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

## 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 + Cards +   
## Education + Gender + Student + Married + Ethnicity, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -181.59 -76.09 -8.09 56.05 328.39   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -519.43345 33.62383 -15.448 < 2e-16 \*\*\*  
## ID 0.03583 0.04356 0.823 0.4112   
## Income -7.88809 0.23231 -33.955 < 2e-16 \*\*\*  
## Limit 0.19170 0.03293 5.821 1.23e-08 \*\*\*  
## Rating 1.13245 0.49323 2.296 0.0222 \*   
## Cards 17.48400 4.36003 4.010 7.28e-05 \*\*\*  
## Education -1.16397 1.60521 -0.725 0.4688   
## GenderFemale -9.98325 10.02803 -0.996 0.3201   
## StudentYes 427.59524 16.79416 25.461 < 2e-16 \*\*\*  
## MarriedYes -7.04296 10.38311 -0.678 0.4980   
## EthnicityAsian 18.65478 14.16076 1.317 0.1885   
## EthnicityCaucasian 10.23033 12.29024 0.832 0.4057   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.26 on 388 degrees of freedom  
## Multiple R-squared: 0.9547, Adjusted R-squared: 0.9534   
## F-statistic: 743 on 11 and 388 DF, p-value: < 2.2e-16

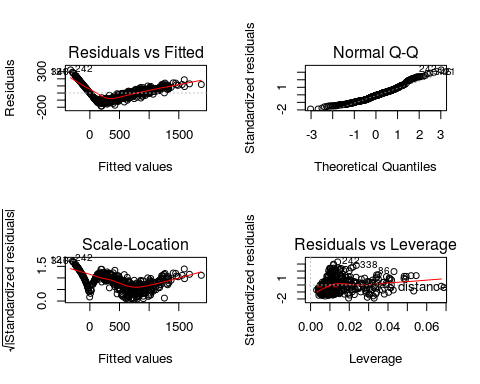
## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student + Married  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 388 3822621   
## 2 391 3849031 -3 -26410 0.8935 0.4445

## Start: AIC=3686.75  
## Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student + Married  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 2803 3851834 3685.0  
## - Education 1 5163 3854194 3685.3  
## - ID 1 8764 3857795 3685.7  
## <none> 3849031 3686.7  
## - Rating 1 48485 3897515 3689.8  
## - Cards 1 164802 4013832 3701.5  
## - Limit 1 344450 4193481 3719.0  
## - Student 1 6432087 10281117 4077.7  
## - Income 1 11365025 15214056 4234.5  
##   
## Step: AIC=3685.04  
## Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 5676 3857510 3683.6  
## - ID 1 8585 3860419 3683.9  
## <none> 3851834 3685.0  
## - Rating 1 46613 3898447 3687.8  
## - Cards 1 168181 4020015 3700.1  
## - Limit 1 354462 4206296 3718.3  
## - Student 1 6507806 10359640 4078.8  
## - Income 1 11372463 15224297 4232.8  
##   
## Step: AIC=3683.63  
## Balance ~ ID + Income + Limit + Rating + Cards + Student  
##   
## Df Sum of Sq RSS AIC  
## - ID 1 8581 3866091 3682.5  
## <none> 3857510 3683.6  
## - Rating 1 49255 3906765 3686.7  
## - Cards 1 168303 4025814 3698.7  
## - Limit 1 349841 4207351 3716.4  
## - Student 1 6516737 10374247 4077.3  
## - Income 1 11367422 15224933 4230.8  
##   
## Step: AIC=3682.52  
## Balance ~ Income + Limit + Rating + Cards + Student  
##   
## Df Sum of Sq RSS AIC  
## <none> 3866091 3682.5  
## - Rating 1 48967 3915058 3685.6  
## - Cards 1 166410 4032502 3697.4  
## - Limit 1 350520 4216611 3715.2  
## - Student 1 6508348 10374439 4075.4  
## - Income 1 11358921 15225012 4228.8

##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Student,   
## data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.08 -77.12 -7.19 53.63 325.13   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -526.15552 19.74661 -26.645 < 2e-16 \*\*\*  
## Income -7.87492 0.23145 -34.024 < 2e-16 \*\*\*  
## Limit 0.19441 0.03253 5.977 5.10e-09 \*\*\*  
## Rating 1.08790 0.48700 2.234 0.026 \*   
## Cards 17.85173 4.33489 4.118 4.66e-05 \*\*\*  
## StudentYes 426.85015 16.57403 25.754 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.06 on 394 degrees of freedom  
## Multiple R-squared: 0.9542, Adjusted R-squared: 0.9536   
## F-statistic: 1640 on 5 and 394 DF, p-value: < 2.2e-16

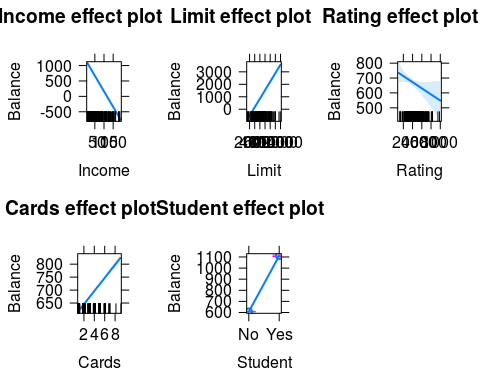
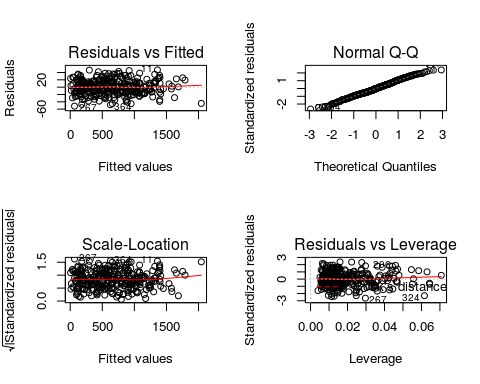
## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ Income + Limit + Rating + Cards + Student  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 388 3822621   
## 2 394 3866091 -6 -43470 0.7354 0.6214

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 112.9874, Df = 1, p = < 2.22e-16



##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Student,   
## data = Credit2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.88 -13.25 -0.86 13.43 47.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.493e+02 5.255e+00 -142.579 <2e-16 \*\*\*  
## Income -1.013e+01 5.500e-02 -184.230 <2e-16 \*\*\*  
## Limit 3.410e-01 7.501e-03 45.466 <2e-16 \*\*\*  
## Rating -2.057e-01 1.109e-01 -1.854 0.0647 .   
## Cards 2.539e+01 9.467e-01 26.823 <2e-16 \*\*\*  
## StudentYes 5.027e+02 3.507e+00 143.346 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.95 on 304 degrees of freedom  
## Multiple R-squared: 0.9977, Adjusted R-squared: 0.9977   
## F-statistic: 2.655e+04 on 5 and 304 DF, p-value: < 2.2e-16

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.1662148, Df = 1, p = 0.6835



## Dealing with zero balance

## Start: AIC=-992.33  
## BalanceF ~ Limit + Student + Rating + Cards + Education + Income  
##   
## Df Sum of Sq RSS AIC  
## - Rating 1 0.0229 32.341 -994.05  
## - Education 1 0.0231 32.341 -994.05  
## - Cards 1 0.1048 32.423 -993.04  
## <none> 32.318 -992.33  
## - Limit 1 0.5668 32.885 -987.38  
## - Student 1 2.2785 34.597 -967.08  
## - Income 1 9.8291 42.148 -888.12  
##   
## Step: AIC=-994.05  
## BalanceF ~ Limit + Student + Cards + Education + Income  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 0.0200 32.361 -995.80  
## - Cards 1 0.0829 32.424 -995.02  
## <none> 32.341 -994.05  
## - Student 1 2.2565 34.598 -969.07  
## - Income 1 9.9392 42.281 -888.86  
## - Limit 1 31.0614 63.403 -726.78  
##   
## Step: AIC=-995.8  
## BalanceF ~ Limit + Student + Cards + Income  
##   
## Df Sum of Sq RSS AIC  
## - Cards 1 0.0872 32.448 -996.72  
## <none> 32.361 -995.80  
## - Student 1 2.2377 34.599 -971.06  
## - Income 1 9.9254 42.287 -890.80  
## - Limit 1 31.0585 63.420 -728.68  
##   
## 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

##   
## 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

## 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

## [1] 5

## Loading required package: rngtools

## Loading required package: pkgmaker

## Loading required package: registry

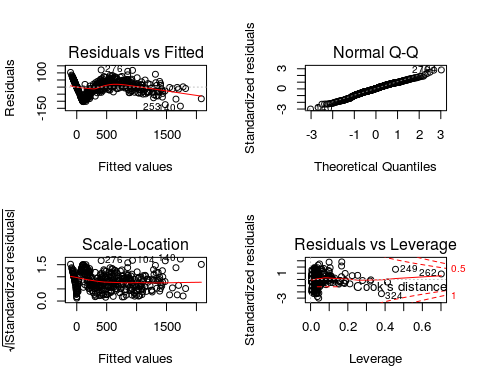
##   
## Attaching package: 'pkgmaker'

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

## [1] 979.25

##   
## Call:  
## lm(formula = Balance ~ Student + Limit + Income + log(Rating) +   
## Cards + Student:Limit + Student:Income + Limit:Income + Student:log(Rating) +   
## Limit:log(Rating) + Limit:Cards + Income:log(Rating) + Income:Cards +   
## Student:Limit:Income + Student:Limit:log(Rating) + Limit:Income:Cards,   
## data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -134.73 -26.93 1.94 33.31 121.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.774e+03 1.998e+02 8.878 < 2e-16 \*\*\*  
## StudentYes -2.150e+03 4.850e+02 -4.432 1.22e-05 \*\*\*  
## Limit -6.070e-01 6.460e-02 -9.396 < 2e-16 \*\*\*  
## Income 1.729e+01 4.307e+00 4.013 7.22e-05 \*\*\*  
## log(Rating) -3.967e+02 4.550e+01 -8.717 < 2e-16 \*\*\*  
## Cards 4.823e+00 7.308e+00 0.660 0.509669   
## StudentYes:Limit 8.563e-01 1.860e-01 4.604 5.64e-06 \*\*\*  
## StudentYes:Income -5.685e+00 1.108e+00 -5.130 4.62e-07 \*\*\*  
## Limit:Income -3.291e-04 1.107e-04 -2.974 0.003129 \*\*   
## StudentYes:log(Rating) 4.278e+02 1.140e+02 3.752 0.000203 \*\*\*  
## Limit:log(Rating) 1.491e-01 8.742e-03 17.055 < 2e-16 \*\*\*  
## Limit:Cards 4.415e-03 1.622e-03 2.722 0.006783 \*\*   
## Income:log(Rating) -4.063e+00 7.982e-01 -5.090 5.61e-07 \*\*\*  
## Income:Cards 3.992e-02 1.636e-01 0.244 0.807319   
## StudentYes:Limit:Income 8.556e-04 1.605e-04 5.329 1.69e-07 \*\*\*  
## StudentYes:Limit:log(Rating) -1.382e-01 2.596e-02 -5.323 1.74e-07 \*\*\*  
## Limit:Income:Cards -3.031e-05 1.756e-05 -1.726 0.085195 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 45.86 on 383 degrees of freedom  
## Multiple R-squared: 0.9904, Adjusted R-squared: 0.99   
## F-statistic: 2482 on 16 and 383 DF, p-value: < 2.2e-16

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 3.921739, Df = 1, p = 0.047666



# Problem 4

The Salaries data in the carData package contains information on academic salaries in 2008 and 2009 in a college in the US. A data dictionary can be found in the help file for the data. This data was collected as part of an on-going effort of the college to monitor salary differences between male and female faculty members. We have been asked to investigate the gender gap in the data, but also what other information that may be relevant to admistrators (i.e. salary growth for years of service, discipline based growth, etc). Investigate if there is a gender gap, but also provide insights on other drivers that you may see of salary in the data. Is your model suitable to make offers based on the infromation provided? Explain your reasoning. Provide insights into any other information you find of interest.

We want to investigate the gender gap in the data as well as provide any other drivers that could influence the gap, and determine whether our model of choice is suitable to make offers.

library(effects)  
library(carData)  
library(car)  
help(Salaries)  
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

dim(Salaries)

## [1] 397 6

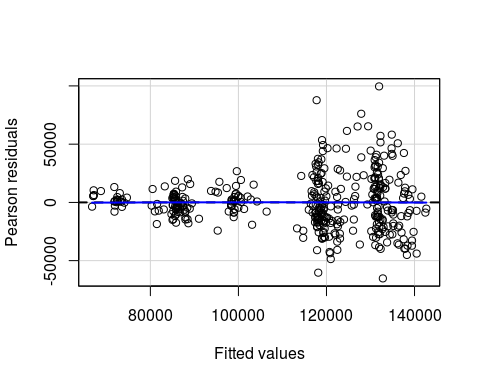
We have 397 observations with 6 variables. Rank is a factor, with 67 professors being Assistants, 64 being Associate, and 266 are full professors. Discipline is another factor: 181 teach a theoretical discipline, and 216 teach applied. yrs.since.phd has a wide range from 1 year all the way to 56. yrs.service also has a wide range, 0 years (assuming these people’s first year was the 08-09 academic calendar year) all the way to 60 years. The 0 could create an issue later on. Sex is a very skewed factor in this data set, 358 are males and only 39 are females. This could create some issues. The nine-month salary is in dollars and ranges from 57,800 - 231,545.

t.test(Salaries$salary[Salaries$sex=="Male"], Salaries$salary[Salaries$sex=="Female"], alternative="greater")

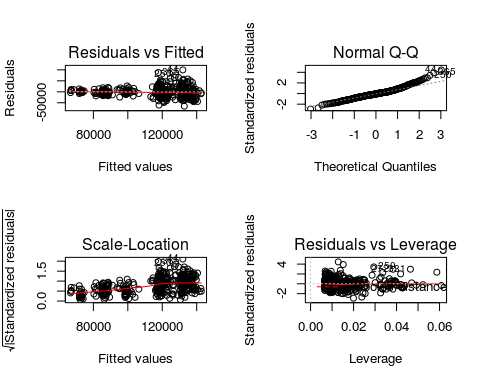
##   
## Welch Two Sample t-test  
##   
## data: Salaries$salary[Salaries$sex == "Male"] and Salaries$salary[Salaries$sex == "Female"]  
## t = 3.1615, df = 50.122, p-value = 0.001332  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 6620.263 Inf  
## sample estimates:  
## mean of x mean of y   
## 115090.4 101002.4

The t-test = null hypothesis is males makes less than or equal to what females make. So the alternative is that males make more than females. the p-value is less than 0.05. Using a hard cut off, we would reject the null hypothesis that males make more than females. the mean of x (males) is 115090.4. the mean for y (females) is 101002.4. But this is just looking at straight averages, and as we saw above, we have way more male observations than female. And we know that there are other variables that might play a role into the conclusion.

mod1\_4<-lm(salary~rank+discipline+yrs.since.phd+yrs.service+sex,data=Salaries)  
residualPlot(mod1\_4)



par(mfrow=c(2,2))  
plot(mod1\_4)



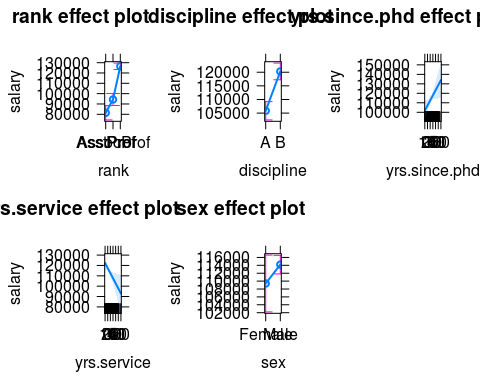
ncvTest(mod1\_4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 66.51281, Df = 1, p = 3.4763e-16

summary(mod1\_4)

##   
## 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(allEffects(mod1\_4))

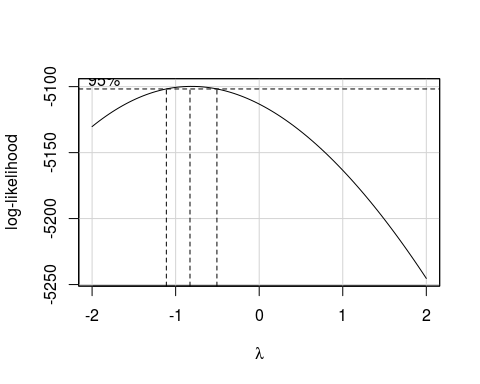


First we will fit a model using salary as our response and the rest of the variables as the predictors. In our residual plot, we see that as the salary increases, so does the variance. Unforuntately this cone-shape is a sign of non-constant variance. We see a similar pattern in the Residuals vs. Fitted plot. The normal Q-Q plot looks ok, has a few points getting away from the line towards the top. The Scale-Location plot has a positive trend showing. The Residuals vs. Leverage plot looks good, nothing is near Cook’s distance.

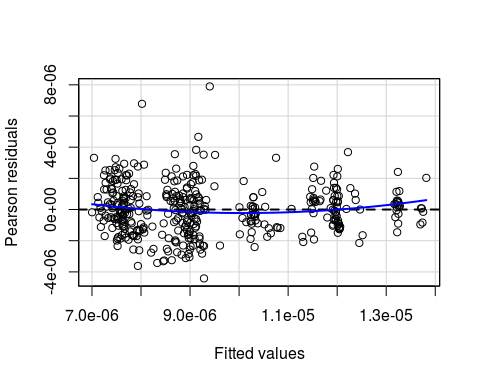
Our ncvTest p-value is close to 0. So we can reject the null hypothesis that the variance of the residuals is constant, and confirm our graphical inference. Even though we know we have NCV in this model, lets see what the summary says. From here we can see that sex is not a significant factor. But what stands out is that the more years of service, the less money they would make in this model, which does not make any sense. So we will keep trying.

The allEffects plot shows this negative correlation between yrs.service and salary, along with a high variance as years increase.

boxCox(mod1\_4)



mod4\_4<-lm(I(1/salary)~rank+discipline+yrs.since.phd+yrs.service+sex,data=Salaries)  
residualPlot(mod4\_4)



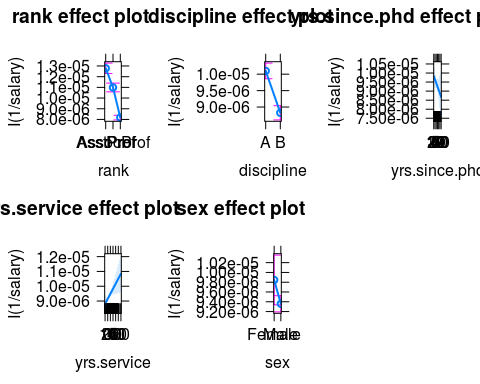
ncvTest(mod4\_4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 8.805049, Df = 1, p = 0.003004

summary(mod4\_4)

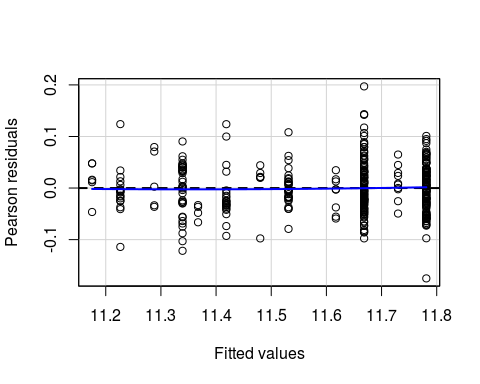
##   
## Call:  
## lm(formula = I(1/salary) ~ rank + discipline + yrs.since.phd +   
## yrs.service + sex, data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.418e-06 -1.070e-06 -1.600e-08 8.598e-07 7.899e-06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.375e-05 3.243e-07 42.392 < 2e-16 \*\*\*  
## rankAssocProf -1.795e-06 2.930e-07 -6.126 2.20e-09 \*\*\*  
## rankProf -4.610e-06 2.995e-07 -15.394 < 2e-16 \*\*\*  
## disciplineB -1.274e-06 1.656e-07 -7.695 1.17e-13 \*\*\*  
## yrs.since.phd -1.926e-08 1.703e-08 -1.131 0.2588   
## yrs.service 3.398e-08 1.498e-08 2.268 0.0239 \*   
## sexMale -4.865e-07 2.727e-07 -1.784 0.0752 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.593e-06 on 390 degrees of freedom  
## Multiple R-squared: 0.5714, Adjusted R-squared: 0.5648   
## F-statistic: 86.65 on 6 and 390 DF, p-value: < 2.2e-16

plot(allEffects(mod4\_4))

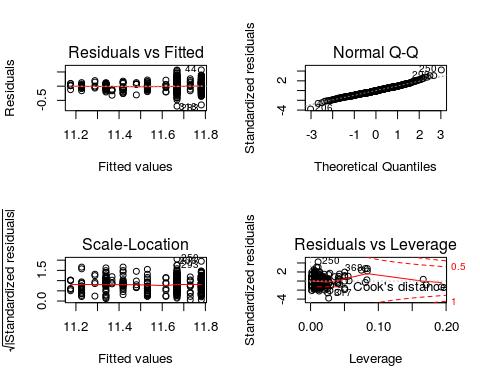


According to boxCox, lambda is close to -1, which tells us to try using the inverse of our repsonse variable. So we try that and check out residuals. We have a very slight u-shaped curve and it looks as if our cone-shape is now just flipped over, but not as bad as before. Our ncvTest p-value is again close to 0. So we can reject the null hypothesis and confirm that we still have NCV. Our summary shows us that our r-squared value is 57%, which we think we can do better than that once we correct the NCV. Our allEffects plots still don’t quite make sense. The yrs.service is now corrected (salary increases as does yrs.service) but now it’s showing that the higher the rank, the lower the salary. So we know this isn’t right.

Salaries$yrs.service.p1 <- Salaries$yrs.service + 1  
mod2\_4<- lm(log(salary) ~ rank+discipline+sex, weights=1/I(yrs.service.p1), data=Salaries)  
residualPlot(mod2\_4)



par(mfrow=c(2,2))  
plot(mod2\_4)



ncvTest(mod2\_4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.005531436, Df = 1, p = 0.94071

anova(mod2\_4)

## Analysis of Variance Table  
##   
## Response: log(salary)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## rank 2 1.89652 0.94826 431.0630 < 2.2e-16 \*\*\*  
## discipline 1 0.16865 0.16865 76.6670 < 2.2e-16 \*\*\*  
## sex 1 0.01621 0.01621 7.3703 0.006924 \*\*   
## Residuals 392 0.86233 0.00220   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(mod2\_4)

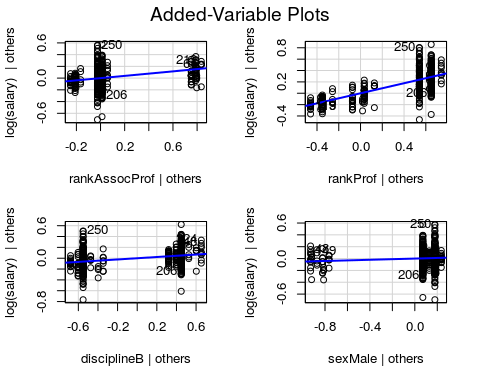
## [1] -268.1798

step(mod2\_4)

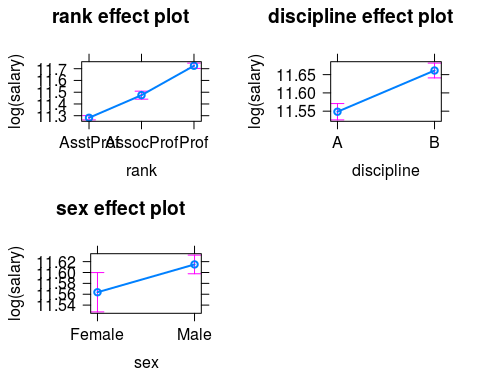
## Start: AIC=-2424.43  
## log(salary) ~ rank + discipline + sex  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.86233 -2424.4  
## - sex 1 0.01621 0.87854 -2419.0  
## - discipline 1 0.14999 1.01232 -2362.8  
## - rank 2 1.96136 2.82369 -1957.5

##   
## Call:  
## lm(formula = log(salary) ~ rank + discipline + sex, data = Salaries,   
## weights = 1/I(yrs.service.p1))  
##   
## Coefficients:  
## (Intercept) rankAssocProf rankProf disciplineB sexMale   
## 11.17524 0.19211 0.44193 0.11274 0.05123

avPlots(mod2\_4)



plot(allEffects(mod2\_4))



summary(mod2\_4)

##   
## Call:  
## lm(formula = log(salary) ~ rank + discipline + sex, data = Salaries,   
## weights = 1/I(yrs.service.p1))  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -0.175150 -0.030313 -0.003326 0.030187 0.197098   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.17524 0.01912 584.348 < 2e-16 \*\*\*  
## rankAssocProf 0.19211 0.01947 9.869 < 2e-16 \*\*\*  
## rankProf 0.44193 0.01481 29.836 < 2e-16 \*\*\*  
## disciplineB 0.11274 0.01365 8.257 2.31e-15 \*\*\*  
## sexMale 0.05123 0.01887 2.715 0.00692 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0469 on 392 degrees of freedom  
## Multiple R-squared: 0.7071, Adjusted R-squared: 0.7041   
## F-statistic: 236.5 on 4 and 392 DF, p-value: < 2.2e-16

mod3\_4<- lm(log(salary) ~ rank+discipline, weights=1/I(yrs.service.p1), data=Salaries)  
anova(mod3\_4, mod2\_4)

## Analysis of Variance Table  
##   
## Model 1: log(salary) ~ rank + discipline  
## Model 2: log(salary) ~ rank + discipline + sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 393 0.87854   
## 2 392 0.86233 1 0.016213 7.3703 0.006924 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(mod3\_4)

## [1] -262.7848

In order to make the dataset nicer, we add one to ever observation of yrs.service to get yrs.service.p1. This will allow us to do more with the variable.

The final model we ended with is using rank + discipline + sex, weighted by yrs.service.p1 to predict log(salary).

The residual plot looks pretty good. a lot of fitted values are on the same value throughout but it looks like pretty good constant variance. There is a similar pattern in the Residuals vs. Fitted plot. The normal Q-Q plot looks pretty good. The Scale-Location plot looks good, no longer has a trend. The Residuals vs. Leverage plot looks good, nothing is too close to Cook’s distance.

Our ncvTest is much better. p-value of 0.94 says we fail to reject the null hypothesis. We do not have enough evidence that this is NCV. So we can assume we have constant variance.

Anova tells us that each of these variables are significant. The higher the rank, the higher the salary. An applied discipline makes more on average than a theoretical discipline. However, this is also showing that sex matters, meaning if you are male then you make more than if you were female, which is very concerning.

Our AIC is -268, lower than any other model we tested with log(salary) as the response. Using backwards stepwise to see if we can make any quick changes to our model, it says that doing nothing will give us the lowest AIC, so based on the predictors included in this model, I still shouldn’t change anything as it is currently.

Our avPlots and allEffects plot tell similar stories. Going from Assistant professor to Associate professor is significant, and a nice increase in salary. Going from Associate to Full professor is an even better increase in salary. Teaching in Discipline B (applied) is also an increase in salary from teaching in Discipline A (theoretical). It appears as if there is an increase from going from female to male employee. The whiskers are overlapping a little so lets look to our summary to check these out.

According to our summary, our intercept is 11.18 when everything else is = 0. Since our coefficients are missing female and assitant, we know these are what’s being included in our intercept. Our coefficients are giving us valuable insights when using this model. Going from Assistant Professor to ASsociate Professor would result in a 19% increase in average salary. Going from Assistant Professor to Full Professor (which I don’t believe can happen) would be a 44% increase in average salary. Teaching an applied discipline instead of a theoretical discipline would be an 11% increase in average salary. And finally, going from female to male would result in an increase of 5% in average salary. Given all of these variables are in the model, all of these variables are significant based on p-value. The fact that sex is considered significant is an issue. Our model r-squared is accounting for 71% of all variance and our model’s p-value is significant, so we decide to go with this model.

Let’s see if we can create a better model by removing sex. In mod3\_4, all we do is remove sex, and leave everything else the same.

ANOVA allows us to compare submodels to determine which one is better. mod3\_4 is a submodel of mod2\_4. Since ‘sex’ is the only variable we removed, we are basically just comparing what the difference in sex does. Our p-value is significant, so we reject the null hypothesis. We confirm that sex is an important variable. In this instance, simpler is not better, and the model containing sex will be the better model. The F-statistic is telling us that it’s better to use variables than to do nothing. We use AIC again to compare the models since our response variables are the same. For mod2\_4, our AIC was -268, for mod3\_4, our AIC is -263. So it confirms that sex is an important variable and that mod2\_4 is the better model to use.

mod2\_4<- lm(log(salary) ~ rank+discipline+sex, weights=1/I(yrs.service.p1), data=Salaries)  
mod3\_4<- lm(log(salary) ~ rank+discipline, weights=1/I(yrs.service.p1), data=Salaries)  
  
library(doParallel)  
library(foreach)  
library(doRNG)   
  
mine<-detectCores()  
mine<-min(c(max(c(1,mine/2)),5))  
cl=makeCluster(mine)  
registerDoParallel(cl)  
registerDoRNG()  
getDoParWorkers()

## [1] 5

m1 = 0  
m2 = 0  
  
m1 = foreach(i=1:1000,.combine="+",.options.RNG=123)%dopar%{  
 set=sample(1:dim(Salaries)[1],300,replace=FALSE)  
 m2a<-lm(log(salary) ~ rank+discipline+sex, weights=1/I(yrs.service.p1), data=Salaries[set,])  
 sum(((Salaries$salary[-set]-exp(predict(m2a,newdata=Salaries[-set,])))^2))  
 }  
sum(m1)

## [1] 5.057975e+13

m2 = foreach(i=1:1000,.combine="+",.options.RNG=123)%dopar%{  
 set=sample(1:dim(Salaries)[1],300,replace=FALSE)  
 m3a<-lm(log(salary) ~ rank+discipline, weights=1/I(yrs.service.p1), data=Salaries[set,])  
 sum(((Salaries$salary[-set]-exp(predict(m3a,newdata=Salaries[-set,])))^2))  
 }  
sum(m2)

## [1] 5.088053e+13

stopCluster(cl)

We can use random splitting to test the models now to make sure the model including sex will be the better predictor. In this method, we are randomly splitting the data, training it, and then testing it however many times we deisgnate, in this case 1000. Predicting over more and more sets will smooth out the average.

We will access the doParallel library and the foreach library to use a for loop while using parallel processing to speed this up. The detectCores() function allows us to find out how many cores are on the machine. Then we divide this by 2, to ensure we are only using half of the available cores. cl will be my cluster and we will registerDoParallel(cl) so I can use the 4 cores. getDoParWorkers verifies the number of cores we are using.

Now I will set my seed and run the foreach loop. We use registerDoRNG() to ensure the seed will be passed the same way to parallel processing. This is saying we will run through this loop 1000 times, training the data on 300 of the observations, and testing it on the remainder. Then we will calculate the residual sum of squares by comparing how bad our predictions were. But to do this, we must use exp() for our repsonse variable so we are comparing correctly.

Again, random splitting confirms that we should use mod2\_4. The residual sum of squares is smaller.

And then we must always use stopCluster afterwards to make sure we clean up.