Colleges Project

Celeste Vargas

3/15/2022

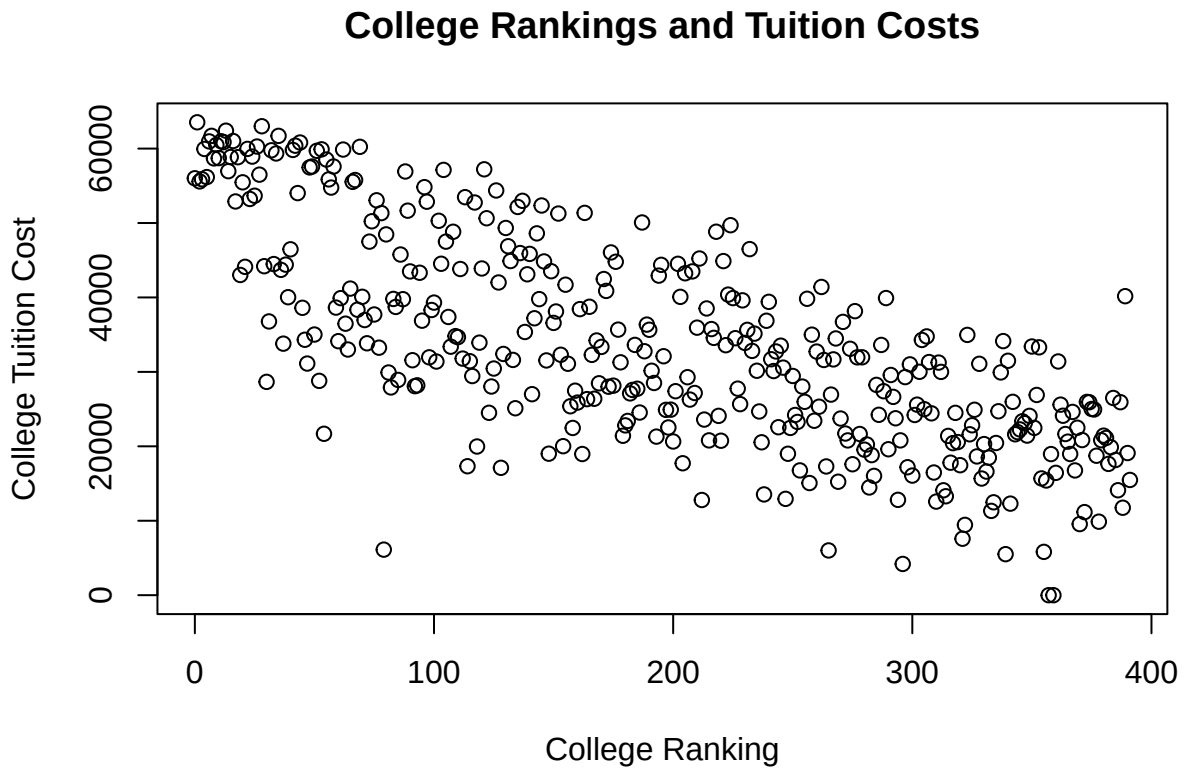
```
colleges <- read.csv("C://Projects//colleges.csv")
```

Question:

Does college ranking indicate higher tuition costs?

So far, there seems to be a relationship between college rankings and tuition costs.

```
plot(colleges$X, colleges$Tuition, main = "College Rankings and Tuition Costs",
     ylab = "College Tuition Cost", xlab = "College Ranking") #Plot tuition against rankings
```



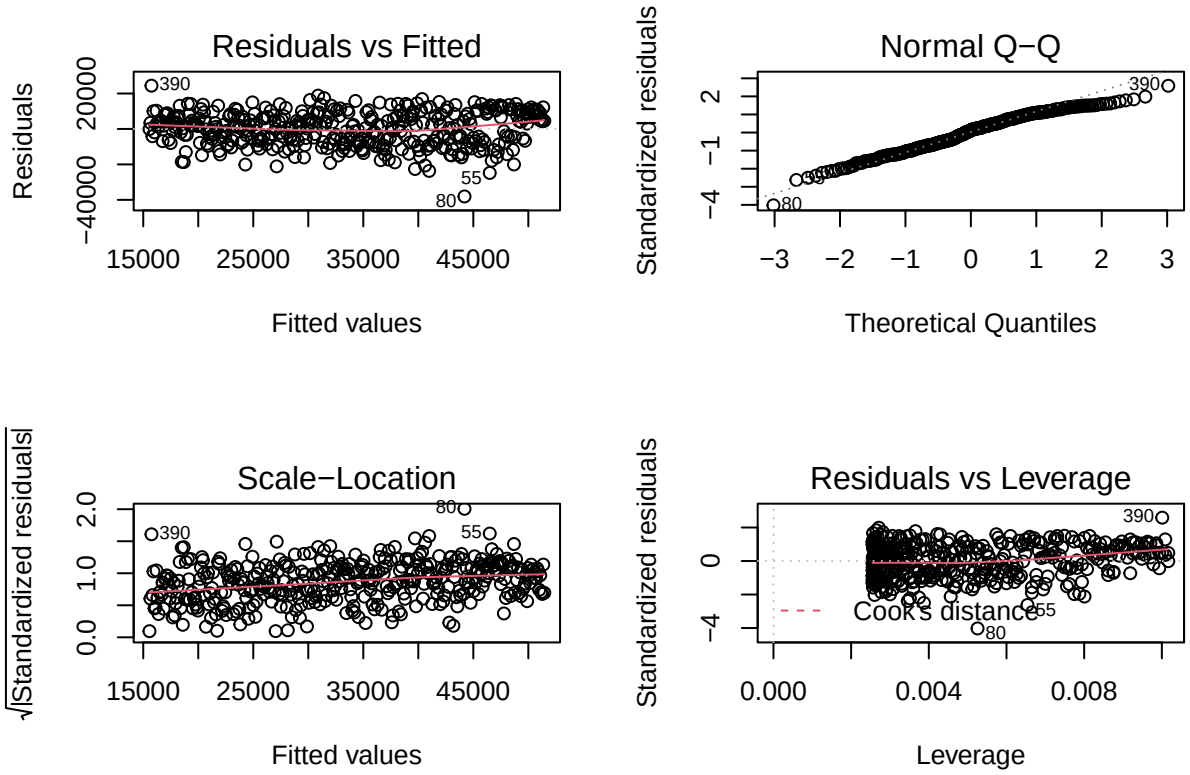
Observations: College rankings seem to be statistically significant, so there seems to be evidence that the lower the college ranking, the less the tuition might cost. The adjusted R squared seems to indicate that the variable of college ranking seem to explain around half of the variability in the data.

```r
lm_colleges <- lm(colleges$Tuition ~ colleges$X) #We will perform a simple linear regression analysis
summary(lm_colleges) #perform t-test on ranking
```

```
##
## Call:
## lm(formula = colleges$Tuition ~ colleges$X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38079  -6891   1224   7556  24420
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51448.532    956.344   53.80   <2e-16 ***
## colleges$X    -91.769      4.234  -21.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9485 on 390 degrees of freedom
## Multiple R-squared:  0.5464, Adjusted R-squared:  0.5453
## F-statistic: 469.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

There are outliers in the data, but the scale-location graph has a pretty horizontal red line, and the Q-Q plot, given that the data set is not really big, it shows that the residuals are close to the middle of the line. These indicate that the model is valid.

```r
par(mfrow=c(2,2))
plot(lm_colleges) #residual analysis
```



We will perform a multiple regression analysis to see if adding the variable of enrollment numbers would explain more variability in the data.

As we can see, the adjusted R square is now 0.6485, meaning adding the new variable does explain more variability in the data, but there could be more variables that do so as well.

Enrollment numbers is also statistically significant when college rankings is also in the model, meaning there exists an association between enrollment numbers and tuition costs.

```
lm_colleges_mult <- lm(colleges$Tuition ~ colleges$X + colleges$Enrollment.Numbers)
summary(lm_colleges_mult)
```

```
##
## Call:
## lm(formula = colleges$Tuition ~ colleges$X + colleges$Enrollment.Numbers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30915.4  -5369.7    994.8   5894.1  21902.8
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.818e+04  1.048e+03   55.51   <2e-16 ***
## colleges$X                  -9.986e+01  3.797e+00  -26.30   <2e-16 ***
## colleges$Enrollment.Numbers -3.972e-01  3.695e-02  -10.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8339 on 389 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.6485
## F-statistic: 361.7 on 2 and 389 DF,  p-value: < 2.2e-16
```

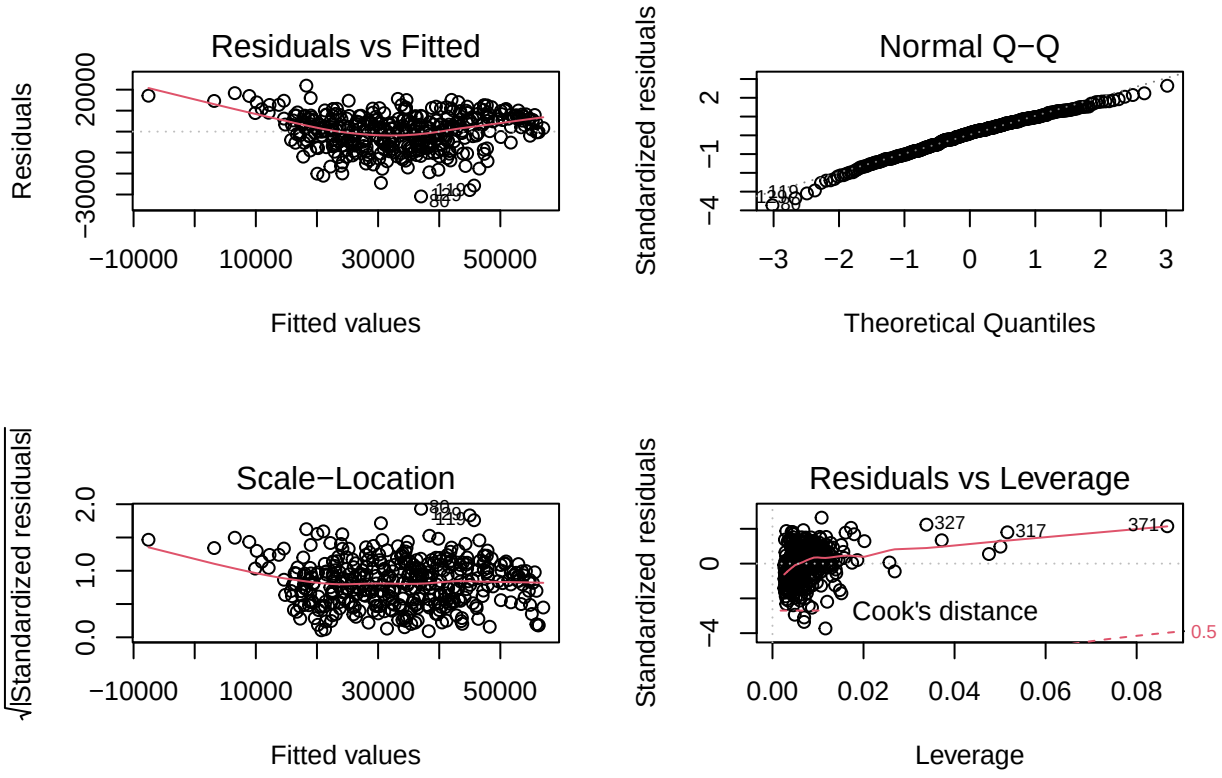
```
library(car)
```

```
## Loading required package: carData
```

```
vif(lm_colleges_mult) #no serious multicolinearity
```

```
##                   colleges$X colleges$Enrollment.Numbers
##                     1.040854                    1.040854
```

```
par(mfrow=c(2,2))
plot(lm_colleges_mult) #residual analysis
```



Seeing the scale location graph, next project we will try to optimize the multiple regression analysis model to try to obtain constant variance.

Conclusion for project 1: Overall, according to the findings of the simple linear regression analysis, there is evidence that a higher college ranking is associated with higher tuition costs.