# Final Exam

Phan Nguyen

09/10/2022

We are going to do a multiple regression analysis on a dataset with 120 rows and six columns. This dataset represents student exam scores, with each row representing a student. For this exercise, the response variable is ScienceScores. Below is a sample of the dataset.

Table 1: Dataset example

```
## Rows: 120
## Columns: 6
## $ TutorService         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0~
## $ Gender               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ InstructionalProgram <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ MathScores           <dbl> 84.16430, 84.37501, 87.29191, 85.87178, 82.80028,~
## $ EnglishScores        <dbl> 80.79973, 79.76153, 87.23620, 81.54955, 78.98671,~
## $ ScienceScores        <dbl> 84.15280, 86.52541, 89.29222, 94.44127, 84.71549,~
```

### 1.Exploratory Analysis on the numerical variables

MathScores, EnglishScores and ScienceScores are variables at interval level. The other variables are either nominal or ordinal.
Therefore we are going to look at only these three variables.

**Descriptive numerical summary table**
Table 2: Descriptive Statistics table for MathScores, EnglishScores, and ScienceScores

```
##               vars   n  mean   sd median trimmed  mad   min   max range  skew
## MathScores       1 120 78.05 7.78  78.28   78.24 8.12 58.35 94.80 36.45 -0.17
## EnglishScores    2 120 73.78 6.28  73.02   73.51 6.39 60.81 90.24 29.44  0.40
## ScienceScores    3 120 78.89 7.97  79.97   79.13 8.86 54.25 96.44 42.19 -0.30
##               kurtosis   se Q0.25 Q0.75
## MathScores       -0.54 0.71 73.01 83.92
## EnglishScores    -0.36 0.57 68.80 77.46
## ScienceScores    -0.50 0.73 72.80 85.31
```

From a glance, all three variables' numerical summary makes logical sense. They have approximately close mean and median, and even minimum and maximum values. EnglishScores has a narrower range than the other two (29.44 compared to 36.45 and 42.19).

EnglishScores skewed right (skewness >0) while the other two variables skewed left. This is also indicated by the fact that its median smaller than its mean, which is not the case for MathScores and Sciencescores. All three variables have heavier tails and lower bell curve, due to their excess kurtosis (i.e kurtosis - 3) lower than 0.

**Descriptive chart summaries**

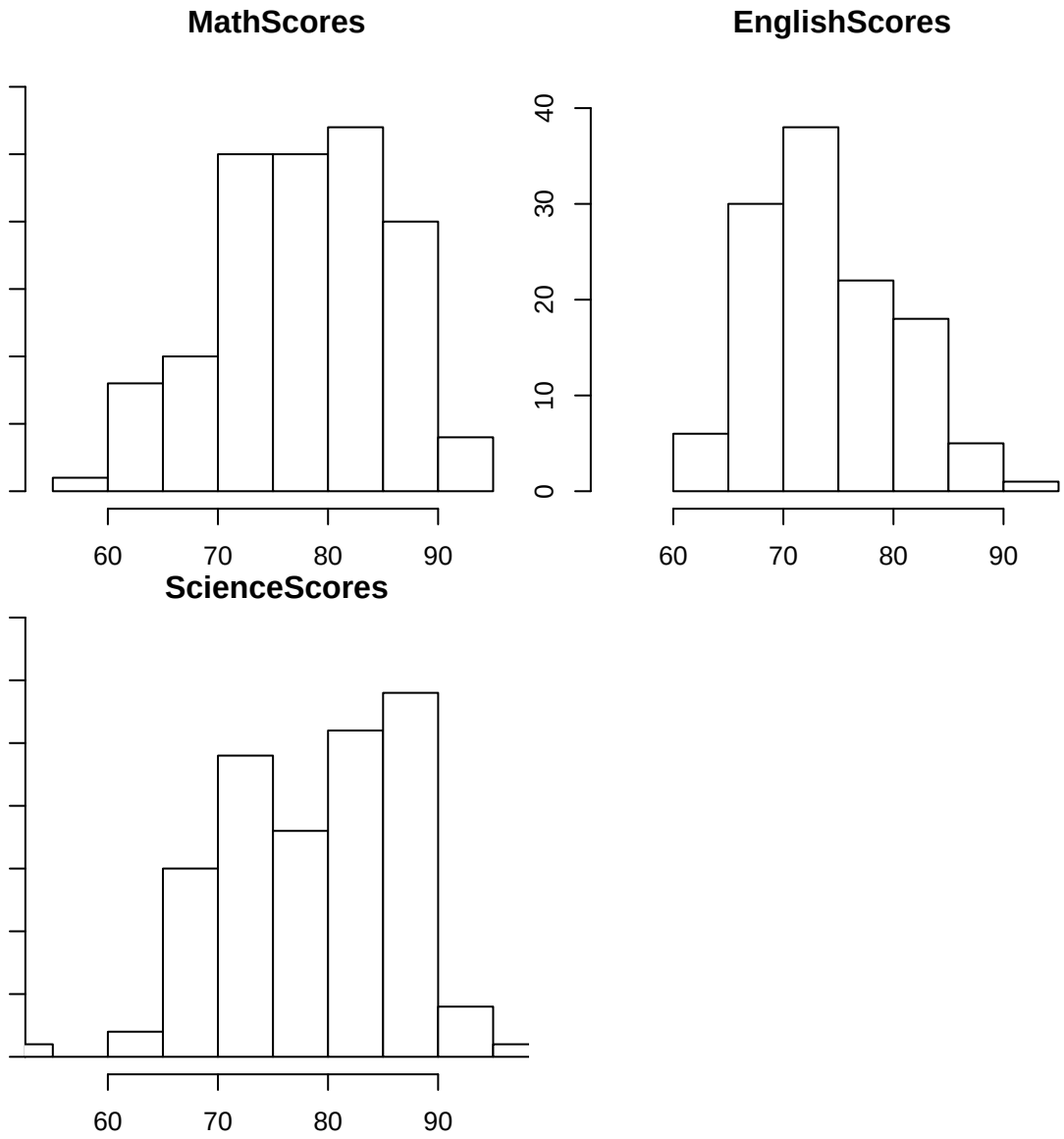


Figure 1: Histogram of MathScores, EnglishScores and ScienceScores

The histogram shows the frequency of data values for the three nterval variable. We can cross check it with the numerical summary we've done on the top. EnglishScores skewed left, while MathScores and Science Scores skewed right. ScienceScores look like it is bimodal.

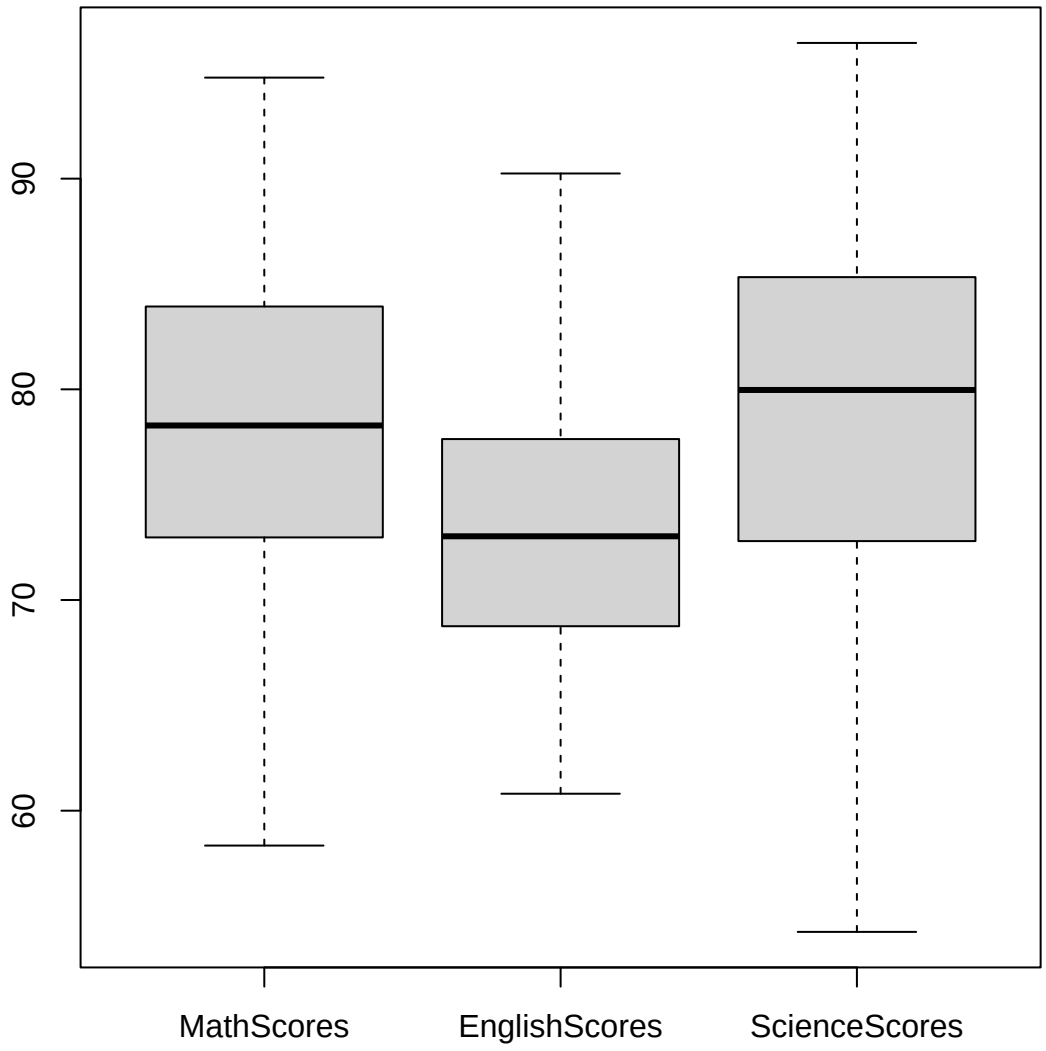


Figure 2: Boxplots of MathScores, EnglishScores, and ScienceScores

EnglishScores have smaller range and interquartile range, lower median and mean than the other two variables. ScienceScores have the biggest range and interquartile range. All three variables look relatively normal.

**2. Scatterplot analysis**

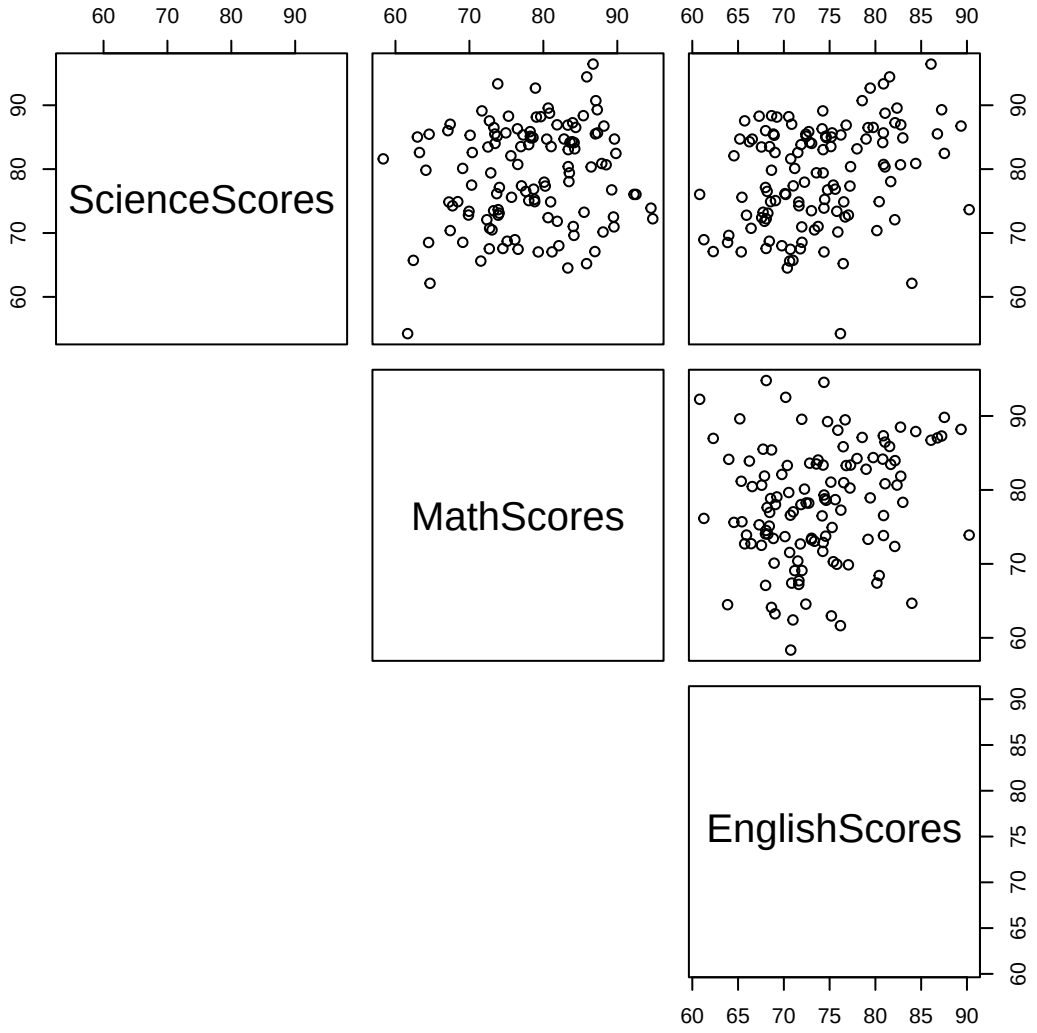


Figure 3: Scatterplots for ScienceScores, MathScores and EnglishScores

There seems to be no correlation between these three variables. The data points are scattered. However, visually, EnglishScores and ScienceScores might have a linear relationship upward.

Table 3: Correlation Matrix for MathScores, EnglishScores, and ScienceScores

```
##               MathScores EnglishScores ScienceScores
## MathScores     1.0000000     0.1956467     0.1590911
## EnglishScores  0.1956467     1.0000000     0.3065704
## ScienceScores  0.1590911     0.3065704     1.0000000
```

These variables don't have strong correlation to one another. Therefore it is safe to say there is no collinearity problem.

### 3. Assumption check

First, we fit the data into a model to check all assumptions. Response variable is ScienceScores.

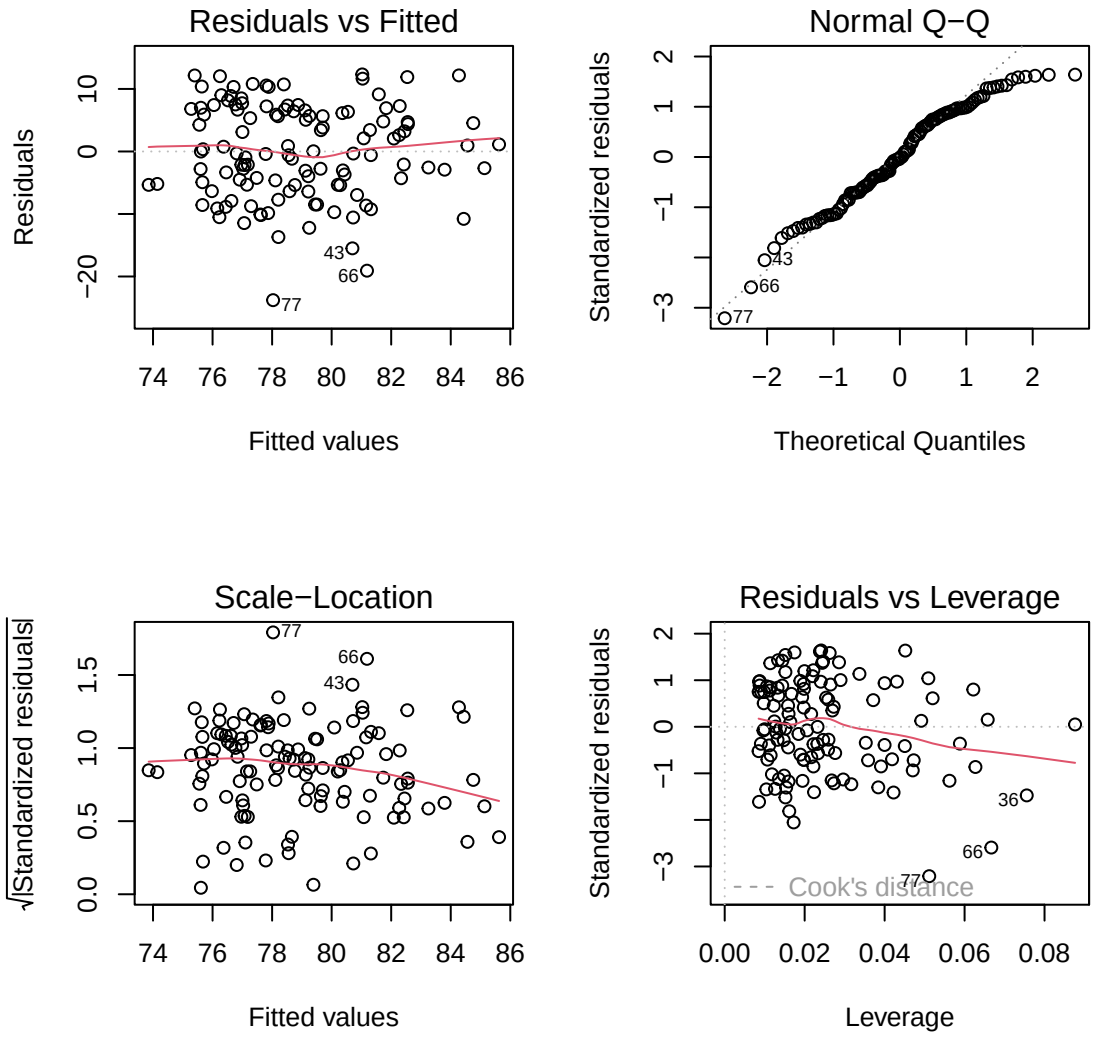


Figure 4: Simple Regression Plots for model

We can check the assumptions for the multiple regression model using these plots.

The residual plot on the left hand side, shows how the predicted versus observed values. Ideally, the red line will stay horizontally at 0. In this case, the red line is approximately horizontal and at line 0. The points are scattered and don't follow any pattern. We can conclude that the **linearity and equal variances assumptions are met**.

Independence of variance can be checked using the Scale-Location plot, on the bottom left hand. The red line supposed to be horizontally at zero for the assumption to be met. We can say that **the indpendence of variance assumption is met**.

The normal Q-Q plots on the top right shows visually a straight line, meaning that **the normality assumption is met**.

Earlier in the analysis, we agreed that there is no multicollinearity problem, meaning the independent variables are not correlated with other.
All the assumptions are met.

### 4. Recoding categorical tables

There are three categorical variables. They are Tutor Service, Gender and Instructional Program. Below is the list of values for each of these variables

```
## $TutorService
## [1] 1 0
##
## $Gender
## [1] 0 1
##
## $InstructionalProgram
## [1] 1 2
```

These variables are binary, they have only 2 unique values. We don't need to recode TutorService and Gender, since they have 0 and 1 already. We need to recode Instructional Program, so that 0 indicates lack of the other program. Here is the recoding:

Table 4: Recoding for InstructionalProgram

| InstructionalProgram | IPdummy |
|---|---|
| 1 | 0 |
| 2 | 1 |