GPA

Chris Parrish

January 18, 2016

Contents

GPA

reference:
- Cannon, et al., Stat2, chapter 04, example 4.2

Data

Import the data.

```
data <- read.csv("FirstYearGPA.csv", header=TRUE)
head(data, 4)
```

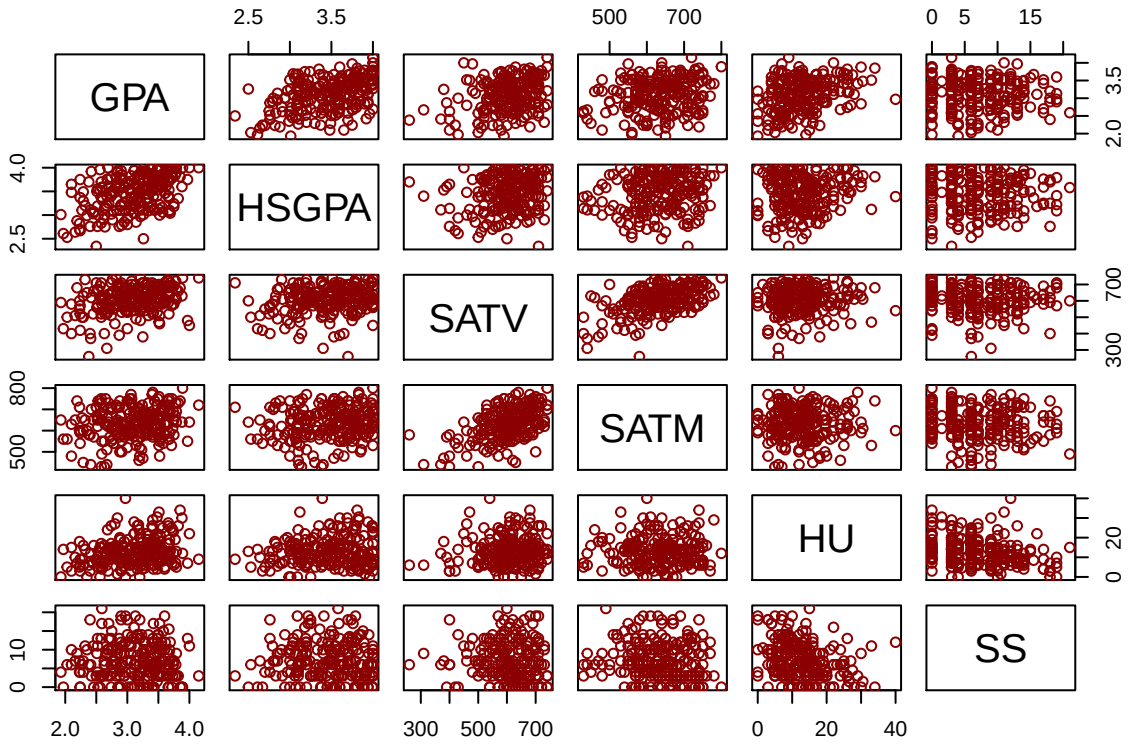
```
##    GPA HSGPA SATV SATM Male HU SS FirstGen White CollegeBound
## 1 3.06  3.83  680  770    1  3  9        1     1            1
## 2 4.15  4.00  740  720    0  9  3        0     1            1
## 3 3.41  3.70  640  570    0 16 13        0     0            1
## 4 3.21  3.51  740  700    0 22  0        0     1            1
```

```
dim(data)
```

```
## [1] 219  10
```

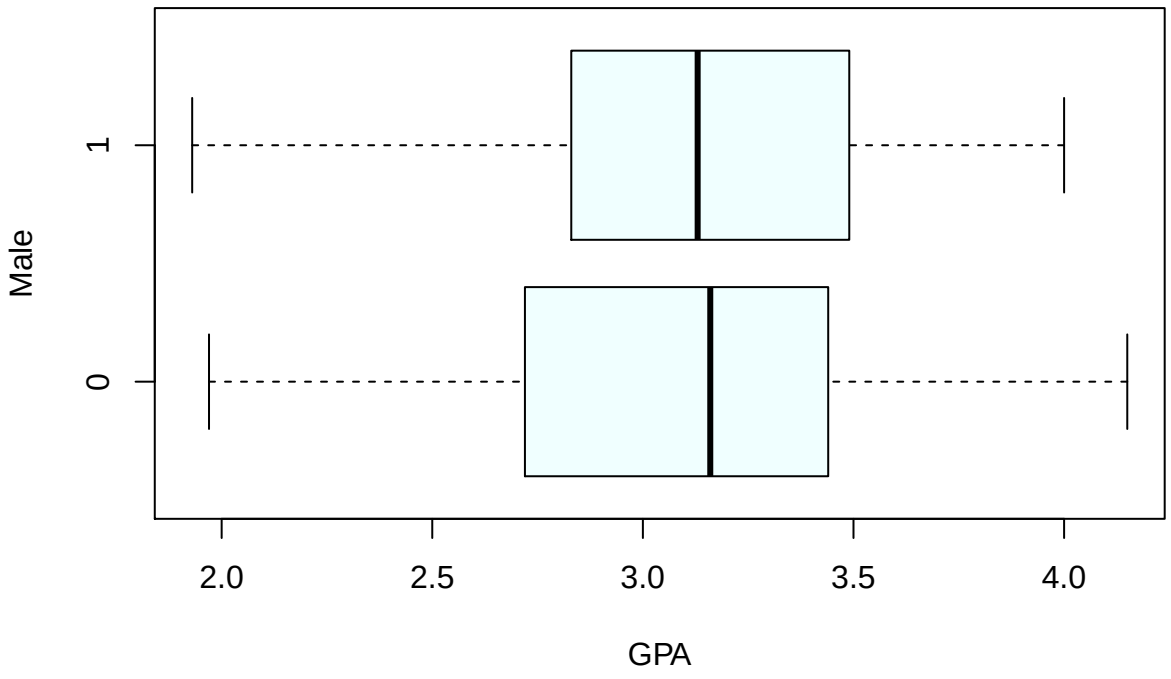
Scatterplot matrix.

```
pairs(~ GPA + HSGPA + SATV + SATM + HU + SS, data=data, col="darkred")
```

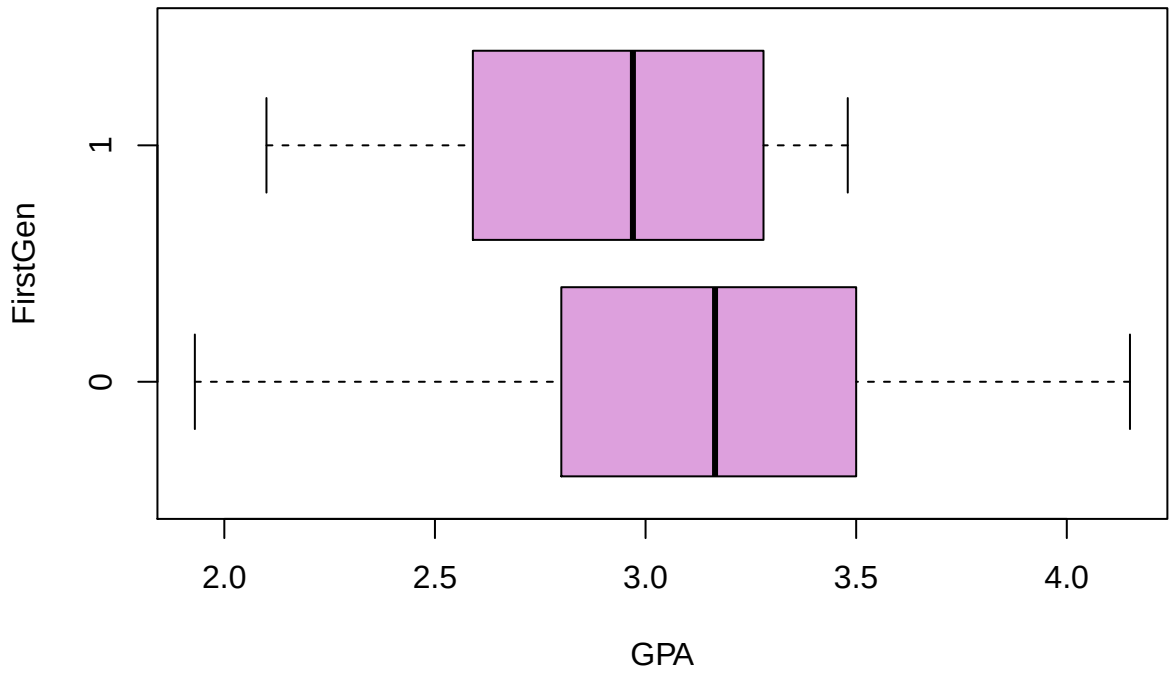


Boxplots.

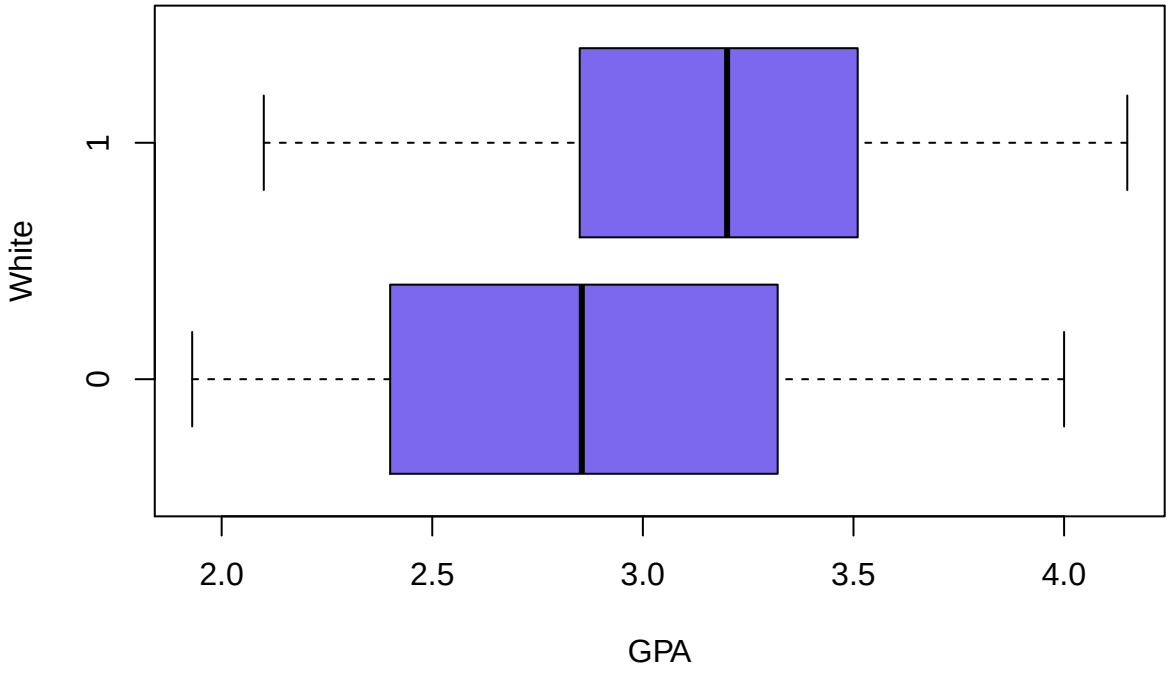
```
boxplot(GPA ~ Male, data=data,
        horizontal=TRUE, col="azure",
        xlab="GPA", ylab="Male")
```



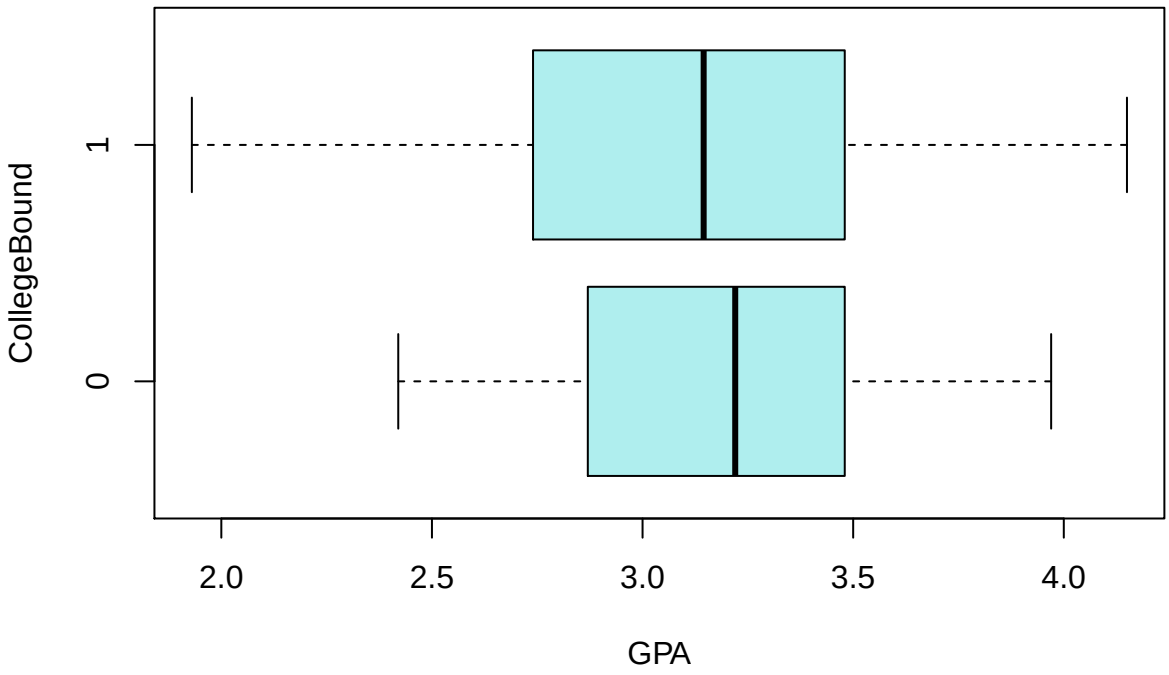

```
boxplot(GPA ~ FirstGen, data=data,
        horizontal=TRUE, col="plum",
        xlab="GPA", ylab="FirstGen")
```




```
boxplot(GPA ~ White, data=data,
        horizontal=TRUE, col="mediumslateblue",
        xlab="GPA", ylab="White")
```




```
boxplot(GPA ~ CollegeBound, data=data,
        horizontal=TRUE, col="paleturquoise",
        xlab="GPA", ylab="CollegeBound")
```



Best subsets

```
full.lm <- lm(GPA ~ HSGPA + SATV + Male + HU + SS + White, data=data)
options(show.signif.stars=FALSE)
summary(full.lm)
```

```
##
## Call:
## lm(formula = GPA ~ HSGPA + SATV + Male + HU + SS + White, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06228 -0.26731  0.05287  0.27230  0.85843
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.5466634  0.2835072   1.928   0.0552
## HSGPA       0.4829491  0.0714659   6.758 1.33e-10
## SATV        0.0006945  0.0003449   2.013   0.0453
## Male        0.0541049  0.0526937   1.027   0.3057
## HU          0.0167958  0.0038181   4.399 1.72e-05
## SS          0.0075702  0.0054421   1.391   0.1657
## White       0.2045215  0.0685954   2.982   0.0032
##
## Residual standard error: 0.3814 on 212 degrees of freedom
## Multiple R-squared:  0.347,  Adjusted R-squared:  0.3285
## F-statistic: 18.78 on 6 and 212 DF,  p-value: < 2.2e-16
```

`leaps`

From Cannon, et al., Student R Manual, chapter 4, p.54.

```
library(leaps)
out <- summary(regsubsets(GPA ~ ., nbest=2, data=data))
names(out)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

Table of included variables

```
subsets <- out$which
included.vars <- matrix(as.numeric(subsets), nrow=16)
rownames(included.vars) <- rownames(subsets)
colnames(included.vars) <- colnames(subsets)
included.vars[ , 2:10]
```

```
##   HSGPA SATV SATM Male HU SS FirstGen White CollegeBound
## 1     1    0    0    0  0  0        0     0            0
## 1     0    0    0    0  1  0        0     0            0
## 2     1    0    0    0  1  0        0     0            0
## 2     1    0    0    0  0  0        0     1            0
## 3     1    0    0    0  1  0        0     1            0
## 3     1    1    0    0  1  0        0     0            0
## 4     1    1    0    0  1  0        0     1            0
## 4     1    0    0    0  1  0        1     1            0
## 5     1    1    0    0  1  1        0     1            0
## 5     1    1    0    1  1  0        0     1            0
## 6     1    1    0    1  1  1        0     1            0
## 6     1    1    0    0  1  1        1     1            0
## 7     1    1    0    1  1  1        1     1            0
## 7     1    1    0    1  1  1        0     1            1
## 8     1    1    0    1  1  1        1     1            1
## 8     1    1    1    1  1  1        1     1            0
```

Table of model statistics.

```
R.sq <- round(100 * out$rsq, 1)
R.sq.adj <- round(100 * out$adjr2, 1)
Cp <- round(out$cp, 1)                        #  Mallow's Cp
best.subsets <- cbind(R.sq, R.sq.adj, Cp)
row.names(best.subsets) <- row.names(subsets)
best.subsets
```

```
##   R.sq R.sq.adj   Cp
## 1 20.0     19.6 42.2
## 1  9.9      9.5 74.5
## 2 27.0     26.3 21.7
## 2 26.8     26.1 22.2
## 3 32.3     31.4  6.5
## 3 30.8     29.8 11.4
## 4 33.7     32.5  3.9
## 4 33.0     31.7  6.4
## 5 34.4     32.8  3.9
## 5 34.1     32.6  4.8
```

```
## 6 34.7     32.9  4.8
## 6 34.6     32.8  5.1
## 7 34.9     32.8  6.1
## 7 34.7     32.6  6.8
## 8 35.0     32.5  8.0
## 8 34.9     32.5  8.0
```

The smallest value for Mallow's Cp is 3.9 for the first model with 4 variables.

Construct that model. All four terms have significant t-tests.

```
best.leap.lm <- lm(GPA ~ HSGPA + SATV + HU + White, data=data)
summary(best.leap.lm)
```

```
##
## Call:
## lm(formula = GPA ~ HSGPA + SATV + HU + White, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06370 -0.26286  0.02436  0.27338  0.87190
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6409767  0.2787933   2.299  0.02246
## HSGPA       0.4761952  0.0710947   6.698 1.83e-10
## SATV        0.0007372  0.0003417   2.157  0.03209
## HU          0.0150566  0.0036383   4.138 5.03e-05
## White       0.2121164  0.0686196   3.091  0.00226
##
## Residual standard error: 0.3824 on 214 degrees of freedom
## Multiple R-squared:  0.3375, Adjusted R-squared:  0.3251
## F-statistic: 27.25 on 4 and 214 DF,  p-value: < 2.2e-16
```

Backward elimination

```
full.lm <- lm(GPA ~ ., data=data)
step(full.lm)
```

```
## Start:  AIC=-410.16
## GPA ~ HSGPA + SATV + SATM + Male + HU + SS + FirstGen + White +
##     CollegeBound
##
##                Df Sum of Sq    RSS     AIC
## - SATM          1    0.0053 30.724 -412.12
## - CollegeBound  1    0.0067 30.726 -412.11
## - FirstGen      1    0.1031 30.822 -411.42
## - Male          1    0.1052 30.824 -411.41
## - SS            1    0.2556 30.975 -410.34
## <none>                      30.719 -410.16
## - SATV          1    0.3309 31.050 -409.81
```

```
## - White         1    1.1545 31.873 -404.08
## - HU            1    2.4409 33.160 -395.41
## - HSGPA         1    6.4345 37.154 -370.51
##
## Step:  AIC=-412.12
## GPA ~ HSGPA + SATV + Male + HU + SS + FirstGen + White + CollegeBound
##
##                Df Sum of Sq    RSS     AIC
## - CollegeBound  1    0.0065 30.731 -414.07
## - FirstGen      1    0.1072 30.832 -413.36
## - Male          1    0.1421 30.866 -413.11
## - SS            1    0.2503 30.975 -412.34
## <none>                      30.724 -412.12
## - SATV          1    0.4532 31.178 -410.91
## - White         1    1.1778 31.902 -405.88
## - HU            1    2.4567 33.181 -397.27
## - HSGPA         1    6.5762 37.301 -371.64
##
## Step:  AIC=-414.07
## GPA ~ HSGPA + SATV + Male + HU + SS + FirstGen + White
##
##            Df Sum of Sq    RSS     AIC
## - FirstGen  1    0.1129 30.844 -415.27
## - Male      1    0.1469 30.878 -415.03
## - SS        1    0.2470 30.978 -414.32
## <none>                  30.731 -414.07
## - SATV      1    0.4677 31.199 -412.77
## - White     1    1.1713 31.902 -407.88
## - HU        1    2.4506 33.181 -399.27
## - HSGPA     1    6.7560 37.487 -372.55
##
## Step:  AIC=-415.27
## GPA ~ HSGPA + SATV + Male + HU + SS + White
##
##         Df Sum of Sq    RSS     AIC
## - Male   1    0.1534 30.997 -416.18
## - SS     1    0.2815 31.125 -415.28
## <none>               30.844 -415.27
## - SATV   1    0.5898 31.434 -413.12
## - White  1    1.2934 32.137 -408.27
## - HU     1    2.8154 33.659 -398.14
## - HSGPA  1    6.6441 37.488 -374.55
##
## Step:  AIC=-416.18
## GPA ~ HSGPA + SATV + HU + SS + White
##
##         Df Sum of Sq    RSS     AIC
## <none>               30.997 -416.18
## - SS     1    0.2951 31.292 -416.11
## - SATV   1    0.7005 31.698 -413.29
## - White  1    1.3133 32.310 -409.10
## - HU     1    2.7987 33.796 -399.25
## - HSGPA  1    6.4968 37.494 -376.51
```

```
##
## Call:
## lm(formula = GPA ~ HSGPA + SATV + HU + SS + White, data = data)
##
## Coefficients:
## (Intercept)        HSGPA         SATV           HU           SS
##   0.5684876    0.4739983    0.0007481    0.0167447    0.0077474
##       White
##   0.2060408
```

Forward selection

reference: variable selection

```
null.lm <- lm(GPA ~ 1, data=data)
step(null.lm, scope=list(lower=null.lm, upper=full.lm), 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
## + SATV          1    4.3741 42.859 -353.22
## + White         1    3.7501 43.483 -350.06
## + SATM          1    1.7840 45.450 -340.37
## + FirstGen      1    1.1580 46.076 -337.37
## <none>                      47.234 -333.94
## + CollegeBound  1    0.1876 47.046 -332.81
## + Male          1    0.1319 47.102 -332.55
## + SS            1    0.0006 47.233 -331.94
##
## Step:  AIC=-380.73
## GPA ~ HSGPA
##
##                Df Sum of Sq    RSS     AIC
## + HU            1    3.3067 34.494 -398.78
## + White         1    3.2292 34.571 -398.28
## + SATV          1    2.1861 35.615 -391.77
## + FirstGen      1    1.6278 36.173 -388.37
## + SATM          1    0.7683 37.032 -383.22
## + Male          1    0.4138 37.387 -381.14
## <none>                      37.801 -380.73
## + CollegeBound  1    0.0342 37.766 -378.93
## + SS            1    0.0008 37.800 -378.73
##
## Step:  AIC=-398.78
## GPA ~ HSGPA + HU
##
##                Df Sum of Sq    RSS     AIC
## + White         1   2.52100 31.973 -413.40
## + SATV          1   1.80435 32.690 -408.54
## + SATM          1   0.86034 33.634 -402.31
```

```
## + FirstGen      1   0.80022 33.694 -401.92
## + Male          1   0.43380 34.060 -399.55
## + SS            1   0.37935 34.115 -399.20
## <none>                      34.494 -398.78
## + CollegeBound  1   0.03905 34.455 -397.02
##
## Step:  AIC=-413.4
## GPA ~ HSGPA + HU + White
##
##                Df Sum of Sq    RSS     AIC
## + SATV          1   0.68060 31.292 -416.11
## + FirstGen      1   0.30945 31.663 -413.53
## <none>                      31.973 -413.40
## + SATM          1   0.28236 31.691 -413.34
## + Male          1   0.27919 31.694 -413.32
## + SS            1   0.27526 31.698 -413.29
## + CollegeBound  1   0.04854 31.924 -411.73
##
## Step:  AIC=-416.11
## GPA ~ HSGPA + HU + White + SATV
##
##                Df Sum of Sq    RSS     AIC
## + SS            1  0.295150 30.997 -416.18
## <none>                      31.292 -416.11
## + Male          1  0.167015 31.125 -415.28
## + FirstGen      1  0.156003 31.136 -415.20
## + SATM          1  0.026915 31.265 -414.30
## + CollegeBound  1  0.013720 31.279 -414.20
##
## Step:  AIC=-416.18
## GPA ~ HSGPA + HU + White + SATV + SS
##
##                Df Sum of Sq    RSS     AIC
## <none>                      30.997 -416.18
## + Male          1  0.153387 30.844 -415.27
## + FirstGen      1  0.119394 30.878 -415.03
## + SATM          1  0.054109 30.943 -414.57
## + CollegeBound  1  0.018808 30.978 -414.32


##
## Call:
## lm(formula = GPA ~ HSGPA + HU + White + SATV + SS, data = data)
##
## Coefficients:
## (Intercept)        HSGPA           HU        White         SATV
##   0.5684876    0.4739983    0.0167447    0.2060408    0.0007481
##          SS
##   0.0077474
```