STAT3612 Assignment 2: Regression Modeling

Date: March 2, 2017

Due date: March 15, 2017

Consider the Times Higher Education (THE) World University Ranking dataset you have explored in Assignment 1, with response variable `total_score` and 7 predictors `citations`, `income`, `international`, `num_students`, `research`, `teaching`, `student_staff_ratio`. Select the world universities which were ranked top 200 during year 2011-2016.

Step 0. Use `R::na.omit()` to remove missing values, then we will have a working subdata `DataX1` with the following numbers of observations over the years.

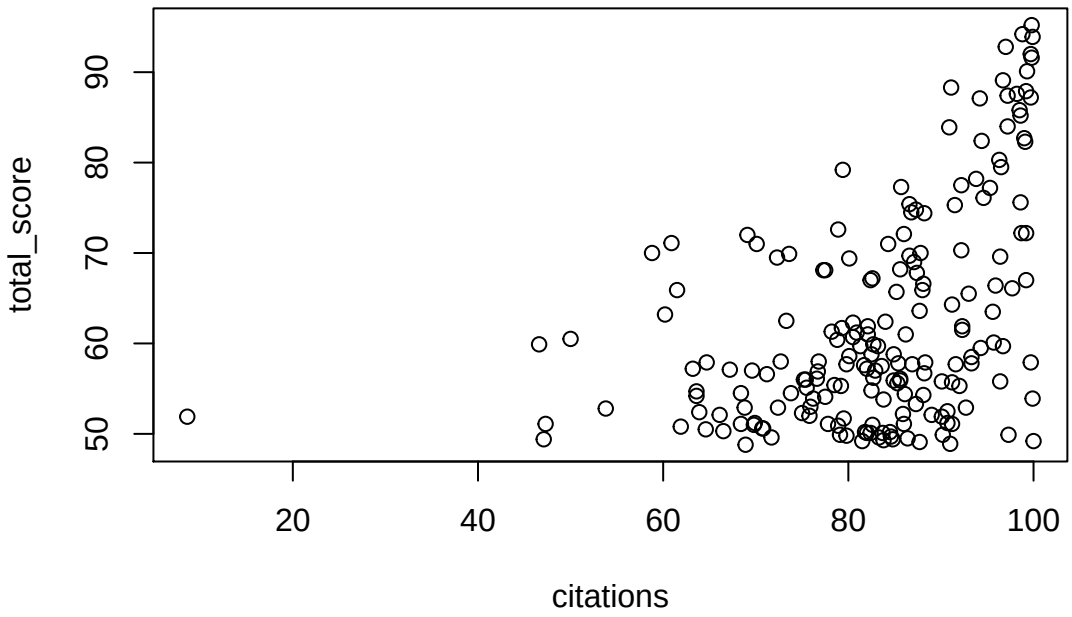
```
## 2011 2012 2013 2014 2015 2016
##  132  172  176  180  186  193
```

Step 1. (30%) **a)** Use `DataX1` to fit a linear regression model (with intercept) for the response `total_score` with four predictors: `teaching`, `research`, `num_students`, `student_staff_ratio`. What is the goodness of fit? Output the coefficients for the significant variables. **b)** Fit another linear model (with intercept) for the response `total_score` with all 7 predictors. What is the goodness of fit? Output the coefficients for the significant variables.

Step 2. (30%) **a)** Use `DataX1` to fit a linear regression model with the potentially significant variables as discovered in Step 1. What do you find about the goodness of fit? **b)** Repeat fitting the same model for each of 6 ranking years (i.e. subsetting data per year). What do you find about their goodness of fit? **c)** Compare the model coefficients of the 7 fitted models, and give a guess about THE world university ranking methodology.

Step 3. (20%) Consider the world ranking score prediction for the year 2016. Given the knowledge of discovered important variables in Step 2, suppose you are allowed to use only p variables for score prediction. For $p = 1, 2, 3$, which model do you recommend in each case?

Step 4. (20%) For year 2016 data with `total_score` and `citations` plotted in the next page, **a)** fit a nonparametric curve using smoothing spline; **b)** there is an outlying point in the plot, would it affect the result? (Both questions can be answered by plots).


total_score
50
60
70
80
90
20
40
60
80
100
citations