Let me predict your GPA

STAT 214 - Fall 2020 - Final Project

Nursima Donuk

12/14/2020

Loading the CSV File and Observing the Data

```r
library(readr)
survey <- read_csv("studentsurvey.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_character(),
##   Gender = col_character(),
##   Award = col_character(),
##   HigherSAT = col_character(),
##   Height = col_double(),
##   Weight = col_double(),
##   Siblings = col_double(),
##   BirthOrder = col_double(),
##   VerbalSAT = col_double(),
##   MathSAT = col_double(),
##   SAT = col_double(),
##   GPA = col_double(),
##   Piercings = col_double()
## )
```

```r
head(survey)
```

```
## # A tibble: 6 x 13
##   Year  Gender Award HigherSAT Height Weight Siblings BirthOrder VerbalSAT
##   <chr> <chr>  <chr> <chr>      <dbl>  <dbl>    <dbl>      <dbl>     <dbl>
## 1 Seni~ M      Olym~ Math          71    180        4          4       540
## 2 Soph~ F      Acad~ Math          66    120        2          2       520
## 3 Firs~ M      Nobel Math          72    208        2          1       550
## 4 Juni~ M      Nobel Math          63    110        1          1       490
## 5 Soph~ F      Nobel Verbal        65    150        1          1       720
## 6 Soph~ F      Nobel Verbal        65    114        2          2       600
## # ... with 4 more variables: MathSAT <dbl>, SAT <dbl>, GPA <dbl>,
## #   Piercings <dbl>
```

```r
str(survey)
```

```
## tibble [335 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Year      : chr [1:335] "Senior" "Sophomore" "FirstYear" "Junior" ...
##  $ Gender    : chr [1:335] "M" "F" "M" "M" ...
##  $ Award     : chr [1:335] "Olympic" "Academy" "Nobel" "Nobel" ...
```

```
##  $ HigherSAT : chr [1:335] "Math" "Math" "Math" "Math" ...
##  $ Height    : num [1:335] 71 66 72 63 65 65 66 74 61 60 ...
##  $ Weight    : num [1:335] 180 120 208 110 150 114 128 235 138 115 ...
##  $ Siblings  : num [1:335] 4 2 2 1 1 2 1 1 2 7 ...
##  $ BirthOrder: num [1:335] 4 2 1 1 1 2 1 1 2 8 ...
##  $ VerbalSAT : num [1:335] 540 520 550 490 720 600 640 660 550 670 ...
##  $ MathSAT   : num [1:335] 670 630 560 630 450 550 680 710 550 700 ...
##  $ SAT       : num [1:335] 1210 1150 1110 1120 1170 1150 1320 1370 1100 1370 ...
##  $ GPA       : num [1:335] 3.13 2.5 2.55 3.1 2.7 3.2 2.77 3.3 2.8 3.7 ...
##  $ Piercings : num [1:335] 0 3 0 0 6 4 8 0 7 2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Year = col_character(),
##   ..   Gender = col_character(),
##   ..   Award = col_character(),
##   ..   HigherSAT = col_character(),
##   ..   Height = col_double(),
##   ..   Weight = col_double(),
##   ..   Siblings = col_double(),
##   ..   BirthOrder = col_double(),
##   ..   VerbalSAT = col_double(),
##   ..   MathSAT = col_double(),
##   ..   SAT = col_double(),
##   ..   GPA = col_double(),
##   ..   Piercings = col_double()
##   .. )
```

We can see that our data has many attributes, we will select a portion of these to build a model that will predict the students `GPA`. The first column we see is the `Year` the student is in, this is a qualitative variable. Then we see `Gender`, female or male. Another qualitative variable is `Award`, the students were asked what type of award would they prefer to win. The next qualitative variable indicates whether the student performed better in the math or verbal section of the SAT. The next two quantitative variables indicate the `Height` and `Weight` of the students in inches and pounds. Next quantitative variables are the number of siblings the student has, followed by the birth order of the student (first-born, second-born etc). Then we have the verbal SAT score, math SAT score, followed by the total SAT score of the student. Finally we have the `GPA` in a 4.0 scale and the number of body `Piercings` the student has.

At the end of this project I would like to test my model to see if it will predict my GPA accurately

Handling Missing Data

```
mean(is.na(survey))
```

```
## [1] 0.001607348
```

We see that there is a very small portion of data missing.

```
survey[!complete.cases(survey),]
```

```
## # A tibble: 7 x 13
##   Year  Gender Award HigherSAT Height Weight Siblings BirthOrder VerbalSAT
##   <chr> <chr>  <chr> <chr>      <dbl>  <dbl>    <dbl>      <dbl>     <dbl>
## 1 Soph~ M      Nobel Math          NA    173        1          1       580
## 2 Soph~ M      Nobel <NA>          72    260        2          3       550
## 3 Soph~ F      Nobel Math          67    140        0         NA       517
```

```
## 4 Juni~ M      Olym~ <NA>         71    192       1         1        640
## 5 Soph~ F      Olym~ <NA>         65    155       1         1        600
## 6 Juni~ F      Olym~ <NA>         67    150       1         1        560
## 7 Soph~ F      Nobel Verbal       NA    110       3         3        800
## # ... with 4 more variables: MathSAT <dbl>, SAT <dbl>, GPA <dbl>,
## #   Piercings <dbl>
```

```r
SSurvey <- na.omit(survey)
mean(is.na(SSurvey))
```

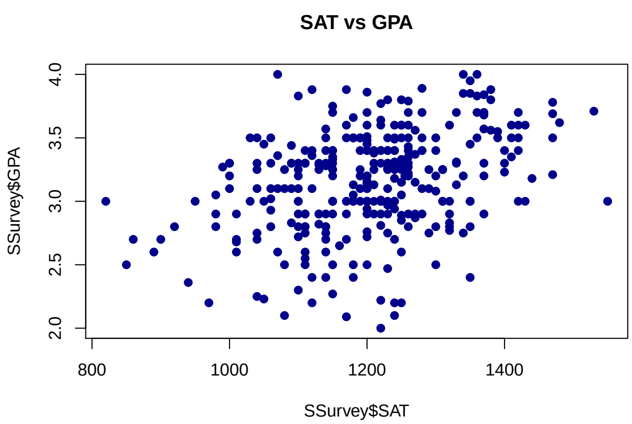
```
## [1] 0
```

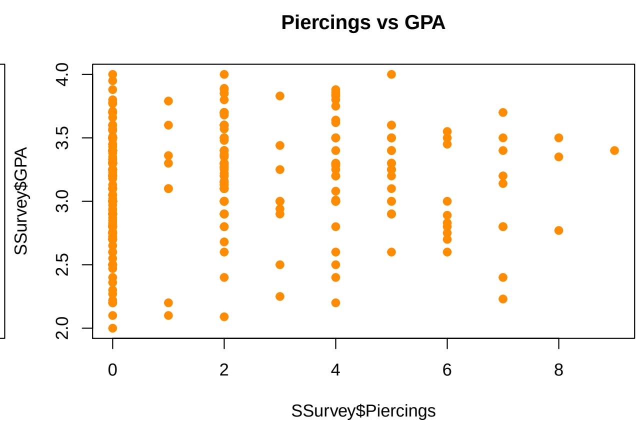
We removed the missing data.

Plots

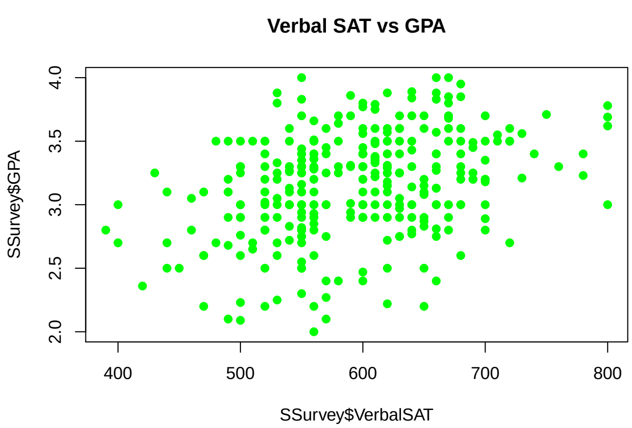
Plotting the different relationships of the independent variables with the dependent variable will give us a sense of what our final model may look like. We will be more familiar with our data as the visualizations are often insightful.

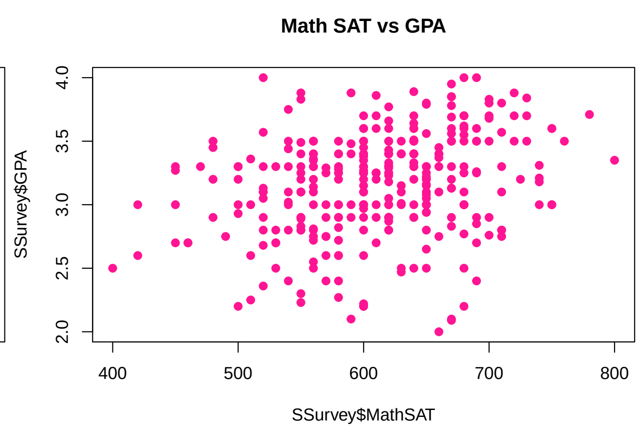
```r
plot(SSurvey$SAT, SSurvey$GPA, main="SAT vs GPA", col = "darkblue", pch=19)
plot(SSurvey$Piercings, SSurvey$GPA, main="Piercings vs GPA", col = "darkorange", pch=19)
```



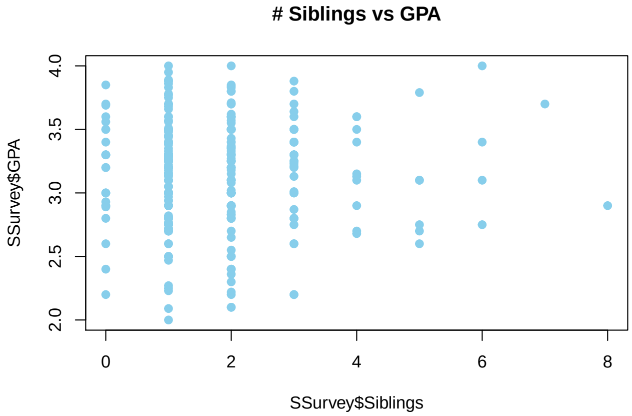



```r
plot(SSurvey$VerbalSAT, SSurvey$GPA, main="Verbal SAT vs GPA", col = "green", pch=19)
plot(SSurvey$MathSAT, SSurvey$GPA, col = "deeppink", main = "Math SAT vs GPA", pch=19)
```

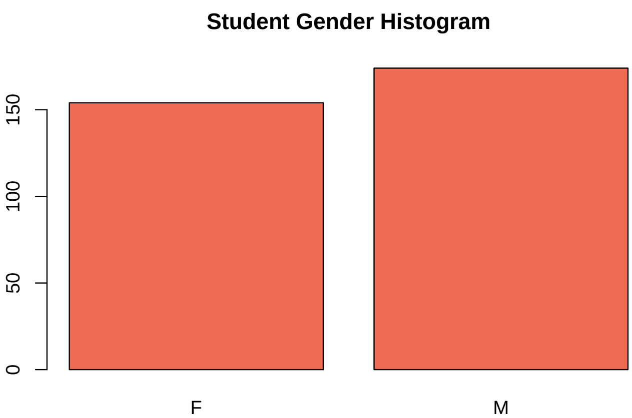


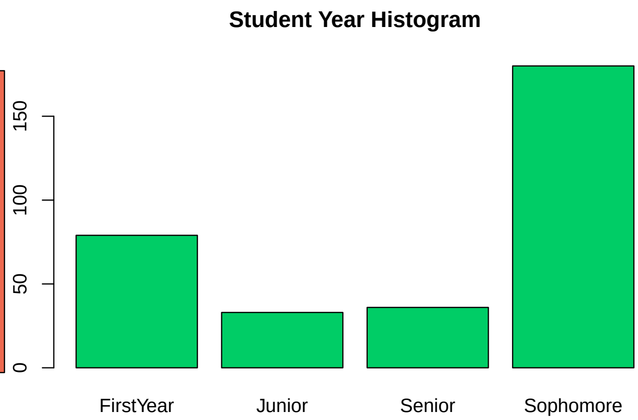



```r
plot(SSurvey$Siblings, SSurvey$GPA, main="# Siblings vs GPA", col = "skyblue", pch=19)
plot(SSurvey$BirthOrder, SSurvey$GPA, col = "darkmagenta", main = "Birth Order vs GPA", pch=19)
```

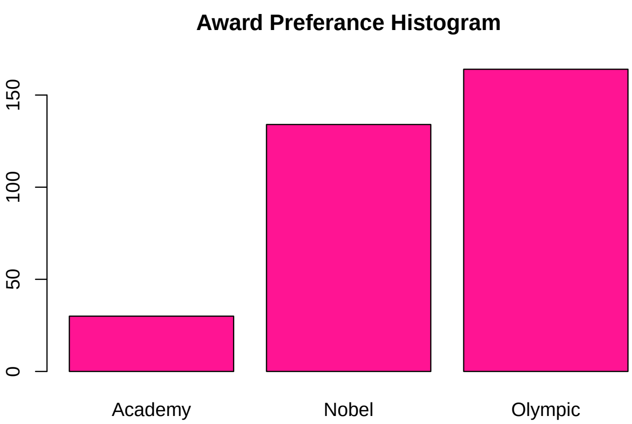


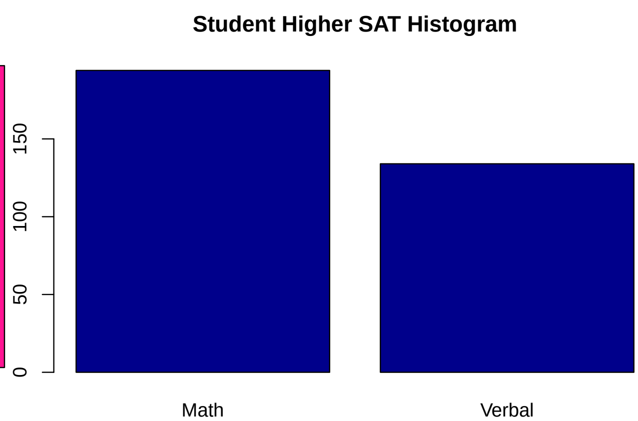

```r
plot(as.factor(SSurvey$Gender), col = "coral2", main = "Student Gender Histogram")
plot(as.factor(SSurvey$Year), col = "springgreen3", main = "Student Year Histogram")
```






```r
plot(as.factor(SSurvey$Award), col = "deeppink", main = "Award Preferance Histogram")
plot(as.factor(SSurvey$HigherSAT), col = "darkblue", main = "Student Higher SAT Histogram")
```

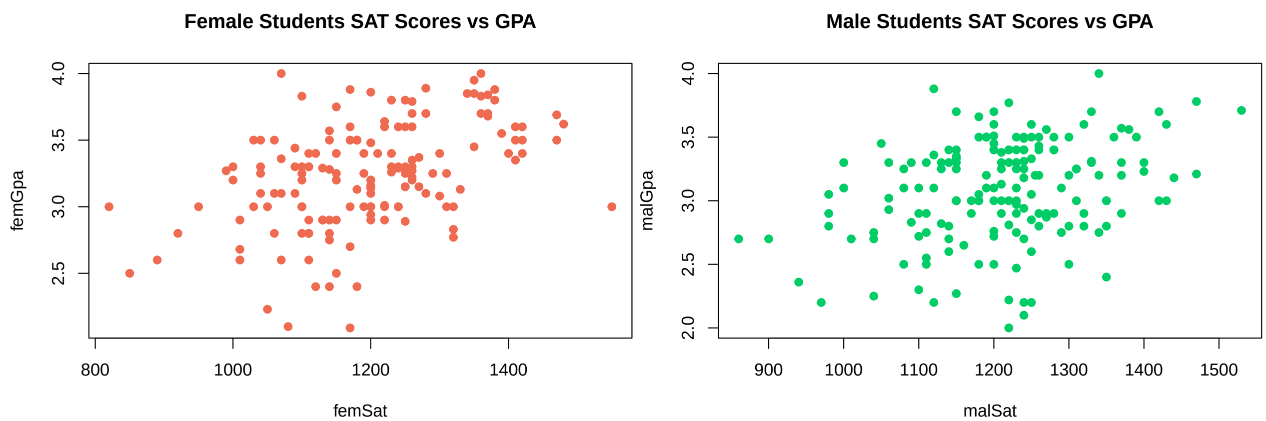





```r
i <- 1
femSat <- c()
femGpa <- c()
malSat <- c()
malGpa <- c()
while(i <= length(SSurvey$Year)) {
  if(SSurvey$Gender[i] == 'F') {
    femSat <- c(femSat, SSurvey$SAT[i])
```

```r
    femGpa <- c(femGpa, SSurvey$GPA[i])
  }
  else {
    malSat <- c(malSat, SSurvey$SAT[i])
    malGpa <- c(malGpa, SSurvey$GPA[i])
  }
  i <- i+1
}
```

```r
plot(femSat, femGpa, col = "coral2", main = "Female Students SAT Scores vs GPA", pch = 19)
plot(malSat, malGpa, col = "springgreen3", main = "Male Students SAT Scores vs GPA", pch = 19)
```



Building a Correlation Matrix

Correlation matrix must contain only quantitative variables.

```r
QuanData <- data.frame(Height = SSurvey$Height,
                       Weight = SSurvey$Weight,
                       Siblings = SSurvey$Siblings,
                       BirthOrder = SSurvey$BirthOrder,
                       VerbalSAT = SSurvey$VerbalSAT,
                       MathSAT = SSurvey$MathSAT,
                       SAT = SSurvey$SAT,
                       Piercings = SSurvey$Piercings)
res <- cor(QuanData)
round(res, 2)
```

```
##            Height Weight Siblings BirthOrder VerbalSAT MathSAT   SAT Piercings
## Height       1.00   0.63     0.04      -0.09      0.06    0.05  0.06     -0.54
## Weight       0.63   1.00     0.05      -0.05     -0.06   -0.01 -0.04     -0.48
## Siblings     0.04   0.05     1.00       0.73     -0.03    0.02 -0.01     -0.07
## BirthOrder  -0.09  -0.05     0.73       1.00      0.00    0.01  0.01     -0.01
## VerbalSAT    0.06  -0.06    -0.03       0.00      1.00    0.45  0.86     -0.01
## MathSAT      0.05  -0.01     0.02       0.01      0.45    1.00  0.84     -0.17
## SAT          0.06  -0.04    -0.01       0.01      0.86    0.84  1.00     -0.10
## Piercings   -0.54  -0.48    -0.07      -0.01     -0.01   -0.17 -0.10      1.00
```

We can observe that the `SAT` score has a high correlation between the `MathSAT` and `VerbalSAT`. Since the `SAT` variable is simply the sum of `MathSAT` and `VerbalSAT`, we can remove those from our model.

Variable Selection - Stepwise Regression

Now we can perform a stepwise regression model to decide which independent variables will be the best predictors of the `GPA`.

```r
# Install development version from GitHub
# install.packages("devtools")
# devtools::install_github("rsquaredacademy/olsrr")
library(olsrr)
library(tidyverse)
```

```r
#The plot method shows the panel of fit criteria for best subset regression methods.
model<- lm(GPA ~ Year + Gender + Award + HigherSAT + Height + Weight + Siblings + BirthOrder + SAT + Pi
k <-ols_step_both_p(model, details = T)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. Year
## 2. Gender
## 3. Award
## 4. HigherSAT
## 5. Height
## 6. Weight
## 7. Siblings
## 8. BirthOrder
## 9. SAT
## 10. Piercings
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - SAT added
##
##                         Model Summary
## --------------------------------------------------------------
## R                        0.362       RMSE              0.374
## R-Squared                0.131       Coef. Var        11.841
## Adj. R-Squared           0.128       MSE               0.140
## Pred R-Squared           0.121       MAE               0.297
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression      6.854          1          6.854    49.096    0.0000
```

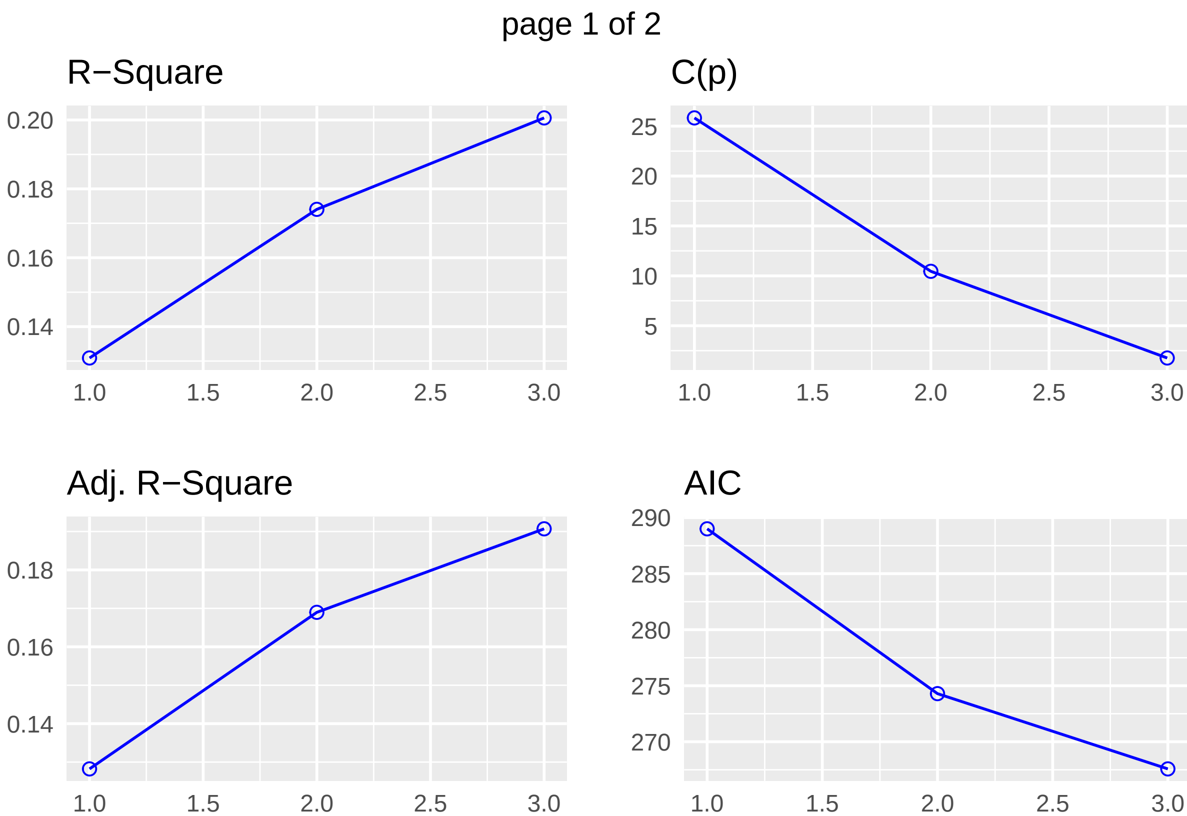
```
## Residual       45.511        326          0.140
## Total          52.365        327
## --------------------------------------------------------------------
##
##                               Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta      t       Sig     lower    upper
## ----------------------------------------------------------------------------------------
## (Intercept)    1.711         0.207                 8.263    0.000    1.304    2.119
##         SAT    0.001         0.000        0.362    7.007    0.000    0.001    0.002
## ----------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - Gender added
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.417       RMSE               0.365
## R-Squared               0.174       Coef. Var         11.561
## Adj. R-Squared          0.169       MSE                0.133
## Pred R-Squared          0.160       MAE                0.290
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                              ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     9.114          2          4.557    34.243    0.0000
## Residual      43.250        325          0.133
## Total         52.365        327
## --------------------------------------------------------------------
##
##                                 Parameter Estimates
## ---------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ---------------------------------------------------------------------------------------------
## (Intercept)     1.742         0.202                  8.611    0.000     1.344     2.141
##         SAT     0.001         0.000        0.376     7.444    0.000     0.001     0.002
##     GenderM    -0.167         0.040       -0.208    -4.121    0.000    -0.246    -0.087
## ---------------------------------------------------------------------------------------------
##
##
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.417       RMSE               0.365
## R-Squared               0.174       Coef. Var         11.561
```

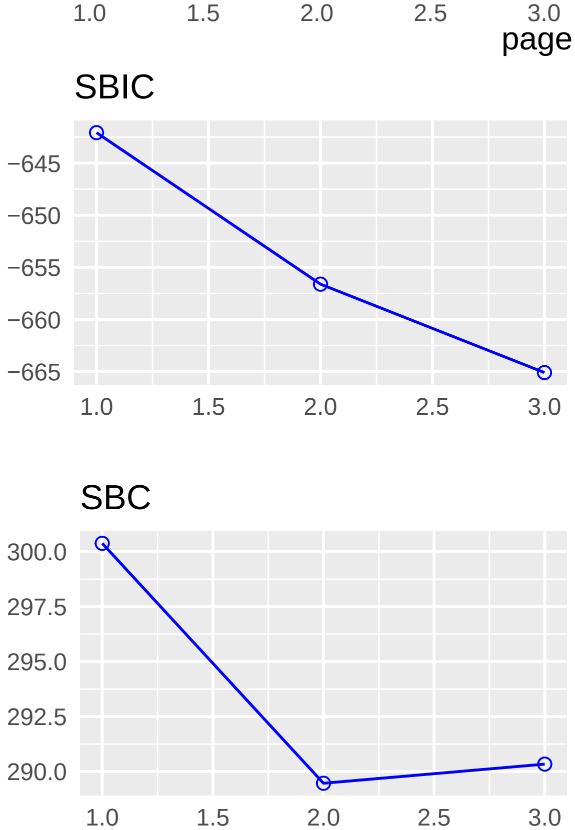
```
## Adj. R-Squared          0.169       MSE                0.133
## Pred R-Squared          0.160       MAE                0.290
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     9.114          2          4.557    34.243    0.0000
## Residual      43.250        325          0.133
## Total         52.365        327
## --------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)     1.742         0.202                 8.611    0.000      1.344     2.141
##         SAT     0.001         0.000        0.376    7.444    0.000      0.001     0.002
##     GenderM    -0.167         0.040       -0.208   -4.121    0.000     -0.246    -0.087
## ----------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 3
##
## - Award added
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.448       RMSE               0.360
## R-Squared               0.201       Coef. Var         11.409
## Adj. R-Squared          0.191       MSE                0.130
## Pred R-Squared          0.176       MAE                0.286
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression    10.504          4          2.626    20.263    0.0000
## Residual      41.860        323          0.130
## Total         52.365        327
## --------------------------------------------------------------------
##
##                                  Parameter Estimates
```

```
## ----------------------------------------------------------------------------------------
##        model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
##  (Intercept)     1.892         0.212                  8.930    0.000     1.475     2.309
##          SAT     0.001         0.000        0.336     6.542    0.000     0.001     0.001
##      GenderM    -0.147         0.041       -0.183    -3.608    0.000    -0.226    -0.067
##   AwardNobel     0.078         0.073        0.096     1.066    0.287    -0.066     0.222
## AwardOlympic    -0.065         0.072       -0.081    -0.894    0.372    -0.207     0.078
## ----------------------------------------------------------------------------------------
##
##
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.448       RMSE              0.360
## R-Squared               0.201       Coef. Var        11.409
## Adj. R-Squared          0.191       MSE               0.130
## Pred R-Squared          0.176       MAE               0.286
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##               Sum of
##              Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression    10.504         4          2.626    20.263    0.0000
## Residual      41.860       323          0.130
## Total         52.365       327
## -------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------
##        model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
##  (Intercept)     1.892         0.212                  8.930    0.000     1.475     2.309
##          SAT     0.001         0.000        0.336     6.542    0.000     0.001     0.001
##      GenderM    -0.147         0.041       -0.183    -3.608    0.000    -0.226    -0.067
##   AwardNobel     0.078         0.073        0.096     1.066    0.287    -0.066     0.222
## AwardOlympic    -0.065         0.072       -0.081    -0.894    0.372    -0.207     0.078
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
```

```
## R                          0.448       RMSE               0.360
## R-Squared                  0.201       Coef. Var         11.409
## Adj. R-Squared             0.191       MSE                0.130
## Pred R-Squared             0.176       MAE                0.286
## -----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -----------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square      F         Sig.
## -----------------------------------------------------------------------
## Regression      10.504           4          2.626    20.263    0.0000
## Residual        41.860         323          0.130
## Total           52.365         327
## -----------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##        model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## -----------------------------------------------------------------------------------------
##  (Intercept)     1.892         0.212                 8.930    0.000     1.475     2.309
##          SAT     0.001         0.000        0.336    6.542    0.000     0.001     0.001
##      GenderM    -0.147         0.041       -0.183   -3.608    0.000    -0.226    -0.067
##   AwardNobel     0.078         0.073        0.096    1.066    0.287    -0.066     0.222
## AwardOlympic    -0.065         0.072       -0.081   -0.894    0.372    -0.207     0.078
## -----------------------------------------------------------------------------------------
```

```r
plot(k)
```





Our stepwise model picked 3 variables to be the best predictors of GPA:

- `SAT`

- Gender
- Award

Time to Build Some Models

The Complete Second Order Model

```
model1 <- lm(GPA ~ SAT + I(SAT^2) + Gender + Award + Gender*Award + SAT*Gender + SAT*Award + SAT*Gender
summary(model1)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + I(SAT^2) + Gender + Award + Gender *
##     Award + SAT * Gender + SAT * Award + SAT * Gender * Award +
##     I(SAT^2) * Gender + I(SAT^2) * Award + I(SAT^2) * Gender *
##     Award, data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18094 -0.21395  0.03033  0.26065  0.97229
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.138e+00  3.767e+00   1.895   0.0591 .
## SAT                          -8.794e-03  6.516e-03  -1.350   0.1781
## I(SAT^2)                      4.518e-06  2.792e-06   1.618   0.1066
## GenderM                      -4.752e+00  7.245e+00  -0.656   0.5124
## AwardNobel                   -6.914e+00  4.826e+00  -1.433   0.1530
## AwardOlympic                 -5.981e+00  5.116e+00  -1.169   0.2433
## GenderM:AwardNobel            8.694e+00  9.627e+00   0.903   0.3672
## GenderM:AwardOlympic          4.013e+00  8.429e+00   0.476   0.6344
## SAT:GenderM                   8.991e-03  1.188e-02   0.757   0.4497
## SAT:AwardNobel                1.283e-02  8.149e-03   1.574   0.1165
## SAT:AwardOlympic              1.090e-02  8.899e-03   1.225   0.2216
## I(SAT^2):GenderM             -4.123e-06  4.850e-06  -0.850   0.3959
## I(SAT^2):AwardNobel          -5.727e-06  3.419e-06  -1.675   0.0949 .
## I(SAT^2):AwardOlympic        -4.853e-06  3.845e-06  -1.262   0.2079
## SAT:GenderM:AwardNobel       -1.555e-02  1.564e-02  -0.994   0.3210
## SAT:GenderM:AwardOlympic     -7.569e-03  1.403e-02  -0.539   0.5900
## I(SAT^2):GenderM:AwardNobel   6.740e-06  6.331e-06   1.065   0.2879
## I(SAT^2):GenderM:AwardOlympic 3.320e-06  5.826e-06   0.570   0.5692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3626 on 310 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.1787
## F-statistic: 5.186 on 17 and 310 DF,  p-value: 4.857e-10
```

Taking Out Quadratic Terms

```
model2 <- lm(GPA ~ SAT + Gender + Award + Gender*Award + SAT*Gender + SAT*Award + SAT*Gender*Award, data
summary(model2)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Gender + Award + Gender * Award + SAT *
```

```
##     Gender + SAT * Award + SAT * Gender * Award, data = SSurvey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1924 -0.2196  0.0194  0.2677  0.9745
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.132e+00  6.444e-01   1.757  0.07984 .
## SAT                       1.712e-03  5.524e-04   3.099  0.00212 **
## GenderM                   6.469e-01  1.195e+00   0.541  0.58873
## AwardNobel                9.223e-01  7.659e-01   1.204  0.22942
## AwardOlympic              4.593e-01  8.107e-01   0.567  0.57144
## GenderM:AwardNobel       -7.279e-01  1.398e+00  -0.521  0.60298
## GenderM:AwardOlympic     -2.591e-01  1.346e+00  -0.193  0.84745
## SAT:GenderM              -5.305e-04  9.766e-04  -0.543  0.58734
## SAT:AwardNobel           -6.717e-04  6.454e-04  -1.041  0.29880
## SAT:AwardOlympic         -3.717e-04  6.944e-04  -0.535  0.59283
## SAT:GenderM:AwardNobel    4.893e-04  1.137e-03   0.430  0.66734
## SAT:GenderM:AwardOlympic  3.891e-05  1.109e-03   0.035  0.97203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3612 on 316 degrees of freedom
## Multiple R-squared:  0.2125, Adjusted R-squared:  0.1851
## F-statistic: 7.752 on 11 and 316 DF,  p-value: 6.852e-12
```

Performing ANOVA test to see if the quadratic terms are useful.

```r
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + I(SAT^2) + Gender + Award + Gender * Award + SAT *
##     Gender + SAT * Award + SAT * Gender * Award + I(SAT^2) *
##     Gender + I(SAT^2) * Award + I(SAT^2) * Gender * Award
## Model 2: GPA ~ SAT + Gender + Award + Gender * Award + SAT * Gender +
##     SAT * Award + SAT * Gender * Award
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    310 40.769
## 2    316 41.237 -6  -0.46748 0.5924 0.7364
```

The high p-value suggests that the quadratic terms do not make the complete second order model significantly better than the reduced one. Therefore we proceed with model 2.

Taking Out QLxQL Interactions

```r
model3 <- lm(GPA ~ SAT + Gender + Award + SAT*Gender + SAT*Award, data = SSurvey)
summary(model3)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Gender + Award + SAT * Gender + SAT *
##     Award, data = SSurvey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.1819 -0.2112  0.0137  0.2655  0.9847
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.1405723  0.5402856   2.111 0.035543 *
## SAT                0.0017497  0.0004531   3.862 0.000136 ***
## GenderM            0.1242680  0.4212553   0.295 0.768189
## AwardNobel         0.8464236  0.6248870   1.355 0.176525
## AwardOlympic       0.6476349  0.6154875   1.052 0.293488
## SAT:GenderM       -0.0002269  0.0003480  -0.652 0.514934
## SAT:AwardNobel    -0.0006453  0.0005160  -1.251 0.211999
## SAT:AwardOlympic  -0.0006029  0.0005139  -1.173 0.241585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3604 on 320 degrees of freedom
## Multiple R-squared:  0.2064, Adjusted R-squared:  0.1891
## F-statistic: 11.89 on 7 and 320 DF,  p-value: 1.76e-13
```

Performing ANOVA test to see if the qualitative-qualitative interaction terms are useful.

```r
anova(model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Gender + Award + Gender * Award + SAT * Gender +
##     SAT * Award + SAT * Gender * Award
## Model 2: GPA ~ SAT + Gender + Award + SAT * Gender + SAT * Award
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    316 41.237
## 2    320 41.554 -4   -0.3172 0.6077 0.6574
```

The large p-value suggests that the qualitative-qualitative interaction terms do not make the model significantly better. Therefore we choose the model with less terms, which is model 3.

Taking Out QNxQL Interactions

```r
model4 <- lm(GPA ~ SAT + Gender + Award, data = SSurvey)
summary(model4)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Gender + Award, data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18260 -0.21412  0.03081  0.25894  0.98061
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.8919555  0.2118622   8.930  < 2e-16 ***
## SAT            0.0011140  0.0001703   6.542 2.37e-10 ***
## GenderM       -0.1465098  0.0406068  -3.608 0.000357 ***
## AwardNobel     0.0781172  0.0732680   1.066 0.287137
## AwardOlympic  -0.0645140  0.0721969  -0.894 0.372210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.36 on 323 degrees of freedom
## Multiple R-squared:  0.2006, Adjusted R-squared:  0.1907
## F-statistic: 20.26 on 4 and 323 DF,  p-value: 6.613e-15
```

Performing ANOVA test to see if the quantitative-qualitative interaction terms are useful.

```
anova(model3, model4)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Gender + Award + SAT * Gender + SAT * Award
## Model 2: GPA ~ SAT + Gender + Award
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    320 41.554
## 2    323 41.860 -3  -0.30642 0.7866 0.5021
```

The large p-value suggests that the quantitative-qualitative interaction terms do not make the model significantly better. Therefore we choose the model with less terms, which is model 4.

Second Attenmp to do Stepwise Regression

```
library(MASS)
```

```
# Fit the full model
full.model <- lm(GPA ~ Year + Gender + Award + Height + Weight + Siblings + SAT + Piercings, data = SSu
summary(full.model)
```

```
##
## Call:
## lm(formula = GPA ~ Year + Gender + Award + Height + Weight +
##     Siblings + SAT + Piercings, data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14305 -0.22475  0.02751  0.25336  0.98942
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.8950956  0.4762769   3.979 8.59e-05 ***
## YearJunior     0.0537131  0.0764157   0.703   0.4826
## YearSenior     0.0931782  0.0740808   1.258   0.2094
## YearSophomore  0.0316394  0.0495161   0.639   0.5233
## GenderM       -0.1830918  0.0710422  -2.577   0.0104 *
## AwardNobel     0.0747016  0.0738866   1.011   0.3128
## AwardOlympic  -0.0617234  0.0729136  -0.847   0.3979
## Height         0.0043641  0.0069810   0.625   0.5323
## Weight        -0.0014211  0.0009009  -1.577   0.1157
## Siblings       0.0061534  0.0167265   0.368   0.7132
## SAT            0.0010596  0.0001736   6.104 3.03e-09 ***
## Piercings     -0.0201167  0.0141271  -1.424   0.1554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3604 on 316 degrees of freedom
## Multiple R-squared:  0.2162, Adjusted R-squared:  0.1889
```

```
## F-statistic: 7.923 on 11 and 316 DF,  p-value: 3.529e-12
```

```r
# Stepwise regression model
step.model <- stepAIC(full.model, direction = "both",
                      trace = T)
```

```
## Start:  AIC=-657.7
## GPA ~ Year + Gender + Award + Height + Weight + Siblings + SAT +
##     Piercings
##
##             Df Sum of Sq    RSS     AIC
## - Year       3    0.2203 41.265 -661.95
## - Siblings   1    0.0176 41.062 -659.56
## - Height     1    0.0508 41.095 -659.30
## <none>                   41.045 -657.70
## - Piercings  1    0.2634 41.308 -657.60
## - Weight     1    0.3232 41.368 -657.13
## - Gender     1    0.8627 41.907 -652.88
## - Award      2    1.2306 42.275 -652.01
## - SAT        1    4.8394 45.884 -623.14
##
## Step:  AIC=-661.95
## GPA ~ Gender + Award + Height + Weight + Siblings + SAT + Piercings
##
##             Df Sum of Sq    RSS     AIC
## - Siblings   1    0.0078 41.273 -663.88
## - Height     1    0.0370 41.302 -663.65
## <none>                   41.265 -661.95
## - Piercings  1    0.2760 41.541 -661.76
## - Weight     1    0.2977 41.563 -661.59
## + Year       3    0.2203 41.045 -657.70
## - Gender     1    0.8279 42.093 -657.43
## - Award      2    1.4185 42.683 -654.86
## - SAT        1    4.9350 46.200 -626.89
##
## Step:  AIC=-663.88
## GPA ~ Gender + Award + Height + Weight + SAT + Piercings
##
##             Df Sum of Sq    RSS     AIC
## - Height     1    0.0374 41.310 -665.59
## <none>                   41.273 -663.88
## - Piercings  1    0.2897 41.562 -663.59
## - Weight     1    0.2939 41.567 -663.56
## + Siblings   1    0.0078 41.265 -661.95
## + Year       3    0.2105 41.062 -659.56
## - Gender     1    0.8546 42.127 -659.16
## - Award      2    1.4167 42.689 -656.81
## - SAT        1    4.9309 46.204 -628.87
##
## Step:  AIC=-665.59
## GPA ~ Gender + Award + Weight + SAT + Piercings
##
##             Df Sum of Sq    RSS     AIC
## <none>                   41.310 -665.59
## - Weight     1    0.2566 41.567 -665.55
```

```
## - Piercings  1    0.3118 41.622 -665.12
## + Height     1    0.0374 41.273 -663.88
## + Siblings   1    0.0083 41.302 -663.65
## + Year       3    0.1968 41.113 -661.15
## - Gender     1    0.8206 42.131 -661.13
## - Award      2    1.3939 42.704 -658.70
## - SAT        1    5.0174 46.327 -629.99
```

```r
summary(step.model)
```

```
##
## Call:
## lm(formula = GPA ~ Gender + Award + Weight + SAT + Piercings,
##     data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15416 -0.22811  0.02819  0.25442  0.97550
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1753927  0.2573008   8.455 9.96e-16 ***
## GenderM      -0.1699487  0.0673011  -2.525    0.012 *
## AwardNobel    0.0847619  0.0730906   1.160    0.247
## AwardOlympic -0.0583115  0.0721222  -0.809    0.419
## Weight       -0.0011628  0.0008234  -1.412    0.159
## SAT           0.0010684  0.0001711   6.244 1.35e-09 ***
## Piercings    -0.0215792  0.0138642  -1.556    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3587 on 321 degrees of freedom
## Multiple R-squared:  0.2111, Adjusted R-squared:  0.1964
## F-statistic: 14.32 on 6 and 321 DF,  p-value: 1.813e-14
```

```r
library(leaps)
```

```r
models <- regsubsets(GPA ~ Year + Gender + Award + HigherSAT + Height + Weight + Siblings + BirthOrder +
                     method = "seqrep")
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(GPA ~ Year + Gender + Award + HigherSAT +
##     Height + Weight + Siblings + BirthOrder + SAT + Piercings,
##     data = SSurvey, nvmax = 5, method = "seqrep")
## 13 Variables  (and intercept)
##                  Forced in Forced out
## YearJunior           FALSE      FALSE
## YearSenior           FALSE      FALSE
## YearSophomore        FALSE      FALSE
## GenderM              FALSE      FALSE
## AwardNobel           FALSE      FALSE
## AwardOlympic         FALSE      FALSE
## HigherSATVerbal      FALSE      FALSE
## Height               FALSE      FALSE
## Weight               FALSE      FALSE
```

```
## Siblings             FALSE      FALSE
## BirthOrder           FALSE      FALSE
## SAT                  FALSE      FALSE
## Piercings            FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: 'sequential replacement'
##          YearJunior YearSenior YearSophomore GenderM AwardNobel AwardOlympic
## 1  ( 1 ) " "        " "        " "           " "     " "        " "
## 2  ( 1 ) " "        " "        " "           "*"     " "        " "
## 3  ( 1 ) " "        " "        " "           "*"     "*"        " "
## 4  ( 1 ) " "        " "        " "           "*"     "*"        " "
## 5  ( 1 ) " "        " "        " "           "*"     "*"        " "
##          HigherSATVerbal Height Weight Siblings BirthOrder SAT Piercings
## 1  ( 1 ) " "             " "    " "    " "      " "        "*" " "
## 2  ( 1 ) " "             " "    " "    " "      " "        "*" " "
## 3  ( 1 ) " "             " "    " "    " "      " "        "*" " "
## 4  ( 1 ) " "             " "    " "    " "      " "        "*" "*"
## 5  ( 1 ) " "             " "    "*"    " "      " "        "*" "*"
```

Both of these stepwise regression models picked:

- SAT
- Gender
- Award
- Weight
- Piercings

Complete Second Order

```
model6 <- lm(GPA ~ SAT + Piercings + Weight + SAT*Piercings + SAT*Weight + Piercings*Weight + SAT*Pierc
summary(model6)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Piercings + Weight + SAT * Piercings +
##     SAT * Weight + Piercings * Weight + SAT * Piercings * Weight +
##     I(SAT^2) + I(Piercings^2) + I(Weight^2) + Gender + Award +
##     Gender * Award + Gender * SAT + Gender * Piercings + Gender *
##     Weight + Gender * SAT * Piercings + Gender * SAT * Weight +
##     Gender * Piercings * Weight + Gender * SAT * Piercings *
##     Weight + Gender * I(SAT^2) + Gender * I(Piercings^2) + Gender *
##     I(Weight^2) + Award * SAT + Award * Piercings + Award * Weight +
##     Award * SAT * Piercings + Award * SAT * Weight + Award *
##     Piercings * Weight + Award * SAT * Piercings * Weight + Award *
##     I(SAT^2) + Award * I(Piercings^2) + Award * I(Weight^2),
##     data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12802 -0.22123  0.01308  0.25054  0.87249
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        6.372e+00  1.021e+01   0.624   0.5329
```

```
## SAT                                 -1.627e-03  1.008e-02  -0.161   0.8719
## Piercings                           -4.587e+00  3.774e+00  -1.215   0.2253
## Weight                              -4.494e-02  6.654e-02  -0.675   0.5000
## I(SAT^2)                             1.802e-07  2.838e-06   0.063   0.9494
## I(Piercings^2)                      -5.319e-03  2.881e-02  -0.185   0.8536
## I(Weight^2)                          5.444e-05  1.147e-04   0.475   0.6354
## GenderM                              5.791e+00  7.522e+00   0.770   0.4420
## AwardNobel                          -3.856e+00  9.459e+00  -0.408   0.6839
## AwardOlympic                        -1.097e+01  9.807e+00  -1.118   0.2644
## SAT:Piercings                        3.434e-03  3.162e-03   1.086   0.2784
## SAT:Weight                           2.181e-05  4.738e-05   0.460   0.6457
## Piercings:Weight                     3.389e-02  2.641e-02   1.283   0.2004
## GenderM:AwardNobel                  -2.162e-01  3.903e-01  -0.554   0.5800
## GenderM:AwardOlympic                -2.421e-01  3.716e-01  -0.652   0.5151
## SAT:GenderM                         -7.382e-03  8.320e-03  -0.887   0.3757
## Piercings:GenderM                   -5.891e+00  4.242e+00  -1.389   0.1660
## Weight:GenderM                      -8.845e-03  4.553e-02  -0.194   0.8461
## I(SAT^2):GenderM                     1.947e-06  2.567e-06   0.758   0.4488
## I(Piercings^2):GenderM              -1.607e-02  3.917e-02  -0.410   0.6820
## I(Weight^2):GenderM                 -9.501e-06  6.192e-05  -0.153   0.8782
## SAT:AwardNobel                       5.490e-03  1.015e-02   0.541   0.5890
## SAT:AwardOlympic                     1.033e-02  1.053e-02   0.981   0.3273
## Piercings:AwardNobel                 3.345e+00  3.427e+00   0.976   0.3299
## Piercings:AwardOlympic               6.525e+00  3.537e+00   1.845   0.0661 .
## Weight:AwardNobel                    1.841e-02  5.578e-02   0.330   0.7416
## Weight:AwardOlympic                  6.725e-02  5.646e-02   1.191   0.2346
## I(SAT^2):AwardNobel                 -2.000e-06  3.280e-06  -0.610   0.5425
## I(SAT^2):AwardOlympic               -2.055e-06  3.221e-06  -0.638   0.5241
## I(Piercings^2):AwardNobel            8.312e-04  3.055e-02   0.027   0.9783
## I(Piercings^2):AwardOlympic          5.428e-03  3.093e-02   0.175   0.8608
## I(Weight^2):AwardNobel              -2.820e-05  1.060e-04  -0.266   0.7905
## I(Weight^2):AwardOlympic            -7.038e-05  1.090e-04  -0.646   0.5190
## SAT:Piercings:Weight                -2.513e-05  2.238e-05  -1.123   0.2624
## SAT:Piercings:GenderM                4.968e-03  3.661e-03   1.357   0.1759
## SAT:Weight:GenderM                   1.038e-05  3.737e-05   0.278   0.7813
## Piercings:Weight:GenderM             2.746e-02  2.459e-02   1.116   0.2652
## SAT:Piercings:AwardNobel            -2.484e-03  2.873e-03  -0.865   0.3880
## SAT:Piercings:AwardOlympic          -5.123e-03  3.001e-03  -1.707   0.0889 .
## SAT:Weight:AwardNobel               -9.445e-06  3.536e-05  -0.267   0.7896
## SAT:Weight:AwardOlympic             -3.794e-05  3.508e-05  -1.081   0.2804
## Piercings:Weight:AwardNobel         -2.506e-02  2.356e-02  -1.064   0.2884
## Piercings:Weight:AwardOlympic       -4.629e-02  2.437e-02  -1.899   0.0585 .
## SAT:Piercings:Weight:GenderM        -2.270e-05  2.114e-05  -1.074   0.2838
## SAT:Piercings:Weight:AwardNobel      1.834e-05  2.002e-05   0.916   0.3604
## SAT:Piercings:Weight:AwardOlympic  3.594e-05  2.096e-05   1.714   0.0876 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3638 on 282 degrees of freedom
## Multiple R-squared:  0.2871, Adjusted R-squared:  0.1734
## F-statistic: 2.524 on 45 and 282 DF,  p-value: 2.298e-06
```

Remove Quadratic Terms

```r
model7 <- lm(GPA ~ SAT + Piercings + Weight + SAT*Piercings + SAT*Weight + Piercings*Weight + SAT*Pierc
summary(model7)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Piercings + Weight + SAT * Piercings +
##     SAT * Weight + Piercings * Weight + SAT * Piercings * Weight +
##     Gender + Award + Gender * Award + Gender * SAT + Gender *
##     Piercings + Gender * Weight + Gender * SAT * Piercings +
##     Gender * SAT * Weight + Gender * Piercings * Weight + Gender *
##     SAT * Piercings * Weight + Award * SAT + Award * Piercings +
##     Award * Weight + Award * SAT * Piercings + Award * SAT *
##     Weight + Award * Piercings * Weight + Award * SAT * Piercings *
##     Weight, data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13560 -0.21771  0.01013  0.24145  0.88004
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.441e+00  7.669e+00   0.579   0.5629
## SAT                              -1.006e-03  6.348e-03  -0.158   0.8742
## Piercings                        -5.030e+00  3.177e+00  -1.583   0.1144
## Weight                           -2.541e-02  5.270e-02  -0.482   0.6301
## GenderM                           3.147e+00  5.789e+00   0.544   0.5872
## AwardNobel                        5.022e-01  6.326e+00   0.079   0.9368
## AwardOlympic                     -5.290e+00  6.393e+00  -0.827   0.4087
## SAT:Piercings                     3.926e-03  2.659e-03   1.477   0.1408
## SAT:Weight                        2.094e-05  4.346e-05   0.482   0.6303
## Piercings:Weight                  3.682e-02  2.264e-02   1.626   0.1050
## GenderM:AwardNobel               -1.886e-01  2.711e-01  -0.696   0.4871
## GenderM:AwardOlympic             -1.164e-01  2.689e-01  -0.433   0.6653
## SAT:GenderM                      -2.779e-03  4.725e-03  -0.588   0.5568
## Piercings:GenderM                -5.865e+00  4.055e+00  -1.447   0.1491
## Weight:GenderM                   -1.212e-02  4.114e-02  -0.295   0.7685
## SAT:AwardNobel                    1.399e-04  5.238e-03   0.027   0.9787
## SAT:AwardOlympic                  4.884e-03  5.334e-03   0.916   0.3606
## Piercings:AwardNobel              3.940e+00  2.923e+00   1.348   0.1787
## Piercings:AwardOlympic            6.895e+00  2.890e+00   2.386   0.0177 *
## Weight:AwardNobel                 5.254e-03  3.948e-02   0.133   0.8942
## Weight:AwardOlympic               3.872e-02  3.938e-02   0.983   0.3263
## SAT:Piercings:Weight             -2.871e-05  1.887e-05  -1.522   0.1292
## SAT:Piercings:GenderM             5.082e-03  3.495e-03   1.454   0.1470
## SAT:Weight:GenderM                1.112e-05  3.363e-05   0.331   0.7411
## Piercings:Weight:GenderM          2.764e-02  2.355e-02   1.174   0.2415
## SAT:Piercings:AwardNobel         -3.118e-03  2.445e-03  -1.275   0.2033
## SAT:Piercings:AwardOlympic       -5.621e-03  2.437e-03  -2.307   0.0218 *
## SAT:Weight:AwardNobel            -6.878e-06  3.265e-05  -0.211   0.8333
## SAT:Weight:AwardOlympic          -3.540e-05  3.277e-05  -1.081   0.2808
## Piercings:Weight:AwardNobel      -2.930e-02  2.033e-02  -1.441   0.1506
## Piercings:Weight:AwardOlympic    -4.945e-02  2.025e-02  -2.442   0.0152 *
## SAT:Piercings:Weight:GenderM     -2.394e-05  2.030e-05  -1.179   0.2393
## SAT:Piercings:Weight:AwardNobel   2.291e-05  1.696e-05   1.351   0.1778
```

```
## SAT:Piercings:Weight:AwardOlympic  4.018e-05  1.702e-05   2.361   0.0189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3583 on 294 degrees of freedom
## Multiple R-squared:  0.2794, Adjusted R-squared:  0.1985
## F-statistic: 3.454 on 33 and 294 DF,  p-value: 7.464e-09
```

Performing ANOVA test to see if the quadratic terms are useful.

```r
anova(model6, model7)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + I(SAT^2) +
##     I(Piercings^2) + I(Weight^2) + Gender + Award + Gender *
##     Award + Gender * SAT + Gender * Piercings + Gender * Weight +
##     Gender * SAT * Piercings + Gender * SAT * Weight + Gender *
##     Piercings * Weight + Gender * SAT * Piercings * Weight +
##     Gender * I(SAT^2) + Gender * I(Piercings^2) + Gender * I(Weight^2) +
##     Award * SAT + Award * Piercings + Award * Weight + Award *
##     SAT * Piercings + Award * SAT * Weight + Award * Piercings *
##     Weight + Award * SAT * Piercings * Weight + Award * I(SAT^2) +
##     Award * I(Piercings^2) + Award * I(Weight^2)
## Model 2: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award + Gender * Award + Gender * SAT + Gender * Piercings +
##     Gender * Weight + Gender * SAT * Piercings + Gender * SAT *
##     Weight + Gender * Piercings * Weight + Gender * SAT * Piercings *
##     Weight + Award * SAT + Award * Piercings + Award * Weight +
##     Award * SAT * Piercings + Award * SAT * Weight + Award *
##     Piercings * Weight + Award * SAT * Piercings * Weight
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1    282 37.330
## 2    294 37.734 -12  -0.40351 0.254 0.9949
```

The high p-value suggests that the quadratic terms do not make the complete second order model significantly better than the reduced one. Therefore we proceed with model 7.

Remove QNxQL Interactions

```r
model8 <- lm(GPA ~ SAT + Piercings + Weight + SAT*Piercings + SAT*Weight + Piercings*Weight + SAT*Pierc
summary(model8)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Piercings + Weight + SAT * Piercings +
##     SAT * Weight + Piercings * Weight + SAT * Piercings * Weight +
##     Gender + Award + Gender * Award, data = SSurvey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1421 -0.2217  0.0278  0.2732  0.9720
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           2.316e+00  1.331e+00   1.740   0.0828 .
## SAT                   9.708e-04  1.096e-03   0.886   0.3763
## Piercings             3.314e-01  5.298e-01   0.626   0.5320
## Weight               -9.880e-04  7.513e-03  -0.132   0.8955
## GenderM               2.294e-02  1.483e-01   0.155   0.8772
## AwardNobel            1.305e-01  9.536e-02   1.369   0.1720
## AwardOlympic          1.740e-02  9.463e-02   0.184   0.8542
## SAT:Piercings        -3.426e-04  4.430e-04  -0.773   0.4399
## SAT:Weight           -5.882e-07  6.225e-06  -0.094   0.9248
## Piercings:Weight     -3.319e-03  3.711e-03  -0.894   0.3718
## GenderM:AwardNobel   -1.601e-01  1.516e-01  -1.056   0.2919
## GenderM:AwardOlympic -2.147e-01  1.486e-01  -1.445   0.1493
## SAT:Piercings:Weight  3.118e-06  3.120e-06   0.999   0.3184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3592 on 315 degrees of freedom
## Multiple R-squared:  0.2238, Adjusted R-squared:  0.1942
## F-statistic: 7.567 on 12 and 315 DF,  p-value: 2.635e-12
```

Performing ANOVA test to see if the quantitative-qualitative interaction terms are useful.

```
anova(model7, model8)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award + Gender * Award + Gender * SAT + Gender * Piercings +
##     Gender * Weight + Gender * SAT * Piercings + Gender * SAT *
##     Weight + Gender * Piercings * Weight + Gender * SAT * Piercings *
##     Weight + Award * SAT + Award * Piercings + Award * Weight +
##     Award * SAT * Piercings + Award * SAT * Weight + Award *
##     Piercings * Weight + Award * SAT * Piercings * Weight
## Model 2: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award + Gender * Award
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1    294 37.734
## 2    315 40.648 -21   -2.9141 1.0812 0.3674
```

The high p-value suggests that the quantitative-qualitative interaction terms do not make the first model significantly better than the reduced one. Therefore we proceed with model 8.

Remove QLxQL Interactions

```
model9 <- lm(GPA ~ SAT + Piercings + Weight + SAT*Piercings + SAT*Weight + Piercings*Weight + SAT*Pierc
summary(model9)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Piercings + Weight + SAT * Piercings +
##     SAT * Weight + Piercings * Weight + SAT * Piercings * Weight +
##     Gender + Award, data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.13592 -0.21552  0.01881  0.26792  0.99166
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.462e+00  1.313e+00   1.875   0.0617 .
## SAT                     9.053e-04  1.081e-03   0.838   0.4028
## Piercings               3.737e-01  5.282e-01   0.707   0.4798
## Weight                 -1.591e-03  7.389e-03  -0.215   0.8296
## GenderM                -1.522e-01  6.821e-02  -2.232   0.0263 *
## AwardNobel              7.317e-02  7.384e-02   0.991   0.3225
## AwardOlympic           -7.130e-02  7.271e-02  -0.981   0.3275
## SAT:Piercings          -3.780e-04  4.419e-04  -0.855   0.3930
## SAT:Weight             -9.431e-08  6.121e-06  -0.015   0.9877
## Piercings:Weight       -3.631e-03  3.701e-03  -0.981   0.3274
## SAT:Piercings:Weight    3.374e-06  3.114e-06   1.084   0.2794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3593 on 317 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.1938
## F-statistic: 8.861 on 10 and 317 DF,  p-value: 7.522e-13
```

Performing ANOVA test to see if the qualitative-qualitative interaction terms are useful.

```r
anova(model8, model9)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award + Gender * Award
## Model 2: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    315 40.648
## 2    317 40.924 -2  -0.27686 1.0728 0.3433
```

The high p-value suggests that the qualitative-qualitative interaction terms do not make the first model significantly better than the reduced one. Therefore we proceed with model 9.

Removing QNxQN Interactions

```r
model10 <- lm(GPA ~ SAT + Piercings + Weight + Gender + Award, data = SSurvey)
summary(model10)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Piercings + Weight + Gender + Award,
##     data = SSurvey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15416 -0.22811  0.02819  0.25442  0.97550
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   2.1753927  0.2573008   8.455 9.96e-16 ***
## SAT           0.0010684  0.0001711   6.244 1.35e-09 ***
## Piercings    -0.0215792  0.0138642  -1.556    0.121
## Weight       -0.0011628  0.0008234  -1.412    0.159
## GenderM      -0.1699487  0.0673011  -2.525    0.012 *
## AwardNobel    0.0847619  0.0730906   1.160    0.247
## AwardOlympic -0.0583115  0.0721222  -0.809    0.419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3587 on 321 degrees of freedom
## Multiple R-squared:  0.2111, Adjusted R-squared:  0.1964
## F-statistic: 14.32 on 6 and 321 DF,  p-value: 1.813e-14
```

Performing ANOVA test to see if the quantitative-quantitative interaction terms are useful.

```r
anova(model9, model10)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Piercings + Weight + SAT * Piercings + SAT * Weight +
##     Piercings * Weight + SAT * Piercings * Weight + Gender +
##     Award
## Model 2: GPA ~ SAT + Piercings + Weight + Gender + Award
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    317 40.924
## 2    321 41.310 -4  -0.38555 0.7466 0.5609
```

The high p-value suggests that the quantitative-quantitative interaction terms do not make the first model significantly better than the reduced one. Therefore we proceed with model 10.

We can observe that model 4 is a reduced version of model10. We will perform a final ANOVA test to see if the `Piercings` and `Weight` variables are significant.

```r
anova(model10, model4)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SAT + Piercings + Weight + Gender + Award
## Model 2: GPA ~ SAT + Gender + Award
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    321 41.31
## 2    323 41.86 -2  -0.55034 2.1382 0.1195
```

We see that the p-value is $> .1$, so we choose our final model to be model 4.

Even though model 10 had a slightly higher adjusted R-squared, the ANOVA test chooses the reduced model to be better.

Our Final Model

Let us take a look at model 4 again.

```r
summary(model4)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + Gender + Award, data = SSurvey)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18260 -0.21412  0.03081  0.25894  0.98061
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8919555  0.2118622   8.930  < 2e-16 ***
## SAT           0.0011140  0.0001703   6.542 2.37e-10 ***
## GenderM      -0.1465098  0.0406068  -3.608 0.000357 ***
## AwardNobel    0.0781172  0.0732680   1.066 0.287137
## AwardOlympic -0.0645140  0.0721969  -0.894 0.372210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.36 on 323 degrees of freedom
## Multiple R-squared:  0.2006, Adjusted R-squared:  0.1907
## F-statistic: 20.26 on 4 and 323 DF,  p-value: 6.613e-15
```

Even tough we have a low adjusted R-squared, we have a low p-value.

We see that we have our intercept, β_0 of 1.8919555

Coefficient for `SAT` is 0.0011140

`Gender` is a qualitative variable, therefore we define it as:

$$\mathrm{G}_{Male} = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

And it has a coefficient of -0.1465098

Then we have Award which is defined as:

$$\mathrm{A}_{Nobel} = \begin{cases} 1 & \text{if Nobel} \\ 0 & \text{if otherwise} \end{cases}$$

$$\mathrm{A}_{Olympic} = \begin{cases} 1 & \text{if Olympic} \\ 0 & \text{if otherwise} \end{cases}$$

A_{Nobel} has a coefficient of 0.0781172 and $\mathrm{A}_{Olympic}$ has a coefficient of -0.0645140

We end if with our prediction equation:

$$\hat{y} = 1.8919555 + 0.001114(\mathrm{SAT}) - 0.1465098(\mathrm{G}_{Male}) + 0.0781172(\mathrm{A}_{Nobel}) - 0.064514(\mathrm{A}_{Olympic})$$

Predicting My GPA

I had a score of 1320 on my SAT. I am a Female, and I would prefer to win an Academy award.

```r
newdat <- data.frame(SAT = 1320,
                     Gender = 'F',
                     Award = 'Academy')
predict(model4, newdata = newdat, interval = 'confidence', level = .95)
```

```
##        fit      lwr      upr
## 1 3.362393 3.221756 3.503029
```

The model is 95% confident that my GPA is between 3.221756 and 3.503029. Even though this seems accurate, my GPA is above this range. This may be due to the data being collected from a certain school district or another reason.

```
predict(model4, newdata = newdat, interval = 'prediction', level = .9)
```

```
##        fit     lwr      upr
## 1 3.362393 2.75695 3.967835
```

Even though this prediction interval may have a far lower and upper bound, my GPA does fall in this range.

Predicting My Friends GPA

My friend got a score of 1330 on the SAT. Is a male and stated that he would rather receive a nobel award.

```
# Used for prediction
newdat <- data.frame(SAT = 1330,
                     Gender = 'M',
                     Award = 'Nobel')
predict(model4, newdata = newdat, interval = 'confidence', level = .95)
```

```
##       fit      lwr      upr
## 1 3.30514 3.226185 3.384094
```

The model is 95% confident that my friend's GPA is between 3.226185 and 3.384094. Even though this seems accurate, my friend's GPA is above this range. This may be due to a similar reason my GPA fell above the range the model predicted.

```
predict(model4, newdata = newdat, interval = 'prediction', level = .9)
```

```
##       fit      lwr      upr
## 1 3.30514 2.707614 3.902666
```

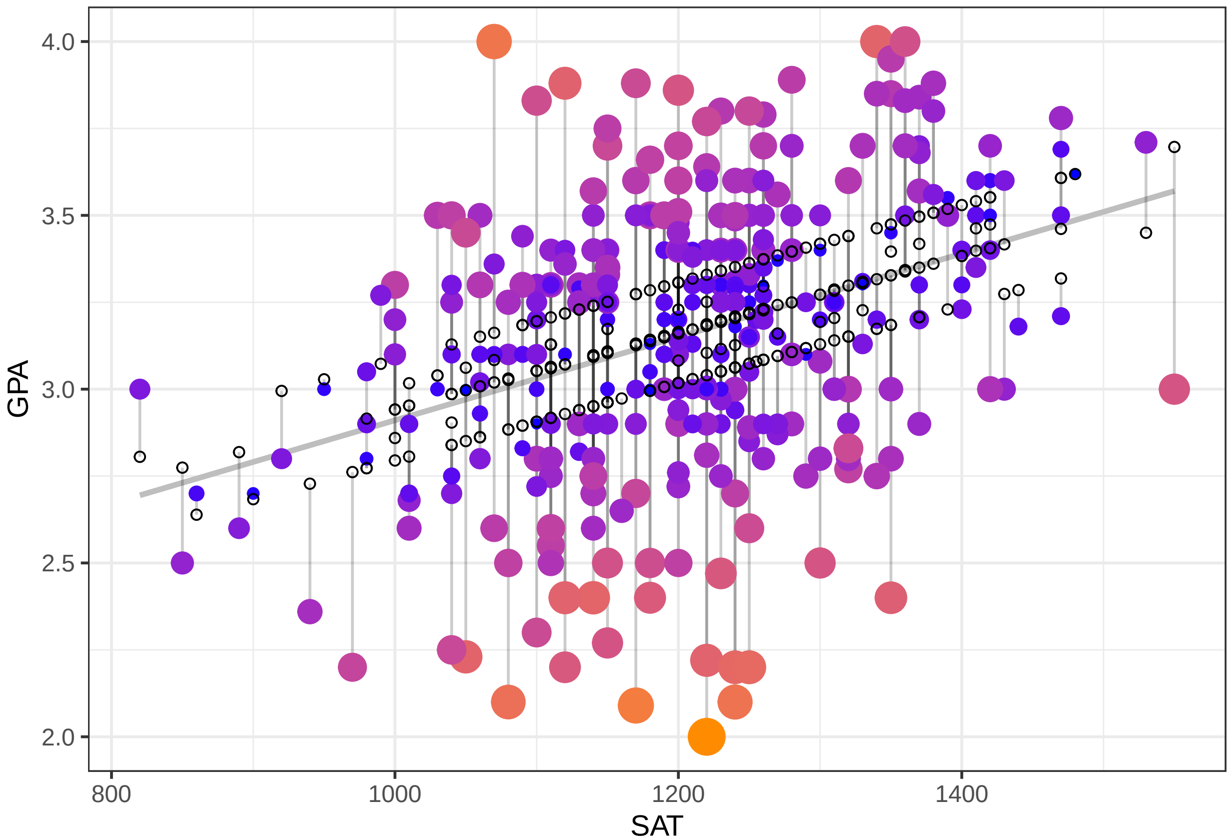
Even though this prediction interval may have a far lower and upper bound, my friend's GPA does fall in this range.

Residual Analysis

Color Coded Residual Plot

The plot shows graphically the size of the residual value using a color code (orange is longer line to blue - smaller line) and size of point. The size of residual is the length of the vertical line from the point to where it meets the regression line. We can observe that, the further the point is from the line, the larger and more orange it gets.

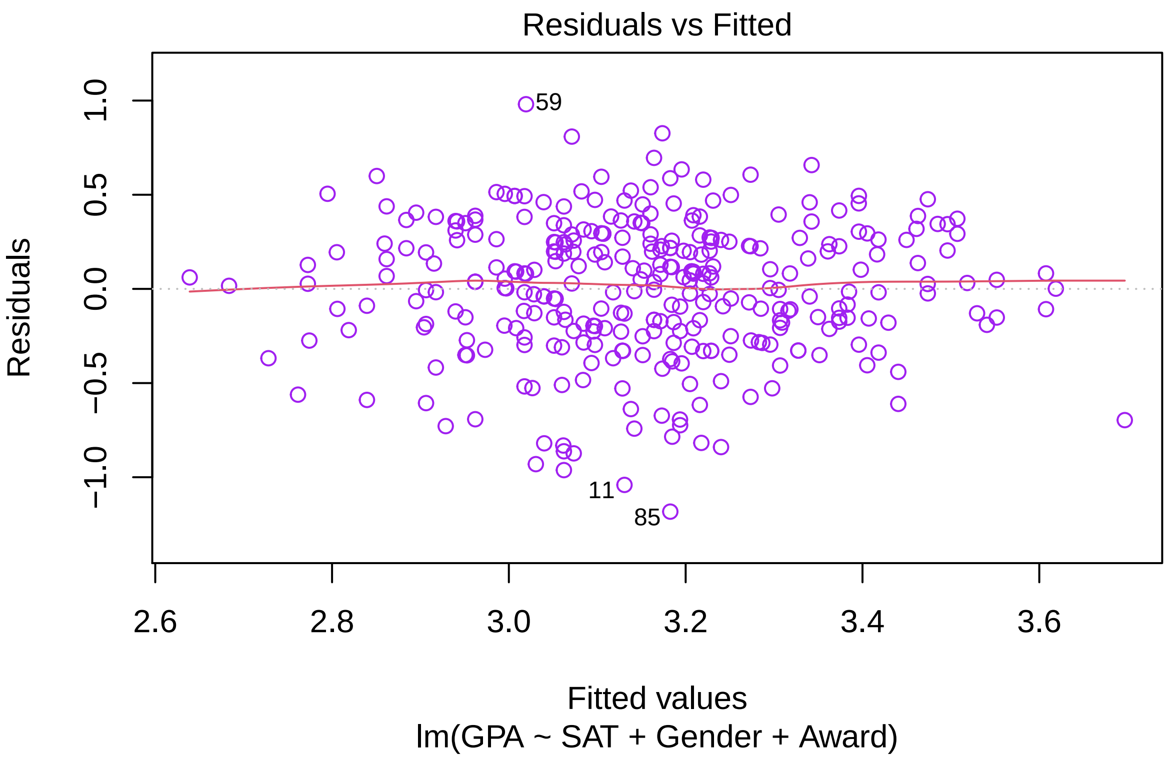
```
d <- SSurvey
d$predicted <- predict(model4)   # Save the predicted values
d$residuals <- residuals(model4) # Save the residual values
ggplot(d, aes(x = SAT, y = GPA)) +
  geom_smooth(method = "lm", se = FALSE, color = "grey") +     # regression line
  geom_segment(aes(xend = SAT, yend = predicted), alpha = .2) +      # draw line from point to line
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +  # size of the points
  scale_color_continuous(low = "blue", high = "darkorange") +     # color of the points mapped to resid
  guides(color = FALSE, size = FALSE) +                           # Size legend removed
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```



Residuals vs Fitted Plot

Residual plots are used to look for underlying patterns in the residuals that may mean that the model has a problem.

```
plot(model4, which=1, col=c("purple"))
```

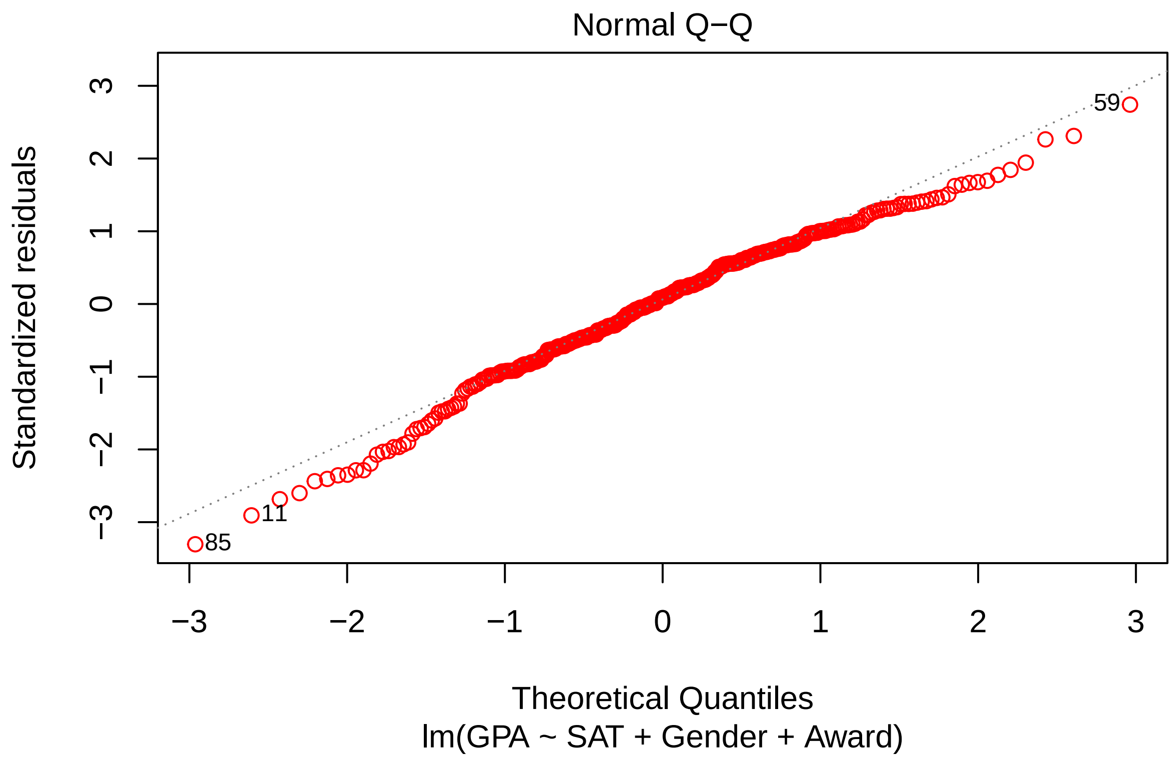


We see that there is slightly more clutter around the middle. The points seem to be equally distributed above and below the line.

Normal Q–Q (quantile-quantile) Plot

One of our assumptions is that the residuals are normally distributed. To check this assumption, we construct the Q-Q plot below.

```
plot(model4, which=2, col=c("red"))
```

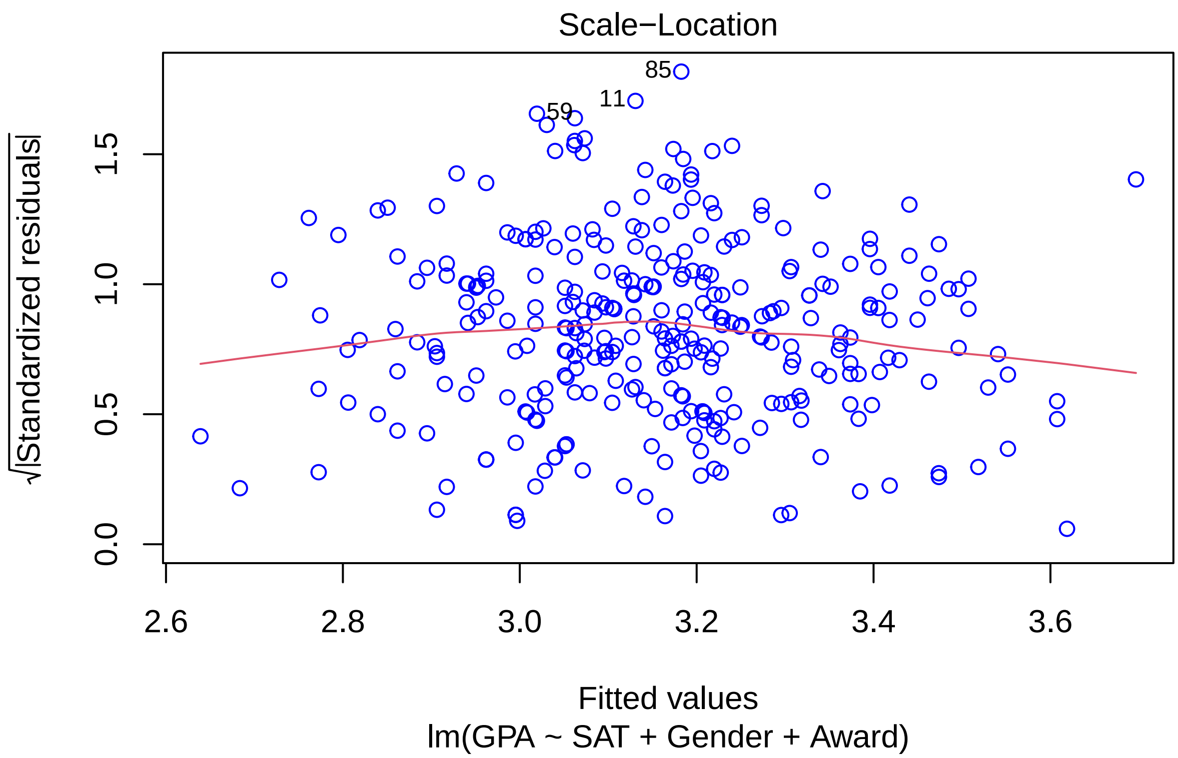


Our plot has a nearly linear trend. This is a good indication that our residuals are nearly normally distributed.

Scale-Location

This plot test the linear regression assumption of equal variance (homoscedasticity) i.e. that the residuals have equal variance along the regression line. It is also called the Spread-Location plot.

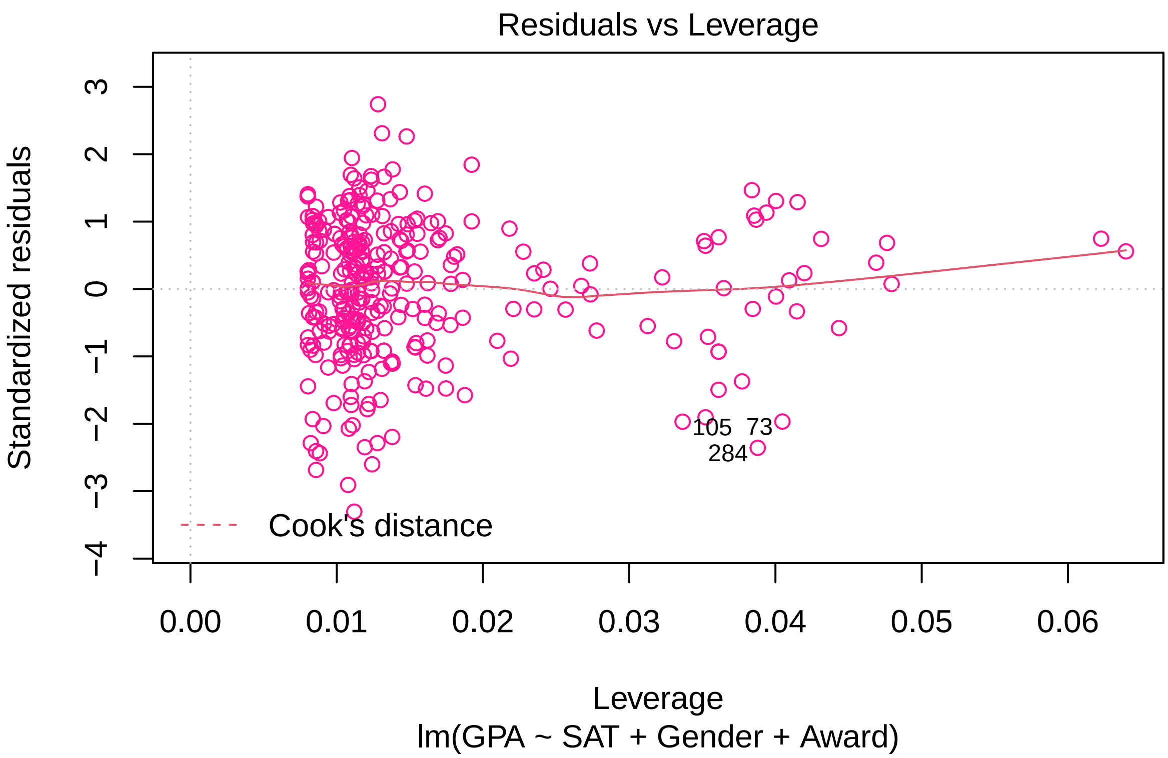
```
plot(model4, which=3, col=c("blue"))
```



Residuals vs Leverage

This plot can be used to find influential cases in the dataset. An influential case is one that, if removed, will affect the model so its inclusion or exclusion should be considered. An influential case may or may not be an outlier and the purpose of this chart is to identify cases that have high influence in the model. Outliers will tend to exert leverage and therefore influence on the model.

```
plot(model4, which=5, col=c("deeppink"))
```



We can see that most of the leverages are low, which is a good indication. Low leverage means that we do not have influential cases.

Conclusion

Our model does not seem to have significant departures from the assumptions. This means that we can use our model. A drawback is the low R-squared, that says only about 19% of the variation in GPA can be explain by our model. The low p value from the global F-test suggests that out model is statistically useful for predicting GPA. As I tested it to predict my GPA, as well as my friend's GPA, the model seems to be fairly accurate. Another source of concern rises from the fact the the model is not curvilinear, maximum GPA is 4.0. Since we have a straight line model, the line will eventually exceed 4 based on the parameters.

The End