Linear Regression with R

Certified Data Analytics (R) Capstone Project – Advanced Diploma in Data Analytics and Machine Learning

Singapore Management University | SMU Academy

WAYNE YONG, June 2020

Table of Contents

Problem Statement

College is an important rite of passage for many American students given its influence on students' career outlook in the future. High school, the education stage before college, supposedly prepares students for college by providing them the foundational background for courses in college.

While high schools are designed to prepare students for college, questions on whether it really does have surfaced in public discourse. The relevance of what's taught in high school to college has been put into question as critics lament that there's a mismatch in the skills taught in high school and the skills needed to succeed in college. Understanding then, the

impact of high school education on college educational attainment would be useful in assessing the effectiveness of high school education.

Therefore, I would be investigating the follow exploratory question:

1) How significant is education attainment in high school a good measure of student's preparedness for college?

install packages

install.packages("tidyverse")

install.packages("knitr")

install.packages("stargazer")

calling the functions

```
library(tidyverse)
library(gvlma)
library(car)
library(ggplot2)
library(stargazer)
library(effects)
```

import

```
getwd()

## [1] "C:/Users/Wayne Yong/Desktop/R Capstone Project (June 2020)"

data<-read.csv("FirstYearGPA.csv")
```

datasets

```
summary(data)

##        X              GPA            HSGPA            SATV
##  Min.   :  1.0   Min.   :1.930   Min.   :2.340   Min.   :260.0
##  1st Qu.: 55.5   1st Qu.:2.745   1st Qu.:3.170   1st Qu.:565.0
##  Median :110.0   Median :3.150   Median :3.500   Median :610.0
##  Mean   :110.0   Mean   :3.096   Mean   :3.453   Mean   :605.1
```

```
##  3rd Qu.:164.5   3rd Qu.:3.480   3rd Qu.:3.760   3rd Qu.:670.0
##  Max.   :219.0   Max.   :4.150   Max.   :4.000   Max.   :740.0
##       SATM            Male              HU              SS
##  Min.   :430.0   Min.   :0.0000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:580.0   1st Qu.:0.0000   1st Qu.: 8.00   1st Qu.: 3.000
##  Median :640.0   Median :0.0000   Median :13.00   Median : 6.000
##  Mean   :634.3   Mean   :0.4658   Mean   :13.11   Mean   : 7.249
##  3rd Qu.:690.0   3rd Qu.:1.0000   3rd Qu.:17.00   3rd Qu.:11.000
##  Max.   :800.0   Max.   :1.0000   Max.   :40.00   Max.   :21.000
##     FirstGen          White       CollegeBound
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:1.0000
##  Median :0.0000   Median :1.00   Median :1.0000
##  Mean   :0.1142   Mean   :0.79   Mean   :0.9224
##  3rd Qu.:0.0000   3rd Qu.:1.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00   Max.   :1.0000
```

```
glimpse(data)
```

```
## Rows: 219
## Columns: 11
## $ X            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
,...
## $ GPA          <dbl> 3.06, 4.15, 3.41, 3.21, 3.48, 2.95, 3.60, 2.87, 3.67,
...
## $ HSGPA        <dbl> 3.83, 4.00, 3.70, 3.51, 3.83, 3.25, 3.79, 3.60, 3.36,
...
## $ SATV         <int> 680, 740, 640, 740, 610, 600, 710, 390, 630, 680, 380
,...
## $ SATM         <int> 770, 720, 570, 700, 610, 570, 630, 570, 560, 670, 470
,...
## $ Male         <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
...
## $ HU           <dbl> 3.0, 9.0, 16.0, 22.0, 30.5, 18.0, 5.0, 10.0, 8.5, 16.
0...
## $ SS           <dbl> 9.0, 3.0, 13.0, 0.0, 1.5, 3.0, 19.0, 0.0, 15.5, 12.0,
...
## $ FirstGen     <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
...
## $ White        <int> 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1,
...
## $ CollegeBound <int> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
...
```

The data frame with 219 observations on the following 12 variables.

GPA || First-year college GPA on a 0.0 to 4.0 scale

HSGPA || High school GPA on a 0.0 to 4.0 scale

SATV || Verbal/critical reading SAT score

SATM || Math SAT score

Male || 1 = male, 0 = female

HU || Number of credit hours earned in humanities courses in high school

SS || Number of credit hours earned in social science courses in high school

FirstGen || 1 = student is the first in her or his family to attend college, 0 =otherwise

White || 1 = white students, 0 = others

CollegeBound || 1 =attended a high school where >=50% students intended to go on to college, 0 =otherwise

tidy

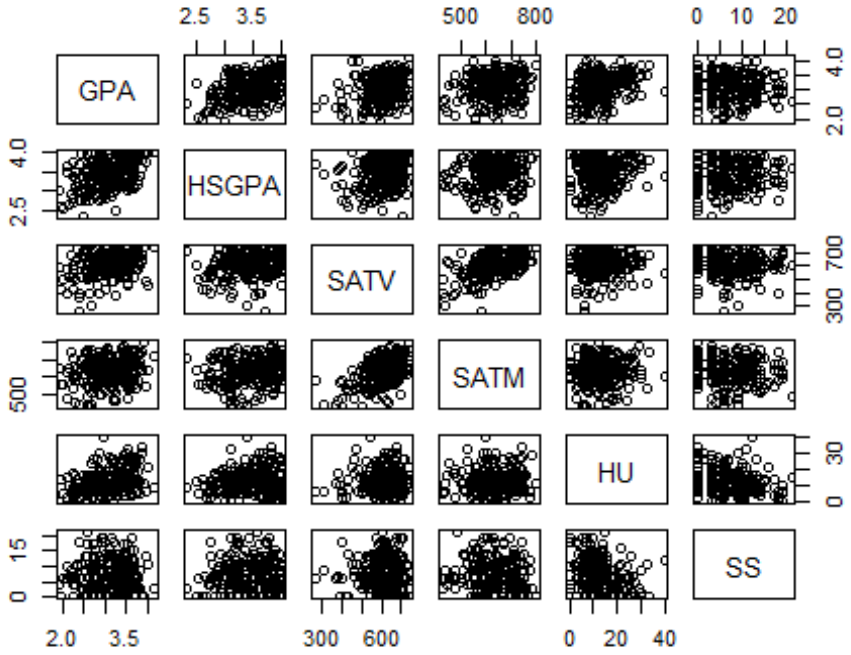
```
data<-data %>%
  #remove X variable from the data frame
  select(-X)

data$Male<-as.factor(data$Male)
data$FirstGen<-as.factor(data$FirstGen)
data$White<-as.factor(data$White)
data$CollegeBound<-as.factor(data$CollegeBound)
```

correlation matrix

```
cor(data[,c(1,2,3,4,6,7)])
```

```
##               GPA       HSGPA        SATV         SATM           HU
SS
## GPA    1.00000000  0.44688735  0.30431137  0.194343851  0.314655754 -0.003
56909
## HSGPA  0.44688735  1.00000000  0.21032124  0.152839634  0.116031169 -0.017
25443
## SATV   0.30431137  0.21032124  1.00000000  0.526943819  0.098748556 -0.026
46987
## SATM   0.19434385  0.15283963  0.52694382  1.000000000 -0.009601549 -0.087
83997
## HU     0.31465575  0.11603117  0.09874856 -0.009601549  1.000000000 -0.306
60787
## SS    -0.00356909 -0.01725443 -0.02646987 -0.087839974 -0.306607866  1.000
00000
```

```
plot(data[,c(1,2,3,4,6,7)])
```



regression model

From the correlation matrix and regression "model1", I can understand from the value that HSGPA and HU are the only continuous variables with at least moderately high correlation coefficient with the dependent variable (GPA).

```
model1<-lm(GPA~.,data=data)
#base R
summary(model1)

##
## Call:
## lm(formula = GPA ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07412 -0.25827  0.05384  0.27675  0.85761
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.5268983  0.3487584   1.511  0.13235
## HSGPA          0.4932945  0.0745553   6.616 3.03e-10 ***
## SATV           0.0005919  0.0003945   1.501  0.13498
## SATM           0.0000847  0.0004447   0.190  0.84912
```

```
## Male1          0.0482478  0.0570277   0.846  0.39850
## HU             0.0161874  0.0039723   4.075 6.53e-05 ***
## SS             0.0073370  0.0055635   1.319  0.18869
## FirstGen1     -0.0743417  0.0887490  -0.838  0.40318
## White1         0.1962316  0.0700182   2.803  0.00555 **
## CollegeBound1  0.0214530  0.1003350   0.214  0.83090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3834 on 209 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3216
## F-statistic: 12.48 on 9 and 209 DF,  p-value: 8.674e-16
```

Removing variables not related to high school education (eg. SATV and SATM scores).

```
model2<-lm(GPA~HSGPA+HU+White,data = data)
#base R
summary(model2)

##
## Call:
## lm(formula = GPA ~ HSGPA + HU + White, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09479 -0.27638  0.02287  0.25411  0.84538
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.933459   0.245673   3.800 0.000189 ***
## HSGPA       0.507404   0.070197   7.228 8.42e-12 ***
## HU          0.015328   0.003667   4.180 4.24e-05 ***
## White1      0.265644   0.064519   4.117 5.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3856 on 215 degrees of freedom
## Multiple R-squared:  0.3231, Adjusted R-squared:  0.3136
## F-statistic: 34.21 on 3 and 215 DF,  p-value: < 2.2e-16
```

For multiple linear regression, a few things need to be checked before I use the regression results to predict the dependent variable from the independent ones. These few things being satisfied, indicates the regression model I used is a good fit.

The points to check are:

1. The relationship between the independent variables (HSGPA) and the dependent variable (GPA) is linear. (ggplot section)

2. Homoscedasticity, ie. the error terms have equal variance.

3. The residuals are normally distributed.
4. There are no outliers affecting the regression results.
5. There is no multicollinearity, ie. there is no correlation among the independent variables.

The graphs plotted below will serve to check these 5 points.

Firstly, to check the two points below, I plot residuals against fitted values.

Homoscedasticity, ie. the error terms have equal variance.

If the scatter plot has all the points scattered evenly about the "residual = 0" line, then these 2 points are satisfied. The plot generated fulfils this.

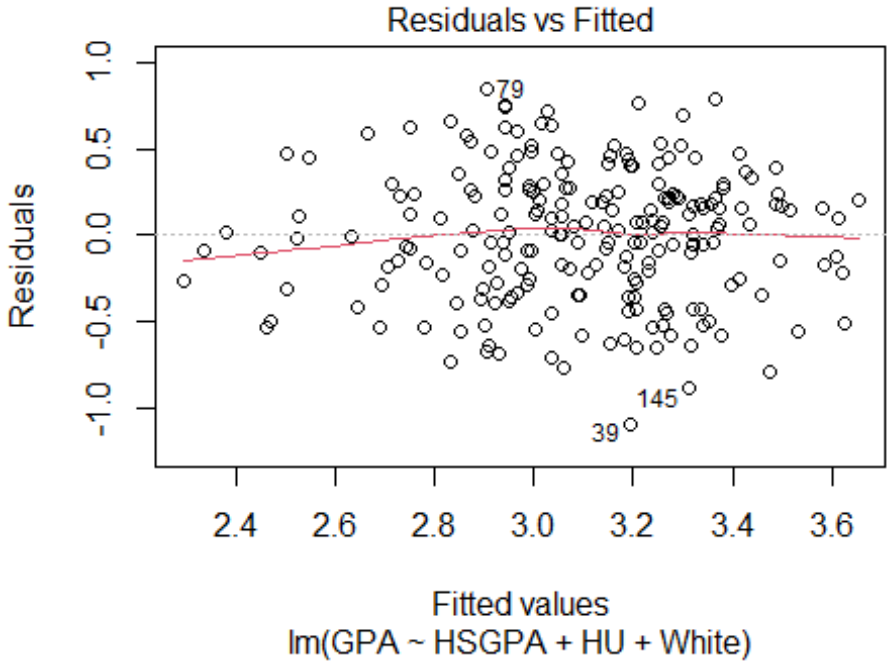
To check whether the residuals are normally distributed, I use QQ-plot. If the QQ-plot is very close to a straight line, then the residuals are approximately normally distributed. The QQ-plot generated is almost a straight line, so this point is satisfied too.

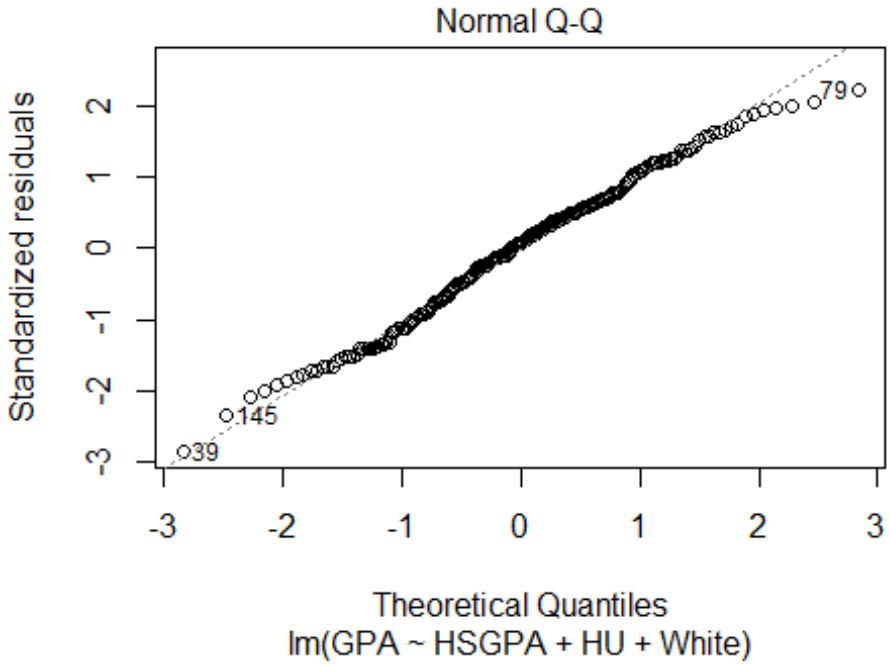
To check that, there are no outliers affecting the regression results, I plot residuals against leverage. All the points in the plot should lie between the dotted curves indicating the Cook distance. The dotted curves are too far away from the points in the plot I generated (the curves don't even show up in the plot), so this point is met.

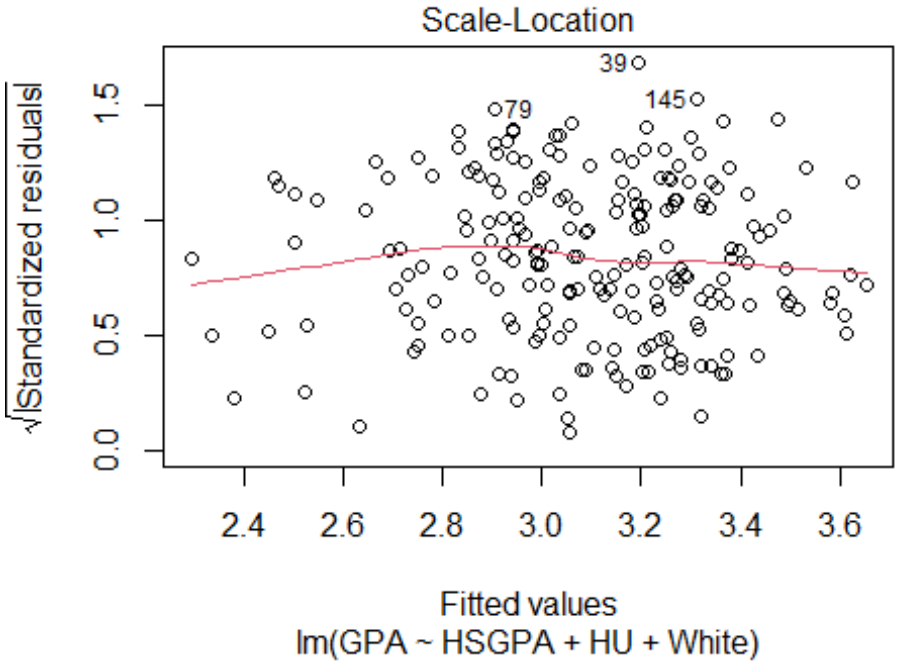
```
anova(model1, model2)
```

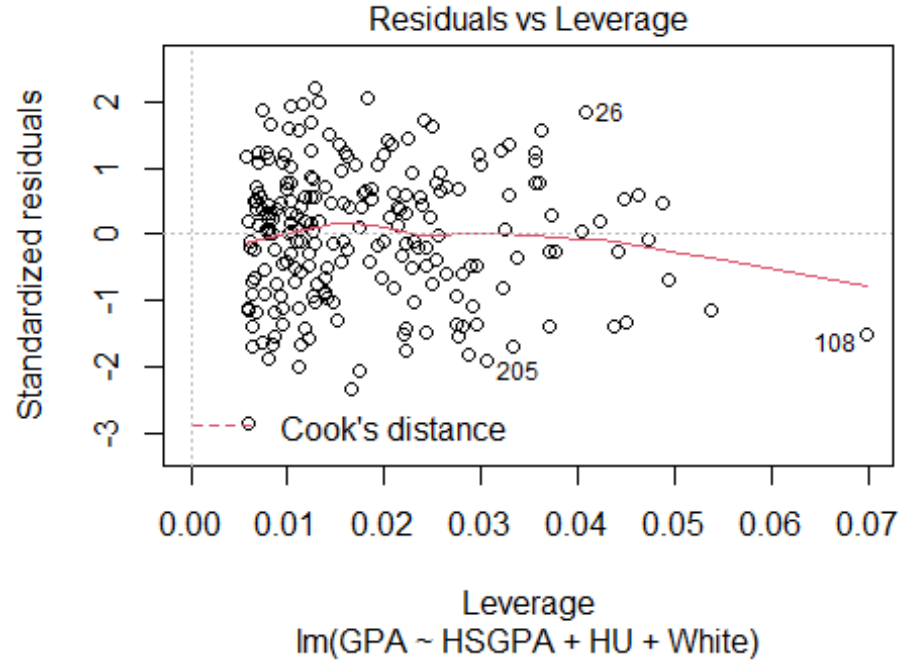
```
## Analysis of Variance Table
##
## Model 1: GPA ~ HSGPA + SATV + SATM + Male + HU + SS + FirstGen + White +
##     CollegeBound
## Model 2: GPA ~ HSGPA + HU + White
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    209 30.719
## 2    215 31.973 -6   -1.2539 1.4218 0.2075
```

```
plot(model2)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
-0.5
-1.0
79
145
39
2.4
2.6
2.8
3.0
3.2
3.4
3.6
Fitted values
lm(GPA ~ HSGPA + HU + White)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
79
145
39
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(GPA ~ HSGPA + HU + White)





assumptions met

```
library(gvlma)

gvlma(model2)

##
## Call:
## lm(formula = GPA ~ HSGPA + HU + White, data = data)
##
## Coefficients:
## (Intercept)        HSGPA           HU       White1
##     0.93346      0.50740      0.01533      0.26564
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model2)
##
##                       Value p-value                Decision
## Global Stat        4.827e+00  0.3055 Assumptions acceptable.
## Skewness           9.002e-01  0.3427 Assumptions acceptable.
## Kurtosis           2.342e+00  0.1259 Assumptions acceptable.
## Link Function      1.585e+00  0.2080 Assumptions acceptable.
## Heteroscedasticity 2.418e-05  0.9961 Assumptions acceptable.
```

no multicollinearity

```
library(car)

vif(model2)

##    HSGPA       HU    White
## 1.014680 1.028845 1.017150
```

To check that there is no multicollinearity (there is no correlation among the independent variables), I use variance inflation factor (VIF) computations. The VIF results are all way less than 10, so multicollinearity is not a concern.

So from what I have done and the results obtained, it is reasonable to believe that the regression model is a good fit. I can now use the model to predict GPA from HSGPA + HU + White.

Plotting Simple Slopes: Standard Deviation

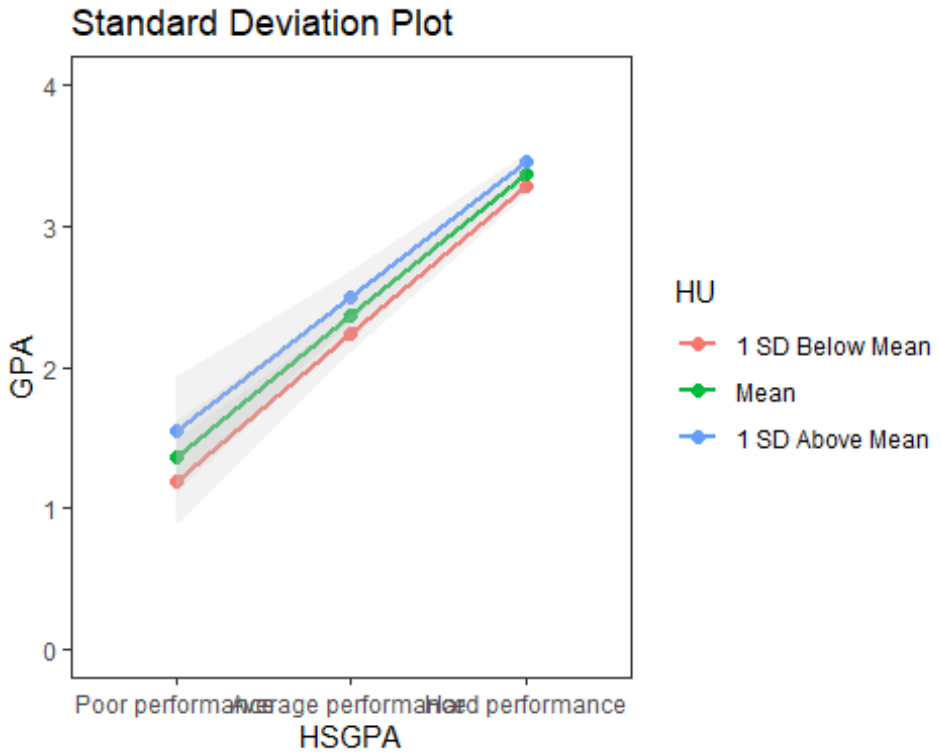
```
library(ggplot2)
library(stargazer)
library(effects)
```

```
model3<-lm(GPA~HSGPA*HU+White,data = data)

#Create a new variable for IQ based on the actual mean/standard deviation in
the data set
HU.SD <- c(mean(data$HU)-sd(data$HU),
           mean(data$HU),
           mean(data$HU)+sd(data$HU))

HU.SD <- round(HU.SD, 2)
HU.SD

## [1]  5.88 13.11 20.33
```

The plot below shows the standard deviation interactions with HSGPA and HU. The graph shows that 1 SD below mean has a much lower GPA as compared to 1 SD above mean. On the other hand, it can be seen that the GPA score converges as HSGPA score increases despite the HU is 1 SD below or 1 SD above mean.

```
Inter.SD <- effect(c("HSGPA*HU"), model3,
                   xlevels=list(HU=c(6.88,13.11,19.33),
                                HSGPA = c(0,2,4)))

Inter.SD <- as.data.frame(Inter.SD)

# Create factors of the different variables for the interaction

Inter.SD$HU<-factor(Inter.SD$HU,
                    levels=c(6.88, 13.11, 19.33),
                    labels=c("1 SD Below Mean", "Mean", "1 SD Above Mean"))

Inter.SD$HSGPA<-factor(Inter.SD$HSGPA,
                           levels=c(0, 2, 4),
                           labels=c("Poor performance", "Average performance
", "Hard performance"))

# Plot
Plot.SD<-ggplot(data=Inter.SD, aes(x=HSGPA, y=fit, group=HU))+
  geom_line(size=1, aes(color=HU), size= 3) +
  geom_point(aes(colour = HU), size=2) +
  geom_ribbon(aes(ymin=fit-se, ymax=fit+se),fill="gray",alpha=.2)+
  ylim(0,4)+
  ylab("GPA")+
  xlab("HSGPA")+
  ggtitle("Standard Deviation Plot")+
  theme_bw()+
   theme(panel.grid.major=element_blank(),
         panel.grid.minor=element_blank(),
         legend.key = element_blank())+
         scale_fill_grey()
```

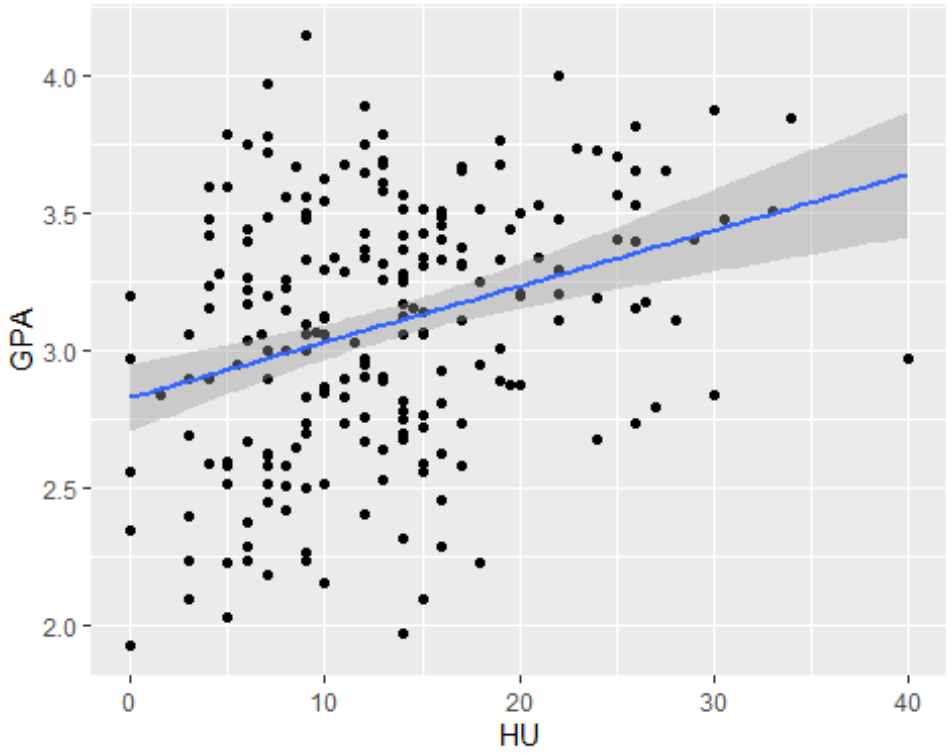
```
Plot.SD
```



Visualize

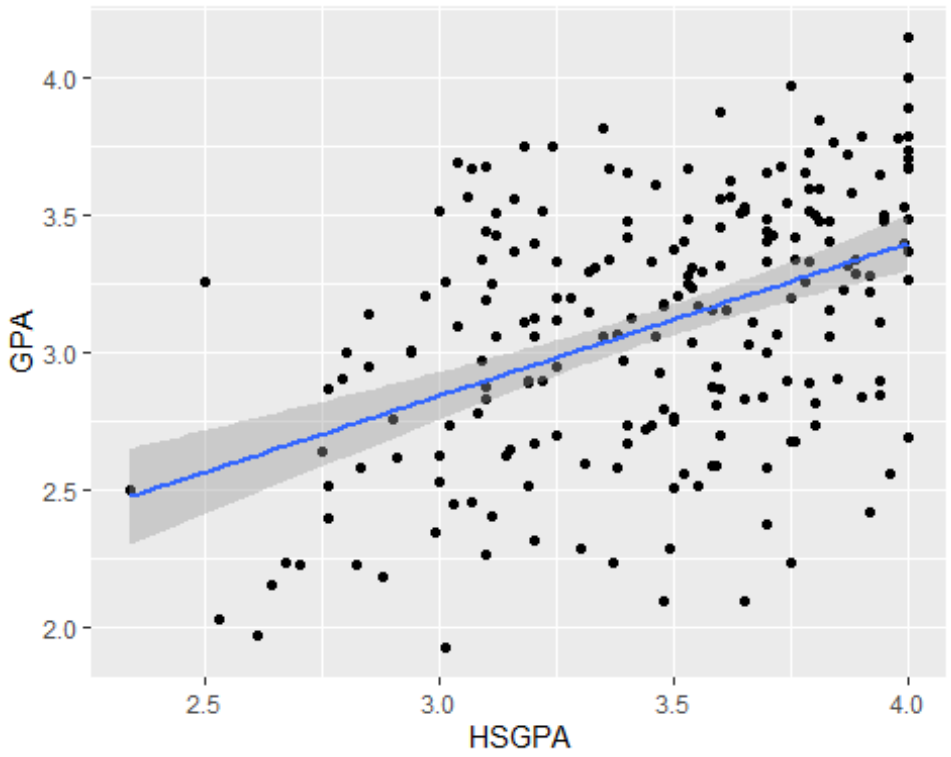
The graph shows that the relationship between the independent variables (GPA) and the dependent variable (HU), according to my linear model explained above.

```
ggplot(data,aes(HU,GPA))+geom_point()+geom_smooth(method = "lm")
```

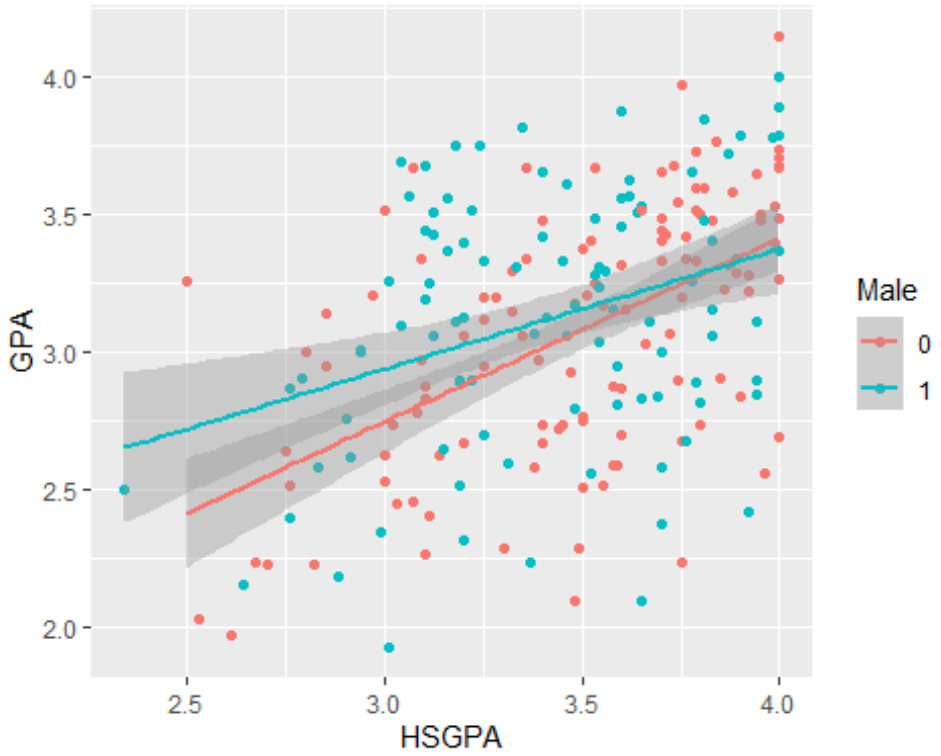


At the onset, plotting first year college GPA (GPA) against High school GPA (HSGPA) will give us a general understanding of the correlation between high school attainment and college attainment. From the scatter plot, we see a positive linear correlation between HSGPA and GPA. Calculating the correlation coefficient (r), r = 0.47, it can be concluded that there's a moderate positive linear correlation between HSGPA and GPA. This suggest that, students who do well in high school are also likely to do well in college.

```
ggplot(data,aes(HSGPA,GPA))+geom_point()+geom_smooth(method = "lm")
```

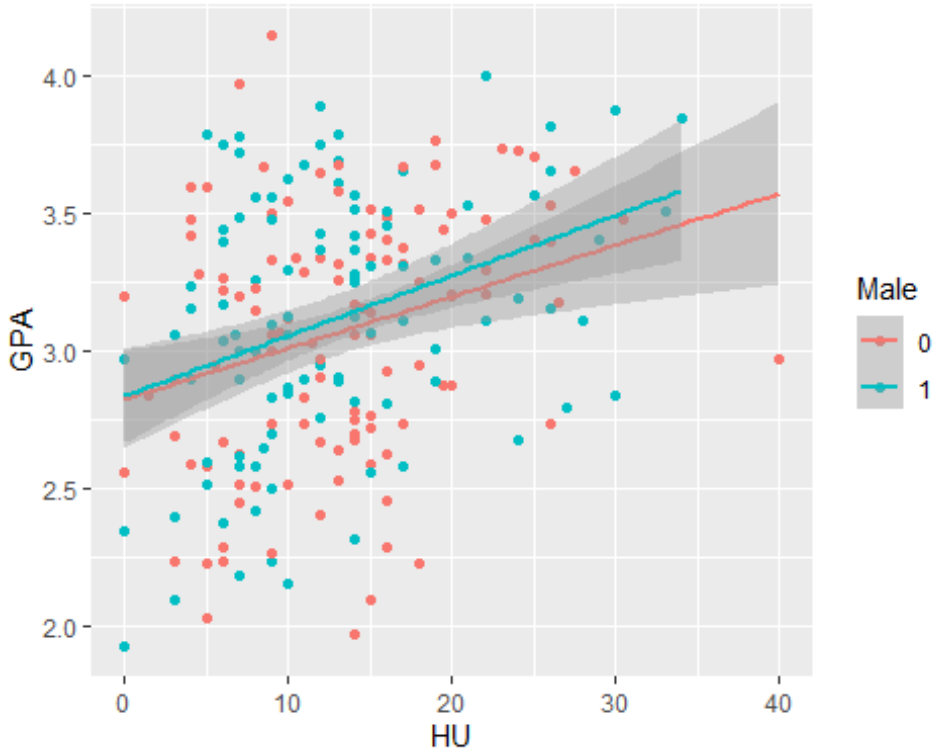



```
ggplot(data,aes(HSGPA,GPA,color=Male))+geom_point()+geom_smooth(method = "lm"
)
```



There is also a moderate positive correlation between GPA and HU, given the correlation coefficient r = 0.31. This shows that students who have earned transferable credit hours in humanities courses in high school tend to do better in college. This suggests that a richer academic high school experience may contribute to better education attainment in college.

```
ggplot(data,aes(HU,GPA,color=Male))+geom_point()+geom_smooth(method = "lm")
```



Implications

HSGPA and HU are factors defining the high school experience. Based on their moderately strong correlation with first year college GPA, we can conclude that indeed high school educational attainment can be useful in predicting college attainment. Academics who wish to investigate factors that affect student's performance in college can therefore take a closer look at their high school experience as a point of study.

Limitation

While my project has shown the correlation between high school attainment and college attainment, it cannot be implied that doing well in high school does enable / cause students to do well in college. That is, correlation does not equal causation. This is because confounders in this correlation has not been considered. There are other factors such as cognitive abilities of students and socio-economic background of students that can potentially have correlations with both high school attainment and college attainment.

These factors have to be considered before we can make a more reasoned argument that education attainment in high school a good measure of student's preparedness for college. Exploring the correlation between these confounding factors and college GPA can be a worthwhile data science investigation. It would be useful in shedding greater clarity to the correlation between high school education attainment and first year college GPA, helping academics better predict college performance from high school experiences.