Team 7 Project

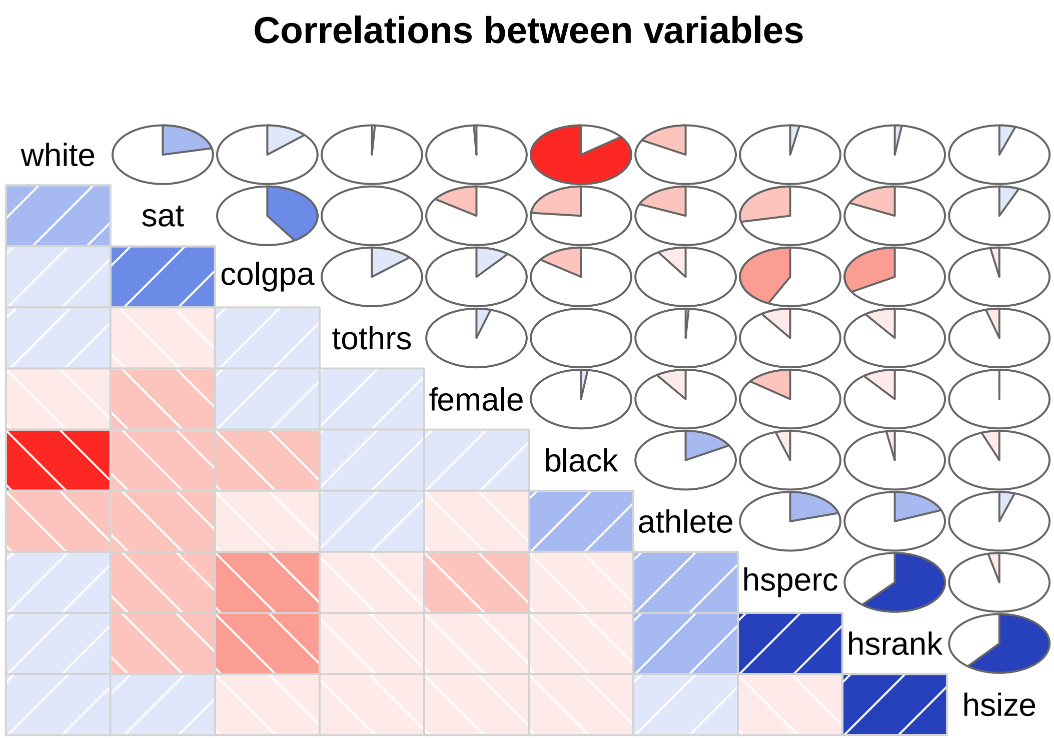
Casey Burgin, Taylor Dyer, Grayson Felt, Heber Jenson, Aleisha Grgich, Peyton Knight

11/21/2019

Introduction We want to investigate the question "How does high school size affect college GPA?" We decided to use the data set GPA2, found in the Wooldridge package. This data set comes from a midsize research university that supports men and women athletics at the Division I level. The original data set has 4,137 observations on 12 variables, including college GPA, combined SAT scores, high school size, and high school rank. There are also several dummy variables, such as athlete, gender, and race. Our main goal is to find out how the size of an individuals high school affects their college GPA; however, we will also investigate some other facets, such as "Does being an athlete affect GPA?" and "Does being male or female have an effect on GPA?" Understanding the effect that certain variables have on college GPA can give us a better understanding of how we can best prepare for college and could help guide changes in high school to create the most productive setting for learning.

Data We are using the data set GPA2 from the Wooldridge library. Our dependent variable is *colgpa*, which is the GPA after fall semester. Our independent variables will include *sat*, which is the combined SAT score; *tothrs*, which is the total hours through fall semester; *athlete*, a dummy variable to tell if student is an athlete; *hsize*, which is the size of their high school graduating class; *hsrank*, which is the rank in their high school graduating class; *female*, a dummy variable to compare gender; and *black*, a dummy variable to compare race.

```
df <- gpa2
df <- df[complete.cases(gpa2),]
df <- select(df, -c(verbmath,hsizesq ))
corrgram(df, order=TRUE, lower.panel=panel.shade,upper.panel=panel.pie,
          text.panel=panel.txt,main="Correlations between variables")
```



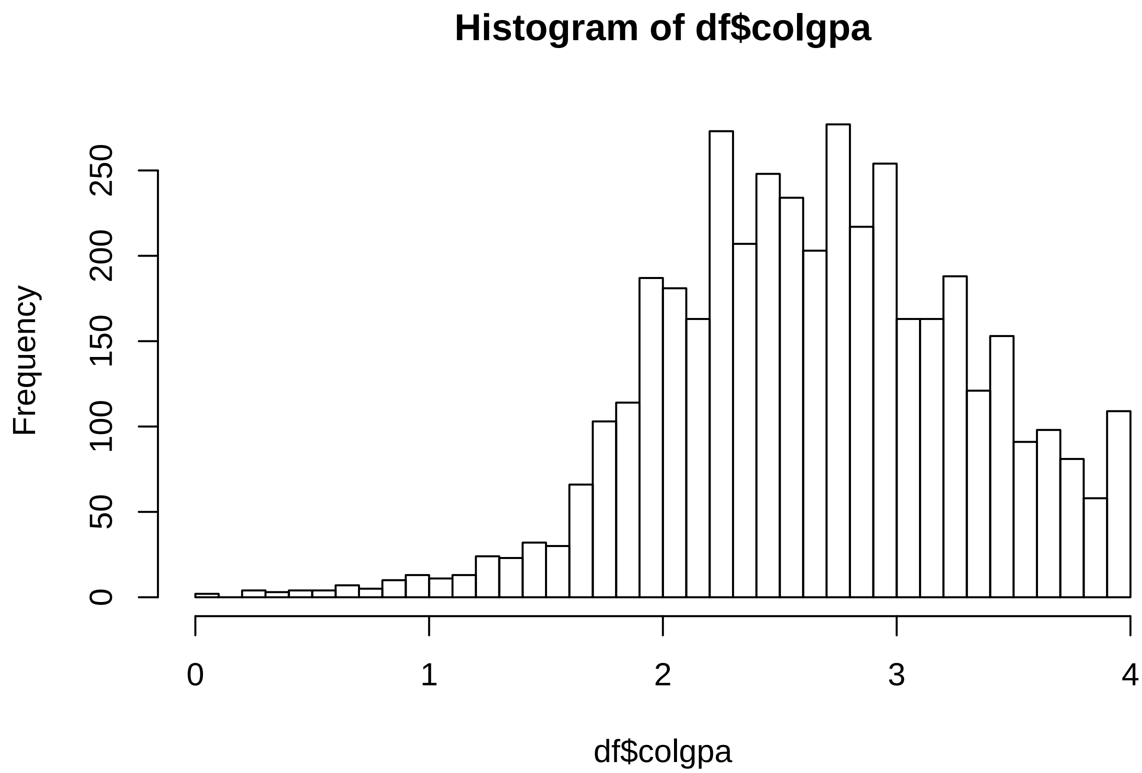

```r
summary(df)
```

```
      sat            tothrs           colgpa          athlete
 Min.   : 470   Min.   :  6.00   Min.   :0.000   Min.   :0.00000
 1st Qu.: 940   1st Qu.: 17.00   1st Qu.:2.210   1st Qu.:0.00000
 Median :1030   Median : 47.00   Median :2.660   Median :0.00000
 Mean   :1030   Mean   : 52.83   Mean   :2.653   Mean   :0.04689
 3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.120   3rd Qu.:0.00000
 Max.   :1540   Max.   :137.00   Max.   :4.000   Max.   :1.00000
     hsize          hsrank           hsperc            female
 Min.   :0.03   Min.   :  1.00   Min.   : 0.1667   Min.   :0.0000
 1st Qu.:1.65   1st Qu.: 11.00   1st Qu.: 6.4328   1st Qu.:0.0000
 Median :2.51   Median : 30.00   Median :14.5833   Median :0.0000
 Mean   :2.80   Mean   : 52.83   Mean   :19.2371   Mean   :0.4496
 3rd Qu.:3.68   3rd Qu.: 70.00   3rd Qu.:27.7108   3rd Qu.:1.0000
 Max.   :9.40   Max.   :634.00   Max.   :92.0000   Max.   :1.0000
     white            black
 Min.   :0.0000   Min.   :0.00000
 1st Qu.:1.0000   1st Qu.:0.00000
 Median :1.0000   Median :0.00000
 Mean   :0.9255   Mean   :0.05535
 3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.00000
```

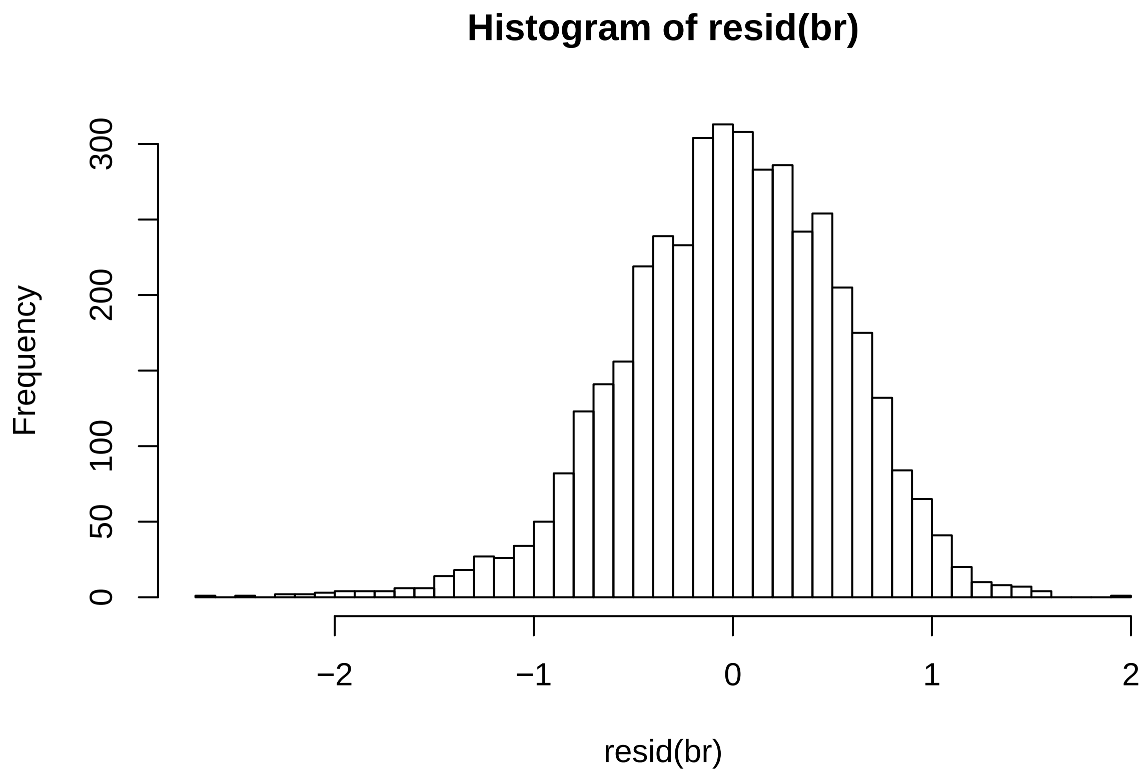
```r
br<- lm(colgpa~.,df)
stargazer(br, type = 'text')
```

```
===============================================
                        Dependent variable:
                    ---------------------------
                              colgpa
-----------------------------------------------
sat                          0.002***
                             (0.0001)

tothrs                       0.002***
                             (0.0002)

athlete                      0.217***
                              (0.042)

hsize                          0.005
                              (0.008)

hsrank                      -0.001***
                             (0.0003)

hsperc                      -0.010***
                              (0.001)

female                       0.146***
                              (0.018)

white                         -0.032
                              (0.062)

black                       -0.331***
                              (0.072)

Constant                     1.227***
                              (0.100)

-----------------------------------------------
Observations                   4,137
R2                             0.315
Adjusted R2                    0.314
Residual Std. Error      0.546 (df = 4127)
F Statistic          210.999*** (df = 9; 4127)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
hist(df$colgpa, breaks = 50)
```




```r
hist(resid(br), breaks = 50)
```



Empirical Framework The basic model we wanted to start with is:

$$\widehat{colgpa} = \beta_0 + \beta_1 sat + \beta_2 tothrs + \delta_0 athlete +$$

$$\beta_3 hsize + \beta_4 hsrank + \beta_5 hsperc + \delta_1 female + \delta_2 white + \delta_3 black + u$$

```
MRM <- lm(colgpa~ .,df)
stargazer(MRM, type = "text", digits = 5)
```

```
================================================
                         Dependent variable:
                    ----------------------------
                               colgpa
------------------------------------------------
sat                         0.00151***
                             (0.00007)

tothrs                      0.00174***
                             (0.00024)

athlete                     0.21731***
                             (0.04226)

hsize                        0.00535
                             (0.00816)
```

```
hsrank                          -0.00129***
                                 (0.00028)

hsperc                          -0.01011***
                                 (0.00087)

female                          0.14639***
                                 (0.01773)

white                            -0.03163
                                 (0.06216)

black                           -0.33144***
                                 (0.07194)

Constant                        1.22675***
                                 (0.09959)

-----------------------------------------------
Observations                       4,137
R2                                0.31513
Adjusted R2                       0.31364
Residual Std. Error     0.54566 (df = 4127)
F Statistic         210.99890*** (df = 9; 4127)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
bptest(MRM)
```

```
    studentized Breusch-Pagan test

data:  MRM
BP = 164.46, df = 9, p-value < 2.2e-16
```

```r
coeftest(MRM, vcov= hccm(MRM, type="hc0"))
```

```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.2267e+00  1.0168e-01  12.0650 < 2.2e-16 ***
sat          1.5077e-03  6.7379e-05  22.3765 < 2.2e-16 ***
tothrs       1.7364e-03  2.4212e-04   7.1716 8.742e-13 ***
athlete      2.1731e-01  3.7711e-02   5.7626 8.884e-09 ***
hsize        5.3529e-03  7.9661e-03   0.6720    0.5016
hsrank      -1.2887e-03  2.5891e-04  -4.9773 6.710e-07 ***
hsperc      -1.0113e-02  8.4887e-04 -11.9130 < 2.2e-16 ***
female       1.4639e-01  1.7569e-02   8.3320 < 2.2e-16 ***
white       -3.1628e-02  6.3574e-02  -0.4975    0.6189
black       -3.3144e-01  7.3907e-02  -4.4845 7.507e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated basic model, using the robust coefficients, is as follows:

$$\widehat{colgpa} = 1.22675 + 0.00151 sat + 0.00174 tothrs + 0.217 athlete$$

$$+0.00535 hsize - 0.00129 hrank - 0.0101 hsperc + 0.146 female - 0.0316 white - 0.331 black$$

We used the OLS with robust errors as our estimation technique. This is because we can make the following assumptions:

1) It is linear in parameters.
2) There are no perfect collinearity issues as seen by our correlation analysis.
3) We assume the zero conditional mean assumption holds true because we are controlling for enough variables.
4) Our data is not homoskedastic as shown by the bptest. This requires us to include the robust errors to correct our OLS estimation.
5) The distribution of the residuals is normally distributed due to the central limit theorem (because we have 4000+ observations).
6) Our equation also passes the multicollinearity test as shown by the VIF test found above.

The functional form is a level-level model. This made the most logical sense, as we are restricted by our dependent variable from performing a log transformation. We later tested different interactions and quadratic functional forms to create the best overall model for estimating colgpa.

Results

We attempted over 20 different estimates containing a variety of interactions and quadratic functions.

$$\widehat{colgpa}\ \beta_0 + \beta_1 sat + \beta_2 tothrs + \delta_0 athlete$$

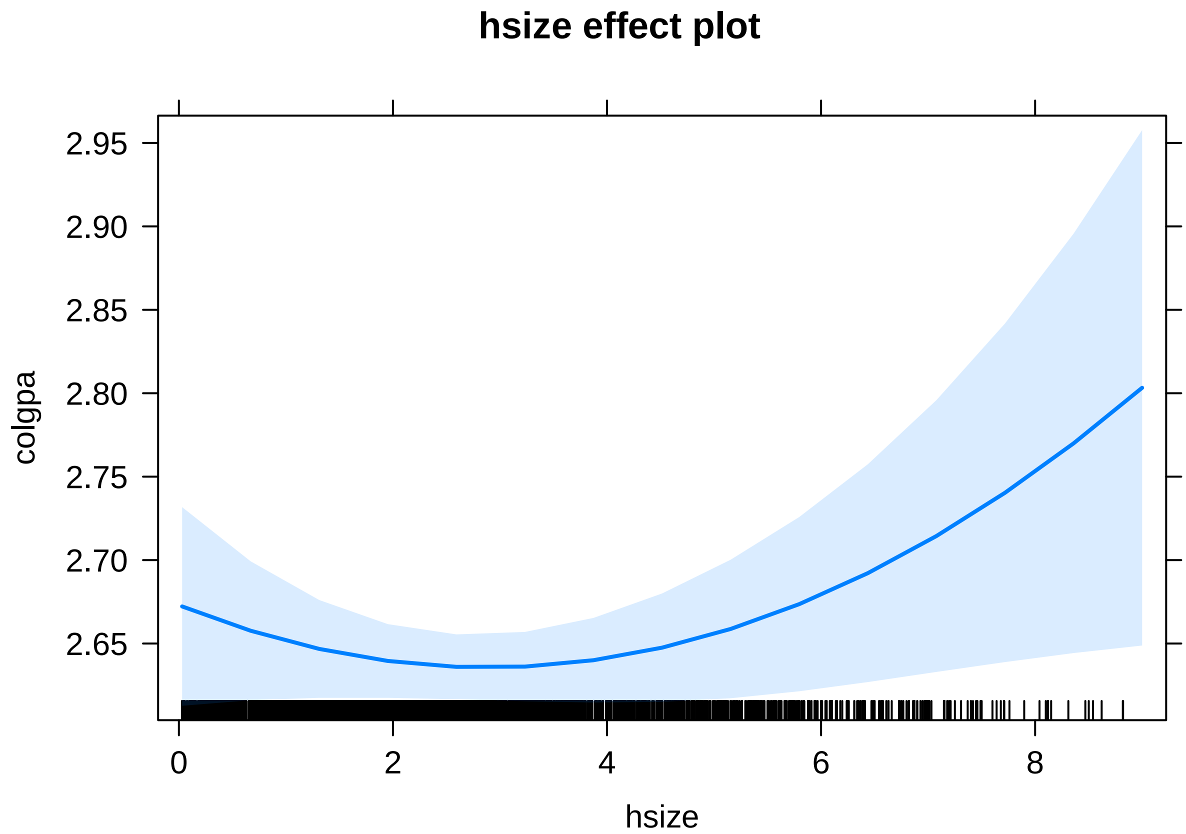
$$+\beta_3 hsrank + \beta_4 hsperc + \delta_1 female + \delta_2 black + \delta_3 (black * athlete) + u$$

This equation shows a variety of interesting information about gender, race, and high school rankings.

```r
A <- lm(colgpa~ I(sat^2)+athlete+I(athlete*black)+hsize+tothrs+
          female+black+I(hsrank*sat)+white,df)
B <- lm(colgpa~ hsrank+I(sat^2)+I(athlete*black)+hsize+tothrs+
          female+black+I(hsrank*sat),df)
C <- lm(colgpa~ I(sat^2)+I(athlete*black)+hsize+tothrs+
          female+black+I(hsrank*sat),df)
D <- lm(colgpa~ I(sat^2)+I(tothrs^2)+I(athlete^2)+I(hsize^2)+
          I(hsperc^2)+I(female^2)+I(black^2),df)
E <- lm(colgpa~sat+tothrs+athlete+hsize+hsrank+hsperc+
          female+black+white,df)
F <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black,df)
G <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete),df)
H <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete),df)
I <- lm(colgpa~sat+I(tothrs^2)+athlete+hsrank+hsperc+
          female+black+I(black*athlete),df)
J <- lm(colgpa~hsize+hsperc+sat+female+athlete,df)
K <- lm(colgpa~female+sat+hsperc+tothrs,df)
L <- lm(colgpa~sat+hsperc+tothrs+female+black+white,df)
M <- lm(colgpa~sat+hsize+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsize^2)+white,df)
N <- lm(colgpa~sat+hsize+tothrs+athlete+hsrank+hsperc+
```

```r
          female+black+I(black*athlete)+I(hsize^2),df)
O <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsize^2),df)
P <- lm(colgpa~sat+I(hsize*black)+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsize^2),df)
Q <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete),df)
R <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(black*sat),df)
S <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(female*athlete),df)
T <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(athlete*tothrs),df)
U <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsrank*hsperc)+I(hsize*hsrank),df)
V <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsrank*hsperc),df)
W <- lm(colgpa~log(sat)+log(tothrs)+athlete+log(hsrank)+log(hsperc)+
          female+black+I(black*athlete),df)
X <- lm(colgpa~log(sat)+log(tothrs)+log(hsrank)+log(hsperc)+
          female+black+I(black*athlete),df)

stargazer(A,B,C, type = "text", digits = 4)
stargazer(D,E,F, type = "text", digits = 4)
stargazer(G,H,I, type = "text", digits = 4)
stargazer(J,K,L, type = "text", digits = 4)
stargazer(M,N,O, type = "text", digits = 4)
stargazer(P,Q,R, type = "text", digits = 4)
stargazer(S,T,U, type = "text", digits = 4)
stargazer(V,W,X, type = "text", digits = 4)
```

```r
N <- lm(colgpa~sat+hsize+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsize^2),df)
plot(effect("hsize",N))
```



This shows that $hsize^2$ should not be used in the model, as two stories are being told with lots of data points on each side of the minimum value.

```
U <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+
          female+black+I(black*athlete)+I(hsrank*hsperc)+I(hsize*hsrank),df)
stargazer(U,type = "text", digits = 5)
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               colgpa
------------------------------------------------
sat                          0.00145***
                             (0.00007)

tothrs                       0.00166***
                             (0.00024)

athlete                      0.10770**
                             (0.04631)

hsrank                      -0.00458***
                             (0.00060)

hsperc                      -0.01075***
                             (0.00083)
```

```
female                      0.14270***
                            (0.01761)

black                      -0.39977***
                            (0.04222)

I(black * athlete)          0.42811***
                            (0.10188)

I(hsrank * hsperc)          0.00004***
                            (0.00001)

I(hsize * hsrank)           0.00023***
                            (0.00007)

Constant                    1.34983***
                            (0.08067)

-----------------------------------------------
Observations                  4,137
R2                           0.32572
Adjusted R2                  0.32409
Residual Std. Error     0.54149 (df = 4126)
F Statistic         199.31410*** (df = 10; 4126)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
bptest(U)
```

```
    studentized Breusch-Pagan test

data:  U
BP = 168.23, df = 10, p-value < 2.2e-16
```

```r
coeftest(U, vcov= hccm(U, type="hc0"))
```

```

t test of coefficients:

                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         1.3498e+00  8.1995e-02  16.4623 < 2.2e-16 ***
sat                 1.4464e-03  6.7637e-05  21.3851 < 2.2e-16 ***
tothrs              1.6605e-03  2.3965e-04   6.9290 4.896e-12 ***
athlete             1.0770e-01  4.0453e-02   2.6624  0.007789 **
hsrank             -4.5761e-03  6.5409e-04  -6.9962 3.056e-12 ***
hsperc             -1.0751e-02  8.1847e-04 -13.1352 < 2.2e-16 ***
female              1.4270e-01  1.7476e-02   8.1652 4.221e-16 ***
black              -3.9977e-01  4.4772e-02  -8.9289 < 2.2e-16 ***
I(black * athlete)  4.2811e-01  8.3713e-02   5.1139 3.298e-07 ***
I(hsrank * hsperc)  4.2861e-05  8.3100e-06   5.1577 2.617e-07 ***
I(hsize * hsrank)   2.3181e-04  7.4743e-05   3.1014  0.001939 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(U)
```

```
              sat             tothrs            athlete 
         1.262195           1.016609           1.352240 
           hsrank             hsperc             female 
        20.911459           2.664830           1.082432 
            black I(black * athlete) I(hsrank * hsperc) 
         1.315365           1.541121           8.035501 
 I(hsize * hsrank) 
        12.078719 
```

This model suffers from multicollinearity. Let's consider V:

```r
V <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+female+black+I(black*athlete)+I(hsrank*hsperc),df)
stargazer(V,type = "text", digits = 5)
```

```
===============================================
                        Dependent variable:    
                    ---------------------------
                              colgpa           
-----------------------------------------------
sat                         0.00146***         
                             (0.00007)          
                                               
tothrs                      0.00168***         
                             (0.00024)          
                                               
athlete                      0.10978**          
                             (0.04635)          
                                               
hsrank                      -0.00303***        
                             (0.00035)          
                                               
hsperc                      -0.01207***        
                             (0.00072)          
                                               
female                      0.14524***         
                             (0.01761)          
                                               
black                       -0.39799***        
                             (0.04227)          
                                               
I(black * athlete)          0.41769***         
                             (0.10194)          
                                               
I(hsrank * hsperc)          0.00004***         
                             (0.00001)          
                                               
Constant                    1.32513***         
                             (0.08039)          
                                               
-----------------------------------------------
Observations                   4,137           
R2                            0.32407          
```

```
Adjusted R2                  0.32259
Residual Std. Error     0.54209 (df = 4127)
F Statistic         219.84850*** (df = 9; 4127)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
bptest(V)
```

```

    studentized Breusch-Pagan test

data:  V
BP = 167.17, df = 9, p-value < 2.2e-16
```

```
coeftest(V, vcov= hccm(V, type="hc0"))
```

```

t test of coefficients:

                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         1.3251e+00  8.1393e-02  16.2806 < 2.2e-16 ***
sat                 1.4615e-03  6.7516e-05  21.6468 < 2.2e-16 ***
tothrs              1.6819e-03  2.3959e-04   7.0197 2.589e-12 ***
athlete             1.0978e-01  4.0242e-02   2.7281  0.006397 **
hsrank             -3.0336e-03  3.7374e-04  -8.1170 6.247e-16 ***
hsperc             -1.2074e-02  7.2465e-04 -16.6618 < 2.2e-16 ***
female              1.4524e-01  1.7479e-02   8.3095 < 2.2e-16 ***
black              -3.9799e-01  4.4792e-02  -8.8854 < 2.2e-16 ***
I(black * athlete)  4.1769e-01  8.2828e-02   5.0429 4.782e-07 ***
I(hsrank * hsperc)  4.3391e-05  7.7464e-06   5.6015 2.264e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(V)
```

```
               sat             tothrs            athlete
          1.256040           1.015817           1.351970
            hsrank             hsperc             female
          7.047767           1.995487           1.080202
             black I(black * athlete) I(hsrank * hsperc)
          1.315136           1.539531           8.031043
```

Although this model doesn't have a VIF score above 10, the interaction between hsrank and hsperc should probably be removed. Now lets look at Q:

```
Q <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+female+black+I(black*athlete),df)
stargazer(Q,type = "text", digits = 5)
```

```

===============================================
                        Dependent variable:
                    ---------------------------
                              colgpa
-----------------------------------------------
sat                         0.00152***
                             (0.00007)
```

```
tothrs                        0.00172***
                               (0.00024)

athlete                       0.13778***
                               (0.04633)

hsrank                        -0.00117***
                               (0.00017)

hsperc                        -0.01056***
                               (0.00068)

female                        0.14705***
                               (0.01768)

black                         -0.37265***
                               (0.04225)

I(black * athlete)            0.42981***
                               (0.10237)

Constant                      1.20861***
                               (0.07844)

-----------------------------------------------
Observations                     4,137
R2                              0.31793
Adjusted R2                     0.31661
Residual Std. Error     0.54448 (df = 4128)
F Statistic         240.52050*** (df = 8; 4128)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
bptest(Q)
```

```
    studentized Breusch-Pagan test

data:  Q
BP = 163.74, df = 8, p-value < 2.2e-16
```

```r
coeftest(Q, vcov= hccm(Q, type="hc0"))
```

```
t test of coefficients:

                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         1.2086e+00  7.8706e-02  15.3560 < 2.2e-16 ***
sat                 1.5159e-03  6.6979e-05  22.6323 < 2.2e-16 ***
tothrs              1.7197e-03  2.4115e-04   7.1312 1.170e-12 ***
athlete             1.3778e-01  4.0096e-02   3.4363 0.0005955 ***
hsrank             -1.1747e-03  1.6247e-04  -7.2302 5.720e-13 ***
hsperc             -1.0558e-02  6.6651e-04 -15.8414 < 2.2e-16 ***
female              1.4705e-01  1.7551e-02   8.3787 < 2.2e-16 ***
black              -3.7265e-01  4.4834e-02  -8.3118 < 2.2e-16 ***
```

```
I(black * athlete)  4.2981e-01  8.4772e-02   5.0702 4.149e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(Q)
```

```
               sat             tothrs            athlete
          1.234440           1.015147           1.338805
            hsrank             hsperc             female
          1.616956           1.758655           1.079898
             black I(black * athlete)
          1.302520           1.538951
```

Everything seems to check out. The BP test does show that there is heteroskedasticity in the data, which means we will need to report the robust estimates. Overall, this model seems to fit our data best. Now let's test our dummy variables to make sure they matter.

```r
Q <- lm(colgpa~sat+tothrs+athlete+hsrank+hsperc+female+black+I(black*athlete),df)
linearHypothesis(Q, c("black=0","I(black * athlete)"), vcov=hccm(Q,type="hc0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## black = 0
## I(black * athlete) = 0
##
## Model 1: restricted model
## Model 2: colgpa ~ sat + tothrs + athlete + hsrank + hsperc + female +
##     black + I(black * athlete)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   4130
## 2   4128  2 35.028 8.229e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(Q, c("athlete","I(black * athlete)"), vcov=hccm(Q,type = "hc0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## athlete = 0
## I(black * athlete) = 0
##
## Model 1: restricted model
## Model 2: colgpa ~ sat + tothrs + athlete + hsrank + hsperc + female +
##     black + I(black * athlete)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   4130
## 2   4128  2 31.386 2.964e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(Q, "female=0",vcov=hccm(Q,type = "hc0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## female = 0
##
## Model 1: restricted model
## Model 2: colgpa ~ sat + tothrs + athlete + hsrank + hsperc + female +
##     black + I(black * athlete)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   4129
## 2   4128  1 70.203 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dummy variables do matter in the equation. We will now construct the LPM.

```r
df <- mutate(df, AAGPA=ifelse(colgpa>mean(colgpa),1,0))
lpm <- lm(AAGPA~sat+tothrs+athlete+hsrank+hsperc+female+black+I(black*athlete),df)
stargazer(lpm,type = "text", digits = 5)
```

```
===============================================
                        Dependent variable:
                    ---------------------------
                               AAGPA
-----------------------------------------------
sat                          0.00088***
                              (0.00005)

tothrs                       0.00056***
                              (0.00020)

athlete                        0.05418
                              (0.03753)

hsrank                       -0.00061***
                              (0.00013)

hsperc                       -0.00775***
                              (0.00055)

female                       0.10221***
                              (0.01432)

black                        -0.28025***
                              (0.03422)

I(black * athlete)           0.30295***
                              (0.08292)
```

```
Constant                         -0.29035***
                                  (0.06354)

-----------------------------------------------
Observations                       4,137
R2                                0.22373
Adjusted R2                       0.22222
Residual Std. Error     0.44101 (df = 4128)
F Statistic         148.71570*** (df = 8; 4128)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
bptest(lpm)
```

```
    studentized Breusch-Pagan test

data:  lpm
BP = 69.271, df = 8, p-value = 6.863e-12
```

WLS

```r
y_hat <- predict(lpm)
summary(y_hat)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.5472  0.3508  0.5200  0.5004  0.6703  1.1793
```

```r
h     <- y_hat * (1-y_hat)
range(h)
```

```
[1] -0.8466225  0.2500000
```

```r
h<- ifelse(h<0,0.01,h)
summary(h)
```

```
     Min.    1st Qu.     Median       Mean    3rd Qu.      Max.
0.0005355 0.1777377 0.2239484 0.1983847 0.2440927 0.2500000
```

```r
w<- 1/h
wls <- lpm <- lm(AAGPA~sat+tothrs+athlete+hsrank+hsperc+female+black+I(black*athlete),df)
stargazer(lpm, wls, type = "text")
```

```
============================================================
                                   Dependent variable:
                               ----------------------------
                                          AAGPA
                                    (1)            (2)
------------------------------------------------------------
sat                              0.001***       0.001***
                                 (0.0001)       (0.0001)

tothrs                           0.001***       0.001***
                                 (0.0002)       (0.0002)

athlete                            0.054          0.054
                                  (0.038)        (0.038)
```

```
hsrank                             -0.001***      -0.001***
                                    (0.0001)       (0.0001)

hsperc                             -0.008***      -0.008***
                                    (0.001)        (0.001)

female                              0.102***      0.102***
                                    (0.014)        (0.014)

black                              -0.280***      -0.280***
                                    (0.034)        (0.034)

I(black * athlete)                  0.303***      0.303***
                                    (0.083)        (0.083)

Constant                           -0.290***      -0.290***
                                    (0.064)        (0.064)

------------------------------------------------------------
Observations                         4,137          4,137
R2                                   0.224          0.224
Adjusted R2                          0.222          0.222
Residual Std. Error (df = 4128)      0.441          0.441
F Statistic (df = 8; 4128)         148.716***    148.716***
============================================================
Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
ywls <- predict(wls)
summary(ywls)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.5472  0.3508  0.5200  0.5004  0.6703  1.1793
```

```
CM <- table(df[, "AAGPA"], predict(wls) >= 0.4)
prop.table(CM,1)
```

```

        FALSE      TRUE
  0 0.4784712 0.5215288
  1 0.1400966 0.8599034
```

```
(PC_PO <- (sum(ywls >= 0.4 & df$AAGPA==1) + sum(ywls <= 0.4 & df$AAGPA==0)) / length(df$AAGPA))
```

```
[1] 0.6693256
```

```
stargazer(Q,type = "text", digits = 5)
```

```

===============================================
                        Dependent variable:
                    ---------------------------
                              colgpa
-----------------------------------------------
sat                         0.00152***
                             (0.00007)
```

```
tothrs                        0.00172***
                               (0.00024)

athlete                       0.13778***
                               (0.04633)

hsrank                        -0.00117***
                               (0.00017)

hsperc                        -0.01056***
                               (0.00068)

female                        0.14705***
                               (0.01768)

black                         -0.37265***
                               (0.04225)

I(black * athlete)            0.42981***
                               (0.10237)

Constant                      1.20861***
                               (0.07844)

-----------------------------------------------
Observations                     4,137
R2                              0.31793
Adjusted R2                     0.31661
Residual Std. Error     0.54448 (df = 4128)
F Statistic         240.52050*** (df = 8; 4128)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
bptest(Q)
```

```
    studentized Breusch-Pagan test

data:  Q
BP = 163.74, df = 8, p-value < 2.2e-16
```

```r
coeftest(Q, vcov= hccm(Q, type="hc0"))
```

```
t test of coefficients:

                      Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept)         1.2086e+00  7.8706e-02  15.3560 < 2.2e-16 ***
sat                 1.5159e-03  6.6979e-05  22.6323 < 2.2e-16 ***
tothrs              1.7197e-03  2.4115e-04   7.1312 1.170e-12 ***
athlete             1.3778e-01  4.0096e-02   3.4363 0.0005955 ***
hsrank             -1.1747e-03  1.6247e-04  -7.2302 5.720e-13 ***
hsperc             -1.0558e-02  6.6651e-04 -15.8414 < 2.2e-16 ***
female              1.4705e-01  1.7551e-02   8.3787 < 2.2e-16 ***
black              -3.7265e-01  4.4834e-02  -8.3118 < 2.2e-16 ***
```

```
I(black * athlete)  4.2981e-01  8.4772e-02   5.0702 4.149e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(Q)
```

```
               sat             tothrs            athlete 
          1.234440           1.015147           1.338805 
            hsrank             hsperc             female 
          1.616956           1.758655           1.079898 
             black I(black * athlete) 
          1.302520           1.538951 
```

The final equation we found to estimate colgpa is:

$$\widehat{colgpa} = 1.209 + 0.00152sat + 0.00172tothrs + 0.13778athlete$$
$$-0.00117hsrank - 0.01065hsperc + 0.14705female - 0.37265black + 0.42981black * athlete$$

You will see the Robust Standard errors in the report above.

The coefficient's interpretations are as follows, assuming that all other variables are held constant: This means that for every unit increase in sat score, colgpa is anticipated to increase by 0.00152 units. For every unit increase to tothrs, colgpa will increase by 0.00172 units. Athletes, on average, are estimated to have a colgpa of .13778 higher than non-athletes. For every unit increase in hsrank, colgpa will decrease by 0.00117 units. (Remember definition of hsrank) For every unit increase in hsperc, colgpa will decrease by 0.01065 units. (Remember definition of hsperc) Females, on average, are estimated to have a colgpa of 0.14705 higher than males. Being black shows an estimate of having a colgpa of .37265 lower than non-blacks. Being a black athlete predicts a colpga of .42981 units lower than non-black, non-athletes.

The R-squared for the estimate says that, according to our data, the variables explain about 31.79% of colgpa. Adjusted R-squared is the better explanation for colgpa, as it factors in the significance of the variables used to estimate colgpa. In our model, we calculated an adjusted R-squared value of 0.31661, or 31.66%. The F-stat measures the overall significance of the variables being used to predict colgpa. The T-stats are similar to the F-stat, except each T-stat only measure one variable's significance.

Conclusion

From our analysis of the data, we conclude that hsize does not have a significant impact on estimating colgpa. This was surprising to us, as we believed that a smaller hsize would lead to a higher colgpa. Instead, we found that race and gender have the largest impact on estimating colgpa. Any data that related to high school was not practical as a significant variable to change the overall colgpa. From our data, we see that for all black students, even across gender, colgpa is lower, only excluding black male athletes.

For further investigation, we would hope to obtain more information from students to better estimate colgpa. Variables we would like to see included in a dataset would be scholarships each student has, marriage status, working hours per week, hours of sleep per week, high school gpa, and socioeconomic status. We feel like these variables would lead to a more accurate estimation of colgpa.