The initial multiple regression model with GPA as our response, and HSGPA, SATV, SATM, and Male as our predictors, is overall decent at predicting the response, since it achieved a p-value of roughly 0, a decent F-statistic of 17.77, and a fairly good adjusted R-squared of 23.53%. However, from the individual p-values, we see that only HSGPA and SATV demonstrate enough evidence for us to reject the null () in favor of the alternative (). In other words, we suspect that SATM and Male are not needed for our multiple regression model.

After excluding these 2 features from our model, and verifying an improvement in the model’s p-value, F-statistic, and adjusted R-squared, we conclude that the equation of our final model is: , where is GPA, is HSGPA, and is SATV.

solution

library(readr)

## Warning: package 'readr' was built under R version 4.1.2

data <- read\_csv("data.csv")

## Rows: 219 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (10): GPA, HSGPA, SATV, SATM, Male, HU, SS, FirstGen, White, CollegeBound  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

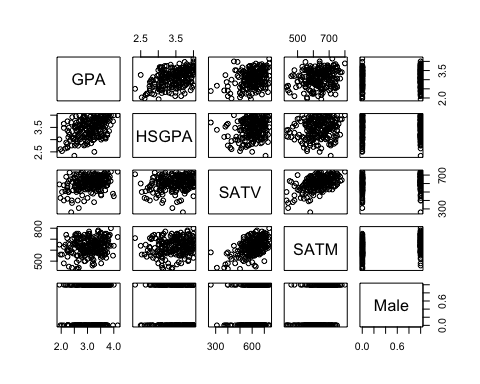
attach(data)  
head(data)

## # A tibble: 6 × 10  
## GPA HSGPA SATV SATM Male HU SS FirstGen White CollegeBound  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 3.06 3.83 680 770 1 3 9 1 1 1  
## 2 4.15 4 740 720 0 9 3 0 1 1  
## 3 3.41 3.7 640 570 0 16 13 0 0 1  
## 4 3.21 3.51 740 700 0 22 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 3 0 1 1

model = lm(GPA ~ HSGPA + SATV + SATM + Male)  
summary(model)

##   
## Call:  
## lm(formula = GPA ~ HSGPA + SATV + SATM + Male)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.95975 -0.27713 0.05058 0.28319 0.89525   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.135e-01 3.283e-01 1.869 0.06301 .   
## HSGPA 5.069e-01 7.623e-02 6.650 2.4e-10 \*\*\*  
## SATV 1.174e-03 3.940e-04 2.979 0.00322 \*\*   
## SATM -5.580e-06 4.626e-04 -0.012 0.99039   
## Male 5.534e-02 6.020e-02 0.919 0.35901   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.407 on 214 degrees of freedom  
## Multiple R-squared: 0.2494, Adjusted R-squared: 0.2353   
## F-statistic: 17.77 on 4 and 214 DF, p-value: 1.298e-12

plot(data[c(1, 2, 3, 4, 5)])



cor(data[c(1, 2, 3, 4, 5)])

## GPA HSGPA SATV SATM Male  
## GPA 1.00000000 0.44688735 0.3043114 0.1943439 0.05284917  
## HSGPA 0.44688735 1.00000000 0.2103212 0.1528396 -0.09031714  
## SATV 0.30431137 0.21032124 1.0000000 0.5269438 0.14555703  
## SATM 0.19434385 0.15283963 0.5269438 1.0000000 0.37099167  
## Male 0.05284917 -0.09031714 0.1455570 0.3709917 1.00000000

anova(model)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)   
## HSGPA 1 9.433 9.4329 56.9351 1.275e-12 \*\*\*  
## SATV 1 2.186 2.1861 13.1947 0.0003514 \*\*\*  
## SATM 1 0.019 0.0193 0.1165 0.7331979   
## Male 1 0.140 0.1400 0.8450 0.3590096   
## Residuals 214 35.455 0.1657   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

improved\_model = lm(GPA ~ HSGPA + SATV)  
summary(improved\_model)

##   
## Call:  
## lm(formula = GPA ~ HSGPA + SATV)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.97894 -0.27639 0.02867 0.30133 0.87956   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6351217 0.2955033 2.149 0.03272 \*   
## HSGPA 0.4975320 0.0750569 6.629 2.66e-10 \*\*\*  
## SATV 0.0012283 0.0003373 3.641 0.00034 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4061 on 216 degrees of freedom  
## Multiple R-squared: 0.246, Adjusted R-squared: 0.239   
## F-statistic: 35.23 on 2 and 216 DF, p-value: 5.711e-14