**DATA ANALYSIS METHODS**

**FINAL PROJECT**

Determination of Credit Limit

**Submitted by:**

Harsh Shah

Amanda Lopes

Ken Baker

SUMMARY

As will be seen in the below analysis, 40% of the credit limit of a customer can be explained by the model that we develop in the following report.

Data Source

http://archive.ics.uci.edu/ml/datasets/credit+approval

THE DATA

We are using a data set which determines the credit limit of the clients based on their payment history. There are 23 variables.

X1:Amount of Credit is the response variable. This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. But this will not work in a linear regression. SO instead we chose to use the amount of given credit as the response variable.

Covariates are as follow:

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:

X6 = the repayment status in September, 2005;

X7 = the repayment status in August, 2005 ;

X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar).

X18 = amount paid in September, 2005;

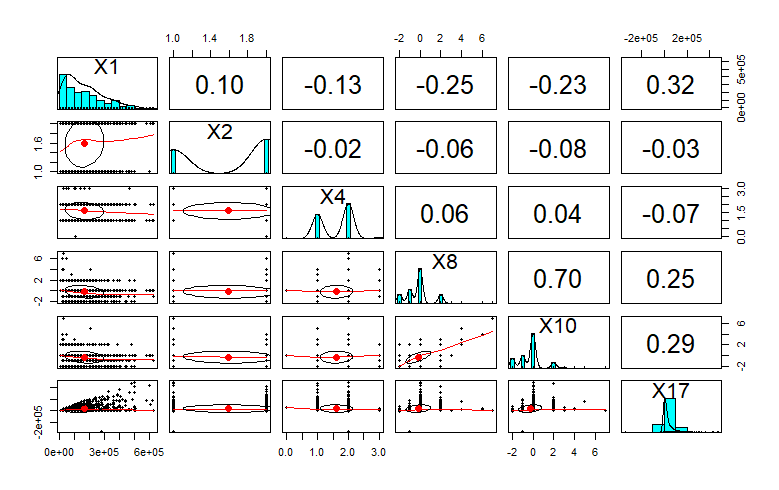
X19 = amount paid in August, 2005;

X23 = amount paid in April, 2005.

PROBLEM STATEMENT

What we are investigating is if there is a relationship between the above-mentioned variables and the credit limit. Below, we will be doing exploratory data analysis and developing a model to help us answer this question.

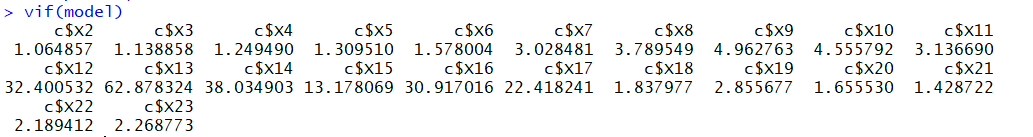
# EXPLORATORY DATA ANALYSIS



The above plot shows the correlation among the different variables and the associated histograms between the particular variables. We can see that most of the variables have normal distribution except for a few.

We can also see the correlation coefficients among different variables. There are a couple of values being very close to zero indicating the relationship is very weak. We will try a few permutations and combinations to come up with a combination that will give a better correlation between the response and independent variables.

Finally, we use the Variance Inflation Factors in determining if multicollinearity is present. These are verified as below:



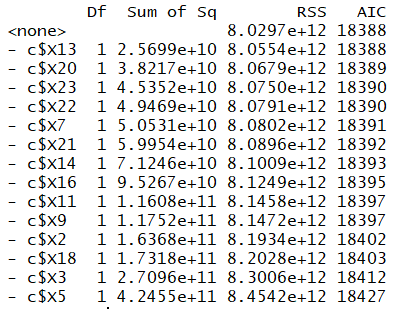
These VIF values measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. We see that Amount of bill statements from April to September (X12-X17) have high VIF values. It means that these regressors may have poorly estimated regression coefficients due to co-relation amongst the predictor variables in the model.

# BUILDING THE MODEL

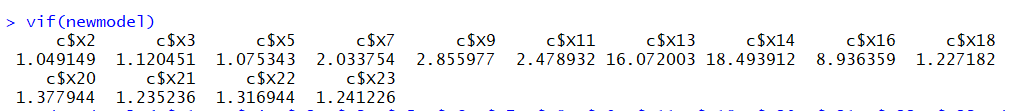
**STEPWISE REGRESSION ANALYSIS**

The first step we aim to study the dependency of other variables on quality, we start with regression of all variables against quality and step by step eliminate those that are insignificant depending on the various values.

Using Akaike Information Criterion (AIC) - The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.



With all other variables on the RHS, we get different values for AIC, based on maximizing the expected entropy of the model. Thus, we eliminate values based on stepwise selection. This reduces the number of regressors of our initial model and we get a final AIC value of 18388. AIC eliminated the following regressors: Gender, Marital status, history of past payment for month of april, repayment status in September, Bill statement and amount of bill payment.

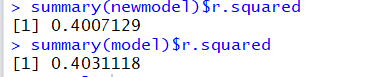


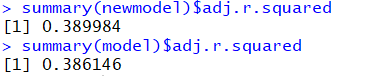
On testing VIF on the reduced model , we find that the number of variables causing correlation has reduced. Thus helping us reduce collinearity.

**ANALYZING OUR MODELS**

WE ran 2 models, we need to rank and analyze them. We are using the test for R2 and adjusted R2. The R2 and adjusted R2 are measures of how well the model explains the response variable. The R2 adjusted however, is needed to partially compensate for the “overfit” of the model, which occurs because of adding more terms would slightly improve the model fit to the given data.

The results of the 2 tests are given below:



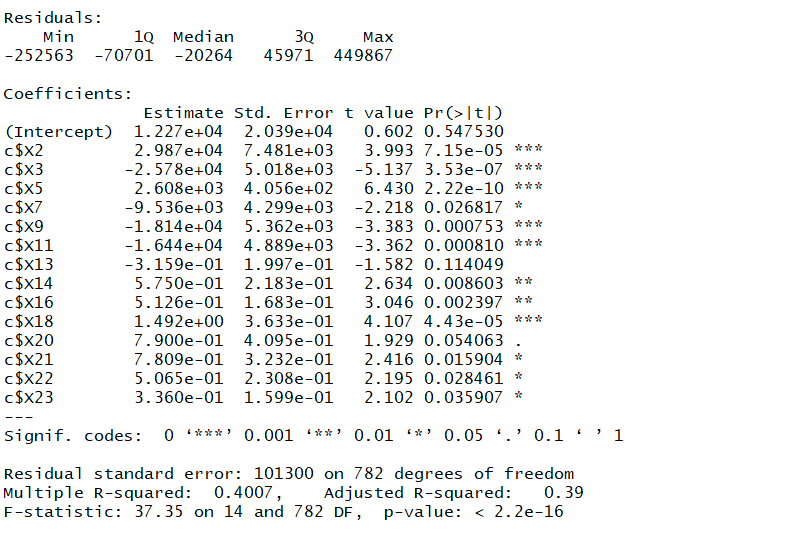


We see that the adjusted R squares for the models are pretty similar in values. newmodel (AIC) has an R2 value of 0.400, which means that 40.07% of variation of the response variable can be explained by the linear regression.

However, both R and R2 are measures of how well the model explains the given data.

**TESING THE SIGNIFICANCE OF REGRESSION**

After selecting our model, the next step is to perform Hypothesis testing on it to test the significance of Regression. In particular, we perform the F test after obtaining the F statistic and the P value. This is done by using the following R code: *summary(newmodel)*



The hypothesis test cases are as below:

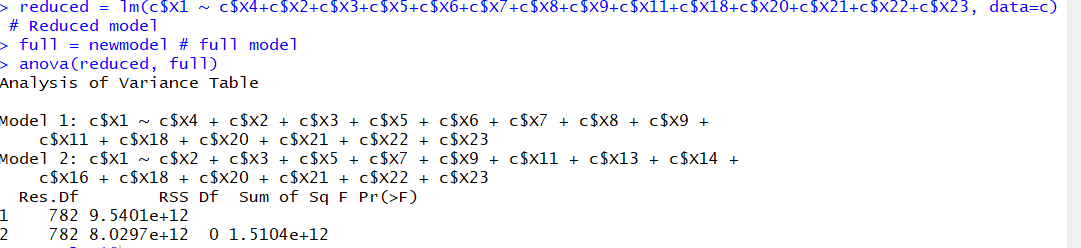
**H0 : β1 = β2 = · · · = βp = 0**

**H1 : βj != 0 for at least one j, j = 1, . . . , p**

We see from the above screen shot that we have an F-statistic of 37.35

We also obtain P-values < 0.05 for all coefficients. This means that we reject the null hypothesis, which again implies that at least one regressor contributes significantly to the model and has a non-zero slope. The linear regression as a whole is significant in explaining the variation in response variable.

Next, we run the partial F test(ANNOVA) on our model. The results are as below:



Testing for Hypothesis:

The partial F test would test the change in the respondent with respect to the subset of the covariate. The result deduced form the partial F test suggest a P value greater than 0.005. Hence, we fail to reject the null hypothesis at the 5% level of confidence. We conclude that the covariates which were not a part of the subset do not contribute significant information on the variation in the response variable, i.e. the wine quality once all the other regressors have been considered.

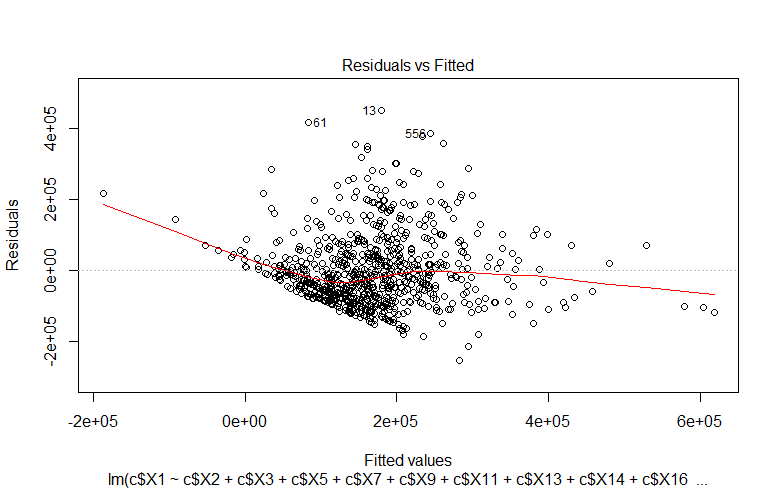
# MODEL ADEQUECY CHECKING

The model assumptions that have been made thus far are as follows:

1. The relationship between response *y* and the regressors is Linear.
2. The errors are Independent.
3. The errors are Normally distributed.
4. Error term, € has Equal variance.

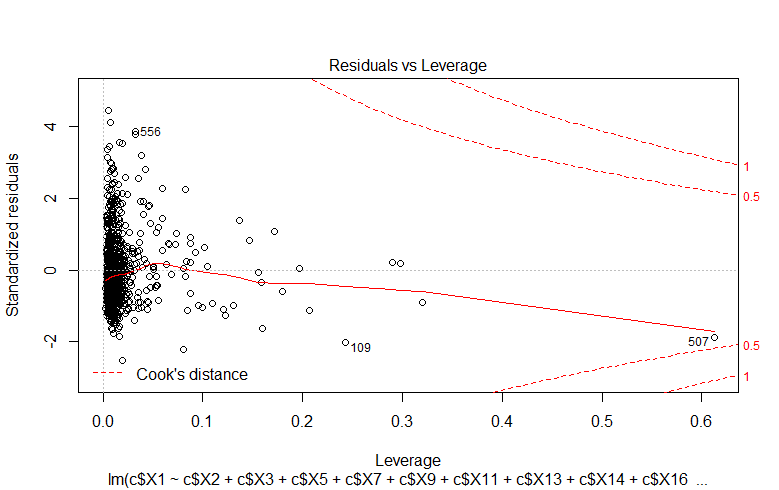
To test these assumptions, we will use the residual plots to investigate the adequacy pf the fit of the regression of a regression model.

**1. Residuals vs Fitted**



The plot on the top right shows if residuals have a non-linear pattern and is a test of Linearity. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn’t capture the non-linear relationship.

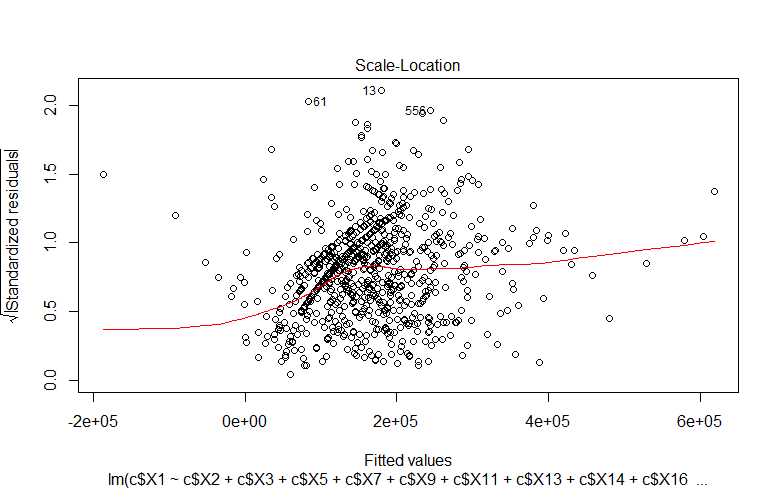
We see that the plot shows Residuals are spread out and have Linearity.

**2. Residuals vs Leverage**

The Residuals VS Leverage plot at the bottom right, is a plot to find influential cases. We look for cases outside of the dashed line ie the Cook’s distance. When cases are outside of the Cook’s distance (meaning they have high Cook’s distance scores), the cases are influential to the regression results.

We find that there is one regressor which is on border line to be an influential cases.

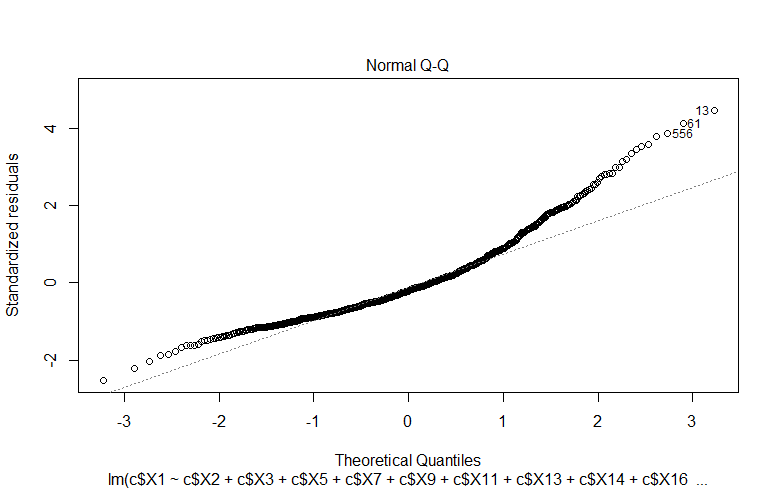
**3. Scale-Location**



This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity).This is the test for equal variance.

As we see, the data is not spread out randomly and thus doesn’t have Equal variance.

**4. Normal Q-Q**



QQ plot is good if residuals are lined well on the straight dashed line. We see that the residuals are pretty well distributed in a straight line, and follow normality closely.

# CONCLUSION

We performed AIC to remove insignificant variables. We plotted the residual plots of Linearity , Normality , independence and equality of the model. Our model can tell 40% of the response variable i.e. the credit limit of each customer based on all the regressors.

References

Google

Lecture PowerPoint