Prediction GPAs

Regression Models

Brady Fisher

Introduction

For this project I will be using data from the FirstYearGPA dataset in the Stat2Data package, with the goal of predicting the GPA in the first year of college for students. To do this I will be using both Simple and Multiple Linear Regression Models.

Loading the Data

```
FirstYearGPA = read.csv("http://www.stat2.org/datasets/FirstYearGPA.csv")
```

1. Fit and Assess the Model.

a) Simple Linear Regression model

Fit the Model

For the Simple Linear Model portion of this project I have been assigned the SAT Verbal score(SATV) for the students as my predictor variable. GPA will be my response variable in my Simple Linear Model. I will create the model:

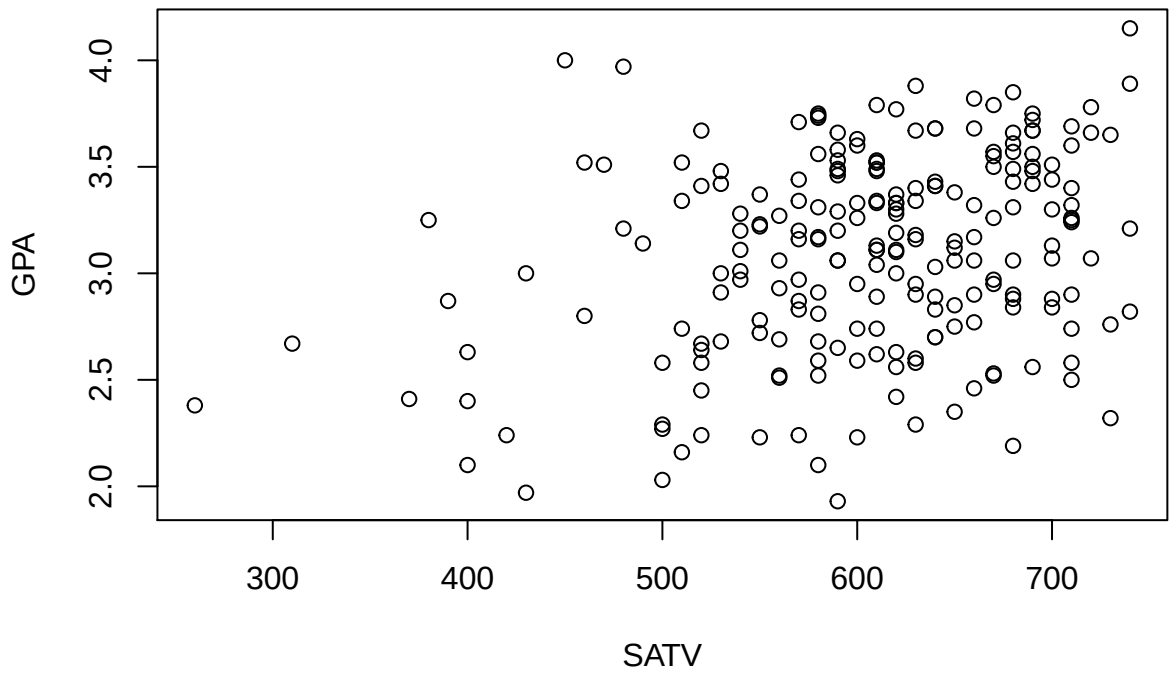
```
simpleModel = lm(GPA ~ SATV, data = FirstYearGPA)
summary(simpleModel)
```

```
##
## Call:
## lm(formula = GPA ~ SATV, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14057 -0.32756  0.03134  0.34259  1.16723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.0684133  0.2204478   9.383  < 2e-16 ***
## SATV        0.0016986  0.0003609   4.706  4.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4444 on 217 degrees of freedom
## Multiple R-squared:  0.09261,    Adjusted R-squared:  0.08842
## F-statistic: 22.15 on 1 and 217 DF,  p-value: 4.499e-06
```

Here I can see the calculated Linear Model Equation is: $\widehat{GPA} = 2.06841 + 0.00170 * \widehat{SATV}$

I will now take a closer look at the data using a scatter plot.

```
plot(GPA~SATV, data = FirstYearGPA)
```

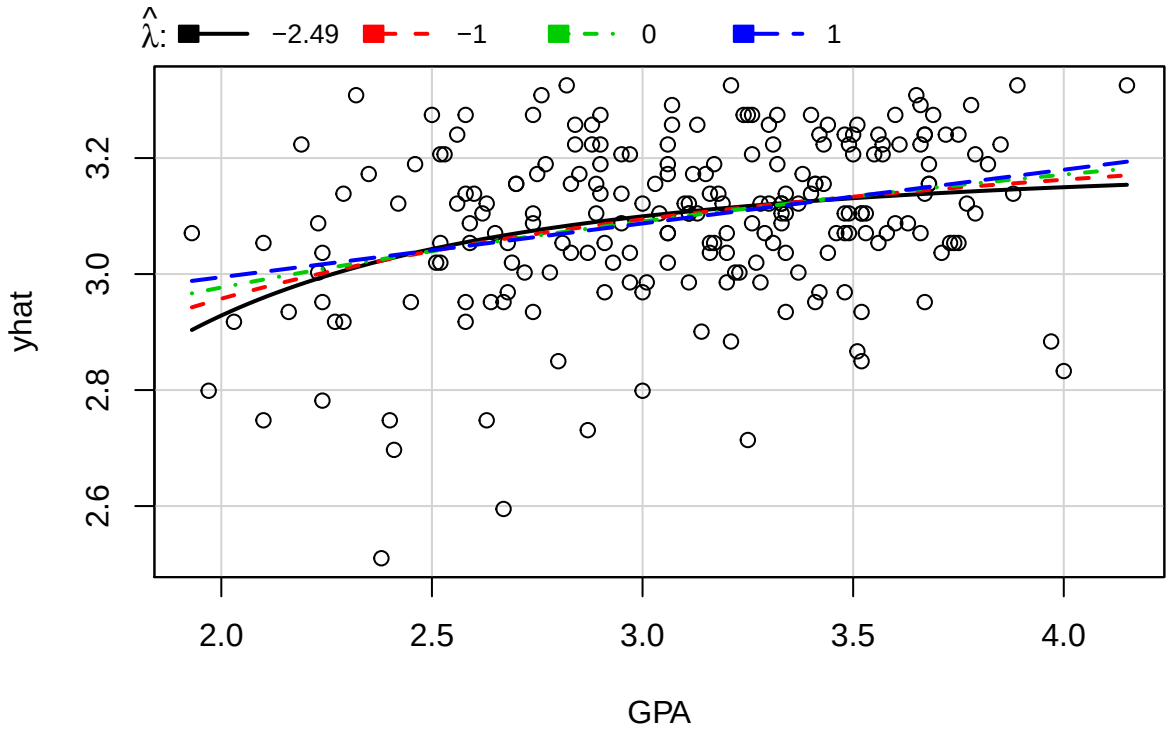


In this plot I see that most of the data is between 500 and 700 SATV with a few lower around 300 and 400 SATV. It does not appear that there needs to be any transformation on the data, but to check I will use the invResPlot function from the car package to assess if a transformation is needed.

```
library("car", lib.loc="~/R/win-library/3.3")
```

```
## Warning: package 'car' was built under R version 3.3.2
```

```
invResPlot(lm(GPA ~ SATV, data = FirstYearGPA))
```

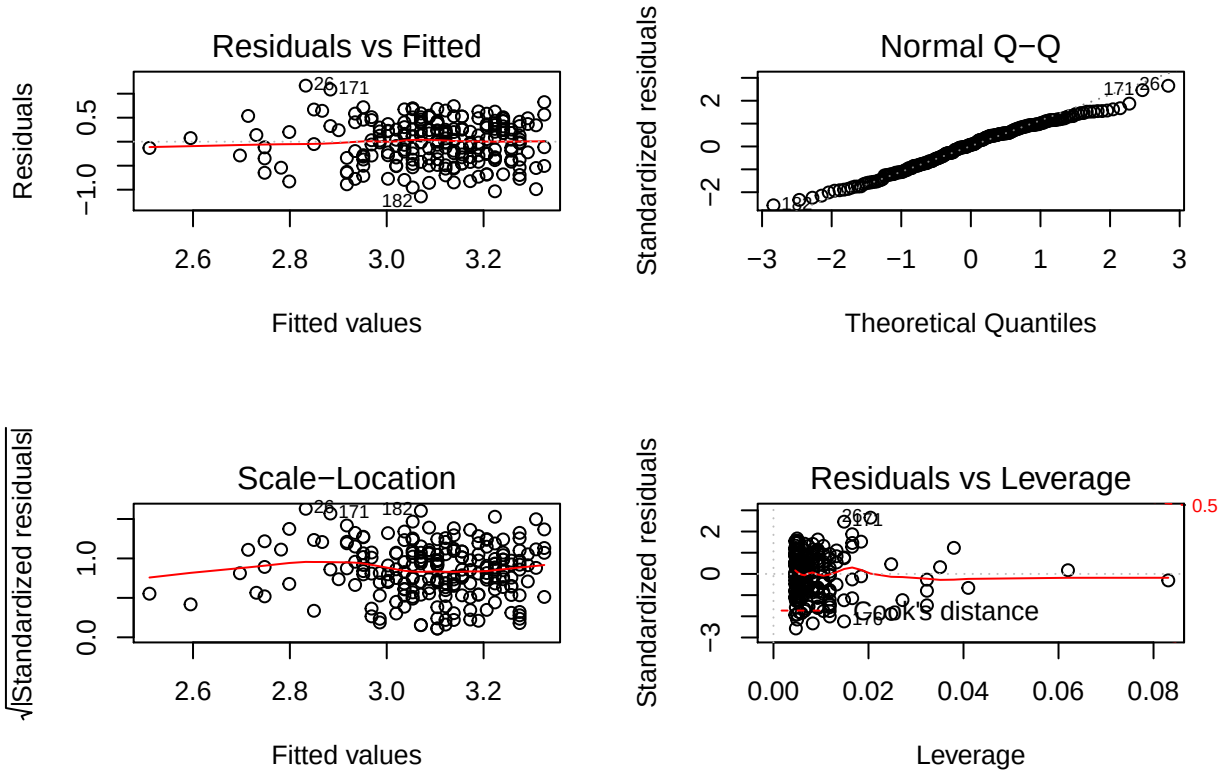



```
##      lambda      RSS
## 1 -2.487753 3.920570
## 2 -1.000000 3.930476
## 3  0.000000 3.946910
## 4  1.000000 3.969021
```

From this calculation I see that the best lambda to use is around -2.49, but the RSS is about the same as a lambda of 1 which represents a simpler linear model. Since a more complex model will have no to minimal benefit here, no transformation is needed.

Assess the Model

```
par(mfrow=c(2,2))
plot(simpleModel)
```



Now to use this Simple Linear Regression Model, I must check the four assumptions of: 1.Linear Relationship 2.Constant variance 3.Independence 4.Conditional Normal Distribution

The first assumption of linear relationship is met, because on the Residual vs Fitted plot the residuals form a nice horizontal band, and there is no apparent pattern.

The second assumption of constant variance is probably met, on the left side of the Residual vs Fitted plot the amount of data points decreases greatly, but for the most part the band on the graph approximately has a constant width.

The third assumption of independence should be met from the study when the data was collected. It would also make sense that every student took their own classes and tests and thus are independent of each other.

The fourth assumption of conditional Normal Distribution is met because the Normal Q-Q forms an approximately straight diagonal line. There is a slight curve at the right tail of the plot that could signify the data having a slight tail.

b) Multiple Linear Regression model

i.

For the Multiple Linear Model portion of this project I have been assigned the SAT Verbal score, the High School GPA of the students, and whether the student is a first generation college student as the predictor variables. GPA will still be the response variable. Now let me define my null model, and full model.

```
nm = lm(GPA ~ 1, data = FirstYearGPA)
fm = lm(GPA ~ SATV * HSGPA * FirstGen, data = FirstYearGPA)
```

Now I will choose the best model using Stepwise Regression twice. First starting with the null model, and

another time starting with the full model. Then if they produce two different models I will compare their adjusted R^2 value to choose the best model.

```
step(nm, scope = list(upper = fm, lower = nm), direction = "both", trace = 0)
```

```
##
## Call:
## lm(formula = GPA ~ HSGPA + SATV + FirstGen, data = FirstYearGPA)
##
## Coefficients:
## (Intercept)        HSGPA         SATV     FirstGen
##    0.715988     0.518540     0.001012    -0.199837
```

```
step(fm, scope = list(upper = fm, lower = nm), direction = "both", trace = 0)
```

```
##
## Call:
## lm(formula = GPA ~ SATV + HSGPA + FirstGen, data = FirstYearGPA)
##
## Coefficients:
## (Intercept)         SATV        HSGPA     FirstGen
##    0.715988     0.001012     0.518540    -0.199837
```

Thus both functions give me the same model of using SATV, HSGPA, and FirstGen to predict GPA. I will create my Multiple Linear Regression Model.

```
multLinModel = lm(GPA ~ SATV + HSGPA + FirstGen, data = FirstYearGPA)
summary(multLinModel)
```

```
##
## Call:
## lm(formula = GPA ~ SATV + HSGPA + FirstGen, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00773 -0.26601  0.02929  0.29703  0.83562
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7159877  0.2950027   2.427  0.01605 *
## SATV         0.0010125  0.0003478   2.911  0.00398 **
## HSGPA        0.5185400  0.0749556   6.918 5.17e-11 ***
## FirstGen    -0.1998374  0.0891536  -2.241  0.02602 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4023 on 215 degrees of freedom
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2529
## F-statistic:  25.6 on 3 and 215 DF,  p-value: 3.349e-14
```

Here the Multiple Linear Model Equation is:

$\widehat{GPA} = 0.7159877 + 0.0010125 * \widehat{SATV} + 0.5185400 * \widehat{HSGPA} + \text{-}0.1998374 * \widehat{FirstGen}$

ii.

Now I am asked to add the terms of whether the student is male(Male), and the number of humanities

credits(HU) the student took to the model, and test if an interaction between Male and HU is needed. Let me define the two models.

```
m1 = lm(GPA ~ SATV + HSGPA + FirstGen + Male + HU, data = FirstYearGPA)
m2 = lm(GPA ~ SATV + HSGPA + FirstGen + Male * HU, data = FirstYearGPA) #with interaction term
```

Here the hypotheses are:

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

with β being the coefficient for the interaction term between Male and HU.

Now I will conduct an F-test using the ANOVA function.

```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ SATV + HSGPA + FirstGen + Male + HU
## Model 2: GPA ~ SATV + HSGPA + FirstGen + Male * HU
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    213 32.205
## 2    212 32.166  1  0.038587 0.2543 0.6146
```

Here the F-statistic is 0.2543, and the p-value is 0.6146. Since the p-value is greater than 0.05, I fail to reject the null hypothesis that the coefficient β for the interaction between Male and HU is 0, and therefore I do not have enough evidence to say the interaction term between Male and HU is significant.

2. Use the model

a)Simple Linear Regression model.

i.

From part 1a I created the simple linear model, now I will look at its p-value for the slope.

```
summary(simpleModel)
```

```
##
## Call:
## lm(formula = GPA ~ SATV, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14057 -0.32756  0.03134  0.34259  1.16723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.0684133  0.2204478   9.383  < 2e-16 ***
## SATV        0.0016986  0.0003609   4.706  4.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4444 on 217 degrees of freedom
## Multiple R-squared:  0.09261,    Adjusted R-squared:  0.08842
## F-statistic: 22.15 on 1 and 217 DF,  p-value: 4.499e-06
```

Here I see the test statistic for the slope is t = 4.706, and the corresponding p-value is 4.5e-06. Since the p-value is much less than 0.05, SATV is a significant predictor of a student's first year GPA.

Now I am tasked with using this model to predict what the first year GPA will be for a student with a SATV of 670.

```
predict(simpleModel, data.frame(SATV = 670))
```

```
##        1
## 3.206455
```

Thus the predicted GPA of a student with a SATV of 670 is 3.206.

ii.

The SATV alone does not seem to be a good predictor of a student's first year GPA. In the summary of the simple regression model I ran in part 2a)i I see that the $R^2 = 0.09261$. This means that only 9.261% of the variability in the GPA can be explained by the SATV.

b)Multiple Regression Model

i.

From part 1b)i I found that the simplest, best model consists of using SATV, HSGPA, and FirstGen as the predictor variables, without any of the interaction terms.

```
multLinModel = lm(GPA ~ SATV + HSGPA + FirstGen, data = FirstYearGPA)
summary(multLinModel)
```

```
##
## Call:
## lm(formula = GPA ~ SATV + HSGPA + FirstGen, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00773 -0.26601  0.02929  0.29703  0.83562
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7159877  0.2950027   2.427  0.01605 *
## SATV         0.0010125  0.0003478   2.911  0.00398 **
## HSGPA        0.5185400  0.0749556   6.918 5.17e-11 ***
## FirstGen    -0.1998374  0.0891536  -2.241  0.02602 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4023 on 215 degrees of freedom
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2529
## F-statistic:  25.6 on 3 and 215 DF,  p-value: 3.349e-14
```

So this model is: $\widehat{GPA} = 0.7159877 + 0.0010125 * \widehat{SATV} + 0.5185400 * \widehat{HSGPA} + -0.1998374 * \widehat{FirstGen}$

I selected this model, because when I compare this model with any-other model including 1 interaction, or excluding 1 of the predictor variables I get a larger AIC. Here AIC is the decision criterion I use to measure how good a model is calculated as $n * log(RSS/n) + 2 * (numberofpraemters)$, where a lower AIC indicates a better model.

ii.

Now I am tasked with predicting what the first year GPA will be for a student who is a first generation college student with a SATV of 670, and a highschool GPA of 3.6.

```
predict(multLinModel, data.frame(SATV = 670, HSGPA = 3.6, FirstGen = 1))
```

```
##       1
## 3.06124
```

Thus the predicted GPA of such a student is 3.0612.