Yan Zhang's Regression GPA Prediction Project

Yan Zhang

October 6, 2015

Simple Linear Regression Model:

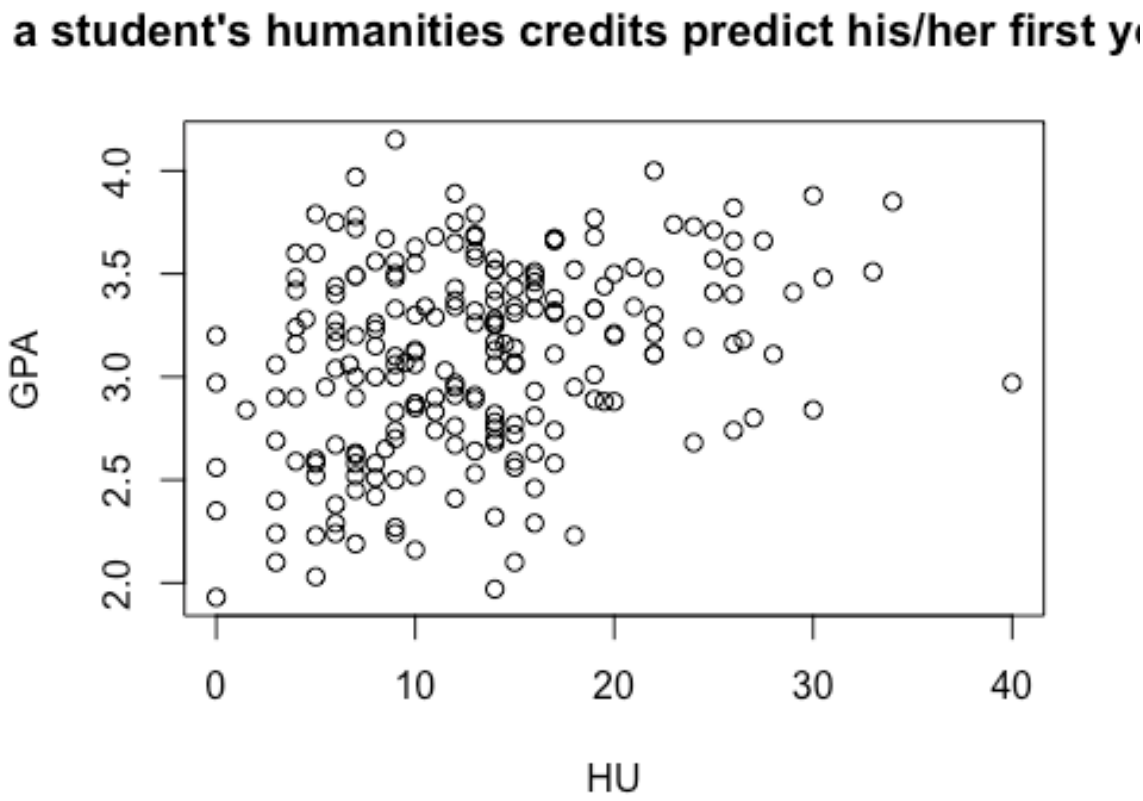
First step is to load my data from the package "Stat2Data":

```
library(Stat2Data)
data(FirstYearGPA)
head(FirstYearGPA)
```

```
##    GPA HSGPA SATV SATM Male   HU   SS FirstGen White CollegeBound
## 1 3.06  3.83  680  770    1  3.0  9.0        1     1             1
## 2 4.15  4.00  740  720    0  9.0  3.0        0     1             1
## 3 3.41  3.70  640  570    0 16.0 13.0        0     0             1
## 4 3.21  3.51  740  700    0 22.0  0.0        0     1             1
## 5 3.48  3.83  610  610    0 30.5  1.5        0     1             1
## 6 2.95  3.25  600  570    0 18.0  3.0        0     1             1
```

Second step, I am going to use a Simple Linear Regression model to compare the student's number of humanities credits (HU) to predict the student's first year GPA (GPA):

```
plot(GPA ~ HU, main = "Can a student's humanities credits predict his/her fir
st year GPA", data = FirstYearGPA)
```



From the plot, the plot has a lot of noise but overall they look they are lie on a straight line, so I am going to check normallity and if the data is constant variance afterwards.

Here is my hypotheses:

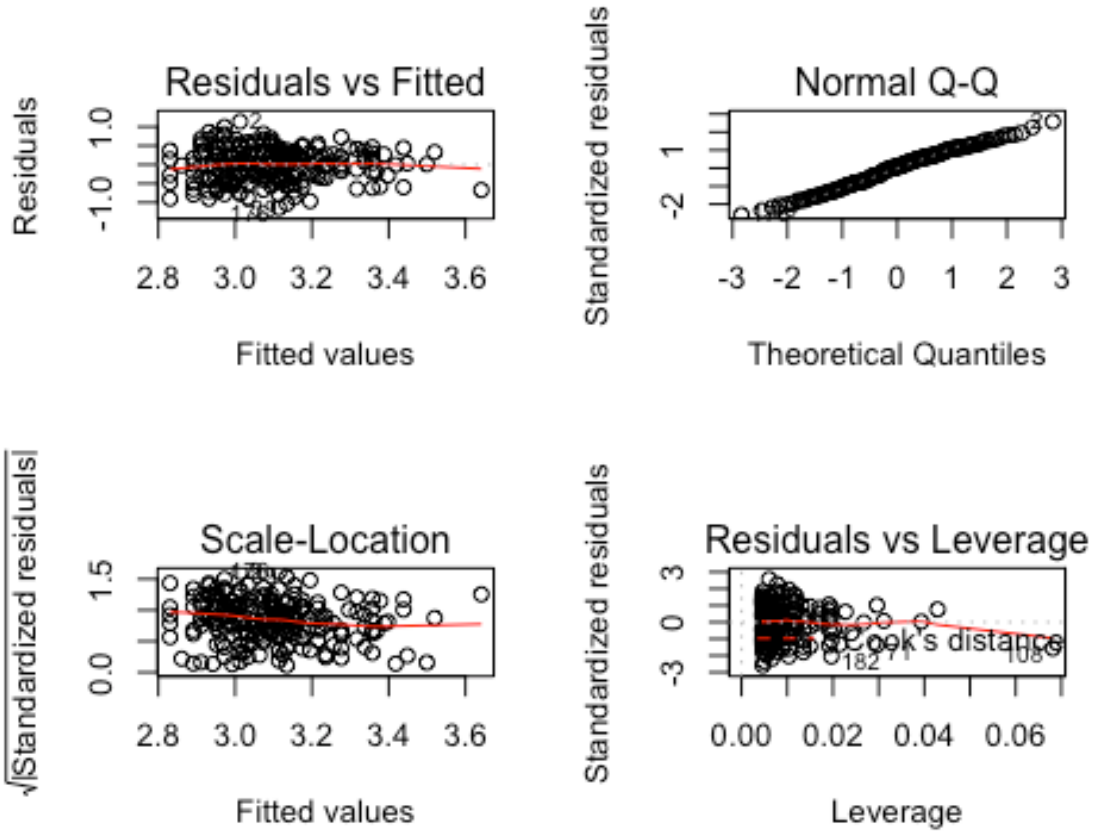
$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

```
lm(m1 <- lm(GPA ~ HU, data = FirstYearGPA))

##
## Call:
## lm(formula = m1 <- lm(GPA ~ HU, data = FirstYearGPA))
##
## Coefficients:
## (Intercept)           HU
##     2.83042      0.02027
```

Predicted GPA = 2.83042 + 0.02027*HU

```
par(mfrow=c(2,2))
plot(m1)
```




```
par(mfrow=c(1,1))
```

The residual line is almost with the 0 line, we roughly have a constance variance. From QQ plot, we can see the plot is almost a linear line so we have a normally distributed data. I assume this data is random so the data is independence.

Since we have done the assumption check, I am going to run the two-sample F test code in R.

```
summary(m1)
```

```
##
## Call:
## lm(formula = GPA ~ HU, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14424 -0.33829  0.03125  0.31753  1.13712
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.830423   0.062105  45.575  < 2e-16 ***
## HU          0.020273   0.004152   4.883 2.02e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4428 on 217 degrees of freedom
## Multiple R-squared:  0.09901,    Adjusted R-squared:  0.09486
## F-statistic: 23.85 on 1 and 217 DF,  p-value: 2.023e-06
```

From the F test above, we can see, the p value is 2.023e-06. The p value is much small than 0.05 so we reject null hypothsis, so we have enough evidence to say two data sets are depedent to each other. R^2 = 0.09901, so 9.9% of variability in humanity is explained by using a simple linear model with first year GPA as the predictor, so humanity is not a good predictor.

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value    Pr(>F)
## HU          1  4.677  4.6765  23.846 2.023e-06 ***
## Residuals 217 42.557  0.1961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mean(FirstYearGPA$GPA)
```

```
## [1] 3.096164
```

```
sd(FirstYearGPA$GPA)
```

```
## [1] 0.4654759
```

```
qt(0.975,217)
```

```
## [1] 1.970956
```

$\hat{y}$ = 2.83042 + 0.02027*16 = 3.15474

SE = 0.4428*sqrt(1 + ((1/219)+((16-3.096164)^2)/((219-1)*(0.4654759^2)))) = 0.9424247

95% CI = 3.15474 + 1.970956 * 0.9424247 = 5.012218 and CI = 3.15474 - 1.970956 * 0.9424247 = 1.297262

The Multiple Regression model:

I am going to use a Multiple Linear Regression model to find out the best model to predict the student's first year GPA (GPA) with reponse variables Humanities credits (HU), Highschool GPA (HSGPA) and whether or not the student was male (Male).

I firstly set up one null model and one full model for the next modeling selection.

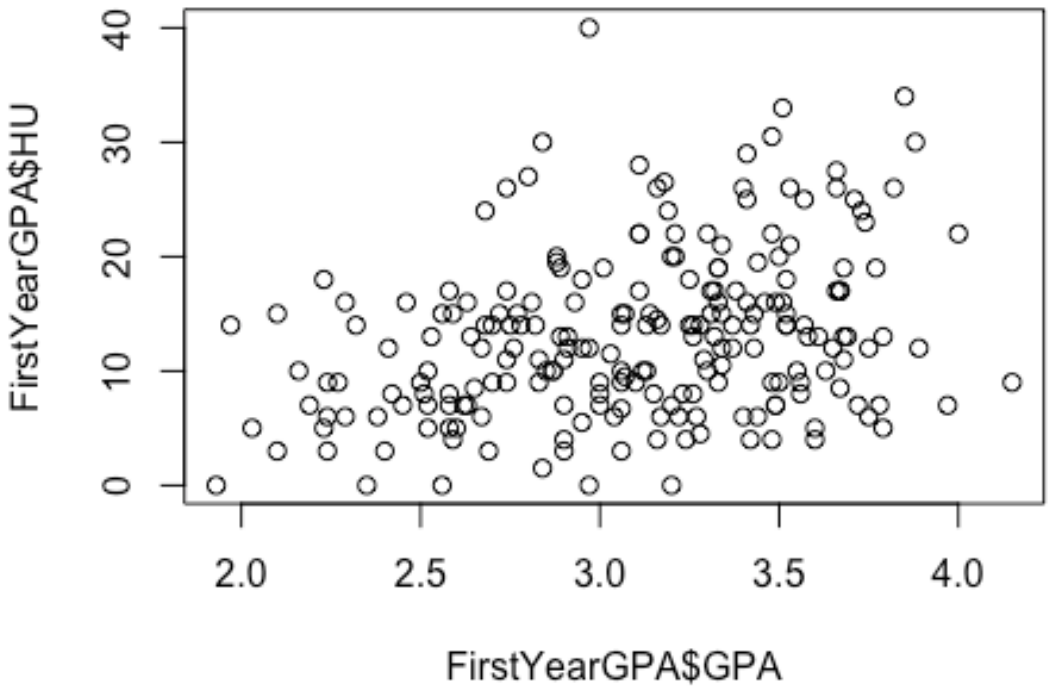
```r
library(MASS) # the package needed to do model selection
modnull <- lm(GPA~1, data=FirstYearGPA) # null model
summary(modnull)
```

```
##
## Call:
## lm(formula = GPA ~ 1, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16616 -0.35116  0.05384  0.38384  1.05384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.09616    0.03145   98.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4655 on 218 degrees of freedom
```

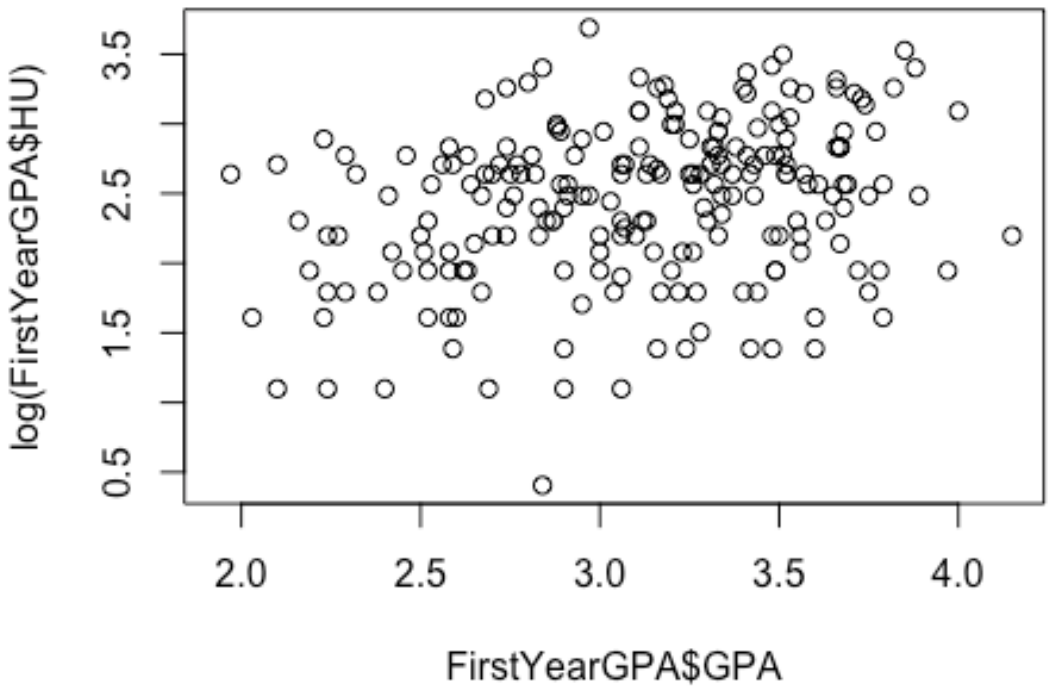
```r
modfull <- lm(GPA~HU+HSGPA+as.factor(Male), data=FirstYearGPA) # full model
summary(modfull)
```

```
##
## Call:
## lm(formula = GPA ~ HU + HSGPA + as.factor(Male), data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00138 -0.28011  0.04876  0.26567  0.87772
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.008136   0.255149   3.951 0.000105 ***
## HU               0.017216   0.003757   4.583 7.78e-06 ***
## HSGPA            0.527301   0.072702   7.253 7.28e-12 ***
## as.factor(Male)1 0.089592   0.054141   1.655 0.099427 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.398 on 215 degrees of freedom
## Multiple R-squared:  0.2789, Adjusted R-squared:  0.2688
## F-statistic: 27.72 on 3 and 215 DF,  p-value: 3.404e-15
```

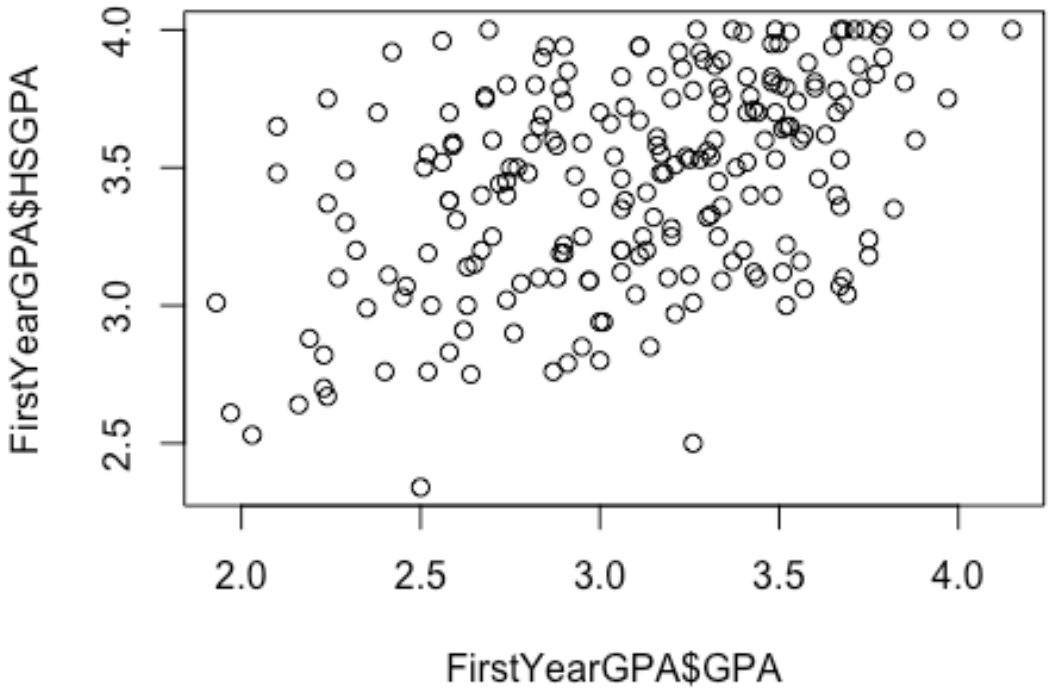
```r
plot(FirstYearGPA$GPA,FirstYearGPA$HU)
```



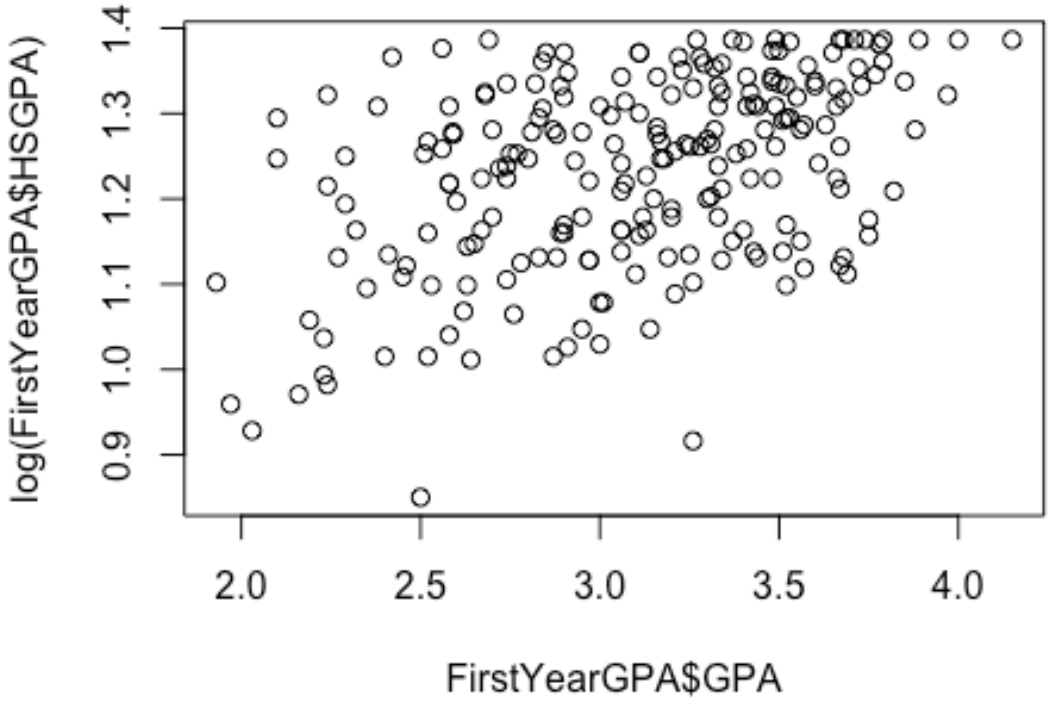

```
plot(FirstYearGPA$GPA,log(FirstYearGPA$HU))
```




```
plot(FirstYearGPA$GPA,FirstYearGPA$HSGPA)
```




```
plot(FirstYearGPA$GPA,log(FirstYearGPA$HSGPA))
```



Before modeling selection, I checked two scatterplot GPA~HU and GPA~HSGPA and since Male is a dummy variable so I don't need to think about transformation for it. From scatterplots above, this data seems uniformly distributed and there aren't strong curvature and extreme values, although there are more dots in the middle in the GPA~HU plot, but overall these two plots don't look exponential. Just in case, I still did a log transfiormation but plots didn't change so much and the linearity still look similar, so I think this data doesn't need to use log transformation.

After I decided I will not conduct a log trnasformation, I will just use my pervious null model and one full model to do the modeling selection:

```
# forward selection, begin with the null model
step(modnull, scope=list(lower=modnull, upper=modfull), direction="forward")
```

```
## Start:  AIC=-333.94
## GPA ~ 1
##
##                   Df Sum of Sq    RSS     AIC
## + HSGPA            1    9.4329 37.801 -380.73
## + HU               1    4.6765 42.557 -354.77
## <none>                         47.234 -333.94
## + as.factor(Male)  1    0.1319 47.102 -332.55
##
```

```
## Step:  AIC=-380.73
## GPA ~ HSGPA
##
##                   Df Sum of Sq    RSS     AIC
## + HU               1    3.3067 34.494 -398.78
## + as.factor(Male)  1    0.4138 37.387 -381.14
## <none>                         37.801 -380.73
##
## Step:  AIC=-398.78
## GPA ~ HSGPA + HU
##
##                   Df Sum of Sq    RSS     AIC
## + as.factor(Male)  1    0.4338 34.060 -399.55
## <none>                         34.494 -398.78
##
## Step:  AIC=-399.55
## GPA ~ HSGPA + HU + as.factor(Male)

##
## Call:
## lm(formula = GPA ~ HSGPA + HU + as.factor(Male), data = FirstYearGPA)
##
## Coefficients:
##      (Intercept)             HSGPA                HU  as.factor(Male)1
##          1.00814           0.52730           0.01722           0.08959
```

```
# backward selection, begin with the full model
step(modfull, scope=list(lower=modnull, upper=modfull), direction="backward")
```

```
## Start:  AIC=-399.55
## GPA ~ HU + HSGPA + as.factor(Male)
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         34.060 -399.55
## - as.factor(Male)  1    0.4338 34.494 -398.78
## - HU               1    3.3268 37.387 -381.14
## - HSGPA            1    8.3337 42.394 -353.61

##
## Call:
## lm(formula = GPA ~ HU + HSGPA + as.factor(Male), data = FirstYearGPA)
##
## Coefficients:
##      (Intercept)                HU             HSGPA  as.factor(Male)1
##          1.00814           0.01722           0.52730           0.08959
```

```
# we can also do both directions at the same time
step(modnull, scope=list(lower=modnull, upper=modfull), direction="both")  #
begin with null model
```

```
## Start:  AIC=-333.94
## GPA ~ 1
```

```
##
##                   Df Sum of Sq    RSS     AIC
## + HSGPA            1    9.4329 37.801 -380.73
## + HU               1    4.6765 42.557 -354.77
## <none>                         47.234 -333.94
## + as.factor(Male)  1    0.1319 47.102 -332.55
##
## Step:  AIC=-380.73
## GPA ~ HSGPA
##
##                   Df Sum of Sq    RSS     AIC
## + HU               1    3.3067 34.494 -398.78
## + as.factor(Male)  1    0.4138 37.387 -381.14
## <none>                         37.801 -380.73
## - HSGPA            1    9.4329 47.234 -333.94
##
## Step:  AIC=-398.78
## GPA ~ HSGPA + HU
##
##                   Df Sum of Sq    RSS     AIC
## + as.factor(Male)  1    0.4338 34.060 -399.55
## <none>                         34.494 -398.78
## - HU               1    3.3067 37.801 -380.73
## - HSGPA            1    8.0631 42.557 -354.77
##
## Step:  AIC=-399.55
## GPA ~ HSGPA + HU + as.factor(Male)
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         34.060 -399.55
## - as.factor(Male)  1    0.4338 34.494 -398.78
## - HU               1    3.3268 37.387 -381.14
## - HSGPA            1    8.3337 42.394 -353.61

##
## Call:
## lm(formula = GPA ~ HSGPA + HU + as.factor(Male), data = FirstYearGPA)
##
## Coefficients:
##      (Intercept)             HSGPA                HU  as.factor(Male)1
##          1.00814           0.52730           0.01722           0.08959
```

```
step(modfull, scope=list(lower=modnull, upper=modfull), direction="both")  # begin with full model
```

```
## Start:  AIC=-399.55
## GPA ~ HU + HSGPA + as.factor(Male)
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         34.060 -399.55
## - as.factor(Male)  1    0.4338 34.494 -398.78
```

```
## - HU                1    3.3268 37.387 -381.14
## - HSGPA             1    8.3337 42.394 -353.61

##
## Call:
## lm(formula = GPA ~ HU + HSGPA + as.factor(Male), data = FirstYearGPA)
##
## Coefficients:
##      (Intercept)                HU             HSGPA  as.factor(Male)1
##          1.00814           0.01722           0.52730           0.08959
```

These three models give me the same model result, so I will use GPA~HSGPA+HU+Male to be my best model but I need to use anova to test out if I need to include intersation in my best model.

Here is my hypotheses:

H_0:: Both models are equally good H_a:: The model with interaction is better

```
# model 1, without interaction term
model1<-lm(GPA~HSGPA+HU+as.factor(Male), data=FirstYearGPA)
# model 2, with interaction term
model2<-lm(GPA~HSGPA+HU+as.factor(Male)+as.factor(Male):HU, data=FirstYearGPA)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ HSGPA + HU + as.factor(Male)
## Model 2: GPA ~ HSGPA + HU + as.factor(Male) + as.factor(Male):HU
##   Res.Df   RSS Df  Sum of Sq     F Pr(>F)
## 1    215 34.06
## 2    214 34.06  1 3.0693e-05 2e-04 0.9889
```

From the anova test above, the p values are over 0.05, which means I don't need to include interaction in my best model so the simplest model is the my best model.

```
summary(model1)
```

```
##
## Call:
## lm(formula = GPA ~ HSGPA + HU + as.factor(Male), data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00138 -0.28011  0.04876  0.26567  0.87772
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.008136   0.255149   3.951 0.000105 ***
## HSGPA             0.527301   0.072702   7.253 7.28e-12 ***
## HU                0.017216   0.003757   4.583 7.78e-06 ***
## as.factor(Male)1 0.089592   0.054141   1.655 0.099427 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.398 on 215 degrees of freedom
## Multiple R-squared:  0.2789, Adjusted R-squared:  0.2688
## F-statistic: 27.72 on 3 and 215 DF,  p-value: 3.404e-15
```

From the funtion above, I got the function for estimated GPA:

Predicted GPA = 1.008136 + 0.527301*HSGPA + 0.017216* HU + 0.089592 * Male = 1.008136 + 0.527301*3.3 + 0.017216* 16 + 0.089592 * 0 = 3.023685

So the predicted GPA is 3.023685