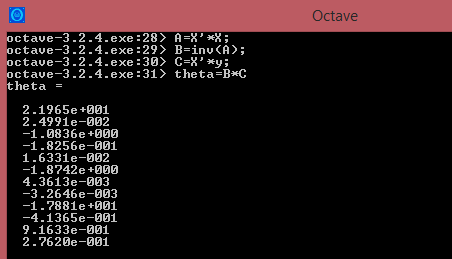
Statistical Analysis of Wine Quality Data

# Linear Regression using Octave

On solving the linear regression using Normal equation method for the equation

Theta0 + Theta1\*fixed.acidity+Theta2.Volatile.acidity+…………..+Theta11\*alcohol, we get the Theta vector as below



# T-test for alcohol level comparison for Good vs Bad Alcohol

Good Alcohol (X):

Quality >6

Mean: 11.52

Standard deviation: 0.918

(Count of Observations) N = 217

Bad Alcohol (Y):

Quality < 5

Mean: 10.22

Standard deviation: 0.998

Mean of good alcohol = 11.52

(Count of Observations) N = 63

Null Hypothesis: Mean(Good Alcohol) = Mean(Bad Alcohol)

Alternate Hypothesis: Mean(Good Alcohol) != Mean(Bad Alcohol)

## Result:

Manual Calculation:

T = Mean1-Mean2/Standard Error

Degrees of Freedom Df = N1+N2-2 = 278

Pooled Variance Sp = ((X-mean(X))^2+(Y-mean(Y))^2)/Df

Pooled Variance = 0.96

Standard Error = sqrt(Sp^2/N1 + Sp^2+N2)

Standard Error = 0.14

T= (11.52-10.22)/0.14 = **9.285**

**This value is greater than T-critical hence both the means are statistically significant and hence the null hypothesis is rejected**

This is also verified from the results from http://www.graphpad.com/

**P value and statistical significance:**   
  The two-tailed P value is less than 0.0001  
  By conventional criteria, this difference is considered to be extremely statistically significant.   
  
**Confidence interval:**  
  The mean of Group One minus Group Two equals 1.30000  
  95% confidence interval of this difference: From 1.02371 to 1.57629   
  
**Intermediate values used in calculations:**  
  t = 9.2623  
  df = 278  
  standard error of difference = 0.140