Final Project R Supplement, Stat 230

Elena Ea, Ben Griesel, Helen Moses

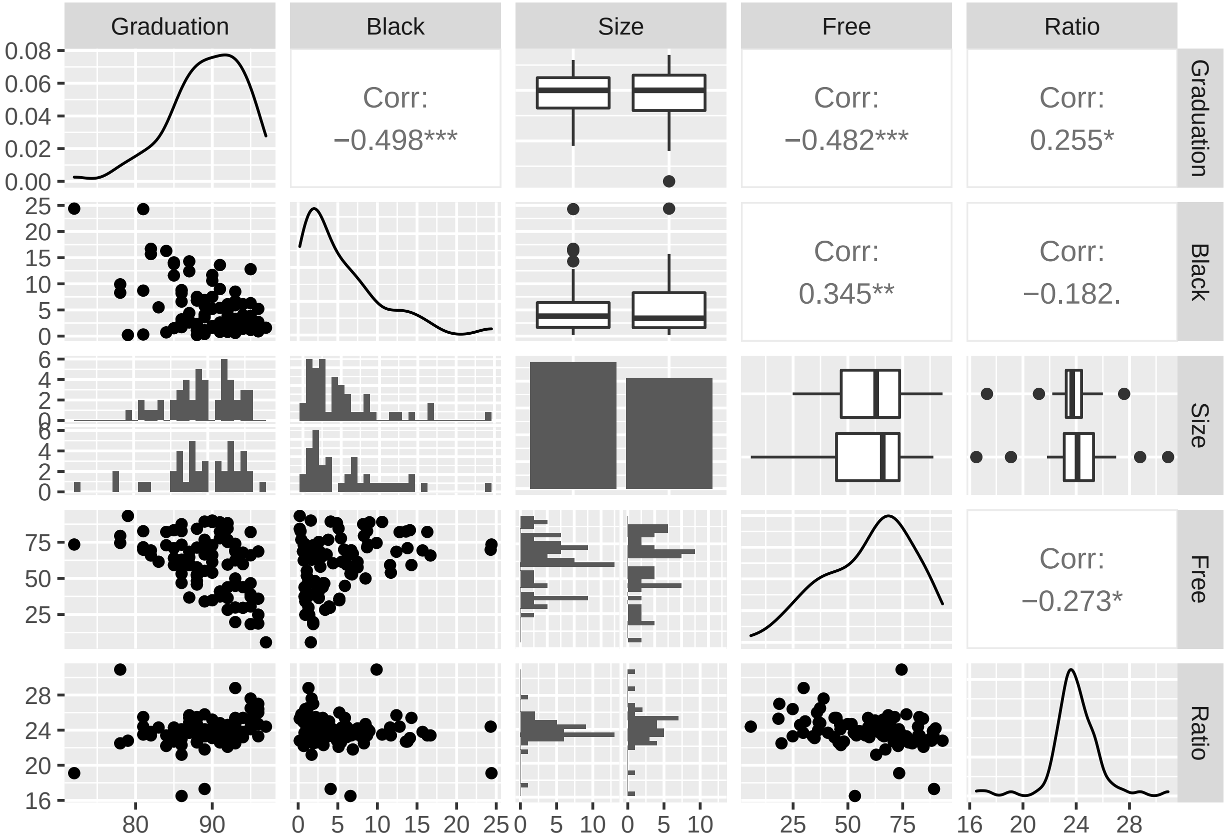
11/15/2022

1. Visualize the Data

```
# Making it so R recognizes our categorical variable as a categorical variable
grad$Size <- as.factor(grad$Size)
```

```
#Plotting all of the pairwise combinations of the variables in a scatterplot format to get a visualizat
ggpairs(grad, columns = c(9,6,10,11,12))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There appears to be some concerns with the variable free. However to confirm this, we first want to check the std errors for each variable in the model.

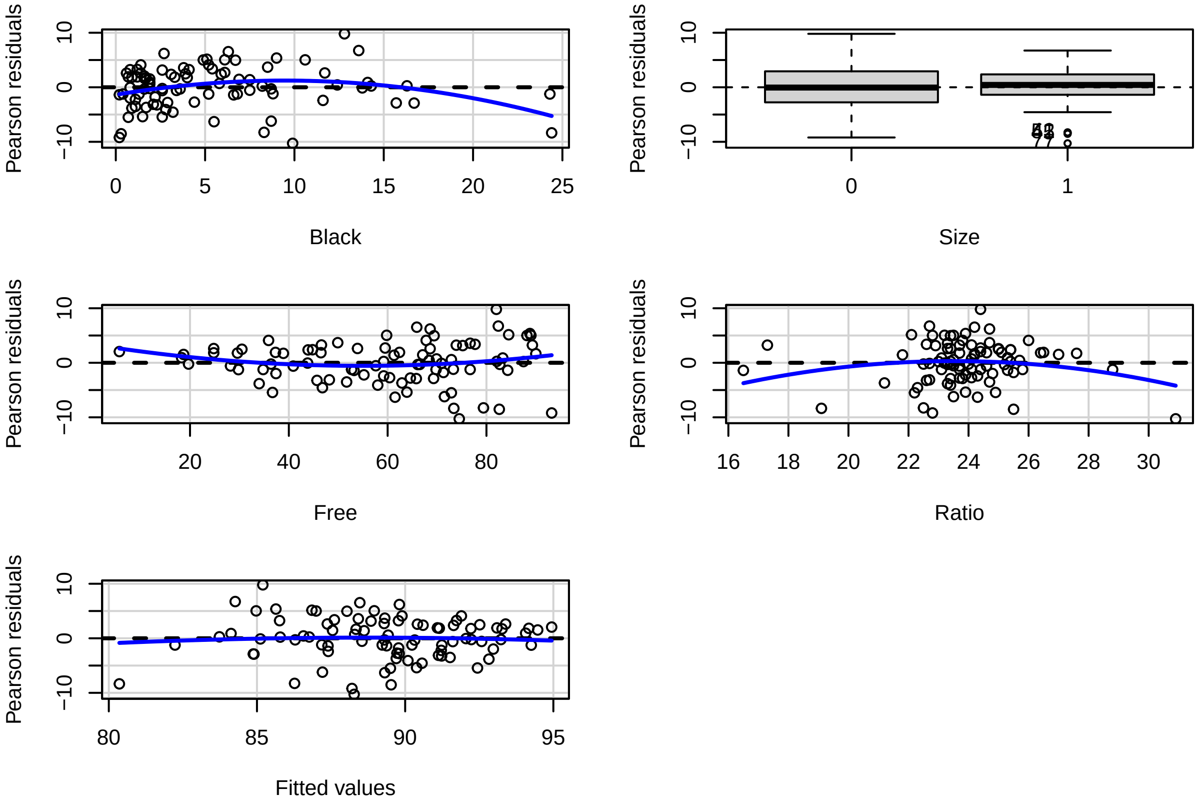
```
# Fitting our primary model
grad_mlr <- lm(Graduation ~ Black + Size + Free + Ratio, data = grad)
summary(grad_mlr)
```

```
##
## Call:
## lm(formula = Graduation ~ Black + Size + Free + Ratio, data = grad)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2788  -2.2648   0.2257   2.5922   9.8021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89.84171    6.11515  14.692  < 2e-16 ***
## Black       -0.34250    0.08850  -3.870 0.000216 ***
## Size1       -0.18834    0.85874  -0.219 0.826938
## Free        -0.07992    0.02321  -3.443 0.000904 ***
## Ratio        0.25794    0.23551   1.095 0.276591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.963 on 83 degrees of freedom
## Multiple R-squared:  0.3664, Adjusted R-squared:  0.3359
## F-statistic:    12 on 4 and 83 DF,  p-value: 9.636e-08
```

It does appear that the smallest standard error is not within two of the largest, which does raise some concerns about the equal variance model assumptions.

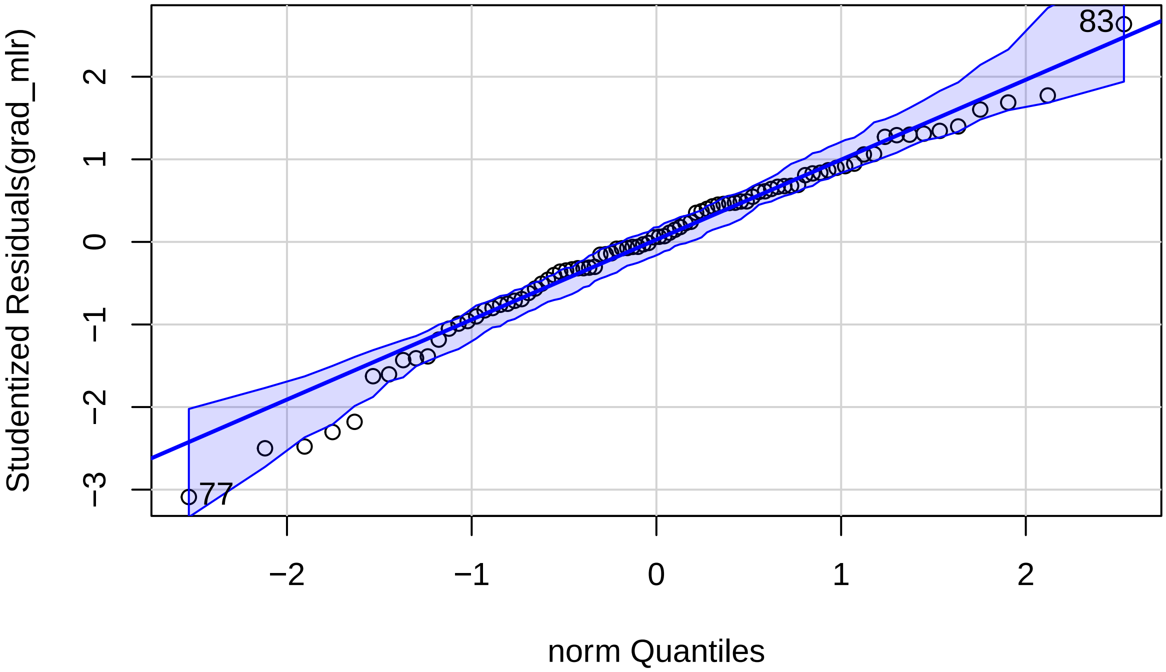
```
# Checking the residual plots
residualPlots(grad_mlr)
```



```
##            Test stat Pr(>|Test stat|)
## Black        -2.7486         0.007359 **
## Size
## Free          1.4204         0.159280
## Ratio        -1.8846         0.063026 .
## Tukey test   -0.4958         0.620011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There appears to be a wedge in the Free variable plot indicating non constant variance.

```
# Checking the normality assumption
qqPlot(grad_mlr, type = "rstandard", distribution = "norm")
```



```
## [1] 77 83
```

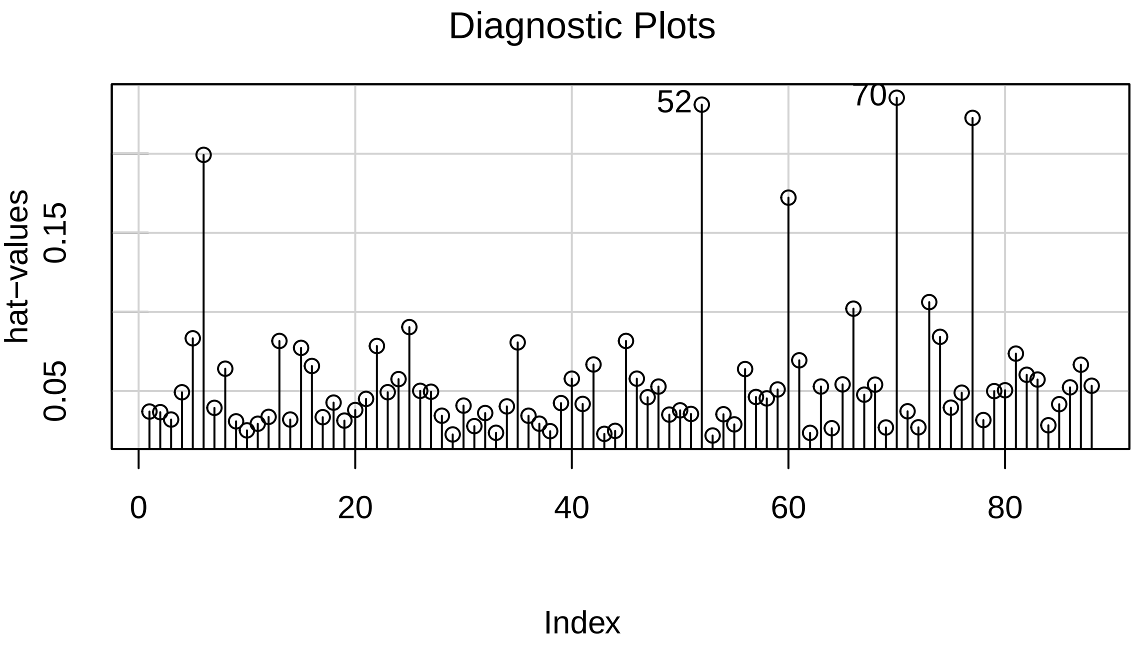
There appears to be a slight skew in the data, but it does not appear to be significant enough to raise any concerns regarding the normality of the errors.

```
#Checking for multicollinearity
vif(grad_mlr)
```

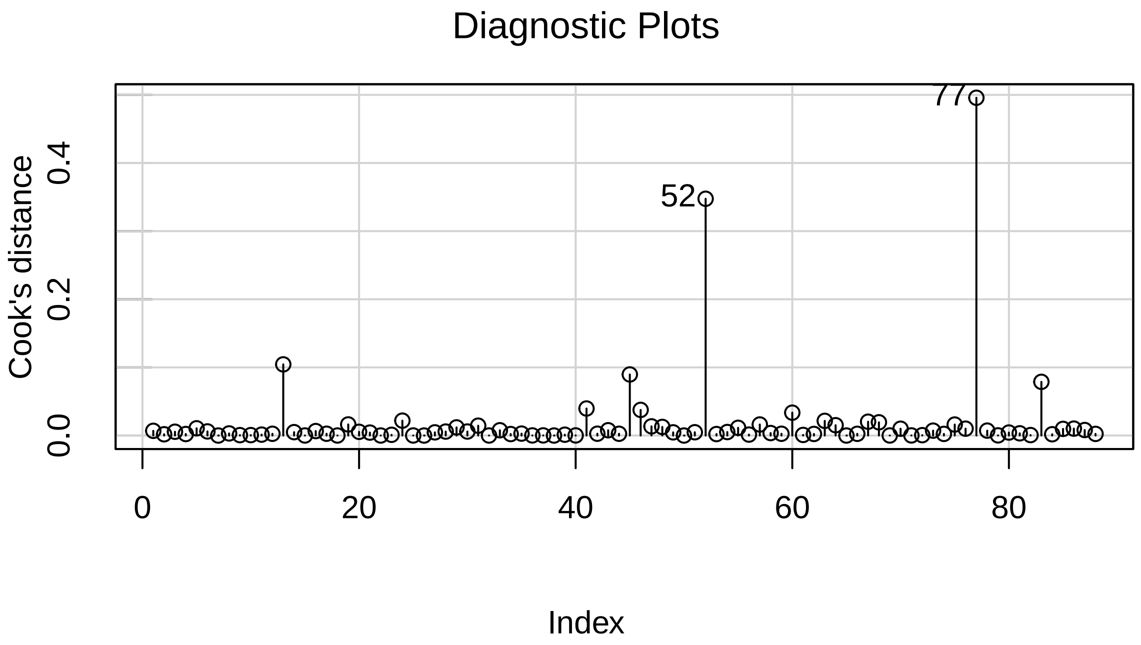
```
##    Black     Size     Free    Ratio 
## 1.161305 1.027984 1.206477 1.101654
```

There do not appear to be any signs of multicollinearity!

```
#Checking to confirm there are no outliers
infIndexPlot(grad_mlr, vars = "hat")
```



```
infIndexPlot(grad_mlr, vars = "Cook")
```



None of the values are close to one, so it does not appear as though there are any significant outliers in our model.

*We did not run any tests for independence because we confirmed that assumption by checking how the data was collected.

2. Transforming the Model

```
# Adding several transformation to be able to try out various different models, to try and remedy the m
grad <- grad |>
  mutate(logBlack = log(Black))
grad <- grad |>
```

```
  mutate(logGraduation = log(Graduation))
grad <- grad |>
  mutate(logFree = log(Free))
grad <- grad |>
  mutate(Free2 = Free^2)
# Centered the Free data
grad$Free_center <- grad$Free- mean(grad$Free)
```

With the time that we had, we could not find any suitable transformations, so we decided to stick with the original model but mention the deficiencies in our report.

3. Are the Variables Significant?

From the R output when we fitted our model, we can find the test statistics and corresponding p-values for each variable. To confirm these results, we also wanted to include a confidence interval test.

```
# Constructing a 95% confidence interval
confint.default(grad_mlr, level = .95)
```

```
##                   2.5 %       97.5 %
## (Intercept) 77.8562440 101.82717616
## Black       -0.5159599  -0.16904455
## Size1       -1.8714342   1.49475426
## Free        -0.1254161  -0.03442345
## Ratio       -0.2036587   0.71952908
```