STUDENT LOAN DEBT

**Springboard Capstone Project: Modeling Student Loan Debt**

Aneesha Vallabhaneni

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 1 |

**Project: Modeling Student Loan Debt**

**Executive Summary**

Problem Statement

* Goal is to build a robust model to explain and predict the student loan debt based on several factors.
* We will be able to predict the time at which federal student loan debt will reach the warning levels so that some alternative measures can be taken to ensure the financial needs of the students are not drastically affected.

The Data

* I have done a thorough analysis of federal student loan debt and the factors influencing it. Based on this analysis I have selected 7 predictor variables that I believe best depict the current student loan landscape
* These predictor variables are: the average tuition fee, total full-time students enrolled, average high school dropout rate, average household income, CPI, average unemployment rate, and stock market index.
* I have compiled the data from different reliable sources, dating back to 1970.

The Model

* I explored the data using several visual techniques for a better understanding of how the data are distributed.
* Necessary actions were taken to help data interpretation.
* Created several tables to compare the outcomes of several potential models from these 7 predictor variables.
* Three variables – average tuition fee, total full-time enrollments, and average household income

– were selected to build the best suited model to predict the student loan debt.

* Confirmed that the best suited model adheres to all the rules of regression analysis and effectively solves the problem.
* Checked for any abnormal observations which are not explained by the model.

Conclusions

* I have built a model from which we can easily find the answers for the problem mentioned in the problem statement. This is done by interpreting the model effectively with regard to the problem.
* The model explains how federal student loan debt varied over the past 50 years.
* The model predicts the trend that federal student loan debt will follow in the coming years.

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 2 |

**Introduction**

The current landscape of student loan debt in the United States is frightening. The aggregate federal student loan debt exceeds $1 trillion, which is none the bit surprising when you consider that roughly two-thirds of college graduates will leave with some amount of debt (Denhart, 2013). The average student will have nearly $26,000 of debt to pay back, while for ten percent of students this value will be as large as $40,000. Some will even owe over $100,000! At this level, student loan debt is the largest form of consumer debt in the United States, second only to mortgages (Denhart, 2013). So what factors contribute to this hefty sum?

The goal of this project is to build an effective model to predict the climbing student loan debt in the United States. I have constructed a dataset from a variety of sources and built a multiple linear regression model on the aggregate federal student loan debt.

**Meet the Data**

The response variable that we are interested in is the aggregate US federal student loan debt in million USD. The predictor variables are listed below. Data are collected from 1970 to 2014 for each variable.

* Tuition = average tuition in American universities, measured in USD
* Enroll = total head count of full-time students enrolled in American universities
* Dropout = average percentage of US high school dropout
* Income = the mean household income for those households whose income is placed in the top 40 - 60%, measured in USD
* CPI = consumer price index:
* UnEmp = annual average US unemployment rate, measured in percentages
* Stock = average NASDAQ Index of the year

It is worth noting that while Loan, Tuition, and Enroll are calculated for an academic year from September to the following August, other variables capture the average values for a calendar year.

Below I have given a justification why a predictor variable is included. It is obvious that the total amount of loan debt is related to the size of college enrollment and the magnitude of tuition. As college tuition increases dramatically for the past decades, we expect to find total amount of loan debt to increase. The larger is the enrollment, the greater quantities of students who take loans exist. Conversely, an increase in high school student dropout leads to a smaller pool of students who intend to attend college. High school dropout negatively influences the amount of loan debt.

Since tuition is considered as part of consumer spending, we expect amount of loan debt to increase as consumer price index (CPI) increases. Intuitively, if household income increases, disposable income might also increase, the amount of loan that people take would decrease. If unemployment rate increases or the stock market performs badly, the economy might have taken a hit; more people go back to school and thus are in greater need of taking larger amount of loans.

All those predictor variables influence the amount of loan debt to some degree, either positively or negatively. We would like to include all of them at first, and then determine which ones to choose by applying our knowledge of multiple linear regression.

The sources of the variables are listed below:

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 3 |

* Loan, Tuition, Enroll, and Dropout: NCES (National Center for Education Statistics)
* Income: US Census Bureau
* CPI and Stock: FRED (Federal Reserve Economic Data)
* UnEmp: BLS (Bureau of Labor Statistics)

Soaring student loan debts create a crisis for not only students, but also the economy at large (Denhart, 2013). The issue of student loan debt took center stage at Congress debate recently (Mitchell, 2014). American families, in general, are interested in the fate of this student loan debt crisis, asking whether the loan debt will continue climbing.

In this project, I plan to explore a few variables that influence student loan debt, and try to forecast the future trend of loan debt, building on multiple linear regression modeling.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## |  | Y | Loan Tuition | | Enroll Dropout Income | | | CPI UnEmp | | Stock |
| ## | 1 | 1970 | 7014.68 | 9625 | 6996129 | 5.7 | 8689 | 39.0 | 5.0 | 83.45 |
| ## | 2 | 1971 | 8485.68 | 9729 | 7148575 | 6.6 | 8965 | 40.7 | 6.0 | 98.21 |
| ## | 3 | 1972 | 8251.56 | 9914 | 7253712 | 7.1 | 9624 | 41.9 | 5.6 | 109.78 |
| ## | 4 | 1973 | 7867.97 | 9443 | 7453467 | 7.1 | 10471 | 44.3 | 4.9 | 106.51 |
| ## | 5 | 1974 | 7732.93 | 8858 | 7805454 | 6.4 | 11147 | 49.4 | 5.6 | 81.48 |
| ## | 6 | 1975 | 6957.83 | 8774 | 8479688 | 6.3 | 11724 | 54.2 | 8.5 | 87.13 |

**Modeling**

**Preliminary Analysis**

Before diving into the model selection process, I explored the dataset using visual tools. In doing so, I was able to get a better understanding of how the data is distributed and how the predictor variables are related to the response variable. Each variable in the dataset was plotted using histograms. This has enabled to determine which variables might require transformation.

From these histograms (see appendix) we can determine that Loan, Tuition, UnEmp, and Stock are all right-skewed. This tells us that re-expression might be necessary. The remaining variables are all *relatively* normal, and may not require transformation. We continued our preliminary analysis byplotting each predictor variable against the response, Loan. This was done so using a matrix scatterplot (see appendix). I have also produced a correlations plot to measure the strength of the relationships among the variables.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Loan | Tuition | Enroll | Dropout | Income | CPI | UnEmp | Stock |
| Loan |  | **1.0000000** | **0.972431574** | **0.9843176** | **-0.835002740** | **0.91654090** | **0.925301514** | 0.141310388 | **0.8735617** |
| Tuition | | 0.9724316 | **1.000000000** | **0.9647447** | **-0.858168505** | **0.94862561** | **0.948139116** | -0.008510763 | **0.9295502** |
| Enroll | | 0.9843176 | 0.964744749 | **1.0000000** | **-0.868093612 0.94108703 0.951251923** 0.169765861 **0.8781773** | | | | |
| Dropout -0.8350027 -0.858168505 -0.8680936 | | | | | **1.000000000** | **-0.93019162** -**0.933442369** | | 0.004666452 | **-0.8094687** |
| Income | | 0.9165409 | 0.948625615 | 0.9410870 | -0.930191622 | **1.00000000** | **0.996245749** | -0.075676340 | **0.9368277** |
| CPI | | 0.9253015 | 0.948139116 | 0.9512519 | -0.933442369 | 0.99624575 | **1.000000000** | -0.008880543 | **0.9157745** |
| UnEmp | | 0.1413104 | -0.008510763 | 0.1697659 | 0.004666452 | -0.07567634 -0.008880543 | | **1.000000000** | -0.2246500 |
| Stock | | 0.8735617 | 0.929550176 | 0.8781773 | -0.809468650 | 0.93682775 | 0.915774482 | -0.224649999 | **1.0000000** |

From these two matrices, we can determine that many of the predictors are strongly related with the the response variable. In fact, all of the predictors except UnEmp are highly correlated with Loan. Specifically, Enroll, Tuition, and Income are all positively correlated with Loan at a level greater than *r* = +0.90. We should keep an eye on these three variables moving forward. We can also see that there is high potential for a multicollinearity issue in this data. A large portion of the variables are highly

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 4 |

correlated with one another. We should be especially concerned with correlations larger than *r* = 0.80 amongst two predictor variables. Each correlation larger than this value is bold-faced in the plot above.

Now that we have an idea about how the data is distributed and about the relationships amongst the variables, we can proceed with the model selection process.

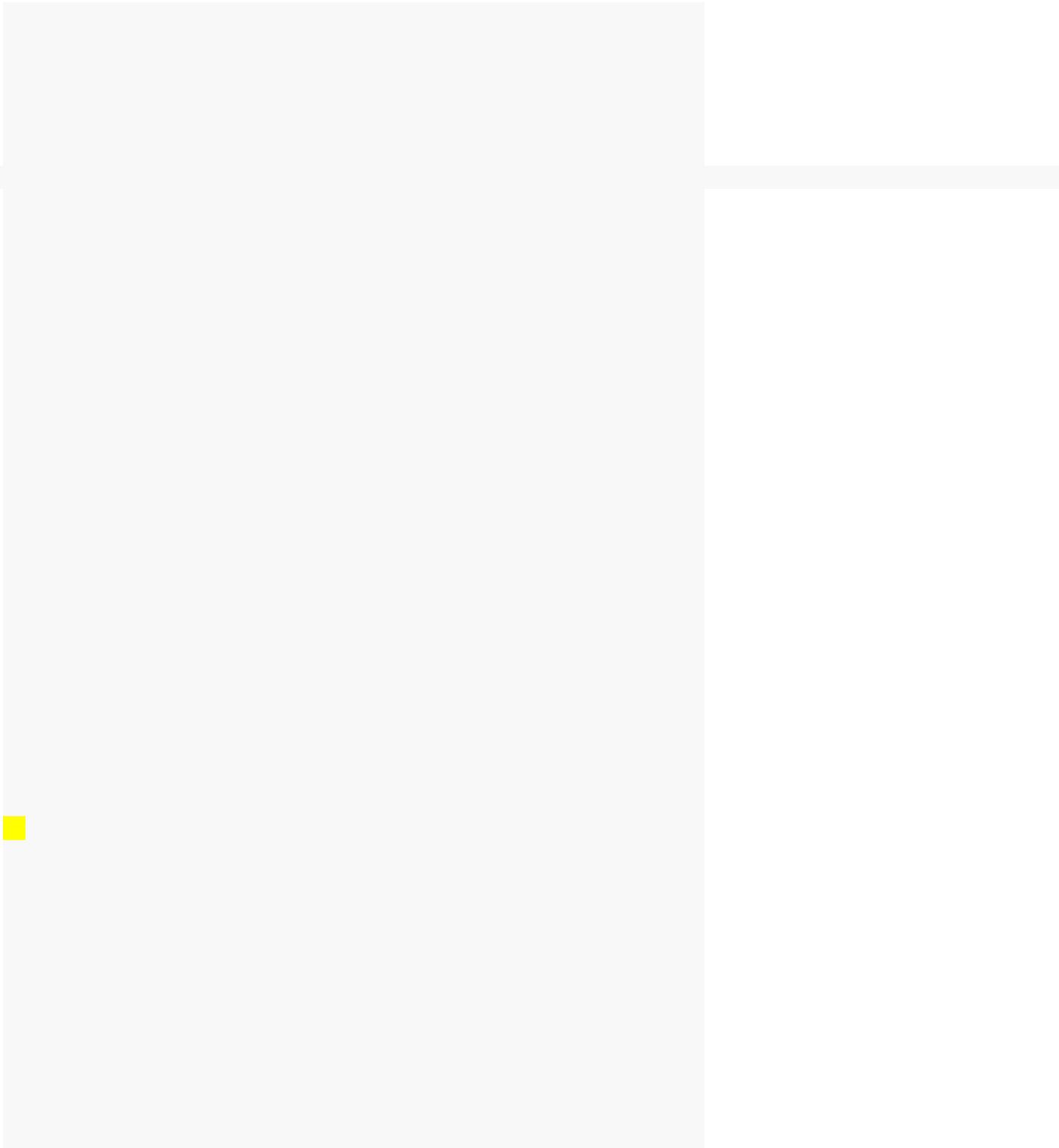
**Model Selection**

I used R studio to assist in the model selection process. Initially I created two tables to show the best outcomes of the “All-Possible Models” method. These tables allowed us to compare the best models in regards to the CP and adjusted R2 values.

* Tuition Enroll Dropout Income CPI UnEmp Stock



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 24.2027710 |
| ## | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 10.6970569 |
| ## | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 0.4164587 |
| ## | 4 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 5 | 2.3175617 |
| ## | 5 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 6 | 4.1114880 |
| ## | 6 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7 | 6.0224789 |
| ## | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 | 8.0000000 |



* Tuition Enroll Dropout Income CPI UnEmp Stock

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0.9681403 |
| ## | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0.9752454 |
| ## | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 0.9810253 |
| ## | 4 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 5 | 0.9805916 |
| ## | 5 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 6 | 0.9801939 |
| ## | 6 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7 | 0.9797087 |
| ## | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 | 0.9791581 |

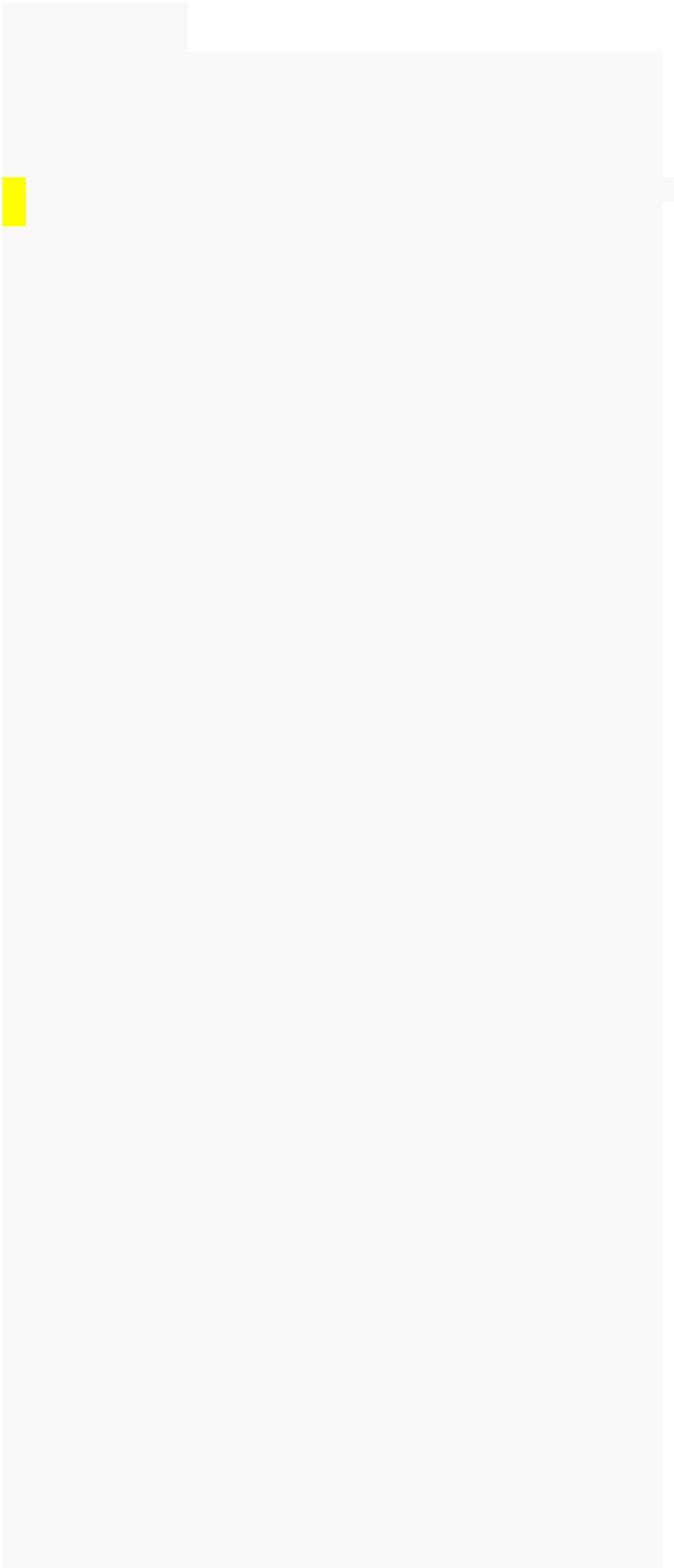
We can see one model that stands out amongst the rest. This model includes three predictors: Tuition, Enroll, and Income. This model has the lowest CP value of the group (0.4165), which is well below the corresponding p value of 4. Additionally, this model has the highest adjusted R2 value (0.9810). These results confirm our initial observations using the correlation and scatterplots. I also used R Studio to perform the “Stepwise” procedure. See appendix for the complete results of this procedure.

* Start: AIC=915.18
* Loan ~ 1
* Step: AIC=764.5
* Loan ~ Enroll
* Step: AIC=754.34
* Loan ~ Enroll + Tuition
* Step: AIC=743.55
* Loan ~ Enroll + Tuition + Income

The results of the “Stepwise” procedure are identical to the results from the “All-Possible Models” procedure. Using this knowledge, we can run a MLR using this model:

Loan ~ Enroll + Tuition + Income

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | | |  |  |  | 5 |  |
| ## Coefficients: | | |  |  |  |  |  |
| ## |  | Estimate Std. Error | | t value Pr(>|t|) | | \*\*\* |  |
| ## (Intercept) -1.000e+05 | | | 4.712e+03 | -21.224 | < 2e-16 |  |
| ## | Tuition | 4.128e+00 | 7.715e-01 | 5.350 | 3.86e-06 | \*\*\* |  |
| ## | Enroll | 9.949e-03 | 1.096e-03 | 9.079 | 2.90e-11 | \*\*\* |  |
| ## | Income | -5.883e-01 | 1.602e-01 | -3.673 | 0.000702 | \*\*\* |  |



* Residual standard error: 4475 on 40 degrees of freedom
* Multiple R-squared: 0.9823, Adjusted R-squared: 0.981
* F-statistic: 742.1 on 3 and 40 DF, p-value: < 2.2e-16

|  |  |  |  |
| --- | --- | --- | --- |
| **VIF** | |  |  |
| ## | Tuition | Enroll | Income |
| ## | 18.25814 | 15.98364 | 11.05937 |

The results of the MLR indicate that the model is valid overall and all three of the predictor variables are linearly related to Loan. However, looking at the VIF values, we can see that there is a multicollinearity issue, as all three corresponding VIF values are larger than 10.00. A series of transformations were performed (see appendix).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **VIF** | |  |  |  |  |  |  |  |
| ## | |  | Tuition |  | Enroll | log(Income) | | |  |
| ## | |  | 14.662244 | 16.309684 | | 4.352248 | | |  |
|  |  |  | log(Tuition) | | Enroll |  | log(Income) | |  |
| ## |  |  |  |
| ## | |  | 10.620169 |  | 9.676720 | |  | 5.022638 |  |
| ## | |  | log(Tuition) | | log(Enroll) | | | log(Income) |  |
| ## | |  | 9.088191 |  | 12.513052 | |  | 6.648182 |  |

After performing log transformations on each of the variables, we were able to compare the models to determine which were the most successful in reducing the multicollinearity issue. The transformations that produced the lowest VIF values were: log(Income) and log(Tuition). By transforming these two variables, we were able to build a model without a severe multicollinearity issue.

Loan ~ log(Tuition) + Enroll + log(Income)

The corresponding VIF values for this model are: 10.62, 9.68, and 5.02, respectively. A VIF value greater than or equal to 10.00 indicates a potential multicollinearity issue, however, we feel that VIF value of 10.62 for log(Tuition) is acceptable moving forward.

Looking at the diagnostic plots for this model (see appendix), it appears that all of the MLR assumptions are supported. In the residuals vs. fits plot, there is an even distribution of points centered around zero. This suggests that the linearity, constant variance, and independence assumptions are supported. Additionally, the Normal-Q-Q plot shows that the standardized residuals follow a mostly-straight line, which indicates that the normality assumption is supported as well. We can run formal tests to check these assumptions further.

**Formal Assumptions Tests**

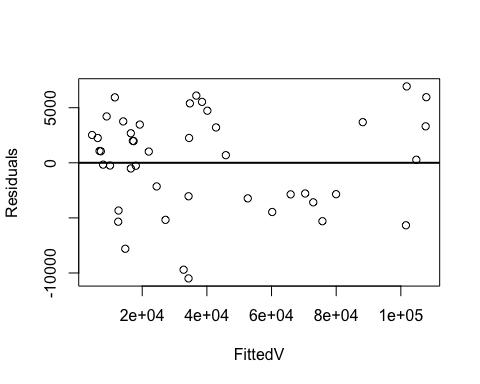
***Linearity***

The linearity assumption of data can be checked from graph of residuals vs fitted from this model. From the plot below, we can see that, in general, there is a linear band of points. Additionally,

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 6 |

the predictor variables are significantly linearly related to the response variable. Taken together, this suggests that the linearity assumption is supported and no remedies are required.

I was unable to run a “lack-of-fit” test to test this assumption formally. This test can be performed only if response variable has multiple values at each combination of the levels of the different predictor variables; there is no such situation in this model, hence we cannot perform this formal test.



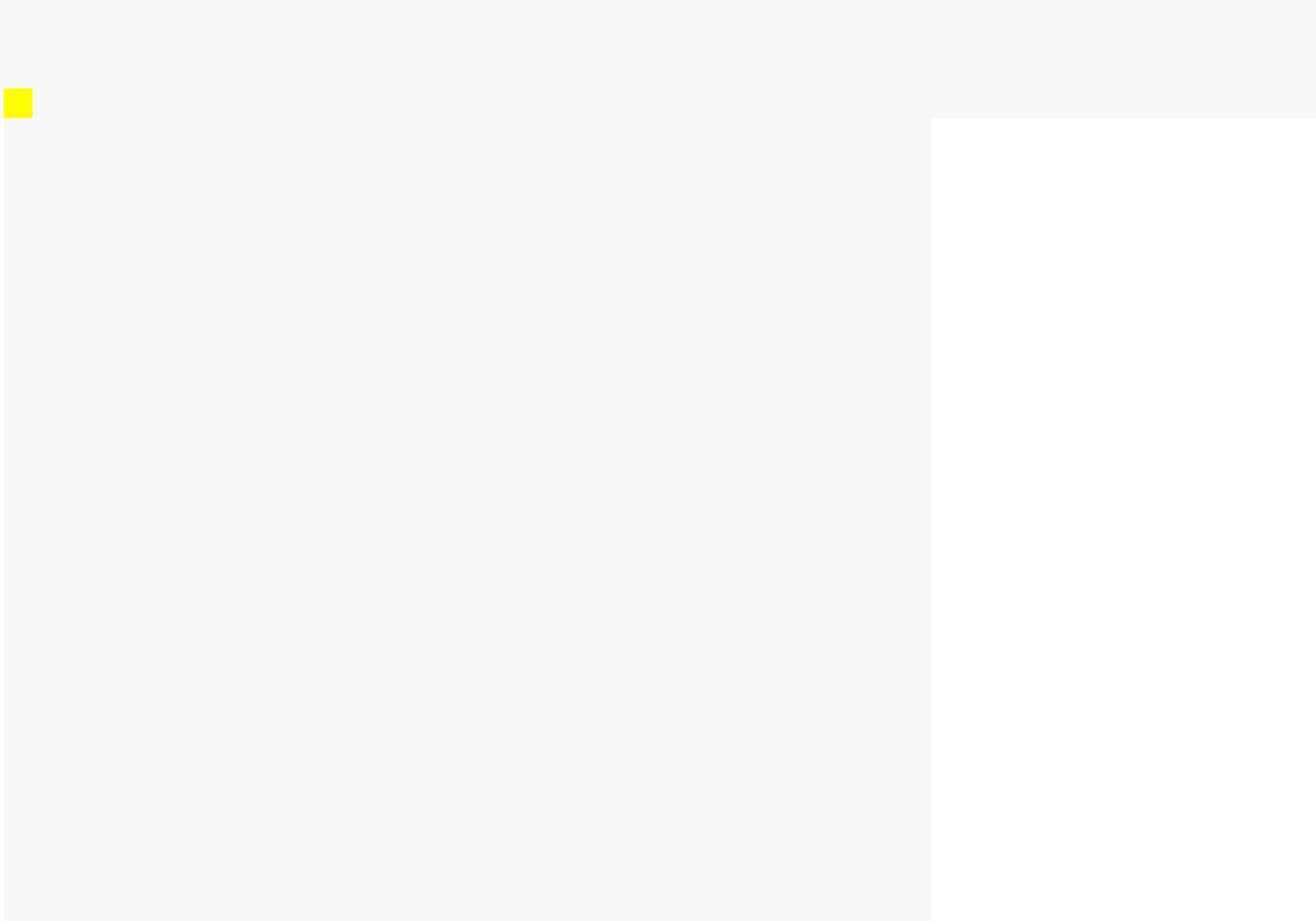
***Constant Variance***

Constant variance assumption of a model can be checked visually by observing the residual vs fitted graph from the model or by calculating Szroeter’s test statistic. Here we can follow both methods to test constant variance assumption. From the graph above, it can be stated that model is following constant variance assumption, as we can see an even distribution of values centered around zero. We can verify this observation using a formal test. The hypotheses from the “Szroeter” test are indicated below:

H0: These is a constant Variance

Ha: There is no constant Variance

Q



## [1] -0.8090251

**pnorm**(Q,lower.tail =FALSE)*#Fail to Reject = Variance is Constant*

## [1] 0.7907496

Since the p-value is greater than level of significance (α=.05), we fail to reject null hypothesis. Hence, we can state that model has met constant variance assumption. No further transformations are needed in regards to constant variance.

***Normality***

Normality assumption test can be carried out for a model either by observing the model’s Normal Q-Q plot or by using a formal test. We can follow both methods in here.

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 7 |

From the graph above, we can see that the model follows normality assumption since the points are distributed about the straight line. A formal hypothesis test can be stated as:

H0: The distribution is normal

Ha: The distribution is not normal

* Anderson-Darling normality test
* A = 0.52504, p-value = 0.1713



Since the p- value from the Anderson-Darling test is greater than the level of significance (α =

.05), we fail to reject the null hypothesis. Thus, we can state that the normality assumption is not violated.

***Independence***

The error terms independence assumption can be tested using a formal hypothesis test. The hypotheses can be given as follows:

H0: The autocorrelation of the disturbances is 0

Ha: The autocorrelation of the disturbances is not 0

* Durbin-Watson test
* data: model3
* DW = 0.51266, p-value = 1.876e-11
* alternative hypothesis: true autocorrelation is greater than 0

Since the p-value from the Durbin-Watson test is less than the level of significance (α = .05), we reject the null hypothesis and state that the independence assumption is violated.

This can be expected in our model, given that the data is time-series. Most of the variables are steadily increasing with the time. I have attempted to remedy this issue by adding time to the model, both as a factor and as a lagged factor, but all the efforts were in vain. Adding these terms leads to a severe multicollinearity issue, reduces adjusted R2, and fails to correct the independence assumption (see appendix). This issue is common when using real-world data. I have chosen to continue with this model, disregarding the independence violation.

**Outliers**

An outlier is a value that deviates from the regression equation more than other observations. To compare the residuals among each other, we need to standardize these residuals. We can identify an outlier as its standardized residual exceeds the absolute value of 2.

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 8 |

From the plot above, we can see that the 22nd and 23rd observations are outliers in the model. Outliers may have effect on the regression equation, and thus the fitted values. I removed these values from our dataset and re-run the MLR model to determine their effect.

* Multiple R-squared: 0.9863, Adjusted R-squared: 0.9852
* F-statistic: 913 on 3 and 38 DF, p-value: < 2.2e-16

Looking at the new model, which excludes the outliers, we can see that there is not much improvement in values of adjusted R 2. Additionally, the p- values are similar with or without the outliers (see appendix) . Given that excluding the outliers has little impact on the model, I chose to leave them in the dataset.

**Interpretation**

Once again, our model is:

Loan = -276700 + 0.0116 \* Enroll + 34250 \* log(Tuition) – 12590 \* log(Income)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | |  |  |  | 9 |
|  | |  |  |  |  |
| ## Coefficients: | |  |  |  |  |
| ## | Estimate Std. Error t value Pr(>|t|) | | | |  |
| ## (Intercept) -2.767e+05 | | 6.111e+04 | -4.528 | 5.25e-05 \*\*\* | |
| ## Enroll | 1.160e-02 | 8.695e-04 | 13.341 | 2.59e-16 | \*\*\* |
| ## logTuition | 3.425e+04 | 7.966e+03 | 4.300 | 0.000107 | \*\*\* |
| ## logIncome | -1.259e+04 | 2.817e+03 | -4.468 | 6.34e-05 | \*\*\* |
| ## |  |  |  |  |  |

* Residual standard error: 4563 on 40 degrees of freedom
* Multiple R-squared: 0.9816, Adjusted R-squared: 0.9803
* F-statistic: 713.1 on 3 and 40 DF, p-value: < 2.2e-16

Using the F-statistic, we can conclude that this model is valid overall, *F(3,40)* = 713.1, *p* < 0 .05 R2 : 98.03% of variation in federal student loan debt is explained by this model (R2 = 0.9803)

Additionally, when t- test is carried out, p- values on coefficients are well below 5% significance level. Thus, we can conclude that the coefficients are significant at 95% confidence level.

Β1 : On average, holding all else constant, when enrollment increases by 10,000, federal student loan debt increases by $116 million (Β1 = 0.0116, *t(40) =* 13.341, *p* < 0.05)

Β2 : On average, holding all else constant, when tuition increases by 1%, federal student loan debt increases by $342.5 million (Β2 = 34250, *t(40)* = 4.300, *p* < 0.05)

Β3 : On average, holding all else constant, when household income increases by 1%, federal student loan debt decreases by $125.9 million (Β3 = -12590, *t(40) =* -4.468, *p* < 0.05)

Generally, the interpretations on coefficients comply with our initial perception of how predictor variables influence the federal student loan debt. The loan debt increases with the increase of enrollment and tuition, and decreases with the increase of household income.

**Conclusions**

**Predictions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***#2012*** | |  |  |  |
| ## | fit |  | lwr | upr |
| ## 1 | 104783 | 101333.6 108232.3 | | |
| ***#1999*** | |  |  |  |
| ## | fit | | lwr | upr |
| ## 1 | 42832.02 39419.59 46244.45 | | | |

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 10 |

To test how well the model can predict real-life values, I picked the data from year 1999 and 2012 to illustrate. We choose the year of 1999 because it is the year before a hike in loan. The year of 2012 represents a decrease in loan debt after recession.

The total loan debt in 2012 was $105060.10 million. This value is within the 95% confidence interval of ($101333.6, $108232.3 million) predicted from our model.

The total loan debt in 1999 was $46040.12 million. We are 99% sure that this value is within the confidence interval of ($39419.59, $46244.45 million).

In conclusion, our model can accurately predict the amount of loan within a selected confidence

level.

This model is used for predicting federal student loan debt based on the factors: tuition, enrollment, and income. We see that there is linear trend between the response variable and these predictors. The model can be used to predict the trend which the federal student loan debt will follow in the coming years.

Student loan debt has skyrocketed over the past 50 years, increasing from just $7 billion in 1970 to more than $950 billion today. If current borrowing patterns continue, student loan levels will reach $2 trillion in 2025! Though a college education remains the surest path to a middle-class life, evidence has begun to mount that the student loan debt may be far more detrimental to financial futures than once thought. This is particularly true for those with the highest levels of debt: students with high tuition fee and students from low-income families.

Looking at this model, we can predict the pattern of federal student debt for the coming years. By predicting this value in advance, precautionary methods can be taken to stabilize the economy and future of many American families.

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 11 |

**References**

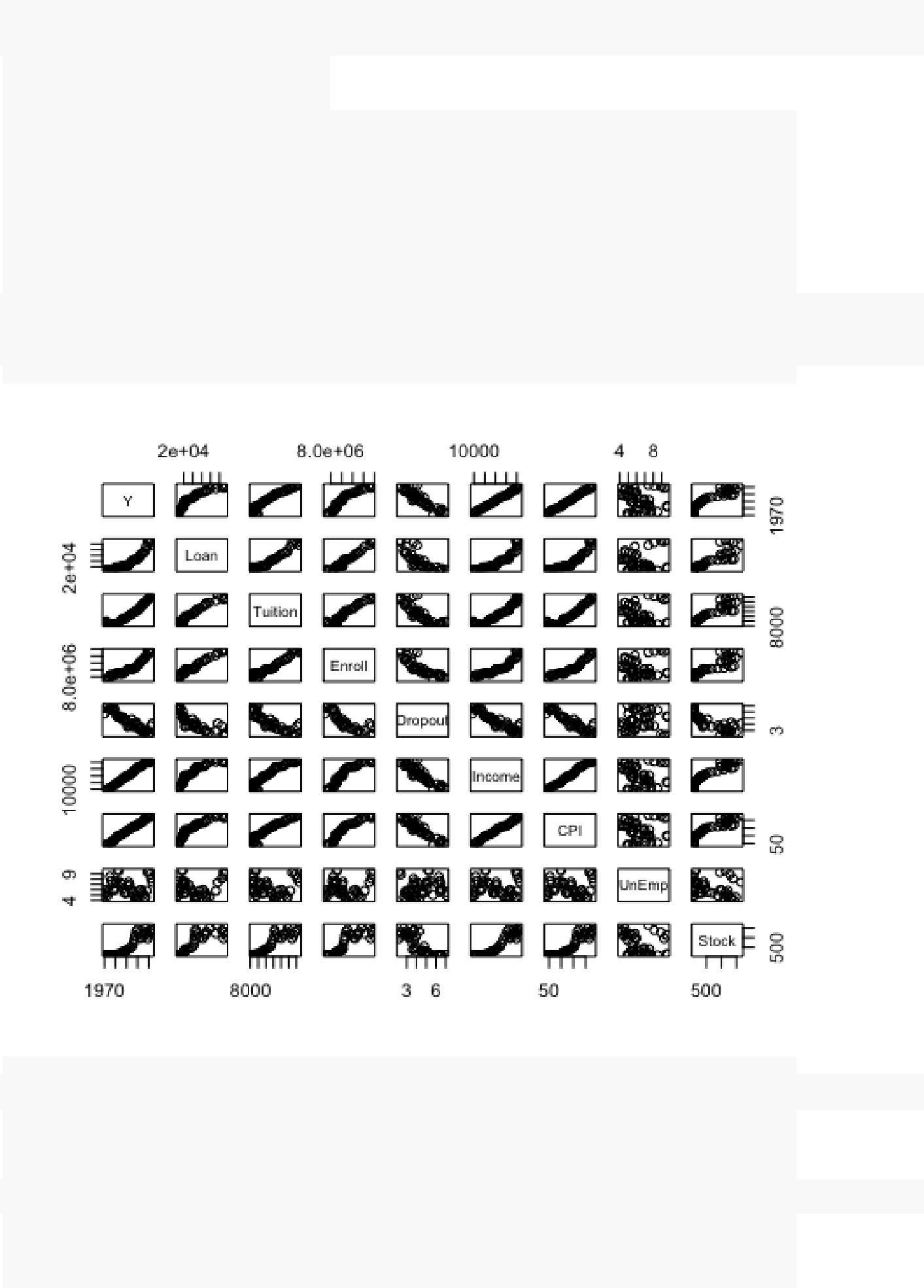
Denhart, C. (2013, August 7). How The $1.2 Trillion College Debt Crisis Is Crippling Students, Parents And The Economy. Retrieved December 4, 2015, from http://www.forbes.com/sites/specialfeatures/2013/08/07/how-the-college-debt-is-crippling-students-parents-and-the-economy/

Mitchell, C. (2014, June 9). Democrats, GOP spar on student loan debt. Retrieved December 5, 2015, from http://www.startribune.com/democrats-gop-spar-on-student-loan-debt/262466751/

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 12 |

**Appendix**

*# WORKING R SCRIPT*



########## PRELIMINARY ###############

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *# Import and Attach* | | | |  |  |  |  |  |  |  |
| Data | | <- **read.csv**("~/Documents/BGSU/Fall 2015/5020 - Regression/Project/Working dataset.csv") | | | | | | | | |
| **head**(Data) | | |  |  |  |  |  |  |  |  |
| ## |  | Y | Loan Tuition | | Enroll Dropout Income | | | CPI UnEmp | | Stock |
| ## | 1 | 1970 | 7014.68 | 9625 | 6996129 | 5.7 | 8689 | 39.0 | 5.0 | 83.45 |
| ## | 2 | 1971 | 8485.68 | 9729 | 7148575 | 6.6 | 8965 | 40.7 | 6.0 | 98.21 |
| ## | 3 | 1972 | 8251.56 | 9914 | 7253712 | 7.1 | 9624 | 41.9 | 5.6 | 109.78 |
| ## | 4 | 1973 | 7867.97 | 9443 | 7453467 | 7.1 | 10471 | 44.3 | 4.9 | 106.51 |
| ## | 5 | 1974 | 7732.93 | 8858 | 7805454 | 6.4 | 11147 | 49.4 | 5.6 | 81.48 |
| ## | 6 | 1975 | 6957.83 | 8774 | 8479688 | 6.3 | 11724 | 54.2 | 8.5 | 87.13 |
|  |  |  |  |  |  |  |  |  |  |  |

**attach**(Data)

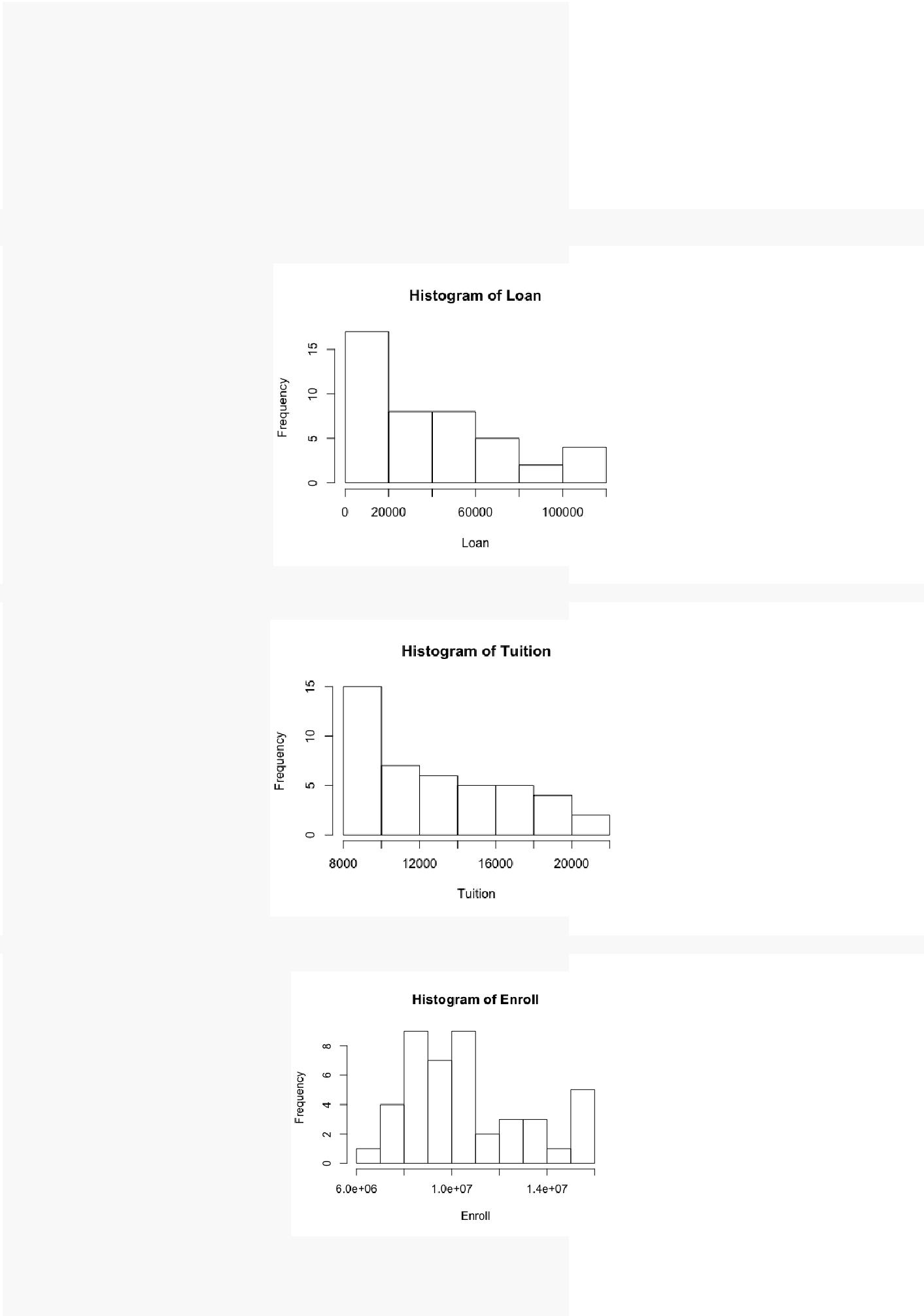
*# Matrix Plots and Correlations* **plot**(Data)

**cor**(Data[**c**(2:9)])

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Loan | Tuition | Enroll | Dropout | Income |
| ## Loan | 1.0000000 | 0.972431574 | 0.9843176 | -0.835002740 | 0.91654090 |
| ## Tuition | 0.9724316 | 1.000000000 | 0.9647447 | -0.858168505 | 0.94862561 |
| ## Enroll | 0.9843176 | 0.964744749 | 1.0000000 | -0.868093612 | 0.94108703 |

* Dropout -0.8350027 -0.858168505 -0.8680936 1.000000000 -0.93019162
* Income 0.9165409 0.948625615 0.9410870 -0.930191622 1.00000000

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | | |  |  | 13 |
|  |  |  |  |  |  |
| ## CPI | 0.9253015 | 0.948139116 | 0.9512519 | -0.933442369 | 0.99624575 |
| ## UnEmp | 0.1413104 -0.008510763 | | 0.1697659 | 0.004666452 -0.07567634 | |
| ## Stock | 0.8735617 | 0.929550176 | 0.8781773 | -0.809468650 | 0.93682775 |
| ## | CPI | UnEmp | Stock | |  |
| ## Loan | 0.925301514 | 0.141310388 | 0.8735617 | |  |
| ## Tuition | 0.948139116 | -0.008510763 | 0.9295502 | |  |
| ## Enroll | 0.951251923 | 0.169765861 | 0.8781773 | |  |
| ## Dropout -0.933442369 | | 0.004666452 -0.8094687 | | |  |
| ## Income | 0.996245749 | -0.075676340 | 0.9368277 | |  |
| ## CPI | 1.000000000 | -0.008880543 | 0.9157745 | |  |
| ## UnEmp | -0.008880543 | 1.000000000 -0.2246500 | | |  |
| ## Stock | 0.915774482 | -0.224649999 | 1.0000000 | |  |
|  |  |  |  |  |  |

*# Individual Histograms*

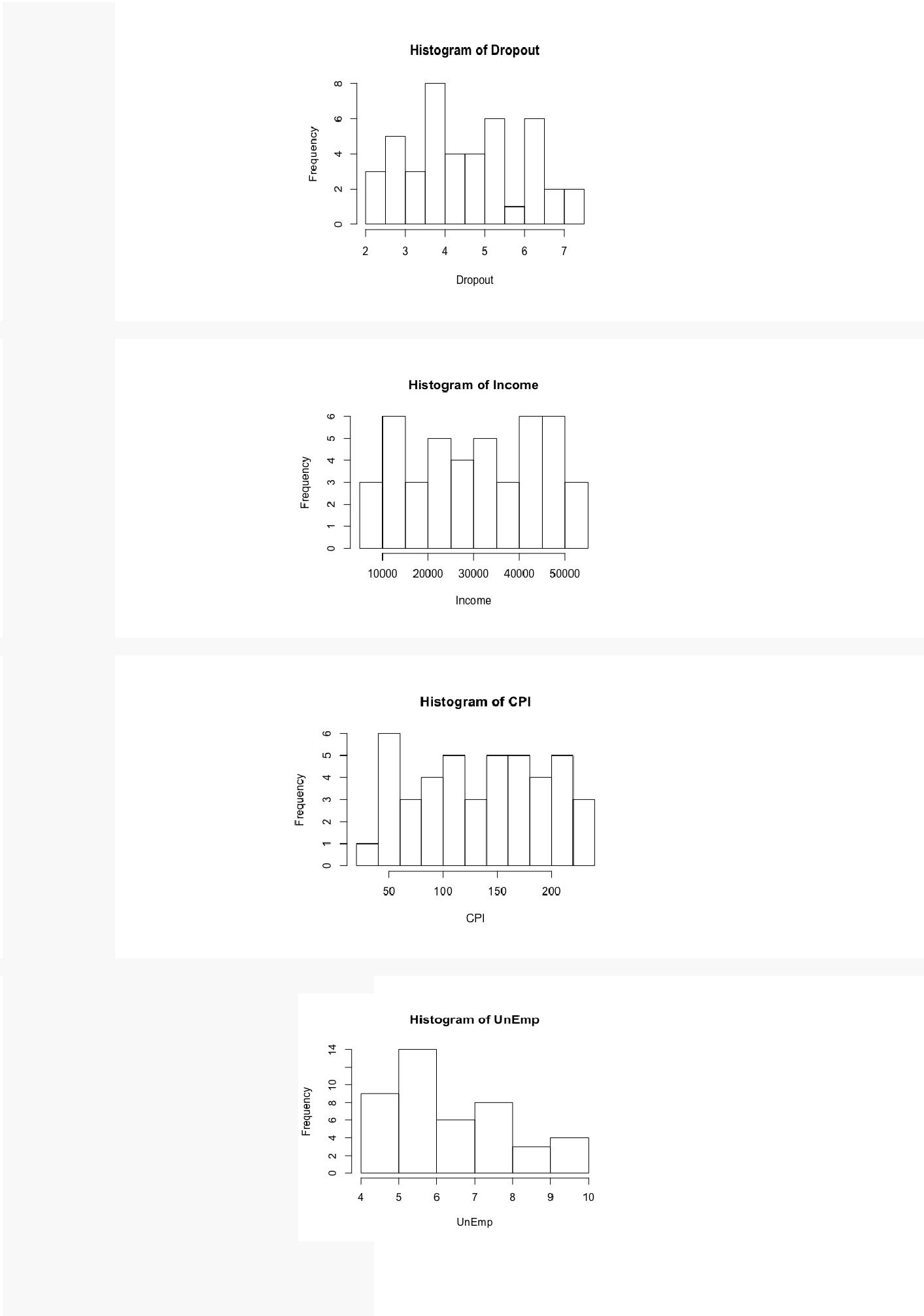
**hist**(Loan)*#Skewed: possible transformation*

**hist**(Tuition)*#Skewed: possible transformation*

**hist**(Enroll)*#Skewed: possible transformation*



|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 14 |
|  |  |
| **hist**(Dropout) |  |

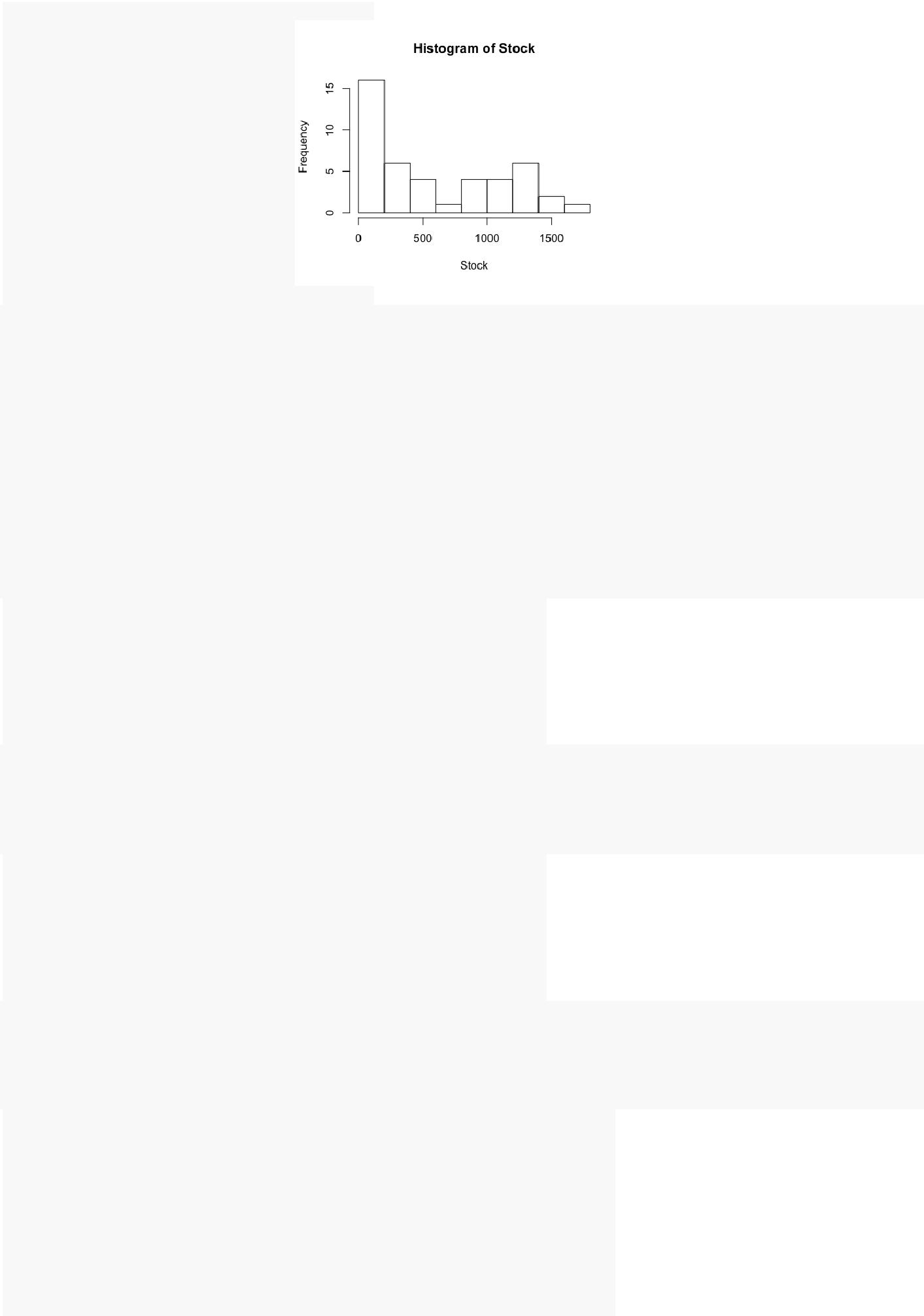


**hist**(Income)

**hist**(CPI)

**hist**(UnEmp)*#Skewed: possible transformation*

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 15 |
|  |  |
| **hist**(Stock)*#Skewed: possible transformation* |  |



########## MODEL SELECTION ###############

*#Define Models*

Data.Mean <- **lm**(Loan~1)

Data.Full <- **lm**(Loan~Tuition+Enroll+Dropout+Income+UnEmp+Stock)

*#install.packages("leaps")*

**library**("leaps")

*#All-Possible*

CP <- **leaps**(x=Data[**c**(3:9)], y=Loan, nbest=TRUE, method=**c**("Cp"), names=**c**("Tuition","Enroll","Dropout","Inco me","CPI","UnEmp","Stock"))

tableCP<-**cbind**(CP$which, CP$size, CP$Cp ) tableCP

* Tuition Enroll Dropout Income CPI UnEmp Stock

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 24.2027710 |
| ## | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 10.6970569 |
| ## | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 0.4164587 |
| ## | 4 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 5 | 2.3175617 |
| ## | 5 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 6 | 4.1114880 |
| ## | 6 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7 | 6.0224789 |
| ## | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 | 8.0000000 |
|  |  |  |  |  |  |  |  |  |  |  |

R2 <- **leaps**(x=Data[**c**(3:9)], y=Loan, nbest=TRUE, method=**c**("adjr2"), names=**c**("Tuition","Enroll","Dropout","I ncome","CPI","UnEmp","Stock"))

tableR2 <-**cbind**(R2$which, R2$size, R2$adjr2) tableR2

* Tuition Enroll Dropout Income CPI UnEmp Stock

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0.9681403 |
| ## | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0.9752454 |
| ## | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 0.9810253 |
| ## | 4 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 5 | 0.9805916 |
| ## | 5 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 6 | 0.9801939 |
| ## | 6 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7 | 0.9797087 |
| ## | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 | 0.9791581 |
|  |  |  |  |  |  |  |  |  |  |  |

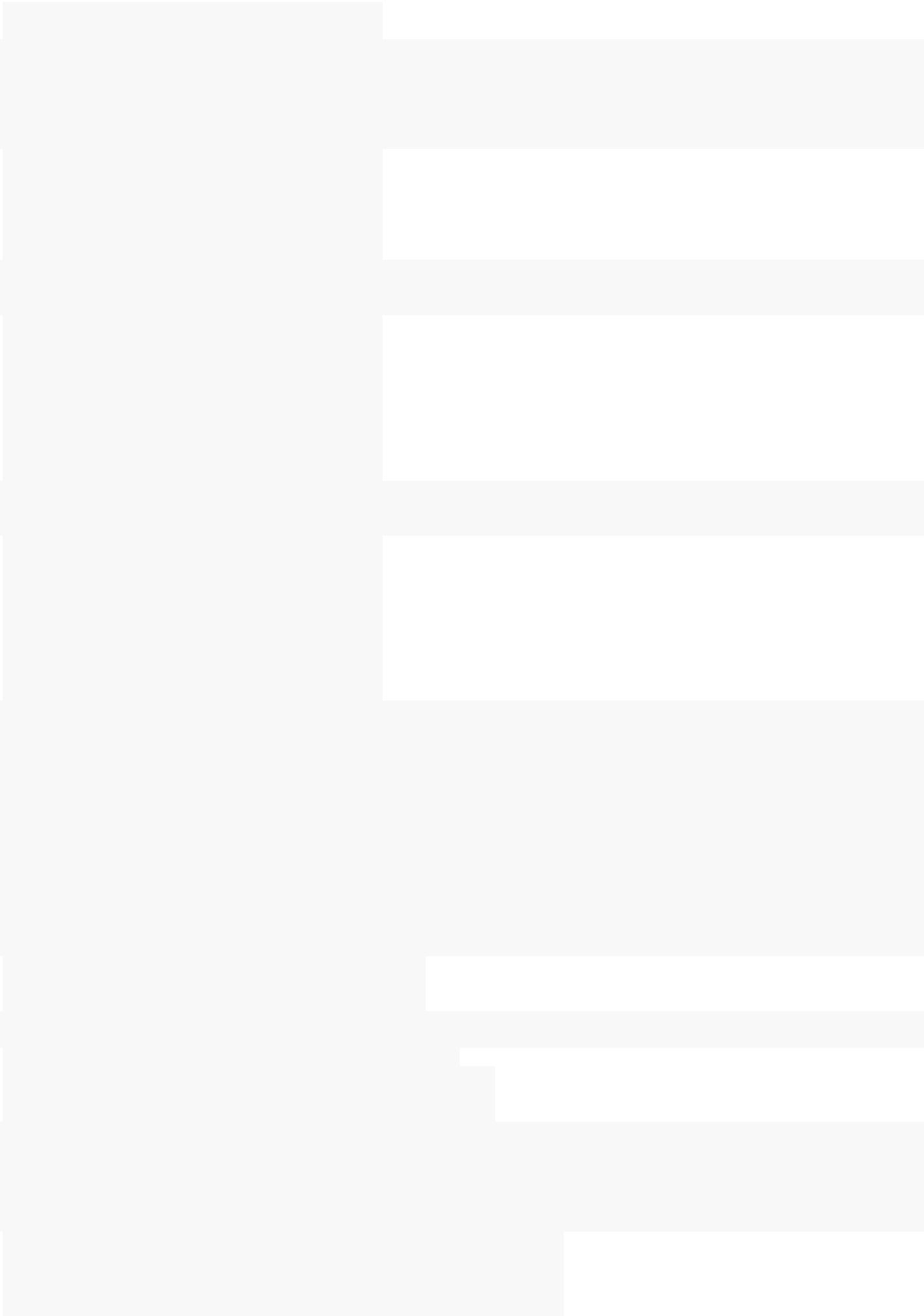
*# Stepwise*

St.Model <- **step**(Data.Mean, scope=**formula**(Data.Full), direction="both")

* Start: AIC=915.18
* Loan ~ 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  |  |  |  |
| ## | Df | Sum of Sq | RSS | AIC |
| ## + Enroll | 1 | 4.3966e+10 | 1.4121e+09 | 764.50 |
| ## + Tuition | 1 | 4.2911e+10 | 2.4675e+09 | 789.06 |
| ## + Income | 1 | 3.8120e+10 | 7.2584e+09 | 836.53 |
| ## + Stock | 1 | 3.4629e+10 | 1.0750e+10 | 853.81 |
| ## + Dropout | 1 | 3.1639e+10 | 1.3739e+10 | 864.61 |

|  |  |  |
| --- | --- | --- |
| STUDENT LOAN DEBT | | 16 |
|  |  |  |
| ## <none> | 4.5378e+10 | 915.18 |
| ## + UnEmp | 1 9.0614e+08 4.4472e+10 | 916.29 |
| ## |  |  |

* Step: AIC=764.5
* Loan ~ Enroll

##

## Df Sum of Sq RSS AIC

* + Tuition 1 3.4104e+08 1.0711e+09 754.34
* + Dropout 1 6.9860e+07 1.3423e+09 764.27

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## <none> |  |  | 1.4121e+09 | 764.50 |
| ## + Income | 1 | 3.8014e+07 | 1.3741e+09 | 765.30 |
| ## + UnEmp | 1 | 3.1085e+07 | 1.3810e+09 | 765.52 |
| ## + Stock | 1 | 1.6627e+07 | 1.3955e+09 | 765.98 |
| ## - Enroll | 1 | 4.3966e+10 | 4.5378e+10 | 915.18 |
| ## |  |  |  |  |

* Step: AIC=754.34
* Loan ~ Enroll + Tuition

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | Df | Sum of Sq | RSS | AIC |
| ## + Income | 1 | 270107008 | 800965927 | 743.55 |
| ## + Dropout | 1 | 130541000 | 940531936 | 750.62 |
| ## + Stock | 1 | 111831834 | 959241101 | 751.49 |
| ## + UnEmp | 1 | 80150064 | 990922871 | 752.92 |
| ## <none> |  |  | 1071072935 | 754.34 |
| ## - Tuition | 1 | 341042185 | 1412115120 | 764.50 |
| ## - Enroll | 1 | 1396448858 | 2467521793 | 789.06 |
| ## |  |  |  |  |

* Step: AIC=743.55
* Loan ~ Enroll + Tuition + Income

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | Df | Sum of Sq | RSS | AIC |
| ## <none> |  |  | 800965927 | 743.55 |
| ## + Stock | 1 | 2175202 | 798790725 | 745.43 |
| ## + UnEmp | 1 | 1585105 | 799380822 | 745.47 |
| ## + Dropout | 1 | 44576 | 800921351 | 745.55 |
| ## - Income | 1 | 270107008 | 1071072935 | 754.34 |
| ## - Tuition | 1 | 573134832 | 1374100759 | 765.30 |
| ## - Enroll | 1 | 1650602144 | 2451568071 | 790.78 |
|  |  |  |  |  |

########## INITIAL CHECKS ###############

*#install.packages("HH")*

**library**("HH")

*#Run Best Model*

best.model <- **lm**(Loan~Tuition+Enroll+Income) **summary**(best.model)

##

* Call:
* lm(formula = Loan ~ Tuition + Enroll + Income)
* Residuals:

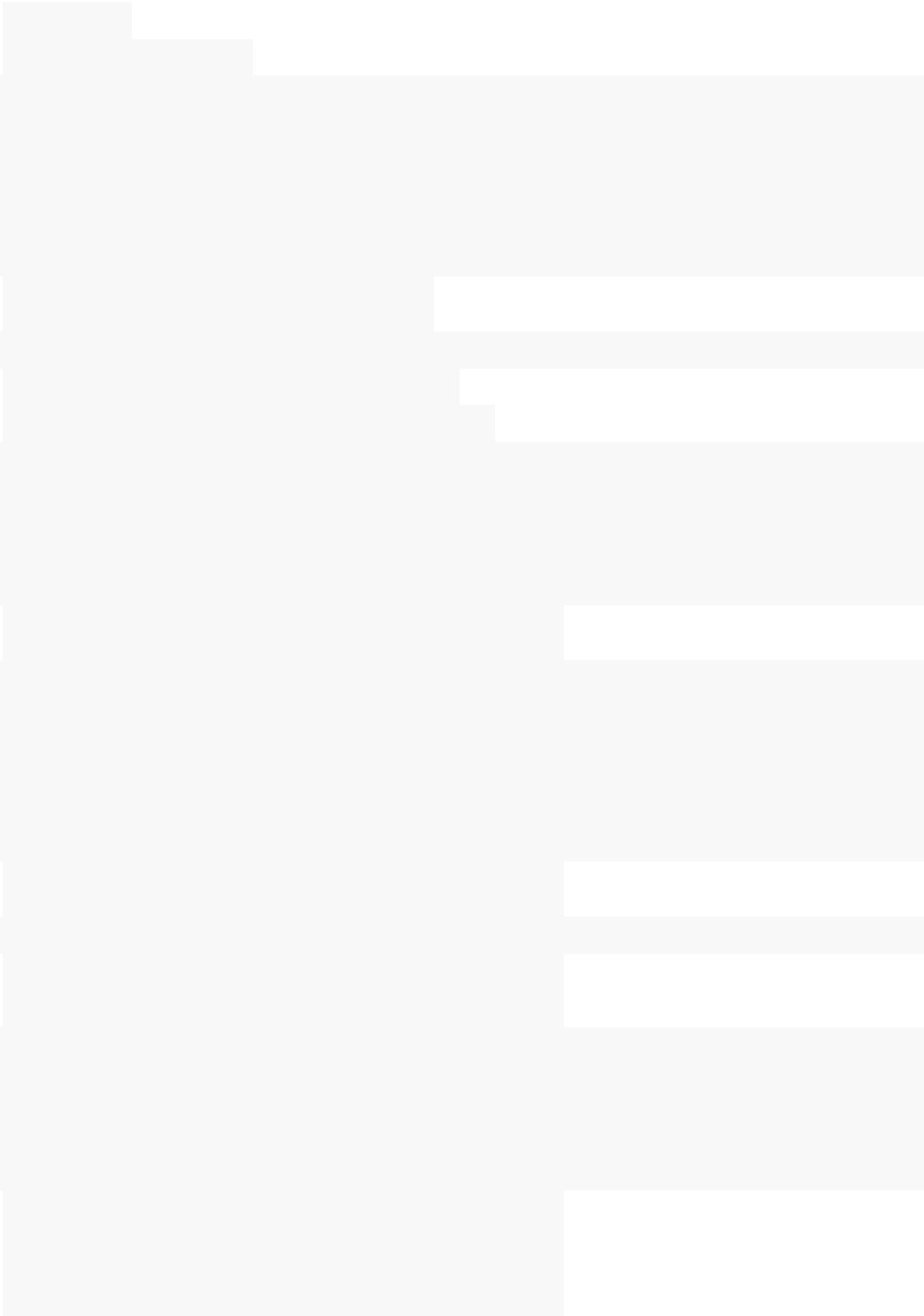
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -10136.0 | | -2726.2 | 541.1 | 3131.5 | 8011.1 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -1.000e+05 | | | 4.712e+03 | -21.224 | < 2e-16 | \*\*\* |
| ## Tuition | | 4.128e+00 | 7.715e-01 | 5.350 | 3.86e-06 | \*\*\* |
| ## | Enroll | 9.949e-03 | 1.096e-03 | 9.079 | 2.90e-11 | \*\*\* |
| ## | Income | -5.883e-01 | 1.602e-01 | -3.673 | 0.000702 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4475 on 40 degrees of freedom
* Multiple R-squared: 0.9823, Adjusted R-squared: 0.981
* F-statistic: 742.1 on 3 and 40 DF, p-value: < 2.2e-16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | | | | 17 |
|  | | |  |  |
| **vif**(best.model) | | |  |  |
| ## | Tuition | Enroll | Income |  |
| ## | 18.25814 | 15.98364 | 11.05937 |  |
|  |  |  |  |  |

*#Transform Income*

Income2 <- **log**(Income)

model2 <- **lm**(Loan~Tuition+Enroll+Income2) **summary**(model2)

##

* Call:
* lm(formula = Loan ~ Tuition + Enroll + Income2)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -9625.6 | | -2956.1 | 783.8 | 3281.4 | 7947.3 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -1.082e+04 | | | 2.149e+04 | -0.504 | 0.617279 |  |
| ## Tuition | | 3.145e+00 | 6.926e-01 | 4.540 | 5.06e-05 | \*\*\* |
| ## | Enroll | 1.007e-02 | 1.109e-03 | 9.079 | 2.90e-11 | \*\*\* |
| ## | Income2 | -9.395e+03 | 2.576e+03 | -3.648 | 0.000756 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4483 on 40 degrees of freedom
* Multiple R-squared: 0.9823, Adjusted R-squared: 0.981
* F-statistic: 739.5 on 3 and 40 DF, p-value: < 2.2e-16

**vif**(model2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | Tuition | Enroll | Income2 |
| ## | 14.662244 | 16.309684 | 4.352248 |
|  |  |  |  |

*#Transform Tuition*

Tuition2 <- **log**(Tuition)

model3 <- **lm**(Loan~Tuition2+Enroll+Income2) **summary**(model3)

##

* Call:
* lm(formula = Loan ~ Tuition2 + Enroll + Income2)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -10491.5 | | -3085.5 | 855.3 | 3348.8 | 6933.1 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -2.767e+05 | | | 6.111e+04 | -4.528 | 5.25e-05 | \*\*\* |
| ## Tuition2 | | 3.425e+04 | 7.966e+03 | 4.300 | 0.000107 | \*\*\* |
| ## | Enroll | 1.160e-02 | 8.695e-04 | 13.341 | 2.59e-16 | \*\*\* |
| ## | Income2 | -1.259e+04 | 2.817e+03 | -4.468 | 6.34e-05 | \*\*\* |

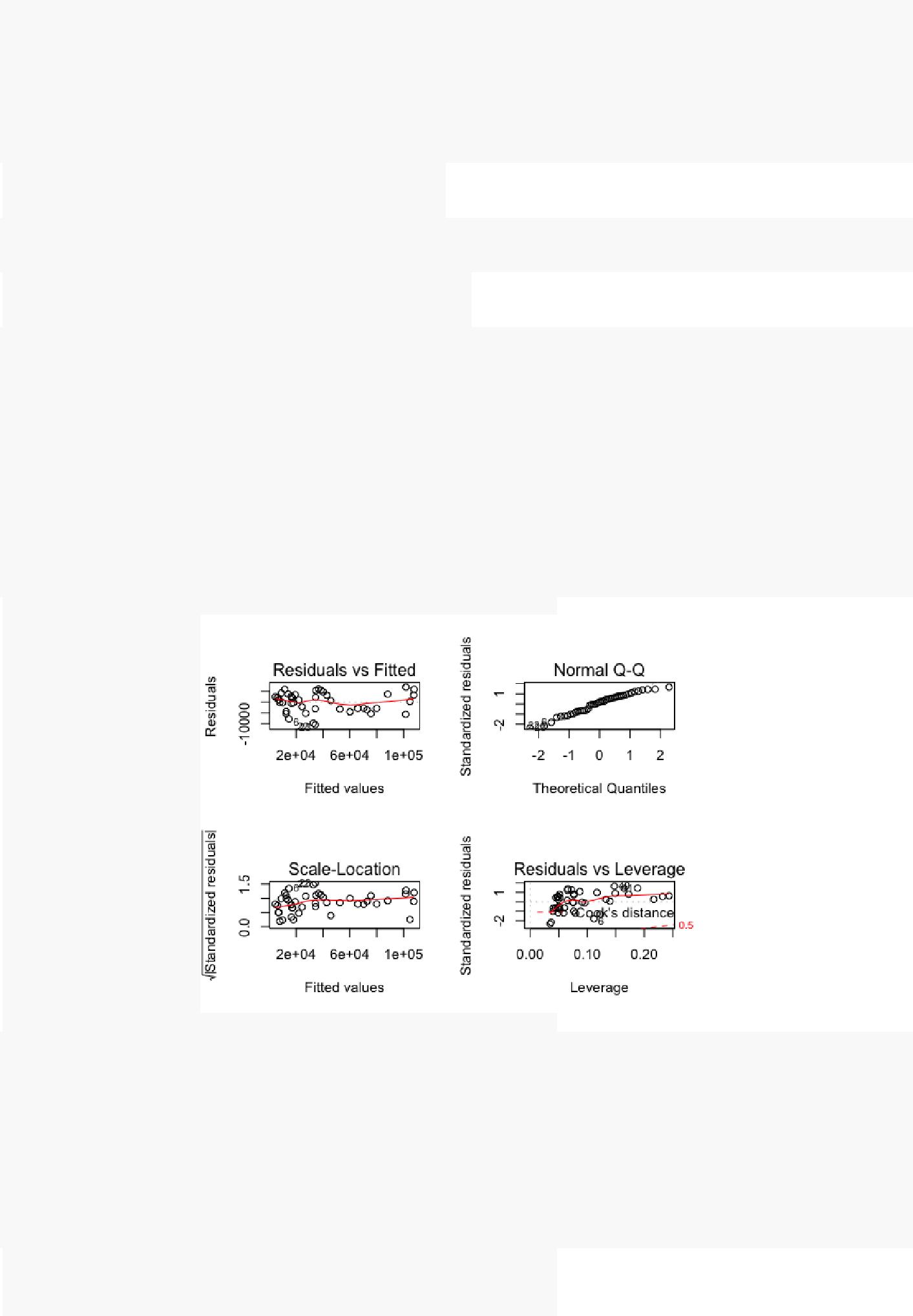
* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4563 on 40 degrees of freedom
* Multiple R-squared: 0.9816, Adjusted R-squared: 0.9803
* F-statistic: 713.1 on 3 and 40 DF, p-value: < 2.2e-16

**vif**(model3)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | Tuition2 | Enroll | Income2 |
| ## | 10.620169 | 9.676720 | 5.022638 |

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 18 |

*#Transform Enroll*



Enroll2 <- **log**(Enroll)

model4 <- **lm**(Loan~Tuition2+Enroll2+Income2) **summary**(model4)

* Call:
* lm(formula = Loan ~ Tuition2 + Enroll2 + Income2)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -11611 | | -4718 | 1046 | 3923 | 9007 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)
* (Intercept) -2381230 135079 -17.628 < 2e-16 \*\*\*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## Tuition2 | | 48722 | 8830 | 5.518 | 2.25e-06 | \*\*\* |
| ## | Enroll2 | 137526 | 13005 | 10.575 | 3.76e-13 | \*\*\* |
| ## | Income2 | -25318 | 3883 | -6.520 | 8.79e-08 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 5468 on 40 degrees of freedom
* Multiple R-squared: 0.9736, Adjusted R-squared: 0.9717
* F-statistic: 492.6 on 3 and 40 DF, p-value: < 2.2e-16

**vif**(model4)

## Tuition2 Enroll2 Income2

* 9.088191 12.513052 6.648182

*#Diagnostic Plot* **par**(mfrow=**c**(2,2))

**plot**(model3)

########### FORMAL TESTS ############

*#install.packages("alr3")* **library**("alr3")

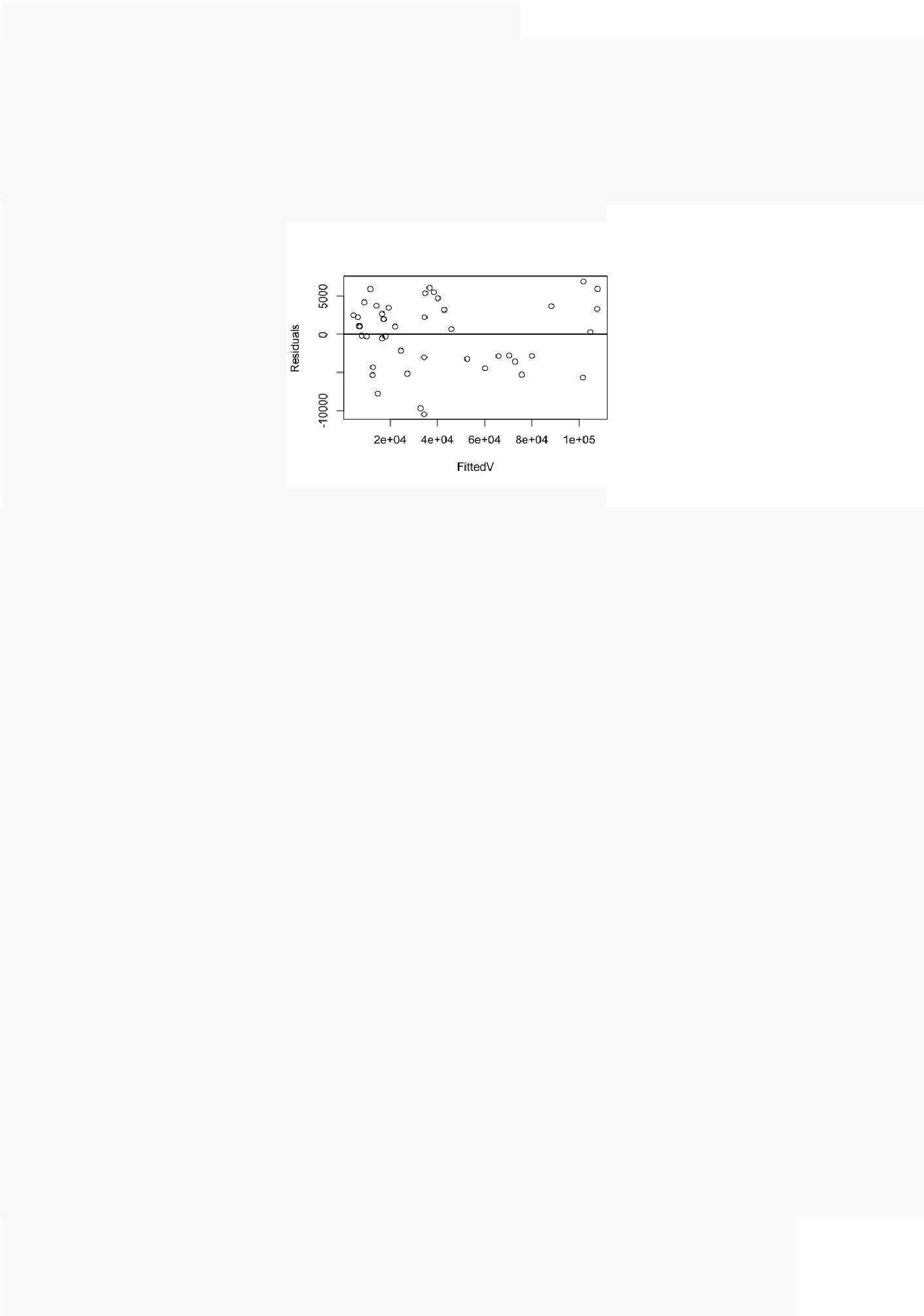
*# Linearity*

**pureErrorAnova**(model3)*#can't perform lack of fit test*

* Analysis of Variance Table
* Response: Loan

## Df Sum Sq Mean Sq F value Pr(>F)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | | |  |  |  | 19 |
|  |  |  |  |  |  |  |
| ## Tuition2 | 1 | 4.0812e+10 | 4.0812e+10 | 1959.874 | < 2.2e-16 | \*\*\* |
| ## Enroll | 1 | 3.3176e+09 | 3.3176e+09 | 159.316 | 1.565e-15 | \*\*\* |
| ## Income2 | 1 | 4.1570e+08 | 4.1570e+08 | 19.963 | 6.336e-05 | \*\*\* |

* Residuals 40 8.3295e+08 2.0824e+07
* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*# Constant Variance*

FitRes=**data.frame**(FittedV=**fitted**(model3), Residuals=**residuals**(model3)) **par**(mfrow=**c**(1,1))

**plot**(FitRes) **abline**(a=0,b=0,lwd=2)

FitRes=FitRes[**order**(FitRes$FittedV, decreasing=T),] *#Res appear to be decreasing across fits* n=**nrow**(Data)

FitRes$ID=**c**(1:n)

h=**sum**(FitRes$ID\*FitRes$Residuals^2)/**sum**(FitRes$Residuals^2) Q=**sqrt**(6\*n/(n^2-1))\*(h-(n+1)/2)

Q

## [1] -0.8090251

**pnorm**(Q,lower.tail =FALSE)*#Fail to Reject = Variance is Constant*

## [1] 0.7907496

*# Independence*

*#install.packages("lmtest")* **library**("lmtest")

**dwtest**(model3)*#Reject = Independence FAILS*

##

* Durbin-Watson test
* data: model3
* DW = 0.51266, p-value = 1.876e-11
* alternative hypothesis: true autocorrelation is greater than 0

*#install.packages("nortest")* **library**("nortest")

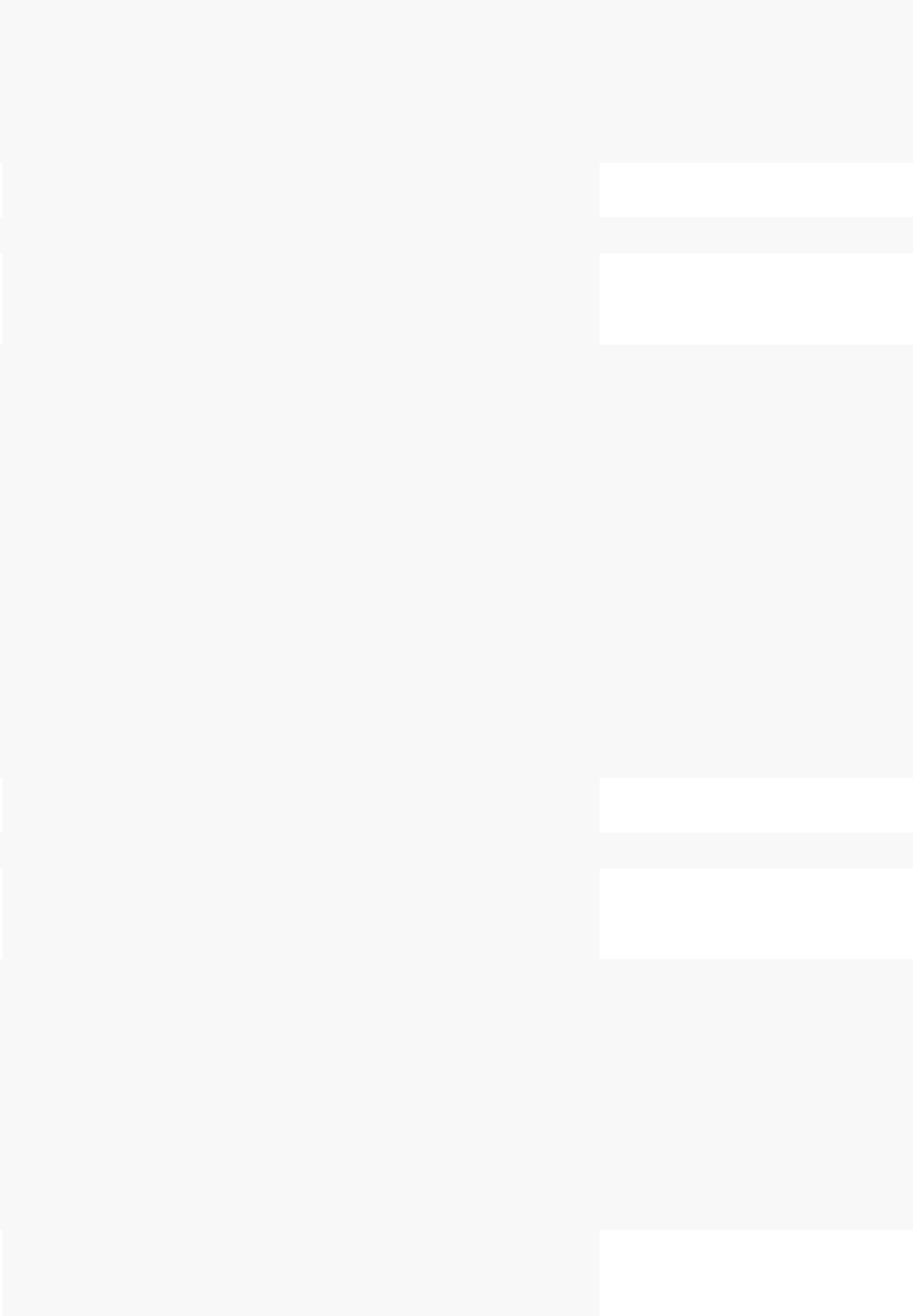
**ad.test**(model3$residuals)*#Fail to Reject = Normality Passes*

##

* Anderson-Darling normality test
* data: model3$residuals
* A = 0.52504, p-value = 0.1713

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 20 |

## Adding Time to the Model



modelT1 <- **lm**(Data$Loan~Tuition2+Data$Enroll+Income2+Data$Y) **summary**(modelT1)

##

* Call:
* lm(formula = Data$Loan ~ Tuition2 + Data$Enroll + Income2 + Data$Y)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -10468.4 | | -3098.4 | 856.5 | 3368.4 | 7002.7 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -3.963e+05 | | | 2.161e+06 | -0.183 | 0.855 |  |
| ## Tuition2 | | 3.339e+04 | 1.752e+04 | 1.906 | 0.064 . | |
| ## Data$Enroll | | 1.151e-02 | 1.825e-03 | 6.308 | 1.93e-07 | \*\*\* |
| ## | Income2 | -1.339e+04 | 1.488e+04 | -0.900 | 0.374 |  |
| ## | Data$Y | 6.874e+01 | 1.242e+03 | 0.055 | 0.956 |  |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4621 on 39 degrees of freedom
* Multiple R-squared: 0.9816, Adjusted R-squared: 0.9798
* F-statistic: 521.5 on 4 and 39 DF, p-value: < 2.2e-16

tut <- **log**(Data$Tuition[3:44]) inc <- **log**(Data$Income[3:44]) loan <- **log**(Data$Loan[3:44]) enr <- **log**(Data$Enroll[3:44]) y1 <- Y[2:43]

data2 <- **data.frame**(loan, tut, inc, y1, enr) **attach**(data2)

modelT2 <- **lm**(loan~tut+enr+inc+y1) **summary**(modelT2)

##

* Call:
* lm(formula = loan ~ tut + enr + inc + y1)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -0.26053 | | -0.08276 | 0.01906 | 0.07888 | 0.26486 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## (Intercept) -108.95688 | | | 58.44931 | -1.864 | 0.0703 . |
| ## tut | | -0.22365 | 0.60663 | -0.369 | 0.7145 |
| ## enr | | 0.31320 | 0.71366 | 0.439 | 0.6633 |
| ## | inc | 0.28145 | 0.40656 | 0.692 | 0.4931 |
| ## | y1 | 0.05696 | 0.03898 | 1.461 | 0.1524 |

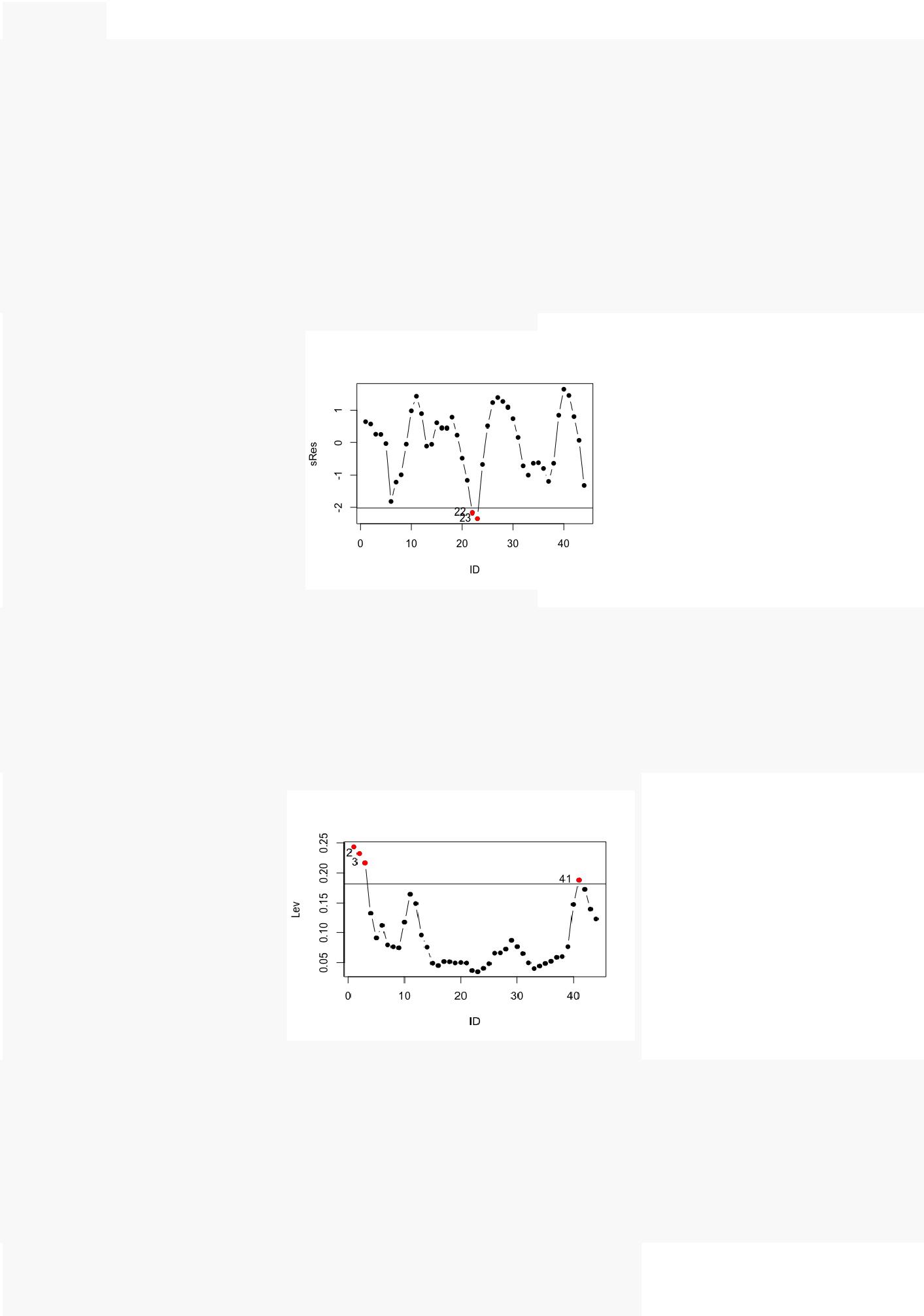
* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 0.136 on 37 degrees of freedom
* Multiple R-squared: 0.977, Adjusted R-squared: 0.9745
* F-statistic: 393 on 4 and 37 DF, p-value: < 2.2e-16

**dwtest**(modelT2)

##

* Durbin-Watson test
* data: modelT2
* DW = 0.55719, p-value = 1.054e-10
* alternative hypothesis: true autocorrelation is greater than 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STUDENT LOAN DEBT | | |  | 21 |
|  | |  |  |  |
| **vif**(modelT2) | |  |  |  |
| ## | tut | enr | inc | y1 |

* 66.83023 52.89154 94.49673 507.13034

######## OUTLIERS, LEV, INFLUENTIAL ###########

*# Outliers*

StdRes=**data.frame**(ID=**c**(1:**nrow**(Data)), sRes= **rstandard**(model3)) tval <- **qt**(0.975, (n-1-3)) Outliers=StdRes[**abs**(StdRes$sRes)>tval,] **plot**(StdRes,pch=16,type="b")

**points**(Outliers,col="red",pch=16) **abline**(h=tval)

**abline**(h=-tval)

**text**(Outliers,labels =Outliers$ID,pos=2)*# Two Outliers*

*# Leverage*

Leverage=**data.frame**(ID=**c**(1:**nrow**(Data)), Lev=**hatvalues**(model3)) LeverageLimit=(2\*(3+1))/**nrow**(Data) HighLeverage=Leverage[Leverage$Lev>LeverageLimit,] **plot**(Leverage,pch=16,type="b")

**points**(HighLeverage,col="red",pch=16) **abline**(h=LeverageLimit)

**text**(HighLeverage,labels =HighLeverage$ID,pos=2)*# 4 High Leverage Points*



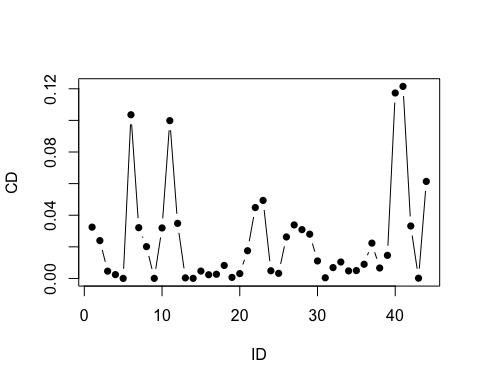
*#Inluential Case*

*# Cook's Distance*

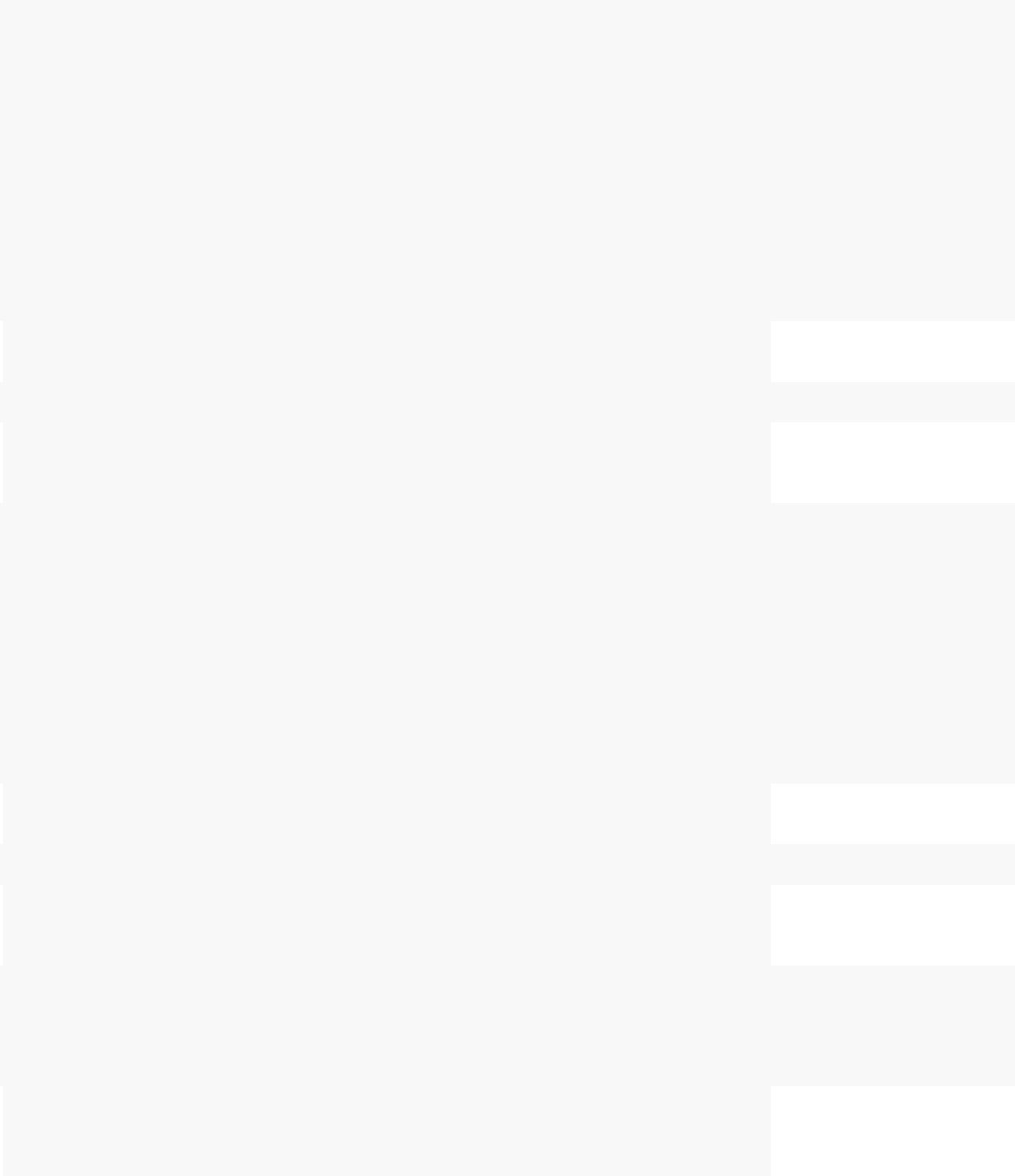
CD=**data.frame**(ID=**c**(1:**nrow**(Data)), CD=**cooks.distance**(model3)) n=**nrow**(Data)

CDLimit=**qf**(0.5,4,(n-3-1)) CDInfluential=CD[CD$CD>CDLimit,] **plot**(CD,pch=16,type="b") **points**(CDInfluential,col="red",pch=16) **abline**(h=CDLimit)

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 22 |



## Remove Outliers Data2 <- Data[-22,] Data3 <- Data2[-22,] **attach**(Data3)



Tuition3 <- **log**(Data3$Tuition)

Income3 <- **log**(Data3$Income)

new.model <- **lm**(Loan~Tuition3+Enroll+Income3)

**summary**(new.model)*#Very small differences between these models: keep outliers in*

##

* Call:
* lm(formula = Loan ~ Tuition3 + Enroll + Income3)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -7893.8 | | -3341.6 | 713.2 | 2936.7 | 6893.9 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -2.751e+05 | | | 5.379e+04 | -5.114 | 9.29e-06 | \*\*\* |
| ## Tuition3 | | 3.268e+04 | 7.026e+03 | 4.651 | 3.93e-05 | \*\*\* |
| ## | Enroll | 1.145e-02 | 7.664e-04 | 14.940 | < 2e-16 | \*\*\* |
| ## | Income3 | -1.108e+04 | 2.513e+03 | -4.410 | 8.22e-05 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4017 on 38 degrees of freedom
* Multiple R-squared: 0.9863, Adjusted R-squared: 0.9852
* F-statistic: 913 on 3 and 38 DF, p-value: < 2.2e-16

**summary**(model3)

##

* Call:
* lm(formula = Loan ~ Tuition2 + Enroll + Income2)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -10491.5 | | -3085.5 | 855.3 | 3348.8 | 6933.1 |
| ## |  |  |  |  |  |

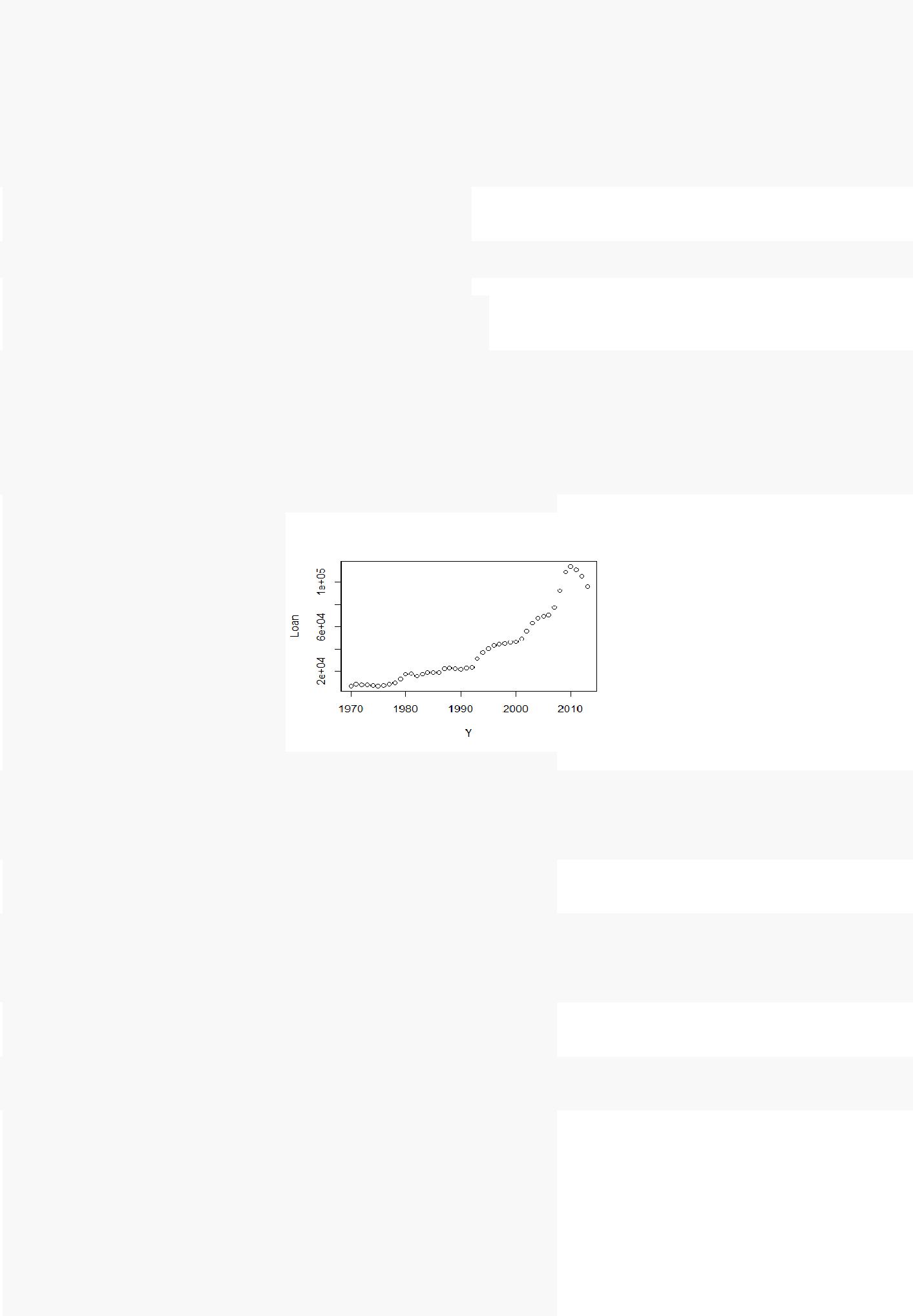
* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -2.767e+05 | | | 6.111e+04 | -4.528 | 5.25e-05 | \*\*\* |
| ## Tuition2 | | 3.425e+04 | 7.966e+03 | 4.300 | 0.000107 | \*\*\* |
| ## | Enroll | 1.160e-02 | 8.695e-04 | 13.341 | 2.59e-16 | \*\*\* |
| ## | Income2 | -1.259e+04 | 2.817e+03 | -4.468 | 6.34e-05 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4563 on 40 degrees of freedom
* Multiple R-squared: 0.9816, Adjusted R-squared: 0.9803
* F-statistic: 713.1 on 3 and 40 DF, p-value: < 2.2e-16

|  |  |
| --- | --- |
| STUDENT LOAN DEBT | 23 |

############# Prediction and Interpretation ########



**summary**(model3)

##

* Call:
* lm(formula = Loan ~ Enroll + logTuition + logIncome)
* Residuals:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Min | 1Q | Median | 3Q | Max |
| ## -10491.5 | | -3085.5 | 855.3 | 3348.8 | 6933.1 |
| ## |  |  |  |  |  |

* Coefficients:
* Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## (Intercept) -2.767e+05 | | | 6.111e+04 | -4.528 | 5.25e-05 | \*\*\* |
| ## Enroll | | 1.160e-02 | 8.695e-04 | 13.341 | 2.59e-16 | \*\*\* |
| ## | logTuition | 3.425e+04 | 7.966e+03 | 4.300 | 0.000107 | \*\*\* |
| ## | logIncome | -1.259e+04 | 2.817e+03 | -4.468 | 6.34e-05 | \*\*\* |

* ---
* Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
* Residual standard error: 4563 on 40 degrees of freedom
* Multiple R-squared: 0.9816, Adjusted R-squared: 0.9803
* F-statistic: 713.1 on 3 and 40 DF, p-value: < 2.2e-16

**plot**(Y, Loan)

***#2012***

**predict**(Model, **data.frame**(Enroll=15372284,logIncome=**log**(51179),logTuition=**log**(20234)),interval="confidence", level =0.95)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | fit | lwr | upr |
| ## | 1 104783 | 101333.6 | 108232.3 |
|  |  |  |  |

***#1999***

**predict**(Model, **data.frame**(Enroll=10818667,logIncome=**log**(40750),logTuition=**log**(14254)),interval="confidence", level =0.95)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | fit | lwr | upr |
| ## | 1 42832.02 | 40281.87 | 45382.17 |
|  |  |  |  |

**predict**(Model, **data.frame**(Enroll=10818667,logIncome=**log**(40750),logTuition=**log**(14254)),interval="confidence", level =0.99)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | fit | lwr | upr |
| ## | 1 42832.02 | 39419.59 | 46244.45 |