Predicting Income from Personal Financial Data

By Brent Schiller, Robert Kimelman, Matt Myers, Isaac Thomas-Markarian

I. Background & Description

In predicting income levels of credit card applicants, we analyzed credit card data sourced from the book “Econometric Analysis”. This dataset includes twelve variables for 1,319 applicants. At first, we sought to predict whether or not a credit card applicant would be accepted using all the independent variables, converting the categorical results into probabilities. However, this model required a logistic regression since the response variable is categorical. Instead, we used this dataset to predict yearly income/10,000 (*income*) from the number of major derogatory reports (*reports*), the age of the applicant (*age*), their monthly expenditure/yearly income (*share*), their average monthly expenditures (*expenditure*), the number of dependents they have including themselves (*dependents*), the number of months they have lived at their current address (*months*), the number of major credit cards they had (*majorcards*), and the number of active credit accounts they had (*active*).

II. Statistical Analysis

A) Initial Data Analysis

We began our analysis by visualizing the data with a histogram of *income*. We observed that the data was skewed right, so we applied a log transformation to the data and obtained a histogram with a more normal distribution. We then created scatter plots for each of the independent variables against *income*. All of the plots appeared to show positive correlations between the two respective variables except for the independent variables *dependents* and *reports* (See Appendix for all of the aforementioned plots).

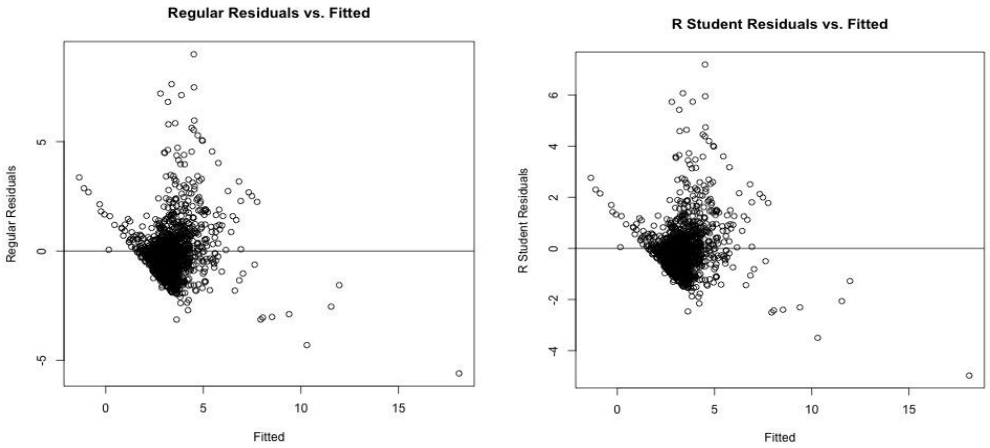
B) Initial Full Model

We found that all independent variables are significant predictors of *income* except for *reports* and *months*, with significance being determined by any variable with a p-value of $< \alpha = 0.05$. The F-statistic also had a p-value < 0.05, leading to the conclusion that the model, as a whole, is significant (See R Appendix for the summary).

C) Model Diagnostics

<u>Constant Variance Assumption</u>

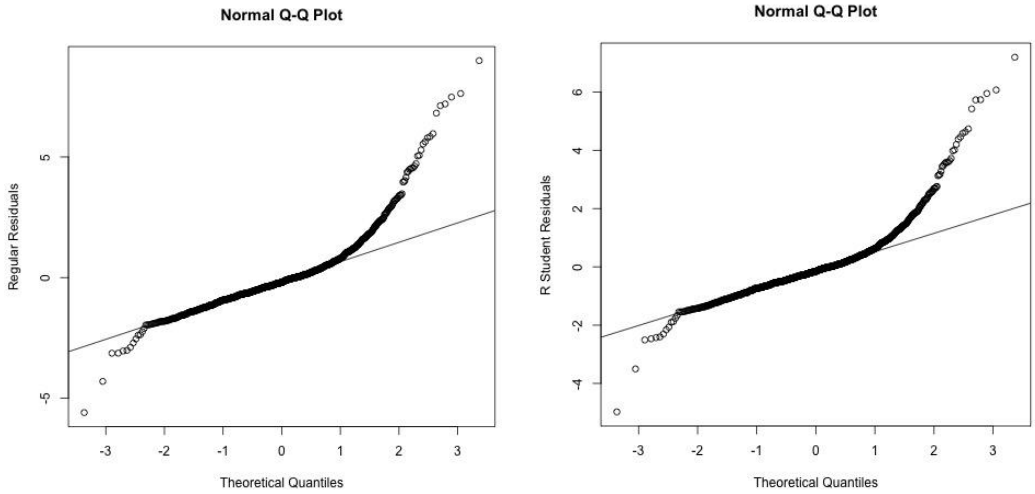
We tested the constant variance assumption of the Gauss-Markov Model using plots of regular residuals and R-student residuals against fitted values:



Both of the above plots indicate a biased (due to a linear relationship in the graph) and homoscedastic (due to relatively even distribution of points) relationship in the data. The constant variance assumption has been met.

Normality Assumption

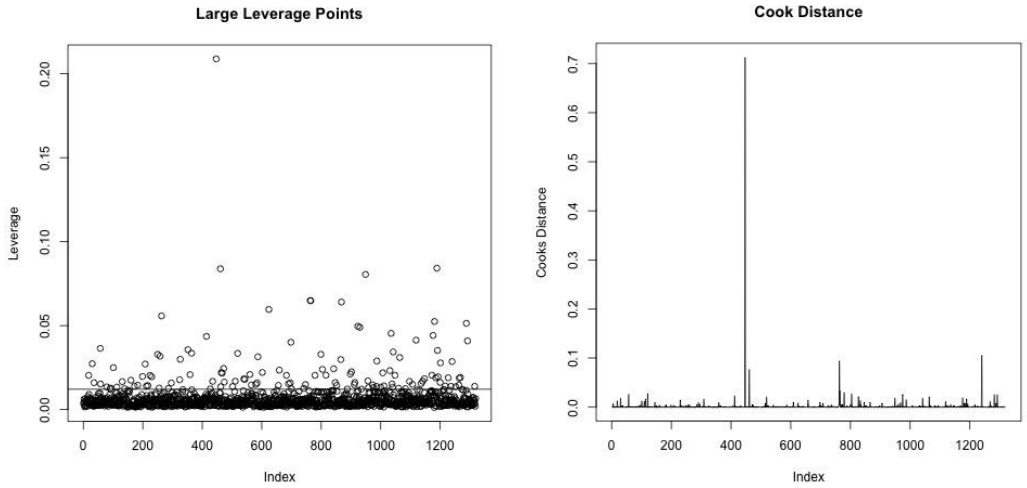
We first tested the normality assumption of the Gauss-Markov model by creating QQ-plots for regular residuals and R-student residuals:



Both plots show a backward-S shape pattern, indicating that the data follows a long-tailed distribution, which is what was also observed in the initial histogram of *income*. We then used more formal tests for normality–the Shapiro-Wilk Test and the Two-Sample Kolmogorov-Smirnov Test–and both yielded very small p-values < 2.2e-16, so we rejected the null hypothesis of normality.

Large Leverage Points, Outliers, and Influential Points

We checked for large leverage points with the criteria that they must be > 2*8/1319. We found 127 large leverage points, as shown above the horizontal line (h=2*8/1319) in the plot below on the left.



We found that the outliers in *income* were 10.0393, 13.5, 9.9999, 10.5, 12.4999, 11.9999, 9.4, 10.9999, 9, 10.032, 9.9999, 11, and 9.9999 (sum = 13) by comparing the jackknife residuals to Bonferroni's critical value at α = 0.05. We then ran a model on the data without these outliers and obtained an R^2 of .4942 and an R^2 adjusted of .4911, larger than both of the respective values in the initial full model. We checked for influential points by plotting the Cook's distance values above, and since there were no Cook's distance values > 1, we concluded that there are no influential points.

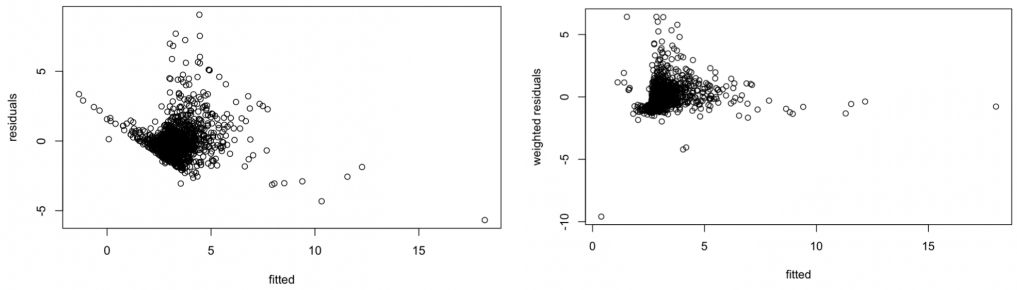
GLS

We then ran Generalized Least Square Estimation. Phi controls the correlation between consecutive errors. The general assumption is that the errors are uncorrelated, but here, we are testing this assumption by exploring **ρ** (rho) correlation between ε_i and ε_{i+1}. Because the confidence interval for **ρ** contains zero, this suggests that there is no strong correlation, and GLS estimation does not add new value. Shown below:

```
Correlation structure:
          lower       est.     upper
Phi -0.006174665 0.04852463 0.1029344
attr(,"label")
[1] "Correlation structure:"
```

Weighted Least Squares

The regression estimates do not change much from the applying the weighted least squares method. There is no new detectable dependence between the errors of the residuals, so there is no objective way to apply weighted least squares estimation. This is shown below:



Multicollinearity

When looking for multicollinearity, we began looking at the correlation table, which is shown below:

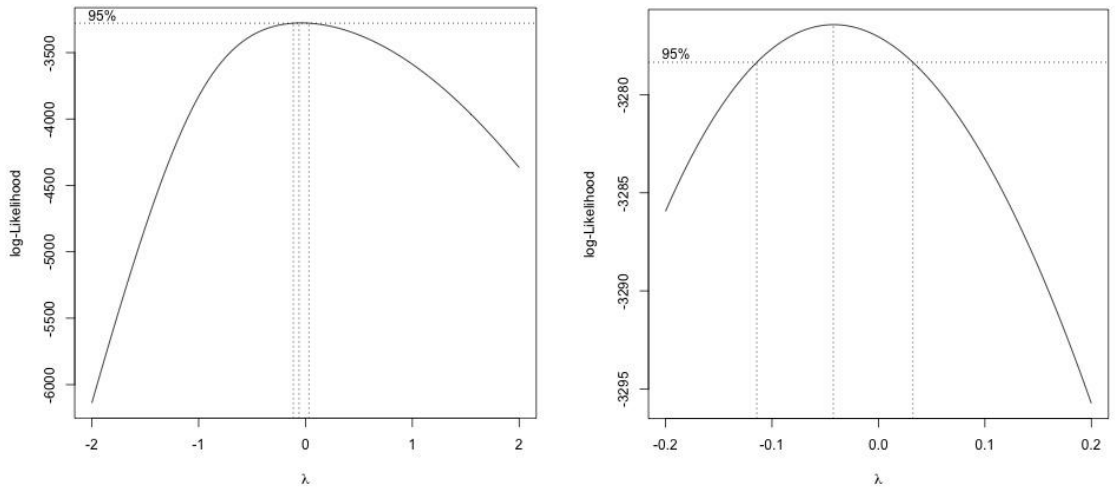
```
> round(cor(projdata),3)
             reports    age income  share expenditure dependents months majorcards active
reports        1.000  0.044  0.011 -0.159      -0.137      0.020  0.049     -0.007  0.208
age            0.044  1.000  0.325 -0.116       0.015      0.212  0.436      0.010  0.181
income         0.011  0.325  1.000 -0.054       0.281      0.318  0.130      0.107  0.181
share         -0.159 -0.116 -0.054  1.000       0.839     -0.083 -0.055      0.051 -0.023
expenditure   -0.137  0.015  0.281  0.839       1.000      0.053 -0.029      0.078  0.055
dependents     0.020  0.212  0.318 -0.083       0.053      1.000  0.047      0.010  0.107
months         0.049  0.436  0.130 -0.055      -0.029      0.047  1.000     -0.041  0.100
majorcards    -0.007  0.010  0.107  0.051       0.078      0.010 -0.041      1.000  0.120
active         0.208  0.181  0.181 -0.023       0.055      0.107  0.100      0.120  1.000
```

Among the predictors, the largest correlations occurred between *share* & *expenditure* (0.839), *months* & *age* (0.436), and *dependents* & *age* (0.212). The two larger correlations seemed more problematic, so we investigated these further with amputations. Before doing such changes, we considered condition indices and variation inflation factors (VIF) of the model (see Appendix). The largest condition index is 6490, which is significantly larger than the benchmark of 30, indicating strong multicollinearity. However, the largest VIF is 3.75, which is less than the benchmark of 10, which indicates no problems.

In the first amputation, we removed *share* (See Appendix for regression output and multicollinearity diagnostics). The largest condition index dropped to 759, which is still much larger than 30, indicating more multicollinearity in the model. The largest VIF decreased to 1.32, which is still smaller than 10. In the next amputation, we removed *months* (See Appendix for regression output and multicollinearity diagnostics). In this amputation, the condition index did not decrease as much, only to 749, which is still well above 30. Lastly, the VIF further decreased to 1.075. The remaining correlations among variables in the model are not large, despite a large condition index.

D) Correcting Model Inadequacies

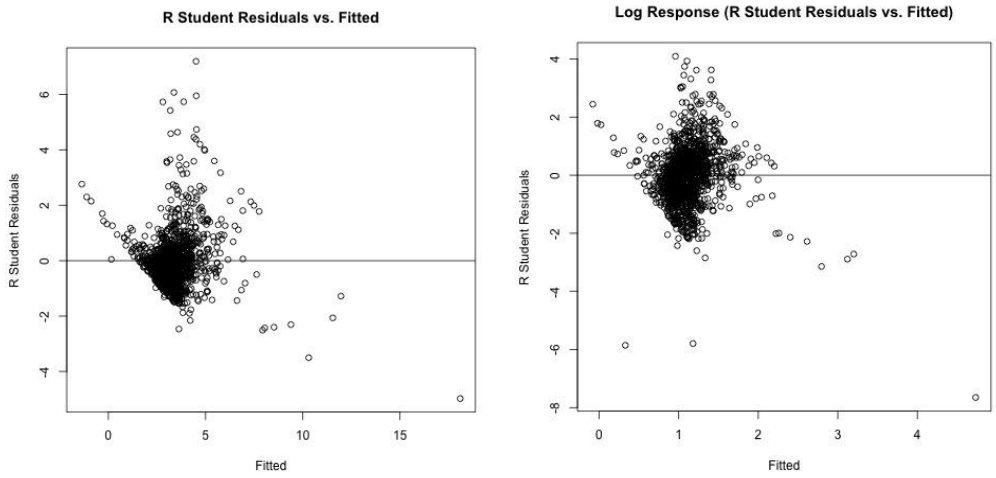
Box-Cox Transformation



Our first plot showing the confidence interval for λ was too broad, so we zoomed in such that the axis containing λ ranged from -0.2 and 0.2. Based on these plots, we determined that a log transformation would be suitable since a transformation that raises Y to a small decimal would be unreasonable.

Constant Variance Assumption

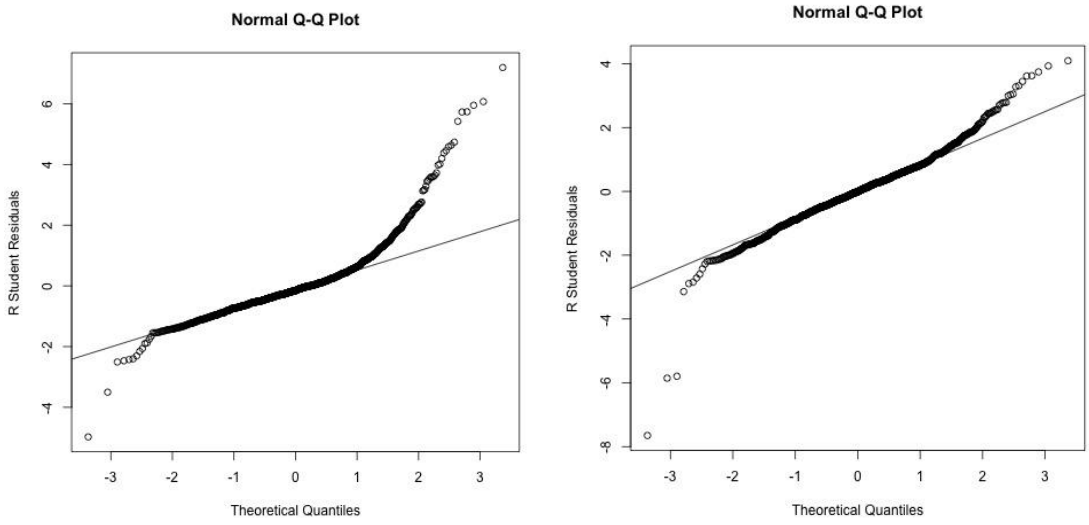
We tested the constant variance assumption of the Gauss-Markov Model using a plot of R-student residuals against fitted values after the transformation.



We compared this plot with the same plot obtained before transformation, and the data becomes slightly less biased and is still homoscedastic; thus, the data still meets the constant variance assumption.

Normality Assumption

We first tested the normality assumption of the Gauss-Markov Model by creating a QQ-plot for R-student residuals (shown below to the right).

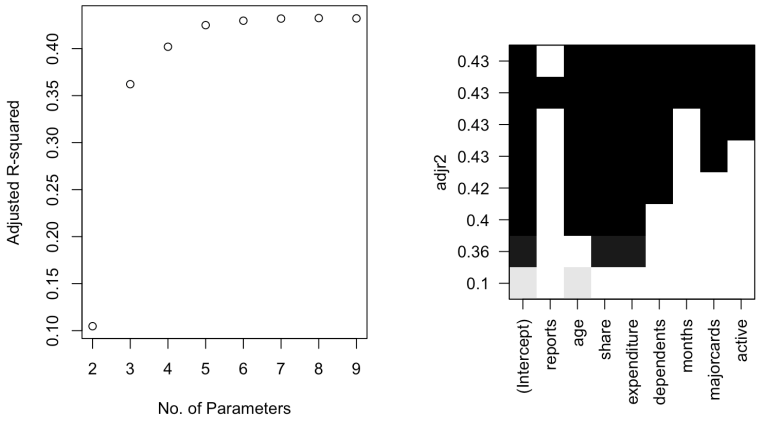


The plot obtained after doing the Box-Cox transformation is more linear than the plot obtained before the transformation (shown above to the left), thus the distribution of the transformed data is more normal. Both the Shapiro-Wilk Test and the Two-Sample Kolmogorov-Smirnov Test yielded the same p-values < 2.2e-16, so we rejected the null hypothesis of normality again.

E) Model Selection

Leaps

We began by considering all possible models using the leaps() function from the leaps R package. The outputs can be found in the Appendix. It is noteworthy that the largest increase in adjusted R-squared between models with variable counts is the difference between the model with one versus the model with two variables. The best one-variable model is the one with *age*, and the best two-variable model is the one with *share* and *expenditure*. Additionally, the model with the highest adjusted R-squared is the seven-variable model, which excludes *reports*. The following shows the best adjusted R-squared models graphically:

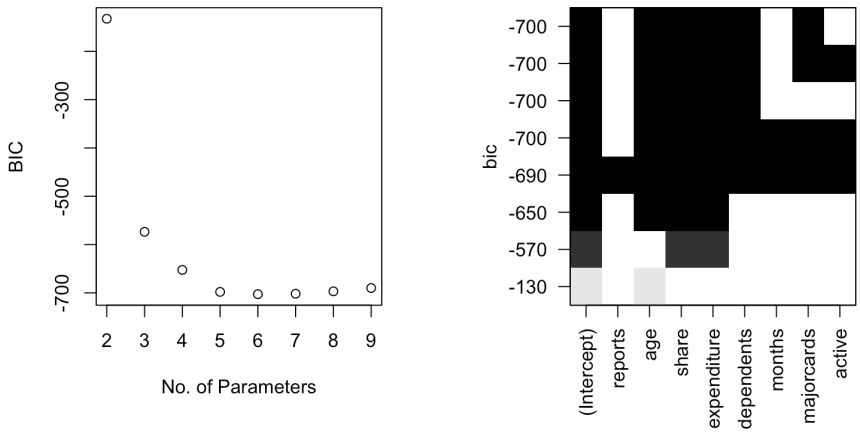


Regsubsets and Cp

In the model selection process, we also ran regsubsets() to examine which models minimized Cp (the output can be found in the Appendix). Of the variables, *expenditure* appeared in the most models to minimize Cp, yet the best standalone variable model was the one with *age*, which is the same as the best leaps model.

BIC

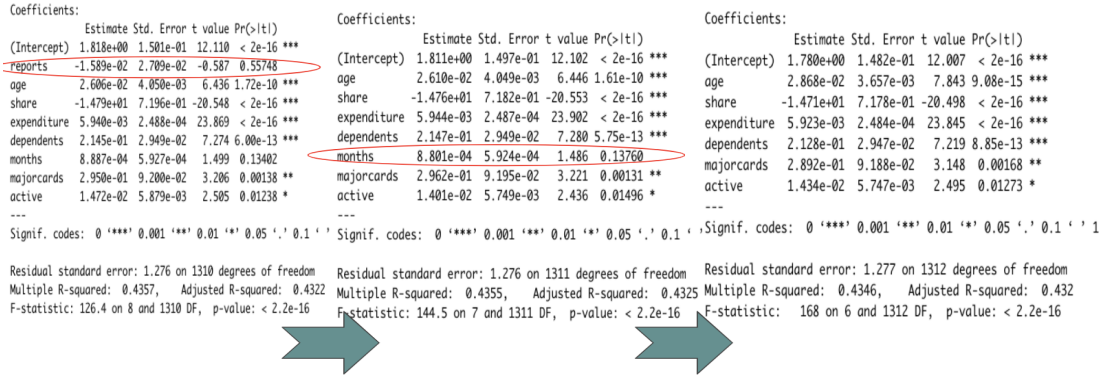
When minimizing BIC, there were some interesting results shown. The best model for minimization was the five-variable model with *age, share, expenditure, dependents,* and *majorcards*. This is represented graphically below:



F) Reduced Model Diagnostics

We employed backward and forward model elimination and selection to eliminate insignificant variables in the creation of our reduced model:

Backward and Forward Model Elimination

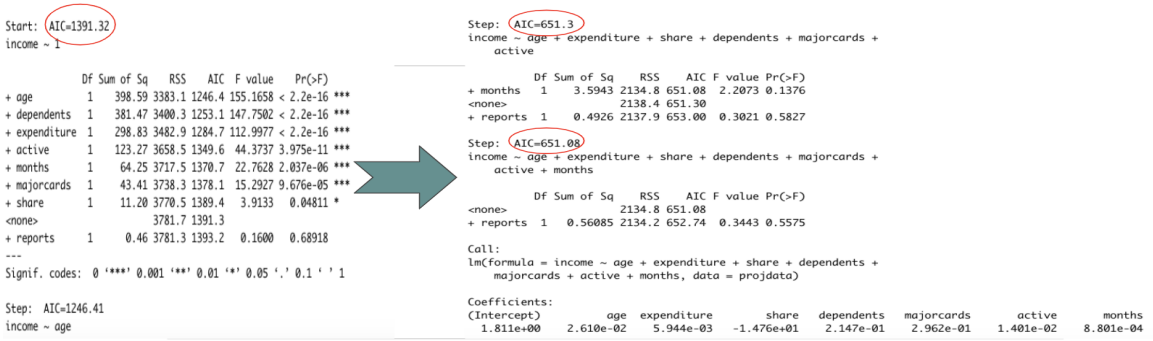


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.818e+00  1.501e-01  12.110  < 2e-16 ***
reports     -1.589e-02  2.709e-02  -0.587  0.55748
age          2.606e-02  4.050e-03   6.436 1.72e-10 ***
share       -1.479e+01  7.196e-01 -20.548  < 2e-16 ***
expenditure  5.940e-03  2.488e-04  23.869  < 2e-16 ***
dependents   2.145e-01  2.949e-02   7.274 6.00e-13 ***
months       8.887e-04  5.927e-04   1.499  0.13402
majorcards   2.950e-01  9.200e-02   3.206  0.00138 **
active       1.472e-02  5.879e-03   2.505  0.01238 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.276 on 1310 degrees of freedom
Multiple R-squared:  0.4357,    Adjusted R-squared:  0.4322
F-statistic: 126.4 on 8 and 1310 DF,  p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.811e+00  1.497e-01  12.102  < 2e-16 ***
age          2.610e-02  4.049e-03   6.446 1.61e-10 ***
share       -1.476e+01  7.182e-01 -20.553  < 2e-16 ***
expenditure  5.944e-03  2.487e-04  23.902  < 2e-16 ***
dependents   2.147e-01  2.949e-02   7.280 5.75e-13 ***
months       8.801e-04  5.924e-04   1.486  0.13760
majorcards   2.962e-01  9.195e-02   3.221  0.00131 **
active       1.401e-02  5.749e-03   2.436  0.01496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 1.276 on 1311 degrees of freedom
Multiple R-squared:  0.4355,    Adjusted R-squared:  0.4325
F-statistic: 144.5 on 7 and 1311 DF,  p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.780e+00  1.482e-01  12.007  < 2e-16 ***
age          2.868e-02  3.657e-03   7.843 9.08e-15 ***
share       -1.471e+01  7.178e-01 -20.498  < 2e-16 ***
expenditure  5.923e-03  2.484e-04  23.845  < 2e-16 ***
dependents   2.128e-01  2.947e-02   7.219 8.85e-13 ***
majorcards   2.892e-01  9.188e-02   3.148  0.00168 **
active       1.434e-02  5.747e-03   2.495  0.01273 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.277 on 1312 degrees of freedom
Multiple R-squared:  0.4346,    Adjusted R-squared:  0.432
F-statistic:   168 on 6 and 1312 DF,  p-value: < 2.2e-16
```

Backward and Forward Model Selection



```
Start:  AIC=1391.32
income ~ 1

             Df Sum of Sq    RSS    AIC  F value    Pr(>F)
+ age         1    398.59 3383.1 1246.4 155.1658 < 2.2e-16 ***
+ dependents  1    381.47 3400.3 1253.1 147.7502 < 2.2e-16 ***
+ expenditure 1    298.83 3482.9 1284.7 112.9977 < 2.2e-16 ***
+ active      1    123.27 3658.5 1349.6  44.3737 3.975e-11 ***
+ months      1     64.25 3717.5 1370.7  22.7628 2.037e-06 ***
+ majorcards  1     43.41 3738.3 1378.1  15.2927 9.676e-05 ***
+ share       1     11.20 3770.5 1389.4   3.9133   0.04811 *
<none>                   3781.7 1391.3
+ reports     1      0.46 3781.3 1393.2   0.1600   0.68918
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=1246.41
income ~ age
```

```
Step:  AIC=651.3
income ~ age + expenditure + share + dependents + majorcards +
    active

             Df Sum of Sq    RSS    AIC F value Pr(>F)
+ months      1    3.5943 2134.8 651.08  2.2073 0.1376
<none>                    2138.4 651.30
+ reports     1    0.4926 2137.9 653.00  0.3021 0.5827

Step:  AIC=651.08
income ~ age + expenditure + share + dependents + majorcards +
    active + months

             Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                    2134.8 651.08
+ reports     1   0.56085 2134.2 652.74  0.3443 0.5575

Call:
lm(formula = income ~ age + expenditure + share + dependents +
    majorcards + active + months, data = projdata)

Coefficients:
(Intercept)          age  expenditure        share   dependents   majorcards       active       months
  1.811e+00    2.610e-02    5.944e-03   -1.476e+01    2.147e-01    2.962e-01    1.401e-02    8.801e-04
```

When using elimination techniques, we found the variables *reports* and *months* to have larger p-values than our chosen α = 0.05. After eliminating the least significant variable, *reports*, we saw the p-value of *months* become larger. After removing *months*, the remaining 6 independent variables all had p-values < 0.05. Additionally, the R-squared value stayed largely the same for the reduced model as it was in the full model. Both models had R-squared values around 0.43. This indicates that the variables *reports* and *months* were insignificant predictors of income and that removing them was not impactful to the model. When using selection techniques based on the reduction of the AIC, we found *reports* to be the only variable to not reduce the AIC. However, removing *months* did not significantly reduce the AIC. Based on the results of the model elimination and selection, we decided to remove *months* and *reports* in the creation of the reduced model.

Constant Variance Assumption

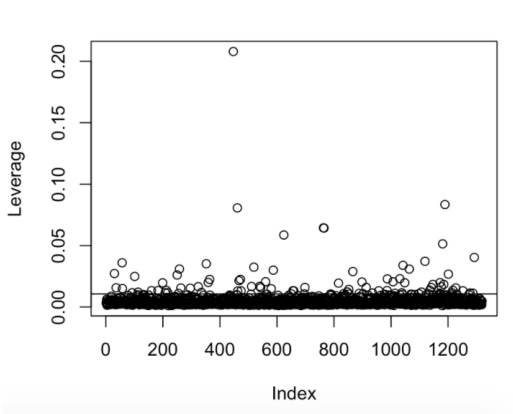
As seen in the Regular Residuals vs. Fitted and the R Student Residuals vs. Fitted scatterplots in the appendix, the constant variance assumption of the Gauss-Markov Model was met. Both plots indicate a biased and homoscedastic relationship.

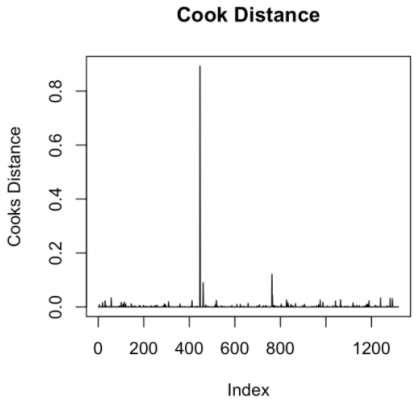
Normality Assumption

To test the normality assumption of the Gauss-Markov model, we created QQ-plots for both regular and R-student residuals. These plots are featured in the appendix. Both plots feature a backward-S shape that is similar to the QQ-plots created for the full model. This shape

indicates a long tailed distribution. Additionally, we conducted the Shapiro-Wilk Test and the Two-Sample Kolmogorov-Smirnov Tests for normality. Both tests yielded the same p-value as that of the full model, < 2.2e-16, which led us to reject the null hypothesis of normality.

Large Leverage Points, Outliers, and Influential Points





We checked for large leverage points with the criteria that they must be > 2*6/1319. We found 128 large leverage points, as shown above the horizontal line (h=2*6/1319) in the plot below. The outliers in income are 10.039, 13.500, 9.999, 10.500, 12.499, 11.999, 9.400, 10.999, 9.000, 10.032, 9.999, 11.000, 9.999. Thus, there are 13 total outliers. The total number of outliers did not change from the full model to the reduced model. While the largest Cook's value did increase with this smaller model, there was still no Cook's distance greater than 1.0. Hence, there are no influential points.

Removing Outliers

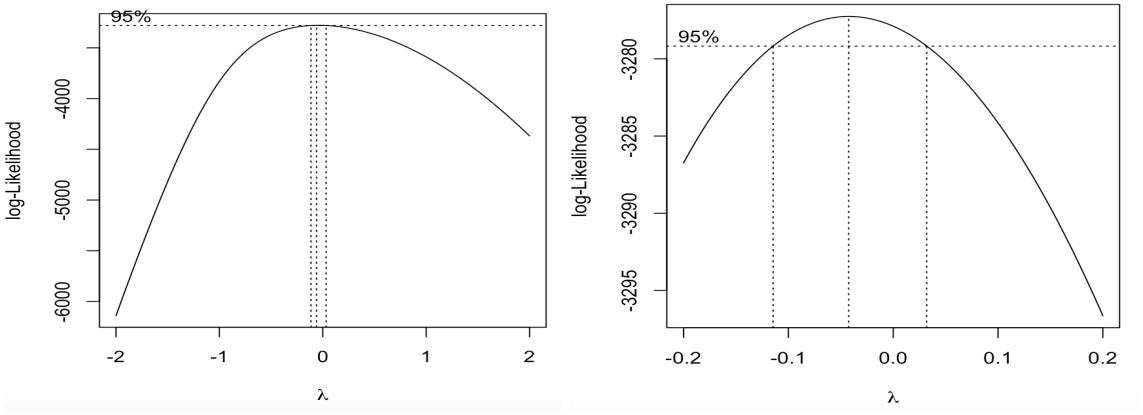
After removing the 13 outliers, the R-Squared for the reduced model rose to 0.4939. The model diagnostics from this model can be found in the Appendix.

Multicollinearity: Pairwise Correlation, Condition Indices, and VIF

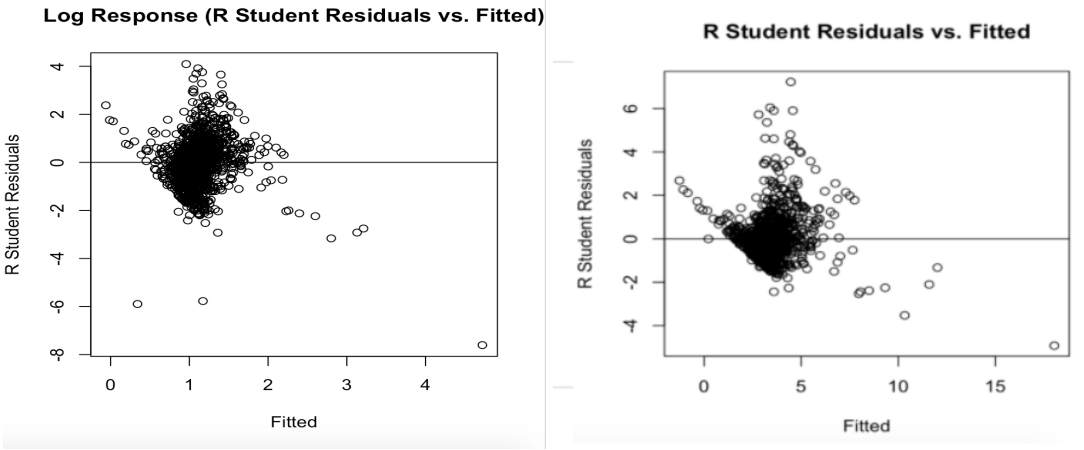
In the pairwise correlation table seen in the appendix, the largest correlation continued to be between *share* and *expenditure* (0.839). The second biggest correlation, between *months* and *age*, was removed from the reduced model. As previously mentioned, the amputation of *share* drastically reduced the R-squared value of the full model so we decided to keep these two highly correlated variables in the final model.

Additionally, we found the largest condition index to be 6457.858. This is problematic as the benchmark is 30. This indicates strong multicollinearity. The largest VIF was found to be 3.73 which is not problematic as it is less than the benchmark of 10.

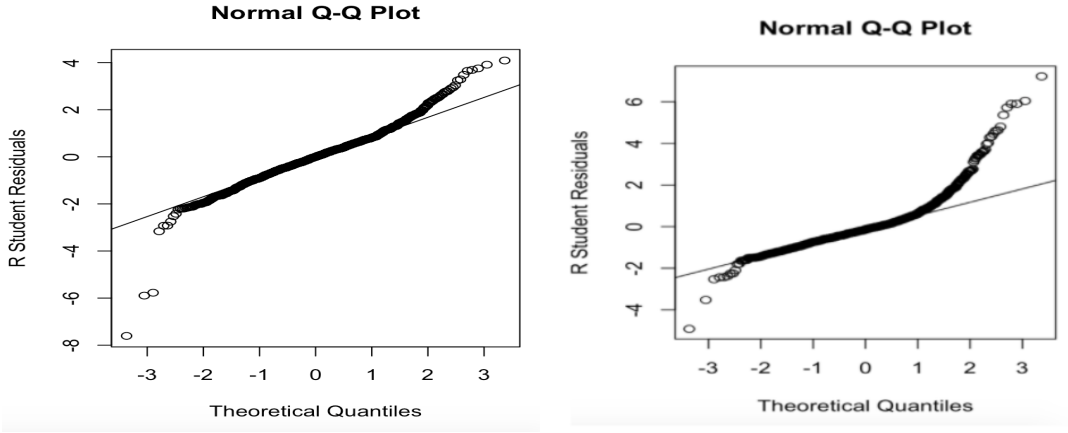
Box-Cox Transformation



To narrow the size of our confidence interval of λ, we narrowed the λ axis to -0.2, 0.2. We determined that a log transformation would be most suitable.



The left scatter plot is for the Box-Cox transformation while the right scatter plot shows the R student residuals for the basic reduced model. Both response plots of the R student residuals vs. the fitted values are biased and homoscedastic. However, the plot for the Box-Cox transformation appears to be more densely packed. Either way, the constant variance assumption is met.



To test the normality assumption, we looked at the two QQ-plots above. The plot on the left features the Box-Cox transformed reduced model QQ-Plot, while the plot on the right is the regular Normal QQ-Plot for the reduced model. The Box-Cox plot appears to be significantly

more linear. While both plots show a backward S-shape indicating a long-tail distribution. The Box-Cox transformation appears to be more normal.

III. Conclusions & Interpretations

The final model to predict income level from individual credit card application data included the following 6 independent variables:

- *Age*: n years
- *Expenditure*: Average monthly expenditure
- *Dependents*: 1+ number of dependents
- *Majorcards*: Number of major cards held
- *Active*: Number of active credit cards held
- *Share*: Monthly expenditure/yearly income

We removed the variables *reports* and *months* because they had p-values greater than our chosen alpha, 0.05, indicating that these two variables were not good predictors of *income*. We found *expenditure* and *share* to be highly correlated. However, an amputation of *share* showed that removing either variable drastically reduced the R-squared of the model. Additionally, we found there to be 13 outliers in both the full and reduced models. Removing these outlier points increased the R-squared of the model by roughly 0.07. For both the full and reduced models, the constant variance assumption was met and the null hypothesis of normality was rejected.

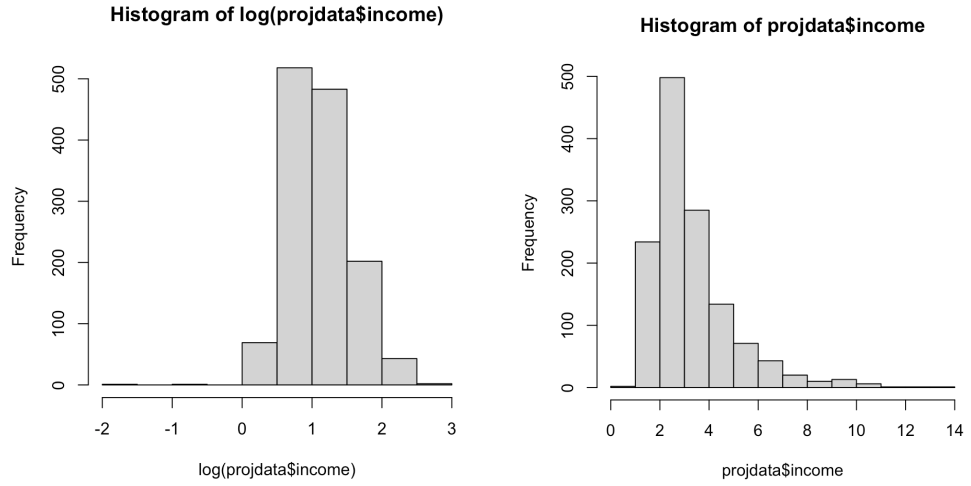
We performed Box-Cox transformations for the full model, the reduced model, and the reduced model without the outliers. While we first found the Box-Cox transformation to be helpful as our initial model was not normal shaped, the Box-Cox reduced model without outliers was actually worse than that without the transformation. This could be a result of removing many of the data points on the far right of the distribution, making it less skewed and more normal. In conclusion, the Box-Cox reduced model with no outliers has lower significance but normalizes the data, however, the reduced model with no outliers is more significant and has a larger R-squared (See a comparison of the models in the Appendix). The R-Squared value for our initial full model is 0.4357 while the reduced model has an R-Squared value of 0.4346. Removing two variables did not drastically affect the R-Squared value. Additionally, removing the outliers from the reduced model improved the R-Squared value to 0.4939 while the Box-Cox transformed reduced model without outliers had an R-Squared value of 0.4619.

Appendix

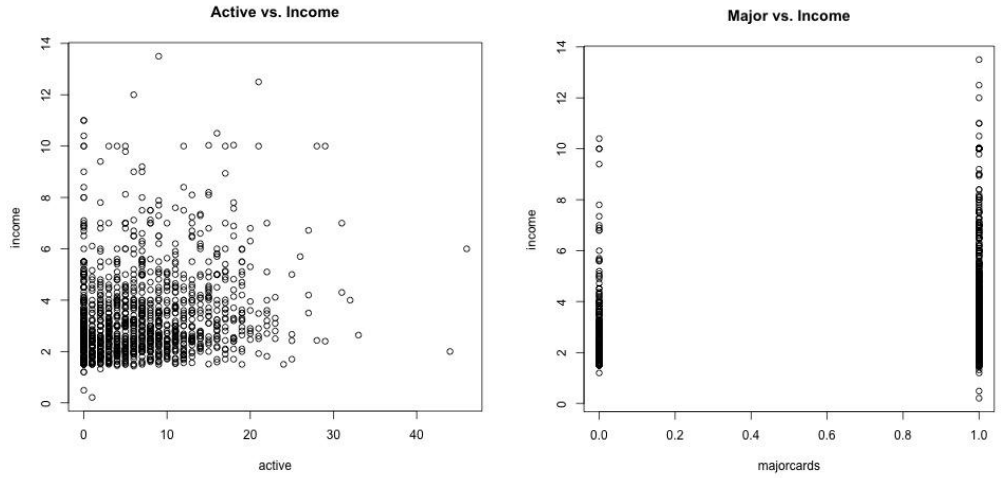
Summary Statistics of Data

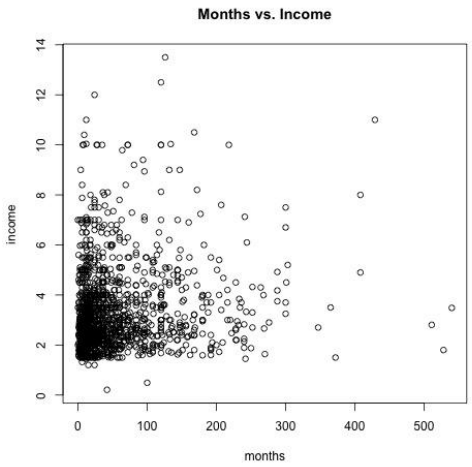
```
> summary(projdata)
    reports            age              income           share            expenditure
 Min.   : 0.0000   Min.   : 0.1667   Min.   : 0.210   Min.   :0.0001091   Min.   :   0.000
 1st Qu.: 0.0000   1st Qu.:25.4167   1st Qu.: 2.244   1st Qu.:0.0023159   1st Qu.:   4.583
 Median : 0.0000   Median :31.2500   Median : 2.900   Median :0.0388272   Median : 101.298
 Mean   : 0.4564   Mean   :33.2131   Mean   : 3.365   Mean   :0.0687322   Mean   : 185.057
 3rd Qu.: 0.0000   3rd Qu.:39.4167   3rd Qu.: 4.000   3rd Qu.:0.0936168   3rd Qu.: 249.036
 Max.   :14.0000   Max.   :83.5000   Max.   :13.500   Max.   :0.9063205   Max.   :3099.505
   dependents        months         majorcards        active
 Min.   :0.0000   Min.   :  0.00   Min.   :0.0000   Min.   : 0.000
 1st Qu.:0.0000   1st Qu.: 12.00   1st Qu.:1.0000   1st Qu.: 2.000
 Median :1.0000   Median : 30.00   Median :1.0000   Median : 6.000
 Mean   :0.9939   Mean   : 55.27   Mean   :0.8173   Mean   : 6.997
 3rd Qu.:2.0000   3rd Qu.: 72.00   3rd Qu.:1.0000   3rd Qu.:11.000
 Max.   :6.0000   Max.   :540.00   Max.   :1.0000   Max.   :46.000
```

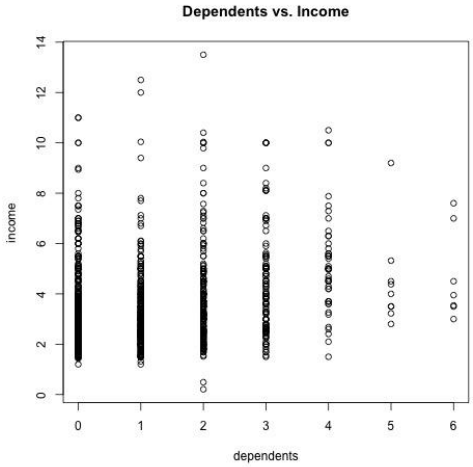
Income Histograms:

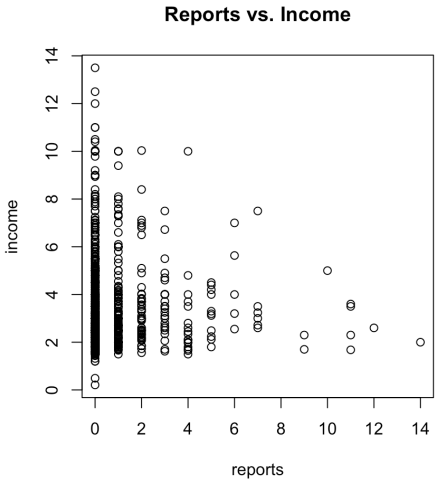


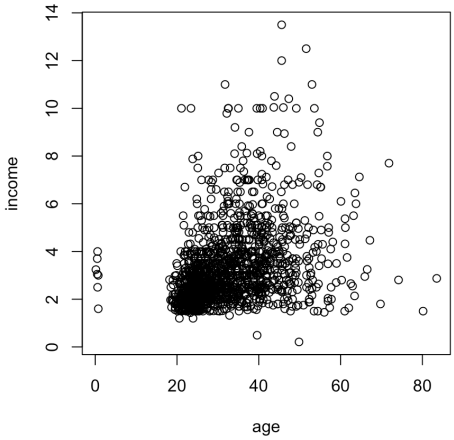
Scatterplots:

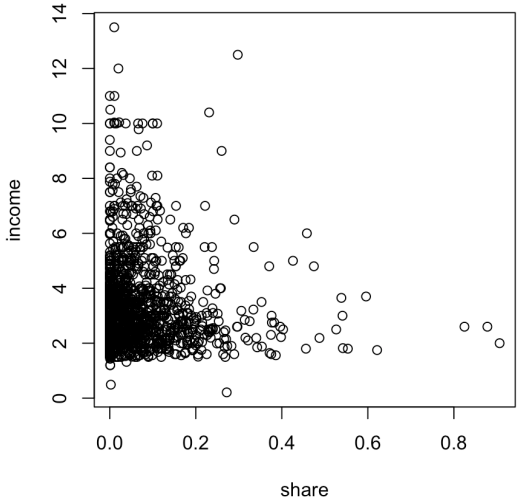


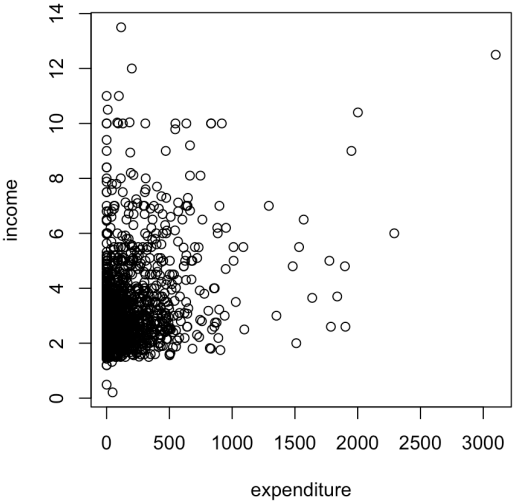

Months vs. Income
income
months
0
2
4
6
8
10
12
14
0
100
200
300
400
500


Dependents vs. Income
income
dependents
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6


Reports vs. Income
income
reports
0
2
4
6
8
10
12
14
0
2
4
6
8
10
12
14


Age vs. Income
income
age
0
2
4
6
8
10
12
14
0
20
40
60
80


Share vs. Income
income
share
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8


Expenditure vs. Income
income
expenditure
0
2
4
6
8
10
12
14
0
500
1000
1500
2000
2500
3000

Initial Full Model

```
> model1 = lm(income~reports+age+share+expenditure+dependents+months+majorcards+active, data=projdata)
> summary(model1)

Call:
lm(formula = income ~ reports + age + share + expenditure + dependents +
    months + majorcards + active, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5979 -0.6891 -0.1938  0.3964  8.9927

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.818e+00  1.501e-01  12.110  < 2e-16 ***
reports     -1.589e-02  2.709e-02  -0.587  0.55748
age          2.606e-02  4.050e-03   6.436 1.72e-10 ***
share       -1.479e+01  7.196e-01 -20.548  < 2e-16 ***
expenditure  5.940e-03  2.488e-04  23.869  < 2e-16 ***
dependents   2.145e-01  2.949e-02   7.274 6.00e-13 ***
months       8.887e-04  5.927e-04   1.499  0.13402
majorcards   2.950e-01  9.200e-02   3.206  0.00138 **
active       1.472e-02  5.879e-03   2.505  0.01238 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.276 on 1310 degrees of freedom
Multiple R-squared:  0.4357,    Adjusted R-squared:  0.4322
F-statistic: 126.4 on 8 and 1310 DF,  p-value: < 2.2e-16
```

Scatterplots:

```
data <- read.csv(file = 'AER_credit_card_data.csv')
mymodel <-
lm(income~reports+age+share+expenditure+dependents+months+majorcards+active, data)
plot(income~active, data, main = "Active vs. Income")
plot(income~majorcards, data, main = "Major vs. Income")
plot(income~months, data, main = "Months vs. Income")
plot(income~dependents, data, main = "Dependents vs. Income")
```

Testing Constant Variance Assumption:

```
plot(fitted(mymodel), residuals(mymodel), xlab = "Fitted", ylab = "Regular Residuals", main =
"Regular Residuals vs. Fitted")
abline(h=0)
plot(fitted(mymodel), rstudent(mymodel), xlab = "Fitted", ylab = "R Student Residuals", main =
"R Student Residuals vs. Fitted")
abline(h=0)
```

Testing Normality Assumption:

```
> shapiro.test(residuals(mymodel))

        Shapiro-Wilk normality test

data:  residuals(mymodel)
W = 0.85216, p-value < 2.2e-16

> ks.test(residuals(mymodel), fitted(mymodel))

        Two-sample Kolmogorov-Smirnov test

data:  residuals(mymodel) and fitted(mymodel)
D = 0.9022, p-value < 2.2e-16
alternative hypothesis: two-sided
```

qqnorm(residuals(mymodel), ylab = "Regular Residuals")
qqline(residuals(mymodel))
hist(residuals(mymodel))
qqnorm(rstudent(mymodel), ylab = "R Student Residuals")
qqline(rstudent(mymodel))
hist(rstudent(mymodel))

Finding Large Leverage Points:
plot(lm.influence(mymodel)$hat, ylab = "Leverage", main= "Large Leverage Points")
*abline(h=2*8/1319)*
*lm.influence(mymodel)$hat[lm.influence(mymodel)$hat>2*8/1319]*

```
> length(lm.influence(mymodel)$hat[lm.influence(mymodel)$hat>2*8/1319])
[1] 127
```

Finding Outliers:
*tcrit = qt(1-.05/(1319*2),1319-9-1)*
outliers = which(abs(rstudent(mymodel)) > tcrit)

```
> for (i in outliers){
+   print(data[i,]$income)
+   }
[1] 10.0393
[1] 13.5
[1] 9.9999
[1] 10.5
[1] 12.4999
[1] 11.9999
[1] 9.4
[1] 10.9999
[1] 9
[1] 10.032
[1] 9.9999
[1] 11
[1] 9.9999
```

Finding Influential Points:

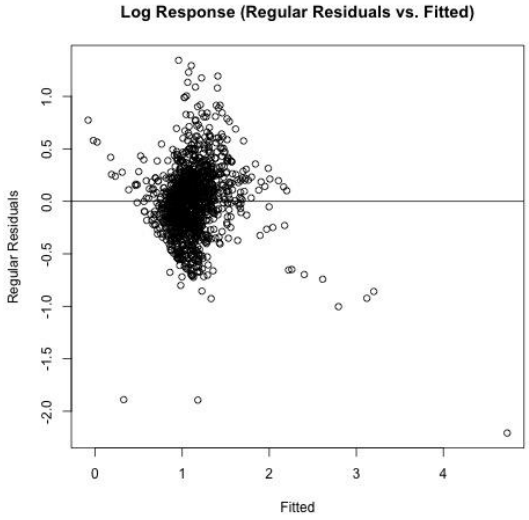
```
cook<-cooks.distance(mymodel)
row_names<-row.names(data)
plot(cook, ylab="Cooks Distance", main = "Cook Distance", type = "l")
ytick<-seq(0, 0.07, by = 0.01)
cook[which.max(cook)]
identify(1:1319, cook, row_names)
```

Full Model with no outliers:

```
> nooutliermodel = lm(income~., data=datanew)
> summary(nooutliermodel)

Call:
lm(formula = income ~ ., data = datanew)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0651 -0.6167 -0.1504  0.4136  5.1826

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.852e+00  1.289e-01  14.364  < 2e-16 ***
reports     -1.504e-02  2.327e-02  -0.647 0.518062
age          2.322e-02  3.482e-03   6.669  3.8e-11 ***
share       -1.568e+01  6.599e-01 -23.763  < 2e-16 ***
expenditure  6.508e-03  2.381e-04  27.330  < 2e-16 ***
dependents   1.857e-01  2.552e-02   7.279  5.8e-13 ***
months       3.105e-04  5.152e-04   0.603 0.546859
majorcards   2.941e-01  7.918e-02   3.715 0.000212 ***
active       1.771e-02  5.054e-03   3.505 0.000472 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.092 on 1297 degrees of freedom
Multiple R-squared:  0.4942,    Adjusted R-squared:  0.4911
F-statistic: 158.4 on 8 and 1297 DF,  p-value: < 2.2e-16
```

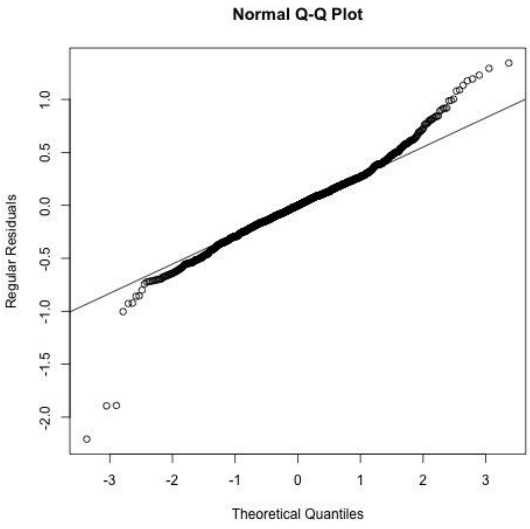
Box-Cox Transformation:

```
library(MASS)
boxcox(mymodel, plotit=T)
boxcox(mymodel, plotit=T, lambda=seq(-0.2,0.2,by=0.05))
mymodel_boxcox<-lm(log(income)~reports+age+share+expenditure+dependents+months+majorcards+active, data)
plot(fitted(mymodel_boxcox), residuals(mymodel_boxcox), xlab = "Fitted", ylab = "Regular Residuals", main = "Log Response (Regular Residuals vs. Fitted)")
abline(h=0)
plot(fitted(mymodel_boxcox), rstudent(mymodel_boxcox), xlab = "Fitted", ylab = "R Student Residuals", main = "Log Response (R Student Residuals vs. Fitted)")
abline(h=0)
```



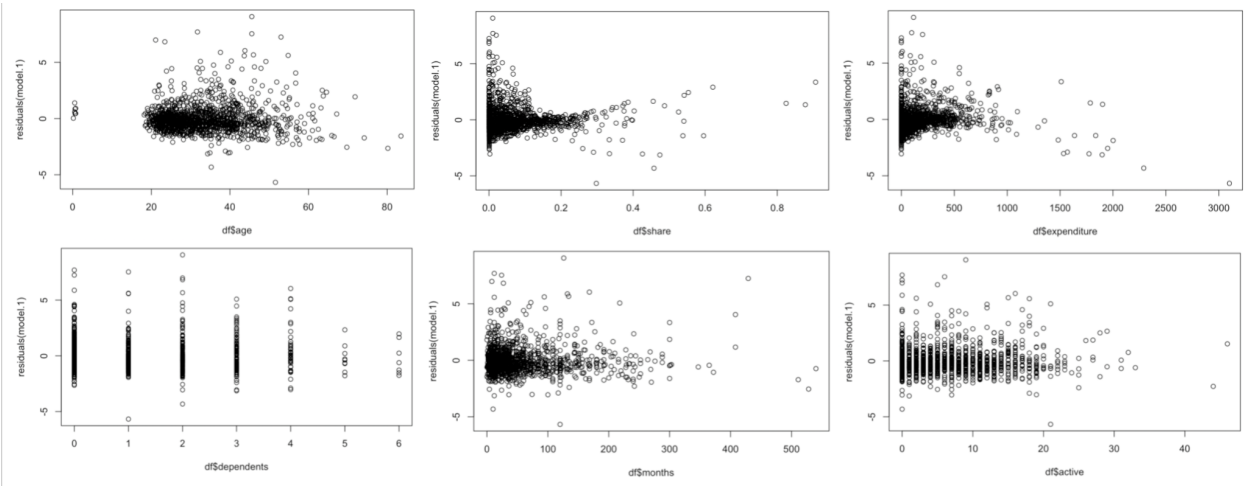

```
> shapiro.test(residuals(mymodel_boxcox))

        Shapiro-Wilk normality test

data:  residuals(mymodel_boxcox)
W = 0.96623, p-value < 2.2e-16

> ks.test(residuals(mymodel_boxcox), fitted(mymodel_boxcox))

        Two-sample Kolmogorov-Smirnov test

data:  residuals(mymodel_boxcox) and fitted(mymodel_boxcox)
D = 0.95148, p-value < 2.2e-16
alternative hypothesis: two-sided
```

qqnorm(residuals(mymodel_boxcox), ylab = "Regular Residuals")
qqline(residuals(mymodel_boxcox))



```
hist(residuals(mymodel_boxcox))
qqnorm(rstudent(mymodel_boxcox), ylab = "R Student Residuals")
qqline(rstudent(mymodel_boxcox))
hist(rstudent(mymodel_boxcox))

#gls
install.packages('nlme')
library(nlme)
g <- gls(income~ age+ share + expenditure + dependents + months + active,
correlation=corARMA(p=1), data=df)
summary(g, cor=T)
intervals(g)
```

Residuals vs Independent Variables



These scatterplots were used to understand the results for the weighted least squares analysis.

Original Model Condition Indices & VIF

```
> x = model.matrix(model1)[,-1]
> e = eigen(t(x)%*%x)
> e$values
[1] 1.445775e+08 9.125765e+06 6.134187e+05 4.966344e+04 2.230432e+03 1.937983e+03 2.506898e+02
[8] 3.431657e+00
> sqrt(e$values[1]/e$values)
[1]    1.000000    3.980299   15.352242   53.955024  254.598470  273.133728  759.419829 6490.801976
> vif(x)
    reports         age       share expenditure  dependents      months  majorcards      active
   1.074391    1.365176    3.753072    3.712326    1.095661    1.248103    1.023275    1.111755
```

Share Amputation Regression Output & Multicollinearity Output

```
> model2 = lm(income~reports+age+expenditure+dependents+months+majorcards+active, data=projdata)
> summary(model2)

Call:
lm(formula = income ~ reports + age + expenditure + dependents +
    months + majorcards + active, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8655 -0.8641 -0.2536  0.5251  9.2863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9267566  0.1651755   5.611 2.46e-08 ***
reports     0.0164162  0.0310860   0.528 0.597526
age         0.0414035  0.0045757   9.049  < 2e-16 ***
expenditure 0.0015970  0.0001510  10.576  < 2e-16 ***
dependents  0.3271754  0.0333122   9.821  < 2e-16 ***
months      0.0003185  0.0006805   0.468 0.639881
majorcards  0.3199355  0.1057415   3.026 0.002529 **
active      0.0223225  0.0067442   3.310 0.000959 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.467 on 1311 degrees of freedom
Multiple R-squared:  0.2538,    Adjusted R-squared:  0.2498
F-statistic: 63.69 on 7 and 1311 DF,  p-value: < 2.2e-16
```

```
> x2 = model.matrix(model2)[,-1]
> e2 = eigen(t(x2)%*%x2)
> e2$values
[1] 1.445775e+08 9.125765e+06 6.134187e+05 4.966344e+04 2.230426e+03 1.937872e+03 2.506174e+02
> sqrt(e2$values[1]/e2$values)
[1]   1.000000   3.980299  15.352241  53.955026 254.598804 273.141568 759.529516
> vif(x2)
    reports         age expenditure  dependents      months  majorcards      active
   1.070771    1.318796    1.034564    1.057812    1.245367    1.023097    1.107356
```

Months Amputation Regression Output & Multicollinearity Output

```
> model3 = lm(income~reports+age+expenditure+dependents+majorcards+active, data=projdata)
> summary(model3)

Call:
lm(formula = income ~ reports + age + expenditure + dependents +
    majorcards + active, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8833 -0.8619 -0.2531  0.5329  9.2986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9163197  0.1636142   5.600  2.6e-08 ***
reports     0.0167367  0.0310692   0.539 0.590192
age         0.0423224  0.0041317  10.243  < 2e-16 ***
expenditure 0.0015948  0.0001509  10.569  < 2e-16 ***
dependents  0.3263528  0.0332559   9.813  < 2e-16 ***
majorcards  0.3174083  0.1055720   3.007 0.002692 **
active      0.0224165  0.0067391   3.326 0.000904 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.467 on 1312 degrees of freedom
Multiple R-squared:  0.2536,    Adjusted R-squared:  0.2502
F-statistic: 74.31 on 6 and 1312 DF,  p-value: < 2.2e-16
```

```
> x3 = model.matrix(model3)[,-1]
> e3 = eigen(t(x3)%*%x3)
> e3$values
[1] 1.433332e+08 1.165841e+06 4.973591e+04 2.230724e+03 1.943706e+03 2.552052e+02
> sqrt(e3$values[1]/e3$values)
[1]   1.00000  11.08802  53.68320 253.48399 271.55526 749.42581
> vif(x3)
    reports         age expenditure  dependents  majorcards      active
   1.070251    1.075907    1.033573    1.054867    1.020428    1.106372
```

Leaps Output

```
> leaps(xdata, ydata, method=c("adjr2"))
$which
      1     2     3     4     5     6     7     8
1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
1 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
1  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2 FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
2 FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
2 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
2 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
2 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
2 FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
2 FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
2 FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
2 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
3 FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
3 FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
3 FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
3 FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE
3 FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
3 FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
3 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
3 FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE
3 FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE
4 FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
4 FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
4 FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE
4 FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
4 FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
4 FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
4  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
4 FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
4  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
4 FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
5 FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
5 FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
5 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
5  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
5 FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
5 FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
5 FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
5 FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
5  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
5 FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
6 FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
6 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
6 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
6  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
6  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
6 FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
6  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
6  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
6 FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
7 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
7  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
7  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
7  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
7  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
7  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"           "5"
[7] "6"           "7"           "8"

$size
 [1] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6 6 6 6 6
[44] 6 6 6 6 6 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 9

$adjr2
 [1]  0.1047204281  0.1001878783  0.0783201672  0.0318602324  0.0162437494  0.0107279185  0.0022054919
 [8] -0.0006377055  0.3621604836  0.1804879109  0.1697006904  0.1689193701  0.1212529479  0.1193903050
[15]  0.1148660466  0.1129107854  0.1103109379  0.1050201661  0.4019993222  0.3956064371  0.3734434162
[22]  0.3713385275  0.3669184121  0.3617106745  0.2379989022  0.1918472526  0.1871845189  0.1867425610
[29]  0.4249763133  0.4067086915  0.4061421922  0.4055453905  0.4022533111  0.4020791272  0.4015443182
[36]  0.4003821899  0.3951660198  0.3807778509  0.4297078988  0.4281140986  0.4254067724  0.4245384284
[43]  0.4109632364  0.4107431542  0.4098555189  0.4069605879  0.4062568822  0.4061335254  0.4319675592
[50]  0.4303545308  0.4292733363  0.4284356541  0.4278462788  0.4249702496  0.4149488428  0.4105378032
[57]  0.4105142778  0.4099997137  0.4324897990  0.4316652359  0.4299219616  0.4281869570  0.4147016461
[64]  0.4097242806  0.2497803543  0.1858949048  0.4322057958
```

Regsubsets Output

```
> rs
Subset selection object
Call: regsubsets.formula(income ~ reports + age + share + expenditure +
    dependents + months + majorcards + active, data = projdata,
    nbest = 2)
8 Variables  (and intercept)
            Forced in Forced out
reports         FALSE      FALSE
age             FALSE      FALSE
share           FALSE      FALSE
expenditure     FALSE      FALSE
dependents      FALSE      FALSE
months          FALSE      FALSE
majorcards      FALSE      FALSE
active          FALSE      FALSE
2 subsets of each size up to 8
Selection Algorithm: exhaustive
         reports age share expenditure dependents months majorcards active
1  ( 1 ) " "     "*" " "   " "         " "        " "    " "        " "
1  ( 2 ) " "     " " " "   " "         "*"        " "    " "        " "
2  ( 1 ) " "     " " "*"   "*"         " "        " "    " "        " "
2  ( 2 ) " "     "*" " "   "*"         " "        " "    " "        " "
3  ( 1 ) " "     "*" "*"   "*"         " "        " "    " "        " "
3  ( 2 ) " "     " " "*"   "*"         "*"        " "    " "        " "
4  ( 1 ) " "     "*" "*"   "*"         "*"        " "    " "        " "
4  ( 2 ) " "     "*" "*"   "*"         " "        " "    "*"        " "
5  ( 1 ) " "     "*" "*"   "*"         "*"        " "    "*"        " "
5  ( 2 ) " "     "*" "*"   "*"         "*"        " "    " "        "*"
6  ( 1 ) " "     "*" "*"   "*"         "*"        " "    "*"        "*"
6  ( 2 ) " "     "*" "*"   "*"         "*"        "*"    "*"        " "
7  ( 1 ) " "     "*" "*"   "*"         "*"        "*"    "*"        "*"
7  ( 2 ) "*"     "*" "*"   "*"         "*"        " "    "*"        "*"
8  ( 1 ) "*"     "*" "*"   "*"         "*"        "*"    "*"        "*"
```

Creating Our Finalized Reduced Model

Backward model elimination:
incomedata <- lm(income~reports+age+share+expenditure+dependents+months+majorcards+active,data=projdata)
summary(incomedata)
incomedata <- update(incomedata, .~. -reports)
summary(incomedata)
incomedata <- update(incomedata, .~. -months)
summary(incomedata)

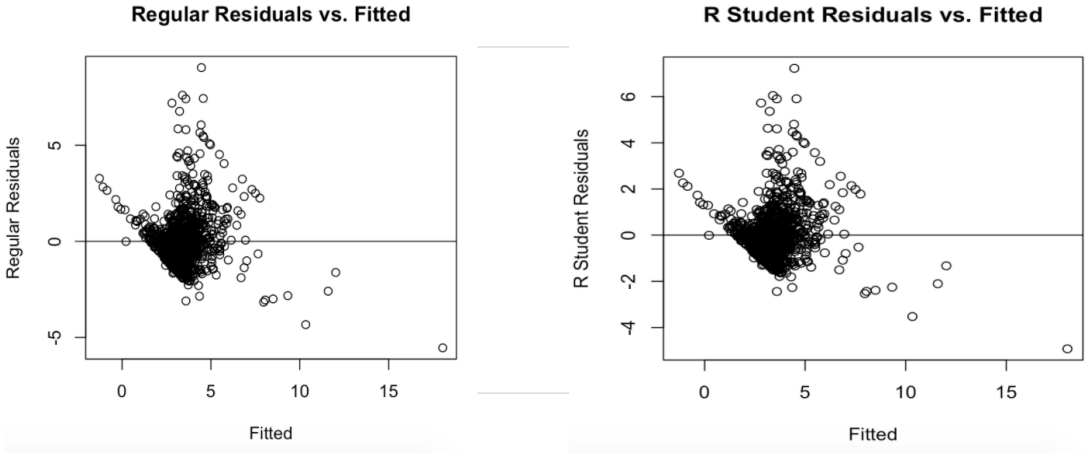
Backward and Forward Model Selection:
full=lm(income~reports+age+share+expenditure+dependents+months+majorcards+active,data=projdata)
step(full,data=projdata, direction="backward",tests="F")
null=lm(income~1, data=projdata)
step(null, scope=list(lower=null, upper=full), direction="forward", tests="F")

reducedmodel <- lm(income~age+share+expenditure+dependents+majorcards+active, data=projdata)
summary(reducedmodel)

```
Call:
lm(formula = income ~ age + share + expenditure + dependents +
    majorcards + active, data = projdata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5427 -0.7015 -0.1869  0.4047  9.0372

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.780e+00  1.482e-01  12.007  < 2e-16 ***
age          2.868e-02  3.657e-03   7.843 9.08e-15 ***
share       -1.471e+01  7.178e-01 -20.498  < 2e-16 ***
expenditure  5.923e-03  2.484e-04  23.845  < 2e-16 ***
dependents   2.128e-01  2.947e-02   7.219 8.85e-13 ***
majorcards   2.892e-01  9.188e-02   3.148  0.00168 **
active       1.434e-02  5.747e-03   2.495  0.01273 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.277 on 1312 degrees of freedom
Multiple R-squared:  0.4346,    Adjusted R-squared:  0.432
F-statistic:   168 on 6 and 1312 DF,  p-value: < 2.2e-16
```

Constant Variance Assumption For the Reduced Model

```
plot(fitted(reducedmodel), residuals(reducedmodel), xlab = "Fitted", ylab = "Regular Residuals",
main = "Regular Residuals vs. Fitted")
abline(h=0)
plot(fitted(reducedmodel), rstudent(reducedmodel), xlab = "Fitted", ylab = "R Student
Residuals", main = "R Student Residuals vs. Fitted")
abline(h=0)
```



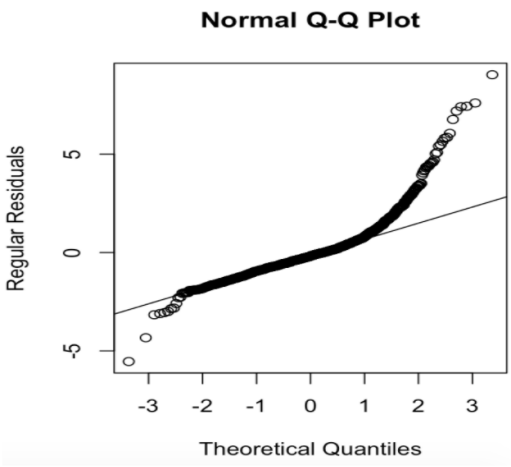
Normality Assumption: Normality Tests

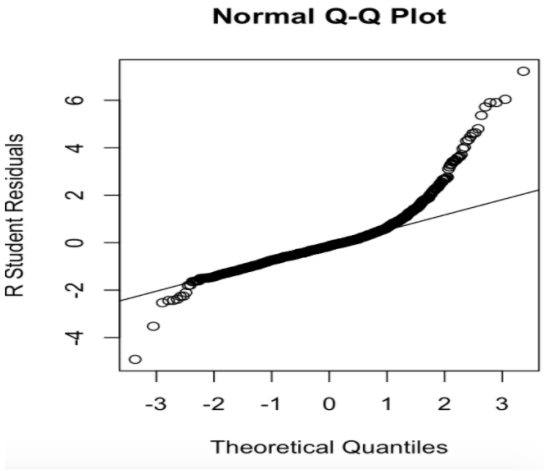
```
shapiro.test(residuals(reducedmodel))
ks.test(residuals(reducedmodel), fitted(reducedmodel))
```

```
        Shapiro-Wilk normality test

data:  residuals(reducedmodel)
W = 0.85179, p-value < 2.2e-16


        Two-sample Kolmogorov-Smirnov test

data:  residuals(reducedmodel) and fitted(reducedmodel)
D = 0.90296, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Normality Assumption: QQ-Plots

```
qqnorm(residuals(reducedmodel), ylab = "Regular Residuals")
qqline(residuals(reducedmodel))
hist(residuals(reducedmodel))
qqnorm(rstudent(reducedmodel), ylab = "R Student Residuals")
qqline(rstudent(reducedmodel))
hist(rstudent(reducedmodel))
```





Large Leverage Points, Outliers, and Influential Points

```
plot(lm.influence(reducedmodel)$hat, ylab = "Leverage")
abline(h=2*6/1319)
lm.influence(reducedmodel)$hat[lm.influence(reducedmodel)$hat>2*7/1319]

jack<-rstudent(reducedmodel)
outliers <- which(jack>abs(qt(0.05/(1319*2), 1319-7-1)))
jack[outliers]




cook<-cooks.distance(reducedmodel)
row_names<-row.names(data)
plot(cook, ylab="Cooks Distance", main = "Cook Distance", type = "l")
```

```
ytick<-seq(0, 0.07, by = 0.01)
cook[which.max(cook)]
identify(1:1319, cook, row_names)
```

Multicollinearity: Pairwise Correlation

```
reducedata=subset(data, select=-c(card,owner,selfemp,reports,months))
```

```
> round(cor(reducedata),3)
              age income  share expenditure dependents majorcards active
age         1.000  0.325 -0.116       0.015      0.212      0.010  0.181
income      0.325  1.000 -0.054       0.281      0.318      0.107  0.181
share      -0.116 -0.054  1.000       0.839     -0.083      0.051 -0.023
expenditure 0.015  0.281  0.839       1.000      0.053      0.078  0.055
dependents  0.212  0.318 -0.083       0.053      1.000      0.010  0.107
majorcards  0.010  0.107  0.051       0.078      0.010      1.000  0.120
active      0.181  0.181 -0.023       0.055      0.107      0.120  1.000
```

Highest correlations (excluding correlations with Income):

- Share & Expenditure (0.839)
- Dependents & Age (0.212)

Multicollinearity Condition Indices and VIF

```
x= model.matrix(reducedmodel)[,-1]
e = eigen(t(x)%*%x)
```

```
> e$values
[1] 1.433332e+08 1.165532e+06 4.963087e+04 1.943948e+03 2.552884e+02 3.436924e+00
> sqrt(e$values[1]/e$values)
[1]    1.00000   11.08949   53.73998  271.53833  749.30370 6457.85801
> vif(x)
       age       share expenditure  dependents  majorcards      active
  1.112773    3.732690    3.697197    1.093562    1.020121    1.062019
```

- Largest condition number is 6457 >> 30 → problematic
- Largest VIF number is 3.73 < 10 → no problem

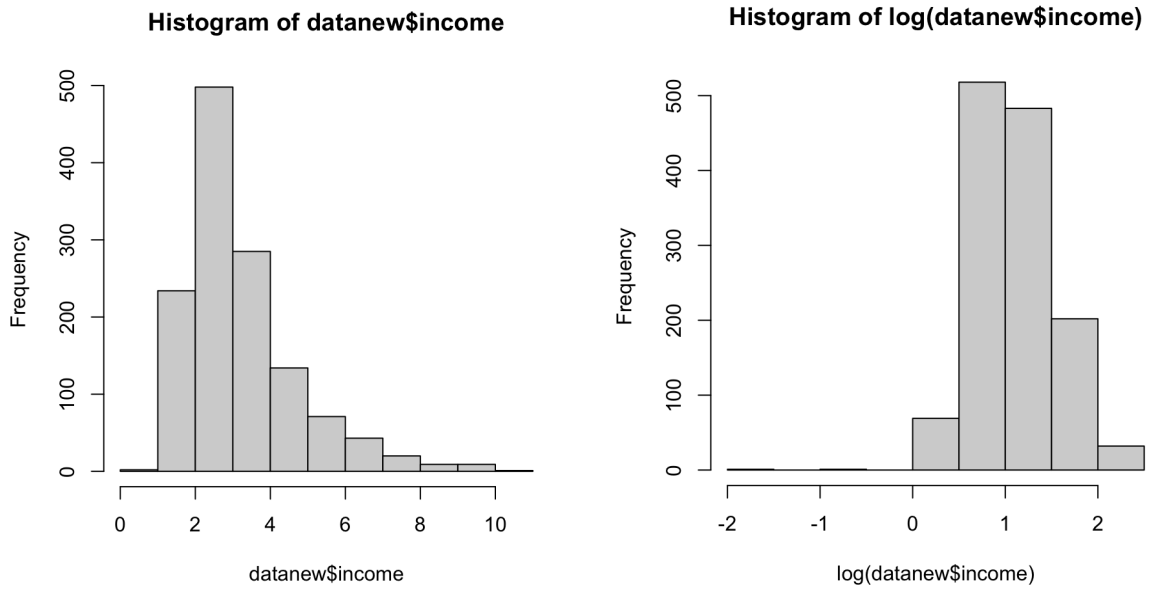
Removing Outliers

```
tcrit = qt(1-.05/(1319*2),1319-7-1)
outliers = which(abs(rstudent(model5)) > tcrit)
datanew = projdata[-outliers,]
finalmodel = lm(income~age+share+expenditure+dependents+majorcards+active,
data=datanew)
summary(finalmodel)
```

Box-Cox Transformations

```
library(MASS)
boxcox(reducedmodel,plotit=T)
boxcox(reducedmodel,plotit=T,lambda=seq(-.2,.2,by=0.05))
reducedmodel_boxcox<-lm(log(income)~age+share+expenditure+dependents+majorcards+acti
ve,data=projdata)
plot(fitted(reducedmodel_boxcox),residuals(reducedmodel_boxcox),xlab="Fitted",ylab="Regular
Residuals",main="Log Response(Regular Residuals vs. Fitted)")
abline(h=0)
plot(fitted(reducedmodel_boxcox),rstudent(reducedmodel_boxcox),xlab="Fitted",ylab="R
Student Residuals", main="Log Response (R Student Residuals vs. Fitted)")
abline(h=0)
shapiro.test(residuals(reducedmodel_boxcox))
ks.test(residuals(reducedmodel_boxcox),fitted(reducedmodel_boxcox))
qqnorm(residuals(reducedmodel_boxcox), ylab = "Regular Residuals")
qqline(residuals(reducedmodel_boxcox))
hist(residuals(reducedmodel_boxcox))
qqnorm(rstudent(reducedmodel_boxcox), ylab = "R Student Residuals")
qqline(rstudent(reducedmodel_boxcox))
hist(rstudent(reducedmodel_boxcox))
```

Histograms of Income after Outlier Removal



Final Model without Box Cox Transformation:

```
> finalmodel = lm(income~age+share+expenditure+dependents+majorcards+active, data=datanew)
> summary(finalmodel)

Call:
lm(formula = income ~ age + share + expenditure + dependents +
    majorcards + active, data = datanew)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0841 -0.6260 -0.1487  0.4251  5.2001

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.836e+00  1.273e-01  14.417  < 2e-16 ***
age          2.412e-02  3.150e-03   7.656 3.72e-14 ***
share       -1.564e+01  6.577e-01 -23.777  < 2e-16 ***
expenditure  6.504e-03  2.375e-04  27.387  < 2e-16 ***
dependents   1.852e-01  2.548e-02   7.269 6.24e-13 ***
majorcards   2.929e-01  7.901e-02   3.707 0.000218 ***
active       1.716e-02  4.936e-03   3.476 0.000526 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.091 on 1299 degrees of freedom
Multiple R-squared:  0.4939,    Adjusted R-squared:  0.4915
F-statistic: 211.3 on 6 and 1299 DF,  p-value: < 2.2e-16
```

Final Model with Box Cox Transformation:

```
> boxcoxmodel = lm(log(income)~age+share+expenditure+dependents+majorcards+active, data=datanew)
> summary(boxcoxmodel)

Call:
lm(formula = log(income) ~ age + share + expenditure + dependents +
    majorcards + active, data = datanew)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8476 -0.1770  0.0021  0.1814  1.0165

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.992e-01  3.590e-02  19.479  < 2e-16 ***
age          6.551e-03  8.879e-04   7.377 2.87e-13 ***
share       -4.101e+00  1.854e-01 -22.120  < 2e-16 ***
expenditure  1.685e-03  6.694e-05  25.172  < 2e-16 ***
dependents   5.033e-02  7.182e-03   7.008 3.88e-12 ***
majorcards   8.507e-02  2.227e-02   3.820  0.00014 ***
active       5.522e-03  1.391e-03   3.969 7.63e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3077 on 1299 degrees of freedom
Multiple R-squared:  0.4619,    Adjusted R-squared:  0.4594
F-statistic: 185.9 on 6 and 1299 DF,  p-value: < 2.2e-16
```