Regression Project

Ian Cran and Alex Her

What Data Set?

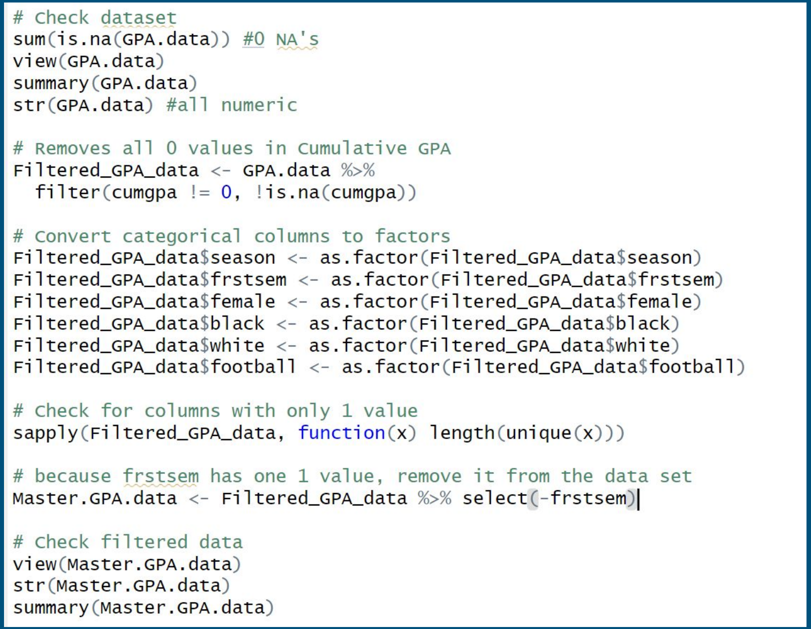
For our project, we are dealing with the GPA dataset. In this dataset, there is data from 366 college students, on 15 columns. These columns being,

- SAT Score
- Total Credit Hours Per Term
- Cumulative GPA
- Athletic Season
- First Semester
- Weighted GPA
- Verbal SAT to math ratio
- Term GPA
- High school class size
- Rank in high school class
- If they are female
- If they are black
- Percentile in high school
- If they play football

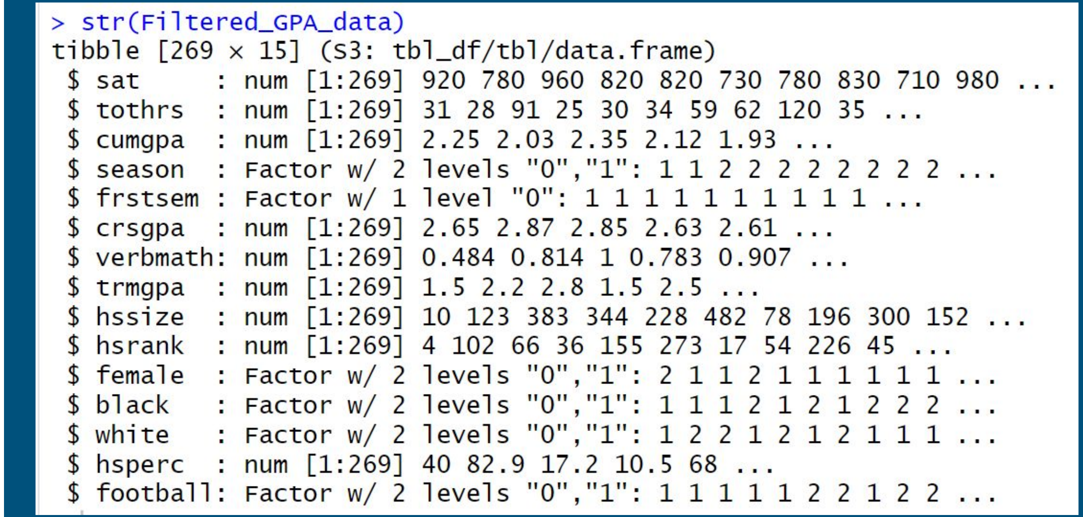
Description and Motivation of Analyses:

- We analyzed student data to see what factors influence cumulative GPA.
- Using regression models, we looked at how factors affects GPA, both individually and together, and whether some factors change the impact of others.
- The goal is to understand what drives academic performance and identify groups of students who may benefit from additional support.

Cleaning data

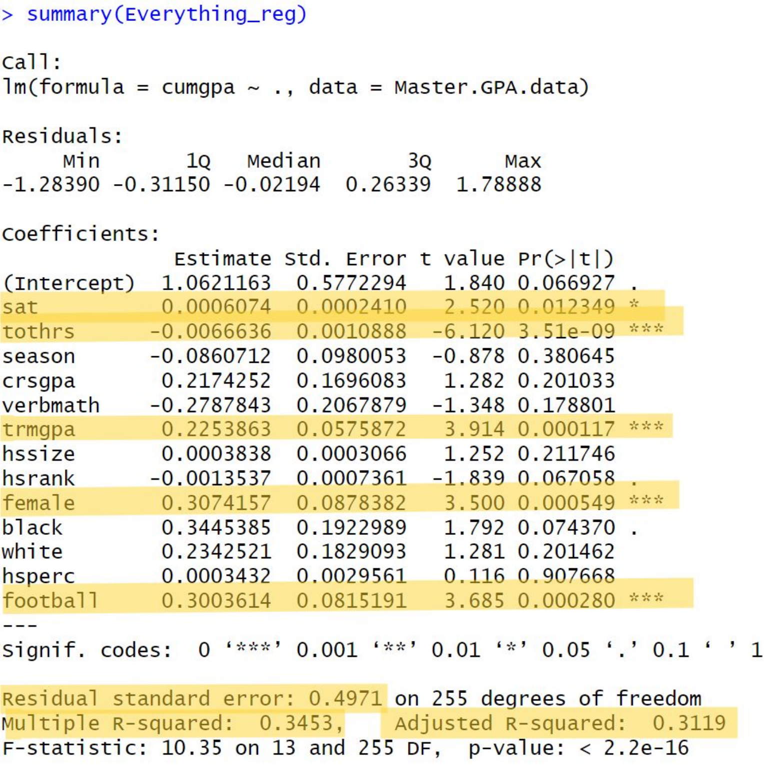


```
# Check dataset
sum(is.na(GPA.data)) #0 NA's
view(GPA.data)
summary(GPA.data)
str(GPA.data) #all numeric

# Removes all 0 values in Cumulative GPA
Filtered_GPA_data <- GPA.data %>%
  filter(cumgpa != 0, !is.na(cumgpa))

# Convert categorical columns to factors
Filtered_GPA_data$season <- as.factor(Filtered_GPA_data$season)
Filtered_GPA_data$frstsem <- as.factor(Filtered_GPA_data$frstsem)
Filtered_GPA_data$female <- as.factor(Filtered_GPA_data$female)
Filtered_GPA_data$black <- as.factor(Filtered_GPA_data$black)
Filtered_GPA_data$white <- as.factor(Filtered_GPA_data$white)
Filtered_GPA_data$football <- as.factor(Filtered_GPA_data$football)

# Check for columns with only 1 value
sapply(Filtered_GPA_data, function(x) length(unique(x)))

# because frstsem has one 1 value, remove it from the data set
Master.GPA.data <- Filtered_GPA_data %>% select(-frstsem)

# Check filtered data
view(Master.GPA.data)
str(Master.GPA.data)
summary(Master.GPA.data)
```



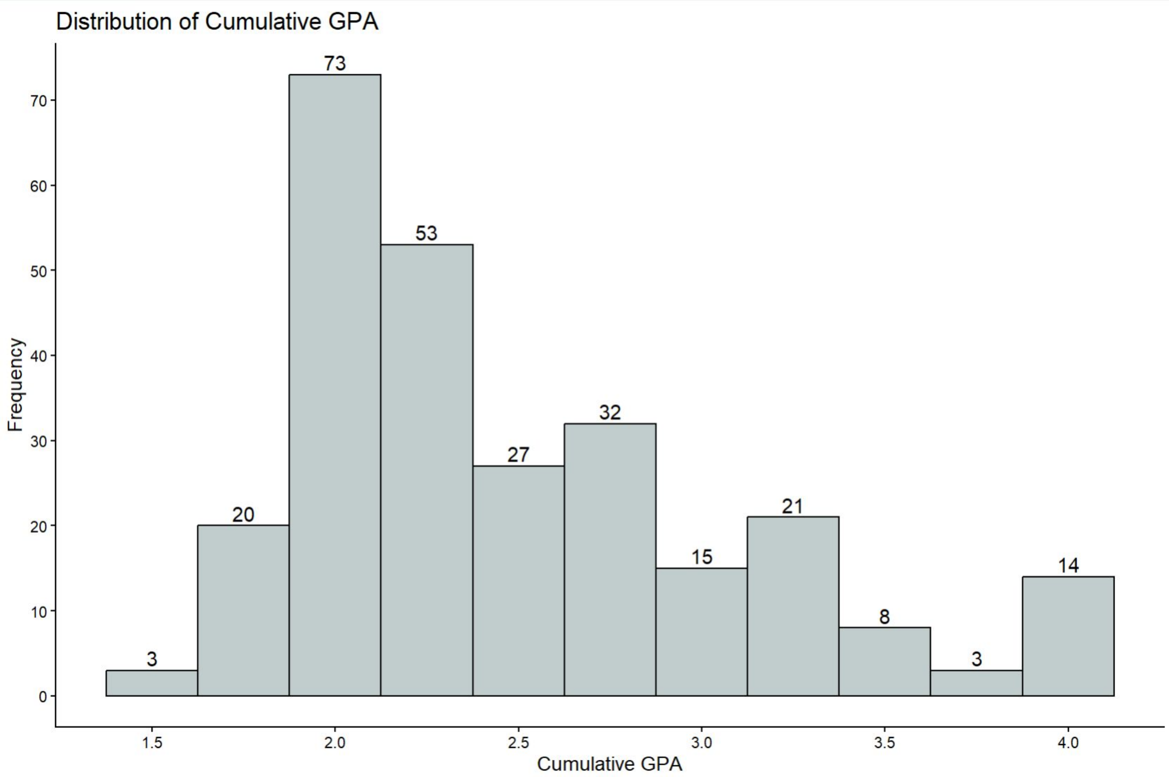
```
> str(Filtered_GPA_data)
tibble [269 × 15] (S3: tbl_df/tbl/data.frame)
 $ sat     : num [1:269] 920 780 960 820 820 730 780 830 710 980 ...
 $ tothrs  : num [1:269] 31 28 91 25 30 34 59 62 120 35 ...
 $ cumgpa  : num [1:269] 2.25 2.03 2.35 2.12 1.93 ...
 $ season  : Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 2 2 2 ...
 $ frstsem : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ crsgpa  : num [1:269] 2.65 2.87 2.85 2.63 2.61 ...
 $ verbmath: num [1:269] 0.484 0.814 1 0.783 0.907 ...
 $ trmgpa  : num [1:269] 1.5 2.2 2.8 1.5 2.5 ...
 $ hssize  : num [1:269] 10 123 383 344 228 482 78 196 300 152 ...
 $ hsrank  : num [1:269] 4 102 66 36 155 273 17 54 226 45 ...
 $ female  : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
 $ black   : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 2 2 ...
 $ white   : Factor w/ 2 levels "0","1": 1 2 2 1 2 1 2 1 1 1 ...
 $ hsperc  : num [1:269] 40 82.9 17.2 10.5 68 ...
 $ football: Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 2 2 ...
```

Identify Significant Variables

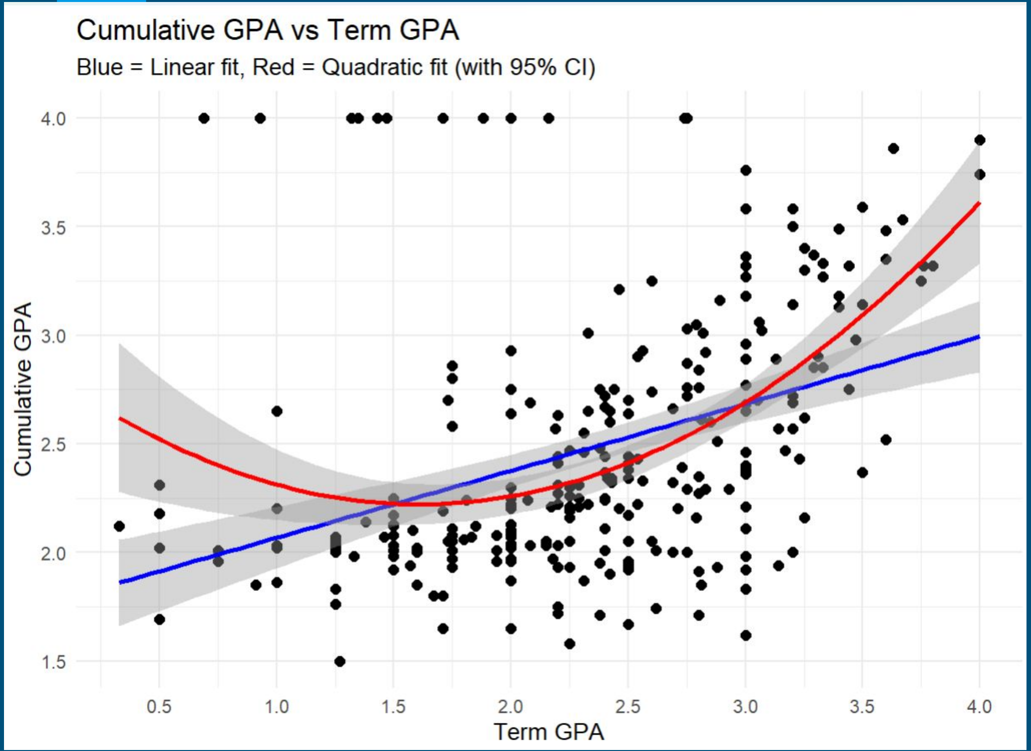


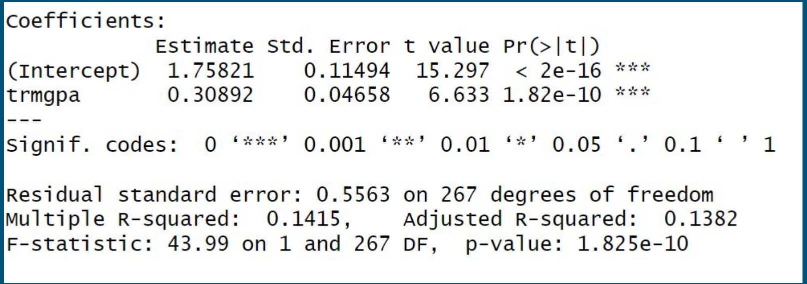
```
> summary(Everything_reg)

Call:
lm(formula = cumgpa ~ ., data = Master.GPA.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.28390 -0.31150 -0.02194  0.26339  1.78888 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.0621163  0.5772294   1.840 0.066927 .  
sat          0.0006074  0.0002410   2.520 0.012349 *  
tothrs      -0.0066636  0.0010888  -6.120 3.51e-09 ***
season      -0.0860712  0.0980053  -0.878 0.380645    
crsgpa       0.2174252  0.1696083   1.282 0.201033    
verbmath    -0.2787843  0.2067879  -1.348 0.178801    
trmgpa       0.2253863  0.0575872   3.914 0.000117 ***
hssize       0.0003838  0.0003066   1.252 0.211746    
hsrank      -0.0013537  0.0007361  -1.839 0.067058 .  
female       0.3074157  0.0878382   3.500 0.000549 ***
black        0.3445385  0.1922989   1.792 0.074370 .  
white        0.2342521  0.1829093   1.281 0.201462    
hsperc       0.0003432  0.0029561   0.116 0.907668    
football     0.3003614  0.0815191   3.685 0.000280 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4971 on 255 degrees of freedom
Multiple R-squared:  0.3453,	Adjusted R-squared:  0.3119 
F-statistic: 10.35 on 13 and 255 DF,  p-value: < 2.2e-16
```

Distribution of Cumulative GPA



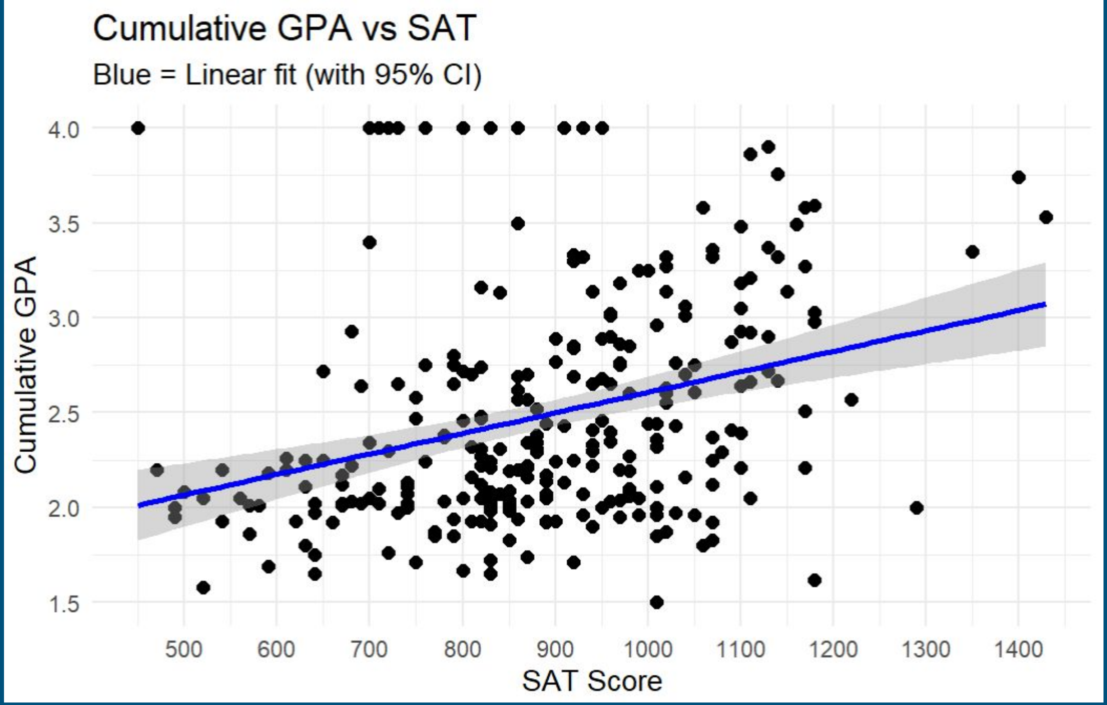
Simple Linear Regression - Cumulative GPA vs Term GPA





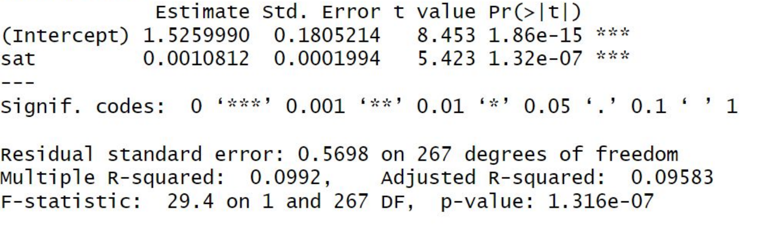
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.75821    0.11494  15.297  < 2e-16 ***
trmgpa       0.30892    0.04658   6.633 1.82e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5563 on 267 degrees of freedom
Multiple R-squared:  0.1415,	Adjusted R-squared:  0.1382
F-statistic: 43.99 on 1 and 267 DF,  p-value: 1.825e-10
```

Simple Linear Regression - Cumulative GPA vs SAT scores



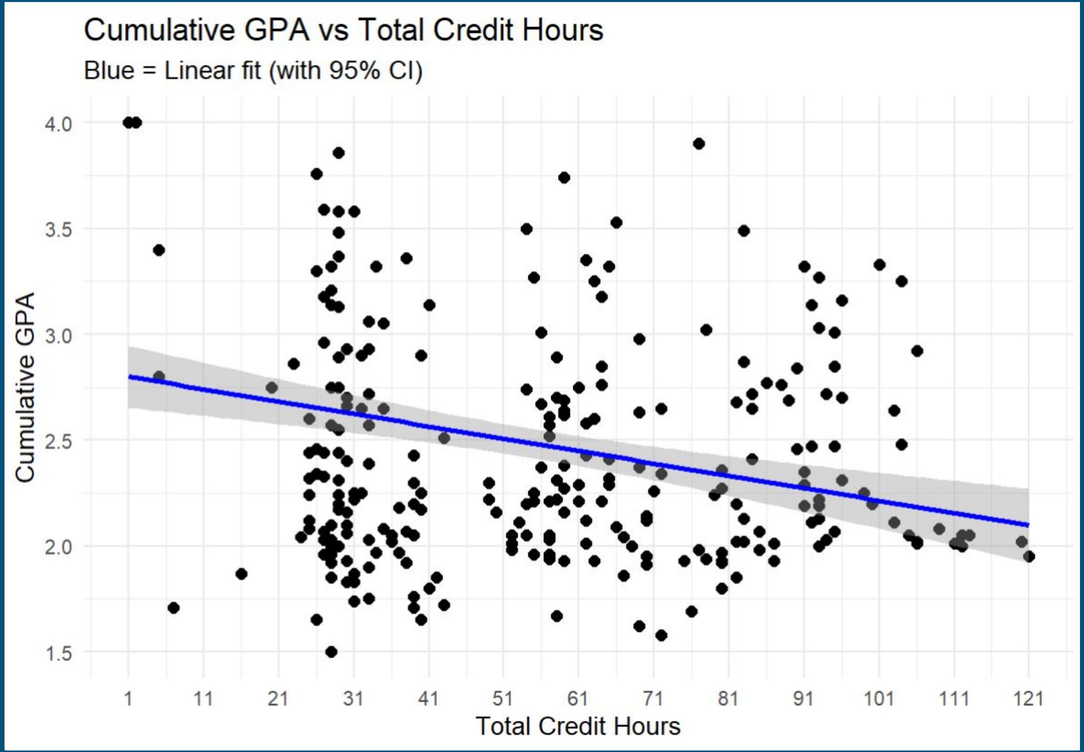
Cumulative GPA vs SAT

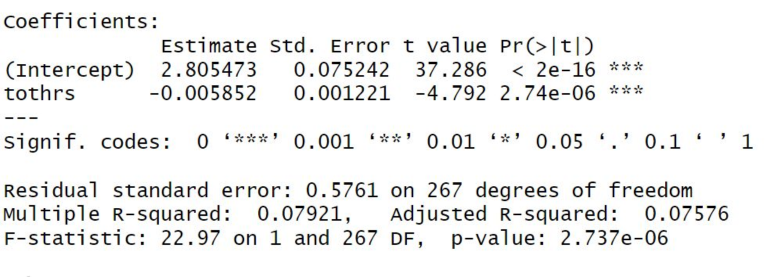
Blue = Linear fit (with 95% CI)



```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.5259990  0.1805214   8.453 1.86e-15 ***
sat         0.0010812  0.0001994   5.423 1.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5698 on 267 degrees of freedom
Multiple R-squared:  0.0992,	Adjusted R-squared:  0.09583 
F-statistic:  29.4 on 1 and 267 DF,  p-value: 1.316e-07
```

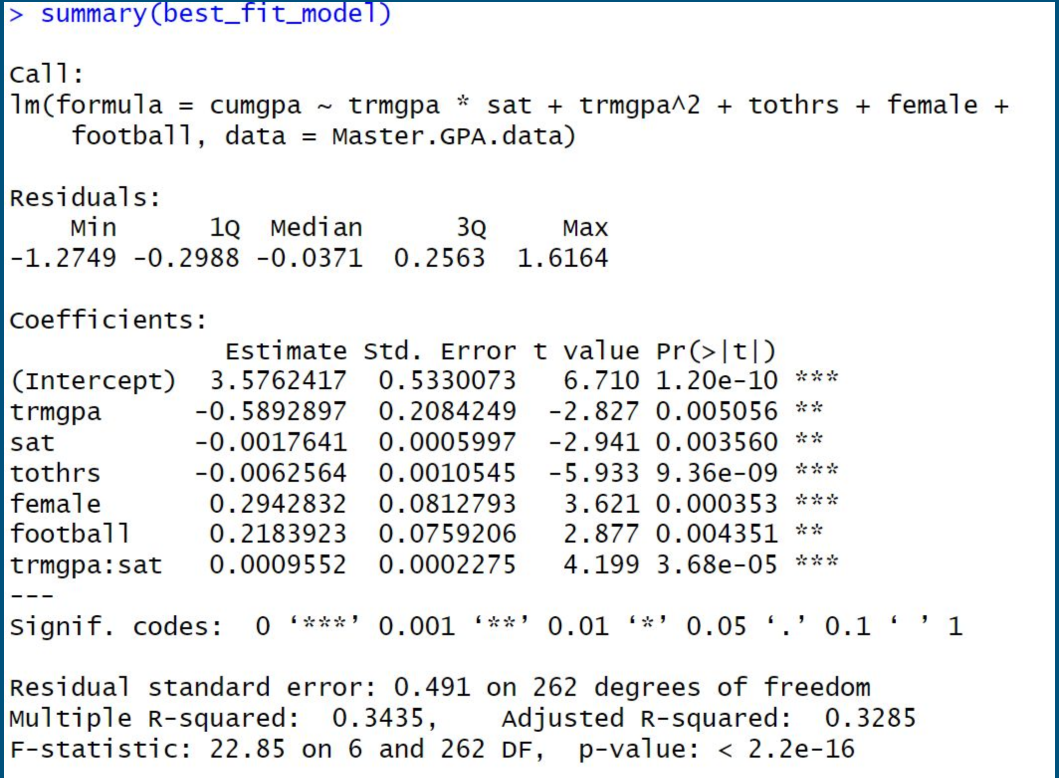
Simple Linear Regression - Cumulative GPA vs Total Credit hours





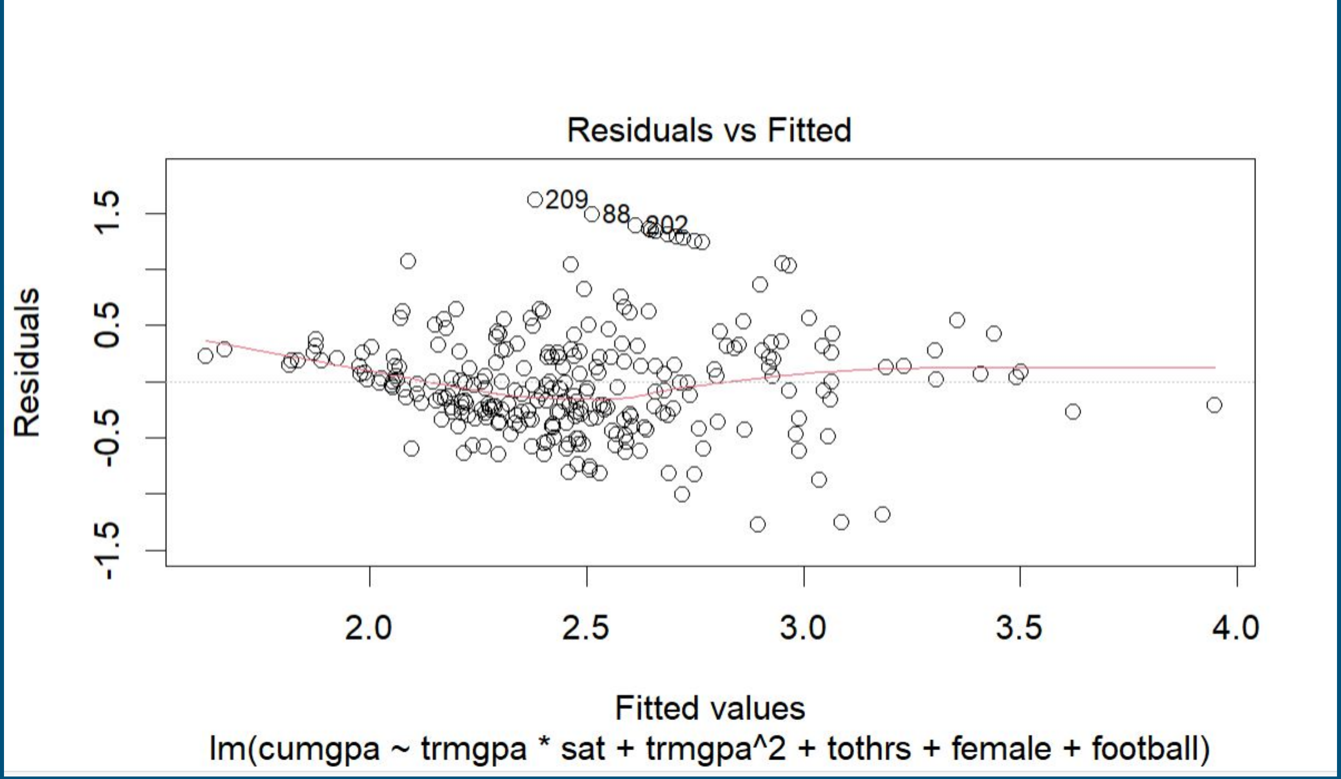
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.805473   0.075242  37.286  < 2e-16 ***
tothrs      -0.005852   0.001221  -4.792 2.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5761 on 267 degrees of freedom
Multiple R-squared:  0.07921,	Adjusted R-squared:  0.07576 
F-statistic: 22.97 on 1 and 267 DF,  p-value: 2.737e-06
```

Create our Best Fit Model

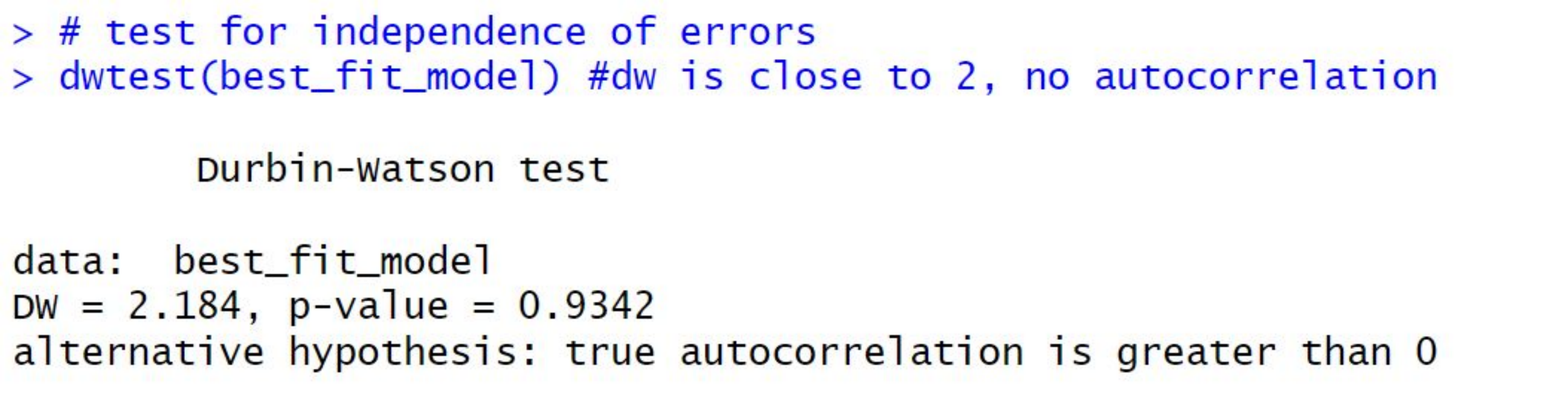


```
> summary(best_fit_model)

Call:
lm(formula = cumgpa ~ trmgpa * sat + trmgpa^2 + tothrs + female + 
    football, data = Master.GPA.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2749 -0.2988 -0.0371  0.2563  1.6164 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.5762417  0.5330073   6.710 1.20e-10 ***
trmgpa      -0.5892897  0.2084249  -2.827 0.005056 ** 
sat         -0.0017641  0.0005997  -2.941 0.003560 ** 
tothrs      -0.0062564  0.0010545  -5.933 9.36e-09 ***
female       0.2942832  0.0812793   3.621 0.000353 ***
football     0.2183923  0.0759206   2.877 0.004351 ** 
trmgpa:sat   0.0009552  0.0002275   4.199 3.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.491 on 262 degrees of freedom
Multiple R-squared:  0.3435,	Adjusted R-squared:  0.3285 
F-statistic: 22.85 on 6 and 262 DF,  p-value: < 2.2e-16
```

Test for Linearity of expectations

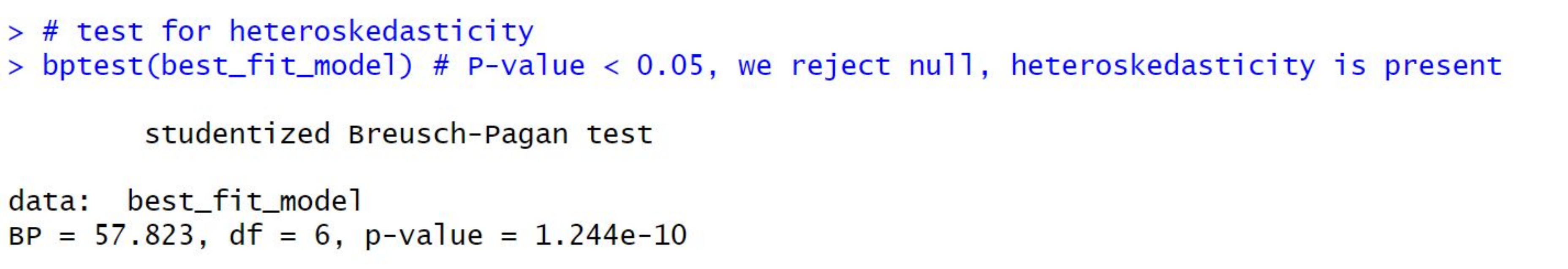


Test for Independence of Errors



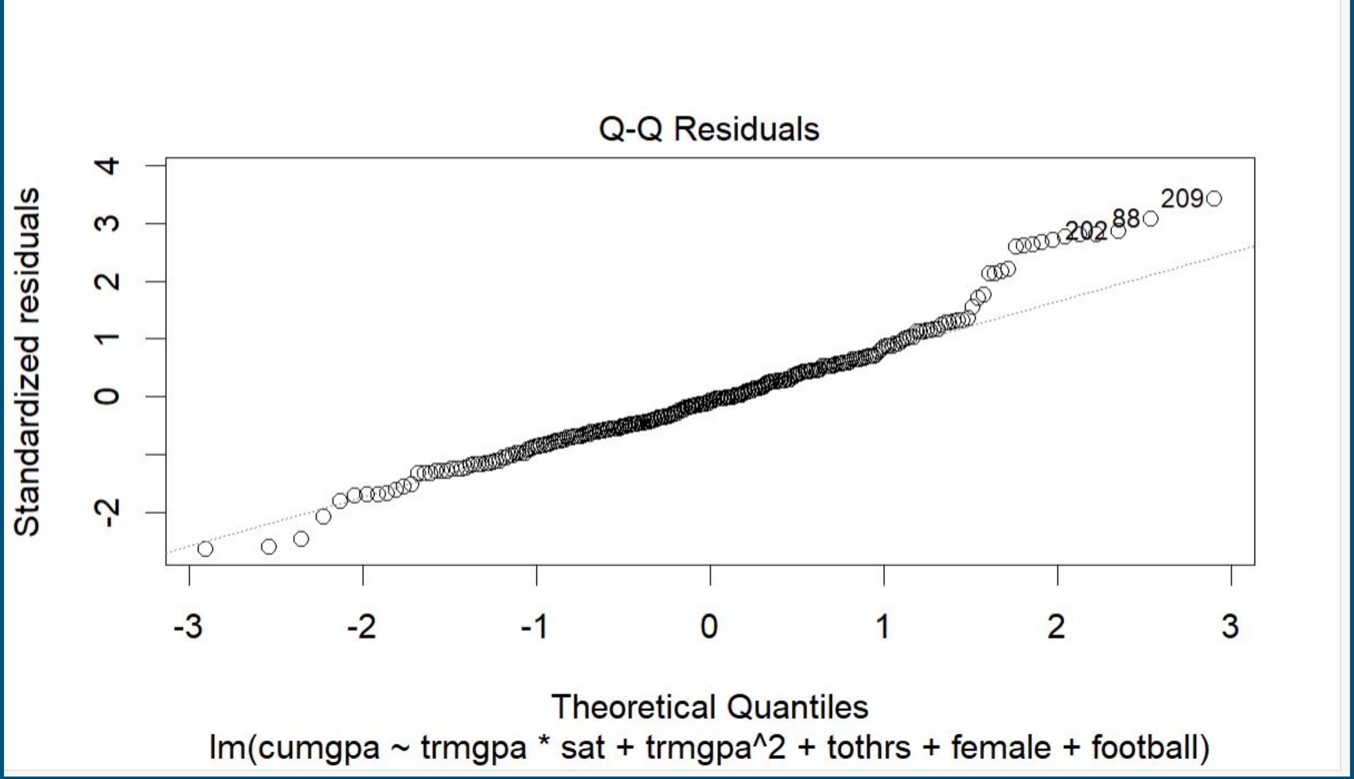
```
> # test for independence of errors
> dwtest(best_fit_model) #dw is close to 2, no autocorrelation

	Durbin-Watson test

data:  best_fit_model
DW = 2.184, p-value = 0.9342
alternative hypothesis: true autocorrelation is greater than 0
```

Test for Heteroskedasticity

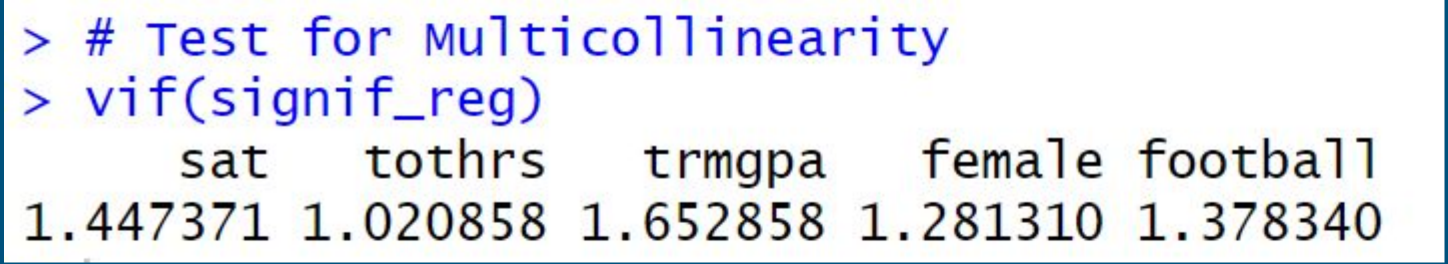


```
> # test for heteroskedasticity
> bptest(best_fit_model) # P-value < 0.05, we reject null, heteroskedasticity is present

	studentized Breusch-Pagan test

data:  best_fit_model
BP = 57.823, df = 6, p-value = 1.244e-10
```

Test for Normality



Test for Multicollinearity



```
> # Test for Multicollinearity
> vif(signif_reg)
     sat   tothrs   trmgpa   female football 
1.447371 1.020858 1.652858 1.281310 1.378340 
```

The best fit model, was not used to test multicollinearity

Conclusion/Next Steps/Recommendations

- Term GPA, SAT scores, and Total Credit Hours are the strongest predictors of cumulative GPA, with their effects varying across students, yet by themselves, they are not a strong predictor of the entire GPA.
- Determine the effects Gender and football has on GPA, then build models to visualize those effects.
- Collect more data/variables to build a better model, ex: Hours Spent studying, Time spent working, Household Income, Neighborhood location, etc...