in the dependent variable that is explained by the independent variables in the regression model.
- It complements RSS, which measures the unexplained variability.

FORMULA

- Is denoted as:

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

TOTAL SUM OF SQUARES (TSS)

- AKA: Sum of Squares Total (SST).
- It represents the total variance (variability) present in the dependent variable, without regard to the explanatory variables in a regression model.
- It is a measure of the dispersion of the data points around their mean.
- Is often decomposed into SSR (explained variability) and RSS (unexplained variability), such that TSS = SSR + RSS:
    - This decomposition helps in understanding the proportion of variance in the dependent variable that is explained by the regression model.

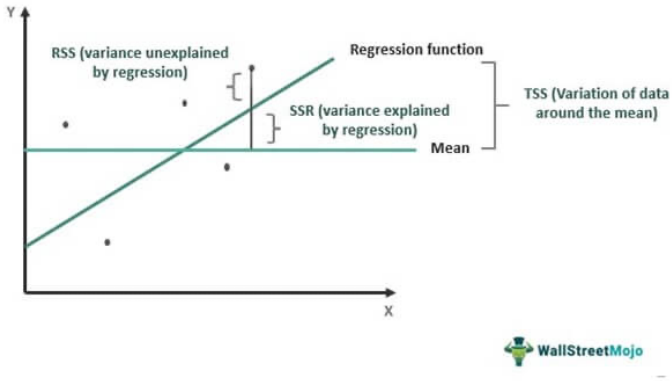
FORMULA

- Is denoted as:

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

$$TSS = RSS + SSR$$

TSS, SSR, and RSS



R-SQUARE (R2)

- AKA: coefficient of determination.
- Is a measure that represents the proportion of variance in the dependent variable that is explained by the independent variables in a regression model.
- It's a commonly used metric to evaluate the goodness-of-fit of a regression model.
- R-squared ranges between 0 and 1:
    - A higher R-squared value indicates a better fit of the regression model to the data, meaning that a larger proportion of the variance in the dependent variable is explained by the independent variables.
    - Conversely, a lower R-squared value suggests that the regression model does not explain much of the variability in the dependent variable.

FORMULA

- Is denoted as:

$$R^2 = 1 - \left(\frac{RSS}{TSS}\right)$$

ADJUSTED R-SQUARE (R2)

- Is a modified version of the R-squared.
- It penalizes overly complex models by adjusting R-squared downward as the number of predictors increases.
- It provides a more conservative estimate of the model's goodness-of-fit, especially when comparing models with different numbers of predictors.
- Adjusted R-squared will always be less than or equal to R-squared.

FORMULA

- Is denoted as:

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

AKAIKE INFORMATION CRITERION (AIC)

- It is a measure used in statistical modeling and model selection to balance the trade-off between the goodness-of-fit of the model (measured by the likelihood function) and its complexity (measured by the number of parameters), penalizing overly complex models.
- Lower values of AIC indicate better trade-offs between model fit and complexity.
- AIC is particularly useful when comparing multiple models fitted to the same dataset:
    - The model with the lowest AIC is considered the best model, as it provides the best balance between goodness-of-fit and parsimony (simplicity).
    - AIC does not provide an absolute measure of model fit; it should be used in combination with other diagnostic tools and considerations when selecting the final model.

FORMULA

- Is denoted as:

$$AIC = n * \ln\left(\frac{RSS}{n}\right) + 2k$$

AMEMIYA's PREDICTION CRITERION (APC)

- Is a measure used in statistical model selection, particularly in the context of autoregressive models.
- APC is used to assess the predictive performance of a time series model. It evaluates how well the model predicts future observations based on a given set of past observations.
- The goal of using APC is to find the model that minimizes this criterion.
- Lower values of APC indicate better predictive performance of the model.

FORMULA

- Is denoted as:

$$APC = \ln(RSS) + \frac{2}{n}k$$

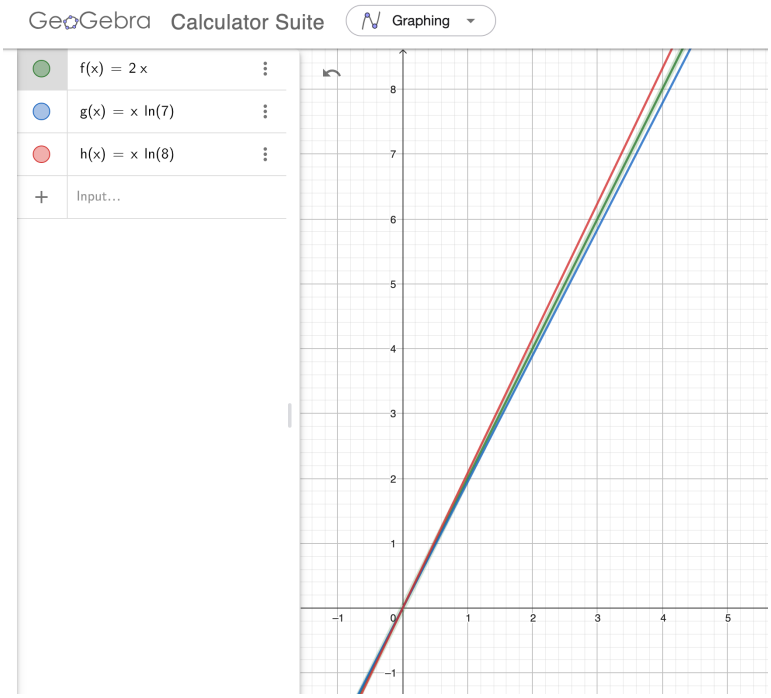
BAYESIAN INFORMATION CRITERION (BIC)

- It's a statistical measure used for model selection among a set of candidate models.
- It balances the trade-off between the goodness-of-fit of the model and the complexity of the model, penalizing overly complex models more severely than simpler ones.
- Similar to AIC, lower values of BIC indicate better trade-offs between model fit and model complexity.
- BIC penalizes complex models more heavily than AIC, as it includes a term that grows logarithmically with the sample size (n).
    - More complex models will have a worse (larger) score and will, in turn, be likely to be selected.
    - BIC is proportional to AIC, with a factor 2 replaced by log *N*, assuming $N > e^2 \approx 7.4$

FORMULA

- Is denoted as:

$$BIC = n * \ln\left(\frac{RSS}{n}\right) + k\ln(n)$$



MEAN SQUARED ERROR (MSE)

- It is a common metric used to evaluate the performance of a predictive model, particularly in regression analysis.
- MSE provides a measure of the average discrepancy between the predicted values and the actual observed values.
- Lower values of MSE indicate better predictive performance of the model, as they imply smaller prediction errors.

FORMULA

- It quantifies the average squared difference between the actual values and the predicted values produced by the model:

$$MSE = \frac{RSS}{n}$$

ROOT MEAN SQUARED ERROR (RMSE)

- It is a measure of the average magnitude of the errors between predicted and observed values in a regression analysis.
- RMSE is preferred over MSE because it is expressed in the same units as the dependent variable, making it easier to interpret.

FORMULA

- RMSE is derived from Mean Squared Error (MSE) by taking the square root of MSE:

$$RMSE = \sqrt{MSE}$$

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

- It is a commonly used metric to evaluate the accuracy of forecasts or predictions, particularly in the context of time series forecasting or demand forecasting.
- It measures the average magnitude of the percentage errors between predicted and observed values.
- It provides a percentage measure of the accuracy of the forecasts, allowing for easy interpretation and comparison across different forecasting models or datasets.

FORMULA

- It quantifies the average percentage difference between the actual and predicted values, making it a relative measure of accuracy.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$