BUDA 525: Team 4 Final Project

Ryan Antonini, Danny Germain, Joshua Meadows, Josh Nelson, Bill Robertson

9/27/2019

# Problem 1

# Problem 2

# Problem 3

## Loading required package: carData

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

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

## 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

##   
## Call:  
## lm(formula = Balance ~ ID + Income + Limit + Rating + cardsF +   
## Education + Gender + Student + Married + Ethnicity, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -184.54 -75.66 -9.41 54.95 326.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -507.54555 35.81376 -14.172 < 2e-16 \*\*\*  
## ID 0.03843 0.04456 0.862 0.38901   
## Income -7.92408 0.23798 -33.297 < 2e-16 \*\*\*  
## Limit 0.19633 0.03397 5.780 1.56e-08 \*\*\*  
## Rating 1.06729 0.50779 2.102 0.03622 \*   
## cardsF2 29.15883 17.54598 1.662 0.09736 .   
## cardsF3 47.11413 18.78843 2.508 0.01257 \*   
## cardsF4 61.73608 20.11055 3.070 0.00230 \*\*   
## cardsF5 77.90630 25.63015 3.040 0.00253 \*\*   
## cardsF6 92.28072 34.85098 2.648 0.00844 \*\*   
## cardsF7 139.22552 55.08664 2.527 0.01189 \*   
## cardsF8 124.59103 103.20110 1.207 0.22808   
## cardsF9 50.56933 103.24161 0.490 0.62455   
## Education -1.34671 1.63188 -0.825 0.40975   
## GenderFemale -11.06096 10.18698 -1.086 0.27826   
## StudentYes 428.18890 16.95558 25.254 < 2e-16 \*\*\*  
## MarriedYes -6.27327 10.58039 -0.593 0.55359   
## EthnicityAsian 19.20353 14.32735 1.340 0.18093   
## EthnicityCaucasian 12.11215 12.74887 0.950 0.34269   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.94 on 381 degrees of freedom  
## Multiple R-squared: 0.9549, Adjusted R-squared: 0.9528   
## F-statistic: 448 on 18 and 381 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = Balance ~ ID + Income + Limit + Rating + cardsF +   
## Education + Student + Married, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -190.42 -74.38 -10.03 54.64 320.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -501.43132 33.98339 -14.755 < 2e-16 \*\*\*  
## ID 0.04506 0.04397 1.025 0.30611   
## Income -7.91198 0.23768 -33.288 < 2e-16 \*\*\*  
## Limit 0.19695 0.03383 5.822 1.23e-08 \*\*\*  
## Rating 1.05269 0.50564 2.082 0.03801 \*   
## cardsF2 26.20862 17.33995 1.511 0.13149   
## cardsF3 43.79288 18.57171 2.358 0.01887 \*   
## cardsF4 61.67159 20.10413 3.068 0.00231 \*\*   
## cardsF5 74.86257 25.20690 2.970 0.00317 \*\*   
## cardsF6 92.71997 34.84147 2.661 0.00811 \*\*   
## cardsF7 139.20064 54.81685 2.539 0.01150 \*   
## cardsF8 130.76655 103.06927 1.269 0.20531   
## cardsF9 54.14897 103.15677 0.525 0.59994   
## Education -1.30012 1.63112 -0.797 0.42590   
## StudentYes 428.10121 16.89403 25.340 < 2e-16 \*\*\*  
## MarriedYes -4.90713 10.50364 -0.467 0.64063   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.93 on 384 degrees of freedom  
## Multiple R-squared: 0.9545, Adjusted R-squared: 0.9528   
## F-statistic: 537.5 on 15 and 384 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student + Married  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 381 3805247   
## 2 384 3834430 -3 -29184 0.974 0.405

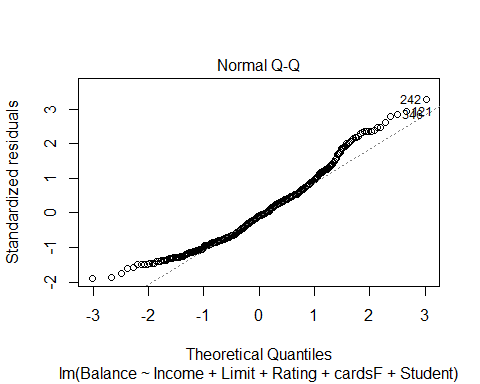
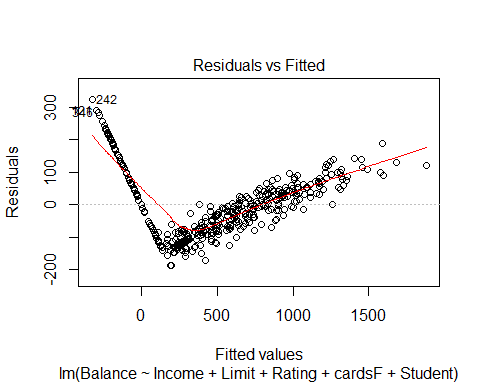
## Start: AIC=3699.23  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student + Married  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 2179 3836610 3697.5  
## - Education 1 6344 3840775 3697.9  
## - ID 1 10486 3844917 3698.3  
## <none> 3834430 3699.2  
## - cardsF 8 179402 4013832 3701.5  
## - Rating 1 43280 3877710 3701.7  
## - Limit 1 338464 4172894 3731.1  
## - Student 1 6412036 10246466 4090.4  
## - Income 1 11064791 14899221 4240.1  
##   
## Step: AIC=3697.45  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 6912 3843522 3696.2  
## - ID 1 10334 3846943 3696.5  
## <none> 3836610 3697.5  
## - Rating 1 41662 3878272 3699.8  
## - cardsF 8 183405 4020015 3700.1  
## - Limit 1 348922 4185532 3730.3  
## - Student 1 6478791 10315401 4091.1  
## - Income 1 11080324 14916934 4238.6  
##   
## Step: AIC=3696.17  
## Balance ~ ID + Income + Limit + Rating + cardsF + Student  
##   
## Df Sum of Sq RSS AIC  
## - ID 1 10483 3854005 3695.3  
## <none> 3843522 3696.2  
## - cardsF 8 182291 4025814 3698.7  
## - Rating 1 44953 3888475 3698.8  
## - Limit 1 343052 4186574 3728.4  
## - Student 1 6482944 10326466 4089.5  
## - Income 1 11073420 14916942 4236.6  
##   
## Step: AIC=3695.26  
## Balance ~ Income + Limit + Rating + cardsF + Student  
##   
## Df Sum of Sq RSS AIC  
## <none> 3854005 3695.3  
## - cardsF 8 178497 4032502 3697.4  
## - Rating 1 44444 3898449 3697.8  
## - Limit 1 344469 4198474 3727.5  
## - Student 1 6472644 10326649 4087.5  
## - Income 1 11063062 14917067 4234.6

##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + cardsF + Student,   
## data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.41 -77.05 -7.76 52.89 323.31   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -513.01919 22.52924 -22.771 < 2e-16 \*\*\*  
## Income -7.90308 0.23712 -33.330 < 2e-16 \*\*\*  
## Limit 0.19669 0.03344 5.881 8.81e-09 \*\*\*  
## Rating 1.05616 0.49995 2.113 0.03528 \*   
## cardsF2 26.02688 17.23430 1.510 0.13181   
## cardsF3 43.88821 18.36755 2.389 0.01735 \*   
## cardsF4 60.77938 19.88926 3.056 0.00240 \*\*   
## cardsF5 74.46820 25.09260 2.968 0.00319 \*\*   
## cardsF6 93.58965 34.59319 2.705 0.00712 \*\*   
## cardsF7 136.84961 54.67609 2.503 0.01273 \*   
## cardsF8 134.51541 102.50884 1.312 0.19022   
## cardsF9 66.16098 102.55960 0.645 0.51925   
## StudentYes 426.95343 16.74713 25.494 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.79 on 387 degrees of freedom  
## Multiple R-squared: 0.9543, Adjusted R-squared: 0.9529   
## F-statistic: 673.5 on 12 and 387 DF, p-value: < 2.2e-16

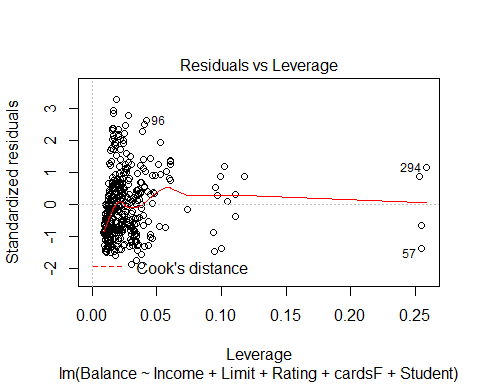
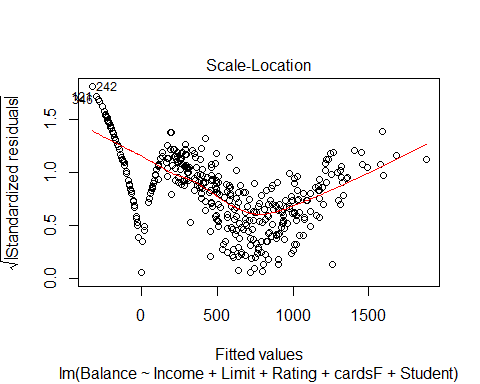
## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ Income + Limit + Rating + cardsF + Student  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 381 3805247   
## 2 387 3854005 -6 -48758 0.8137 0.5598

plot(mod3\_3)

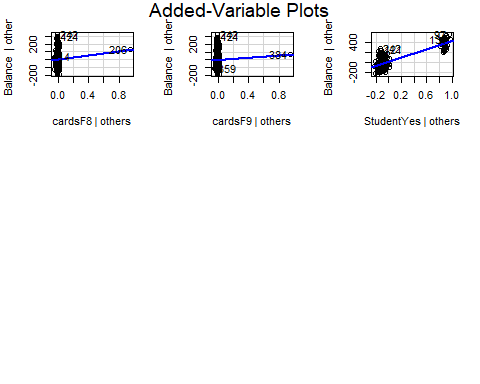
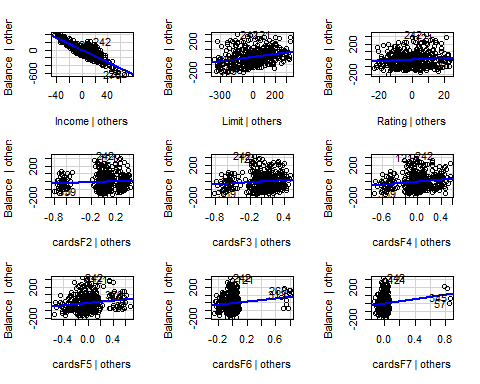
## Warning: not plotting observations with leverage one:  
## 206, 384



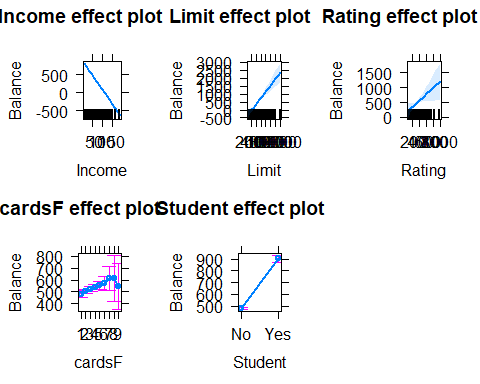
## Warning: not plotting observations with leverage one:  
## 206, 384



avPlots(mod3\_3)



plot(allEffects(mod3\_3))



#Running Diagnostics

There is some serious NCV that needs delt with, coming from entries where Balance=0

Credit2<- Credit[Credit$Balance!=0,]  
mod3\_4 <- step(lm(Balance~ID+Income+Limit+Rating+cardsF+Education+Gender+Student+Married+Ethnicity,data=Credit2))

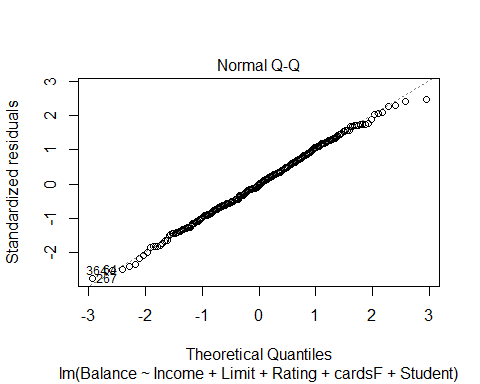
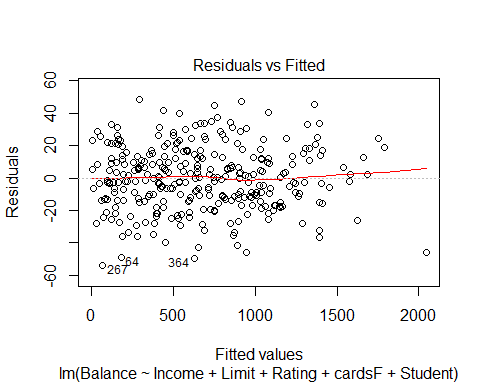
## Start: AIC=1877.77  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married + Ethnicity  
##   
## Df Sum of Sq RSS AIC  
## - Ethnicity 2 117 117280 1874.1  
## - Gender 1 2 117165 1875.8  
## - Married 1 27 117190 1875.8  
## - Education 1 196 117360 1876.3  
## - ID 1 479 117642 1877.0  
## <none> 117163 1877.8  
## - Rating 1 1280 118443 1879.1  
## - cardsF 8 286412 403576 2245.2  
## - Limit 1 768412 885576 2502.8  
## - Student 1 7847230 7964394 3183.7  
## - Income 1 13015563 13132727 3338.8  
##   
## Step: AIC=1874.08  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married  
##   
## Df Sum of Sq RSS AIC  
## - Gender 1 0 117281 1872.1  
## - Married 1 23 117303 1872.1  
## - Education 1 198 117479 1872.6  
## - ID 1 430 117710 1873.2  
## <none> 117280 1874.1  
## - Rating 1 1222 118503 1875.3  
## - cardsF 8 286969 404249 2241.7  
## - Limit 1 773659 890940 2500.7  
## - Student 1 7921893 8039173 3182.6  
## - Income 1 13029570 13146850 3335.1  
##   
## Step: AIC=1872.08  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student + Married  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 23 117303 1870.1  
## - Education 1 198 117479 1870.6  
## - ID 1 431 117712 1871.2  
## <none> 117281 1872.1  
## - Rating 1 1222 118503 1873.3  
## - cardsF 8 287259 404539 2239.9  
## - Limit 1 773673 890954 2498.7  
## - Student 1 7938234 8055515 3181.2  
## - Income 1 13029718 13146999 3333.1  
##   
## Step: AIC=1870.14  
## Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 216 117519 1868.7  
## - ID 1 435 117738 1869.3  
## <none> 117303 1870.1  
## - Rating 1 1271 118574 1871.5  
## - cardsF 8 288387 405691 2238.8  
## - Limit 1 783493 900797 2500.1  
## - Student 1 8048153 8165456 3183.4  
## - Income 1 13056303 13173607 3331.7  
##   
## Step: AIC=1868.71  
## Balance ~ ID + Income + Limit + Rating + cardsF + Student  
##   
## Df Sum of Sq RSS AIC  
## - ID 1 437 117956 1867.9  
## <none> 117519 1868.7  
## - Rating 1 1196 118715 1869.9  
## - cardsF 8 289251 406770 2237.6  
## - Limit 1 786629 904148 2499.2  
## - Student 1 8089010 8206529 3183.0  
## - Income 1 13074216 13191735 3330.1  
##   
## Step: AIC=1867.86  
## Balance ~ Income + Limit + Rating + cardsF + Student  
##   
## Df Sum of Sq RSS AIC  
## <none> 117956 1867.9  
## - Rating 1 1123 119079 1868.8  
## - cardsF 8 289222 407178 2235.9  
## - Limit 1 786837 904793 2497.5  
## - Student 1 8128724 8246680 3182.5  
## - Income 1 13081429 13199386 3328.3

summary(mod3\_4)

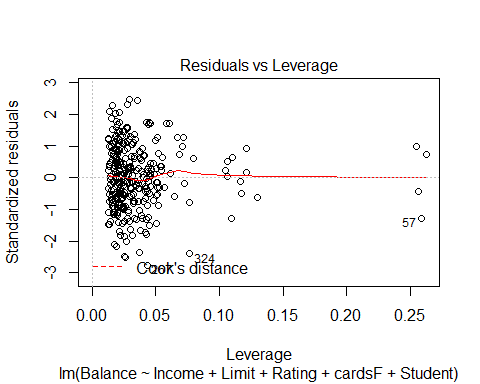
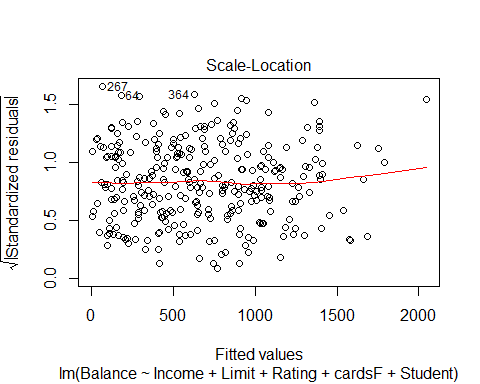
##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + cardsF + Student,   
## data = Credit2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.75 -13.07 0.00 13.23 48.43   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.224e+02 5.756e+00 -125.509 < 2e-16 \*\*\*  
## Income -1.013e+01 5.582e-02 -181.487 < 2e-16 \*\*\*  
## Limit 3.399e-01 7.636e-03 44.510 < 2e-16 \*\*\*  
## Rating -1.899e-01 1.129e-01 -1.682 0.0937 .   
## cardsF2 2.426e+01 3.874e+00 6.263 1.32e-09 \*\*\*  
## cardsF3 4.692e+01 4.122e+00 11.384 < 2e-16 \*\*\*  
## cardsF4 7.646e+01 4.391e+00 17.412 < 2e-16 \*\*\*  
## cardsF5 9.627e+01 5.595e+00 17.206 < 2e-16 \*\*\*  
## cardsF6 1.232e+02 7.330e+00 16.810 < 2e-16 \*\*\*  
## cardsF7 1.607e+02 1.110e+01 14.476 < 2e-16 \*\*\*  
## cardsF8 2.127e+02 2.060e+01 10.329 < 2e-16 \*\*\*  
## cardsF9 1.901e+02 2.064e+01 9.207 < 2e-16 \*\*\*  
## StudentYes 5.031e+02 3.517e+00 143.064 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.93 on 297 degrees of freedom  
## Multiple R-squared: 0.9978, Adjusted R-squared: 0.9977   
## F-statistic: 1.108e+04 on 12 and 297 DF, p-value: < 2.2e-16

plot(mod3\_4)

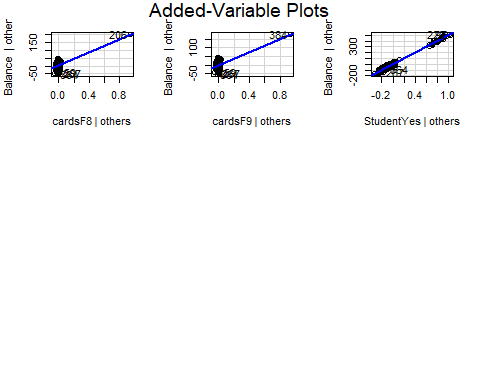
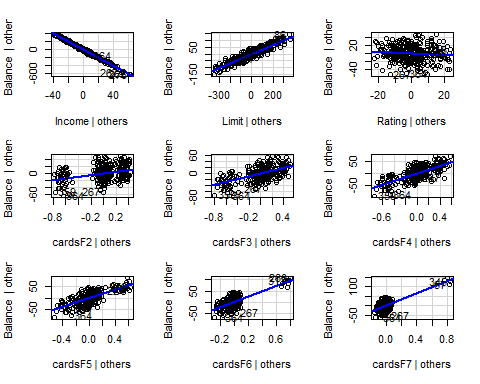
## Warning: not plotting observations with leverage one:  
## 153, 299



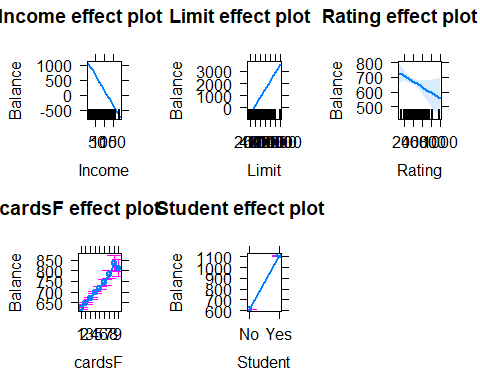
## Warning: not plotting observations with leverage one:  
## 153, 299



avPlots(mod3\_4)



plot(allEffects(mod3\_4))



ncvTest(mod3\_4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.03635902, Df = 1, p = 0.84878

#Running Diagnostics

By removing the 90 cases where balance is zero, we can do an extremely good job at prediciting balance with only 5 predictors. Next, we create a new variable that reports whether or not the person has a balance, and use this variable as a response.

library(car)  
Credit3<-Credit  
Credit3$BalanceF<- as.numeric(Credit3$Balance>0)  
mod3\_5<- lm(BalanceF~Limit+Student+Rating+cardsF+Age+Education+Gender+Married+Ethnicity+Income+ID,data=Credit3)  
summary(mod3\_5)

##   
## Call:  
## lm(formula = BalanceF ~ Limit + Student + Rating + cardsF + Age +   
## Education + Gender + Married + Ethnicity + Income + ID, data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.57982 -0.24068 -0.00183 0.24467 0.54516   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.254e-01 1.122e-01 1.118 0.2643   
## Limit 2.514e-04 9.817e-05 2.561 0.0108 \*   
## StudentYes 2.595e-01 4.905e-02 5.291 2.06e-07 \*\*\*  
## Rating -8.131e-04 1.468e-03 -0.554 0.5799   
## cardsF2 3.764e-02 5.081e-02 0.741 0.4593   
## cardsF3 2.028e-02 5.441e-02 0.373 0.7095   
## cardsF4 5.731e-02 5.816e-02 0.985 0.3251   
## cardsF5 4.864e-02 7.435e-02 0.654 0.5134   
## cardsF6 1.200e-01 1.009e-01 1.189 0.2350   
## cardsF7 5.322e-02 1.593e-01 0.334 0.7384   
## cardsF8 2.370e-01 2.988e-01 0.793 0.4283   
## cardsF9 3.265e-01 2.986e-01 1.093 0.2750   
## Age -1.675e-04 8.687e-04 -0.193 0.8472   
## Education -2.058e-03 4.717e-03 -0.436 0.6628   
## GenderFemale 3.068e-02 2.944e-02 1.042 0.2981   
## MarriedYes 3.047e-02 3.067e-02 0.994 0.3210   
## EthnicityAsian -4.864e-02 4.146e-02 -1.173 0.2415   
## EthnicityCaucasian 6.866e-03 3.685e-02 0.186 0.8523   
## Income -7.282e-03 6.963e-04 -10.458 < 2e-16 \*\*\*  
## ID 1.310e-04 1.289e-04 1.016 0.3104   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2888 on 380 degrees of freedom  
## Multiple R-squared: 0.5455, Adjusted R-squared: 0.5228   
## F-statistic: 24.01 on 19 and 380 DF, p-value: < 2.2e-16

mod3\_6<-step(mod3\_5)

## Start: AIC=-974.07  
## BalanceF ~ Limit + Student + Rating + cardsF + Age + Education +   
## Gender + Married + Ethnicity + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - cardsF 8 0.2914 31.990 -986.41  
## - Age 1 0.0031 31.702 -976.04  
## - Education 1 0.0159 31.715 -975.87  
## - Rating 1 0.0256 31.724 -975.75  
## - Ethnicity 2 0.2095 31.908 -975.44  
## - Married 1 0.0824 31.781 -975.04  
## - ID 1 0.0860 31.785 -974.99  
## - Gender 1 0.0906 31.789 -974.93  
## <none> 31.699 -974.07  
## - Limit 1 0.5471 32.246 -969.23  
## - Student 1 2.3351 34.034 -947.64  
## - Income 1 9.1234 40.822 -874.90  
##   
## Step: AIC=-986.41  
## BalanceF ~ Limit + Student + Rating + Age + Education + Gender +   
## Married + Ethnicity + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - Rating 1 0.0000 31.990 -988.41  
## - Age 1 0.0017 31.992 -988.39  
## - Education 1 0.0236 32.014 -988.12  
## - Ethnicity 2 0.2209 32.211 -987.66  
## - Married 1 0.0681 32.058 -987.56  
## - Gender 1 0.0731 32.063 -987.50  
## - ID 1 0.0813 32.071 -987.40  
## <none> 31.990 -986.41  
## - Limit 1 0.5066 32.497 -982.13  
## - Student 1 2.3196 34.310 -960.41  
## - Income 1 9.6885 41.679 -882.59  
##   
## Step: AIC=-988.41  
## BalanceF ~ Limit + Student + Age + Education + Gender + Married +   
## Ethnicity + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - Age 1 0.0017 31.992 -990.39  
## - Education 1 0.0238 32.014 -990.12  
## - Ethnicity 2 0.2215 32.212 -989.65  
## - Married 1 0.0686 32.059 -989.56  
## - Gender 1 0.0731 32.063 -989.50  
## - ID 1 0.0814 32.072 -989.40  
## <none> 31.990 -988.41  
## - Student 1 2.3295 34.320 -962.30  
## - Income 1 9.6953 41.686 -884.52  
## - Limit 1 30.7898 62.780 -720.73  
##   
## Step: AIC=-990.39  
## BalanceF ~ Limit + Student + Education + Gender + Married + Ethnicity +   
## Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 0.0240 32.016 -992.09  
## - Ethnicity 2 0.2203 32.212 -991.65  
## - Married 1 0.0707 32.063 -991.51  
## - Gender 1 0.0727 32.065 -991.48  
## - ID 1 0.0804 32.072 -991.39  
## <none> 31.992 -990.39  
## - Student 1 2.3384 34.330 -964.17  
## - Income 1 9.9923 41.984 -883.67  
## - Limit 1 30.9557 62.948 -721.66  
##   
## Step: AIC=-992.09  
## BalanceF ~ Limit + Student + Gender + Married + Ethnicity + Income +   
## ID  
##   
## Df Sum of Sq RSS AIC  
## - Ethnicity 2 0.2242 32.240 -993.30  
## - Married 1 0.0664 32.082 -993.26  
## - Gender 1 0.0736 32.089 -993.17  
## - ID 1 0.0801 32.096 -993.09  
## <none> 32.016 -992.09  
## - Student 1 2.3160 34.332 -966.16  
## - Income 1 9.9766 41.992 -885.59  
## - Limit 1 30.9543 62.970 -723.52  
##   
## Step: AIC=-993.3  
## BalanceF ~ Limit + Student + Gender + Married + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 0.0497 32.290 -994.69  
## - Gender 1 0.0706 32.311 -994.43  
## - ID 1 0.1009 32.341 -994.05  
## <none> 32.240 -993.30  
## - Student 1 2.2404 34.480 -968.43  
## - Income 1 10.0383 42.278 -886.88  
## - Limit 1 31.1434 63.383 -724.90  
##   
## Step: AIC=-994.69  
## BalanceF ~ Limit + Student + Gender + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - Gender 1 0.0730 32.363 -995.78  
## - ID 1 0.1038 32.394 -995.40  
## <none> 32.290 -994.69  
## - Student 1 2.2021 34.492 -970.30  
## - Income 1 10.0130 42.303 -888.65  
## - Limit 1 31.1475 63.437 -726.57  
##   
## Step: AIC=-995.78  
## BalanceF ~ Limit + Student + Income + ID  
##   
## Df Sum of Sq RSS AIC  
## - ID 1 0.0858 32.448 -996.72  
## <none> 32.363 -995.78  
## - Student 1 2.2500 34.613 -970.90  
## - Income 1 10.0702 42.433 -889.42  
## - Limit 1 31.2680 63.631 -727.35  
##   
## Step: AIC=-996.72  
## BalanceF ~ Limit + Student + Income  
##   
## Df Sum of Sq RSS AIC  
## <none> 32.448 -996.72  
## - Student 1 2.2175 34.666 -972.28  
## - Income 1 10.0220 42.470 -891.06  
## - Limit 1 31.2380 63.686 -729.00

summary(mod3\_6)

##   
## Call:  
## lm(formula = BalanceF ~ Limit + Student + Income, data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56004 -0.24852 0.01882 0.23604 0.52837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.425e-01 3.408e-02 4.181 3.57e-05 \*\*\*  
## Limit 1.987e-04 1.018e-05 19.525 < 2e-16 \*\*\*  
## StudentYes 2.484e-01 4.775e-02 5.202 3.17e-07 \*\*\*  
## Income -7.373e-03 6.667e-04 -11.059 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2863 on 396 degrees of freedom  
## Multiple R-squared: 0.5348, Adjusted R-squared: 0.5313   
## F-statistic: 151.7 on 3 and 396 DF, p-value: < 2.2e-16

We have found the optimal factors to use. Next we see if the interactions between these factors can help us.

mod3\_7<-step(lm(BalanceF~Student\*Limit\*Income,data=Credit3))

## Start: AIC=-1065.42  
## BalanceF ~ Student \* Limit \* Income  
##   
## Df Sum of Sq RSS AIC  
## - Student:Limit:Income 1 0.037525 26.824 -1066.9  
## <none> 26.787 -1065.4  
##   
## Step: AIC=-1066.86  
## BalanceF ~ Student + Limit + Income + Student:Limit + Student:Income +   
## Limit:Income  
##   
## Df Sum of Sq RSS AIC  
## <none> 26.824 -1066.9  
## - Student:Income 1 1.1339 27.958 -1052.3  
## - Student:Limit 1 2.8001 29.624 -1029.2  
## - Limit:Income 1 3.2015 30.026 -1023.8

summary(mod3\_7)

##   
## Call:  
## lm(formula = BalanceF ~ Student + Limit + Income + Student:Limit +   
## Student:Income + Limit:Income, data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.80561 -0.18521 -0.00653 0.19275 0.52459   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.697e-01 5.047e-02 -3.362 0.000849 \*\*\*  
## StudentYes 8.381e-01 1.074e-01 7.802 5.55e-14 \*\*\*  
## Limit 2.572e-04 1.148e-05 22.398 < 2e-16 \*\*\*  
## Income -1.101e-03 1.224e-03 -0.899 0.369158   
## StudentYes:Limit -2.014e-04 3.144e-05 -6.405 4.30e-10 \*\*\*  
## StudentYes:Income 7.424e-03 1.822e-03 4.076 5.55e-05 \*\*\*  
## Limit:Income -8.907e-07 1.301e-07 -6.849 2.88e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2613 on 393 degrees of freedom  
## Multiple R-squared: 0.6154, Adjusted R-squared: 0.6096   
## F-statistic: 104.8 on 6 and 393 DF, p-value: < 2.2e-16

Since these models are not being used like normal linear model, we use random splitting to test which one most effeciently predicts when the balance will be zero.

library(doParallel)

## Loading required package: foreach

## Loading required package: iterators

## Loading required package: parallel

mine<-detectCores()  
mine<-min(c(max(c(1,mine-1)),5))  
cl<-makeCluster(mine)  
registerDoParallel(cl)  
getDoParWorkers()

## [1] 5

library(doRNG)

## Loading required package: rngtools

## Loading required package: pkgmaker

## Loading required package: registry

##   
## Attaching package: 'pkgmaker'

## The following object is masked from 'package:base':  
##   
## isFALSE

library(foreach)  
  
foreach(i=1:1000,.combine="+",.options.RNG=623)%dopar% {  
 set=sample(1:dim(Credit3)[1],300,replace=FALSE)  
 M1<-lm(BalanceF~Student\*Limit+Student\*Income+Income\*Limit,data=Credit3[set,])  
 Predict <- predict(M1,newdata=Credit3[-set,])  
 myPredict<- ifelse(Predict >0.5,"1","0")  
 mytable <- table(Credit3[-set,]$BalanceF,myPredict)  
 eff<-sum(diag(mytable))/sum(mytable)  
 return(eff)  
}

## [1] 978.98

foreach(i=1:1000,.combine="+",.options.RNG=623)%dopar% {  
 set=sample(1:dim(Credit3)[1],300,replace=FALSE)  
 M1<-lm(BalanceF~Student+Limit+Income,data=Credit3[set,])  
 Predict <- predict(M1,newdata=Credit3[-set,])  
 myPredict<- ifelse(Predict >0.5,"1","0")  
 mytable <- table(Credit3[-set,]$BalanceF,myPredict)  
 print(mytable)  
 eff<-sum(diag(mytable))/sum(mytable)  
 return(eff)  
}

## [1] 961.23

stopCluster(cl)

Model 3\_7 predicts correctly 97.8 percent of the time whether or not the balance is zero, while model 3\_6 predicts correctly only 96.2 percent of the time. We can now use Model 3\_7 to predict whether or not the balance is zero, then predict the value of the balance, when appropriate, using Model 3\_4. (This method can be further improved by using a generalized linear model with a binomial distribution instead of a normal one but that is outside the scope of this class).

# Problem 4

library(carData)  
head(Salaries)

## rank discipline yrs.since.phd yrs.service sex salary  
## 1 Prof B 19 18 Male 139750  
## 2 Prof B 20 16 Male 173200  
## 3 AsstProf B 4 3 Male 79750  
## 4 Prof B 45 39 Male 115000  
## 5 Prof B 40 41 Male 141500  
## 6 AssocProf B 6 6 Male 97000

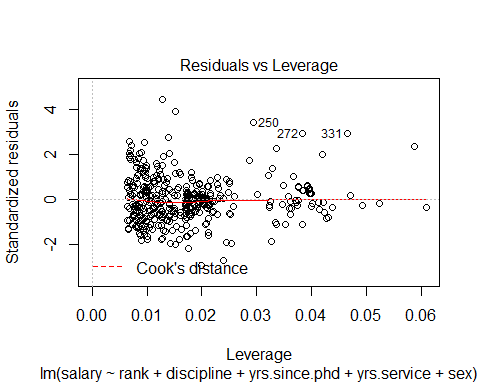
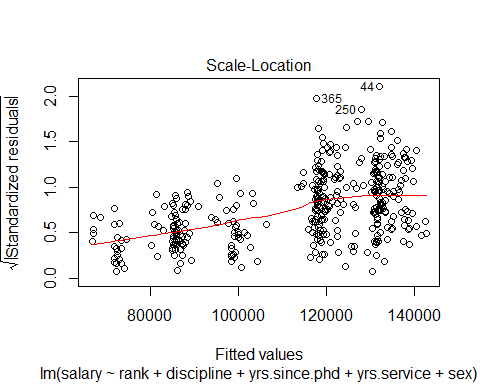
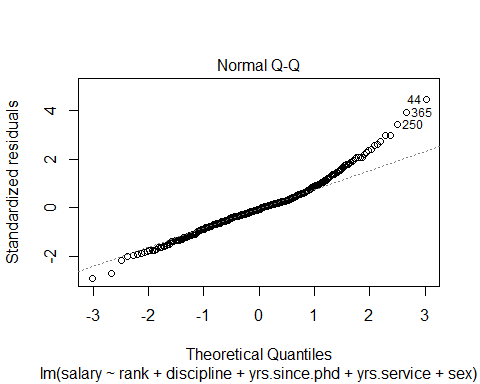
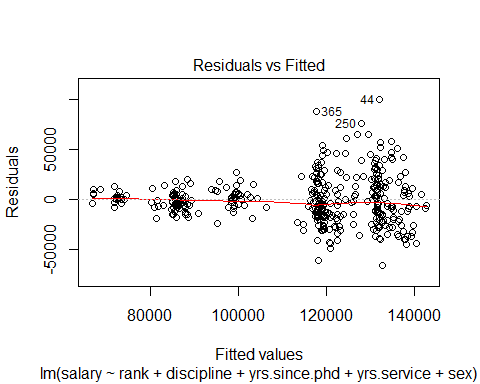
summary(Salaries)

## rank discipline yrs.since.phd yrs.service sex   
## AsstProf : 67 A:181 Min. : 1.00 Min. : 0.00 Female: 39   
## AssocProf: 64 B:216 1st Qu.:12.00 1st Qu.: 7.00 Male :358   
## Prof :266 Median :21.00 Median :16.00   
## Mean :22.31 Mean :17.61   
## 3rd Qu.:32.00 3rd Qu.:27.00   
## Max. :56.00 Max. :60.00   
## salary   
## Min. : 57800   
## 1st Qu.: 91000   
## Median :107300   
## Mean :113706   
## 3rd Qu.:134185   
## Max. :231545

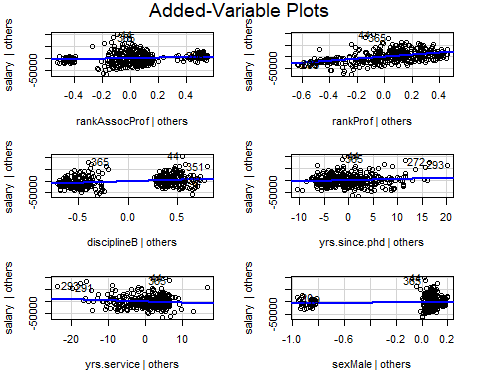
mod4\_1<-lm(salary~rank+discipline+yrs.since.phd+yrs.service+sex, data=Salaries)  
summary(mod4\_1)

##   
## Call:  
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service +   
## sex, data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65248 -13211 -1775 10384 99592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65955.2 4588.6 14.374 < 2e-16 \*\*\*  
## rankAssocProf 12907.6 4145.3 3.114 0.00198 \*\*   
## rankProf 45066.0 4237.5 10.635 < 2e-16 \*\*\*  
## disciplineB 14417.6 2342.9 6.154 1.88e-09 \*\*\*  
## yrs.since.phd 535.1 241.0 2.220 0.02698 \*   
## yrs.service -489.5 211.9 -2.310 0.02143 \*   
## sexMale 4783.5 3858.7 1.240 0.21584   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22540 on 390 degrees of freedom  
## Multiple R-squared: 0.4547, Adjusted R-squared: 0.4463   
## F-statistic: 54.2 on 6 and 390 DF, p-value: < 2.2e-16

plot(mod4\_1)



avPlots(mod4\_1)



plot(allEffects(mod4\_1))

