---

output: pdf_document: latex_engine: pdflatex —

Regression

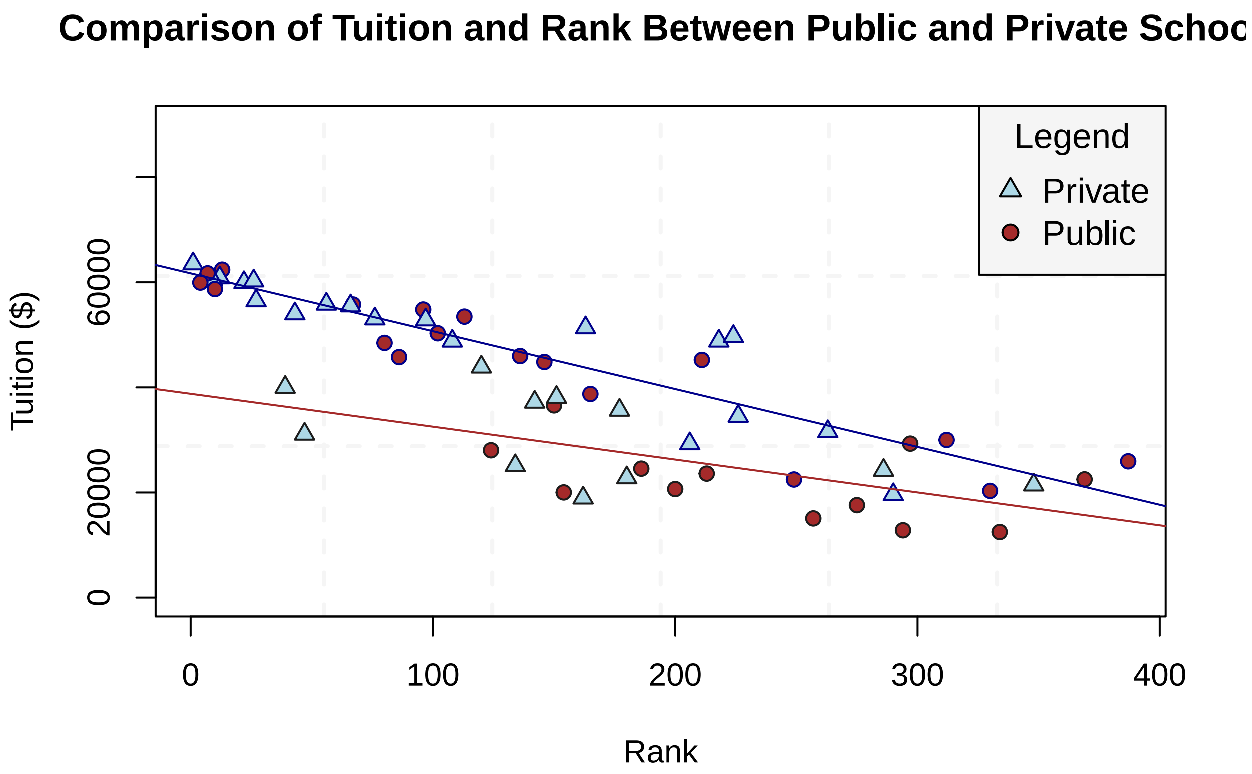
Set up

```
library(dplyr)
library(caret)
library(stargazer)
library(tinytex)

# Load Data from previous section
obs_60_final<- read.csv('C:\\Users\\roryq\\Downloads\\Stat 1223\\obs_60_final.csv')


# Filter by private or public schools
Private_60 = obs_60_final[which(obs_60_final$institutionalControl == "private"),]
Private_60<- Private_60 %>% select(Tuition,Expend,Median_Income, number_Undergrads,Rank)
Public_60 = obs_60_final[which(obs_60_final$institutionalControl == "public"),]
Public_60<- Public_60 %>% select(Tuition,Expend,Median_Income, number_Undergrads,Rank)
```

Check for Interaction Terms

```
# Check for interaction terms


# Create 2 linear regression models one with private and one
  # with public to compare expenditure per student and tuition levels
model_pri60 = lm(Tuition ~ Rank, data = Private_60)
model_pub60 = lm(Tuition ~ Rank, data = Public_60)

plot.new() # Add grid to look pretty
grid(nx = 6, # X-axis divided in two sections
 ny = 3, # Y-axis divided in three sections
 lty = 2, col = "gray96", lwd = 2)
par(new = TRUE)

# Scatterplot with groups
# Specify colors to be used in scatterplot
colors = c("darkblue", "gray11")
plot(obs_60_final$Rank, obs_60_final$Tuition, pch = c(24,21),
col = colors[factor(obs_60_final$institutionalControl)],bg=c("lightblue", "brown") ,
xlab = "Rank", ylab = "Tuition ($)" , ylim= c(0,90000),
main= "Comparison of Tuition and Rank Between Public and Private School")

abline(model_pri60, col = "darkblue")  # Plot the regression line for private colleges

abline(model_pub60, col = "brown")   # Plot the regression line for public colleges
```

```
# Add legend
legend("topright", title="Legend", legend=
c("Private","Public "),pt.bg=c("light blue", "brown"),bg=
"whitesmoke", pch= c(24,21),cex=1.1)
```



```
rbind(confint(model_pri60,'Rank',level=0.975),confint(model_pub60,'Rank',level=0.975))
```

```
##         1.25 %   98.75 %
## Rank -129.4187 -90.79403
## Rank -102.3979 -22.75579
```

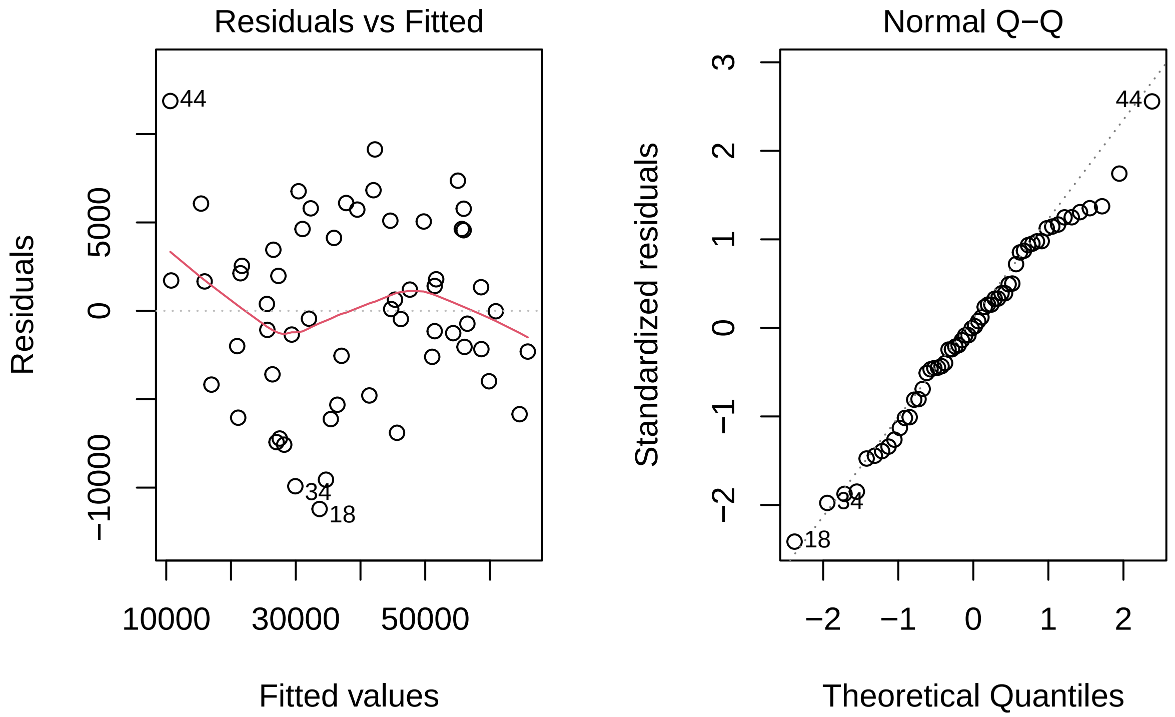
- From the graphs the regression lines intersect, again suggesting an interaction term. However with further inspection we can see the confidence intervals for the slopes of the two regressions overlap, thins indicates that there isnt a significant difference between them for our purposes.

Fit Full Model

```
# Create linear regression model with all factors we are interested
 model = lm(Tuition ~ Rank+S.F.Ratio+Unemployment+Diversity_Rank_Race+ Expend+perc.alumni
+institutionalControl+number_Undergrads+
Median_Income+Grad.Rate+ Crime.Rate+Cost_of_Living+AVG_C_two_I , data = obs_60_final)

# Print model summary
summary(model)
```

```
## 
## Call:
## lm(formula = Tuition ~ Rank + S.F.Ratio + Unemployment + Diversity_Rank_Race + 
##     Expend + perc.alumni + institutionalControl + number_Undergrads + 
##     Median_Income + Grad.Rate + Crime.Rate + Cost_of_Living + 
##     AVG_C_two_I, data = obs_60_final)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11209.1  -3345.0     37.4   4450.9  11867.8 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 3.880e+04  1.558e+04   2.491 0.016575 *  
## Rank                       -9.456e+01  1.398e+01  -6.766 2.52e-08 ***
## S.F.Ratio                   7.410e+01  2.988e+02   0.248 0.805261    
## Unemployment                1.624e+05  9.704e+04   1.674 0.101324    
## Diversity_Rank_Race        -1.226e+00  1.200e+00  -1.022 0.312542    
## Expend                      1.936e-01  1.577e-01   1.228 0.226014    
## perc.alumni                -9.598e+01  1.050e+02  -0.914 0.365768    
## institutionalControlpublic -1.136e+04  2.752e+03  -4.127 0.000161 ***
## number_Undergrads          -3.789e-01  2.013e-01  -1.883 0.066349 .  
## Median_Income               2.763e-01  1.393e-01   1.983 0.053611 .  
## Grad.Rate                  -2.929e+01  7.151e+01  -0.410 0.684125    
## Crime.Rate                 -4.661e+04  5.923e+04  -0.787 0.435571    
## Cost_of_Living             -1.470e-01  1.459e-01  -1.007 0.319266    
## AVG_C_two_I                 1.066e+04  1.269e+04   0.840 0.405353    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5818 on 44 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8908, Adjusted R-squared:  0.8585 
## F-statistic: 27.61 on 13 and 44 DF,  p-value: < 2.2e-16
```

```r
# Check model assumptions
par(mfrow= c(1,2))
plot(model, which= c(1,2))
```



- Residuals appear randomly dispersed around zero, implying there is no heteroskewdasticity
- QQ plot appears to follow a straight line, although extreme outliers at the top of the range begin to affect the very top of the plot, showing that our observations are approximately normal with a slight left skew

Model Selection

```
# Model selection
# Use Forward and Backward Stepwise Regression Selection (AIC)

min_model = lm(Tuition ~ 1, data = obs_60_final)
max_model = formula(lm(Tuition ~ Rank + S.F.Ratio + Unemployment + Diversity_Rank_Race +
Expend+ institutionalControl+number_Undergrads+Median_Income
+Grad.Rate+Crime.Rate+Cost_of_Living, data = obs_60_final))
best_model = step(min_model, direction = "both", scope = max_model)
```

```
## Start:  AIC=1159.49
## Tuition ~ 1
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                          Df  Sum of Sq        RSS    AIC
## + Rank                    1 8786615224 4.8526e+09 1062.1
## + Grad.Rate             1 6543660673 7.0956e+09 1084.1
## + institutionalControl  1 6210508037 7.4287e+09 1086.8
## + S.F.Ratio             1 5841327150 7.7979e+09 1089.6
## + Expend                1 5730376759 7.9089e+09 1090.4
## + Cost_of_Living        1 2080042695 1.1559e+10 1112.4
## + number_Undergrads     1 2048877599 1.1590e+10 1112.5
## + Unemployment          1 1265208252 1.2374e+10 1116.3
## + Diversity_Rank_Race   1 1053308676 1.2586e+10 1117.3
## + Median_Income         1  896111172 1.2743e+10 1118.0
## <none>                              1.3639e+10 1120.0
## + Crime.Rate            1   58027120 1.3581e+10 1121.8
##
## Step:  AIC=1096.98
## Tuition ~ Rank

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit

##                        Df  Sum of Sq        RSS    AIC
## + institutionalControl  1 2566878018 2.2858e+09 1020.4
## + number_Undergrads     1 1993859369 2.8588e+09 1033.4
## + S.F.Ratio             1  599239985 4.2534e+09 1056.4
## + Cost_of_Living        1  516535540 4.3361e+09 1057.5
## + Grad.Rate             1  499040782 4.3536e+09 1057.8
## + Median_Income         1  442716657 4.4099e+09 1058.5
## + Expend                1  380271112 4.4724e+09 1059.3
## + Diversity_Rank_Race   1  160368441 4.6923e+09 1062.1
## + Unemployment          1    2279317 4.8504e+09 1064.0
## + Crime.Rate            1     482356 4.8522e+09 1064.0
## <none>                              4.8908e+09 1097.0
## - Rank                  1 9442427717 1.4333e+10 1159.5
##
## Step:  AIC=1053.45
## Tuition ~ Rank + institutionalControl

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit

##                        Df  Sum of Sq        RSS    AIC
## + Median_Income         1  415807153 1869957307 1010.8
## + Cost_of_Living        1  242401872 2043362587 1015.9
## + number_Undergrads     1   81995697 2203768762 1020.3
## + Expend                1   55795938 2229968521 1021.0
## + Crime.Rate            1   17654348 2268110112 1021.9
## + Diversity_Rank_Race   1   13692198 2272072262 1022.0
## + Grad.Rate             1   11002050 2274762410 1022.1
## + S.F.Ratio             1    3866790 2281897670 1022.3
## + Unemployment          1     580703 2285183756 1022.4
## <none>                              2290077189 1053.5
## - institutionalControl  1 2600751657 4890828847 1097.0
## - Rank                  1 5354017257 7644094447 1123.8
##
## Step:  AIC=1043.31
## Tuition ~ Rank + institutionalControl + Median_Income
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                        Df  Sum of Sq        RSS    AIC
## + Diversity_Rank_Race   1   88696636 1781260671 1009.9
## + number_Undergrads     1   80471866 1789485441 1010.2
## + Unemployment          1   62708007 1807249300 1010.8
## + Expend                1   36644244 1833313063 1011.6
## + Crime.Rate            1   31952914 1838004393 1011.8
## + S.F.Ratio             1   26231231 1843726076 1011.9
## + Grad.Rate             1    4266839 1865690468 1012.6
## + Cost_of_Living        1    1584200 1868373106 1012.7
## <none>                               1870561072 1043.3
## - Median_Income         1  419516118 2290077189 1053.5
## - institutionalControl  1 2562373052 4432934123 1093.1
## - Rank                  1 5021704222 6892265293 1119.6
##
## Step:  AIC=1042.43
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                        Df  Sum of Sq        RSS    AIC
## + Unemployment          1   90042221 1691218450 1008.9
## + number_Undergrads     1   68996034 1712264637 1009.6
## + Expend                1   32649244 1748611427 1010.9
## + S.F.Ratio             1   16625754 1764634917 1011.4
## + Crime.Rate            1   10865977 1770394694 1011.6
## + Grad.Rate             1      59081 1781201590 1011.9
## + Cost_of_Living        1      30303 1781230368 1011.9
## <none>                               1782767956 1042.4
## - Diversity_Rank_Race   1   87793115 1870561072 1043.3
## - Median_Income         1  492521434 2275289390 1055.1
## - institutionalControl  1 2647980574 4430748530 1095.0
## - Rank                  1 5109484417 6892252373 1121.6
##
## Step:  AIC=1041.29
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                        Df  Sum of Sq        RSS    AIC
## + number_Undergrads     1   96771778 1594446672 1007.5
## + Expend                1   49741762 1641476688 1009.2
## + S.F.Ratio             1   33380839 1657837611 1009.8
## + Cost_of_Living        1    9440520 1681777929 1010.6
## + Crime.Rate            1    3103079 1688115371 1010.8
## + Grad.Rate             1      10108 1691208342 1010.9
## <none>                               1691963378 1041.3
## - Unemployment          1   90804578 1782767956 1042.4
## - Diversity_Rank_Race   1  115638270 1807601648 1043.3
## - Median_Income         1  583214309 2275177687 1057.1
## - institutionalControl  1 2622018821 4313982199 1095.5
## - Rank                  1 3874111872 5566075250 1110.7
##
## Step:  AIC=1039.75
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment + number_Undergrads
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                        Df  Sum of Sq        RSS    AIC
## + Expend                1   33830118 1560616555 1008.3
## + S.F.Ratio             1    8250107 1586196566 1009.2
## + Grad.Rate             1    6958332 1587488340 1009.2
## + Crime.Rate            1    5702423 1588744249 1009.3
## + Cost_of_Living        1    5024640 1589422032 1009.3
## <none>                               1594947456 1039.8
## - number_Undergrads     1   97015922 1691963378 1041.3
## - Diversity_Rank_Race   1  106456269 1701403725 1041.6
## - Unemployment          1  117399840 1712347296 1042.0
## - Median_Income         1  590406576 2185354033 1056.6
## - institutionalControl  1  632111004 2227058460 1057.8
## - Rank                  1 3918980495 5513927952 1112.2
##
## Step:  AIC=1040.76
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment + number_Undergrads + Expend
```

```
# View best model
best_model
```

```
##
## Call:
## lm(formula = Tuition ~ Rank + institutionalControl + Median_Income +
##     Diversity_Rank_Race + Unemployment + number_Undergrads +
##     Expend, data = obs_60_final)
##
## Coefficients:
##                (Intercept)                          Rank
##                  4.303e+04                    -8.751e+01
## institutionalControlpublic                 Median_Income
##                 -1.129e+04                     1.639e-01
##        Diversity_Rank_Race                  Unemployment
##                 -1.909e+00                     1.731e+05
##          number_Undergrads                        Expend
##                 -2.744e-01                     9.929e-02
```

```
# Model validation
# Use Leave One Our Cross Validation

ctrl = trainControl(method = "LOOCV")
```

```
model1 = train(Tuition ~ Rank + institutionalControl + Median_Income +
Diversity_Rank_Race + Unemployment + number_Undergrads + Expend,
data = obs_60_final, method = "lm", trControl = ctrl)
model1$results
```

```
##   intercept     RMSE  Rsquared      MAE
## 1      TRUE 5988.835 0.8513063 4797.676
```

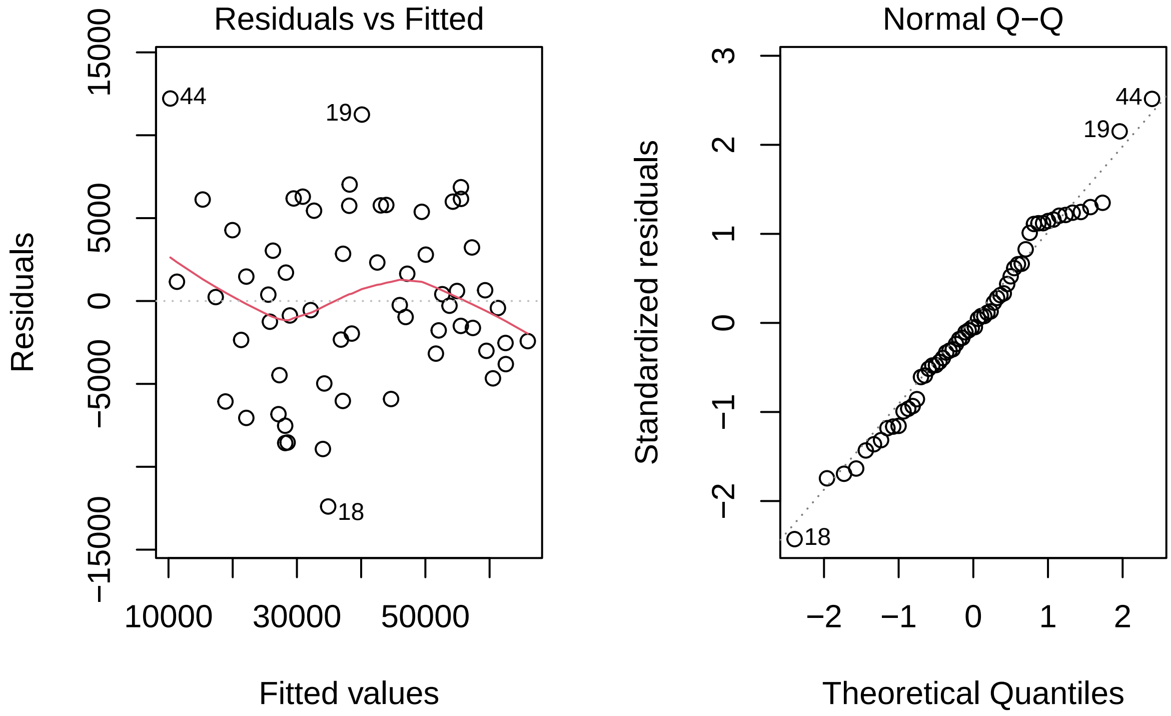
```
# print summary of best model
stargazer(best_model, se = list(best_model$rse))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sat, Oct 26, 2024 - 2:24:56 PM

Table 1:

	Dependent variable:
	Tuition
Rank	-87.514^{***}
	(10.174)
institutionalControlpublic	$-11{,}292.310^{***}$
	(2,497.191)
Median_Income	0.164^{***}
	(0.037)
Diversity_Rank_Race	-1.909^{*}
	(1.017)
Unemployment	$173{,}051.400^{**}$
	(84,461.880)
number_Undergrads	−0.274
	(0.166)
Expend	0.099
	(0.107)
Constant	$43{,}034.530^{***}$
	(6,028.025)
Observations	60
R^2	0.891
Adjusted R^2	0.876
Residual Std. Error	5,493.054 (df = 52)
F Statistic	60.432^{***} (df = 7; 52)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

```
# Check model assumptions
par(mfrow= c(1,2))
plot(best_model, which= c(1,2))
```



+ The selected model preforms better in the QQ plot upper ranges. Residuals appear randomly dispersed around zero

Predicting Tuition

```
# Impute data for University of Pittsburgh
# Select mean for diversity rank because data not available
point<-data.frame(Rank=67,
                  institutionalControl="public"
                  ,Median_Income=34022
                  ,Diversity_Rank_Race= as.numeric(mean(obs_60_final$Diversity_Rank_Race))
                  , Unemployment= 0.04
                  ,number_Undergrads=19928
                  ,Expend=15000)
```

```
pred<-predict(best_model,point);pred
```

```
##       1
## 29843.4
```

- Pitt yearly tuition is in state tuition is $22,000 per year and out of state tuition is 37,320

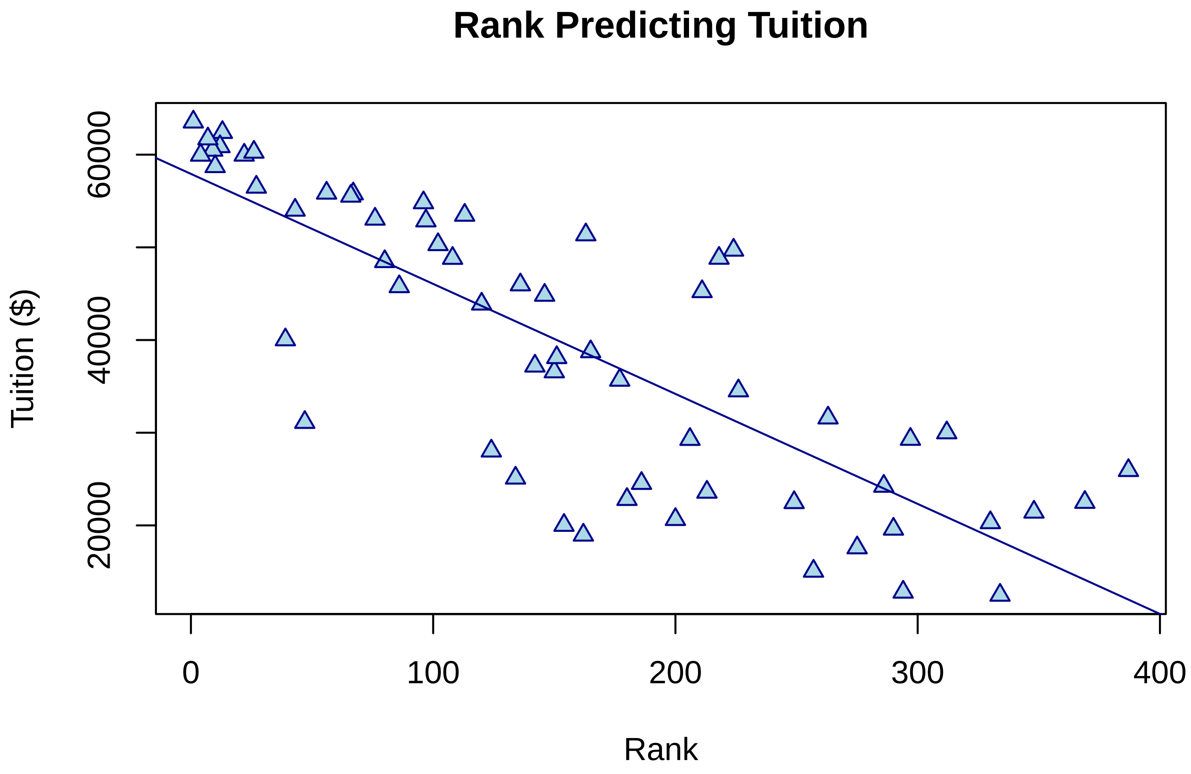
- The predicted tuition according to out model was $29,843
- Pitt is below market price for in state students and above market price for out of state students according to our model

The Power of Prestige

```
summary(lm(Tuition ~ Rank, data = obs_60_final))
```

```
##
## Call:
## lm(formula = Tuition ~ Rank, data = obs_60_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21226  -4591   2242   5438  18368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57924.08    2102.08   27.56  < 2e-16 ***
## Rank         -118.69      11.22  -10.58 3.66e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9183 on 58 degrees of freedom
## Multiple R-squared:  0.6588, Adjusted R-squared:  0.6529
## F-statistic:   112 on 1 and 58 DF,  p-value: 3.659e-15
```

```
plot(obs_60_final$Rank, obs_60_final$Tuition, pch = 24, col = "darkblue",bg="lightblue"
,xlab = "Rank", ylab = "Tuition ($)" ,  main= "Rank Predicting Tuition")

abline(lm(Tuition ~ Rank, data = obs_60_final), col = "darkblue")
```



```
# Plot the regression line for Rank
```