Final Project Report

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, 1, 0, 0~
## $ Gender               <dbl> 0, 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, 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.Summary

In the earlier version of our report, we have looked at all the variables, explored descriptive statistics and visualizations, correlations matrix. MathScores and EnglishScores are our two numerical explanatory variables. We also have recoded InstructionalProgram into IPdummy, so that it can be used in our regression. Therefore, TutorService, Gender & IPdummy will be our categorical variables

2. Assumption check

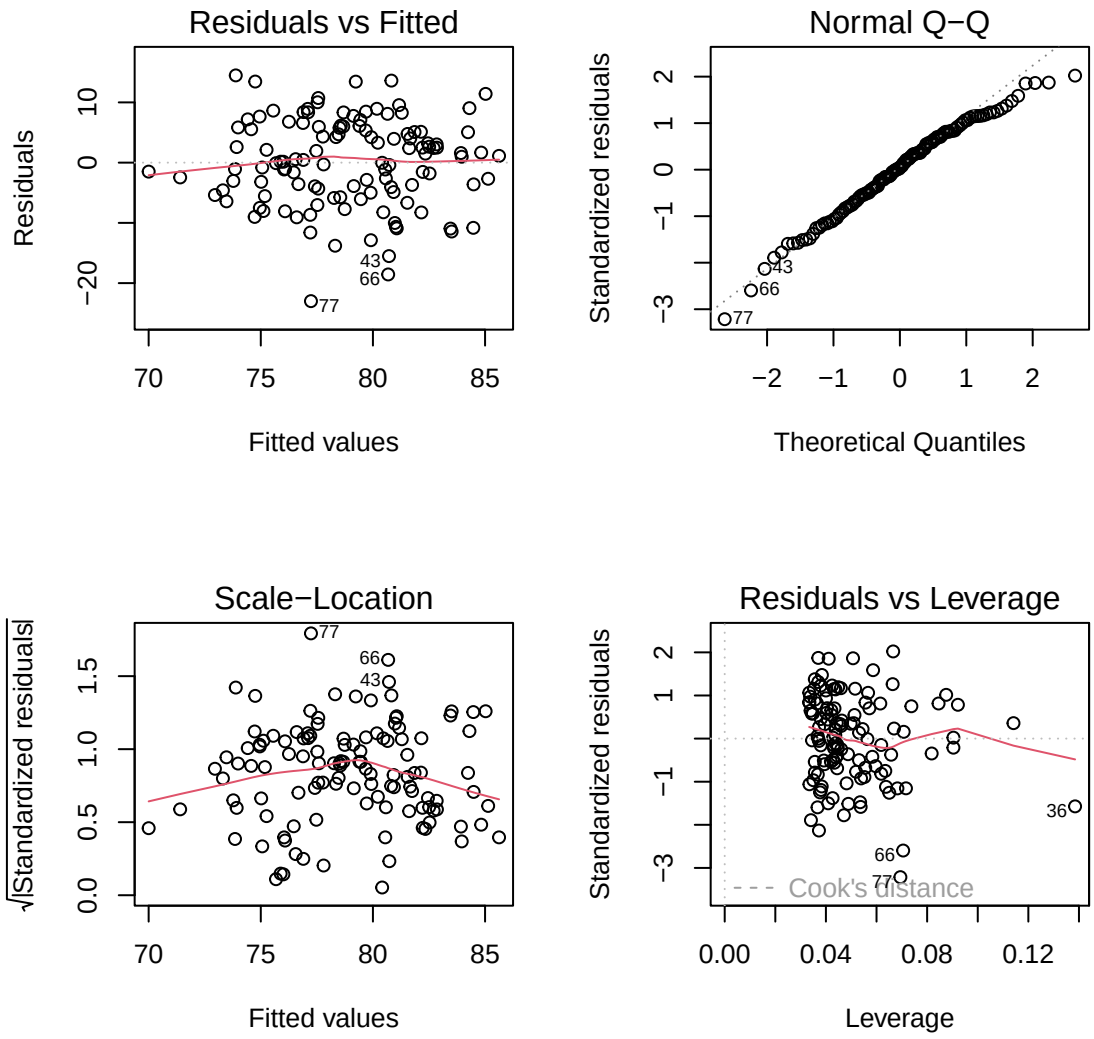

Figure 1: Simple Regression Plots for model

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.

3. Multiple Regression and Variable Selection

We are going to use backward stepwise variable selection method.

Table 2: Details for backward Stepwise and Regression Model Output

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . TutorService
## 2 . Gender
## 3 . MathScores
## 4 . EnglishScores
## 5 . IPdummy
##
## We are eliminating variables based on p value...
##
##
## No more variables satisfy the condition of p value = 0.3
##
##
## Variables Removed:
##
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.414       RMSE                 7.414
## R-Squared               0.171       Coef. Var            9.398
## Adj. R-Squared          0.135       MSE                 54.975
## Pred R-Squared          0.077       MAE                  5.904
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
```

```
##                 Sum of
##                Squares         DF    Mean Square      F        Sig.
## ----------------------------------------------------------------------
## Regression    1296.559          5        259.312    4.717     6e-04
## Residual      6267.097        114         54.975
## Total         7563.657        119
## ----------------------------------------------------------------------
##
##                                   Parameter Estimates
## ------------------------------------------------------------------------------------------
##         model      Beta    Std. Error    Std. Beta      t        Sig       lower     upper
## ------------------------------------------------------------------------------------------
##   (Intercept)    14.180        14.167                 1.001    0.319    -13.886    42.246
##  TutorService    -1.563         1.375       -0.098    -1.137    0.258     -4.287     1.161
##        Gender     2.575         1.512        0.162     1.703    0.091     -0.420     5.569
##    MathScores     0.293         0.123        0.286     2.387    0.019      0.050     0.536
## EnglishScores     0.528         0.127        0.416     4.154    0.000      0.276     0.779
##       IPdummy     4.785         1.947        0.301     2.458    0.015      0.929     8.641
## ------------------------------------------------------------------------------------------
```

```
## [1] "No variables have been removed from the model."
```

The backward stepwise method have not removed any variables

The model has F-test of 4.717, with p-value of approximately 0. **The model is statistically significant**. The Adjusted R-squared is around 0.135, meaning 13.5% of the variation in ScienceScores can be explained by variation in the independent variables.

All variables are significant with p-values less than 0.05, except for TutorSerivce and Gender.

Looking at coefficients, we see that people who has TutorService on average score less on Science than people who don't by 1.563. People with Gender 1 scores averagely 2.575 higher in Science than people with Gender 0 (we don't know which is which). An 1 unit increase in Mathscores and EnglishScores averagely increase ScienceScores by 0.293 and 0.528, respectively. And people who has InstructionalProgram 2 has averagely 4.785 higher score than people wit InstructionalProgram 1.

The final regression equation is
ScienceScores = 14.180 - 1.563 TutorService + 2.575 Gender + 0.293 MathScores + 0.528 English-Scores + 4.785 InstructionalProgram