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

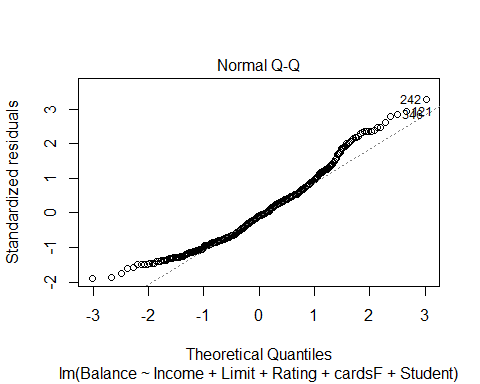
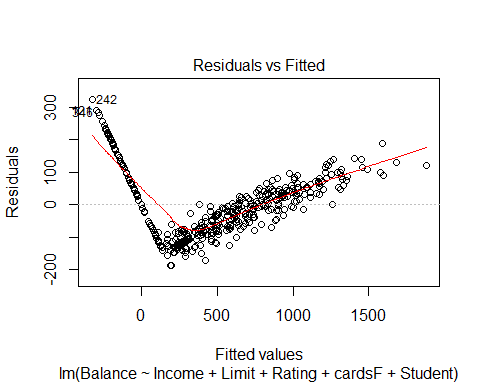
## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Student + Married  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1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

##   
## 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

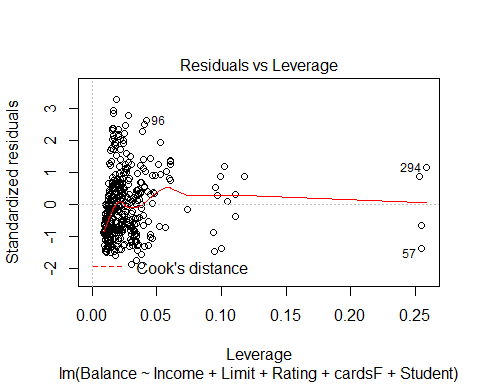
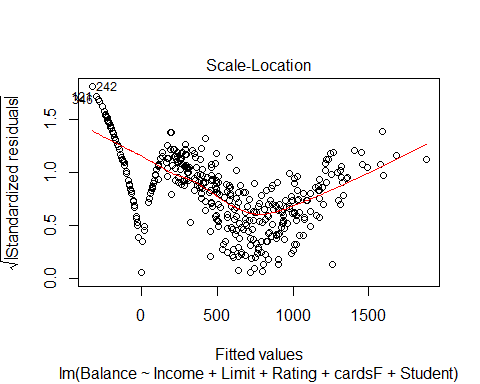
## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + cardsF + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ Income + Limit + Rating + cardsF + Student  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 381 3805247   
## 2 387 3854005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

plot(mod3\_3)

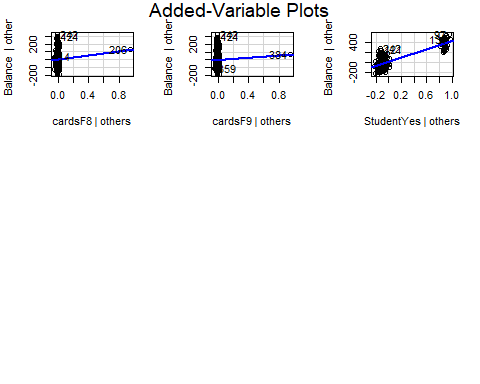
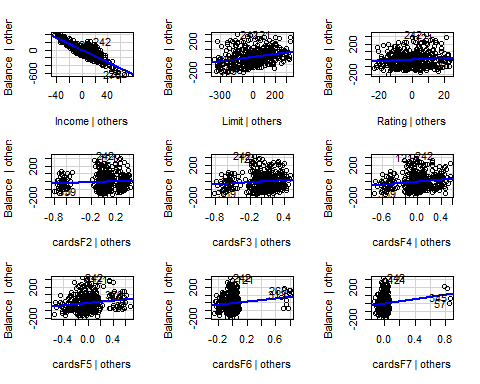
## Warning: not plotting observations with leverage one:  
## 206, 384



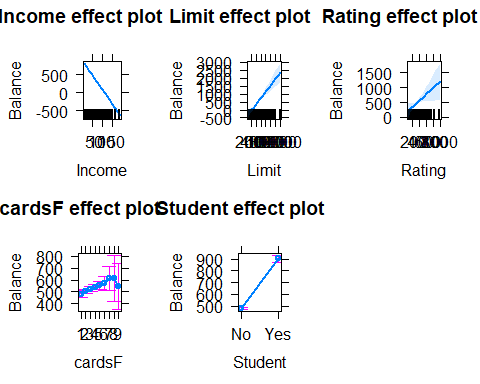
## Warning: not plotting observations with leverage one:  
## 206, 384



avPlots(mod3\_3)



plot(allEffects(mod3\_3))



#Running Diagnostics

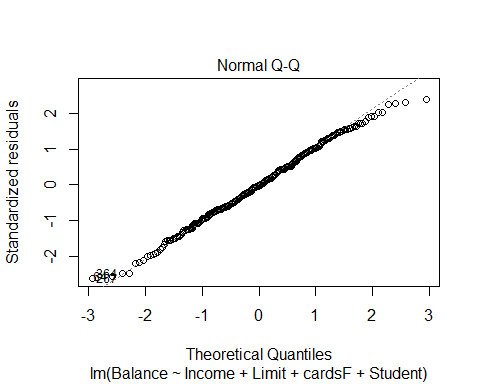
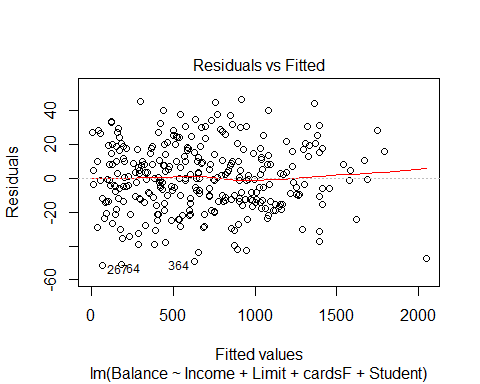
There is some serious NCV that needs delt with, coming from entries where Balance=0

Credit2<- Credit[Credit$Balance!=0,]  
mod3\_4 <- lm(Balance~Income+Limit+cardsF+Student,data=Credit2)  
summary(mod3\_4)

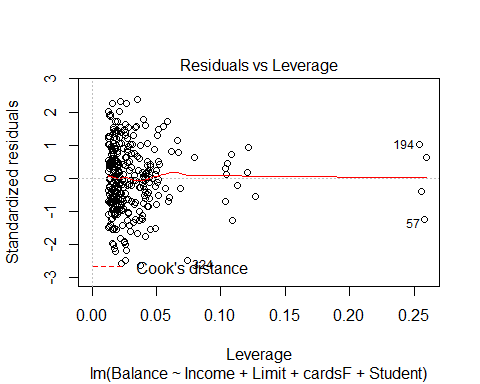
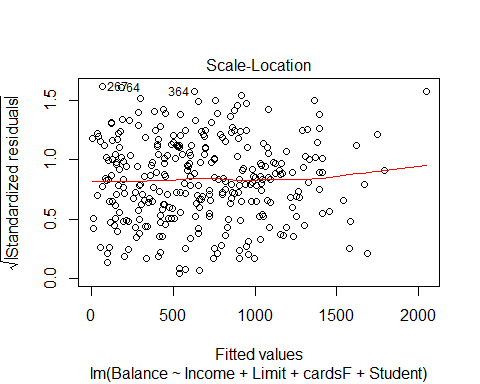
##   
## Call:  
## lm(formula = Balance ~ Income + Limit + cardsF + Student, data = Credit2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.300 -13.319 -0.442 14.219 46.639   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.273e+02 4.992e+00 -145.687 < 2e-16 \*\*\*  
## Income -1.013e+01 5.599e-02 -180.961 < 2e-16 \*\*\*  
## Limit 3.271e-01 1.032e-03 316.866 < 2e-16 \*\*\*  
## cardsF2 2.303e+01 3.816e+00 6.035 4.7e-09 \*\*\*  
## cardsF3 4.439e+01 3.849e+00 11.534 < 2e-16 \*\*\*  
## cardsF4 7.362e+01 4.066e+00 18.108 < 2e-16 \*\*\*  
## cardsF5 9.208e+01 5.025e+00 18.325 < 2e-16 \*\*\*  
## cardsF6 1.198e+02 7.070e+00 16.949 < 2e-16 \*\*\*  
## cardsF7 1.549e+02 1.058e+01 14.641 < 2e-16 \*\*\*  
## cardsF8 2.060e+02 2.027e+01 10.165 < 2e-16 \*\*\*  
## cardsF9 1.830e+02 2.028e+01 9.026 < 2e-16 \*\*\*  
## StudentYes 5.026e+02 3.512e+00 143.117 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.99 on 298 degrees of freedom  
## Multiple R-squared: 0.9978, Adjusted R-squared: 0.9977   
## F-statistic: 1.202e+04 on 11 and 298 DF, p-value: < 2.2e-16

plot(mod3\_4)

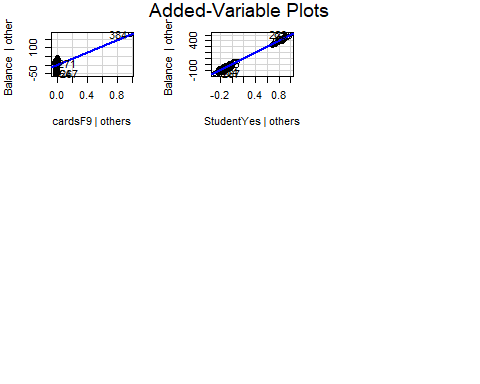
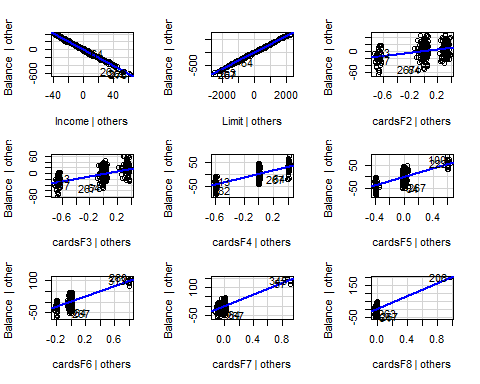
## Warning: not plotting observations with leverage one:  
## 153, 299



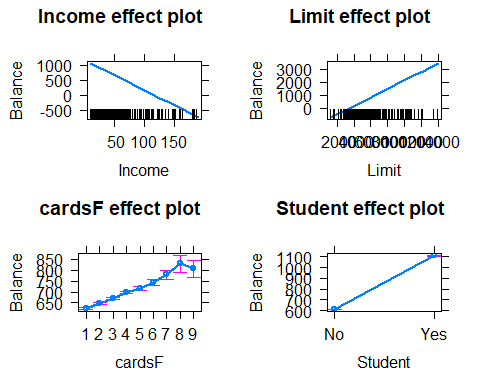
## Warning: not plotting observations with leverage one:  
## 153, 299



avPlots(mod3\_4)



plot(allEffects(mod3\_4))



ncvTest(mod3\_4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.05727568, Df = 1, p = 0.81085

#Running Diagnostics

By removing the 90 cases where balance is zero, we can do an extremely good job at prediciting balance with only 4 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

Anova(mod3\_5,type=2)

## Anova Table (Type II tests)  
##   
## Response: BalanceF  
## Sum Sq Df F value Pr(>F)   
## Limit 0.547 1 6.5585 0.01082 \*   
## Student 2.335 1 27.9931 2.061e-07 \*\*\*  
## Rating 0.026 1 0.3070 0.57985   
## cardsF 0.291 8 0.4367 0.89882   
## Age 0.003 1 0.0372 0.84724   
## Education 0.016 1 0.1904 0.66285   
## Gender 0.091 1 1.0858 0.29807   
## Married 0.082 1 0.9873 0.32104   
## Ethnicity 0.210 2 1.2559 0.28601   
## Income 9.123 1 109.3694 < 2.2e-16 \*\*\*  
## ID 0.086 1 1.0315 0.31044   
## Residuals 31.699 380   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod3\_6<-lm(BalanceF~Limit+Student+Income,data=Credit3)  
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

#testing goodness of fit. Needs to be updated to cross-fold validation  
Credit3$Predict <- ifelse(mod3\_6$fitted.values >0.5,"1","0")  
mytable <- table(Credit3$BalanceF,Credit3$Predict)  
print(mytable)

##   
## 0 1  
## 0 79 11  
## 1 2 308

efficiency <- sum(diag(mytable))/sum(mytable)  
efficiency

## [1] 0.9675

Model 3\_6 predicts correctly 97 percent of the time whether or not the balance is zero. We can now use Model 3\_6 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 linear 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

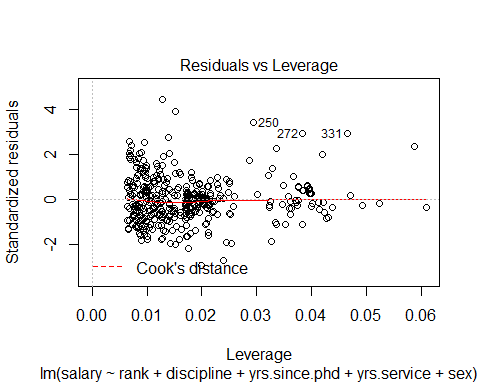
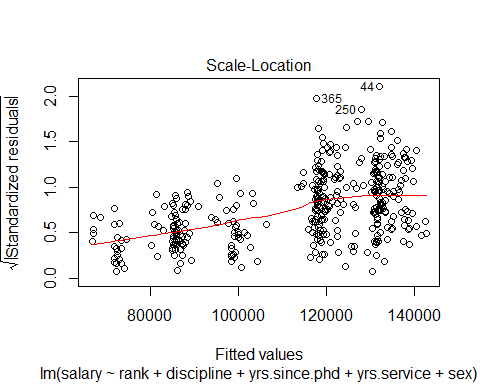
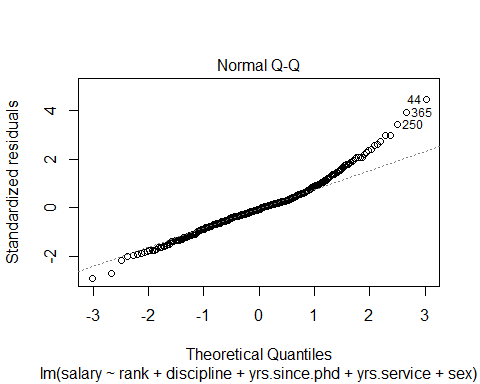
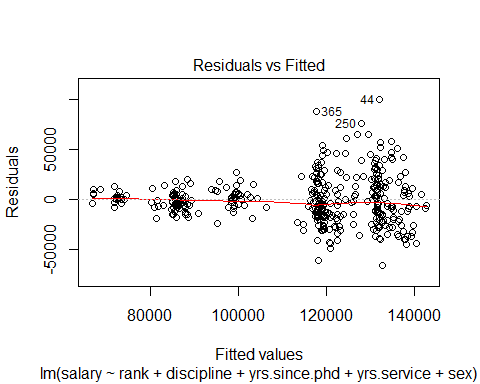
summary(Salaries)

## 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

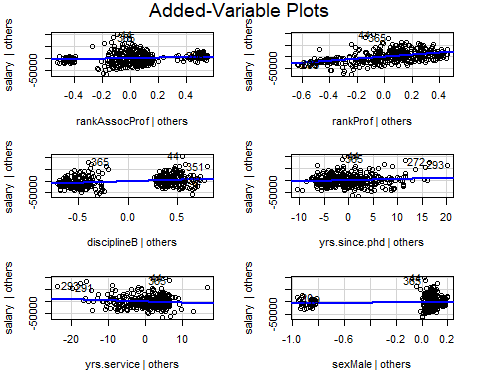
mod4\_1<-lm(salary~rank+discipline+yrs.since.phd+yrs.service+sex, data=Salaries)  
summary(mod4\_1)

##   
## 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))

