DAR ES SALAAM INSTITUTE OF TECHNOLOGY (D.I.T)

DEPARTMENT OF COMPUTER STUDIES

NAME: SAMWEL JOHN PETRO

MODULE CODE: 08104

MODULE: DATA MINING

ASSIGNMENT 2

1.2. Collinearity: Is the situation where two or more predictor are strongly linear related meaning that one predictor can be approximated expressed as linear combination of others.

Effects of collinearity on estimated coefficient

Collinearity led to coefficient estimates become unstable and highly sensitive to small changes in data.

1.3.not all variables are significant, yes lack of significant can be caused by multicollinearity.

1.4. Forward selection: Is a stepwise model selection procedure that starts with no predictor variables in the model then the variables are added one at a time.

While

Backward selection: is a stepwise model selection procedure that starts with all candidate's predictor variables in the model then removed one least significant variable at a time.

1.5. stepwise selection: Is an iterative variable selection procedure that combines idea of forward selection and backward elimination.

The procedure works as follow

1. start with an initial model (null for forward selection)
2. Add the most significant variable
3. After addition, reassess all variables
4. Remove any variable that has become insignificant
5. Repeat until no further improvement

2.1. Linear regression: Is used when there is a continuous random variable. It is used in prediction of values.

While

Logistic regression: is used when the response variable is binary (0/1). Used in classification.

2.4. Parameter estimates and their significant

a) Passenger class has a negative coefficient and statically significant indicating that passenger from lower class has low probability of survival compared to those from high class
b) Sex has positive coefficient and is highly significant showing that female passengers have high odds of surviving compared to male passengers
c) Age has negative coefficient and statically significant showing that older passengers were less likely to survive

3.1. Principal Component Analysis: Is a method that reduce many correlated variables into small number of new, uncorrelated variables that keep most of the important information in the data.

Application of Principal component analysis

1. Dimensionally reduction
2. Feature Extraction
3. Noise reduction
4. Data visualization

- PCA is useful when the explanatory variable exhibit multicollinearity since it removes correlation by transforming the variables into independent component which is significant since it reduces overfitting and improve model stability.

3.2. Mathematical description of PCA

Let X be the data matrix containing the original variables

First the data are centered by subtracting the mean of each variable(μ)

$$X_c = X - \mu$$

Then correlation matrix is computed

$$\sum = \frac{1}{n-1} X_C^T X_C$$

PCA is then performed by solving the eigenvalue problem

$$\sum v_i = \lambda_i v_i$$

Where: λ_i are eigenvalues (variance by each principal component) and v_i are the eigenvectors (principal component)

Finally the original data are projected onto principal component

$$Z = X_c V$$

Where V is matrix of eigenvectors and Z is transformed data

3.3 . First principal component is similar to the market because of high average correlation and all stocks load positively also it captures maximum variance.

3.5. unusual stocks in PCA show up a distant because their returns vary more than other stocks also they move different than the market.