BUDA 525 Final— Team Chambers

Collin Edwards, Olga Ilyaguyeva, Adam Weisenstine

10/07/2024

### Problem 3

In the Credit data in the ISLR package it contains 400 customers and information on their credit history. For full information of the data look at the help file. A company has approached us to better understand factors that influence the Balance variable which is average credit card balance in USD. Using the information in the model, discuss the influential factors and discuss the factors you choose to put in the model. Do you have any concerns about the use of certain variables in the model? Discuss how your model was created and any insights you can provide based on the results. HINT: Adding Gender and/or Ethnicity could be controversial or illegal in some uses of this this model you should discuss your decision on these variables and how it effects the organizations ability to use your model for prediction or inference.

library(ISLR)  
head(Credit)

## ID Income Limit Rating Cards Age Education Gender Student Married Ethnicity  
## 1 1 14.891 3606 283 2 34 11 Male No Yes Caucasian  
## 2 2 106.025 6645 483 3 82 15 Female Yes Yes Asian  
## 3 3 104.593 7075 514 4 71 11 Male No No Asian  
## 4 4 148.924 9504 681 3 36 11 Female No No Asian  
## 5 5 55.882 4897 357 2 68 16 Male No Yes Caucasian  
## 6 6 80.180 8047 569 4 77 10 Male No No Caucasian  
## Balance  
## 1 333  
## 2 903  
## 3 580  
## 4 964  
## 5 331  
## 6 1151

data("Credit")  
#Data Exploration  
help(Credit)  
dim(Credit)

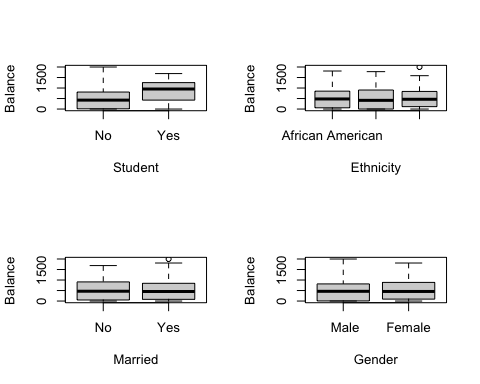
## [1] 400 12

summary(Credit)

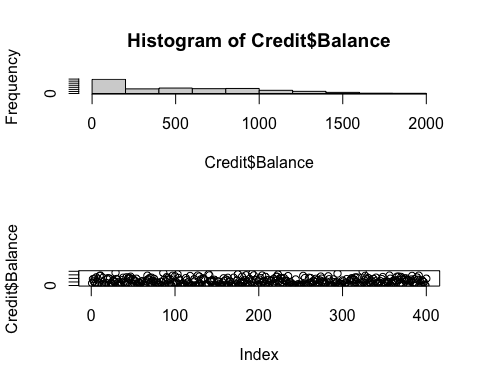
## ID Income Limit Rating   
## Min. : 1.0 Min. : 10.35 Min. : 855 Min. : 93.0   
## 1st Qu.:100.8 1st Qu.: 21.01 1st Qu.: 3088 1st Qu.:247.2   
## Median :200.5 Median : 33.12 Median : 4622 Median :344.0   
## Mean :200.5 Mean : 45.22 Mean : 4736 Mean :354.9   
## 3rd Qu.:300.2 3rd Qu.: 57.47 3rd Qu.: 5873 3rd Qu.:437.2   
## Max. :400.0 Max. :186.63 Max. :13913 Max. :982.0   
## Cards Age Education Gender Student   
## Min. :1.000 Min. :23.00 Min. : 5.00 Male :193 No :360   
## 1st Qu.:2.000 1st Qu.:41.75 1st Qu.:11.00 Female:207 Yes: 40   
## Median :3.000 Median :56.00 Median :14.00   
## Mean :2.958 Mean :55.67 Mean :13.45   
## 3rd Qu.:4.000 3rd Qu.:70.00 3rd Qu.:16.00   
## Max. :9.000 Max. :98.00 Max. :20.00   
## Married Ethnicity Balance   
## No :155 African American: 99 Min. : 0.00   
## Yes:245 Asian :102 1st Qu.: 68.75   
## Caucasian :199 Median : 459.50   
## Mean : 520.01   
## 3rd Qu.: 863.00   
## Max. :1999.00

First we wanted to look at the data. There are 400 observations in this data collection, and there are 12 variables that are all connected to credit card balance details. We observed a wide variation between the min and max values for several variables, including Income, Limit, and Balance, based on the data summary. Next, our suggestion that in order to obtain a well-fitting linear model, we might need to alter these variables. Most of the observations are for non-students, and the data is around half male and half female.

par(mfrow=c(2,2))  
boxplot(Balance~Student, data=Credit)  
boxplot(Balance~Ethnicity, data=Credit)  
boxplot(Balance~Married, data=Credit)  
boxplot(Balance~Gender, data=Credit)



par(mfrow=c(2,1))  
hist(Credit$Balance)  
plot(Credit$Balance)



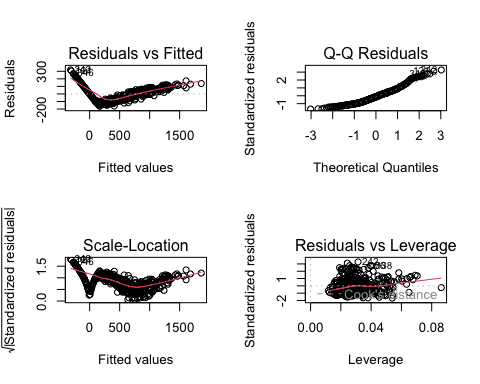
When we looked at credit card balances, we noticed that students have higher median balances, while non-students show a wider range. This makes sense because students usually have limited income and might take longer to pay off debts. We believe the Student variable will be important in our model.

Balances are pretty similar across different Ethnicity groups, which is good since including this might be controversial. We believe in doing data analysis ethically and don’t want to build a model that discriminates. So, we’re hoping not to use this variable.

We also didn’t see big differences based on Married status or Gender. We don’t expect these to be important factors in our model, and we’d prefer not to include Gender for similar reasons as Ethnicity.

Another thing we found is that the Balance variable has a lot of zeros. This might make it hard to build a good model, and we need to figure out how to transform Balance. Because of the zeros, we can’t use log transformations or Box-Cox analysis. We’ll come back to this later, but let’s start building the model now. ### Problem 3 **Model Selection Process**

m0<-lm(Balance~Income+Rating+Limit+Cards+Age+Education+Student+Married+Ethnicity+Gender, data = Credit)  
par(mfrow=c(2,2))  
plot(m0)



Credit$newbalance<-Credit$Balance+.01  
summary(Credit[12:13])

## Balance newbalance   
## Min. : 0.00 Min. : 0.01   
## 1st Qu.: 68.75 1st Qu.: 68.76   
## Median : 459.50 Median : 459.51   
## Mean : 520.01 Mean : 520.02   
## 3rd Qu.: 863.00 3rd Qu.: 863.01   
## Max. :1999.00 Max. :1999.01

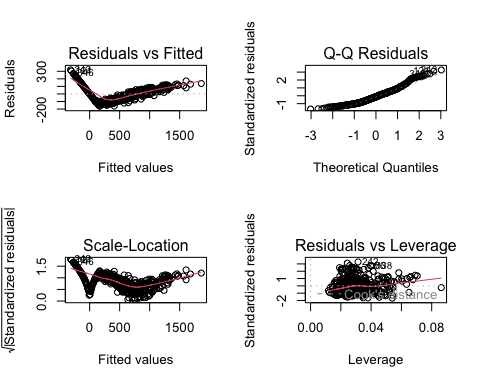
We started by building a model with all the variables. We used the lm() function to build the model, and we used the summary() function to look at the results. We found that the model is significant, but the R-squared value is low. This means that the model doesn’t explain much of the variation in Balance. Right away, we observed notable curvature in our residuals plot. To address this issue, we wanted to apply the boxCox() transformation to Balance to see if it would improve our model.

However, the presence of zero balances posed some limitations. To overcome this, we decided to increase all Balance observations by 0.01. This small adjustment allows us to leverage the necessary tools to build the best possible model, and we don’t anticipate that adding a penny will significantly alter the insights from the data. We created a new variable called newbalance, which is simply the current Balance plus $0.01. Now, we can use additional R tools to help us determine the optimal way to transform newbalance.

m1<-lm(newbalance~Income+Rating+Limit+Cards+Age+Education+Student+Married+Ethnicity+Gender, data = Credit)  
summary(m1)

##   
## Call:  
## lm(formula = newbalance ~ Income + Rating + Limit + Cards + Age +   
## Education + Student + Married + Ethnicity + Gender, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -161.64 -77.70 -13.49 53.98 318.20   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -479.19787 35.77394 -13.395 < 2e-16 \*\*\*  
## Income -7.80310 0.23423 -33.314 < 2e-16 \*\*\*  
## Rating 1.13653 0.49089 2.315 0.0211 \*   
## Limit 0.19091 0.03278 5.824 1.21e-08 \*\*\*  
## Cards 17.72448 4.34103 4.083 5.40e-05 \*\*\*  
## Age -0.61391 0.29399 -2.088 0.0374 \*   
## Education -1.09886 1.59795 -0.688 0.4921   
## StudentYes 425.74736 16.72258 25.459 < 2e-16 \*\*\*  
## MarriedYes -8.53390 10.36287 -0.824 0.4107   
## EthnicityAsian 16.80418 14.11906 1.190 0.2347   
## EthnicityCaucasian 10.10703 12.20992 0.828 0.4083   
## GenderFemale -10.65325 9.91400 -1.075 0.2832   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 98.79 on 388 degrees of freedom  
## Multiple R-squared: 0.9551, Adjusted R-squared: 0.9538   
## F-statistic: 750.3 on 11 and 388 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(m1)



m1test<-step(m1)

## Start: AIC=3686.22  
## newbalance ~ Income + Rating + Limit + Cards + Age + Education +   
## Student + Married + Ethnicity + Gender  
##   
## Df Sum of Sq RSS AIC  
## - Ethnicity 2 14084 3800814 3683.7  
## - Education 1 4615 3791345 3684.7  
## - Married 1 6619 3793349 3684.9  
## - Gender 1 11269 3798000 3685.4  
## <none> 3786730 3686.2  
## - Age 1 42558 3829288 3688.7  
## - Rating 1 52314 3839044 3689.7  
## - Cards 1 162702 3949432 3701.0  
## - Limit 1 331050 4117780 3717.7  
## - Student 1 6326012 10112742 4077.1  
## - Income 1 10831162 14617892 4224.5  
##   
## Step: AIC=3683.7  
## newbalance ~ Income + Rating + Limit + Cards + Age + Education +   
## Student + Married + Gender  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 4545 3805359 3682.2  
## - Education 1 4757 3805572 3682.2  
## - Gender 1 10760 3811574 3682.8  
## <none> 3800814 3683.7  
## - Age 1 45650 3846464 3686.5  
## - Rating 1 49473 3850287 3686.9  
## - Cards 1 166806 3967621 3698.9  
## - Limit 1 340250 4141064 3716.0  
## - Student 1 6372573 10173387 4075.5  
## - Income 1 10838891 14639705 4221.1  
##   
## Step: AIC=3682.18  
## newbalance ~ Income + Rating + Limit + Cards + Age + Education +   
## Student + Gender  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 5399 3810759 3680.7  
## - Gender 1 11019 3816378 3681.3  
## <none> 3805359 3682.2  
## - Age 1 43545 3848904 3684.7  
## - Rating 1 46929 3852289 3685.1  
## - Cards 1 170729 3976088 3697.7  
## - Limit 1 352112 4157472 3715.6  
## - Student 1 6461978 10267338 4077.2  
## - Income 1 10860901 14666261 4219.8  
##   
## Step: AIC=3680.75  
## newbalance ~ Income + Rating + Limit + Cards + Age + Student +   
## Gender  
##   
## Df Sum of Sq RSS AIC  
## - Gender 1 10861 3821620 3679.9  
## <none> 3810759 3680.7  
## - Age 1 43983 3854741 3683.3  
## - Rating 1 49520 3860279 3683.9  
## - Cards 1 170898 3981656 3696.3  
## - Limit 1 347654 4158413 3713.7  
## - Student 1 6472072 10282831 4075.8  
## - Income 1 10855780 14666539 4217.8  
##   
## Step: AIC=3679.89  
## newbalance ~ Income + Rating + Limit + Cards + Age + Student  
##   
## Df Sum of Sq RSS AIC  
## <none> 3821620 3679.9  
## - Age 1 44472 3866091 3682.5  
## - Rating 1 49263 3870883 3683.0  
## - Cards 1 172930 3994549 3695.6  
## - Limit 1 347898 4169518 3712.7  
## - Student 1 6462599 10284218 4073.9  
## - Income 1 10845018 14666638 4215.8

m1test

##   
## Call:  
## lm(formula = newbalance ~ Income + Rating + Limit + Cards + Age +   
## Student, data = Credit)  
##   
## Coefficients:  
## (Intercept) Income Rating Limit Cards Age   
## -493.7242 -7.7951 1.0912 0.1937 18.2119 -0.6241   
## StudentYes   
## 425.6099

We begin by fitting a model that includes all available variables as predictors of newbalance. A preliminary examination of the model summary provides insights into how these variables influence the model. At the p < .05 significance level, we observe several variables with very low p-values. However, Education, Married, Ethnicity, and Gender have high p-values, indicating that they do not significantly affect the model.

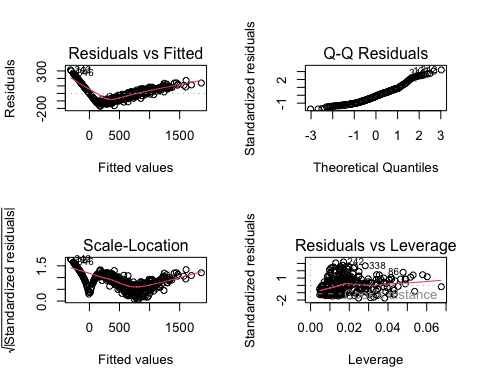
To refine the model, we will do a backward stepwise selection to determine whether these four variables should be removed. The results from this procedure suggest removing Ethnicity, Education, Married, and Gender from the model.

With a more accurate model containing what we believe to be the most influential predictor variables, we can now review the model summary and consider appropriate transformations to further enhance the model’s performance.

m2<-lm(newbalance~Income+Rating+Limit+Cards+Age+Student, data = Credit)  
summary(m2)

##   
## Call:  
## lm(formula = newbalance ~ Income + Rating + Limit + Cards + Age +   
## Student, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -170.00 -77.85 -11.84 56.87 313.52   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -493.72419 24.82476 -19.888 < 2e-16 \*\*\*  
## Income -7.79508 0.23342 -33.395 < 2e-16 \*\*\*  
## Rating 1.09119 0.48480 2.251 0.0250 \*   
## Limit 0.19369 0.03238 5.981 4.98e-09 \*\*\*  
## Cards 18.21190 4.31865 4.217 3.08e-05 \*\*\*  
## Age -0.62406 0.29182 -2.139 0.0331 \*   
## StudentYes 425.60994 16.50956 25.780 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 98.61 on 393 degrees of freedom  
## Multiple R-squared: 0.9547, Adjusted R-squared: 0.954   
## F-statistic: 1380 on 6 and 393 DF, p-value: < 2.2e-16

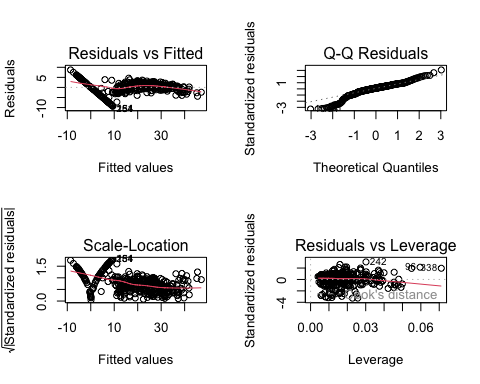
par(mfrow=c(2,2))  
plot(m2)

 After removing Ethnicity, Education, Married, and Gender, we refit the model with the remaining variables: Income, Rating, Limit, Cards, Age, and Student. The summary of the updated model shows that all variables are statistically significant with p-values below the 0.05 threshold. The residuals plot shows improvement, although there is still some minor curvature, suggesting further transformations may be necessary.

m3<-lm(sqrt(newbalance)~sqrt(Income)+sqrt(Rating)+sqrt(Limit)+sqrt(Cards)+Age+Student, data = Credit)  
summary(m3)

##   
## Call:  
## lm(formula = sqrt(newbalance) ~ sqrt(Income) + sqrt(Rating) +   
## sqrt(Limit) + sqrt(Cards) + Age + Student, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.305 -1.147 0.353 1.615 8.701   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -33.895933 1.160785 -29.201 < 2e-16 \*\*\*  
## sqrt(Income) -2.940520 0.094224 -31.208 < 2e-16 \*\*\*  
## sqrt(Rating) 2.724146 0.442032 6.163 1.77e-09 \*\*\*  
## sqrt(Limit) 0.305004 0.105274 2.897 0.00398 \*\*   
## sqrt(Cards) 0.442398 0.423366 1.045 0.29669   
## Age -0.015086 0.008513 -1.772 0.07715 .   
## StudentYes 10.232160 0.480579 21.291 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.875 on 393 degrees of freedom  
## Multiple R-squared: 0.9497, Adjusted R-squared: 0.9489   
## F-statistic: 1235 on 6 and 393 DF, p-value: < 2.2e-16

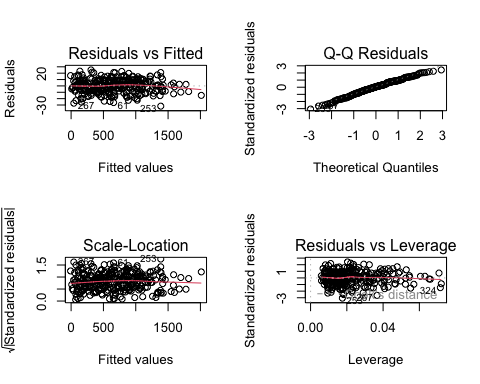
par(mfrow=c(2,2))  
plot(m3)

 To address the remaining curvature in the residuals plot, we applied a square root transformation to newbalance, Income, Rating, Limit, and Cards. The updated model summary shows that all variables are statistically significant with p-values below the 0.05 threshold. The residuals plot shows improvement, with less curvature than the previous model. The R-squared value has increased, indicating that the model explains more of the variation in newbalance. We believe this model is the best fit for the data and provides valuable insights into the factors that influence credit card balances. Now we will focus on handling those zero balances…

Credit$Balance[Credit$Balance == 0] <- NA  
Credit2 <- na.omit(Credit)  
m4<-lm(Balance~Income+Rating+Limit+Cards+Age+Student, data = Credit2)  
summary(m4)

##   
## Call:  
## lm(formula = Balance ~ Income + Rating + Limit + Cards + Age +   
## Student, data = Credit2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.784 -6.311 0.769 7.711 25.113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.985e+02 3.294e+00 -212.053 <2e-16 \*\*\*  
## Income -9.993e+00 2.931e-02 -340.893 <2e-16 \*\*\*  
## Rating -1.501e-01 5.831e-02 -2.574 0.0105 \*   
## Limit 3.365e-01 3.943e-03 85.333 <2e-16 \*\*\*  
## Cards 2.550e+01 4.973e-01 51.278 <2e-16 \*\*\*  
## Age -1.000e+00 3.539e-02 -28.260 <2e-16 \*\*\*  
## StudentYes 5.010e+02 1.843e+00 271.823 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.48 on 303 degrees of freedom  
## Multiple R-squared: 0.9994, Adjusted R-squared: 0.9994   
## F-statistic: 8.031e+04 on 6 and 303 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(m4)

 By removing the zero balances, our model greatly improved. The residual plots showed no signs of curvature, and the Normal Q-Q plot shows that the model fits well, even in the tails of the distribution. All predictor variables remained significant, confirming that Income, Rating, Limit, Cards, Age, and Student are important factors in predicting an individual’s credit card balance.

### Problem 3 Conclusions and Recommended Pathforward

In conclusion, our final model (m4) indicates that Income, Rating, Limit, Cards, Age, and Student status are significant predictors of an individual’s credit balance. By excluding variables like Gender and Ethnicity, we avoided potential ethical or legal concerns related to discrimination in predictive modeling. Additionally, removing the zero-balance observations improved the model’s fit and predictive power, making this the most reliable model for understanding and predicting credit card balances in this dataset.