STAT 306 2024 S2 Group Project Group A2

Introduction

In today's financial landscape, understanding the factors that influence approved loan amounts is becoming increasingly critical. Financial institutions are leveraging data-driven decision-making to not only provide accessibility to borrowers but also to minimize risk. In this analysis, we focus on identifying and explaining the key variables that significantly influence the amount of money approved for loans. By exploring the relationships between various financial metrics and loan amounts, we aim to gain deeper insights into the underlying factors that drive lending decisions.

The dataset used in this analysis, Loan-Approval-Prediction-Dataset, was taken from Kaggle and includes various factors of interest for each loan. Below is an overview of the variables used in this analysis:

Possible Qualitative Predictors:

- **loan_id:** Loan ID of the applicant, numerical but is qualitative due to its nature as a unique identifier
- **education:** Educational level of the applicant (Graduate/Not Graduate)
- **self_employed:** Self-employment status of the applicant (Yes/No)
- **loan_status:** Current status of the loan application (Approved/Rejected), filter by 'Approved' loans only

Possible Quantitative Predictors:

- **no_of_dependents:** Number of dependents of the applicant (ranges from 0 onwards)
- **income_annum:** Annual income of the applicant ($)
- **loan_term**: Loan Term (in years > 0)
- **cibil_score:** Credit score of the applicant (ranges from 300 to 900, inclusive)
- **residential_assets_value:** Total value of residential assets owned by the applicant ($)
- **commercial_assets_value:** Total value of commercial assets owned by the applicant ($)
- **luxury_assets_value:** Total value of luxury assets owned by the applicant ($)
- **bank_asset_value:** Total value of bank/financial assets of the applicant ($)

Response Variable:

- **loan_amount:** Loan amount (in $), key point of interest based on how the predictors change

As this analysis aims to see which variables are most significant in increasing the amount of money for approved loans, we needed to filter out all the rejected loans prior to making our analysis. From an outsider's standpoint, it may seem obvious that someone who is more "successful" from a monetary standpoint may be able to get larger loans approved. This may be due to their high credit scores, net worth, high educational status, or some combination of many predictors. With that being said, while it may be intuitive to think that financially successful individuals may be able to fetch larger loans, we anticipate that there may be multicollinearity amongst some of the financial metrics which make them successful.

As such, this analysis aims to explore and uncover such findings and ascertain what metrics are most significant in determining the monetary amount of approved loans. Multicollinearity, variance inflation factors, Cook's distance, and other statistical concepts will be used in determining which variables are most significant. We will use the R programming language as a tool for analytics, especially for graphical outputs.

Data Cleaning and Preliminary Investigation

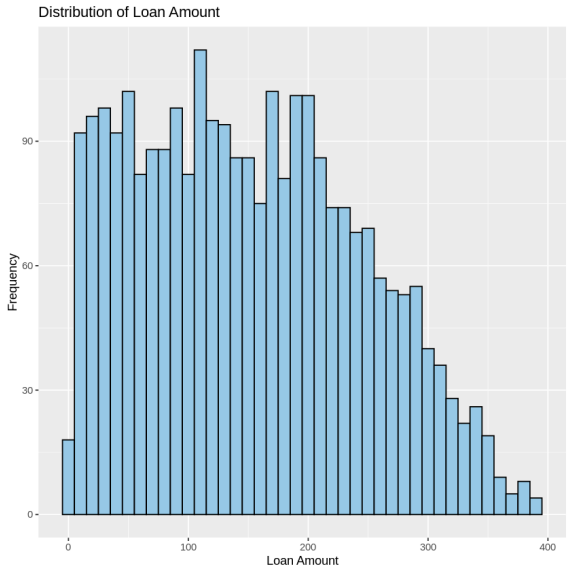
A preliminary assessment of our data showed some inefficiencies that could hinder our analysis process, hence the need to clean our data and prepare it for easier interpretation. As mentioned earlier, we are only interested in how the loan prices of approved loans are affected, so we filtered for just approved loans (under the loan_status column). Additionally, we found that loan ID (an identifier of each loan) and loan status would not be helpful in this analysis given that identifiers do provide any meaningful characteristics about the data itself and the loan status was simply used to filter out our tuples of interest.

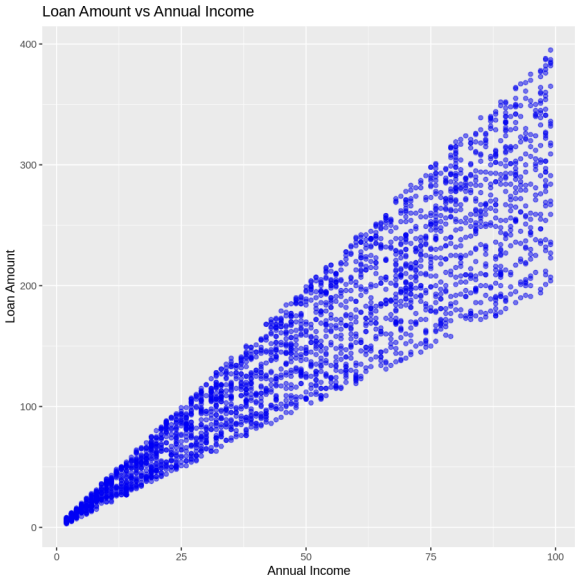
On the quantitative side of our data, we found that there were five predictor variables (annual income, residential assets value, commercial assets value, luxury assets value, bank assets value), as well as the response variable (loan amount), that were in terms of $ (USD). Many of the figures were well over $100000 across all tuples. As such, we decided upon dividing each value by 100000, scaling the units from $ to $100000. This is also a common practice in the industry as it makes analysis easier with smaller figures and interpretability is not made any more difficult.

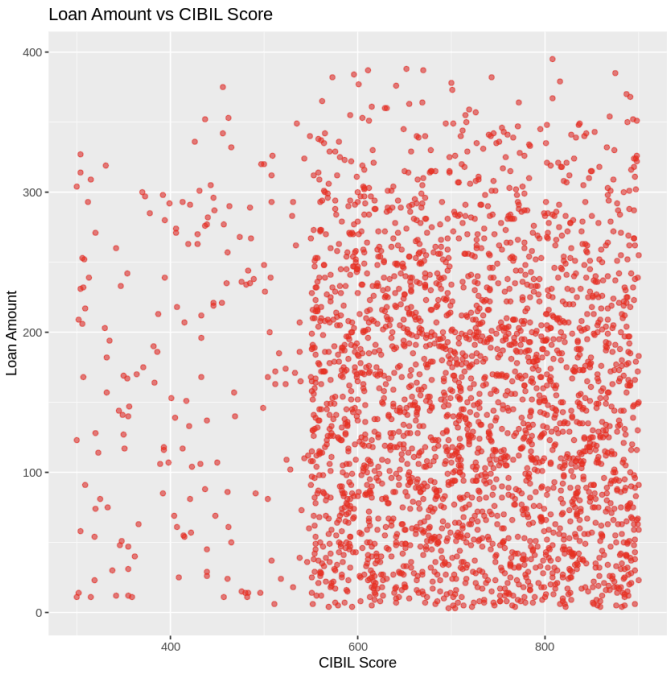
Continuing with our preliminary assessment, we constructed some scatter plots between predictor variables and income value. The plots were as follows:

- **distribution of loan amount - histogram (shown below)**
- **annual income against loan amount (shown below)**
- **credit score against loan amount (shown below)**

- number of dependents against loan amount
- loan term against loan amount
- residential assets value against loan amount
- commercial assets value against loan amount
- luxury assets value against loan amount
- bank asset value against loan amount
- Bar plot for education
- Bar plot for self employed







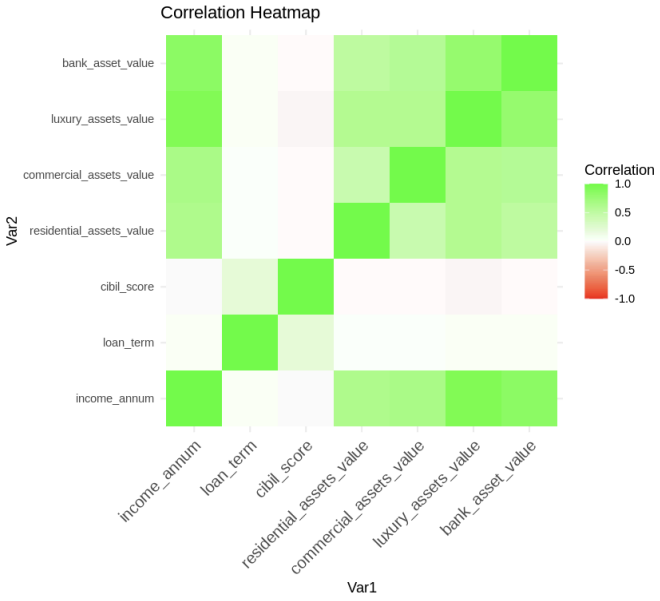
The plot of annual income against loan amount revealed a strong positive correlation between the two variables, although it is noteworthy that variance increased rapidly as annual income progressed over $1 million dollars per year. Similarly, the plots for luxury assets value and bank asset value against loan amount revealed a moderate positive correlation. The other graphs did not particularly reveal any convincing relationship between the predictors and loan amount. However, interestingly enough, there was indication that the data we were examining was based on approved loans, such as the credit score against loan amount plot which had drastically more data points for credit scores greater than 550. The same plot also showed that credit scores greater than 550 had many greater approved loan amounts compared to approved loans for people with lower credit scores, further confirming our early suspicions that credit score could be a significant variable in predicting for greater loan amounts.

Analysis

Our approach was to gradually cut any explanatory variables which didn't prove to be significant as a predictor. We started by creating a full model which included every explanatory variable. Below is the summary for this model in R:

```
Call:
lm(formula = loan_amount ~ ., data = loan_data)

Residuals:
    Min      1Q  Median      3Q     Max
-98.395 -18.153   0.171  19.698  99.118

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              26.468960   4.060459   6.519 8.46e-11 ***
no_of_dependents         -0.610683   0.379292  -1.610   0.1075
educationNot Graduate    -0.464121   1.285379  -0.361   0.7181
self_employedYes         -0.131761   1.286044  -0.102   0.9184
income_annum              2.934869   0.078605  37.337  < 2e-16 ***
loan_term                -0.248972   0.111818  -2.227   0.0261 *
cibil_score              -0.031683   0.005249  -6.036 1.80e-09 ***
residential_assets_value  0.007954   0.012794   0.622   0.5342
commercial_assets_value   0.049574   0.019493   2.543   0.0110 *
luxury_assets_value       0.003205   0.019054   0.168   0.8664
bank_asset_value          0.032963   0.037911   0.869   0.3847
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.09 on 2645 degrees of freedom
Multiple R-squared:  0.8718,    Adjusted R-squared:  0.8713
F-statistic:  1798 on 10 and 2645 DF,  p-value: < 2.2e-16
```

While the R^2 and the adjusted R^2 values for this model are high, both standing at a very strong 0.87, most of the explanatory variables are not statistically significant, evident in the rightmost column of the coefficients table. Statistical significance in this context will be defined as 0.05, with values below this threshold being considered statistically significant. We will further calculate the variance inflation factor (VIF) values, in order to observe the presence of collinearity. The correlation heatmap to the right displays potential correlation between the explanatory variables and large VIF values will help support that, with large VIF values typically being defined as greater than 3-5.



Explanatory Variable	VIF Value
no_of_dependents	1.003
education	1.002
self_employed	1.003

income_annum	11.978
loan_term	1.048
cibil_score	1.048
residential_assets_value	1.678
commercial_assets_value	1.777
luxury_assets_value	7.379
bank_asset_value	3.736

Annual income and luxury assets value have high values of VIF, especially annual income, confirming our preconceptions in the previous section. With that being said, the correlation heatmap suggests the possibility of collinearity between the asset variables (residential, commercial, luxury, bank) as seen by the deep green square on the top right of the correlation heatmap. One possible way to address this is to transform these explanatory variables into one single variable, in this instance, the average of the assets variables so that we have one explanatory variable that aggregates the total value of all assets in one's possession. We repeated the previous process with a new model replacing the asset variables with the new average of assets variable. However, the VIF value for this new average of assets was significantly high as well, sitting at approximately 7.77. As such, we made the decision to remove that variable altogether. We refitted another model with the remaining six explanatory variables, and subsequently calculated the VIF values again:

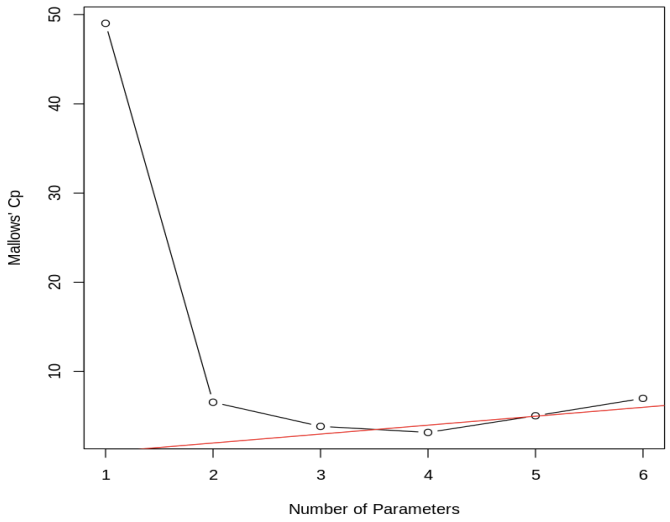
```
Call:
lm(formula = loan_amount ~ no_of_dependents + education + self_employed +
    income_annum + loan_term + cibil_score, data = loan_data)

Residuals:
     Min       1Q   Median       3Q      Max
-100.165  -17.879    0.238   19.499   97.083

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          26.426662   4.061199   6.507 9.13e-11 ***
no_of_dependents     -0.617758   0.379353  -1.628   0.1035
educationNot Graduate -0.462188   1.285983  -0.359   0.7193
self_employedYes     -0.282722   1.285789  -0.220   0.8260
income_annum          3.039452   0.022747 133.621  < 2e-16 ***
loan_term            -0.247170   0.111850  -2.210   0.0272 *
cibil_score          -0.031620   0.005248  -6.025 1.92e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.11 on 2649 degrees of freedom
Multiple R-squared:  0.8714,    Adjusted R-squared:  0.8711
F-statistic:  2991 on 6 and 2649 DF,  p-value: < 2.2e-16
```

Explanatory Variable	VIF Value
no_of_dependents	1.002
education	1.002
self_employed	1.001
income_annum	1.002
loan_term	1.047
cibil_score	1.046

As seen above, VIF values for every explanatory variable are small ranging from 1.001 to 1.047, possibly indicating that the issue of collinearity appears to be resolved. Next, we needed to decide which explanatory variables could produce the best linear model to explain the monetary amount of approved loans. As such, we used the "regsubsets" command in R to find the best model for each number of parameters. Moreover, we calculate the Mallows' Cp statistic, in order to help select a model:

A matrix: 6 × 7 of type lgl

	(Intercept)	no_of_dependents	educationNot Graduate	self_employedYes	income_annum	loan_term	cibil_score
1	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
3	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
4	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
5	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

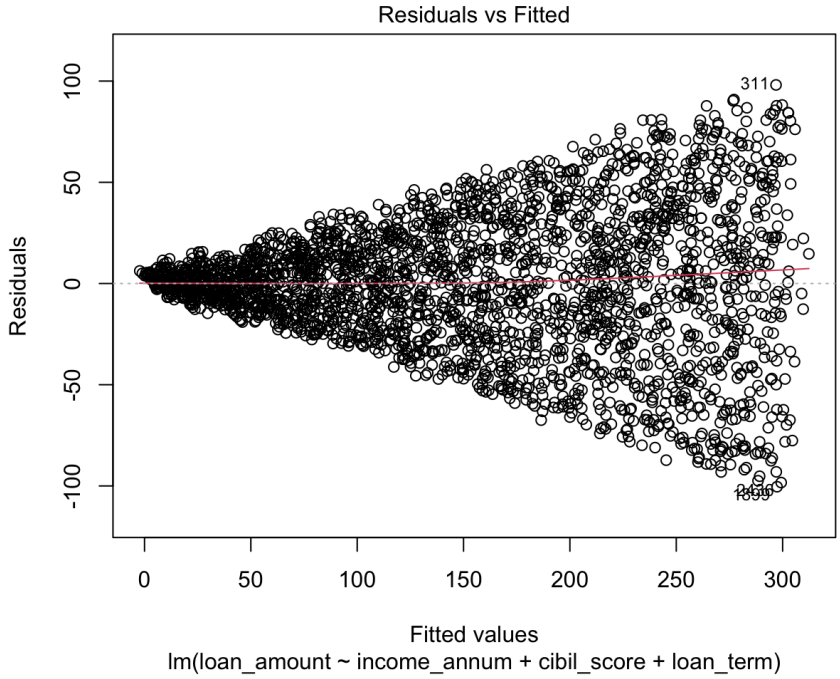


From the Mallows' Cp plot on the right, we can say that the model with 3 parameters appears to be the best model. This is because it has a Mallows' Cp value close to its p value while having a low number of parameters. Models with 4, 5, and 6 parameters also have a Mallows' Cp value close to their respective p values, however we need to weigh the benefit of only slightly higher precision at the expense of potentially introducing overfitting or collinearity from new parameters. As such, the model with the explanatory variables annual

income, loan term, and credit score was chosen to be the best model for explaining the monetary amount of approved loans. We fitted this model similar to previous methods, with the summary of this linear model of three parameters being:

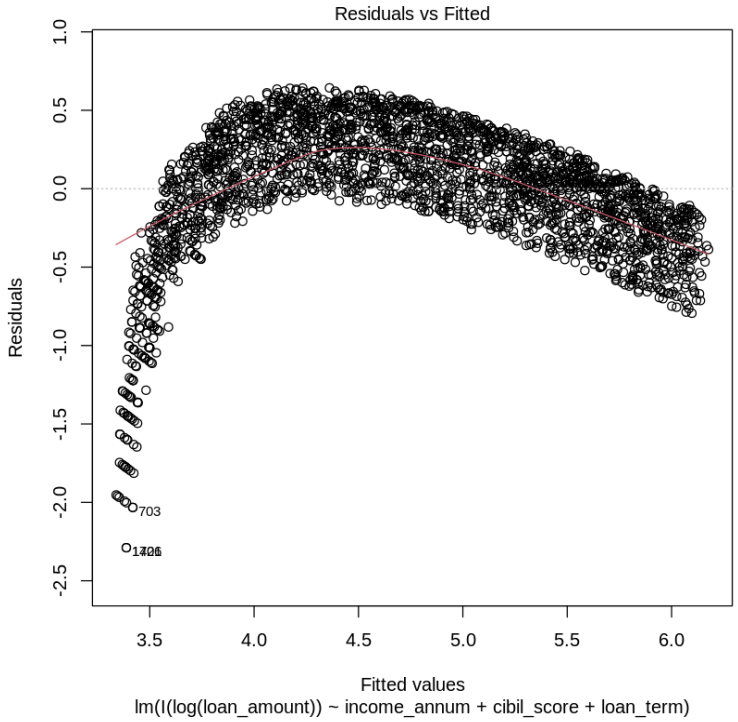
```
Call:
lm(formula = loan_amount ~ income_annum + cibil_score + loan_term,
    data = loan_data)

Residuals:
     Min       1Q   Median       3Q      Max
-100.383  -18.170    0.317   19.564   98.096

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.637829   3.874480   6.359 2.38e-10 ***
income_annum  3.039545   0.022745 133.637  < 2e-16 ***
cibil_score  -0.031857   0.005245  -6.074 1.43e-09 ***
loan_term    -0.242348   0.111764  -2.168   0.0302 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.11 on 2652 degrees of freedom
Multiple R-squared:  0.8712,    Adjusted R-squared:  0.8711
F-statistic:  5981 on 3 and 2652 DF,  p-value: < 2.2e-16
```

Similar to the full model, the R^2 and the adjusted R^2 values are both approximately 0.87, but with the added advantage of every parameter being less than 0.05, or statistically significant. The next step is to visualize the residuals, and evaluate different diagnostic plots in order to see if there are any concerns with our linear model, adjusting using transformations as necessary. We started with a plot of the residuals against the fitted values:



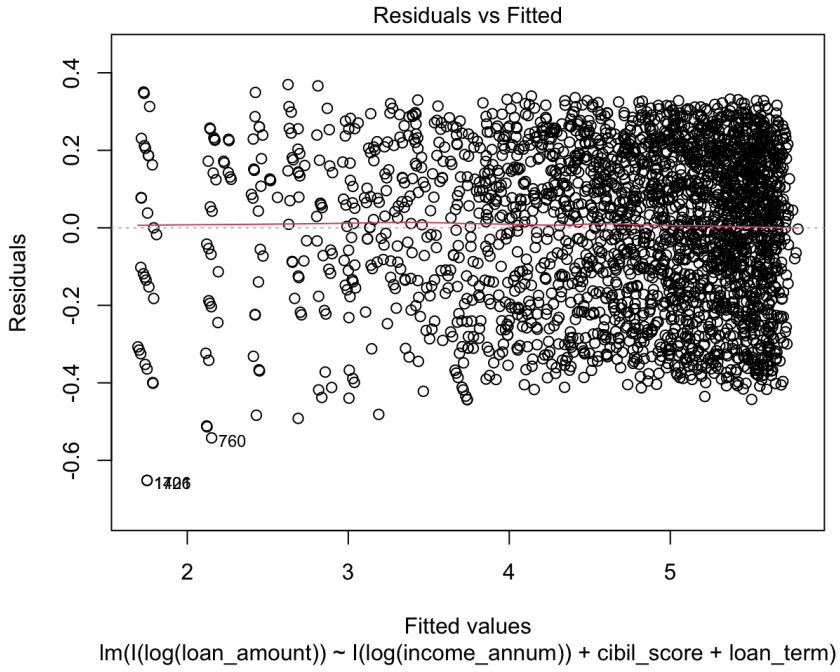
The residuals plotted against fitted values shows that there is heteroscedasticity (non-constant variance) which is a big concern. There is fanning out of the residuals, meaning that the residuals get larger as the fitted value increases. One common industry method to control heteroscedasticity is taking the log of the response variable, loan amount in this case, so that's what we tried. We repeated the previous process of investigating model diagnostics with our newly transformed model:

```
Call:
lm(formula = I(log(loan_amount)) ~ income_annum + cibil_score +
    loan_term, data = loan_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28953 -0.19606  0.04846  0.29207  0.64297

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.494e+00  4.784e-02  73.043  < 2e-16 ***
income_annum  2.807e-02  2.808e-04  99.947  < 2e-16 ***
cibil_score  -2.149e-04  6.475e-05  -3.319 0.000916 ***
loan_term    -1.082e-03  1.380e-03  -0.784 0.432861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4088 on 2652 degrees of freedom
Multiple R-squared:  0.7908,    Adjusted R-squared:  0.7906
F-statistic:  3341 on 3 and 2652 DF,  p-value: < 2.2e-16
```



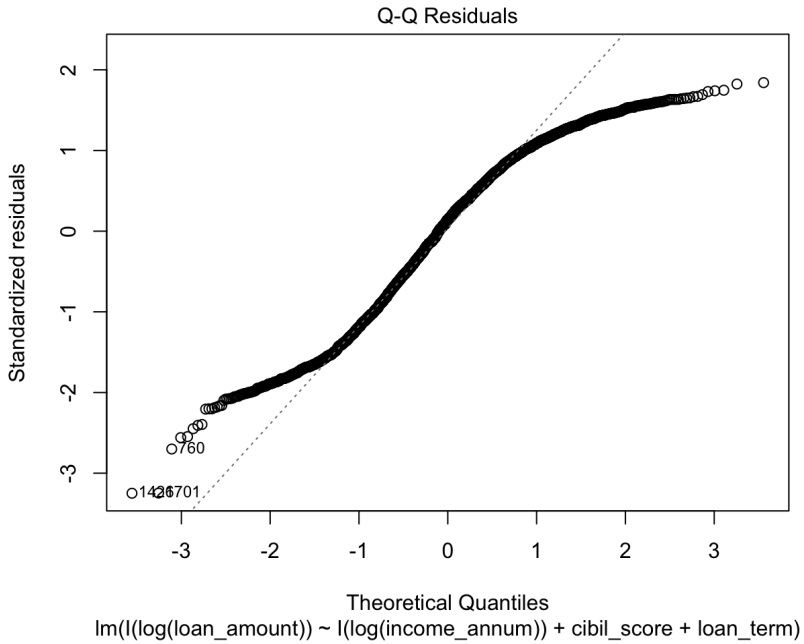
The residuals plotted against fitted values shows a quadratic pattern, which is another indication of a poorly fitted model. The next step would be taking the log of both the response variable and explanatory variables, gradually taking the log of one variable at a time. As such, we started by fitting a model including the log of the loan amount and the log of the annual income variables.

```
Call:
lm(formula = I(log(loan_amount)) ~ I(log(income_annum)) + cibil_score +
    loan_term, data = loan_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65191 -0.15670  0.02862  0.17216  0.36967

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.242e+00  2.795e-02  44.454  < 2e-16 ***
I(log(income_annum))  1.010e+00  4.537e-03 222.615  < 2e-16 ***
cibil_score          -2.564e-04  3.186e-05  -8.047 1.27e-15 ***
loan_term            -1.145e-03  6.789e-04  -1.686   0.0919 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2011 on 2652 degrees of freedom
Multiple R-squared:  0.9493,    Adjusted R-squared:  0.9493
F-statistic: 1.657e+04 on 3 and 2652 DF,  p-value: < 2.2e-16
```

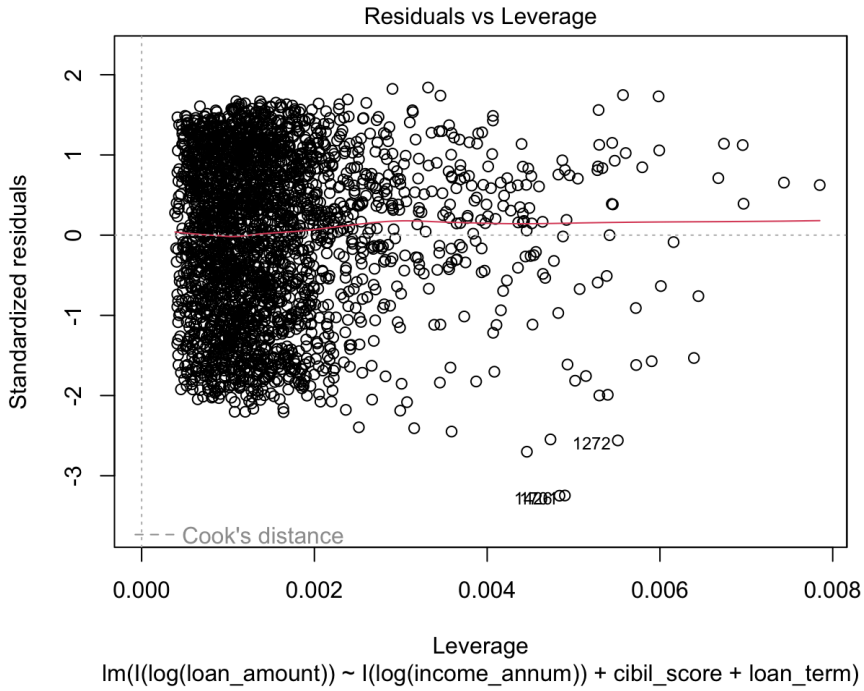


The issue of heteroscedasticity appears to be resolved, since there is no fanning of the residuals in the residuals plotted against fitted values. In other words, the residuals now have

homoscedasticity (constant variance). The plot above displays a very ideal patternless cloud of randomly scattered data points about 0.



We continued with model diagnostics with a QQ plot and leverage plotted against standardized residuals. The QQ plot suggests that there is slight skewness at the tails, but overall does not pose any major concerns since the majority of the residuals follow a straight line, allowing us to assume normality of the residuals.



From the Leverage vs Standardised Residuals plot, we can say that there are no influential points that greatly impact the linear model since all the points have an acceptable value of Cook's Distance. Thus, all the residual plots posit that the logistic model fitted is a good fit and poses no glaring concerns.

Conclusion

This analysis aimed to identify the most significant factors influencing the monetary amount of approved loans. Specifically, we explored how various predictors, such as annual income, credit score, loan term, education, self employment, and assets values impact the loan amount, while addressing potential multicollinearity among these variables.

The final and best model fitted was:

$$log(loan\ amount) = \beta_0 + \beta_1 log(income) + \beta_2 cibil\ score + \beta_3 loan\ term$$

$$log(loan\ amount) = 1.242 + 1.010\ log(income) - 0.00026\ cibil\ score - 0.0011\ loan\ term$$

Explanatory Variables:

- Annual Income: The coefficient of the log of income per annum is 1.010, suggesting that a 1% increase in income leads to a 1.01% increase in the loan amount. This result aligns with the intuition that higher income levels enable borrowers to qualify for larger loans.

- CIBIL Score: The coefficient for the CIBIL score is -0.00026, which was unexpected, as it implies that an increase in the CIBIL score is associated with a decrease in the loan amount. A plausible explanation for this negative relationship is that borrowers with higher CIBIL scores may be more financially cautious and tend to borrow smaller amounts that they are confident in repaying.
- Loan Term: The coefficient of loan term is -0.0011, indicating that an increase in the loan term by one year results in a slight decrease in the loan amount. This could be because longer-term loans are often viewed as riskier due to the extended period over which the borrower's financial situation could change.

This model yielded an adjusted R^2 value of 0.9493, indicating that the three variables of log(annual income), credit score, and loan term explain approximately 94.93% of the variance in the loan amount. This is the highest adjusted R^2 value among all models considered throughout this analysis, demonstrating the strong predictive power of the selected variables.

The analysis effectively identified annual income, credit score, and loan term as significant predictors of the loan amount. While higher income was associated with larger loans, the unexpected negative relationship between credit score and loan amount highlights the complex behaviour of financially successful individuals, who may opt for smaller, more manageable loans. These insights are valuable for lenders in refining their loan approval processes and for borrowers in understanding how different financial metrics influence loan amounts. For future reference, it would be reasonable to study why these negative relationships exist.