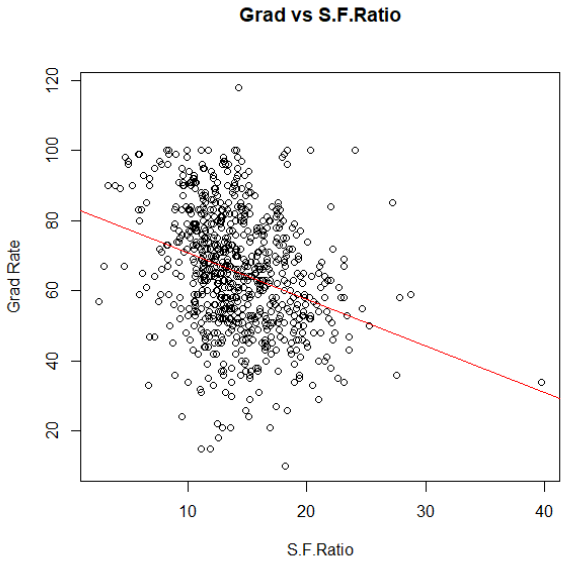
Partners: Damani and Sahil
CIS 3920 Data Mining for Business Analytics
Professor Charles Terng
March 18th, 2021

Assignment #3: Linear Regression

1. **Build a simple linear regression model to examine the relationship between graduation rates ("Grad.Rate") and student-faculty ratio ("S.F.Ratio").**

    a. **Is there a relationship? Why?**
        - Yes, there is a negative or inverse relationship because when one variable decreases, the other increases. Also, the S.F.Ratio has a negative slope of -1.3310.

    b. **Plot graduation rates against student-faculty ratio to visualize the relationship.**



    c. **As the student-faculty ratio increases, what can we expect to happen to graduation rates?**
        - As the student-faculty ratio increases, we can expect the graduation rates to decrease.

d. **Test if this relationship is linear. Build a model that also includes a quadratic term: Use I(S.F.Ratio^2) or poly(S.F.Ratio, 2) in the model. What do the results tell us about the relationship?**

- The relationship between Grad.Rate ~ S.F.Tatio is linear because it matches the criteria for the following linear equation: $Y_i = f(\mathbf{X}_i) + \varepsilon_i$ *. But, while testing the model that includes the quadratic I(S.F.Ratio^2), we came to realize that this model fails to satisfy the conditions for linearity. Hence, the relationship for this new model is one that encompasses "non-linearity in the data**" and "likely [a] curvilinear relationship**". *From Slide 2 **From Slide 72 (https://docs.google.com/presentation/d/1Y3fxsTLhtRBV0FU1VxCIcZIL8Jd_H0xsgdsTgIDb-u4/edit#slide=id.gbf15a9faf5_0_389)

Model #1

```
Call:
lm(formula = Grad.Rate ~ S.F.Ratio, data = College)

Residuals:
    Min      1Q  Median      3Q     Max
-54.443 -11.094   0.284  11.612  52.817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.2168     2.1713  38.786   <2e-16 ***
S.F.Ratio    -1.3310     0.1484  -8.971   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.36 on 775 degrees of freedom
Multiple R-squared:  0.09407,   Adjusted R-squared:  0.0929
F-statistic: 80.48 on 1 and 775 DF,  p-value: < 2.2e-16
```

Model #2

```
Call:
lm(formula = Grad.Rate ~ I(S.F.Ratio^2) + S.F.Ratio, data = College)

Residuals:
    Min      1Q  Median      3Q     Max
-54.472 -10.939   0.588  11.528  53.226

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    89.79312    4.61787  19.445  < 2e-16 ***
I(S.F.Ratio^2)  0.02535    0.01853   1.368 0.171702
S.F.Ratio      -2.11206    0.58988  -3.580 0.000364 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.35 on 774 degrees of freedom
Multiple R-squared:  0.09626, Adjusted R-squared:  0.09392
F-statistic: 41.22 on 2 and 774 DF,  p-value: < 2.2e-16
```

2. **Build a multiple linear regression model for graduation rates on predictors "Private", "Top25perc", "Outstate", and "Room.Board":**

$$Grad.Rate = \beta_1 Private + \beta_2 Top25perc + \beta_3 Outstate + \beta_4 Room.Board + \epsilon$$

   a. **Evaluate the model's quality of fit based on F-statistic and R^2. What are these metrics telling us?**
      - The F statistic indicates if our model is useful compared to a null model (p-value < 2.2e-16). Adjusted R^2 assesses our model fit in consideration of other variables (r^2 = 0.3856), this means 38.56% of the variation is explained by the model.

```
Call:
lm(formula = Grad.Rate ~ Private + Top25perc + Outstate + Room.Board,
    data = College)

Residuals:
    Min      1Q  Median      3Q     Max
-54.843  -8.438   0.060   7.944  57.321

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.474e+01  2.749e+00   9.001  < 2e-16 ***
Private     3.974e+00  1.341e+00   2.963  0.00314 **
Top25perc   2.449e-01  2.883e-02   8.493  < 2e-16 ***
Outstate    1.361e-03  1.998e-04   6.811 1.95e-11 ***
Room.Board  1.373e-03  5.831e-04   2.355  0.01876 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.46 on 772 degrees of freedom
Multiple R-squared:  0.3888,  Adjusted R-squared:  0.3856
F-statistic: 122.8 on 4 and 772 DF,  p-value: < 2.2e-16
```

   b. **Do we expect to see higher graduation rates for private or non-private institutions? Why?**
      - We expect higher graduation rates for private institutions because not only is the p-value of this model < 0.05 which indicates it is statistically significant but historically they have been higher.

c. **For a unit increase in Top25perc (% of new freshmen students coming from top 25% of their high school class), give a 95% confidence interval for the range we expect graduation rates to change.**
   - The range for Graduation Rate would be from 1.9345… to 30.13711….

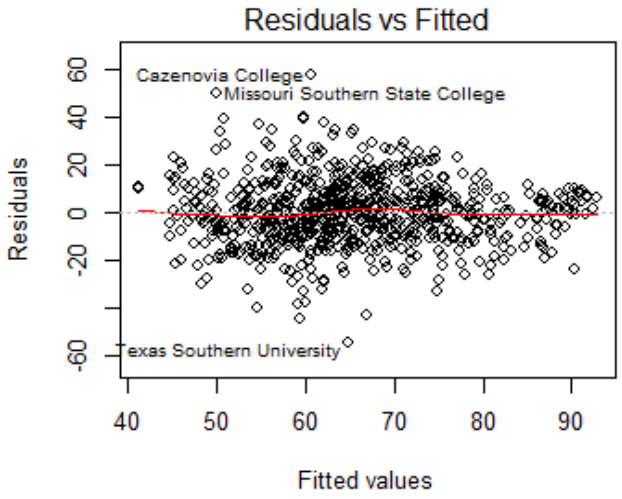
```
confint(q2_model)
                   2.5 %       97.5 %
(Intercept) 1.934509e+01 30.137115072
Private     1.340890e+00  6.606214247
Top25perc   1.882886e-01  0.301486951
Outstate    9.688043e-04  0.001753387
Room.Board  2.287256e-04  0.002518064
```

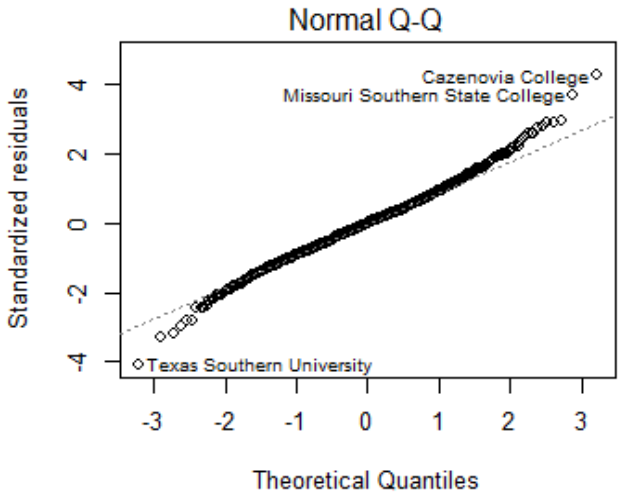
d. **Predict the graduation rate of a public college with a $25,000 out-of-state tuition, $4,000 costs for room and board, and 55% of their new students having graduated in the top 25% of their high school. Provide a prediction interval for the prediction as well. What does the range tell us about our prediction's usefulness?**
   - The graduation rate will be 81.7%, with an interval of 54.3%

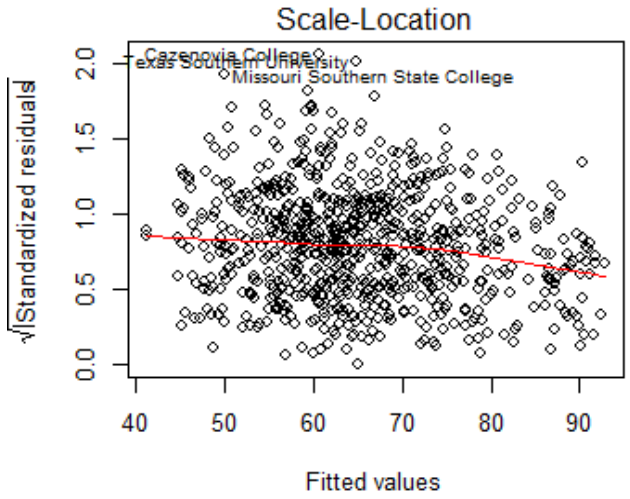
```
predict(q2_model,new_obs,interval='prediction')
       fit      lwr      upr
1 81.70446 54.32017 109.0887
```

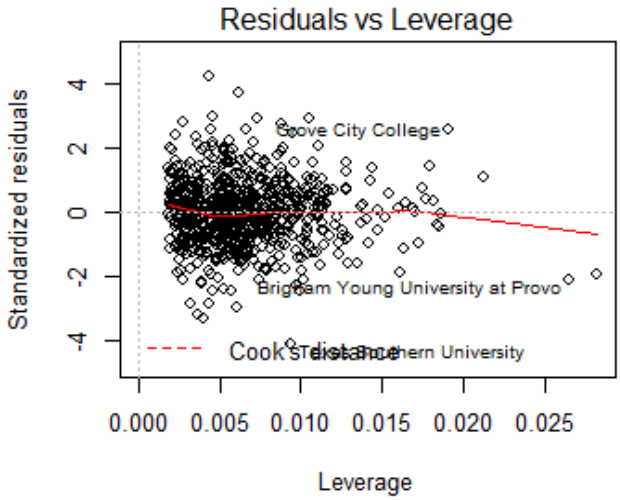
e. **Use the plot( ) function on the model to examine diagnostic plots. Looking at the residual vs fit plot, give your assessment on whether any assumptions have been violated or not. Explain your decision based on the plot.**

- From looking at the residual vs fit plot, there is linearity because the residuals are around the horizontal line without distinct patterns so no assumptions were violated.









3. **Build a regression model for graduation rates using all the predictors. *Tip: lm(y ~ .) uses the period symbol "." as shorthand for "all", instead of typing out all variables.***

a. **What is the adjusted R^2 for the full model?**
   - The adjusted R^2 is 0.4495 ~~ 45%

```
Residual standard error: 12.75 on 759 degrees of freedom
Multiple R-squared:  0.4615,  Adjusted R-squared:  0.4495
F-statistic: 38.27 on 17 and 759 DF,  p-value: < 2.2e-16
```

b. **Apply forward selection to find the best subset of variables for a reduced model. Use adjusted R^2 ("adjr2") as the metric for choosing the optimal model. How many variables are in the reduced model compared to the full? Which variables were left out?**
   - There are 13 variables in the optimal model compared to the full, the ones that got left out were: Accept, Enroll, Books, Ph.D., Terminal, & S.F.Ratio.

```
summary(model_fwd)$which[13,]
(Intercept)     Private        Apps      Accept      Enroll   Top10perc
Top25perc F.Undergrad P.Undergrad
       TRUE        TRUE        TRUE       FALSE       FALSE        TRUE
TRUE        TRUE        TRUE
   Outstate  Room.Board       Books    Personal         PhD    Terminal
S.F.Ratio perc.alumni      Expend
       TRUE        TRUE       FALSE        TRUE        TRUE        TRUE
FALSE        TRUE        TRUE
```

c. **Build a new model based on the selected variables. Compare the adjusted R^2 between the reduced model and the full. Is there a noticeable difference? What does that tell us?**
   - R^2 (reduced model) = 0.4504; R^2 (full model) = 0.4495, there isn't really much of a difference between the two and it tells us that even without the other variables from the full model compared to reduced, there's still the same amount of variation and the other variables are insignificant.

```
Residual standard error: 12.73 on 765 degrees of freedom
Multiple R-squared:  0.4582,  Adjusted R-squared:  0.4504
F-statistic: 58.82 on 11 and 765 DF,  p-value: < 2.2e-16
```

d. Plot the ranking of subsets in the feature selection process. We see that there isn't much difference between the best subset and one with far fewer variables. Pick the subset with fewest predictors that has at least an adjusted R^2 of 0.45. Which predictors are in this model?

- The predictors in this model are Private, Enroll, Apps, Top10perc, Top25perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Personal, Ph.D, Terminal, perc.alumni, & Expend.