Assignment 3
Lunwei Wu
1001303957
STA302H1f-Summer 2017

Introduction:

The dataset i found was from https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/FirstYearGPA.html. And it consists 219 observations with one dependent Y(GPA) :First-year college GPA on a 0.0-4.0 scale, and following 9 variables:

X1(HSGPA):	High school GPA on a 0.0-4.0 scale
X2(SATV):	Verbal/critical reading SAT score
X3(SATM):	Math SAT score
X4(Male)	1= male, 0= female
X5(HU):	Number of credit hours earned in humanities courses in high school
X6(SS):	Number of credit hours earned in social science courses in high school
X7(FirstGen):	1= student is the first in her or his family to attend college, 0=otherwise
X8(White):	1= white students, 0= others
X9(CollegeBound):	1=attended a high school where >=50% students intended to go on to college, 0=otherwise

For this assignment, I want to investigate the linear regression and correlation between Y(GPA) and X. Also, I would like to use the model I made to make prediction of Y(GPA) by X. Keeping a good GPA is important for every student since GPA is often viewed by professors and potential employers for an indication of how a person's study ability and productivity. The motivation behind this investigation was to help people improving study efficiency by knowing what factors are related to Y(GPA) and to understand why does Y(GPA) vary from individual. The methods I would use for analysis are:

1. Using pairwise scatter and correlation plot to observe the correlation between Y(GPA) and all variables.
2. Finding linear regression between Y(GPA) and all variables and analyze their fitted Residuals plot, Standardized residuals plot and Normal Q-Q plot.
3. Using boxcox test to transform Y(GPA) into its better fitted model and analyze the differences between Y(GPA) and transformed Y(GPA)
4. Comparing models and choose the better one to do further analysis.
5. Using hypothesis test and confidence interval to find b0 and b1 for the linear regression model

Analysis：

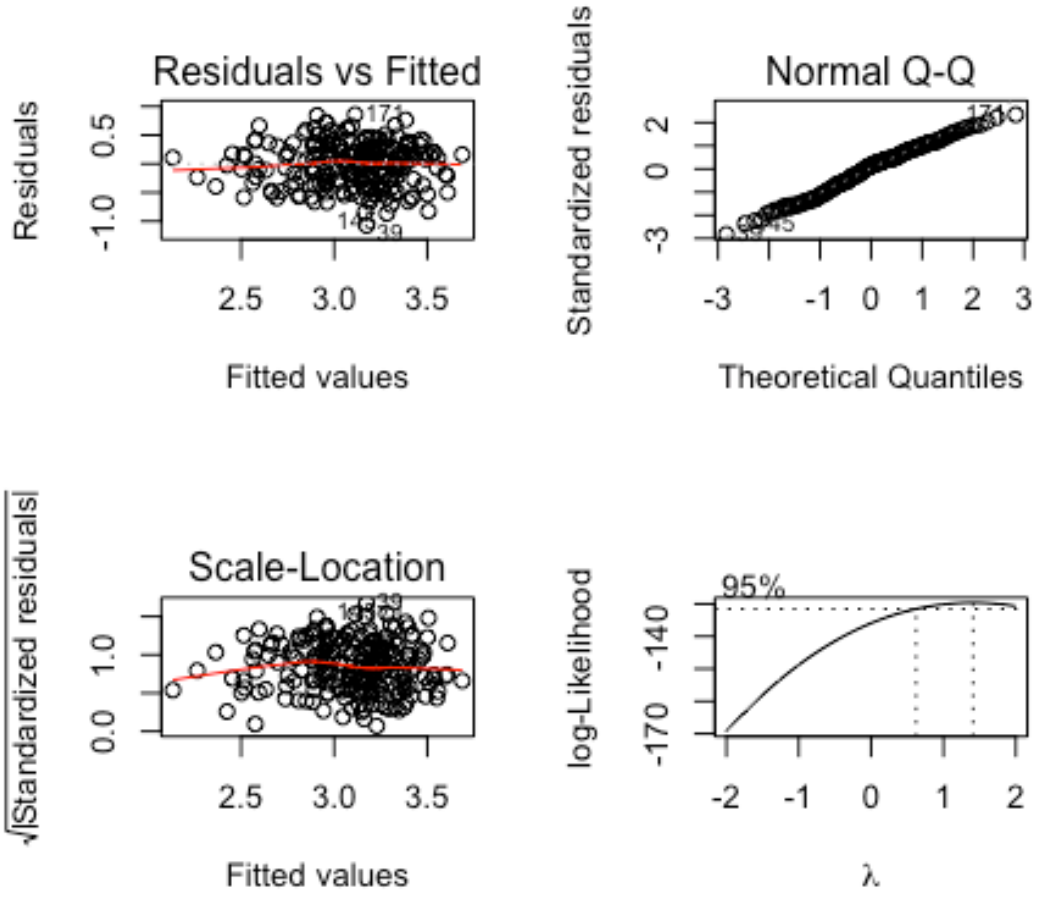
From the pairwise scatter and correlation plot, we have the following predictors have evidence of relationship with the response Y(GPA):

1. X1(HSGPA): r = 0.4469(moderate correlation), p-value < 0.0001
2. X2(SATV) : r=0.3043(weak correlation), p-value<0.0001
3. X5(HU): r=0.3147(weak correlation), p-value <0.0001
4. X8(White): r=0.2818(weak correlation), p-value<0.0001

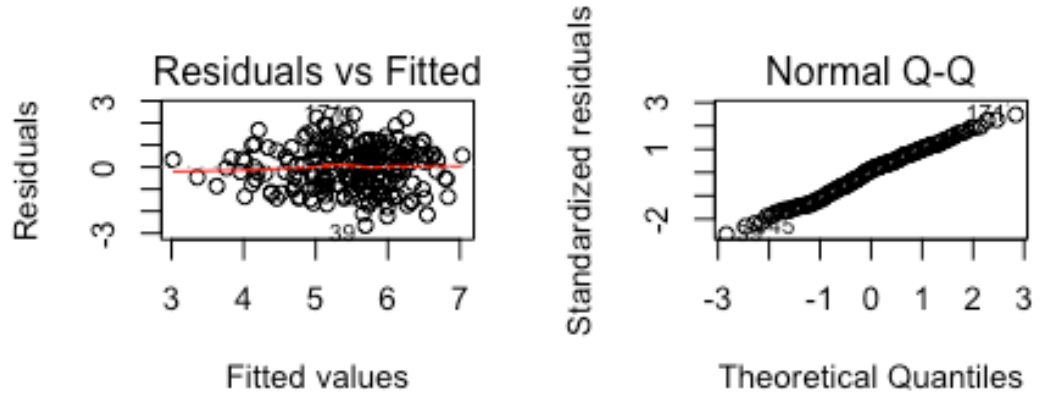
Notice that there are 4 of predictor variables, which are X1(HSGPA), X2(SATV), X5(HU), and X8(White) respectively, appear to have linear relationship with the response Y(GPA), since the pairwise scatter and correlation plot shows fairly small p-value for correlations.

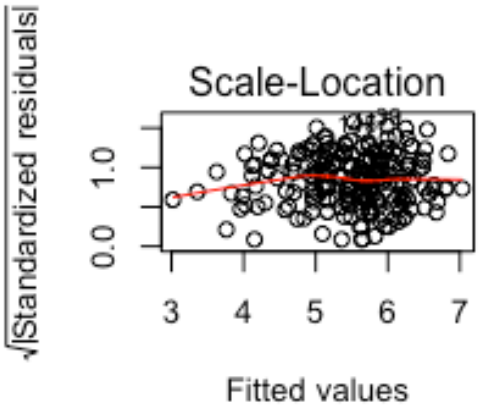
Also notice that the following predictors have evidence of relationship:

1. X1(HSGPA) and X2(SATV): r = 0.2103 (week correlation), p-value = 0.00175
2. X2(SATV) and X3(SATM): r = 0.5269 (moderate correlation), p-value <0.0001
3. X3(SATM) and X4(Male): r = 0.371(week correlation), p-value < 0.0001
4. X5(HU) and X6(SS): r = -0.3066(week correlation), p-value <0.0001
5. X7(FirstGen) and X8(White): r = -0.2379(week correlation), p-value = 0.000382



Checking both Residuals plot and Scale plot for Y(GPA) and its variables, we find that there is no clear pattern. Residuals get randomly spread as the increase of fitted values. There is a constant variance of Y(GPA). This indicates a linear relationship between Y(GPA) and its variables. Furthermore, from either before or standardized residual plot, there are 3 observations (39, 145, 171) have larger residuals. And their standardized residuals are close to 1.6. On the other hand, Normal Q-Q plot shows a slightly left skewed tail with several observations that have large residuals, but overall it follows an approximately straight line. The normality assumption looks fine.





Using “box-cox” method we find the MLE of the power parameter is 1.41, which is close to 1.5. So I choose to transform Y(GPA) to Y(GPA)^1.5. Comparing both models, we find no clear difference in Residuals plot and Standardized plot. The constant variance assumption is slightly improved after the transformation. Further examining the Q-Q plot, the normality assumption looks fine. Besides the plot, the transformed model gets improved in SSE, AIC and R square. SSE for the transformed model is 30.68389, which is smaller than 30.71902(before transformation). AIC for the transformed model is 208.8588, which is smaller than 213.3375(before transformation). R square for transformed model is 0.343, which is smaller than 0.3496(before transformation). Overall, transformed model is better than the original model. So I will use transformed model to do the further analysis.

From the summary of transformed model, we find that the p-value for the slope of X1(HSGPA), X5(HU) and X8(White) are less than 0.0001, which are significantly less than the significant level 0.05. So, if we built a hypothesis test with H0: b1 = 0, and H1: b1 not equal to 0 for any of the variable I mentioned above, there is enough

evidence to reject our H0 since its p-value is significantly less than 0.05. This indicates the linear relation ship between Y(GPA) and X1(HSGPA), X5(HU), X8(White). Furthermore, if we look the 95% confidence interval for b1, b5 and b8, we find that there are 95% of chance that b1 lies in between 0.9014 to 1.6709, b5 lies in between 0.02 to 0.061 and b8 lies in 0.1306 to 0.8533, which all three of the interval don't contain 0. Therefore, our estimated model using Y(GPA)^1.5 as the response could be written as Y(GPA)^1.5 = -1.155354 +1.2861982X1(HSGPA)+0.0413376X5(HU)+0.4920089X8(White)

Conclusion:

Given the data analyzed, we first found that First-year college GPA(Y) has correlation with High school GPA(X1), Verbal/critical reading SAT score(X2), Number of credit hours earned in humanities courses in high school(X5) and whether the student is white(X8). Also, we found the linear relationship exists between First-year college GPA(Y) and all the variables from their randomly spread residual plot and standardized residual plot. Further analysis shows the significant linear relation between First-year college GPA(Y) with High school GPA(X1), Number of credit hours earned in humanities courses in high school(X5) and whether the student is white(X8). In conclusion, First-year college GPA is associated with your High school GPA, which is reasonable in common sense. People with good GPA in high school usually tend to maintain their good GPA in the college. Also, data shows that the number of credit hour earned in humanities courses in high school and the fact whether the student is white could also affects GPA in first year college.

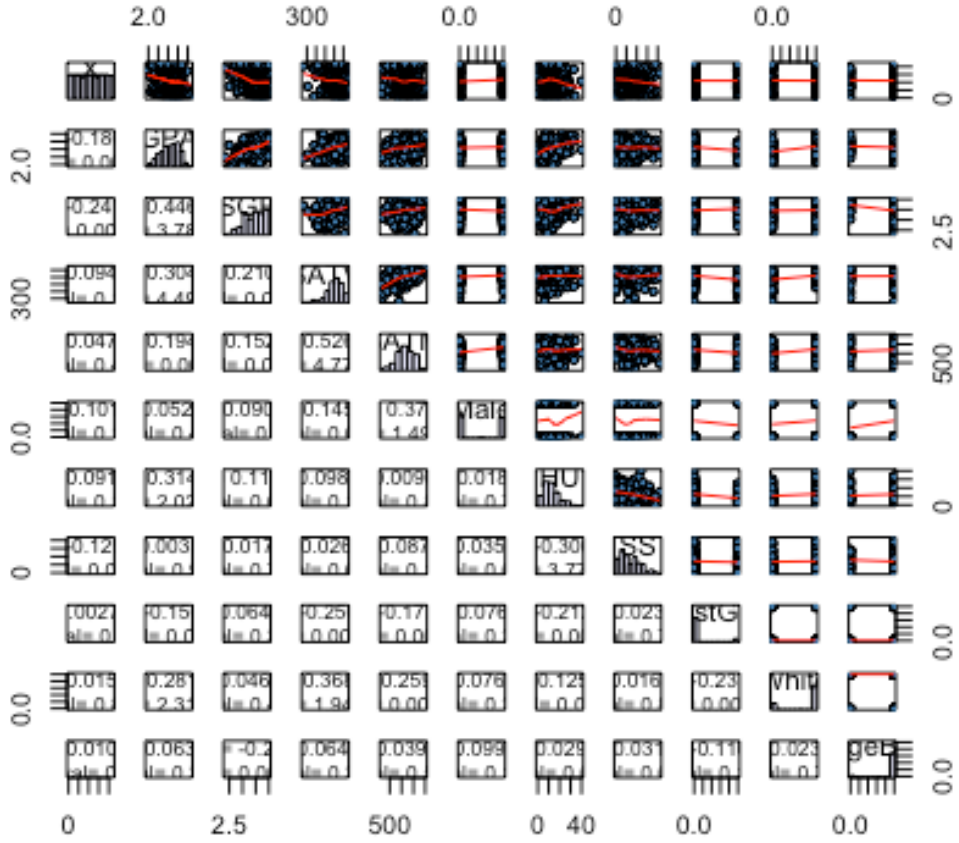
Appendix:

```
+ a3 = read.csv("/Users/jesse/Desktop/FirstYearGPA.csv",header=T)
```

#code for pairwise scatter correlation plot

```
> panel.hist <- function(x, ...){
+     usr <- par("usr"); on.exit(par(usr))
+     par(usr = c(usr[1:2], 0, 1.5) )
+     h <- hist(x, plot = FALSE)
+     breaks <- h$breaks; nB <- length(breaks)
+     y <- h$counts; y <- y/max(y)
+     rect(breaks[-nB], 0, breaks[-1], y, col="lavender", ...)
+ }
> panel.cor <- function(x, y, digits=4, prefix="", cex.cor, ...){
+     usr <- par("usr");
+     on.exit(par(usr))
+     par(usr = c(0, 1, 0, 1))
+     txt1 <- format( cor(x,y), digits=digits )
+     txt2 <- format(cor.test(x,y)$p.value, digits=digits)
+     text(0.5,0.5, paste("r=",txt1, "\n P.val=",txt2), cex=0.8)
+ }
```

- ```
  pairs(a3, lower.panel=panel.cor, cex =0.7, pch = 21,
  bg="steelblue",diag.panel=panel.hist, cex.labels = 1.1, font.labels=0.9,
  upper.panel=panel.smooth)
  ```
  ```



➢

#code for linear regression

> mod1 <-

lm(a3$GPA~a3$HSGPA+a3$SATV+a3$SATM+a3$Male+a3$HU+a3$SS+a3$First

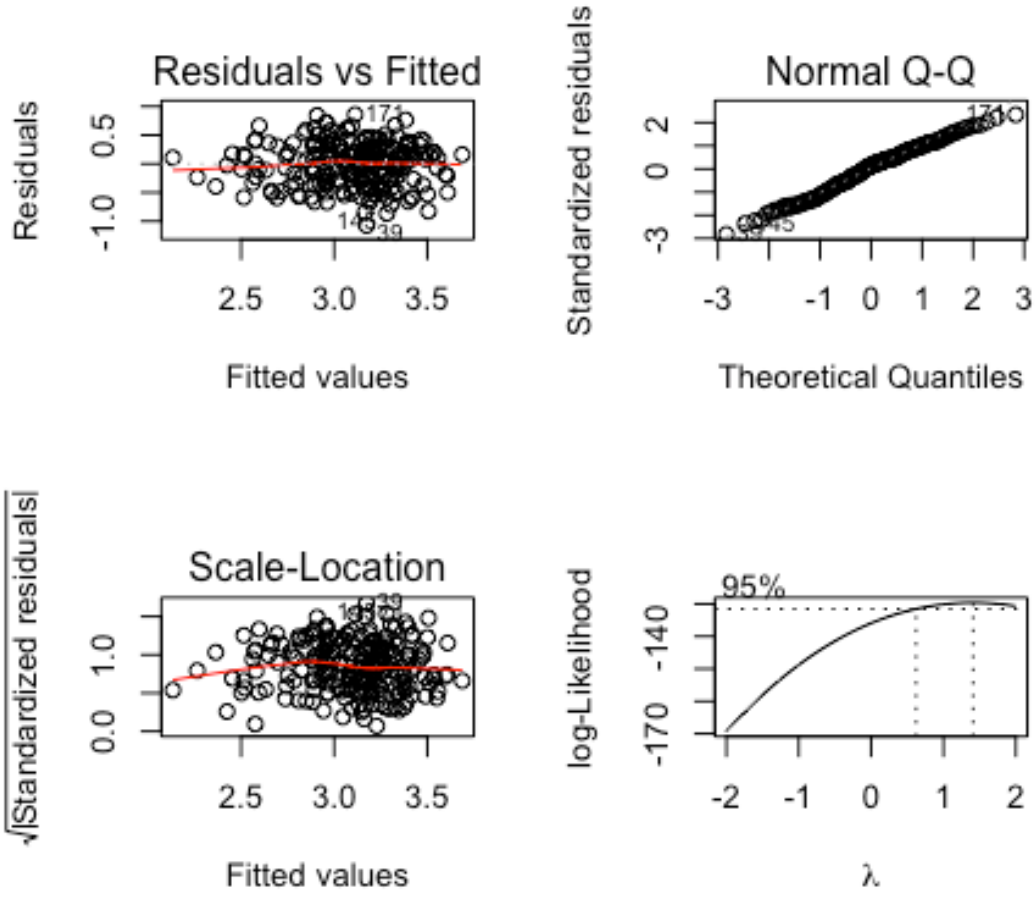
Gen+a3$White+a3$CollegeBound, data = a3)

> par(mfrow=c(2,2))

> plot(mod1,1)

> plot(mod1,2)

> plot(mod1,3)



> summary(mod1)

Call:

lm(formula = a3$GPA ~ a3$HSGPA + a3$SATV + a3$SATM + a3$Male +
    a3$HU + a3$SS + a3$FirstGen + a3$White + a3$CollegeBound,
    data = a3)

Residuals:

Min	1Q	Median	3Q	Max
-1.07412	-0.25827	0.05384	0.27675	0.85761

Coefficients:

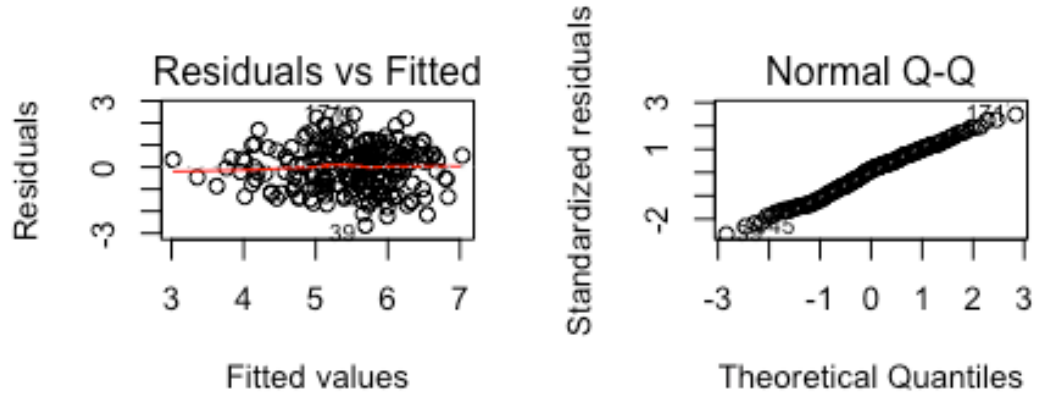
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.5268983	0.3487584	1.511	0.13235	
a3$HSGPA	0.4932945	0.0745553	6.616	3.03e-10	***

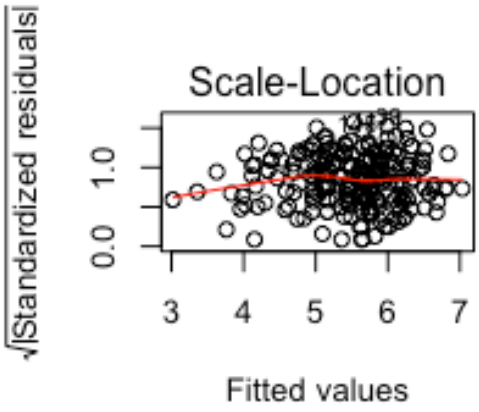
```
a3$SATV                 0.0005919    0.0003945      1.501    0.13498
a3$SATM                 0.0000847    0.0004447      0.190    0.84912
a3$Male                 0.0482478    0.0570277      0.846    0.39850
a3$HU                   0.0161874    0.0039723      4.075 6.53e-05 ***
a3$SS                   0.0073370    0.0055635      1.319    0.18869
a3$FirstGen            -0.0743417    0.0887490     -0.838    0.40318
a3$White                0.1962316    0.0700182      2.803    0.00555 **
a3$CollegeBound         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
```

#code for box-cox test

```
> install.packages("MASS")
> library(MASS)
> bc=boxcox(mod1,lambda=seq(-2,2,by=0.01))
> bc$x[which.max(bc$y)]
[1] 1.41
> mod2 <-
lm(a3$GPA^1.5~a3$HSGPA+a3$SATV+a3$SATM+a3$Male+a3$HU+a3$SS+a3$F
irstGen+a3$White+a3$CollegeBound, data = a3)
> par(mfrow=c(2,2))
> plot(mod2,1)
> plot(mod2,2)
> plot(mod2,3)
```





> summary(mod2)

Call:

lm(formula = a3$GPA^1.5 ~ a3$HSGPA + a3$SATV + a3$SATM + a3$Male +
    a3$HU + a3$SS + a3$FirstGen + a3$White + a3$CollegeBound,
    data = a3)

Residuals:

Min	1Q	Median	3Q	Max
-2.6471	-0.7141	0.1190	0.7233	2.3801

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.1555354	0.9130188	-1.266	0.20706	
a3$HSGPA	1.2861982	0.1951793	6.590	3.52e-10	***

```
a3$SATV                 0.0014604    0.0010327      1.414    0.15882
a3$SATM                 0.0002871    0.0011641      0.247    0.80545
a3$Male                 0.1259103    0.1492936      0.843    0.39998
a3$HU                   0.0413376    0.0103990      3.975 9.69e-05 ***
a3$SS                   0.0170844    0.0145648      1.173    0.24213
a3$FirstGen            -0.2111317    0.2323370     -0.909    0.36454
a3$White                0.4920089    0.1833014      2.684    0.00785 **
a3$CollegeBound         0.0590452    0.2626682      0.225    0.82236
---
Signif. codes:    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.004 on 209 degrees of freedom
Multiple R-squared:    0.343,   Adjusted R-squared:    0.3147
F-statistic: 12.13 on 9 and 209 DF,    p-value: 2.334e-15

> SSE2 = sum( (a3$GPA-(mod2$fitted)^(1/1.5))^2)
> SSE1 = sum( (a3$GPA-(mod1$fitted))^2)
> SSE1
[1] 30.71902
> SSE2
[1] 30.68389
> c(SSE1,SSE2)
[1] 30.71902 30.68389
> c(AIC(mod1),AIC(mod2))
[1] 208.8588 634.8588
> confint(mod2)
                                2.5 %          97.5 %
(Intercept)         -2.9554420159 0.644371117
a3$HSGPA             0.9014256988 1.670970666
a3$SATV             -0.0005755115 0.003496301
a3$SATM             -0.0020077901 0.002581956
a3$Male             -0.1684039497 0.420224562
a3$HU                0.0208371277 0.061837990
```

```
a3$SS                -0.0116283955 0.045797269
a3$FirstGen          -0.6691560550 0.246892727
a3$White              0.1306523351 0.853365534
a3$CollegeBound -0.4587735197 0.576863888
```