Project 2- Advanced Statistical Analysis Project Using R by Michael Massaad

Michael Massaad

2024-04-02

This Project aims to analyse consumer spending patterns across different regions using multiple linear regression, hypothesis testing, and model selection.

1. Data generation and visualization

```
set.seed(123) # For reproducibility
# Generating random data
n <- 100 # Number of observations
regions <- c("North", "South", "East", "West")
spending_categories <- c("Food", "Clothing", "Electronics")
data <- data.frame(
  Region = sample(regions, n, replace = TRUE),
  Income = rnorm(n, mean = 50000, sd = 10000),

  Age = sample(20:70, n, replace = TRUE),
  Spending = rnorm(n, mean = 300, sd = 50),
  Category = sample(spending_categories, n, replace = TRUE)
)

dim(data)
```

```
## [1] 100   5
```

```
str(data)
```

```
## 'data.frame':    100 obs. of  5 variables:
##  $ Region  : chr  "East" "East" "East" "South" ...
##  $ Income  : num  52533 49715 49571 63686 47742 ...
##  $ Age     : int  45 61 68 36 48 45 46 40 26 45 ...
##  $ Spending: num  277 316 286 355 308 ...
##  $ Category: chr  "Electronics" "Food" "Food" "Electronics" ...
```

```
head(data, 10)
```

```
##    Region   Income Age Spending    Category
```

```
## 1    East 52533.19  45 276.8856 Electronics
## 2    East 49714.53  61 316.0978        Food
## 3    East 49571.30  68 285.7999        Food
## 4   South 63686.02  36 354.5928 Electronics
## 5    East 47742.29  48 308.0108 Electronics
## 6   South 65164.71  45 435.4084 Electronics
## 7   South 34512.47  46 314.3145 Electronics
## 8   South 55846.14  40 298.3745    Clothing
## 9    East 51238.54  26 250.5161        Food
## 10  North 52159.42  45 310.5905        Food
```

```
summary(data)
```

```
##     Region              Income           Age           Spending
##  Length:100         Min.   :26908   Min.   :21.00   Min.   :134.5
##  Class :character   1st Qu.:43390   1st Qu.:36.00   1st Qu.:257.0
##  Mode  :character   Median :49497   Median :47.00   Median :305.9
##                     Mean   :49463   Mean   :45.59   Mean   :298.7
##                     3rd Qu.:55267   3rd Qu.:57.25   3rd Qu.:339.1
##                     Max.   :71873   Max.   :69.00   Max.   :435.4
##    Category
##  Length:100
##  Class :character
##  Mode  :character
##
##
##
```

1.1 Data plots and distribution analysis

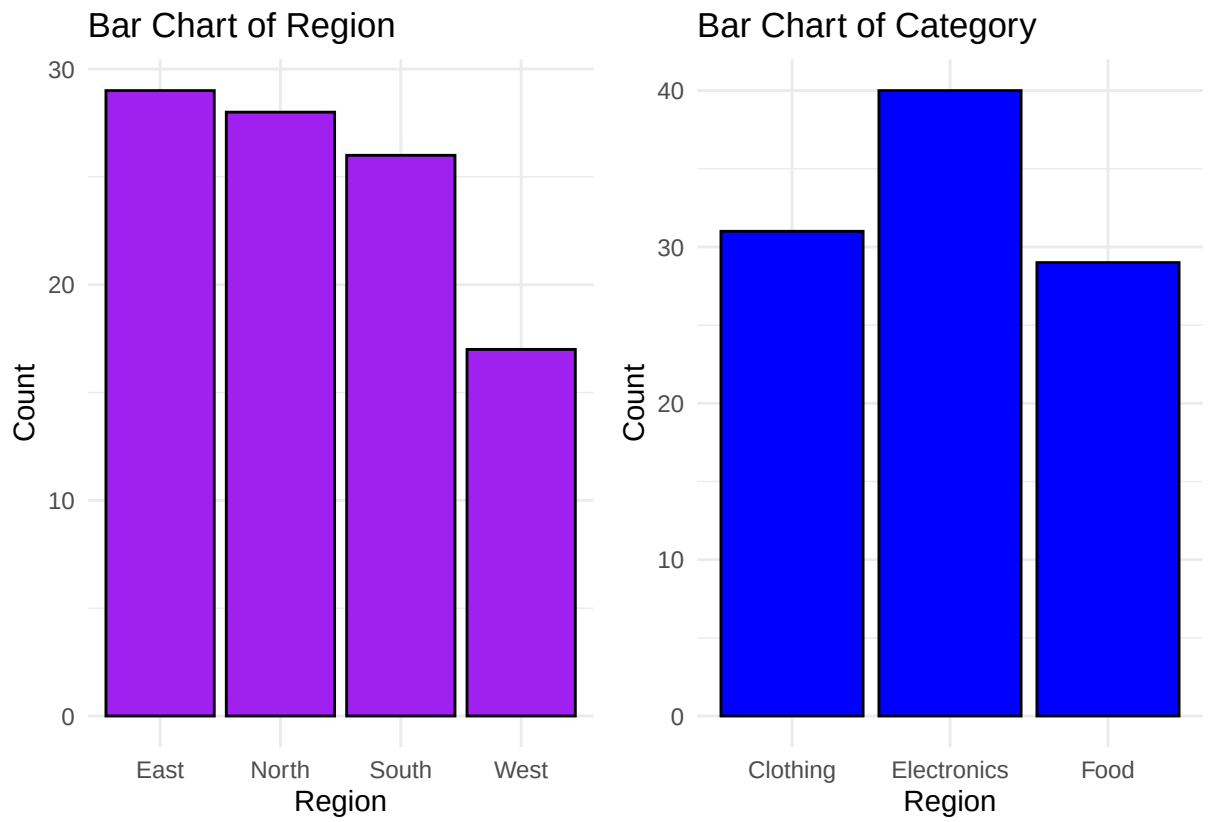
1.1.1 Histograms of Income, Age and Spending variables

```
library(patchwork)
library(ggplot2)

p1 = ggplot(data, aes(x = Income)) + geom_histogram(fill = "red", color = "black") +
  theme_minimal() + labs(title = "Histogram of Income", x = "Income", y = "Frequency")

p2 = ggplot(data, aes(x = Age)) + geom_histogram(fill = "lightgreen", color = "black") +
  theme_minimal() + labs(title = "Histogram of Age", x = "Age", y = "Frequency")

p3 = ggplot(data, aes(x = Spending)) + geom_histogram(fill = "gray", color = "black") +
  theme_minimal() + labs(title = "Histogram of Spending", x = "Spending", y = "Frequency")

p1 | p2 | p3
```

In observing all the histograms, we can observe that the histogram for the Income is approximately symmetric, the histogram for the Age is also approximately symmetric (since the distribution on both the right and the left is approximately equal), the histogram for the Spending is approximately symmetric (since we can observe that most of the data is distributed in the center of the graph).

1.1.2 Bar chart for the Region and Category variables

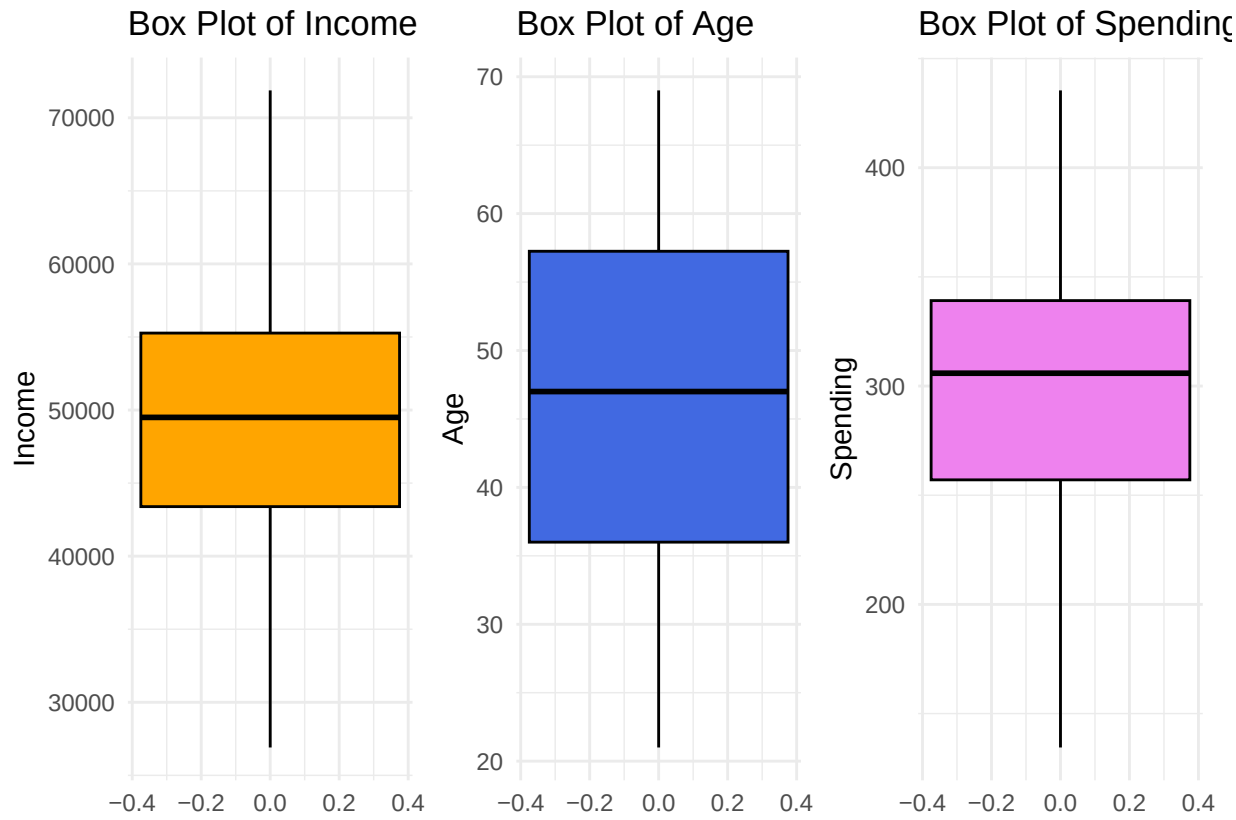
```
bc1 = ggplot(data, aes(x = Region)) + geom_bar(fill = "purple", color = "black") +
  theme_minimal() + labs(title = "Bar Chart of Region", x = "Region", y = "Count")

bc2 = ggplot(data, aes(x = Category)) + geom_bar(fill = "blue", color = "black") +
  theme_minimal() + labs(title = "Bar Chart of Category", x = "Region", y = "Count")

bc1 | bc2
```



1.1.3 Boxplots for the Income, Age, and Spending variables

```
library(patchwork)
library(ggplot2)

b1 = ggplot(data, aes(y = Income)) +
  geom_boxplot(
    fill = "orange",
    color = "black",
    outlier.color = "red"
  ) + theme_minimal() + labs(title = "Box Plot of Income", y = "Income")

b2 = ggplot(data, aes(y = Age)) +
  geom_boxplot(
    fill = "royalblue",
    color = "black",
    outlier.color = "red"
  ) + theme_minimal() + labs(title = "Box Plot of Age", y = "Age")

b3 = ggplot(data, aes(y = Spending)) +
  geom_boxplot(
    fill = "violet",
    color = "black",
    outlier.color = "red"
```

```
  ) + theme_minimal() + labs(title = "Box Plot of Spending", y = "Spending")

b1 | b2 | b3
```



1.2 Is it normally distributed?

In observing the boxplots for the Income, Age and Spending variables, we can observe that there are no outliers for any of the variables. Also, from observing the histograms and the boxplots of all 3 variables, we can conclude that they are all normally distributed which can be deduced from all 3 histograms being approximately symmetric, and which can also be reflected in the boxplots, where the mean (represented within the colored square in each plot), is approximately in the center of each distribution. Another indicator that the Income and Spending variables are normally distributed is the fact that they are generated by the rnorm command in the supplied code.

1.3 Is there a difference in average expenditure across any pairs of categories?

Before doing the two sided test for the equality of the means, we need to know if the variances are equal or not to use the right formula. In observing the histogram of the category variable, we can observe that it is approximately symmetric, therefore it is normally distributed. As seen in the previous questions, we know that the Spending variable is normally distributed. Therefore, both variables that will be used, being the category and spending variables, they meet the assumptions to proceed with the testing for variance equality.

1.3.1 Test hypothesis for the equality of means between the food and clothing category

$H_0 : mu_{foodSpending} = mu_{clothingSpending}$ $H_1 : mu_{foodSpending}$ is different from $mu_{clothingSpending}$

1. F-test for variance equality of food and clothing

```
foodSpending = data$Spending[data$Category == 'Food']
clothingSpending = data$Spending[data$Category == 'Clothing']
var.test(foodSpending, clothingSpending)
```

```
##
##  F test to compare two variances
##
## data:  foodSpending and clothingSpending
## F = 0.49932, num df = 28, denom df = 30, p-value = 0.06822
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2386861 1.0545995
## sample estimates:
## ratio of variances
##           0.499316
```

The p-value of the F-test is $p = 0.06822$ which is greater than the significance level 0.05. Therefore, we can conclude that there isn't any significant difference in the variances, therefore we will use the t-test with the pooled variances (var.equal=TRUE).

2. t-test between food and clothing

```
foodSpending = data$Spending[data$Category == 'Food']
clothingSpending = data$Spending[data$Category == 'Clothing']

t.test(foodSpending, clothingSpending, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  foodSpending and clothingSpending
## t = 1.3884, df = 58, p-value = 0.1703
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.173795 45.184413
## sample estimates:
## mean of x mean of y
##  307.9900  289.4847
```

In observing the p-value, which is $p = 0.1703 > 0.05$, we do not reject the null hypothesis, therefore the means/average expenditure of the Food and Clothing category are not different.

1.3.2 Test hypothesis for the equality of means between the clothing and electronics category

$H_0 : mu_{electronicsSpending} = mu_{clothingSpending}$ $H_1 : mu_{electronicsSpending}$ is different from $mu_{clothingSpending}$

1. F-test for variance equality of clothing and electronics

```r
electronicsSpending = data$Spending[data$Category == 'Electronics']
var.test(clothingSpending, electronicsSpending)
```

```
##
##  F test to compare two variances
##
## data:  clothingSpending and electronicsSpending
## F = 1.0176, num df = 30, denom df = 39, p-value = 0.9477
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.521101 2.049532
## sample estimates:
## ratio of variances
##           1.017641
```

The p-value of the F-test is $p = 0.9477$ which is greater than the significance level 0.05. Therefore, we can conclude that there isn't any significant difference in the variances, therefore we will use the t-test with the pooled variances (var.equal=TRUE).

2. t-test between clothing and electronics

```r
t.test(clothingSpending, electronicsSpending, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  clothingSpending and electronicsSpending
## t = -0.67463, df = 69, p-value = 0.5022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -37.65976  18.62568
## sample estimates:
## mean of x mean of y
##  289.4847  299.0018
```

In observing the p-value, which is $p = 0.5022 > 0.05$, we do not reject the null hypothesis, therefore the means/average expenditure of the Clothing and Electronics category are not different.

1.3.3 Test hypothesis for the equality of means between the food and electronics category

$H_0 : mu_{electronicsSpending} = mu_{foodSpending}$ $H_1 : mu_{electronicsSpending}$ is different from $mu_{foodSpending}$

1. F-test for variance equality of food and electronics

```r
var.test(foodSpending, electronicsSpending)
```

```
##
##  F test to compare two variances
##
## data:  foodSpending and electronicsSpending
```

```
## F = 0.50812, num df = 28, denom df = 39, p-value = 0.0646
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2577432 1.0430470
## sample estimates:
## ratio of variances
##          0.5081246
```

The p-value of the F-test is $p = 0.0646$ which is greater than the significance level 0.05. Therefore, we can conclude that there isn't any significant difference in the variances, therefore we will use the t-test with the pooled variances (var.equal=TRUE).

2. t-test between food and electronics

```
t.test(foodSpending, electronicsSpending, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  foodSpending and electronicsSpending
## t = 0.70403, df = 67, p-value = 0.4839
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.49447  34.47100
## sample estimates:
## mean of x mean of y
##  307.9900  299.0018
```

In observing the p-value, which is $p = 0.4839 > 0.05$, we do not reject the null hypothesis, therefore the means/average expenditure of the Food and Electronics category are not different.

Conclusion

Through our results for the tests between category pairs, we are able to conclude that there is no difference in average expenditure between the categories.

2. Multiple Linear Regression and Analysis

Getting dataset and description of dataset

```
library(readxl)
dataset = read_excel("sat.xls")
dim(dataset)
```

```
## [1] 105   5
```

```r
str(dataset)
```

```
## tibble [105 x 5] (S3: tbl_df/tbl/data.frame)
##  $ high_GPA: num [1:105] 3.45 2.78 2.52 3.67 3.24 2.1 2.82 2.36 2.42 3.51 ...
##  $ math_SAT: num [1:105] 643 558 583 685 592 562 573 559 552 617 ...
##  $ verb_SAT: num [1:105] 589 512 503 602 538 486 548 536 583 591 ...
##  $ comp_GPA: num [1:105] 3.76 2.87 2.54 3.83 3.29 2.64 2.86 2.03 2.81 3.41 ...
##  $ univ_GPA: num [1:105] 3.52 2.91 2.4 3.47 3.47 2.37 2.4 2.24 3.02 3.32 ...
```

```r
head(dataset, 10)
```

```
## # A tibble: 10 x 5
##    high_GPA math_SAT verb_SAT comp_GPA univ_GPA
##       <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
##  1     3.45      643      589     3.76     3.52
##  2     2.78      558      512     2.87     2.91
##  3     2.52      583      503     2.54     2.4
##  4     3.67      685      602     3.83     3.47
##  5     3.24      592      538     3.29     3.47
##  6     2.1       562      486     2.64     2.37
##  7     2.82      573      548     2.86     2.4
##  8     2.36      559      536     2.03     2.24
##  9     2.42      552      583     2.81     3.02
## 10     3.51      617      591     3.41     3.32
```

```r
summary(dataset)
```

```
##     high_GPA        math_SAT        verb_SAT        comp_GPA
##  Min.   :2.030   Min.   :516.0   Min.   :480.0   Min.   :2.030
##  1st Qu.:2.670   1st Qu.:573.0   1st Qu.:548.0   1st Qu.:2.870
##  Median :3.170   Median :612.0   Median :591.0   Median :3.210
##  Mean   :3.076   Mean   :623.1   Mean   :598.6   Mean   :3.128
##  3rd Qu.:3.480   3rd Qu.:675.0   3rd Qu.:645.0   3rd Qu.:3.490
##  Max.   :4.000   Max.   :718.0   Max.   :732.0   Max.   :4.000
##     univ_GPA
##  Min.   :2.080
##  1st Qu.:3.010
##  Median :3.290
##  Mean   :3.173
##  3rd Qu.:3.470
##  Max.   :3.810
```

Questions

1. Perform the multiple linear regression usig all the predictors and interpret the output;

```r
model = lm(univ_GPA ~ high_GPA + math_SAT + verb_SAT + comp_GPA, dataset)
summary(model)
```

```
##
## Call:
## lm(formula = univ_GPA ~ high_GPA + math_SAT + verb_SAT + comp_GPA,
##     data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54760 -0.09442 -0.01051  0.11025  0.37766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5595710  0.1867474   2.996  0.00344 **
## high_GPA     0.0720545  0.0555757   1.297  0.19778
## math_SAT    -0.0007338  0.0005629  -1.304  0.19537
## verb_SAT     0.0008045  0.0004434   1.814  0.07263 .
## comp_GPA     0.7568120  0.0489260  15.469  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1519 on 100 degrees of freedom
## Multiple R-squared:  0.8891, Adjusted R-squared:  0.8846
## F-statistic: 200.3 on 4 and 100 DF,  p-value: < 2.2e-16
```

```
model
```

```
##
## Call:
## lm(formula = univ_GPA ~ high_GPA + math_SAT + verb_SAT + comp_GPA,
##     data = dataset)
##
## Coefficients:
## (Intercept)     high_GPA     math_SAT     verb_SAT     comp_GPA
##   0.5595710    0.0720545   -0.0007338    0.0008045    0.7568120
```

In observing the summary of the model, we can observe that the $R^2_{adjusted} = 0.8846$, which indicates that the model is a good fit for the data. In addition, we can observe that the intercept is 0.5595710, the coefficient of the high_GPA is 0.0720545, the coefficient of the mat_SAT is -0.0007338, the coefficient of the verb_SAT is 0.0008045, and the the coefficient of the comp_GPA is 0.7568120. Therefore we have the multiple linear regression model being : $\hat{Y} = univ_GPA = 0.0720545 * (high_GPA) + -0.0007338 * (mat_SAT) + 0.0008045 * (verb_SAT) + 0.7568120 * (comp_GPA) + 0.5595710$. Looking at the residuals from the summary, we can see that they are approximately symmetrical (small difference between the absolute values of each side), which means that the model is also approximately symmetrical. In observing the values of the coefficients, we can see that they are very small, therefore the high_GPA, mat_SAT, verb_SAT and comp_GPA variables do not have a major impact on the univ_GPA. However, in observing the p-values of each coefficient, only the coefficient of the comp_GPA is statistically significant (i.e p<0.001), which indicates that all the work is being done by the comp_GPA variable, which can be confirmed by the fact that the coefficient for that variable is relatively large compared to the coefficients of the other variables.

2. Test the hypothesis H0 : B2 (mat_SAT) = 0 versus H1: B2 is not equal 0;

In observing the p-value of the mat_SAT coefficient, we can see that 0.19537 > alpha = 0.05, we fail to reject our null hypothesis, so the coefficient is 0. Therefore the predictor B2 doesn't contribute to the response variable Y = univ_GPA.

3. Is University GPA Linearly Related to High School GPA?

```
HS_model = lm(univ_GPA ~ high_GPA, dataset)
summary(HS_model)
```

```
##
## Call:
## lm(formula = univ_GPA ~ high_GPA, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69040 -0.11922  0.03274  0.17397  0.91278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.09682    0.16663   6.583 1.98e-09 ***
## high_GPA     0.67483    0.05342  12.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2814 on 103 degrees of freedom
## Multiple R-squared:  0.6077, Adjusted R-squared:  0.6039
## F-statistic: 159.6 on 1 and 103 DF,  p-value: < 2.2e-16
```

In observing the summary of the model, we have the simple linear regression: $\hat{Y} = univ_GPA = 0.67483 * (high_GPA) + 1.09682$. In addition, we can observe that both the intercept and the high_GPA coefficient are statistically significant (i.e p<0.001). In observing the $R^2_{adjusted} = 0.6039$, we can observe that the model is not a good representation of the response variable, therefore we can conclude that the High School GPA variable is not linearly related to the University GPA.

4. Select the best model using stepAIC function from the MASS library;

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:patchwork':
##
##     area
```

```
selectedModel = stepAIC(model)
```

```
## Start:  AIC=-390.87
## univ_GPA ~ high_GPA + math_SAT + verb_SAT + comp_GPA
##
##            Df Sum of Sq    RSS     AIC
## - high_GPA  1    0.0388 2.3463 -391.12
## - math_SAT  1    0.0392 2.3467 -391.10
## <none>                  2.3075 -390.87
## - verb_SAT  1    0.0760 2.3835 -389.47
```

```
## - comp_GPA  1    5.5213 7.8288 -264.60
##
## Step:  AIC=-391.12
## univ_GPA ~ math_SAT + verb_SAT + comp_GPA
##
##            Df Sum of Sq     RSS     AIC
## - math_SAT  1    0.0223  2.3687 -392.12
## <none>                   2.3463 -391.12
## - verb_SAT  1    0.1005  2.4468 -388.71
## - comp_GPA  1    8.6721 11.0184 -230.71
##
## Step:  AIC=-392.12
## univ_GPA ~ verb_SAT + comp_GPA
##
##            Df Sum of Sq     RSS     AIC
## <none>                   2.3687 -392.12
## - verb_SAT  1    0.0895  2.4582 -390.23
## - comp_GPA  1    9.6341 12.0028 -223.73
```

```
summary(selectedModel)
```

```
##
## Call:
## lm(formula = univ_GPA ~ verb_SAT + comp_GPA, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53395 -0.10577 -0.00479  0.10607  0.34548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3795490  0.1428461   2.657  0.00915 **
## verb_SAT    0.0006057  0.0003084   1.964  0.05229 .
## comp_GPA    0.7770930  0.0381520  20.368  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1524 on 102 degrees of freedom
## Multiple R-squared:  0.8861, Adjusted R-squared:  0.8839
## F-statistic: 396.8 on 2 and 102 DF,  p-value: < 2.2e-16
```

Using the stepAIC, we have the best model being: $\hat{Y} = univ_GPA = 0.0006057 * (verb_SAT) + 0.7770930 * (comp_GPA) + 1.09682$.

5. Compare the selected model (obtained from 3.) and the full model (ie. with all the predictors) using adjusted R^2 and AIC;

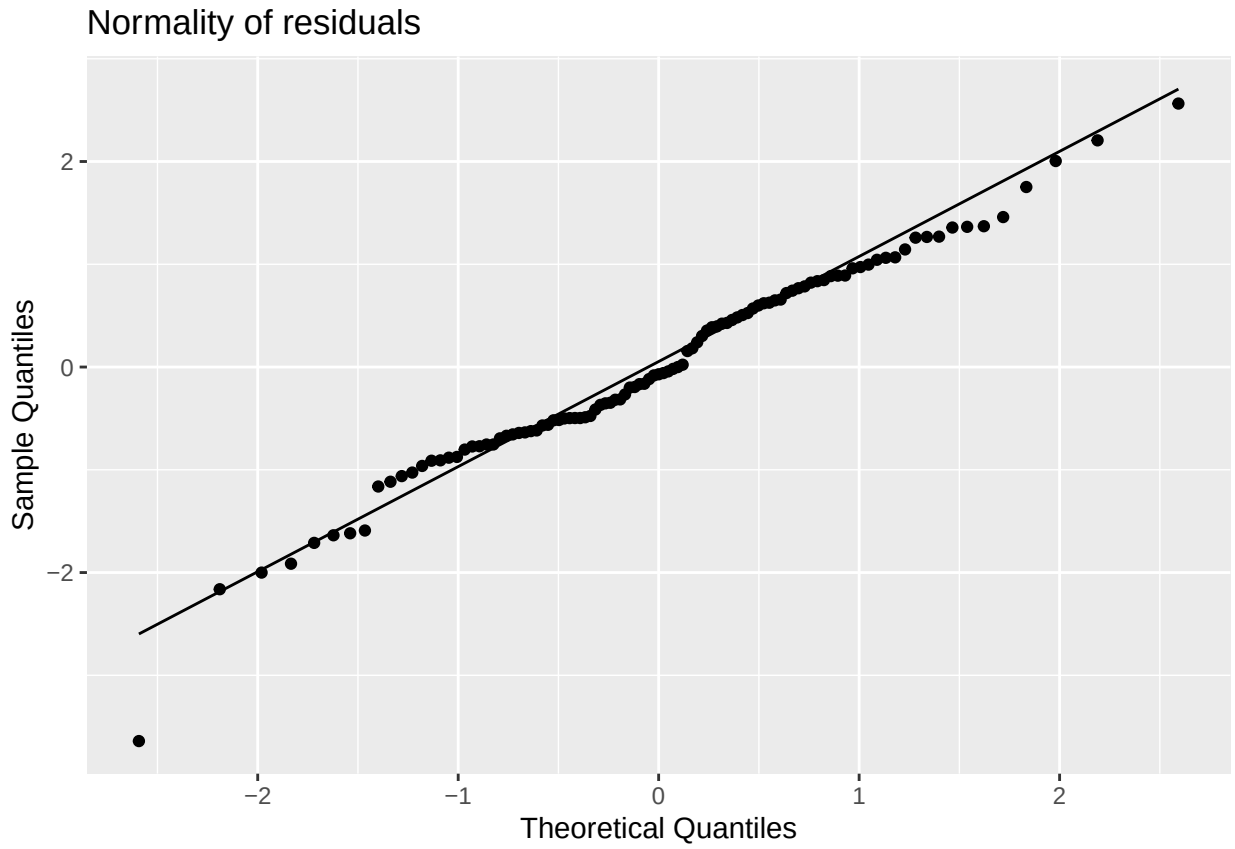
```
AIC(model, HS_model)
```

```
##          df       AIC
## model     6 -90.89049
## HS_model  3  35.71546
```

In observing the summaries of both models, we can see that the adjusted R^2 for the full model is 0.8846 and the one for the selected model is 0.6039, which is worse than the adjusted R^2 for the full model. In observing the results from the AIC, we can observe that the full model has a smaller value of AIC than the HS_model.

6. Perform all the steps of model adequacy of the selected model (HS_model obtained in question 3). 6.1 QQplot

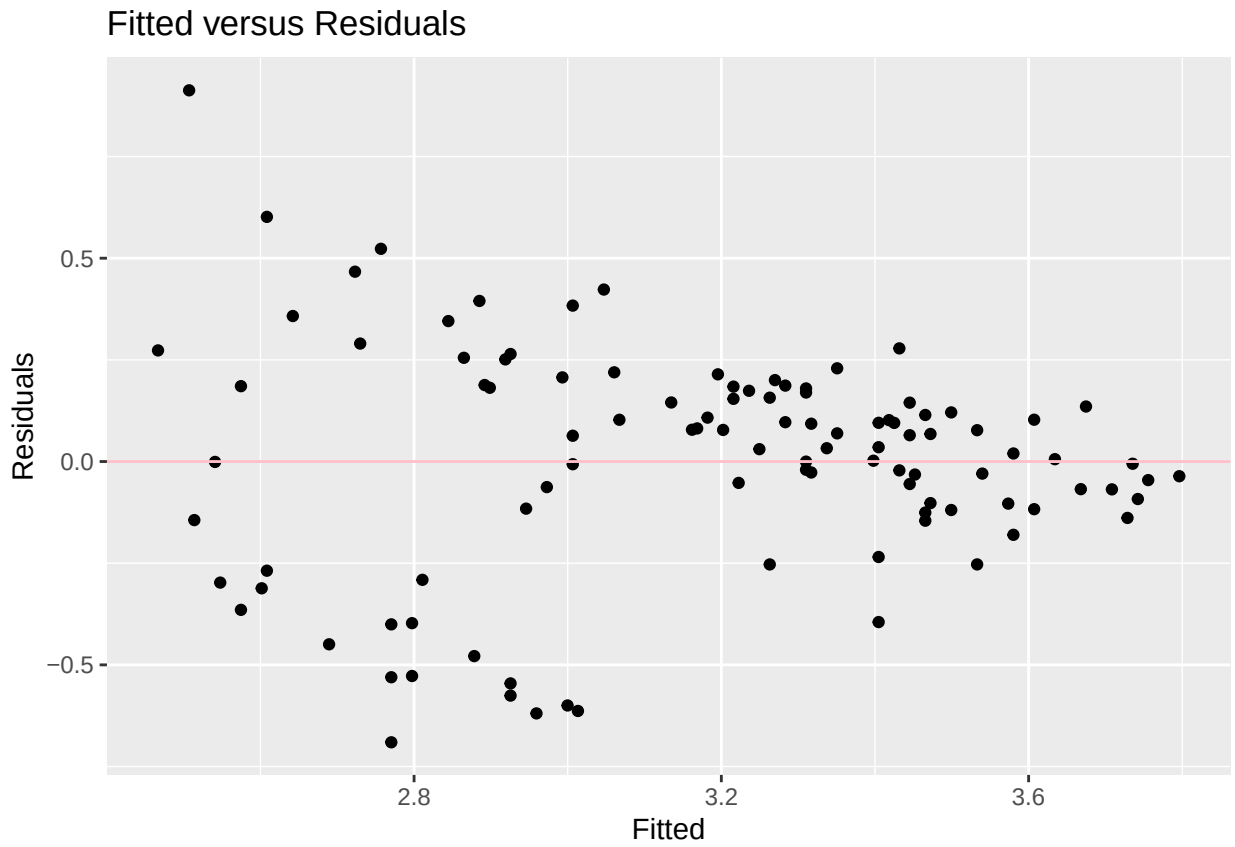
```
qq = ggplot(HS_model, aes(sample = rstandard(model))) + geom_qq() + stat_qq_line() +
  labs(title = "Normality of residuals", x = "Theoretical Quantiles",
       y = "Sample Quantiles")
qq
```



In observing the QQplot of the residuals, since most of the points are on/around the line, we can conclude that the assumption of the normality of the residuals is confirmed.

6.2 Residual plot

```
rp = ggplot(HS_model, aes(.fitted, .resid)) + geom_point() + geom_hline(yintercept = 0,
color = "pink") + labs(title = "Fitted versus Residuals", x = "Fitted", y = "Residuals")
rp
```



In observing the fitted versus residuals plot, we can see that the point are not evenly spread, and they form the shape of a funnel, therefore we can conclude that the variance is not constant, so the selected model does not satisfy one of the assumptions, therefore we can conclude that this is not a good model. This can also be supported by the value of the $R^2_{adjusted}$ for the model obtained in question 3 being low (with a value of 0.6039).