THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

# FACO Score Analysis

## Linear Regression

Olatunji Akinbule

*EMSE-6765 Data Analysis For Engineers And Scientists*

December 4, 2018

*FALL 2018*

## Abstract

Credit scoring is a quintessential appraisal tool for financial institutions including collecting, analyzing and classifying different credit elements in order to make accurate credit decisions.

According to *Abdou, H. & Pointon, J. (2011)* the quality of bank loans is the key determinant of competition, survival and profitability. One of the most important kits, to classify a bank's customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring.

At the heart of our objective is to model credit worthiness based on key determinants in order to transform relevant data into numerical measures that provide an accurate numeric classification (FACO Scores) of the risk involved with providing credit to borrowers.

The provided dataset with 1280 samples containing customer data was non-normal which presented a bit of a challenge because statistical tests such as the F and t tests that will be employed to authenticate the utility of the resultant regression models assume an underlying normality in our distributions. Substantial deviations from normality render such parametric statistical tests inaccurate. As a result the research employed a subset of 108 samples, specifically customers who had mortgages with an employment length between 5 to 7 years.

A comparative assessment of the utility of multiple regression models based on critical assumptions for multiple regression model such as heteroscedasticity, R-squared values and multicollinearity.

**Keywords**: *Multicollinearity, Heteroscedasticity, Durbin-Watson Statistic, R-Squared, Adjusted R-Squared*

## Introduction

Institutions associate the creditworthiness of individuals with a scoring model which is simply a numerical representation of their trustworthiness as borrowers. Lenders use this scoring systems in determining the likelihood of a borrower to repay their loans, and at what interest rate. The higher the credit score, the less risky a borrower and the more likely to secure loans at a favorable interest rate.

According to FICO the percentages of credit rating scores are calculated using the pieces of data grouped in categories as shown below in the diagram.


The Loan Startup Incorporated (**LSI**) provides crowdsourced loans to qualified individuals through an online portal. The objective of this research is to analyze LSI's customer loan application data and create an enhanced and competitive Credit Score Metric that would improve customer satisfaction, increase profitability and present opportunities to expand into related markets using Linear Regression.

The outcome of this research will be developing the optimal predictive model for the customer FACO scores.

# Data and Methods

## Data Description

| Variable | Description |
|---|---|
| **annual_inc** | The annual income provided by the borrower during registration. |
| **delinq_2yrs** | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| **dti** | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income. |
| **emp_length** | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| **home_ownership** | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| **inq_last_6mths** | The number of inquiries by creditors during the past 6 months. |
| **int_rate** | Interest Rate on the loan |
| **loan_amnt** | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| **open_acc** | The number of open credit lines in the borrower's credit file. |
| **revol_bal** | Total credit revolving balance |
| **revol_util** | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| **term** | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| **total_acc** | The total number of credit lines currently in the borrower's credit file |
| **faco** | Borrower's FACO score |

## Data Cleaning

### *Original Data*

| faco | loan_ar | term | int_rate | emp_le | home_ | annual_ | dti | delinq_ | inq_las | open_a | revol_b | revol_u | total_a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 721.9975 | 9600 | 36 month | 7.66 | 6 years | MORTGAG | 120000 | 12.32 | 0 | 3 | 16 | 15953 | 28.7 | 32 |
| 715.1567 | 2000 | 36 month | 7.29 | < 1 year | RENT | 44000 | 6 | 0 | 0 | 5 | 1034 | 68.9 | 6 |
| 730.3597 | 7000 | 60 month | 9.63 | 3 years | MORTGAG | 63600 | 5.83 | 0 | 0 | 8 | 60684 | 35.2 | 29 |
| 757.7201 | 6000 | 36 month | 5.79 | < 1 year | MORTGAG | 120000 | 10.5 | 0 | 2 | 14 | 38190 | 7.6 | 37 |
| 659.8972 | 7200 | 36 month | 14.17 | 10+ years | MORTGAG | 79632 | 12.25 | 1 | 1 | 6 | 2391 | 47.8 | 17 |
| 698.1811 | 2000 | 36 month | 7.29 | 8 years | MORTGAG | 60000 | 20.38 | 0 | 0 | 7 | 16309 | 70.9 | 23 |
| 723.6576 | 5200 | 36 month | 7.29 | 6 years | RENT | 60197 | 18.68 | 0 | 0 | 4 | 14036 | 80.7 | 10 |
| 713.1789 | 20000 | 36 month | 10.74 | 3 years | RENT | 155000 | 4.78 | 0 | 0 | 11 | 37200 | 66.2 | 20 |
| 676.2812 | 8000 | 36 month | 10.37 | 6 years | MORTGAG | 99996 | 7.56 | 0 | 0 | 8 | 22347 | 84.2 | 32 |
| 786.7513 | 3600 | 36 month | 5.42 | 10+ years | MORTGAG | 78000 | 8.54 | 0 | 1 | 6 | 6281 | 17.2 | 23 |
| 678.4863 | 8000 | 36 month | 12.68 | 6 years | MORTGAG | 123000 | 17.45 | 1 | 2 | 10 | 36970 | 83.5 | 21 |
| 670.9862 | 8850 | 36 month | 10.37 | 4 years | MORTGAG | 60000 | 12.12 | 0 | 0 | 7 | 8680 | 49.9 | 10 |
| 779.0993 | 5300 | 36 month | 5.42 | 10+ years | MORTGAG | 50205 | 9.44 | 0 | 2 | 13 | 11456 | 22.4 | 21 |
| 768.0234 | 11000 | 36 month | 5.42 | n/a | OWN | 36000 | 11.37 | 0 | 0 | 14 | 2231 | 9.4 | 23 |

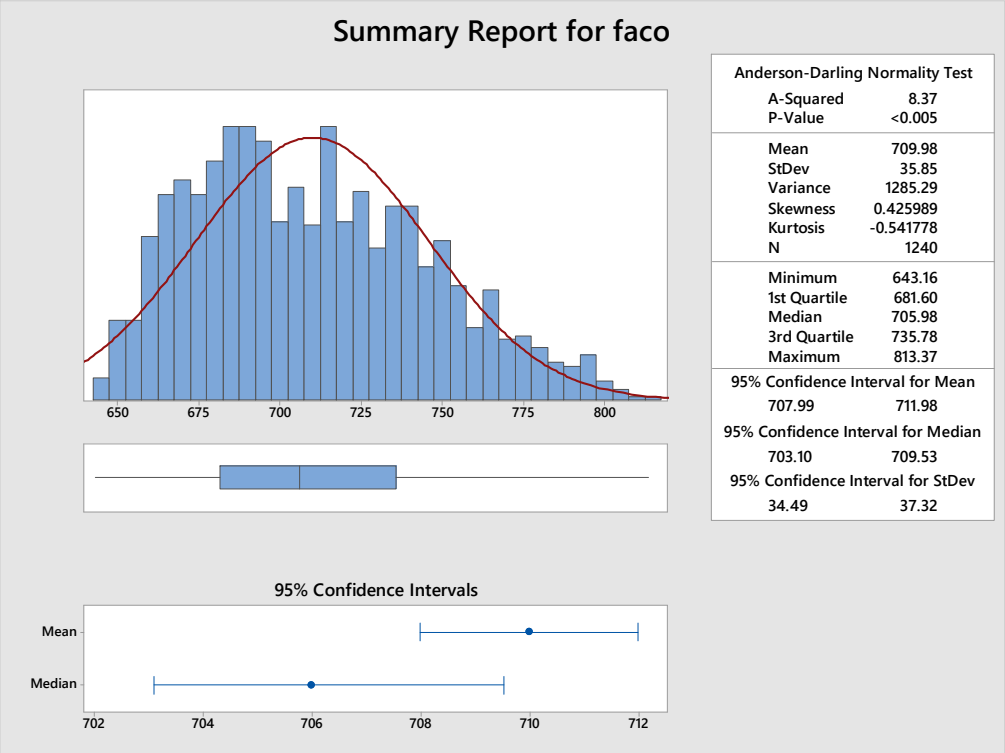
### *PreProcessed Data*

| log(fac | loan_ar | term(m | int_rate | emp(1-4ye | emp(>8yea | RENT | OWN | annual_ | dti | delinq_ | inq_las | open_a | revol_b | revol_u | total_a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.858536 | 9600 | 36 | 7.66 | 0 | 0 | 0 | 0 | 120000 | 12.32 | 0 | 3 | 16 | 15953 | 28.7 | 32 |
| 2.854401 | 2000 | 36 | 7.29 | 1 | 0 | 1 | 0 | 44000 | 6 | 0 | 0 | 5 | 1034 | 68.9 | 6 |
| 2.863537 | 7000 | 60 | 9.63 | 1 | 0 | 0 | 0 | 63600 | 5.83 | 0 | 0 | 8 | 60684 | 35.2 | 29 |
| 2.879509 | 6000 | 36 | 5.79 | 1 | 0 | 0 | 0 | 120000 | 10.5 | 0 | 2 | 14 | 38190 | 7.6 | 37 |
| 2.819476 | 7200 | 36 | 14.17 | 0 | 1 | 0 | 0 | 79632 | 12.25 | 1 | 1 | 6 | 2391 | 47.8 | 17 |
| 2.843968 | 2000 | 36 | 7.29 | 0 | 0 | 0 | 0 | 60000 | 20.38 | 0 | 0 | 7 | 16309 | 70.9 | 23 |
| 2.859533 | 5200 | 36 | 7.29 | 0 | 0 | 1 | 0 | 60197 | 18.68 | 0 | 0 | 4 | 14036 | 80.7 | 10 |
| 2.853198 | 20000 | 36 | 10.74 | 1 | 0 | 1 | 0 | 155000 | 4.78 | 0 | 0 | 11 | 37200 | 66.2 | 20 |
| 2.830127 | 8000 | 36 | 10.37 | 0 | 0 | 0 | 0 | 99996 | 7.56 | 0 | 0 | 8 | 22347 | 84.2 | 32 |
| 2.895837 | 3600 | 36 | 5.42 | 0 | 1 | 0 | 0 | 78000 | 8.54 | 0 | 1 | 6 | 6281 | 17.2 | 23 |
| 2.831541 | 8000 | 36 | 12.68 | 0 | 0 | 0 | 0 | 123000 | 17.45 | 1 | 2 | 10 | 36970 | 83.5 | 21 |
| 2.826714 | 8850 | 36 | 10.37 | 1 | 0 | 0 | 0 | 60000 | 12.12 | 0 | 0 | 7 | 8680 | 49.9 | 10 |
| 2.891593 | 5300 | 36 | 5.42 | 0 | 1 | 0 | 0 | 50205 | 9.44 | 0 | 2 | 13 | 11456 | 22.4 | 21 |

## Variable Coding

| **Variable Coding** | | |
|---|---|---|
| | Dummy Variable 1 | Dummy Variable 2 |
| *Home Ownership* | | |
| Mortgage | 1 | 0 |
| Rent | 0 | 1 |
| Own | 0 | 0 |
| *Employment Length* | | |
| 1-4 years | 1 | 0 |
| 5-7 years | 0 | 1 |
| > 8 years | 0 | 0 |

## Data Exploration




In a check for normality using the Anderson-Darling test, the objective was to assess if the FACO scores followed a normal distribution. The two hypothesis for the Anderson Darling test are shown below:

$H_0$: FACO scores follow the normal distribution

$H_1$: FACO scores do not follow the normal distribution.

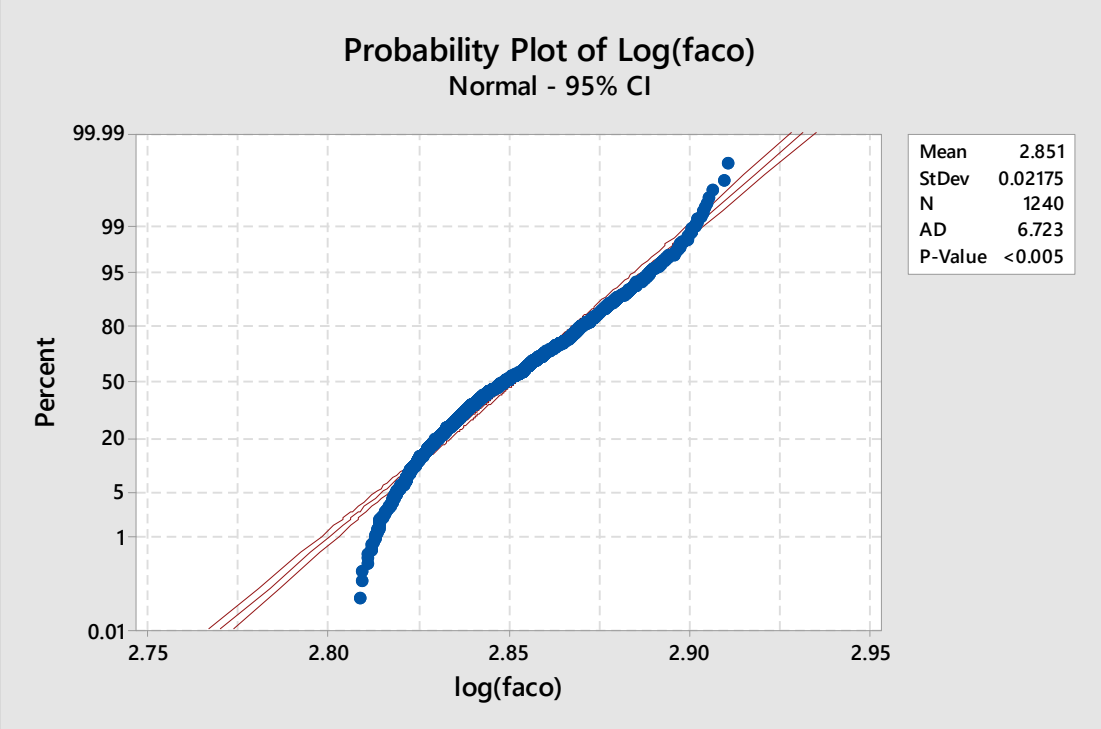
Since the reported probability value is low (< 0.05) i.e the probability of getting a result that is more extreme if the null hypothesis is true, consequently one can reject the null hypothesis and conclude that the FACO scores do not follow a normal distribution. Since non-normality affects the probability of making a wrong decision, whether it be rejecting the null hypothesis when it's true o

A key project requirement is to perform data transformations that will provide more normality or employ a more "normal" subset primarily because the forecasts, confidence intervals yielded by a regression model on a non-normal data may be (at best) inefficient or (at worst) seriously biased or misleading. As a result some data transformations and or exploration of the dataset that will satisfy the normality assumption will be the next step.

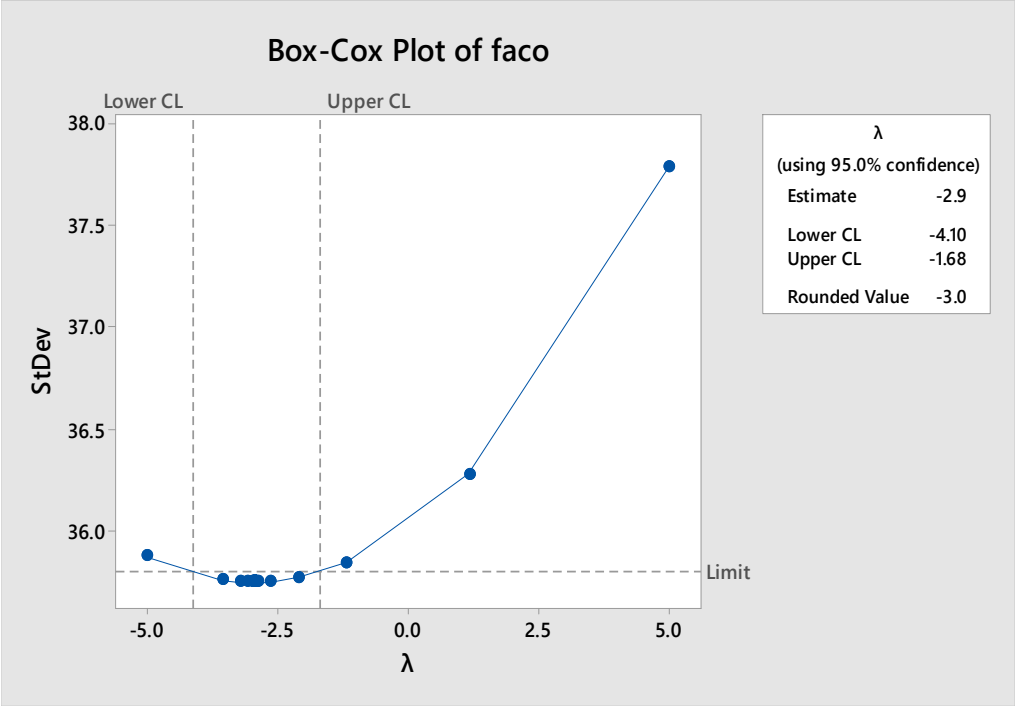
## Data Transformations

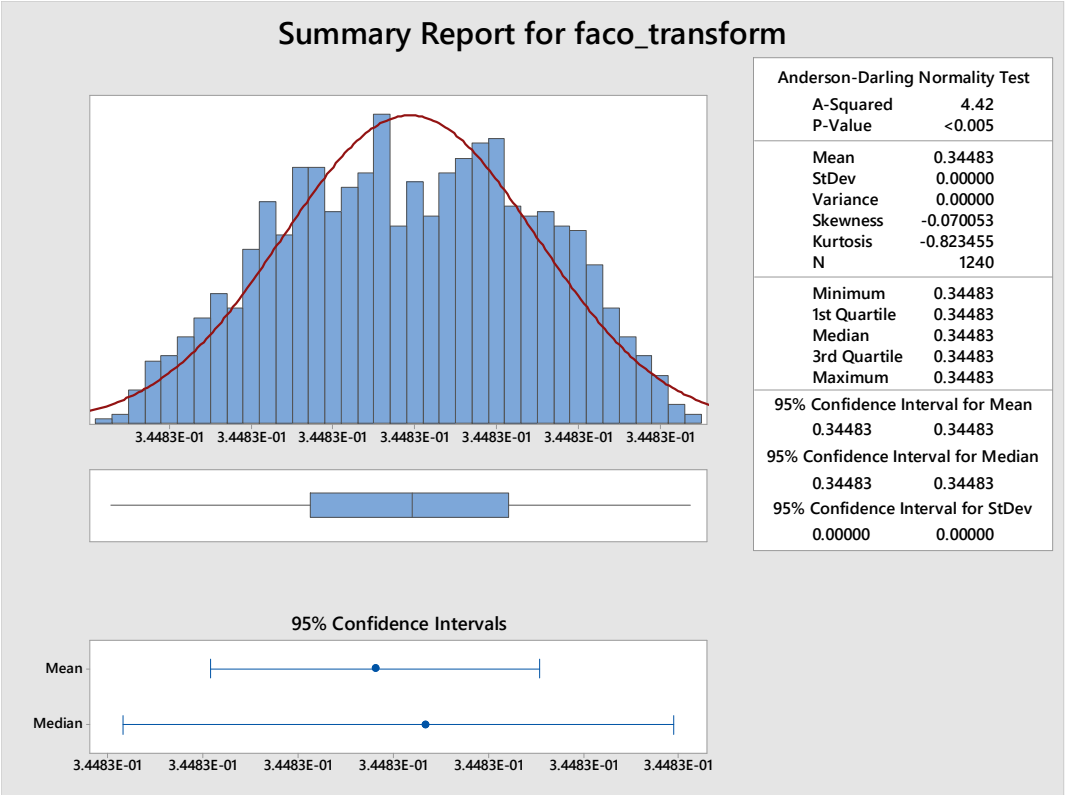
In many cases where the data do not fit a normal distribution, transformations exist that will make the data "more normal". The log transformation (or the Box-Cox power transformation) is very effective for skewed data. The arcsin transformation can be used for binomial proportions.

Probability Plot of Log(faco)



## Box –Cox Transformation

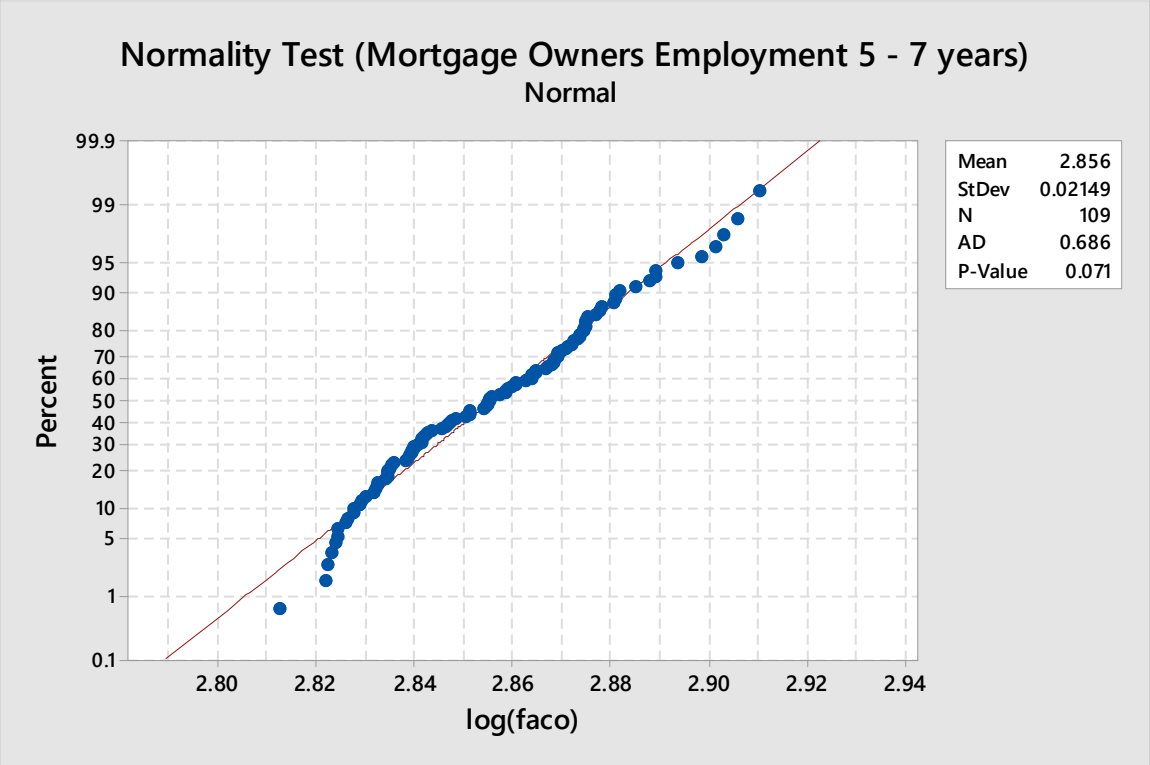


Although my attempt to reduce non-normality of the dataset using the Box Cox transformation provided a reduced A-square value from 8.37 to 4.42, the transformation did not produce normality of the dataset remains insignificant at a level of 0.05.

## Data Subset Selection

From a thorough data exploration, a subset of the provided population – individuals who had mortgages and had worked between 5-7 years- provided a sample size with a more normal distribution having a P-Value of 7% and as such cannot reject the null hypothesis that this subset is normal. We don't have sufficient evidence to reject the null hypothesis at an alpha level of 0.05.



# Model Development

## Variable Selection

### Correlation Analysis

Correlation Matrix

In order to obtain a general idea of what the optimal variable selection of our model composition, I developed a correlation matrix based on the original data features.

In the Table below the yellow highlighted cells represent correlations with absolute values larger than 0.15, an arbitrarily selected threshold indicating variables with a sufficiently substantial correlation with the dependent variable log (faco) for consideration in developing the model.

The int_rate, annual_income, dti, delinq_2yr, revol_bal and revol_utility have relatively higher correlations with log(faco) and so they are highlighted in orange.

In addition, the substantial inter-correlation values between some of the prospective variables such as revol_util and int_rate(0.5400) which might be an indication of potential multi-collinearity in regression.

| | *log(faco)* | *oan_amnt* | *rm(month* | *int_rate* | *nnual_inc* | *dti* | *elinq_2yrs* | *_last_6mt* | *open_acc* | *revol_bal* | *revol_util* | *total_acc* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| log(faco) | 1 | | | | | | | | | | | |
| loan_amn | 0.028239 | 1 | | | | | | | | | | |
| term(mon | -0.10005 | 0.260432 | 1 | | | | | | | | | |
| int_rate | -0.59815 | 0.515068 | 0.581615 | 1 | | | | | | | | |
| annual_in | 0.16898 | 0.299643 | -0.11372 | 0.027208 | 1 | | | | | | | |
| dti | -0.25818 | -0.01241 | 0.105389 | 0.169073 | -0.20279 | 1 | | | | | | |
| delinq_2y | -0.2637 | -0.001 | 0.065409 | 0.158072 | 0.010698 | 0.014789 | 1 | | | | | |
| inq_last_6 | -0.01927 | -0.13244 | -0.01798 | 0.115662 | -0.06038 | 0.003112 | 0.087185 | 1 | | | | |
| open_acc | 0.004927 | 0.12747 | -0.01542 | -0.00114 | 0.284522 | 0.298986 | -0.01566 | -0.04218 | 1 | | | |
| revol_bal | -0.25087 | 0.35344 | 0.132471 | 0.312145 | 0.410351 | 0.238258 | 0.055614 | -0.16661 | 0.173221 | 1 | | |
| revol_util | -0.65336 | 0.114107 | 0.133215 | 0.540016 | -0.09123 | 0.296447 | 0.116537 | -0.08355 | -0.14879 | 0.467378 | 1 | |
| total_acc | 0.114263 | 0.170752 | 0.080859 | 0.011658 | 0.222723 | 0.264166 | 0.006569 | 0.068115 | 0.755152 | 0.187616 | -0.07727 | 1 |
| | | | | | | | | | | | | |
| | | | | | Threshold | 0.15 | | | | | | |

Interactions

From my assessment of the data description, I am exploring the existence of interactions among some of these variables and their significance, if any, to the development of a robust model. For example, is there a non-trivial interaction between the credit utilization rate (revol_util) and the reported annual income?

As a result, I generated additional variables in Table X below from a select combination of provided variables to represent interactions.

| | log(faco) | loan_amnt | rm(month | int_rate | nnual_inc | dti | elinq_2yrs | _last_6mt | open_acc | revol_bal | revol_util | total_acc | nnual_inc | _util*revol | l_bal*int_ | _bal*open | nnual_inc | r*inq_last | pen_acc*d | nt*annual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| log(faco) | 1 | | | | | | | | | | | | | | | | | | | |
| loan_amnt | 0.028239 | 1 | | | | | | | | | | | | | | | | | | |
| term(months) | -0.10005 | 0.260432 | 1 | | | | | | | | | | | | | | | | | |
| int_rate | -0.59815 | 0.515068 | 0.581615 | 1 | | | | | | | | | | | | | | | | |
| annual_inc | 0.16898 | 0.299643 | -0.11372 | 0.027208 | 1 | | | | | | | | | | | | | | | |
| dti | -0.25818 | -0.01241 | 0.105389 | 0.169073 | -0.20279 | 1 | | | | | | | | | | | | | | |
| delinq_2yrs | -0.2637 | -0.001 | 0.065409 | 0.158072 | 0.010698 | 0.014789 | 1 | | | | | | | | | | | | | |
| inq_last_6mths | -0.01927 | -0.13244 | -0.01798 | 0.115662 | -0.06038 | 0.003112 | 0.087185 | 1 | | | | | | | | | | | | |
| open_acc | 0.004927 | 0.12747 | -0.01542 | -0.00114 | 0.284522 | 0.298986 | -0.01566 | -0.04218 | 1 | | | | | | | | | | | |
| revol_bal | -0.25087 | 0.35344 | 0.132471 | 0.312145 | 0.410351 | 0.238258 | 0.055614 | -0.16661 | 0.173221 | 1 | | | | | | | | | | |
| revol_util | -0.65336 | 0.114107 | 0.133215 | 0.540016 | -0.09123 | 0.296447 | 0.116537 | -0.08355 | -0.14879 | 0.467378 | 1 | | | | | | | | | |
| total_acc | 0.114263 | 0.170752 | 0.080859 | 0.011658 | 0.222723 | 0.264166 | 0.006569 | 0.068115 | 0.755152 | 0.187616 | -0.07727 | 1 | | | | | | | | |
| dti*annual_income | -0.08592 | 0.279129 | -0.00285 | 0.158583 | 0.628906 | 0.555655 | 0.034968 | -0.07306 | 0.50327 | 0.511706 | 0.165762 | 0.385395 | 1 | | | | | | | |
| revol_util*revol_bal | -0.32779 | 0.314273 | 0.088993 | 0.355922 | 0.381006 | 0.205806 | 0.030065 | -0.14489 | 0.041375 | 0.941457 | 0.58495 | 0.067977 | 0.447856 | 1 | | | | | | |
| revol_bal*int_rate | -0.30068 | 0.444808 | 0.231218 | 0.467774 | 0.410695 | 0.182634 | 0.053034 | -0.11811 | 0.120612 | 0.959738 | 0.47927 | 0.130078 | 0.45839 | 0.948752 | 1 | | | | | |
| revol_bal*open_acc | -0.18472 | 0.335103 | 0.118057 | 0.262006 | 0.475848 | 0.264697 | 0.047185 | -0.14138 | 0.453676 | 0.917998 | 0.351994 | 0.411472 | 0.591966 | 0.806148 | 0.851546 | 1 | | | | |
| dti*annual_income | -0.08592 | 0.279129 | -0.00285 | 0.158583 | 0.628906 | 0.555655 | 0.034968 | -0.07306 | 0.50327 | 0.511706 | 0.165762 | 0.385395 | 1 | 0.447856 | 0.45839 | 0.591966 | 1 | | | |
| delinq_2yr*inq_last_6mnths | -0.22745 | -0.04339 | 0.008346 | 0.162409 | -0.03224 | -0.01016 | 0.789762 | 0.23109 | -0.05917 | -0.06078 | 0.128235 | 0.004408 | -0.03633 | -0.02582 | -0.03791 | -0.0668 | -0.03633 | 1 | | |
| open_acc*dti | -0.1695 | 0.027021 | 0.07827 | 0.109104 | 0.013036 | 0.825584 | 0.001038 | -0.02544 | 0.745382 | 0.243809 | 0.121511 | 0.570695 | 0.638698 | 0.147028 | 0.176767 | 0.428872 | 0.638698 | -0.02437 | 1 | |
| loan_amnt*annual_income | 0.080646 | 0.746515 | 0.077583 | 0.339103 | 0.777839 | -0.10483 | 0.009401 | -0.02713 | 0.238022 | 0.512147 | 0.040237 | 0.240882 | 0.552832 | 0.476566 | 0.570269 | 0.551661 | 0.552832 | -0.03294 | 0.03034 | 1 |
| | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | Threshold | 0.15 | | | | | | | | | | | |

Fig Correlation Matrix

Based on the correlation matrix on an arbitrary threshold value chosen in order to select the explanatory variables that account for the greatest observed change from the list of dependent variables. A threshold value of 0.15 selects 6 predictor variables. These 6 variables demonstrate significant correlation to the dependent variable log(faco) which provides a good starting point for this regression analysis.

From Fig 5. The selected variables are interest rate, annual_income, dti, delinq_2years, revol_util and revol_bal.

Based on the Correlation Matrix I selected the following as explanatory variables:

*int_rate, home_ownership_MORTGAGE, home_ownership_RENT, dti, delinq_2yrs, revol_util, total_acc*

# Linear Regression

## Model Fitting

**Equation**

$$y = A\mathbf{X}^T + e$$

Keys

Description

| Variable | Description | Dimension |
|---|---|---|
| y | Dependent variable matrix | 1 x 1 |
| A | Coefficient Matrix | 1 x n |
| X | Independent variable matrix | 1 x n |
| e | Error term | 1 x 1 |
| n | Number of dependent variables | |

**Model Selection and Diagnosis Analysis.**

Models are selected based on the following criteria:

a. R-squared and adjusted R-squared values of the model to examine the variances coverage of the model;
b. Standard Error for coefficients to measure the precision of the coefficient predications
c. P-values for the coefficients to test the statistical significances for coefficients
d. Normality for residuals from the model

Subsequently models will be adjusted and selected based on diagnosis analysis:

a. Multicollinearity measured by the Variance inflation Factor (VIF) for coefficients
b. Heteroscedasticity measured by the residual vs fitted values (RVF)
c. Durbin-Watson Statistic (DW) for auto-correlation
d. Normality for X
e. Independence estimated by residual vs observation order (RVO) plot;

Models

After cautiously examining a plethora of variable combinations, a model with an R-sq of 72% and an adjRSq of 70% is selected. The independent variable matrix is represented below.

$$X^T = \begin{bmatrix} revol\_util \\ int\_rate \\ annual\_inc \\ open\_acc \\ total\_acc \\ loan\_amnt \\ delinq\_2yrs \\ inq\_last6\_mnths \\ openacc \\ term(months) \end{bmatrix}$$

Model 1 (Chosen)

**Analysis of Variance**

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 8 | 0.036012 | 0.004501 | 32.50 | 0.000 |
| loan_amnt | 1 | 0.004849 | 0.004849 | 35.01 | 0.000 |
| term(months) | 1 | 0.003782 | 0.003782 | 27.31 | 0.000 |
| int_rate | 1 | 0.011064 | 0.011064 | 79.89 | 0.000 |
| annual_inc | 1 | 0.000689 | 0.000689 | 4.98 | 0.028 |
| delinq_2yrs | 1 | 0.000850 | 0.000850 | 6.14 | 0.015 |
| inq_last_6mths | 1 | 0.000811 | 0.000811 | 5.85 | 0.017 |
| open_acc | 1 | 0.000572 | 0.000572 | 4.13 | 0.045 |
| revol_util | 1 | 0.001662 | 0.001662 | 12.00 | 0.001 |
| Error | 100 | 0.013850 | 0.000138 | | |
| Total | 108 | 0.049862 | | | |

**Model Summary**

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.0117685 | 72.22% | 70.00% | 65.69% |

**Coefficients**

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 2.87412 | 0.00634 | 453.24 | 0.000 | |
| loan_amnt | 0.000001 | 0.000000 | 5.92 | 0.000 | 1.75 |
| term(months) | 0.000667 | 0.000128 | 5.23 | 0.000 | 1.78 |
| int_rate | -0.004796 | 0.000537 | -8.94 | 0.000 | 3.57 |
| annual_inc | 0.000000 | 0.000000 | 2.23 | 0.028 | 1.24 |
| delinq_2yrs | -0.00739 | 0.00298 | -2.48 | 0.015 | 1.04 |
| inq_last_6mths | 0.002360 | 0.000976 | 2.42 | 0.017 | 1.19 |
| open_acc | -0.000638 | 0.000314 | -2.03 | 0.045 | 1.12 |
| revol_util | -0.000198 | 0.000057 | -3.46 | 0.001 | 1.83 |

**Fits and Diagnostics for Unusual Observations**

| Obs | log(faco) | Fit | Resid | Std Resid | | |
|---|---|---|---|---|---|---|
| 43 | 2.88170 | 2.85817 | 0.02352 | 2.07 | R | |
| 47 | 2.89826 | 2.87354 | 0.02472 | 2.22 | R | |
| 65 | 2.84235 | 2.83882 | 0.00352 | 0.36 | | X |
| 73 | 2.86882 | 2.85509 | 0.01373 | 1.36 | | X |
| 85 | 2.90137 | 2.87603 | 0.02534 | 2.32 | R | |
| 87 | 2.90276 | 2.87901 | 0.02374 | 2.06 | R | |
| 90 | 2.83434 | 2.85506 | -0.02073 | -2.13 | R | X |
| 91 | 2.91029 | 2.87965 | 0.03064 | 2.75 | R | |
| 99 | 2.82418 | 2.82110 | 0.00308 | 0.30 | | X |

*R Large residual*
*X Unusual X*

## Durbin-Watson Statistic

Durbin-Watson Statistic = 1.77443

## Regression Equation

log(faco) = 2.87412 + 0.000001 loan_amnt + 0.000667 term(months) - 0.004796 int_rate
+ 0.000000 annual_inc - 0.00739 delinq_2yrs + 0.002360 inq_last_6mths
- 0.000638 open_acc - 0.000198 revol_util

The observations from the initial regression results above are as follows:

- The variance inflation (VIF) values are obtained from Minitab for each independent variable are mostly within the range of 1 except for int_rate which is 3.75 indicating the predictors may be correlated.
- The Durbin-Watson statistic is 1.77443, indicating the presence of positive auto correlation. I would prefer a value much closer to 2 which indicates no auto correlation in the sample
- From Table above showing the *Analysis of Variances* the least point estimates for all coefficient are relevant based on the statistical significance of the p-values for the t-statistics of each variable at an alpha level of 0.05. However, it is noteworthy that the coefficient of the annual_inc is 0 despite being statistically significant. I suspect this is a Minitab approximation of what might be a very low coefficient value as removing the feature substantially reduced the $R^2$ values.
- Also, the standard error the coefficients for each for the selected independent variables are precise estimates less than an alpha level of 0.05 leading to the conclusion that the coefficient is significantly different from 0.
- The residual analysis appears to support the assumption of normality for the analysis
- The normal probability plot of the residuals shows some deviation from normality. However, these deviations do not invalidate the assumption of normality for the residuals with a P-value of 0.243 which is statistically significant even at 0.05.
- There is no apparent heteroscedasticity in the plot of the residual versus fitted values for Log (Faco) lending credence to constant variance in residuals.

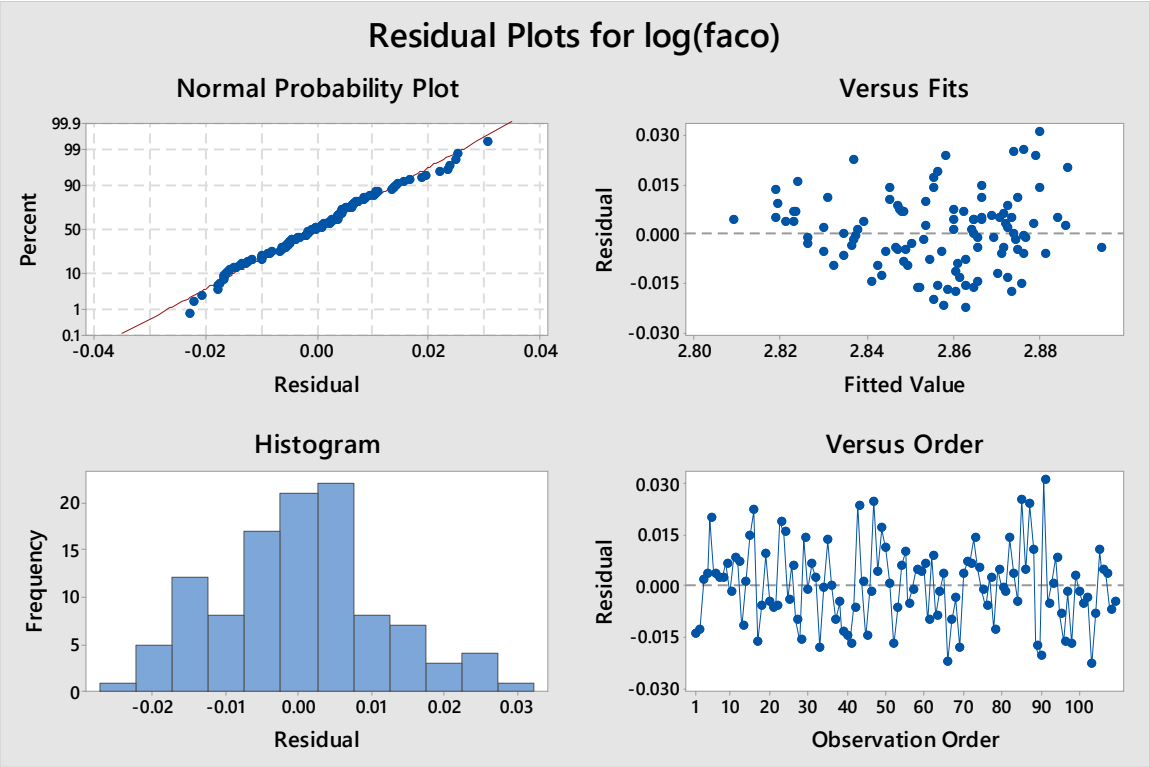
The results so far show that this is an auspicious model to start with and is the initial model chosen in this research

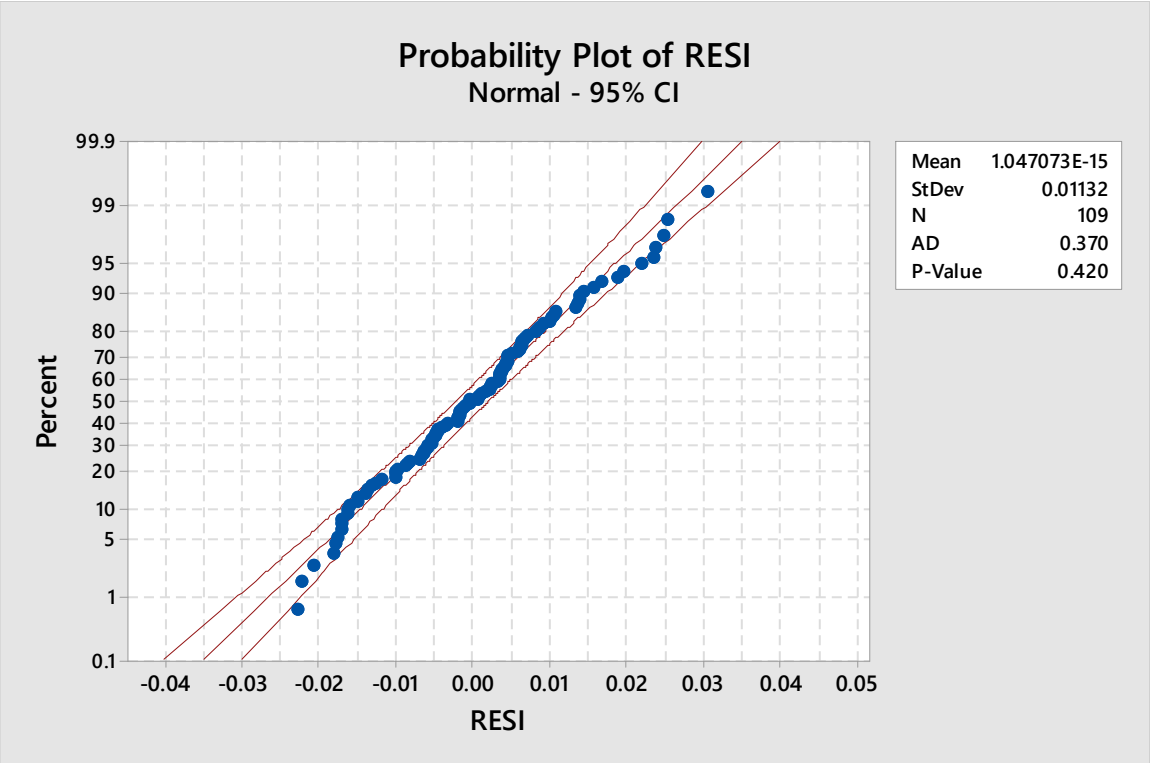
Next Steps:

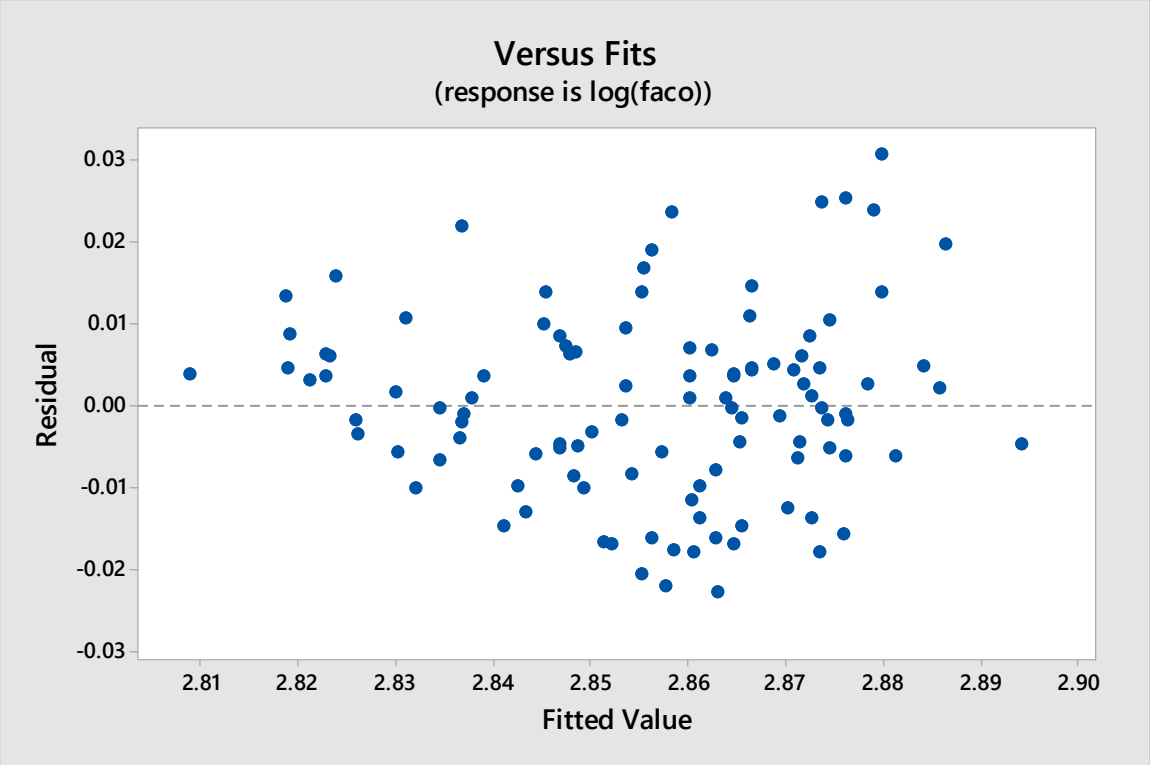
Eliminate explanatory variables to cater for possible multicollinearity from the VIF values

Look out for possible adjustments to a higher Durbin-Watson statistic

The following figures support the observations listed above:






Diagnostic Analysis – Model 1:

The Analysis of the initial regression model indicates that the model described in the regression equation is within reason.

The analysis of the residuals versus fitted values indicates that the majority of the values fall within expected thresholds. Only two on the extreme appear suspicious but these are not enough to invalidate the model.

Based on the likely logical interaction between the number of open credit lines in the borrower's file and the total number of accounts. The research proceeds to explore any feasibility in the validity of this interaction.

The following table displays the new independent variable along with the other independent variables that comprise the regression model which will be referred to as the adjusted model

Model 2

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 6 | 0.034993 | 0.005832 | 40.01 | 0.000 |
| loan_amnt | 1 | 0.006124 | 0.006124 | 42.01 | 0.000 |
| term(months) | 1 | 0.003514 | 0.003514 | 24.11 | 0.000 |
| int_rate | 1 | 0.011287 | 0.011287 | 77.43 | 0.000 |
| delinq_2yrs | 1 | 0.000794 | 0.000794 | 5.45 | 0.022 |
| inq_last_6mths | 1 | 0.000836 | 0.000836 | 5.73 | 0.018 |
| revol_util | 1 | 0.001621 | 0.001621 | 11.12 | 0.001 |
| Error | 102 | 0.014868 | 0.000146 | | |
| Total | 108 | 0.049862 | | | |

## Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.0120734 | 70.18% | 68.43% | 66.19% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 2.87324 | 0.00510 | 562.83 | 0.000 | |
| loan_amnt | 0.000001 | 0.000000 | 6.48 | 0.000 | 1.61 |
| term(months) | 0.000633 | 0.000129 | 4.91 | 0.000 | 1.72 |
| int_rate | -0.004834 | 0.000549 | -8.80 | 0.000 | 3.55 |
| delinq_2yrs | -0.00714 | 0.00306 | -2.33 | 0.022 | 1.04 |
| inq_last_6mths | 0.002392 | 0.000999 | 2.39 | 0.018 | 1.18 |
| revol_util | -0.000192 | 0.000058 | -3.33 | 0.001 | 1.77 |

## Fits and Diagnostics for Unusual Observations

| Obs | log(faco) | Fit | Resid | Std Resid | | |
|---|---|---|---|---|---|---|
| 5 | 2.90584 | 2.87587 | 0.02997 | 2.55 | R | |
| 43 | 2.88170 | 2.85419 | 0.02751 | 2.33 | R | |
| 65 | 2.84235 | 2.83930 | 0.00305 | 0.30 | | X |
| 66 | 2.83524 | 2.86049 | -0.02526 | -2.12 | R | |
| 73 | 2.86882 | 2.85479 | 0.01403 | 1.36 | | X |
| 85 | 2.90137 | 2.87011 | 0.03126 | 2.73 | R | |
| 91 | 2.91029 | 2.87596 | 0.03432 | 2.90 | R | |
| 99 | 2.82418 | 2.82397 | 0.00021 | 0.02 | | X |

*R Large residual*
*X Unusual X*

## Durbin-Watson Statistic

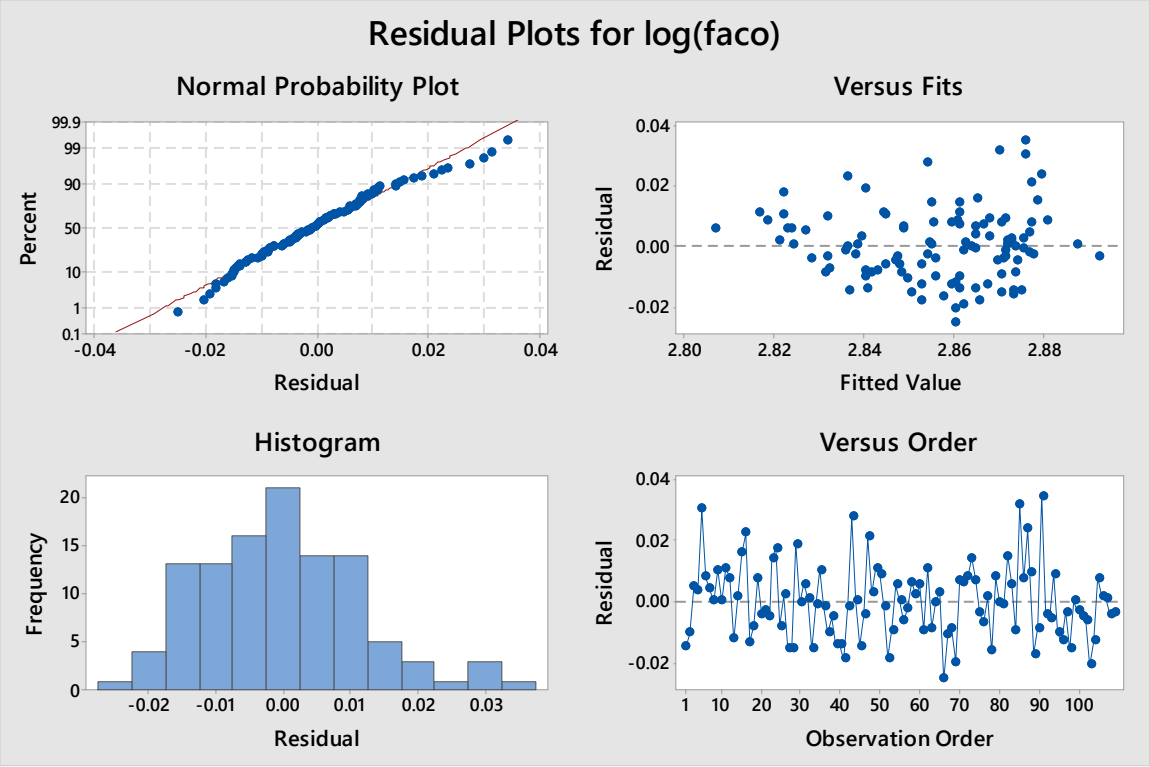
Durbin-Watson Statistic = 1.71976

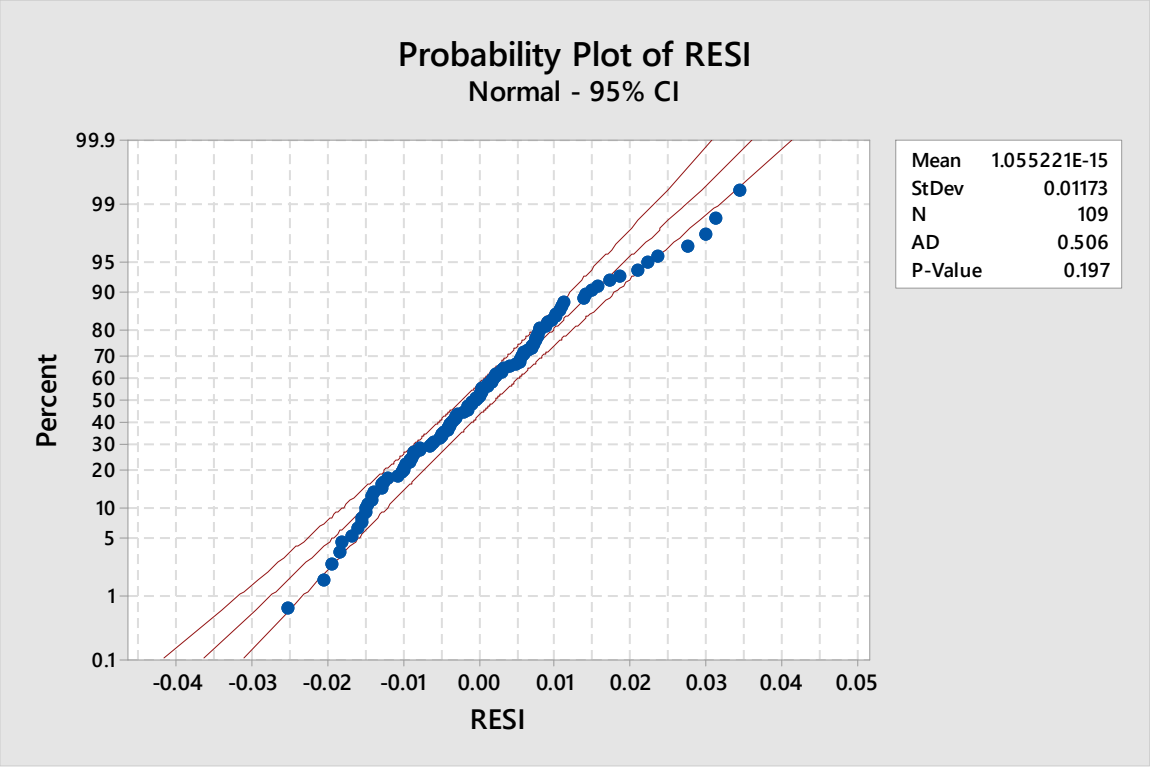
## Regression Equation

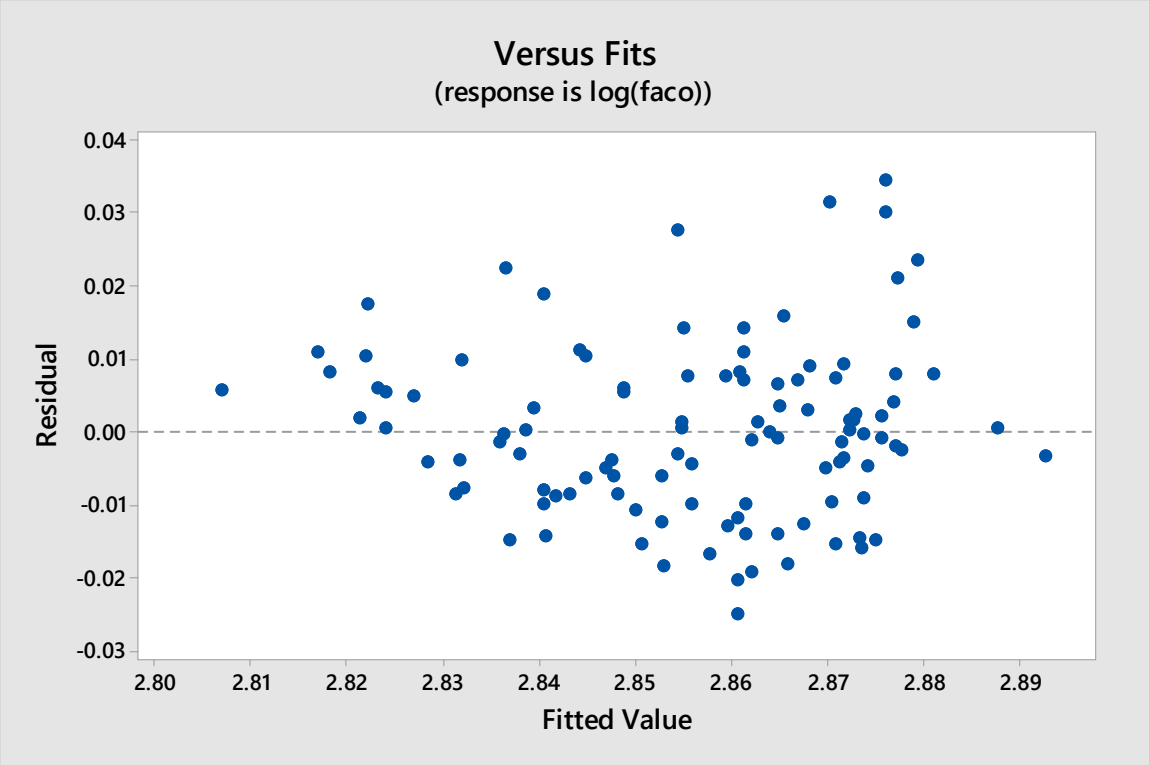
log(faco) = 2.87324 + 0.000001 loan_amnt + 0.000633 term(months) - 0.004834 int_rate
- 0.00714 delinq_2yrs + 0.002392 inq_last_6mths - 0.000192 revol_util

The observations from the Model 2 regression results above are as follows:

- The R-squared value has reduced from 72.22% to 70.18%
- The Adjusted R-squared value has reduced from 70% to 68.43%
- The variance inflation (VIF) values are obtained from Minitab for each independent variable are mostly within the range of 1 except for int_rate which is 3.55 indicating the predictors may be correlated.
- The Durbin-Watson statistic drops from 1.77443 to 1.71976, indicating a higher presence of positive auto correlation. I would prefer an increase to a value much closer to 2 which indicates no auto correlation in the sample
- From Table above showing the *Analysis of Variances* the least point estimates for all coefficient are relevant based on the statistical significance of the p-values for the t-statistics of each variable at an alpha level of 0.1. An anomaly worthy of note is the (0.0000) coefficient of the annual income (annual_inc) despite the statistical relevance of its Pvalues. This is only an approximation of the infinitesimal value (6.88 E -8) by Minitab.
- In addition, the standard error the coefficients for each for the selected independent variables are precise estimates less than an alpha level of 0.05 leading to the conclusion that the effect of the coefficients is significantly different from 0.
- The residual analysis appears to support the assumption of normality for the analysis
- The normal probability plot of the residuals shows some deviation from normality. However these deviations do not invalidate the assumption of normality for the residuals with a P-value of 0.197 which is statistically significant even at 0.05.
- There is no apparent heteroscedasticity in the plot of the residual versus fitted values for Log (Faco) lending credence to constant variance in residuals.


Residual Plots for log(faco)
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order


Probability Plot of RESI
Normal - 95% CI
Mean 1.055221E-15
StDev 0.01173
N 109
AD 0.506
P-Value 0.197
Percent
RESI


Versus Fits
(response is log(faco))
Residual
Fitted Value
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
2.80
2.81
2.82
2.83
2.84
2.85
2.86
2.87
2.88
2.89

Model 3

$$X^T = \begin{bmatrix} revol\_util \\ int\_rate \\ annual\_inc \\ open\_acc \\ loan\_amnt \\ delinq\_2yrs \\ inq\_last6\_mnths \\ openacc * totalacc \end{bmatrix}$$

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 6 | 0.030509 | 0.005085 | 26.80 | 0.000 |
| int_rate | 1 | 0.004209 | 0.004209 | 22.18 | 0.000 |
| annual_inc | 1 | 0.001337 | 0.001337 | 7.05 | 0.009 |
| delinq_2yrs | 1 | 0.001120 | 0.001120 | 5.90 | 0.017 |
| open_acc | 1 | 0.002802 | 0.002802 | 14.77 | 0.000 |
| revol_util | 1 | 0.006822 | 0.006822 | 35.95 | 0.000 |
| open_acc * total_acc | 1 | 0.002215 | 0.002215 | 11.67 | 0.001 |
| Error | 102 | 0.019353 | 0.000190 | | |
| Total | 108 | 0.049862 | | | |

## Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.0137743 | 61.19% | 58.90% | 54.17% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 2.90654 | 0.00677 | 429.55 | 0.000 | |
| int_rate | -0.001892 | 0.000402 | -4.71 | 0.000 | 1.46 |
| annual_inc | 0.000000 | 0.000000 | 2.65 | 0.009 | 1.10 |
| delinq_2yrs | -0.00844 | 0.00347 | -2.43 | 0.017 | 1.03 |
| open_acc | -0.002951 | 0.000768 | -3.84 | 0.000 | 4.87 |
| revol_util | -0.000360 | 0.000060 | -6.00 | 0.000 | 1.47 |
| open_acc * total_acc | 0.000042 | 0.000012 | 3.42 | 0.001 | 4.79 |

## Fits and Diagnostics for Unusual Observations

| Obs | log(faco) | Fit | Resid | Std Resid | | |
|---|---|---|---|---|---|---|
| 17 | 2.84637 | 2.87436 | -0.02800 | -2.18 | R | |
| 19 | 2.86271 | 2.83592 | 0.02679 | 2.02 | R | |
| 47 | 2.89826 | 2.86613 | 0.03213 | 2.40 | R | |
| 49 | 2.87191 | 2.87571 | -0.00380 | -0.32 | | X |
| 50 | 2.87694 | 2.84968 | 0.02727 | 2.07 | R | |
| 55 | 2.85506 | 2.86271 | -0.00766 | -0.65 | | X |
| 65 | 2.84235 | 2.82000 | 0.02235 | 1.87 | | X |
| 90 | 2.83434 | 2.85669 | -0.02236 | -1.94 | | X |
| 98 | 2.83500 | 2.86485 | -0.02985 | -2.24 | R | |
| 99 | 2.82418 | 2.83359 | -0.00941 | -0.78 | | X |

*R Large residual*
*X Unusual X*

## Durbin-Watson Statistic

Durbin-Watson Statistic = 1.85645

## Regression Equation

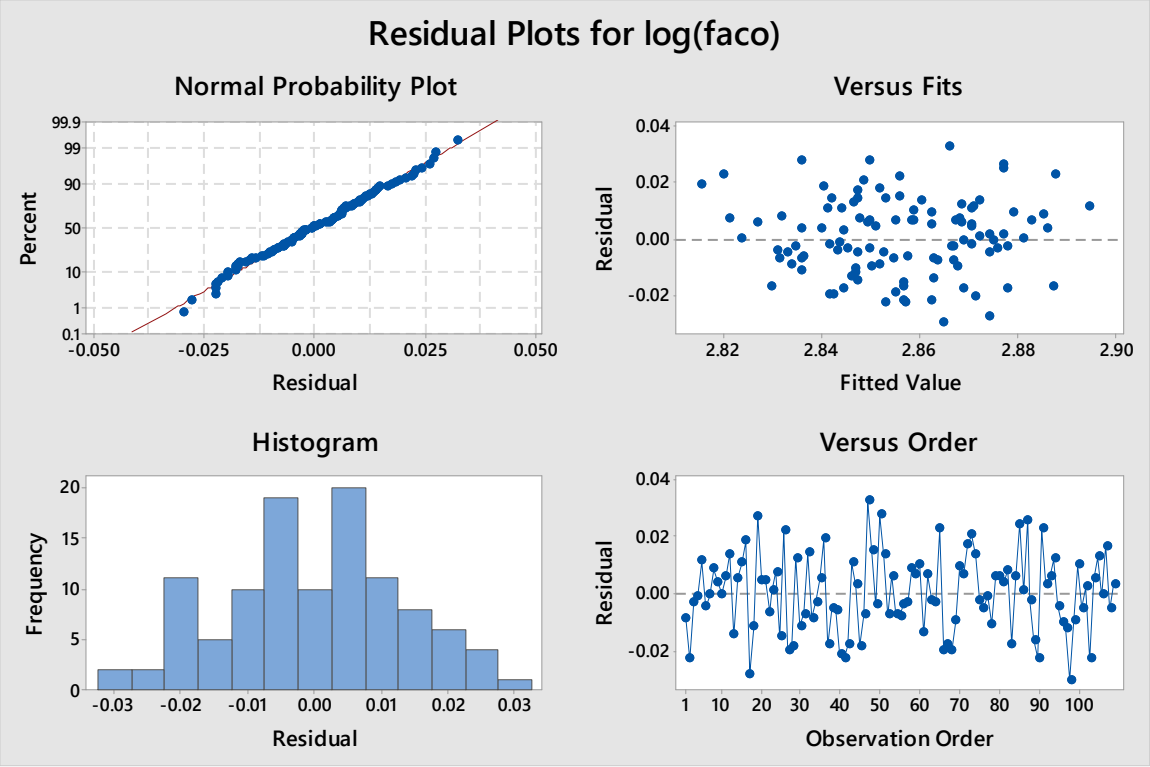
log(faco) = 2.90654 - 0.001892 int_rate + 0.000000 annual_inc - 0.00844 delinq_2yrs
- 0.002951 open_acc - 0.000360 revol_util + 0.000042 open_acc * total_acc

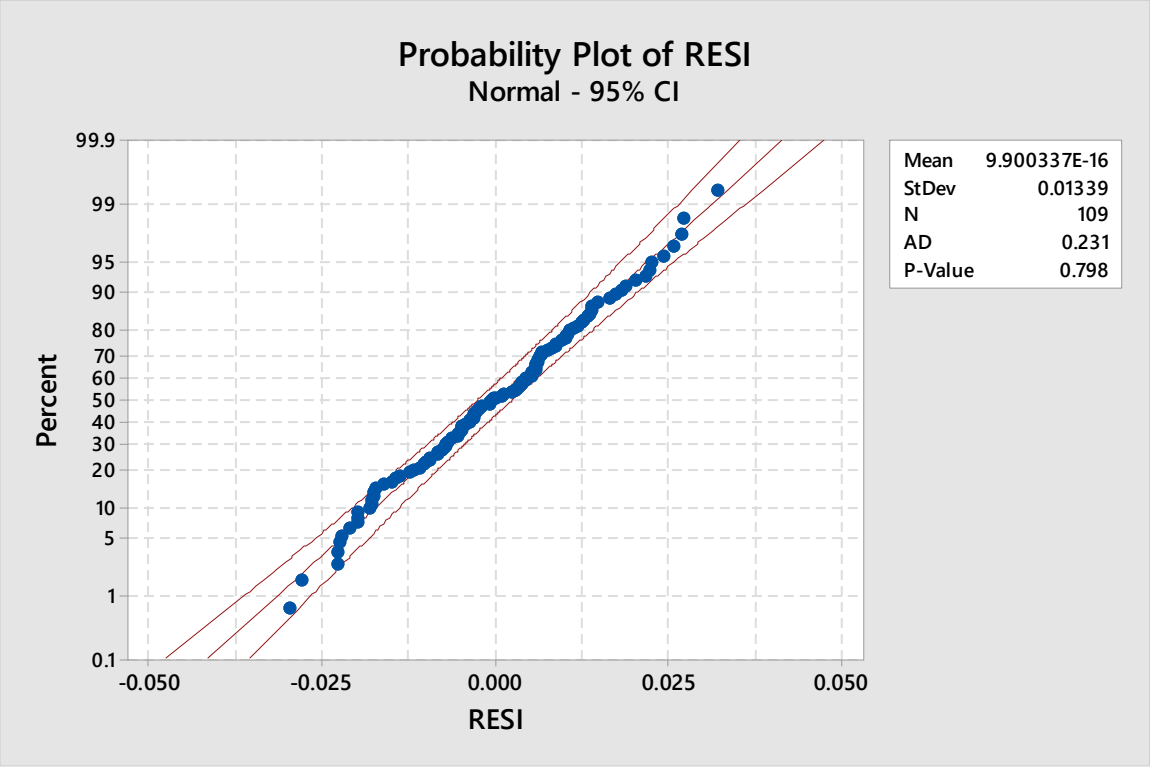
Diagnostic Analysis Model 3

The observations resulting from the adjusted regression results above are as follows:

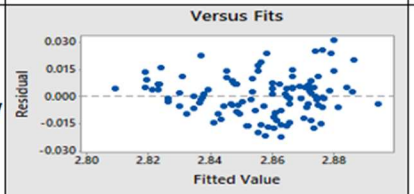
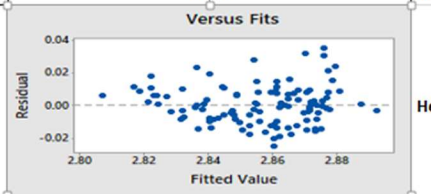
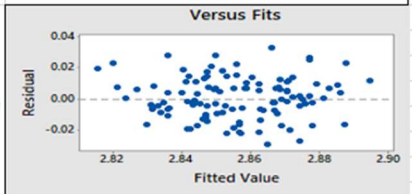
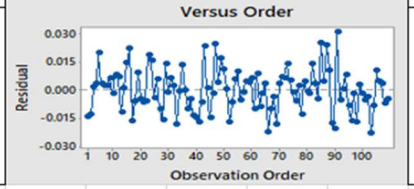
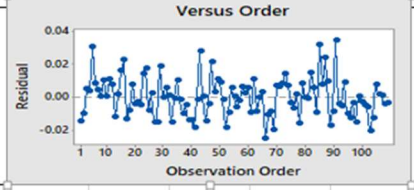
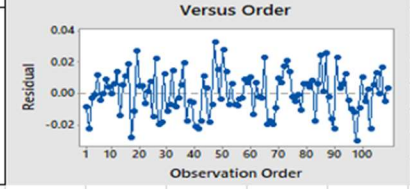
- The R-squared value has is 61.19%
- The Adjusted R-squared value is 58.90%
- The VIF values are all within acceptable 1.X values excluding open_acc and open_acc*total_acc which is less than 5 indicating that they are not terribly egregious predictors. According to Montgomery and Peck, if the VIF is falls in the range of 5-10 then the regression coefficients are poorly estimated.
- The P-value remained the same at 0.00
- The F-Value decreased from 32.50 to 36.80
- The Durbin-Watson Statistic = 1.85645 which is closer to 2 which means much lesser autocorrelation than in Model1
- From Table above showing the *Analysis of Variances* the least point estimates for all coefficient are relevant based on the statistical significance of the p-values for the t-statistics of each variable at an alpha level of 0.05
- There are no outstanding outliers in the probability plot for the residuals
- There is no apparent heteroscedasticity in the plot of residuals versus fitted values for log(faco)

In addition For Model 3 The research explored the probable interaction between open_acc and total_accounts. As noticed the credit ratings can generally be reduced with a higher number of open accounts could the consideration of the total number of successfully closed accounts have a positive value in the regression analysis.


Residual Plots for log(faco)
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order


Probability Plot of RESI
Normal - 95% CI
Mean 9.900337E-16
StDev 0.01339
N 109
AD 0.231
P-Value 0.798
Percent
RESI

Parsimony Analysis

| | Model1 | | | | | Model 2 | | | | | Model 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model Summary** | R-Sq | 72.22% | | | | R-Sq | 70.18% | | | | R-Sq | 61.19% | | | |
| | adjR-sq | 70.00% | | | | adjR-sq | 68.43% | | | | adjR-sq | 58.90% | | | |
| | predR-sq | 65.69% | | | | predR-sq | 66.19% | | | | predR-sq | 54.17% | | | |
| | DW-Statistics | 1.77433 | | | | DW-Statistics | 1.79176 | | | | DW-Statistics | 1.8564 | | | |
| | F-Stat | 32.5 | | | | F-Stat | 40.1 | | | | F-Stat | 26.8 | | | |
| | Standard Error | 0.0117685 | | | | Standard Error | 0.0120734 | | | | Standard Error | 0.0137743 | | | |
| **Coefficients** | **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** | **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** | **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| | Constant | 2.87412 | 0.00634 | 453.24 | 0 | | Constant | 2.87324 | 0.0051 | 562.83 | 0 | | Constant | 2.90654 | 0.00677 | 429.55 | 0 | |
| | loan_amnt | 0.000001 | 0 | 5.92 | 0 | 1.75 | loan_amnt | 0.000001 | 0 | 6.48 | 0 | 1.61 | int_rate | -0.00189 | 0.000402 | -4.71 | 0 | 1.46 |
| | term(months) | 0.000667 | 0.000128 | 5.23 | 0 | 1.78 | term(months) | 0.000633 | 0.000129 | 4.91 | 0 | 1.72 | annual_inc | 0 | 0 | 2.65 | 0.009 | 1.1 |
| | int_rate | -0.0048 | 0.000537 | -8.94 | 0 | 3.57 | int_rate | -0.00483 | 0.000549 | -8.8 | 0 | 3.55 | delinq_2yrs | -0.00844 | 0.00347 | -2.43 | 0.017 | 1.03 |
| | annual_inc | 0 | 0 | 2.23 | 0.028 | 1.24 | delinq_2yrs | -0.00714 | 0.00306 | -2.33 | 0.022 | 1.04 | open_acc | -0.00295 | 0.000768 | -3.84 | 0 | 4.87 |
| | delinq_2yrs | -0.00739 | 0.00298 | -2.48 | 0.015 | 1.04 | inq_last_6mths | 0.002392 | 0.000999 | 2.39 | 0.018 | 1.18 | revol_util | -0.00036 | 0.00006 | -6 | 0 | 1.47 |
| | inq_last_6mths | 0.00236 | 0.000976 | 2.42 | 0.017 | 1.19 | revol_util | -0.00019 | 0.000058 | -3.33 | 0.001 | 1.77 | open_acc * total_ac | 0.000042 | 0.000012 | 3.42 | 0.001 | 4.79 |
| | open_acc | -0.00064 | 0.000314 | -2.03 | 0.045 | 1.12 | | | | | | | | | | | | |
| | revol_util | -0.0002 | 0.000057 | -3.46 | 0.001 | 1.83 | | | | | | | | | | | | |
| **Residuals** | Normality p-value | 0.42 | | | | | Normality p-value | 0.197 | | | | | Normality p-value | 0.798 | | | | |
| | **Heteroscedasticity** |  | | | | | **Heteroscedasticity** |  | | | | | **Heteroscedasticity** |  | | | | |
| | **Independence** |  | | | | | **Independence** |  | | | | | **Independence** |  | | | | |



Red Text – Represents the optimal of each comparative metric listed in Fig X above

Observations

Checking Significance of the Increase in R-squared values from Model 2 to Model 1

The assumption of normality holds for both Model1 and Model 2 hence using the equation below one can assess if this is an improvement.

$$R_f^2 : R^2\text{-value of the full model}, R_r^2 : R^2\text{-value of restricted model}$$

$$df_f : \text{Degees of Freedom of Residual/Error Term in full model}$$
$$df_r : \text{Degees of Freedom of Residual/Error Term in restricted model}$$

$$F = \frac{(R_f^2 - R_r^2)/(df_r - df_f)}{(1-R_f^2)/df_f} \sim F_{(df_r - df_f), df_f}$$

$$H_0 : \text{No model improvement}, H_1 : \text{Model Improvement}$$

| | R-squared Values | Degrees of freedom |
|---|---|---|
| Full (Model 1) | $R^2_f$ = 72.22% | $df_f$ = 100 |
| Restricted (Model 2) | $R^2_r$ = 70.18% | $df_r$ = 102 |
| F ≈ 3.743<br>(72.22% – 70.18%)/ (102– 100)<br>(1- 72.22%)/ 100 | $F_{2,100,\ 0.01}$ = 4.79 | |
| F ≈ 3.743 < 4.79 | Fail to reject Ho:<br>No Model Improvement | |

Conclusion:

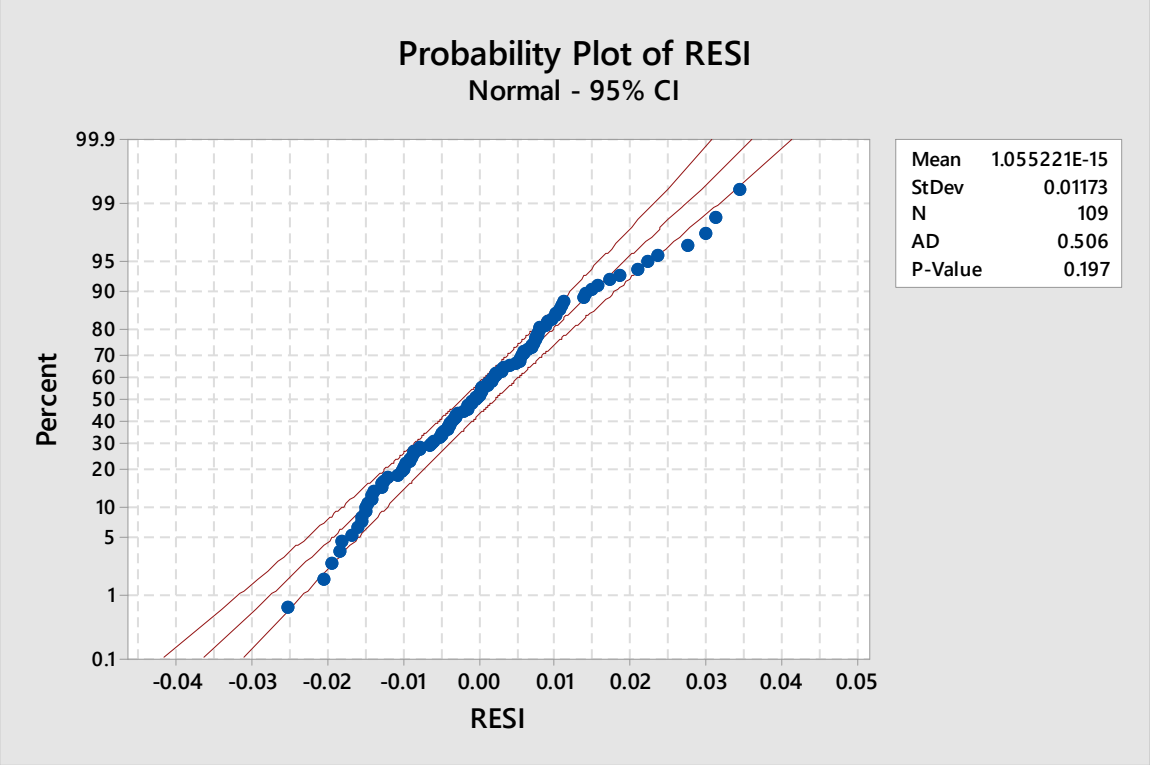
The F Test at alpha level of 0.01 shows no model improvement ruling out Model 1 as the choice model.

Comparing Model 2 against Model 3:

While Model 3 performed better in dealing with multicollinearity because it reported the highest Durbin Watson Statistic (1.8564) closest to 2. This positive development came at the cost of a relatively high Variance Inflation Factor to both *open_acc* and *open_acc* * *total_acc* variables at 4.87 and 4.89 respectively.

Since Model 2 has better overall VIF values, lower standard error metric and a higher R squared Value it will serve as our choice model.

Model Improvement: Outlier Analysis: Model 2




Some outliers

## Model Improvement: Outlier Analysis (Model 2) – Final Model

|DFIT| greater than 2 * √(p+1)/n are considered large

Identified outliers

### Fits and Diagnostics for Unusual Observations

| Obs | log(faco) | Fit | Resid | Std Resid | | |
|---|---|---|---|---|---|---|
| 5 | 2.90584 | 2.87587 | 0.02997 | 2.55 | R | |
| 43 | 2.88170 | 2.85419 | 0.02751 | 2.33 | R | |
| 65 | 2.84235 | 2.83930 | 0.00305 | 0.30 | | X |
| 66 | 2.83524 | 2.86049 | -0.02526 | -2.12 | R | |
| 73 | 2.86882 | 2.85479 | 0.01403 | 1.36 | | X |
| 85 | 2.90137 | 2.87011 | 0.03126 | 2.73 | R | |
| 91 | 2.91029 | 2.87596 | 0.03432 | 2.90 | R | |
| 99 | 2.82418 | 2.82397 | 0.00021 | 0.02 | | X |

*R Large residual*
*X Unusual X*

## PREDICTIONS

### Prediction for log(faco)

#### Regression Equation

log(faco) = 2.87324 + 0.000001 loan_amnt + 0.000633 term(months) - 0.004834 int_rate - 0.00714 delinq_2yrs + 0.002392 inq_last_6mths - 0.000192 revol_util

#### Settings

| Variable | Setting |
|---|---|
| loan_amnt | 9800 |
| term(months) | 60 |
| int_rate | 9.63 |
| delinq_2yrs | 0 |
| inq_last_6mths | 0 |
| revol_util | 28.5 |

#### Prediction

| Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|
| 2.87208 | 0.0025759 | (2.86697, 2.87719) | (2.84760, 2.89657) |

Showing the 95% prediction interval and looking at the estimated values from the adjusted model

| **XTX** | | | | | | |
|---|---|---|---|---|---|---|
| 109 | 1350850 | 4980 | 1249.18 | 14 | 112 | 4925.9 |
| 1350850 | 22445343750 | 64135800 | 17092555.25 | 173200 | 1256500 | 63445120 |
| 4980 | 64135800 | 242640 | 60034.32 | 672 | 5088 | 229611.6 |
| 1249.18 | 17092555.25 | 60034.32 | 16031.7824 | 186.8 | 1346.56 | 62676.32 |
| 14 | 173200 | 672 | 186.8 | 18 | 19 | 763.2 |
| 112 | 1256500 | 5088 | 1346.56 | 19 | 288 | 4755.8 |
| 4925.9 | 63445120 | 229611.6 | 62676.317 | 763.2 | 4755.8 | 300027 |

| (XTX)-1 | | | | | | |
|---|---|---|---|---|---|---|
| 0.178783 | -1.61775E-06 | -0.003168762 | 0.002909071 | 0.001477064 | -0.00990529 | -0.00062 |
| -1.6E-06 | 2.82922E-10 | 3.20165E-08 | -4.32654E-07 | 3.2397E-07 | 4.14157E-07 | 2.52E-08 |
| -0.00317 | 3.20165E-08 | 0.000113943 | -0.000294412 | 6.77165E-05 | 0.000178234 | 1.66E-05 |
| 0.002909 | -4.32654E-07 | -0.000294412 | 0.002070071 | -0.001286169 | -0.00134385 | -0.00014 |
| 0.001477 | 3.2397E-07 | 6.77165E-05 | -0.001286169 | 0.06416066 | -0.00102625 | -2.3E-05 |
| -0.00991 | 4.14157E-07 | 0.000178234 | -0.001343854 | -0.001026247 | 0.006845684 | 0.000113 |
| -0.00062 | 2.52234E-08 | 1.65596E-05 | -0.000138825 | -2.28417E-05 | 0.000113476 | 2.28E-05 |

## Conclusions and Recommendations

### Regression Equation

log(faco) = 2.87324 + 0.000001 loan_amnt + 0.000633 term(months) - 0.004834 int_rate - 0.00714 delinq_2yrs + 0.002392 inq_last_6mths - 0.000192 revol_util

A simple interpretation of the equation above is as follows:

- FACO scores increase by 0.000001 unit when the average loan_amnt is increased by $1 and all other variables remain unchanged
- FACO scores increase by 0.000633 unit when the term months is increased by 1 month, other variables remaining unchanged
- FACO scores decreases by -0.00483 units when the interest rate is increased by 1%, all other variables remaining unchanged
- FACO scores decrease by -0.00714 units when the deliq_2yrs is increased by 1 instance, all other variables remaining unchanged
- FACO scores decrease by 0.000192 unit when the revol_util increases by 1%, all other variables remaining unchanged

One of the clear indicators of the pertinence of this model to general scoring metrics is the need to keep the revolving utility down as indicated by the negative intercept of our equation in order to maintain a higher credit score.

The highest factor that contributes to a reduction in Credit rating is the delinquency in 2 years feature with the highest absolute negative slope estimate value at (-0.00714)

## References

Abdou, H. & Pointon, J. (2011) 'Credit scoring, statistical techniques and evaluation criteria: a review of the literature ', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88

My FICO Scores https://www.myfico.com/credit-education/whats-in-your-credit-score/