datacamp

COURSE GLOSSARY

Introduction to Regression with statsmodels in Python

Coefficient of determination (R-squared): The proportion of variance in the response variable that is explained by the explanatory variable(s), ranging from 0 (no explanatory power) to 1 (perfect fit)

Confusion matrix: A 2×2 table for binary classification showing counts of true negatives, false positives, false negatives, and true positives used to evaluate model predictions

Explanatory variable: A variable used to explain or predict the response, also called an independent variable or "x

Extrapolation: Using a fitted model to predict responses at explanatory-variable values outside the range of the training data, which can yield unreliable or nonsensical results if the modeled relationship does not hold there

Fitted values: The model's predicted responses for the observations used to fit the model, often available as a fittedvalues attribute or column

Influence (Cook's distance): A metric combining residual size and leverage to quantify how much the fitted model would change if a particular observation were removed, with larger values indicating more influential points

Intercept: The predicted value of the response variable when all explanatory variables equal zero, represented as the constant term in a linear model

Dimension table: A table that provides descriptive context about facts, containing attributes (like customer, product, or geography) that are used to filter, group, and label measures

KPI (Key Performance Indicator): A measurable value that indicates how effectively an organization or team is achieving key business objectives and is used to track progress and inform decisions

Odds ratio and log-odds (logit): The odds ratio is the probability of an event divided by the probability it does not occur, and the log-odds (logit) is the natural logarithm of the odds, which linearizes multiplicative effects and is the scale used in logistic regression

Ordinary least squares (OLS): A common method for fitting linear regression that chooses coefficients to minimize the sum of squared residuals between observed and predicted response values

Predict (prediction): The act of using a fitted model and new explanatory-variable values to compute estimated response values or probabilities with the model's predict method

Q-Q plot (quantile-quantile plot): A diagnostic plot that compares the quantiles of sample residuals to the theoretical quantiles of a normal distribution to assess normality of residuals

regplot: A seaborn plotting function that draws a scatter plot with an optional fitted trend line (linear or logistic) and confidence interval, commonly used to visualize regression fits

Regression model: A statistical model that describes the relationship between one response variable and one or more explanatory variables and can be used to quantify that relationship or make predictions

Residual standard error (RSE), MSE and RMSE: Metrics that quantify typical prediction error where MSE is the mean squared residual, RMSE is its square root, and RSE is a similar root measure that adjusts by the model's residual degrees of freedom to give error in the response units

Residuals: The differences between observed responses and fitted values for each observation, representing the model's errors for those data points

Response variable: The outcome or target variable you want to predict or explain, also called the dependent variable or "y

Scatter plot: A graphical display of two numeric variables where each point represents an observation's values on the x and y axes, useful for visualizing relationships

Sensitivity and specificity: Sensitivity (true positive rate) is the proportion of actual positives correctly predicted, and specificity (true negative rate) is the proportion of actual negatives correctly predicted

Simple linear regression: A regression model with one explanatory variable where the relationship between x and y is modeled as a straight line defined by an intercept and a slope

Simple logistic regression: A regression model for a binary (two-category) response that models the log-odds of the outcome as a linear function of a single explanatory variable, producing S-shaped probability predictions

Slope (coefficient): The parameter that quantifies the expected change in the response variable for a one-unit increase in an explanatory variable, holding other variables constant

Transformation and back-transformation: Applying a mathematical change (e.g., log, square root, power) to a variable to improve model fit or meet assumptions, and back-transformation is undoing that change to interpret predictions on the original scale

Trend line: A line fitted through data points (often by regression) that summarizes the central tendency of the relationship between variables, such as a linear or logistic curve