Stats101A, Spring 2023 - Take Home Final

Luke Villanueva - 206039397

06/13/23

Problem 1

```r
# get Y transforms
# boxcox
car::boxCox(m)
summary(car::powerTransform(m))
# transform is not log, and possibly no transform needed
# Inv Resp plot
car::inverseResponsePlot(m)
# data follows lambda = 0 well

# conclusion: no transform on Y

# get X transforms
# check mmps, removing categorical vars
car::mmps(update(m,.~. -Gender -Rank))
# no polynomial transforms seem to be needed

# check diagnostic plots
plot(m)

# try for interaction models
m1 <- lm(Salary~. + Gender:Begin.Salary + Gender:StartYr + Gender:Expernc + Begin.Salary:Expernc, salar
summary(m1)
car::vif(m1)

# remove Gender:StartYr, Gender:Begin.Salary
m1 <- update(m1,.~. -Gender:StartYr -Gender:Begin.Salary)

car::vif(m1)

# remove Begin.Salary:Expernc
m1 <- update(m1,.~.-Begin.Salary:Expernc)

car::vif(m1)

# remove Gender:Expernc
m1 <- update(m1, .~.-Gender:Expernc)

car::vif(m1)
```

```r
# check Y transform
car::boxCox(m1)
summary(car::powerTransform(m1))
```

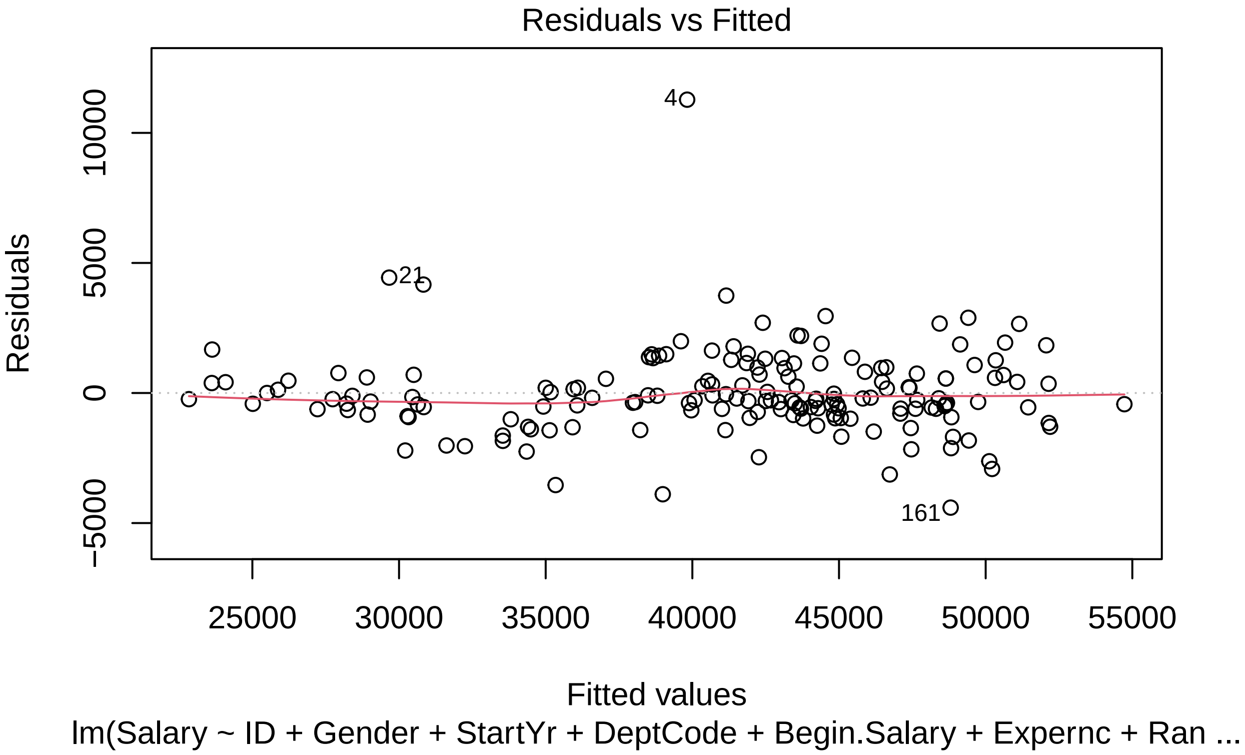
a.

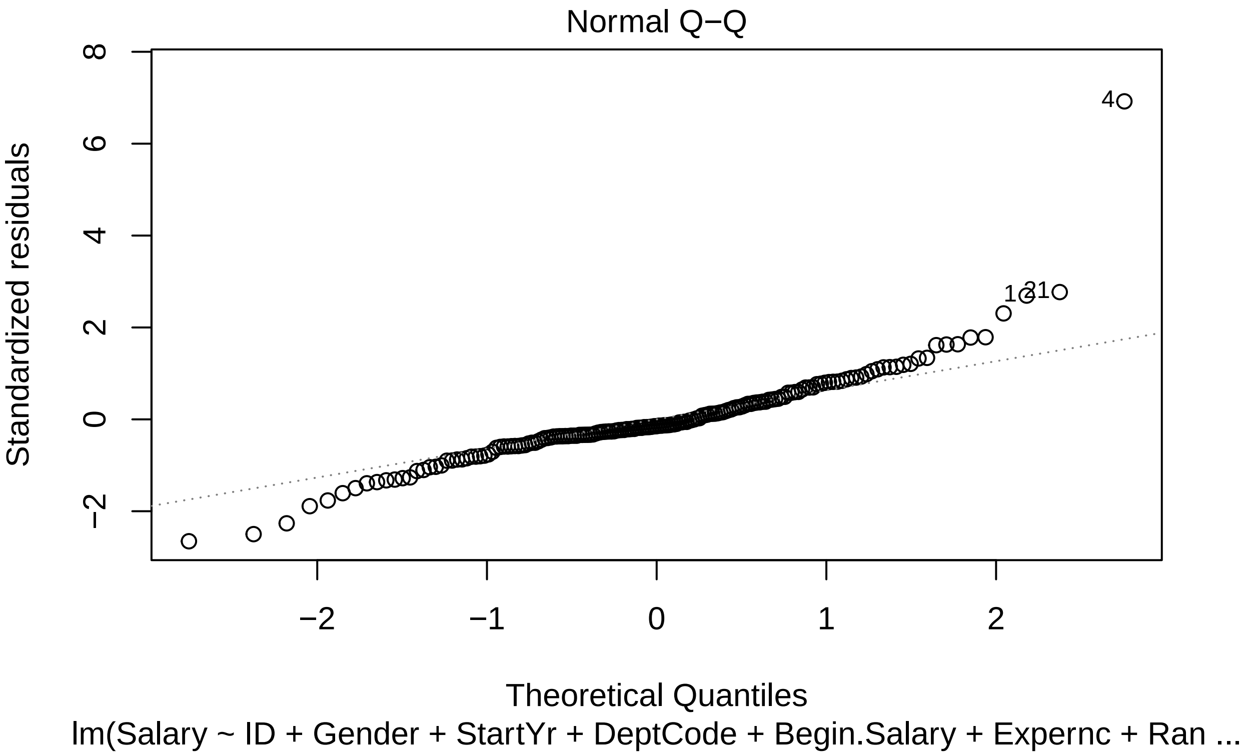
$$Y_{SalaryOfFemaleAssoProf} = 3394000 + 0.4709X_{ID} + 449.5X_{GenderMale} - 1710X_{StartYr} + 18.23X_{DeptCode}$$
$$+1.386X Begin.Salary + 52.54X_{Expernc} - 1632X_{RankAsstProf} - 2011X_{RankInstruct} + 2955X_{RankProfessr}$$

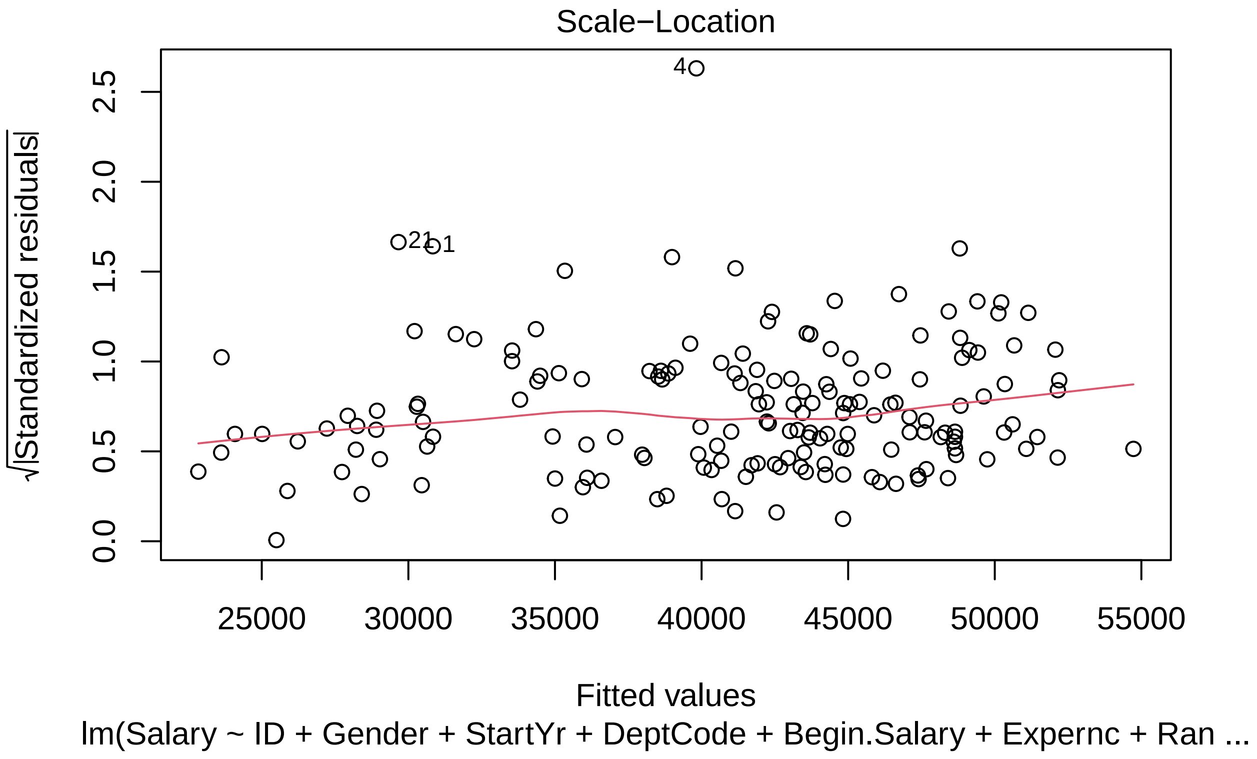
b.

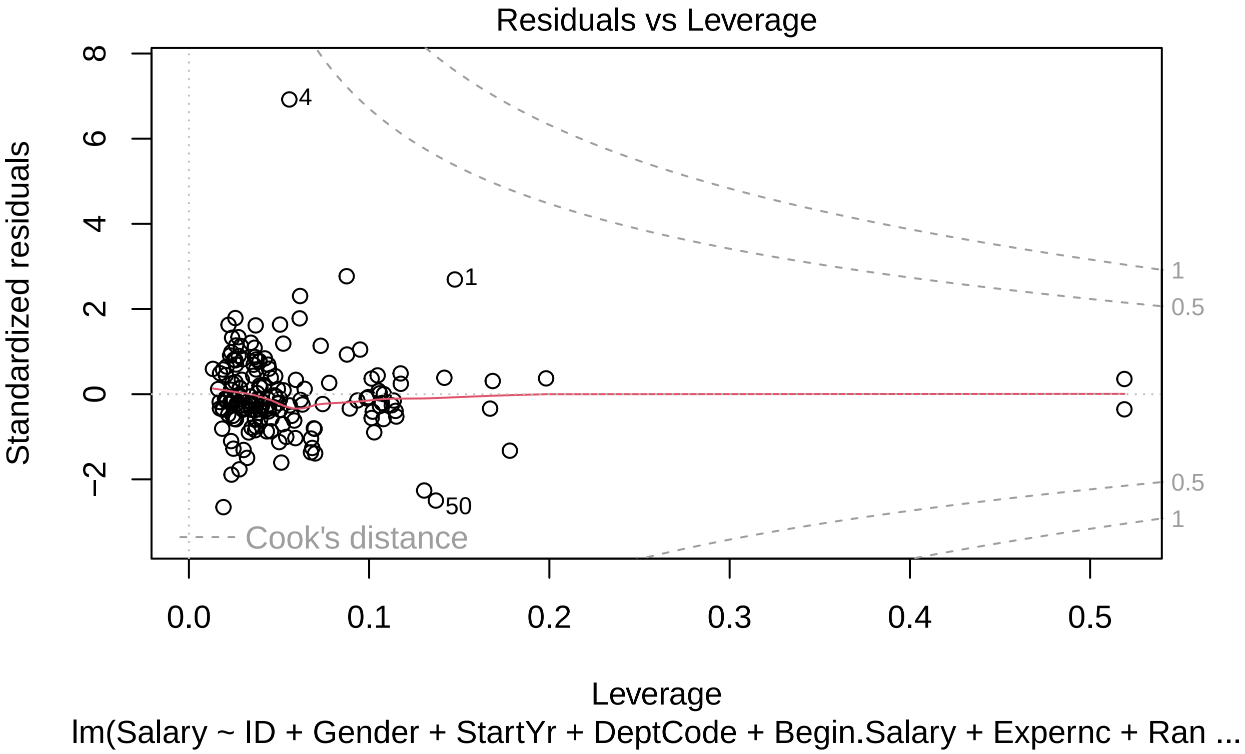
For model validity, we have to check linearity, normality, homeoscedasticity, independence.

```r
plot(m)
```




Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
4
12
21
−2
−1
0
1
2
Theoretical Quantiles
lm(Salary ~ ID + Gender + StartYr + DeptCode + Begin.Salary + Expernc + Ran ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
4
21
1
25000
30000
35000
40000
45000
50000
55000
Fitted values
lm(Salary ~ ID + Gender + StartYr + DeptCode + Begin.Salary + Expernc + Ran ...



Linearity: The residuals vs fitted plot shows that the points are generally scattered. The only reason why the plot is squished-looking is because of the outliers. But generally, the scattered points are random. There is no obvious linear trend with the data as well. So, this implies that linearity can be assumed. Variance, on the other hand, can be questioned, so the scale-location plot is needed.

Normality: The QQ plot shows a good number of points straying off of the line, meaning normality can be assumed to not be too strong. However, because there are 171 observations, the sample size is large enough to believe the t-stat tests, p-values, and predicted values. Though, prediction intervals are not to be believed.

```
nrow(salary)
```

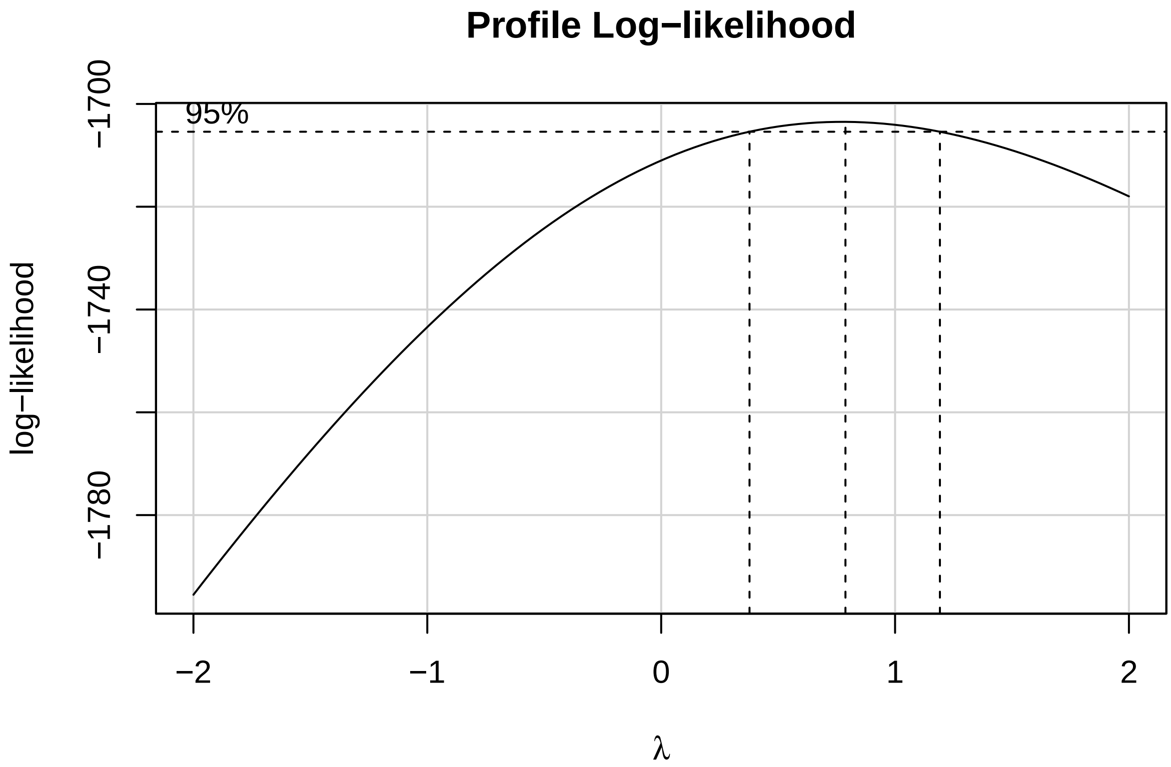
```
## [1] 171
```

Homeoscedasticity: The scale-location plot showcases a good generally random spread of points, but there is a slight linear trend upwards. It seems to be negligible, but possible inconsistent variance should be kept in mind while going forward with analysis.

Sampling Independence: The sampling is independent because each observation is a different person.

Other Models' Validity: Models with transformed variables were considered as well. First, transformations on the Y and X variables were considered. Using BoxCox and Inverse Response Plot, possible lambdas were considered.

```
car::boxCox(m)
```



Via BoxCox graph, we can see that the 95% confidence interval shows lambda being around 1, decently implying that the best transformation for the Y variable would be no transformation.

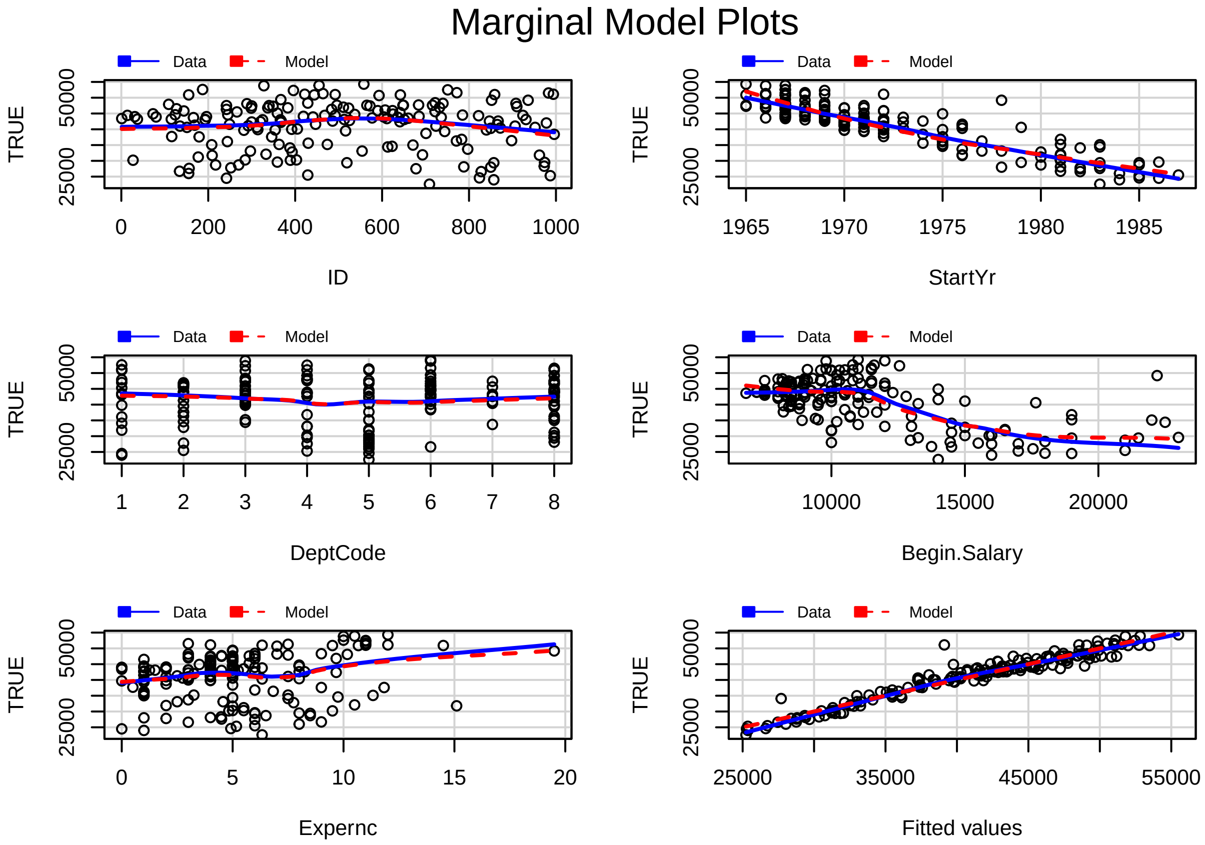
```
summary(car::powerTransform(m))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.7744           1        0.369       1.1798
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 15.04148  1 0.00010517
##
## Likelihood ratio test that no transformation is needed
##                             LRT df    pval
## LR test, lambda = (1) 1.154636  1 0.28258
```

Similarly, the hypothesis tests for response variable transformations conclude to reject the null hypothesis that the transformation uses log. In addition, it fails to reject that there even needs to be a transformation. And so, we can conclude that there is little evidence that there needs to be a transformation on Y.

Moreover, the X variables were checked as well. Since there are 0's in the Expernc column, marginal model plots are the only tests viable to use, for BoxCox cannot be used with 0 values.

```
car::mmps(update(m,.~. -Gender -Rank))
```



From the marginal model plots, it's clear that each variable individually contributes to the linearity of the model. And so, there is little evidence that there needs to be any transformation on any variable.

Finally, interaction variables were added and tested, but the variables provided too high of VIF, so collinearity was an obvious issue with interaction variables.

In conclusion, the normal multiple linear regression model was chosen.

c.

```
# leverage
# high leverage for salary
highLev <- 2 * (length(names(m$coefficients))-1
 + 1) / nrow(salary)
```

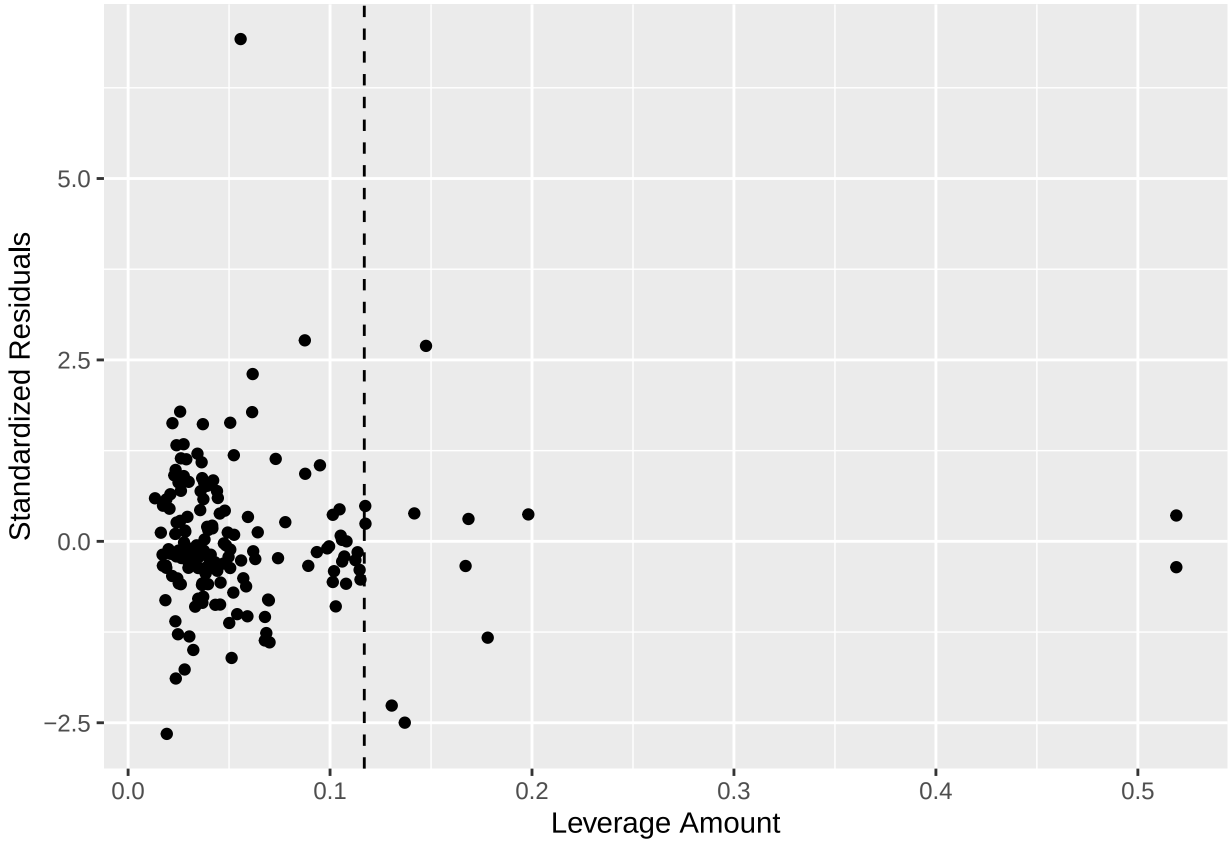
```
# plot residual vs leverage, with leverage point marked
ggplot() + geom_point(aes(hatvalues(m),rstandard(m))) + geom_vline(xintercept = highLev)

# get points with high leverage
which(hatvalues(m) > highLev)

# these points are high leverage points
```

```r
# determine which points are bad leverage points

# any observation
levSalary <- salary[hatvalues(m) > highLev,]
# make new column of index of observation
levSalary$Index <- which(hatvalues(m) > highLev)
```

```r
# which high lev points have high rstandard
which(abs(rstandard(m)[levSalary$Index]) > 2)

# Thus, observations 1, 50, 156 are bad high leverage points

# influential points

# big cooks distance 4/(n-2)
bigCook <- 4/(nrow(salary)-2)

cook <- cooks.distance(m)

cook[cook > bigCook]
```

For the found model, there needs to be analysis on leverage and influential points.

```r
ggplot() + geom_point(aes(hatvalues(m),rstandard(m))) + geom_vline(xintercept = highLev, linetype = "da
```



Given the prior graph, there are multiple points that surpass the data's calculated high leverage amount, which is about 0.1169591.

The following result is the list of observations' rows on `salary.csv` that have a high leverage:

```r
names(which(hatvalues(m) > highLev))
```

```
##  [1] "1"   "5"   "7"   "20"  "22"  "50"  "143" "152" "156" "160" "170" "171"
```

To discern if these are bad leverage points, we need to filter out which observation has a high standardized residual amount. The following result are the high leverage points with high standardized residuals:

```r
names(which(abs(rstandard(m)[levSalary$Index]) > 2))
```

```
## [1] "1"   "50"  "156"
```

And so, these are observations that must be considered because they negatively contribute to the linearity of the model.

Next, influential points must be analyzed. Influence can be measured by finding Cook's Distance for each observation. The following result is the list of observations that have a large Cook's Distance:

```r
# influential points

# big cooks distance 4/(n-2)
bigCook <- 4/(nrow(salary)-2)

cook <- cooks.distance(m)

cook[cook > bigCook]
```

```
##          1          4          6         21         22         50        156 
## 0.12552789 0.28268360 0.03494617 0.07358213 0.03818093 0.09909710 0.07690569
```

Observations that have high bad leverage and high influence are data points that severely negatively impact the linear model. These points are observations: 1, 50, and 156.

Another aspect that must be analyzed is the collinearity of the model.

```r
car::vif(m)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## ID           1.042506  1        1.021032
## Gender       1.212622  1        1.101191
## StartYr      6.749974  1        2.598071
## DeptCode     1.029708  1        1.014745
## Begin.Salary 6.563276  1        2.561889
## Experenc     1.623114  1        1.274015
## Rank         2.723650  3        1.181749
```

Based on the results, Begin.Salary and StartYr are variables that have a high GVIF. These variables should be considered to be dropped from the model because they could negatively impact the significance of other variables.

Problem 2

```r
# using algebra manipulation, bic conversion is this function
BICToAIC <- function(bic,n,p)
{bic - (log(n) - 2)*p}

# make function to use regsubsets obj and provide aic
# and bic analysis
aicbicanalysis <- function(regsubObj, data)
{

  # get summary
  sumModel <- summary(regsubObj)

  n <- nrow(data)

  # convert bic to aic
  aic <- BICToAIC(sumModel$bic,n,seq_along(sumModel$bic))

  # get bic
  bic <- sumModel$bic

  aic
  bic

  # get model with lowest aic and bic
  print(paste0("Best AIC Model: ",which(aic == min(aic))))
  print(paste0("Best BIC Model: ", which(bic == min(bic))))
  print(sumModel)

}
```

```r
# forward regression
fModel <- leaps::regsubsets(Salary~.,salary,nvmax = 9 ,method="f")
aicbicanalysis(fModel, salary)
```

```
## [1] "Best AIC Model: 6"
## [1] "Best BIC Model: 3"
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., salary, nvmax = 9, method = "f")
## 9 Variables  (and intercept)
##              Forced in Forced out
## ID               FALSE      FALSE
## GenderMale       FALSE      FALSE
## StartYr          FALSE      FALSE
## DeptCode         FALSE      FALSE
## Begin.Salary     FALSE      FALSE
## Experncc         FALSE      FALSE
## RankAsstProf     FALSE      FALSE
## RankInstruct     FALSE      FALSE
## RankProfessr     FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: forward
##           ID  GenderMale StartYr DeptCode Begin.Salary Expernc RankAsstProf
## 1  ( 1 ) " " " "        "*"     " "      " "          " "     " "         
## 2  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 3  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 4  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 5  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 6  ( 1 ) " " "*"        "*"     " "      "*"          " "     "*"         
## 7  ( 1 ) " " "*"        "*"     " "      "*"          "*"     "*"         
## 8  ( 1 ) "*" "*"        "*"     " "      "*"          "*"     "*"         
## 9  ( 1 ) "*" "*"        "*"     "*"      "*"          "*"     "*"         
##           RankInstruct RankProfessr
## 1  ( 1 ) " "          " "         
## 2  ( 1 ) " "          " "         
## 3  ( 1 ) " "          "*"         
## 4  ( 1 ) " "          "*"         
## 5  ( 1 ) "*"          "*"         
## 6  ( 1 ) "*"          "*"         
## 7  ( 1 ) "*"          "*"         
## 8  ( 1 ) "*"          "*"         
## 9  ( 1 ) "*"          "*"
```

```r
# backward regression
bModel <- leaps::regsubsets(Salary~.,salary,nvmax = 9,method = "b")
aicbicanalysis(bModel,salary)
```

```
## [1] "Best AIC Model: 6"
## [1] "Best BIC Model: 3"
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., salary, nvmax = 9, method = "b")
## 9 Variables  (and intercept)
##               Forced in Forced out
## ID                FALSE      FALSE
## GenderMale        FALSE      FALSE
## StartYr           FALSE      FALSE
## DeptCode          FALSE      FALSE
## Begin.Salary      FALSE      FALSE
## Expernc           FALSE      FALSE
## RankAsstProf      FALSE      FALSE
## RankInstruct      FALSE      FALSE
## RankProfessr      FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: backward
##           ID  GenderMale StartYr DeptCode Begin.Salary Expernc RankAsstProf
## 1  ( 1 ) " " " "        "*"     " "      " "          " "     " "         
## 2  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 3  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 4  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 5  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 6  ( 1 ) " " "*"        "*"     " "      "*"          " "     "*"         
## 7  ( 1 ) " " "*"        "*"     " "      "*"          "*"     "*"         
## 8  ( 1 ) "*" "*"        "*"     " "      "*"          "*"     "*"         
## 9  ( 1 ) "*" "*"        "*"     "*"      "*"          "*"     "*"         
##           RankInstruct RankProfessr
## 1  ( 1 ) " "          " "         
## 2  ( 1 ) " "          " "         
## 3  ( 1 ) " "          "*"         
## 4  ( 1 ) " "          "*"         
## 5  ( 1 ) "*"          "*"         
## 6  ( 1 ) "*"          "*"         
## 7  ( 1 ) "*"          "*"         
## 8  ( 1 ) "*"          "*"         
## 9  ( 1 ) "*"          "*"         
```

```r
# exhaustive regression
eModel <- leaps::regsubsets(Salary~.,salary,nvmax = 9,method = "e")
aicbicanalysis(eModel,salary)
```

```
## [1] "Best AIC Model: 6"
## [1] "Best BIC Model: 3"
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., salary, nvmax = 9, method = "e")
## 9 Variables  (and intercept)
##              Forced in Forced out
## ID               FALSE      FALSE
## GenderMale       FALSE      FALSE
## StartYr          FALSE      FALSE
## DeptCode         FALSE      FALSE
## Begin.Salary     FALSE      FALSE
## Expernc          FALSE      FALSE
## RankAsstProf     FALSE      FALSE
## RankInstruct     FALSE      FALSE
## RankProfessr     FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          ID  GenderMale StartYr DeptCode Begin.Salary Expernc RankAsstProf
## 1  ( 1 ) " " " "        "*"     " "      " "          " "     " "         
## 2  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 3  ( 1 ) " " " "        "*"     " "      "*"          " "     " "         
## 4  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 5  ( 1 ) " " " "        "*"     " "      "*"          " "     "*"         
## 6  ( 1 ) " " "*"        "*"     " "      "*"          " "     "*"         
## 7  ( 1 ) " " "*"        "*"     " "      "*"          "*"     "*"         
## 8  ( 1 ) "*" "*"        "*"     " "      "*"          "*"     "*"         
## 9  ( 1 ) "*" "*"        "*"     "*"      "*"          "*"     "*"         
##          RankInstruct RankProfessr
## 1  ( 1 ) " "          " "         
## 2  ( 1 ) " "          " "         
## 3  ( 1 ) " "          "*"         
## 4  ( 1 ) " "          "*"         
## 5  ( 1 ) "*"          "*"         
## 6  ( 1 ) "*"          "*"         
## 7  ( 1 ) "*"          "*"         
## 8  ( 1 ) "*"          "*"         
## 9  ( 1 ) "*"          "*"         
```

Based on the best subsetting algorithm and using forward, backward, and exhaustive regression, the model with the best AIC is the 6 variable model. This model uses: GenderMale, StartYr, Begin.Salary, RankAsstProf, RankInstruct, RankProfessr.

Problem 3

```
bestModel <- lm(Salary ~ Gender + StartYr + Begin.Salary + Rank, salary)
summary(bestModel)
```

```
##
## Call:
## lm(formula = Salary ~ Gender + StartYr + Begin.Salary + Rank,
##     data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4319.7  -846.7  -184.3   731.1 11467.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.437e+06  1.041e+05  33.015  < 2e-16 ***
## GenderMale    4.605e+02  2.891e+02   1.593   0.1131
## StartYr      -1.731e+03  5.296e+01 -32.693  < 2e-16 ***
## Begin.Salary  1.442e+00  7.423e-02  19.419  < 2e-16 ***
## RankAsstProf -1.504e+03  6.799e+02  -2.213   0.0283 *
## RankInstruct -2.163e+03  1.296e+03  -1.670   0.0969 .
## RankProfessr  3.097e+03  6.202e+02   4.993 1.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1673 on 164 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.9502
## F-statistic: 541.8 on 6 and 164 DF,  p-value: < 2.2e-16
```

Based on the given model from Problem 2, if a 10% significance were to be used, then the mean salary between each department is statistically significantly different. This implies that Professors have the highest mean salaries because the estimated value of RankProfessr is higher than the Intercept, and the Intercept represents the AssoProf rank. In addition, the AssoProf is higher than the rest of the ranks. This implies that Professors have the highest mean salaries. However, this statistical conclusion fails if a 5% significance were to be used.

Problem 4

Based on the model from Problem 2, using a 10% significance, the most important variables determining salary would be department level, starting year, and beginning salary because their p-values fall within the 10% significance. This implies that their effects on salary are non-zero. Gender is a considerable variable as well, but the p-value is slightly higher than the 10% significance. The effect that Gender has on salary is still debatable because the test fails to reject that Gender's slope is 0 using a 10% significance.

Problem 5

```
glimpse(salary)
```

The following are the confidence intervals of the average salaries of Professors and Assistant Professors given the the averages of being Male, starting 1972, and beginning salary of $11,289:

```
# Professr
predict(bestModel,list(Gender = "Male", StartYr = 1972, Begin.Salary = 11289, Rank = "Professr"), interv
```

```
##        fit      lwr      upr
## 1 42787.75 42419.82 43155.68
```

```
# AsstProf
predict(bestModel,list(Gender = "Male", StartYr = 1972, Begin.Salary = 11289, Rank = "AsstProf"), interv
```

```
##        fit      lwr      upr
## 1 38186.59 36978.55 39394.63
```

The estimated average Professor salary is $42,787.75, with an interval [$42,419.82 , $43,155.68]

The estimated average Assistant Professor salary is $38,186.59, with an interval [$36,978.55 , $39,394.63]