End of Term Project

Setting up

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/this is for R"
)
getwd()
```

```
## [1] "/Users/jasmine/Library/Mobile Documents/com~apple~CloudDocs/
this is for R"
```

```
library(haven)
library(ggplot2)
library(descr)
library(carData)
library(texreg)

library(cowplot)
library(coefplot)
library(car)
library(carData)
#To load the Data set I will be using
d <- read.csv("world.csv", stringsAsFactors = TRUE)
```

Question One

Pick a dataset among *gss*, *nes* and *world*. Inspect it, have a look at the variables it contains and at the codebook. Select an outcome and a predictor variable. These will be the central elements of your assignment. Remember that the outcome variable needs to be interval, ratio or high-level ordinal - what we call a continuous variable. Feel free to recode variables where you need to. Formulate the working and the null hypotheses. (15 points)

```
#Names of variables
names(d)
```

```
##  [1] "country"        "dem_economist"  "confidence"     "dem_scor
e14"
##  [5] "durable"        "enpp3_democ08"  "effectiveness"  "gdp_10_t
hou"
##  [9] "gender_unequal" "gender_equal3"  "hdi"            "literacy
"
## [13] "pop_age"        "pop_total"      "spendeduc"      "spendhea
lth"
## [17] "womyear2"       "women13"        "votevap00s"
```

```
#Structure of variables
str(d)
```

```
## 'data.frame':    167 obs. of  19 variables:
##  $ country       : Factor w/ 167 levels "Afghanistan",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ dem_economist : int  0 0 0 0 1 0 1 1 0 0 ...
##  $ confidence    : num  NA 49.3 52.1 NA 7.3 ...
##  $ dem_score14   : num  2.77 5.67 3.83 3.35 6.84 4.13 9.01 8.54 2
.83 2.87 ...
##  $ durable       : int  4 3 5 3 17 2 99 54 5 25 ...
##  $ enpp3_democ08 : Factor w/ 3 levels "1-3 parties",..: NA 1 NA N
A 1 3 1 2 NA NA ...
##  $ effectiveness : num  13.7 35.5 32.6 19.1 35 ...
##  $ gdp_10_thou   : num  NA 0.1535 0.1785 0.0857 0.2797 ...
##  $ gender_unequal: num  0.797 0.545 0.594 NA 0.534 0.57 0.296 0.3
0.553 0.512 ...
##  $ gender_equal3 : Factor w/ 3 levels "High","Low","Medium": NA N
A NA NA 1 NA 1 1 NA 2 ...
##  $ hdi           : num  0.349 0.719 0.677 0.403 0.775 0.695 0.937
0.851 0.713 0.801 ...
##  $ literacy      : num  28.1 NA 69.9 67.4 97.2 99.4 99 98 98.8 86
.5 ...
##  $ pop_age       : num  16.9 30 26.2 17.4 30.4 32 37.8 41.8 28.4
28.1 ...
##  $ pop_total     : num  29.1 3.2 35.4 19 40.7 3.1 21.5 8.4 8.9 0.
8 ...
##  $ spendeduc     : num  NA 2.9 4.3 2.6 4.9 3 4.7 5.4 1.9 2.9 ...
##  $ spendhealth   : num  1.8 2.9 3.6 2 5.1 2.1 6 7.7 1 2.6 ...
##  $ womyear2      : Factor w/ 2 levels "1944 or before",..: NA 1 2
2 2 1 1 1 1 2 ...
##  $ women13       : num  NA 15.7 NA NA 37.4 10.7 24.7 27.9 NA NA .
..
##  $ votevap00s    : num  NA 59.6 NA NA 70.9 ...

#Summary of data set
summary(d)

##         country    dem_economist      confidence      dem_score14
##  Afghanistan:  1   Min.   :0.0000   Min.   : 6.495   Min.   :1.08
0
##  Albania    :  1   1st Qu.:0.0000   1st Qu.:38.889   1st Qu.:3.61
0
##  Algeria    :  1   Median :0.0000   Median :49.508   Median :5.79
0
##  Angola     :  1   Mean   :0.4551   Mean   :48.900   Mean   :5.54
8
##  Argentina  :  1   3rd Qu.:1.0000   3rd Qu.:59.523   3rd Qu.:7.39
5
##  Armenia    :  1   Max.   :1.0000   Max.   :99.862   Max.   :9.93
0
##  (Other)    :161                    NA's   :99
##     durable            enpp3_democ08 effectiveness      gdp_10_th
ou
##  Min.   :  0.00   1-3 parties :28    Min.   :  7.801   Min.   :0.
0090
##  1st Qu.:  4.00   4-5 parties :23    1st Qu.: 28.132   1st Qu.:0.
```

```
0436
##  Median :  9.00   6-11 parties:30    Median : 39.007   Median :0.
1656
##  Mean   : 23.11   NA's        :86    Mean   : 46.036   Mean   :0.
6258
##  3rd Qu.: 31.25                      3rd Qu.: 62.884   3rd Qu.:0.
4894
##  Max.   :191.00                      Max.   :100.000   Max.   :4.
7354
##  NA's   :19                          NA's   :14        NA's   :18
##  gender_unequal   gender_equal3      hdi             literacy
##  Min.   :0.1740   High  :23     Min.   :0.1400   Min.   : 21.80
##  1st Qu.:0.3885   Low   :24     1st Qu.:0.4675   1st Qu.: 67.90
##  Median :0.5940   Medium:25     Median :0.6690   Median : 88.70
##  Mean   :0.5466   NA's  :95     Mean   :0.6325   Mean   : 80.61
##  3rd Qu.:0.6990                 3rd Qu.:0.7810   3rd Qu.: 98.20
##  Max.   :0.8530                 Max.   :0.9380   Max.   :100.00
##  NA's   :32                     NA's   :9        NA's   :10
##     pop_age        pop_total          spendeduc       spendhealth
##  Min.   :15.00   Min.   :   0.300   Min.   : 0.600   Min.   : 0.2
00
##  1st Qu.:20.30   1st Qu.:   3.875   1st Qu.: 3.300   1st Qu.: 1.9
00
##  Median :26.30   Median :  10.000   Median : 4.400   Median : 3.2
50
##  Mean   :28.09   Mean   :  41.484   Mean   : 4.488   Mean   : 3.6
66
##  3rd Qu.:36.60   3rd Qu.:  29.075   3rd Qu.: 5.375   3rd Qu.: 5.1
25
##  Max.   :44.70   Max.   :1354.100   Max.   :13.600   Max.   :11.5
00
##  NA's   :1       NA's   :1          NA's   :17       NA's   :3
##            womyear2     women13        votevap00s
##  1944 or before:56   Min.   : 2.70   Min.   :27.59
##  After 1944    :90   1st Qu.:12.18   1st Qu.:54.24
##  NA's          :21   Median :20.80   Median :65.12
##                      Mean   :21.10   Mean   :64.87
##                      3rd Qu.:27.65   3rd Qu.:77.48
##                      Max.   :44.70   Max.   :98.39
##                      NA's   :77      NA's   :94
```

```
#Check for missing data
sum(is.na(d))
```

```
## [1] 596
```

```
colSums(is.na(d))
```

```
##        country  dem_economist     confidence    dem_score14
durable
##              0              0             99              0
19
##  enpp3_democ08  effectiveness    gdp_10_thou gender_unequal  gend
er_equal3
```

```
##              86              14              18              32
95
##             hdi        literacy        pop_age       pop_total
spendeduc
##               9              10               1               1
17
##    spendhealth        womyear2        women13      votevap00s
##               3              21              77              94

#Making a new data set
newdataset <- read.csv ("world.csv", stringsAsFactors = TRUE)

#To subset new data set without missing data for the variables I wil
l be

utilizing
newdataset <- d[with(d, {
  !(is.na(dem_score14) |
      is.na(literacy)|is.na(gender_unequal)|is.na(spendeduc))
}), ]

#Tables for both variables
table(newdataset$literacy)

##
## 28.7 34.7 35.1 39.3 40.1 46.4 47.8 47.9 48.6 48.7 49.9 50.2 51.2
52.3 57.5 57.9
##    1    1    1    1    1    1    1    1    2    1    1    1    1
1    1    1
## 59.3 60.9   61 62.7 63.2 66.8 67.5 67.9 69.1 69.9 70.4 71.4   73
73.6 74.3   77
##    1    1    1    1    1    1    1    1    1    1    1    1    1
1    1    1
## 77.9 78.8 79.6 80.6 81.1 81.2 81.6 82.6 83.8 84.4 84.8   85 85.1
86.1 86.4 86.5
##    1    1    1    1    1    1    1    1    1    1    1    1    1
1    1    1
## 86.7   87 87.4 88.6 88.7   89 89.9 90.3 90.4 90.7   91 91.6 91.9
92.5 92.6 92.9
##    1    1    1    1    1    1    1    1    2    1    1    1    1
1    2    1
##   93 93.3   94 94.9 95.7   96 97.1 97.2 97.3 97.6 97.8 97.9   98
98.1 98.2 98.4
##    1    2    1    1    1    1    1    1    1    1    1    2    2
1    1    1
## 98.6 98.7 98.8   99 99.1 99.4 99.5 99.6 99.7 99.8  100
##    1    1    1   14    1    4    2    2    1    3    4

table(newdataset$dem_score14)

##
## 1.49 1.74 1.82 1.98 2.21 2.37 2.64 2.78 2.79 2.83 2.87 2.89    3
3.05 3.09 3.16
##    1    1    1    1    1    1    1    1    1    1    1    1    1
1    1    1
```

```
## 3.17 3.18 3.25 3.33 3.39 3.41 3.45 3.52 3.53 3.76 3.78  3.8 3.83
4 4.02 4.13
##    1    1    1    1    1    2    1    1    1    2    1    1    1
1    1    1
## 4.17 4.56 4.64 4.66 4.77 4.78 4.95 5.07 5.12 5.13 5.22 5.24 5.32
5.39 5.42 5.65
##    1    1    1    1    1    1    1    1    1    1    1    1    1
1    1    1
## 5.66 5.78 5.79 5.81 5.82 5.87 6.03 6.15 6.24 6.26 6.31 6.32 6.33
6.39 6.49 6.53
##    1    1    2    1    1    1    1    1    1    1    1    1    1
1    1    1
## 6.54 6.55 6.62 6.66 6.67 6.68 6.73 6.77 6.84  6.9 6.93 6.95 6.99
7.08 7.35 7.38
##    1    1    1    1    1    2    1    1    1    1    1    1    1
1    1    1
##  7.4 7.45 7.47 7.48 7.54 7.63 7.74 7.79  7.8 7.82 7.85 7.87 7.92
7.93 8.03 8.04
##    1    1    1    1    1    1    1    1    1    1    1    1    1
1    1    1
## 8.05 8.06 8.08 8.11 8.17 8.54 8.64 8.72 8.88 8.92 9.01 9.03 9.08
9.09 9.11 9.26
##    1    1    1    1    2    1    1    1    1    1    1    1    1
1    1    1
## 9.58 9.73 9.93
##    1    1    1
```

This project will be utilising R-studio to analyse the World Dataset. The hypothesis within this project will be that there is a relationship between the predictor variable and the outcome variable; that literacy rate (predictor variable) of a country has a direct impact on a countries democracy score (outcome variable). Both the outcome variable and the predictor variable are Ratio variables, which are continuous.

On the contrary the null hypothesis asserts there is no relationship between the two variables and as a result there is a null effect. The hypothesis is determined by there being a statistically significance below or equal to 0.05. However, to accept the null hypothesis the statistically significance needs to be above 0.05.

Question 2

Describe the two variables. Create appropriate visualisations for each variable, accompanied by the appropriate descriptive statistics (hint: it all depends on the level of measurement). (15 points

```
#Summarising the variables
summary(newdataset$literacy)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.70   73.45   90.85   83.60   99.00 100.00
```

The range of literacy rate is from 28.7-100, however we can tell from the median that low scores are more uncommon since the median is 90.85%. By comparing the median & mean we can tell how evenly the data is distruste

d, the mean is 83.6 and the median is 90.85, they are quite close together , showing that most countries who are on this dataset have a high liter acy rate.

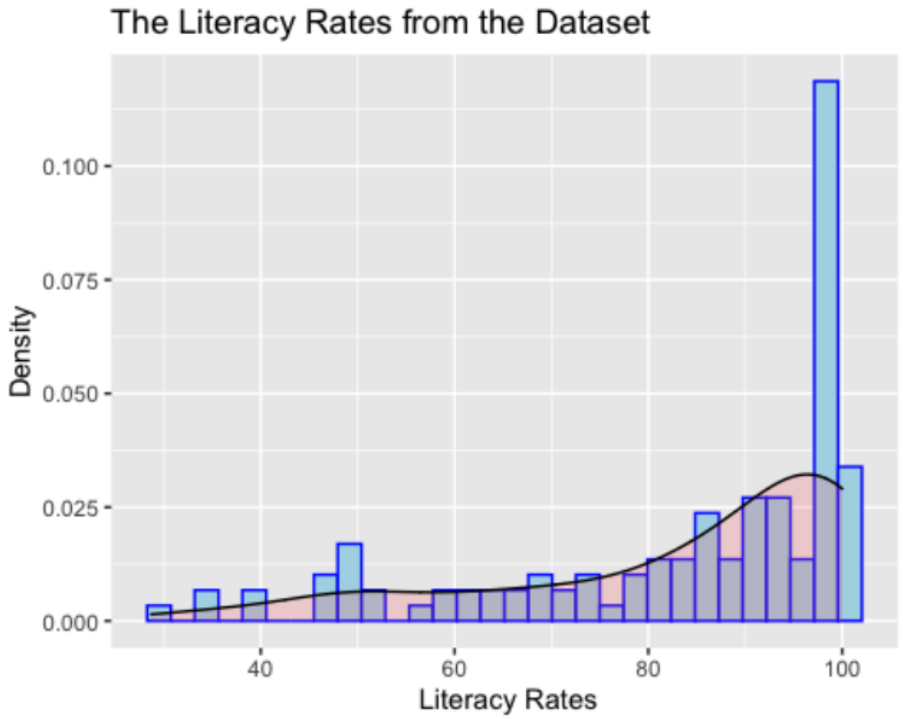
```
summary(newdataset$dem_score14)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.490   3.822   6.250   5.878   7.657   9.930
```

This shows that the measurement of democracy score ranges from 1.49 to 9.93. Which is quite significan 2 median and mean are again quite similar the median being 6.25 and t he mean 5.878, which illustrates that most countries have some form of a democracy.

```
ggplot(newdataset, aes(literacy)) +
  geom_histogram(aes(y=..density..), color="blue", fill="lightblue")
+
  geom_density(alpha=.2, fill="#FF6666")+
  labs(title = "The Literacy Rates from the Dataset ",
       x= ' Literacy Rates',
       y = 'Density')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`
.
```

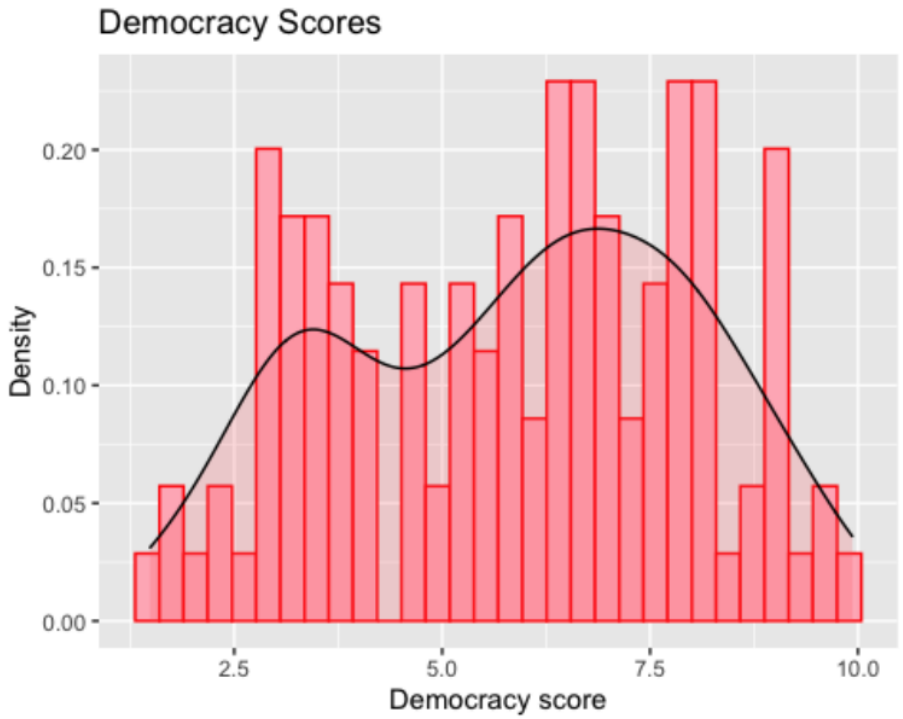


The reason for using a histogram instead of a bar plot is due to 'Literacy Rate' being 'a ra variable which is a continuous variable. The histogr 3 am has a negative since there is a lager more prominent tail on the lef t side, the literacy rate seems to be dense towards the end around 80% to

100%. This shows a Non-Normal Distribution of data since there is a negat ive skew. This meets our expectations of literacy rate since as stated. previously, the literacy rate of a country is commonly higher than lower.

```
ggplot(newdataset, aes(dem_score14)) +
  geom_histogram(aes(y=..density..), color="red", fill="lightpink")+
  geom_density(alpha=.2, fill="#FF6666")+
  labs(title = "Democracy Scores",
       x= 'Democracy score',
       y = 'Density')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`
.
```



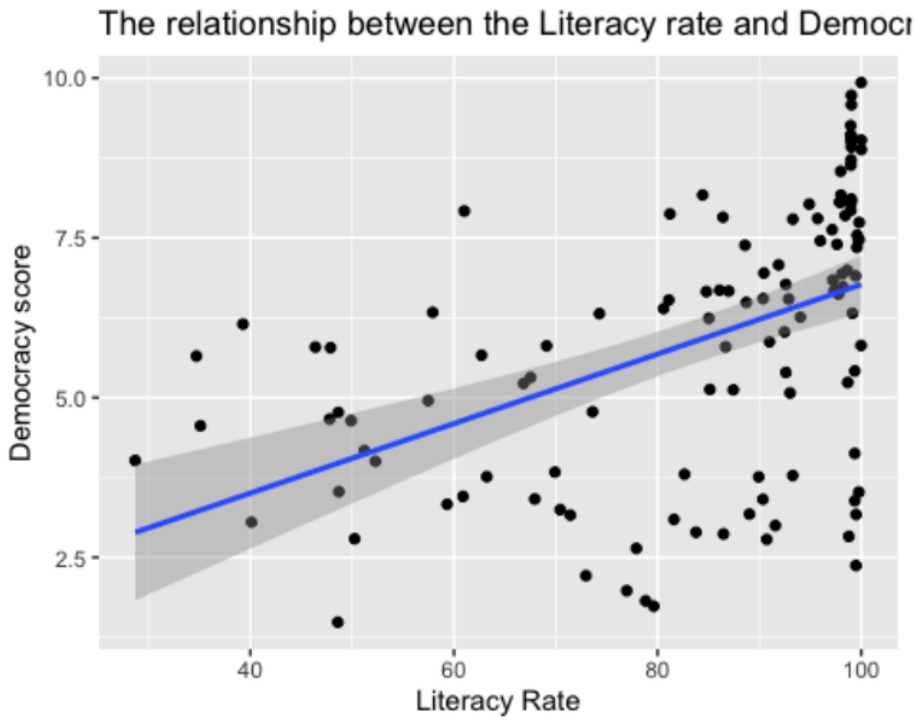
'Democracy score' is also a ratio variable as result a histogram is best suited to visualize the data. Unlike literacy rate this can be seen as more irregular, since it is a bimodal shape, having two peaks ranging from 3 – 7.5 Illustrating the non-normal distribution because of having two peaks. The peaks show that there is a large probability a country having some form of democracy since the first peak is from 2.5 - 4 however the second peak is larger and more denser ranging from 5.5-8. The first peak shows there's a split despite being less dense and larger than the second peak, it is still prominent showing some countries have a partial democracy and still have room for improvement. This is support in the mean and median that we analyzed prior.

Good

Question 3

Thinking about the type of variable you selected, create a graph that will illustrate the relationship between your dependent and independent variables. Remember that visualisations have to be nice to look at, represent the data truthfully, be clear and informative. In other words, do not forget to add titles, labels and so on. (15 points)

```
#A scatter plot to show the relationship between the two variables
ggplot(newdataset,aes(x= literacy, y = dem_score14))+
  geom_point(position = 'jitter')+
  geom_smooth(method = 'lm', se = T)+
  labs(x='Literacy Rate',
       y='Democracy score',
       title = 'The relationship between the Literacy rate and Democ
racy score')

## `geom_smooth()` using formula 'y ~ x'
```



Scatterplots are used to describe the relation between the two continuous variables. The scatterplot enables us to see a clear relation between the two variables, since as the Literacy rate the predictor variable increase so does the Democracy score outcome variable. The cluster is located when the literacy rate is 100% and the democracy score ranges from 7.5 -9.9. Suggesting there is a strong link between the two variables. The blue

line seen is the line of regression, which has a positive incline meaning there is a positive correlation overall. However, we can also see a few outliers scattered across the plot.

Question 4

Test the hypothesis you formulated in Step 1 using a t-test or a non-parametric test, depending on which one is appropriate (hint: remember it depends on whether the variable is normally distributed or not). Report the test statistics, and its associated p-value. Use the .05 cut off point for statistical significance and interpret the results. (15 points)

Since both variables are continuous to test the hypothesis, I will run a correlation test.

```
#A correlation test
cor.test(newdataset$literacy, newdataset$dem_score14, method = 'spea
rman', exact = FALSE)

##
##  Spearman's rank correlation rho
##
## data:  newdataset$literacy and newdataset$dem_score14
## S = 130837, p-value = 1.155e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5456726
```

The use of spearman's rank correlation is because the variables are not normally disturbed. The correlation coefficient between x and y is 0.5456726, this shows a direct relationship between the two which is moderate.

Good

The p-value is 1.155e-10, there is a low probability of accepting the null hypothesis as a result we reject the null hypothesis which states there is no correlation between the two variables.

Question 5

Test the hypothesis you formulated in Step 1 using a regression model. Present the regression results in a table and interpret them. Use the .05 cut off point for statistical significance

```
#regression results in a table
m1<-lm(dem_score14~literacy,newdataset)
summary(m1)

##
## Call:
## lm(formula = dem_score14 ~ literacy, data = newdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3727 -1.2366  0.3688  1.3345  3.2721
```

The residuals are the distance from the data to the fitted line, it should be symmetric-al, meaning the min and max should be the same distance to 0. Our min is -4.3727 & the max is 3.2721 which is similar distance to 0. Also, the 1stQ & 3rd Q should have the same distance to 0 which is symmetrical with a difference of 0.0979. Which suggests that the model is statistically significant

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.328824   0.791703   1.678   0.0959 .
## literacy    0.054411   0.009242   5.887 3.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.889 on 118 degrees of freedom
## Multiple R-squared:  0.227, Adjusted R-squared:  0.2205
## F-statistic: 34.66 on 1 and 118 DF, p-value: 3.777e-08
```

The F-value measures the significance of the overall model.

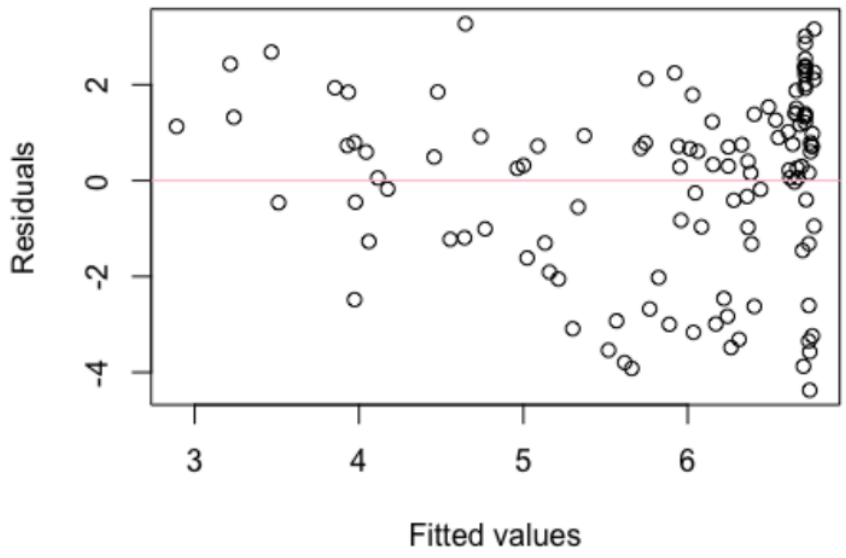
The p-value for the F is statically significant, which means that the model is better than a model where the effect of the predictor would be 0.

```
screenreg(m1)
```

```
##
## ========================
##                 Model 1
## ------------------------
## (Intercept)       1.33
##                  (0.79)
## literacy          0.05 ***
##                  (0.01)
## ------------------------
## R^2               0.23
## Adj. R^2          0.22
## Num. obs.       120
## ========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The R^2 has a value of 0.23, and Adjusted R^2 has a value of 0.22. Suggesting that the model explains 22-23% of the variation in the outcome variable. The intercept value is 1.33, this is the value of democracy score when literacy rate is 0. Suggesting the lower the literacy rate the lower the democracy score thus approving the hypotheses. The regression coefficient is 0.05***, which statically significant than the cut-off point 0.05. We can reject the null hypothesis that states that there is no relationship between the Democracy score and the literacy rate, based on the p-value of the regression coefficient, which is smaller than the .05 cut-off point.

```
#Plotting
plot(y=m1$residuals, x=m1$fitted.values, ylab='Residuals', xlab='Fitted values')
abline(h=0, col="Pink")
```

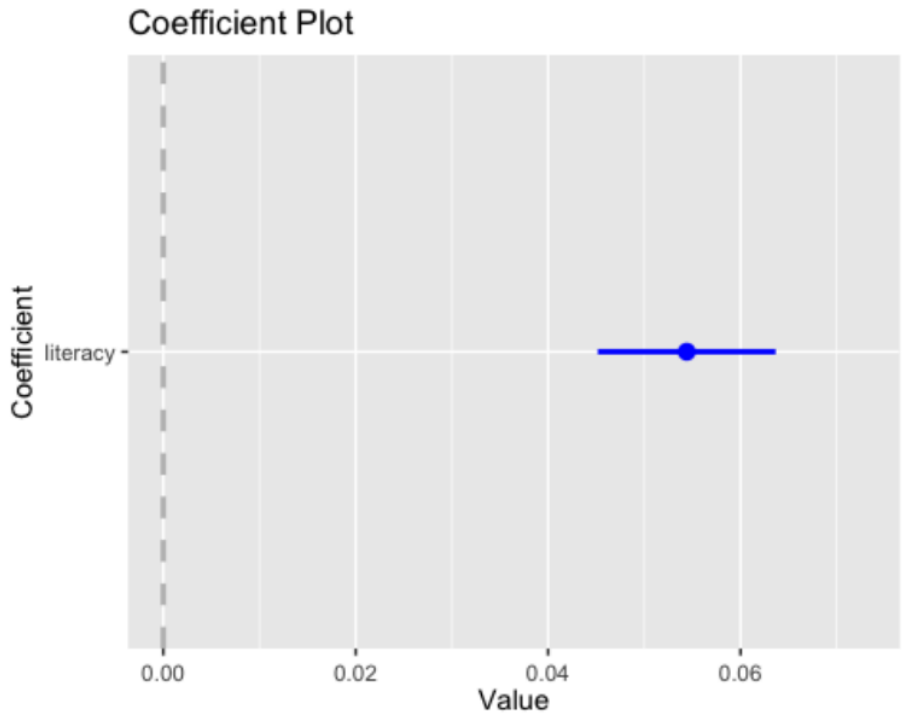


The variability is similar throughout the model. With no curvature which suggests this residual plot gives no indication that more hypothesis is false.

```
confint(m1,level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -0.23896267 2.89661136
## literacy     0.03610826 0.07271338
```

```
coefplot(m1,inncerCI=2,outerCI=2,intercept = FALSE)
```



Since literacy rate does not touch zero this shows that is it significant. Thus, again rejecting the null hypotheses.

Question 6

Expand on the relationship you tested above, by choosing another two variables that could improve your model. Feel free to recode variables.

```
str(d)
```

```
## 'data.frame':    167 obs. of  19 variables:
##  $ country       : Factor w/ 167 levels "Afghanistan",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ dem_economist : int  0 0 0 0 1 0 1 1 0 0 ...
##  $ confidence    : num  NA 49.3 52.1 NA 7.3 ...
##  $ dem_score14   : num  2.77 5.67 3.83 3.35 6.84 4.13 9.01 8.54 2
.83 2.87 ...
##  $ durable       : int  4 3 5 3 17 2 99 54 5 25 ...
##  $ enpp3_democ08 : Factor w/ 3 levels "1-3 parties",..: NA 1 NA N
A 1 3 1 2 NA NA ...
##  $ effectiveness : num  13.7 35.5 32.6 19.1 35 ...
##  $ gdp_10_thou   : num  NA 0.1535 0.1785 0.0857 0.2797 ...
##  $ gender_unequal: num  0.797 0.545 0.594 NA 0.534 0.57 0.296 0.3
0.553 0.512 ...
##  $ gender_equal3 : Factor w/ 3 levels "High","Low","Medium": NA N
A NA NA 1 NA 1 1 NA 2 ...
##  $ hdi           : num  0.349 0.719 0.677 0.403 0.775 0.695 0.937
```

```
0.851 0.713 0.801 ...
##  $ literacy      : num  28.1 NA 69.9 67.4 97.2 99.4 99 98 98.8 86
.5 ...
##  $ pop_age       : num  16.9 30 26.2 17.4 30.4 32 37.8 41.8 28.4
28.1 ...
##  $ pop_total     : num  29.1 3.2 35.4 19 40.7 3.1 21.5 8.4 8.9 0.
8 ...
##  $ spendeduc     : num  NA 2.9 4.3 2.6 4.9 3 4.7 5.4 1.9 2.9 ...
##  $ spendhealth   : num  1.8 2.9 3.6 2 5.1 2.1 6 7.7 1 2.6 ...
##  $ womyear2      : Factor w/ 2 levels "1944 or before",..: NA 1 2
2 2 1 1 1 1 2 ...
##  $ women13       : num  NA 15.7 NA NA 37.4 10.7 24.7 27.9 NA NA .
..
##  $ votevap00s    : num  NA 59.6 NA NA 70.9 ...

summary(d)

##         country    dem_economist      confidence      dem_score14
##  Afghanistan:  1   Min.   :0.0000   Min.   : 6.495   Min.   :1.08
0
##  Albania    :  1   1st Qu.:0.0000   1st Qu.:38.889   1st Qu.:3.61
0
##  Algeria    :  1   Median :0.0000   Median :49.508   Median :5.79
0
##  Angola     :  1   Mean   :0.4551   Mean   :48.900   Mean   :5.54
8
##  Argentina  :  1   3rd Qu.:1.0000   3rd Qu.:59.523   3rd Qu.:7.39
5
##  Armenia    :  1   Max.   :1.0000   Max.   :99.862   Max.   :9.93
0
##  (Other)    :161                    NA's   :99
##     durable            enpp3_democ08 effectiveness      gdp_10_th
ou
##  Min.   :  0.00   1-3 parties :28    Min.   :  7.801   Min.   :0.
0090
##  1st Qu.:  4.00   4-5 parties :23    1st Qu.: 28.132   1st Qu.:0.
0436
##  Median :  9.00   6-11 parties:30    Median : 39.007   Median :0.
1656
##  Mean   : 23.11   NA's        :86    Mean   : 46.036   Mean   :0.
6258
##  3rd Qu.: 31.25                      3rd Qu.: 62.884   3rd Qu.:0.
4894
##  Max.   :191.00                      Max.   :100.000   Max.   :4.
7354
##  NA's   :19                          NA's   :14        NA's   :18
##  gender_unequal   gender_equal3      hdi            literacy
##  Min.   :0.1740   High  :23     Min.   :0.1400   Min.   : 21.80
##  1st Qu.:0.3885   Low   :24     1st Qu.:0.4675   1st Qu.: 67.90
##  Median :0.5940   Medium:25     Median :0.6690   Median : 88.70
##  Mean   :0.5466   NA's  :95     Mean   :0.6325   Mean   : 80.61
##  3rd Qu.:0.6990                 3rd Qu.:0.7810   3rd Qu.: 98.20
##  Max.   :0.8530                 Max.   :0.9380   Max.   :100.00
##  NA's   :32                     NA's   :9        NA's   :10
```

```
##     pop_age         pop_total           spendeduc       spendhealth
##  Min.   :15.00   Min.   :   0.300   Min.   : 0.600   Min.   : 0.2
00
##  1st Qu.:20.30   1st Qu.:   3.875   1st Qu.: 3.300   1st Qu.: 1.9
00
##  Median :26.30   Median :  10.000   Median : 4.400   Median : 3.2
50
##  Mean   :28.09   Mean   :  41.484   Mean   : 4.488   Mean   : 3.6
66
##  3rd Qu.:36.60   3rd Qu.:  29.075   3rd Qu.: 5.375   3rd Qu.: 5.1
25
##  Max.   :44.70   Max.   :1354.100   Max.   :13.600   Max.   :11.5
00
##  NA's   :1       NA's   :1          NA's   :17       NA's   :3
##            womyear2     women13        votevap00s
##  1944 or before:56   Min.   : 2.70   Min.   :27.59
##  After 1944    :90   1st Qu.:12.18   1st Qu.:54.24
##  NA's          :21   Median :20.80   Median :65.12
##                      Mean   :21.10   Mean   :64.87
##                      3rd Qu.:27.65   3rd Qu.:77.48
##                      Max.   :44.70   Max.   :98.39
##                      NA's   :77      NA's   :94
```

6a) Create hypotheses for each new variable (and your outcome variable)

The hypothesis within this section will be there is a relationship between gender inequality and the democracy score. Both variables are ratio variables meaning they are continuous. The null hypothesis states there is no relationship between the two variables and as a result is a null effect. The hypothesis is determined by there being a statistically significance below or equal to 0.05. However, to reject the null hypothesis the statistically significance needs to be above 0.05.

The hypothesis within this section will be there is a relationship between expenditure on the education system and the democracy score. Both variables are ratio variables meaning they are continuous. The null hypothesis states there is no relationship between the two variables and as a result is a null effect. The hypothesis is determined by there being a statistically significance below or equal to 0.05. However, to reject the null hypothesis the statistically significance needs to be above 0.05.

```
#Create tables for both new variables
table(newdataset$gender_unequal)

##
## 0.174 0.209 0.212 0.228 0.234 0.236  0.24 0.248 0.251 0.255  0.26
0.273 0.279
##     1     1     1     1     1     1     1     1     1     1     1
1     1
##  0.28 0.284 0.289 0.296   0.3  0.31 0.316 0.317 0.318  0.32 0.325
0.332 0.344
##     1     1     1     1     1     2     1     1     1     1     1
1     1
## 0.345 0.352 0.359 0.382 0.399   0.4 0.405 0.409 0.429 0.442 0.451
0.463 0.464
##     1     1     1     1     1     1     1     1     1     1     1
1     1
## 0.466 0.473 0.478 0.493 0.501 0.504 0.505 0.508 0.512 0.515 0.523
```

```
0.53 0.534
##     1     2     1     1     1     1     1     1     1     1     1
1     1
## 0.553  0.56 0.561 0.568  0.57 0.575 0.576 0.586 0.594 0.597 0.614
0.615 0.616
##     1     1     1     1     1     1     1     1     1     1     1
1     1
## 0.621 0.623 0.627 0.631 0.634 0.635 0.638 0.643 0.645 0.646  0.65
0.653 0.658
##     1     1     1     1     1     1     1     1     1     1     1
1     1
## 0.663 0.668 0.671 0.672 0.674 0.678  0.68 0.685 0.687 0.693 0.705
0.713 0.714
##     1     1     1     2     2     1     1     1     1     1     1
1     1
## 0.715 0.716 0.718 0.721 0.727 0.729 0.731 0.734 0.738 0.742 0.744
0.748 0.752
##     1     1     1     1     1     1     1     1     2     1     1
1     1
## 0.756 0.758 0.759  0.76 0.763 0.765 0.766 0.768 0.799 0.807 0.853
##     1     1     1     1     1     1     1     1     1     1     1

table(newdataset$spendeduc)

##
##  0.9    1  1.3  1.4  1.6  1.8  1.9    2  2.2  2.3  2.4  2.6  2.7
2.8  2.9    3
##    1    1    1    1    1    1    2    1    1    1    1    1    3
3    5    1
##  3.1  3.2  3.3  3.4  3.5  3.6  3.7  3.8  3.9    4  4.1  4.2  4.3
4.4  4.5  4.6
##    1    2    1    2    2    4    4    8    3    2    2    3    2
4    1    2
##  4.7  4.8  4.9    5  5.1  5.2  5.3  5.4  5.5  5.6  5.7  5.9  6.1
6.2  6.3  6.4
##    2    2    7    4    3    2    4    3    2    1    2    1    1
1    1    1
##  6.5  6.6  6.7    7  7.1  7.2  7.5  7.9  8.1  8.2 12.4 13.6
##    1    1    2    1    1    2    1    2    1    1    1    1
```

6b) Present univariate analysis on the new variables (descriptive statistics and visualisations)

```
summary(newdataset$gender_unequal)

##    Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
## 0.1740 0.3762    0.575   0.5392  0.6855   0.8530
```

The range of gender inequality is from 0.1740 to 0. 853. By comparing the median & mean we can tell how evenly the data is distrusted, the mean is 53.92 & the median is 57.5, they are similar. This shows that gender inequality id spread out evenly.

```r
summary(newdataset$spendeduc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   3.400   4.300   4.498   5.300   13.600
```

The range of expenditure of the education system is 0.9 to 13.6, suggesting some countries do not spend/ value the education system. Since the mean is larger than the median, we can infer that this creates a positive skew.

```r
#to see what type of variables they are
class(newdataset$spendeduc)
```

```
## [1] "numeric"
```

This is a ratio variable, which is a continuous variable.

```r
class(newdataset$gender_unequal)
```

```
## [1] "numeric"
```

This is a ratio variable, which is a continuous variable.

```r
ggplot(newdataset, aes(gender_unequal)) +
  geom_histogram(aes(y=..density..), color="blue", fill="lightblue")
+
  geom_density(alpha=.2, fill="#FF6666")+
  labs(title = "Gender inequality",
       x= 'Gender inequality',
       y='Density')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`
.
```

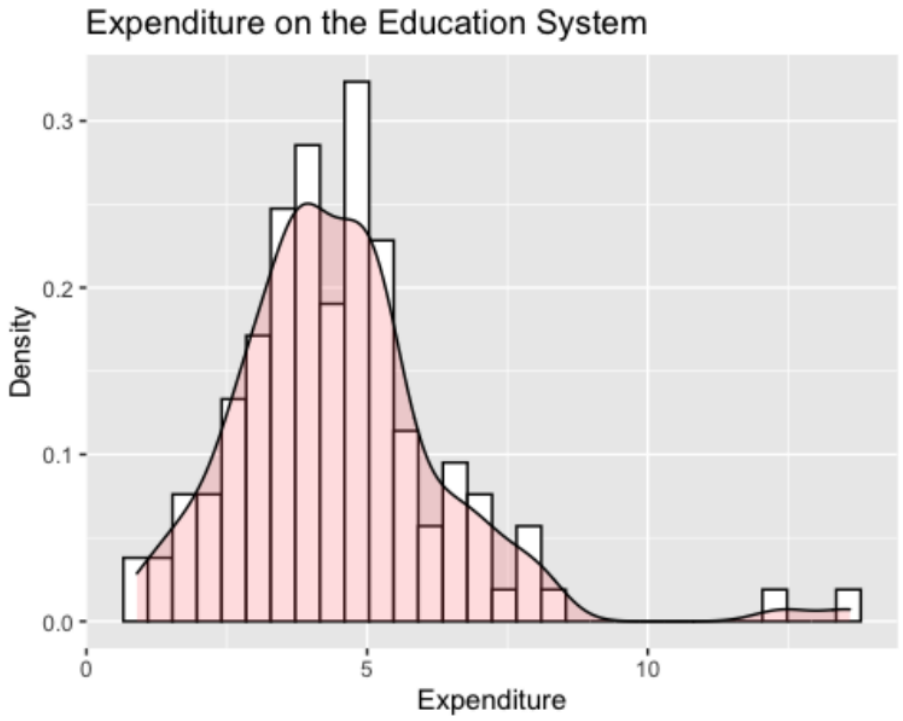
'Gender inequality' is also a ratio variable as result a histogram is best suited to visualize the data. It has bimodal shape, having two peaks ranging from 0.3- 0.7. Illustrating the non-normal distribution because of having two peaks. The first peak is less dense and tall compared to the second peak which is larger and dense. The data shows there is more gender inequality. This is support in the mean and median that we analyzed prior.

```
ggplot(newdataset,aes(x= gender_unequal, y = dem_score14))+
  geom_point(position = 'jitter')+
  geom_smooth(method = 'lm', se = T)+
  labs(x='gender inequality ',
       y='Democracy score',
       title = 'The relationship between the Democracy score and ine
quality amongst gender')

## `geom_smooth()` using formula 'y ~ x'
```

Scatterplots are used to describe the relation between the two continuous variables. The scatterplot enables us to see a clear relation between gender inequality and democracy score. The line of regression shows a negative correlation with a negative decline. There is no prominent cluster found, however there is a descending pattern.

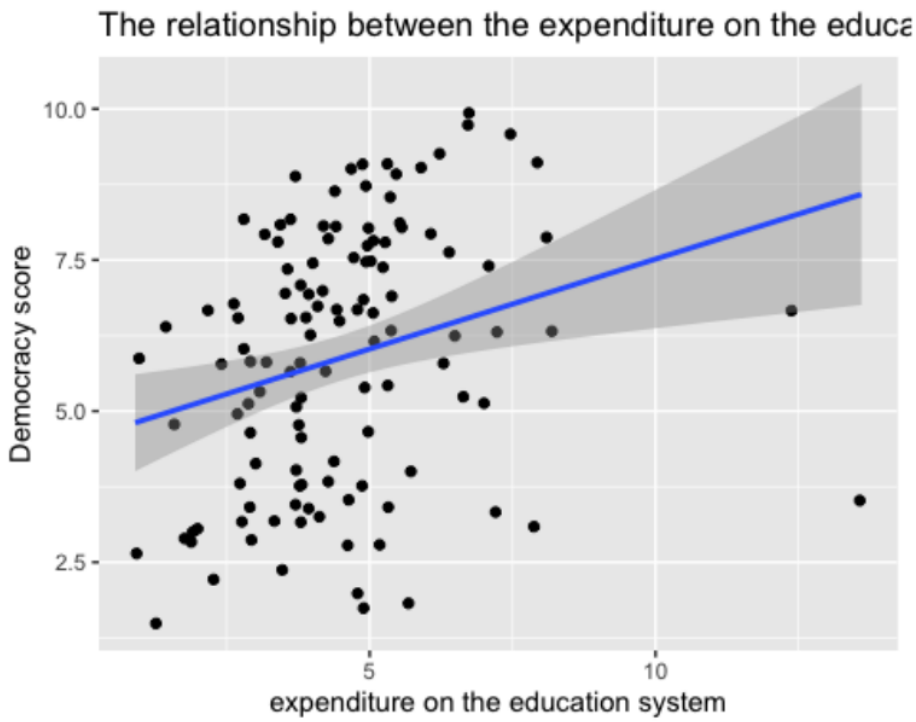
```
ggplot(newdataset, aes(spendeduc)) +
  geom_histogram(aes(y=..density..), color="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666")+
  labs(title = "Expenditure on the Education System",
       x='Expenditure',
       y= 'Density')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`
.
```



'Expenditure on the education system' is also a ratio variable as result a histogram is best suited to visualize the data. It has a positive skew since it peaks at the beginning, causing the right tail longer than the left. This is support what we analyzed prior, since the mean is higher the median.

```
ggplot(newdataset,aes(x= spendeduc, y = dem_score14))+
  geom_point(position = 'jitter')+
  geom_smooth(method = 'lm', se = T)+
  labs(x='expenditure on the education system',
       y='Democracy score',
       title = 'The relationship between the expenditure on the educ
ation system  and Democracy score')

## `geom_smooth()` using formula 'y ~ x'
```



As the histogram suggested there is a large cluster at the beginning of the scatterplot. However, identifying the relationship between the two variables is difficult due to the cluster, showing no clear relation or pattern.

6C

Run a regression model that includes the new variables. Present the regression results in a table and interpret them. Use the .05 cut off point for statistical significance. (10 points)

```
m2<-lm(dem_score14~literacy+spendeduc+gender_unequal,newdataset)
summary(m2)

##
## Call:
## lm(formula = dem_score14 ~ literacy + spendeduc + gender_unequal,
```

```
##     data = newdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8858 -0.9561  0.2369  1.2584  3.7490
```

The residuals are symmetrical, and the median is close to 0 **which suggests that the model is statistically significant.**

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.928704   1.598211   6.212 8.42e-09 ***
## literacy        -0.004008   0.011732  -0.342    0.733
## spendeduc        0.117775   0.079876   1.474    0.143
## gender_unequal -7.874439   1.237544  -6.363 4.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 116 degrees of freedom
## Multiple R-squared:  0.4481, Adjusted R-squared:  0.4338
## F-statistic: 31.39 on 3 and 116 DF,  p-value: 6.25e-15
```

```
screenreg(m2)
```

```
##
## ==========================
##                  Model 1
## --------------------------
## (Intercept)       9.93 ***
##                  (1.60)
## literacy         -0.00
##                  (0.01)
## spendeduc         0.12
##                  (0.08)
## gender_unequal   -7.87 ***
##                  (1.24)
## --------------------------
## R^2               0.45
## Adj. R^2          0.43
## Num. obs.       120
## ==========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The R^2 has a value of 0.45, and Adjusted R^2 has a value of 0.43. Suggesting that the model explains 43-45% of the variation in the outcome variable.

The intercept value is 9.93*, shows us the value of the dependent variable when the independent variable is 0.**

The gender inequality variable has a statistically significant since the p-value is -7.87*, thus supporting the hypothesis and rejecting the null. This is similar to the literacy variable since they p-value is -0.00. However, expenditure on the education system is not statistically significant since the p value is 0.12.**

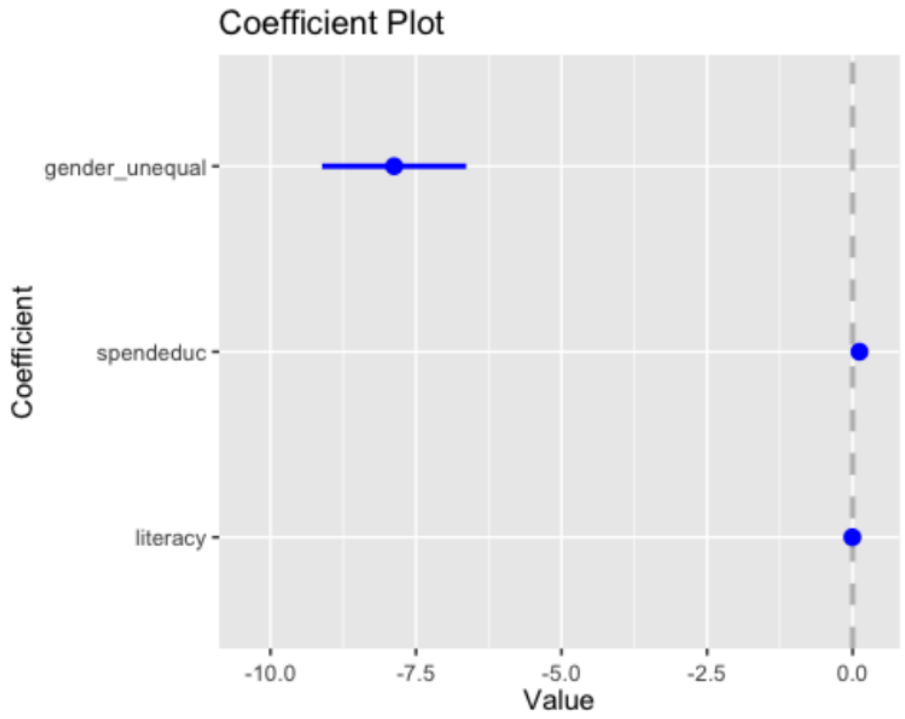
incorrect

As result, the democracy score improves when there is less gender inequality this has an apparent correlation. Similarly, to the literacy rate of a country effects the democracy but not as much. Finally, the expenditure on the education system has no correlation to the democracy score of a country.

```
confint(m2,level = 0.95)
```

```
##                       2.5 %      97.5 %
## (Intercept)      6.76324575 13.09416151
## literacy        -0.02724469  0.01922799
## spendeduc       -0.04042910  0.27598002
## gender_unequal -10.32555039 -5.42332861
```

```
coefplot(m2,inncerCI=2,outerCI=2,intercept = FALSE)
```



Since both the literacy rate and expenditure on the education is touching the 0 this shows they are insignificant. These two variables accept the null hypothesis.

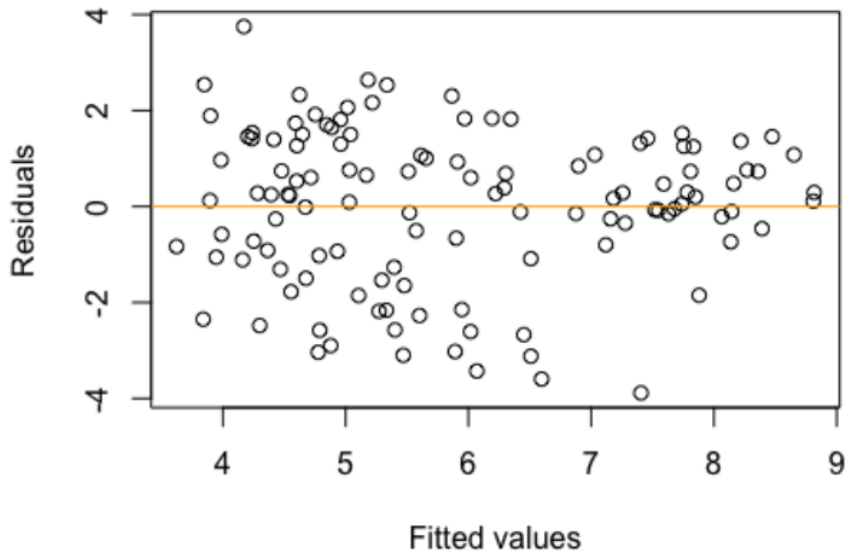
However, gender inequality does not touch 0 it shows it is statically significant and this variable alone rejects the null hypothesis.

```
#plot the residuals against the fitted values.
names(m2)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual
```

```
"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
plot(y=m2$residuals, x=m2$fitted.values, ylab='Residuals', xlab='Fit
ted values')
abline(h=0, col='orange')
```



The variability is similar throughout the model. With no curvature which suggests this residual plot gives no indication of problems.

```
#The Ramsay TEST
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

resettest(m2, power = 2:3, type = 'fitted')

##
##  RESET test
##
## data:  m2
## RESET = 2.5375, df1 = 2, df2 = 114, p-value = 0.08352
```

This rejects the null hypothesis, since the p-value is <0.05 , which is 0.08352.

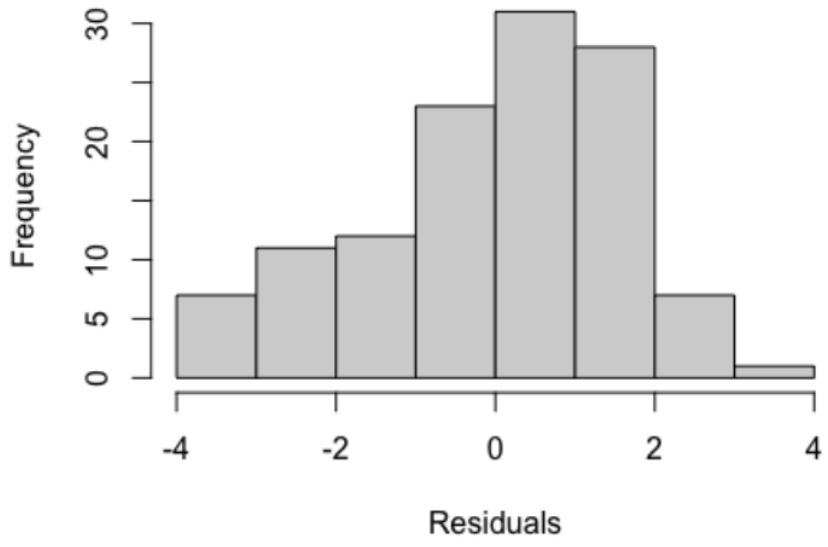
Since the residuals are normally distributed, I. have chosen to use homoskedasticity

Unclear

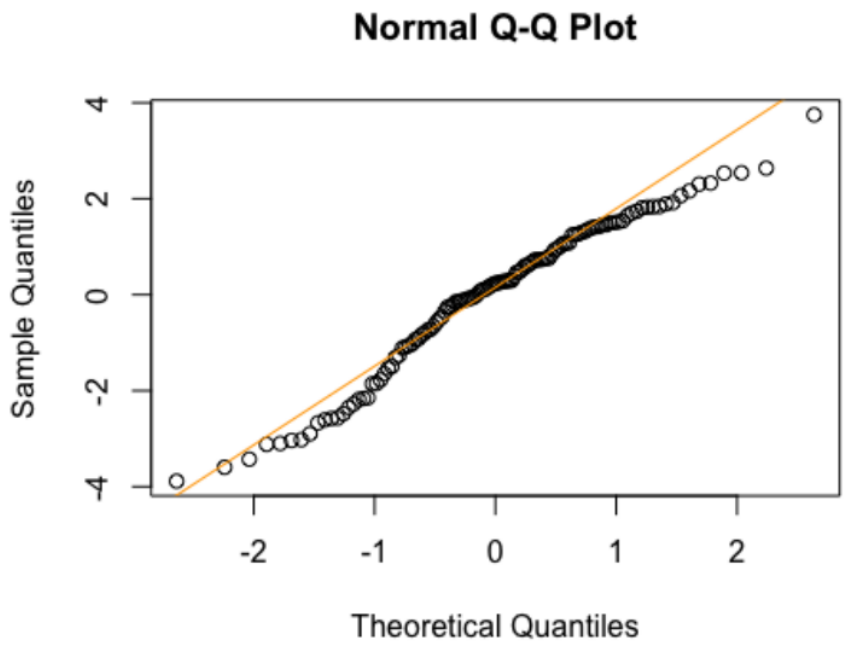
```
#homosekedasticity - constant variance of errors
bptest(m2, studentize=FALSE)

##
##  Breusch-Pagan test
##
## data:  m2
## BP = 6.7432, df = 3, p-value = 0.08055

shapiro.test(m2$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.96942, p-value = 0.007805

ncvTest(m2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.246098, Df = 1, p = 0.071594
```

The p-values throughout test are all less than <0.05, as result we reject null hypothesis. Suggesting that data is not normally distributed.

```r
#Histogram of the residuals
hist(m2$residuals, xlab='Residuals', main='')
```




```r
#Q-Q Plot
qqnorm(m2$residuals)
qqline(m2$residuals, col='Orange')
```

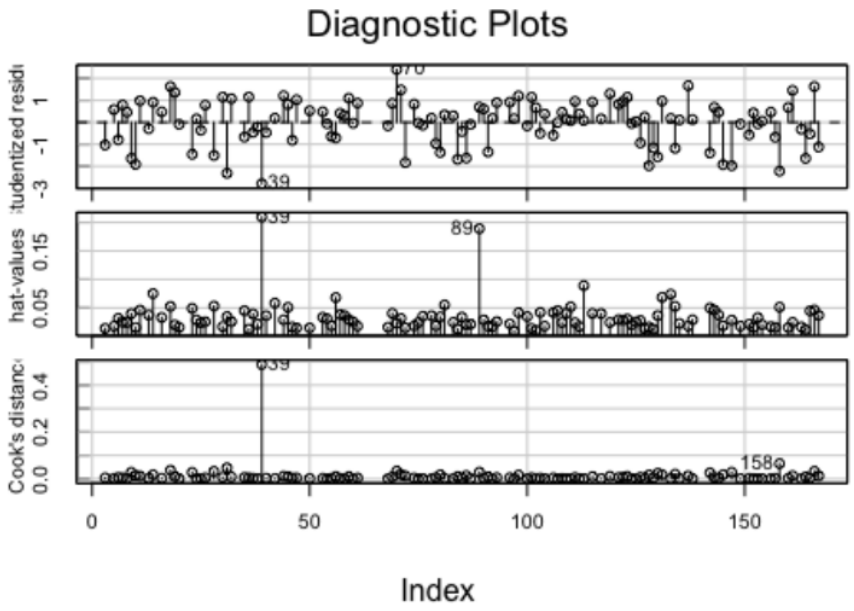


From the QQ plot we can we a selected portion of the residuals lie on the line; however, the end of the tail is below the line. This suggests our residuals may not be normally distributed

```
#t test of coefficients
library(sandwich)
coeftest(m2,vcov=vcovHC)

##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     9.9287036  1.7316175  5.7338 7.931e-08 ***
## literacy       -0.0040084  0.0112381 -0.3567    0.7220
## spendeduc       0.1177755  0.1356743  0.8681    0.3871
## gender_unequal -7.8744395  1.2212265 -6.4480 2.707e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


#Multicollinearity
d3<- newdataset[, c('literacy','gender_unequal','spendeduc')]

cor(d3,use = 'complete.obs')

##                  literacy gender_unequal  spendeduc
## literacy        1.0000000     -0.7406781  0.2120941
```

```
## gender_unequal -0.7406781      1.0000000 -0.2543434
## spendeduc       0.2120941     -0.2543434  1.0000000

# variation inflation test for multicollinearity
vif(m2)

##       literacy      spendeduc gender_unequal
##       2.218303       1.070590       2.265042

Since all my variables are above 0.8, this means that all three vari
ables

Linearly relate. This shows that the data has multicollinearity.

#to find out if there is any influential outliers
influenceIndexPlot(m2, vars = c ('Studentized','hat', 'cook'))
```

Incorrect



Studentized plot illustrates that there are outliers.

In the hat value plot, there is a variation of points and with some whom are high and have high leverage points such as 89.

In the cooks plot we can observe there are none influential with two outliers, this means that we do not have any influential data points and we do not need to make any corrections.

In the plots, 39 has the highest leverage and highest Cook's *d*-value.

6D

Compare the new regression model to the model from Step 5, using the appropriate statistical test. Report the results and interpret them. Is the second regression model more informative? (5 points)

```
#anova - comparing two models
anova(m2,m1)
```

```
## Analysis of Variance Table
##
## Model 1: dem_score14 ~ literacy + spendeduc + gender_unequal
## Model 2: dem_score14 ~ literacy
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    116 300.58
## 2    118 420.96 -2   -120.39 23.23 3.276e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
d2<-d[complete.cases(d$dem_score14,d$literacy,d$gender_unequal,d$spe
ndeduc),]


#the simple regression analysis
m3<-lm(dem_score14~literacy,d2)
summary(m3)
```

```
##
## Call:
## lm(formula = dem_score14 ~ literacy, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3727 -1.2366  0.3688  1.3345  3.2721
```

The median is close to zero and the residuals are somewhat symmetrical. As result we can reject the null hypothesis.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.328824   0.791703   1.678   0.0959 .
## literacy    0.054411   0.009242   5.887 3.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.889 on 118 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2205
## F-statistic: 34.66 on 1 and 118 DF,  p-value: 3.777e-08
```

```
screenreg(m3)
```

```
##
## ========================
##                 Model 1
## ------------------------
## (Intercept)      1.33
##                 (0.79)
## literacy         0.05 ***
##                 (0.01)
## ------------------------
## R^2              0.23
## Adj. R^2         0.22
## Num. obs.      120
## ========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

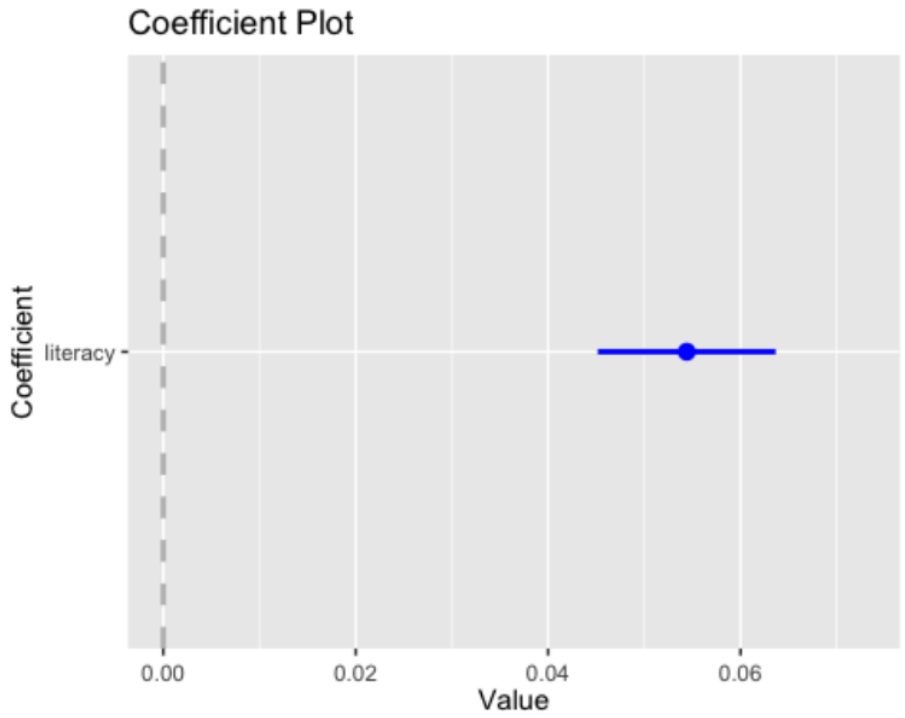
*R^2 is 0.23 and Adjusted R^2 is 0.22, being that there is an outcome variance from 22%-23%. The p-value is statically significant, being 0.05***. The intercept value is 1.33, this is the value of democracy score when literacy rate is 0. Suggesting the lower the literacy rate the lower the democracy score thus approving the hypotheses. We can reject the null hypothesis that states that there is no relationship between the Democracy score and the literacy rate, based on the p-value of the regression coefficient, which is smaller than the .05 cut-off point.*

```
confint(m3,level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.23896267 2.89661136
## literacy     0.03610826 0.07271338
```

```
coefplot(m3,inncerCI=2,outerCI=2,intercept = FALSE)
```

The intercept value is 1.33, this is the value of democracy score when literacy rate is 0. Suggesting the lower the literacy rate the lower the democracy score thus approving the hypotheses. The regression coefficient is 0.05*, which statically significant than the cut-off points 0.05. We can reject the null hypothesis that states that there is no relationship between the Democracy score and the literacy rate, based on the p-value of the regression coefficient, which is smaller than the .05 cut-off point.**



The coefficient plot again highlights the statically significance of the literacy rate since it does not touch zero. As result, we reject the null hypothesis.

```
#the simple regression analysis
m4<-lm(dem_score14~literacy+gender_unequal+spendeduc,d2)
summary(m4)
```

```
##
## Call:
## lm(formula = dem_score14 ~ literacy + gender_unequal + spendeduc,
##     data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8858 -0.9561  0.2369  1.2584  3.7490
```

The median is close to zero and the residuals are somewhat symmetrical. As result we can reject the null hypothesis.

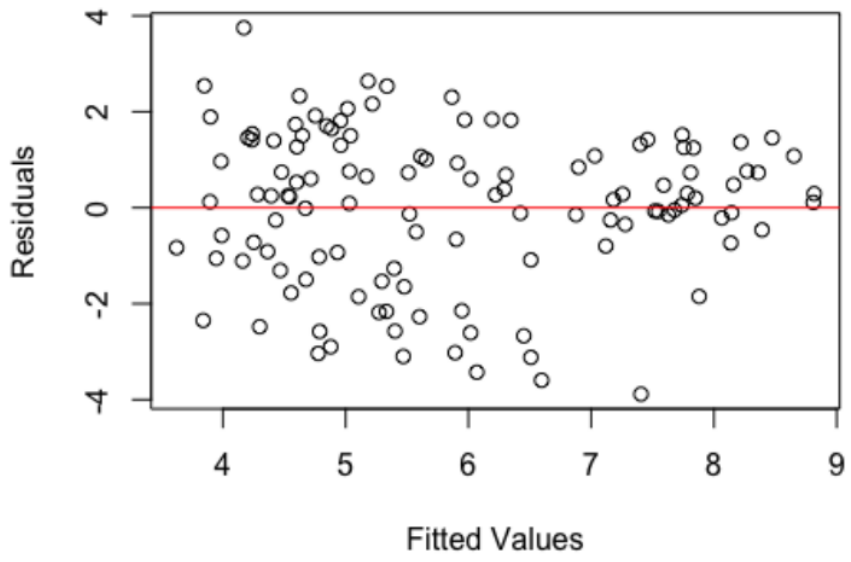
```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.928704   1.598211   6.212 8.42e-09 ***
## literacy       -0.004008   0.011732  -0.342    0.733
## gender_unequal -7.874439   1.237544  -6.363 4.08e-09 ***
## spendeduc       0.117775   0.079876   1.474    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 116 degrees of freedom
```

```
## Multiple R-squared:  0.4481, Adjusted R-squared:  0.4338
## F-statistic: 31.39 on 3 and 116 DF,  p-value: 6.25e-15
```

```
screenreg(m4)
```

```
##
## ==========================
##                   Model 1
## --------------------------
## (Intercept)        9.93 ***
##                   (1.60)
## literacy          -0.00
##                   (0.01)
## gender_unequal    -7.87 ***
##                   (1.24)
## spendeduc          0.12
##                   (0.08)
## --------------------------
## R^2                0.45
## Adj. R^2           0.43
## Num. obs.        120
## ==========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

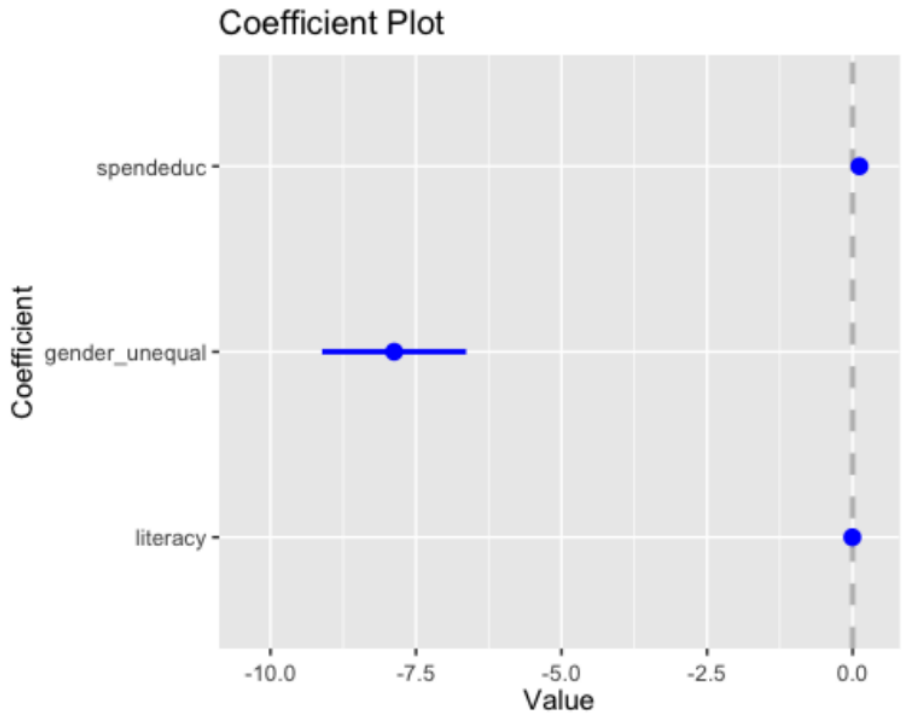
```
plot(y=m4$residuals, x=m4$fitted.values, ylab = 'Residuals', xlab =
'Fitted Values')
abline(h=0, col='red')
```




```
confint(m4,level = 0.95)
```

```
##                       2.5 %      97.5 %
## (Intercept)      6.76324575 13.09416151
## literacy        -0.02724469  0.01922799
## gender_unequal -10.32555039 -5.42332861
## spendeduc       -0.04042910  0.27598002
```

```
coefplot(m4,inncerCI=2,outerCI=2,intercept = FALSE)
```




```
#Run the ANOVA test and interpret
anova(m3,m4)

## Analysis of Variance Table
##
## Model 1: dem_score14 ~ literacy
## Model 2: dem_score14 ~ literacy + gender_unequal + spendeduc
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    118 420.96
## 2    116 300.58  2    120.39 23.23 3.276e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

end of term project

GRADEMARK REPORT

FINAL GRADE

85 /100

GENERAL COMMENTS

Instructor

You did a very good job overall with this project. There is enough evidence to show that you understand hypothesis testing, regression, and you know how to use R. Well done on that front. However, there is still room for improvement, particularly in terms of accuracy in the interpretation. There were a few sections where you confused your results and interpreted them the wrong way around (see in-text comments). Also, you missed your Anova interpretation. While minor, these mistakes can have an impact on the overall interpretation of your results, so keep that in mind.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

Comment 1

You needed to choose variables that were different from the example on Moodle!

PAGE 7

Comment 2

Careful with the word significant, as this has a specific meaning in statistics.

Comment 3

A negative skew, I'm assuming?

RUBRIC: PO687 - 2020 9.50 / 10

PART 1 (15%) 10 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4
(6)

SCALE 5
(8)

SCALE 6
(10)

PART 2 (15%) 10 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4
(6)

SCALE 5
(8)

SCALE 6
(10)

PART 3 (15%) 10 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4

(6)

SCALE 5
(8)

SCALE 6
(10)

PART 4 (15%) 10 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4
(6)

SCALE 5
(8)

SCALE 6
(10)

PART 5 (15%) 10 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4
(6)

SCALE 5
(8)

SCALE 6
(10)

PART 6 (25%) 8 / 10

SCALE 1
(0)

SCALE 2
(2)

SCALE 3
(4)

SCALE 4
(6)

SCALE 5
(8)

SCALE 6
(10)