REGRESSION ANALYSIS

DEFINITION:

Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables.

USES:

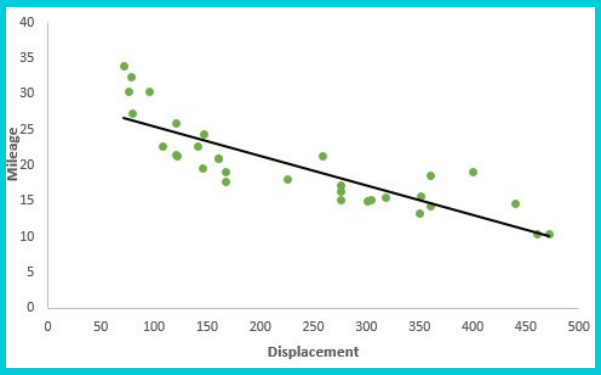
- Assess the significance of the predictor variable in explaining the variability or behavior of the response variable
- Predict the values of the response variable given the values of the predictor variable.
- Forecast the values of dependent variable based on relationship of dependent and independent variable taking in account the past values.

TYPES OF REGRESSION:

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Ridge Regression
- LASSO Regression

LINEAR REGRESSION

It is the simplest form of regression. It is a technique in which the dependent variable is continuous in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.
When you have only 1 independent variable and 1 dependent variable, it is called simple linear regression.
When you have more than 1 independent variable and 1 dependent variable, it is called Multiple linear regression.

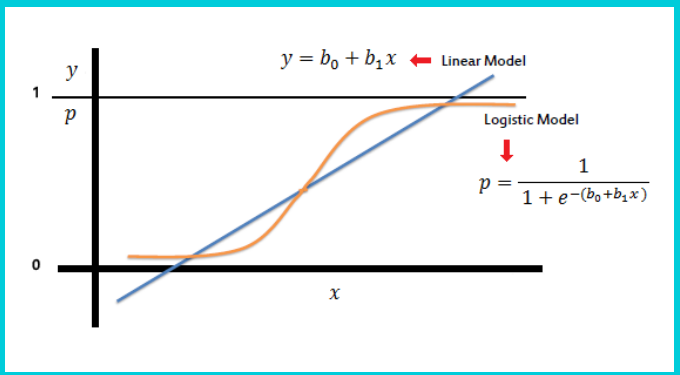


Its equation is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$

LOGISTIC REGRESSION

Generally logistic regression Is used when the response variable is dichotomous (yes or no questions, ex: will the student pass or fail, will the contestant win or lose, etc). But logistic regression can be used in cases where the response variables have more than two categories. In linear regression, the response variable is continuous. In logistic regression, the response variable is categorical.

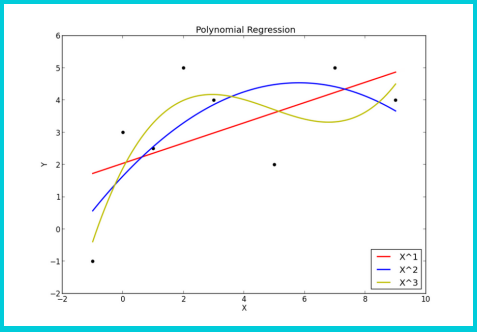


Its equation is given by:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + --- + b_k X_k)}}$$

POLYNOMIAL REGRESSION

It is a technique to fit a nonlinear equation by taking polynomial functions of independent variable. In the given figure the red curve fits the data better than the green curve. Where the red curve represents polynomial regressio



Its equation is given by:

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

RIDGE REGRESSION

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the **linear regression** estimates, ridge regression reduces the standard errors.

The equation to estimate parameters using ridge regression is:

$$\text{Min } (\sum \varepsilon^2 + \lambda \sum \beta^2) = \text{Min } \sum (y - (\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k))^2 + \lambda \sum \beta^2$$

LASSO REGRESSION

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares.

The equation to estimate parameters using LASSO regression is:

$$\text{Min } (\sum \varepsilon^2 + \lambda \sum |\beta|) = \text{Min } \sum (y - (\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k))^2 + \lambda \sum |\beta|$$

SELECTION CRITERIA FOR BEST MODEL

When you have several models with similar predictive power, choose the simplest because it is the most likely to be the best model.
Typically, you want to select models that have larger adjusted R-squared values. Adjusted R-squared increases only when a new variable improves the model by more than chance. Low-quality variables can cause it to decrease.
In regression, p-values less than the significance level indicate that the term is statistically significant.
“Reducing the model” is the process of including all candidate variables in the model, and then repeatedly removing the single term with the highest non-significant p-value until your model contains only significant terms.
During the specification process, check the residual plots. Residuals plots are an easy way to avoid biased models and can help you make adjustments.

DATE August 08, 2018

SUBJECT Regression Analysis

AUTHOR **SAMUEL**