INDIVIDUAL TASK COVER SHEET

JAMES COOK UNIVERSITY AUSTRALIA

MA5820 Statistical Methods for Data Scientists

Assessment Task	Assessment 3: CAPSTONE REPORT
College	

Student: Please sign, date, and attach this cover sheet to the front of your assessment task for all hard copy submissions.

Student Family Name	**Student Given Name**	**JCU Student Number**							
MILLS	ADAM	1	3	8	1	3	6	1	7
Assessment Title	Assessment 3: CAPSTONE REPORT								
Due Date	19 Aug. 2020								
Lecturer Name	Yvette Everingham								
Tutor Name									

Student Declaration

1. This assignment is my original work and no part has been copied/ reproduced from any other person's work or from any other source, except where acknowledgement has been made (see *Learning, Teaching and Assessment Policy 5.1*).
2. This work has not been submitted previously for assessment and received a grade OR concurrently for assessment, either in whole or part, for this subject (unless part of integrated assessment design/approved by the Subject Coordinator), any other subject or any other course (see *Learning, Teaching and Assessment Policy 5.9*).
3. This assignment has not been written for me.
4. We hold a copy of this assignment and can produce a copy if requested.
5. This work may be used for the purposes of moderation and identifying plagiarism.
6. We give permission for a copy of this marked assignment to be retained by the College for benchmarking and course review and accreditation purposes.

Learning, Teaching and Assessment Policy 5.1. A student who submits work containing plagiarised material for assessment will be subject to the provisions of the Student Academic Misconduct Requirements Policy.

Note the definition of plagiarism and self plagiarism in the Learning, Teaching and Assessment Policy:

Plagiarism: reproduction without acknowledgement of another person's words, work or expressed thoughts from any source. The definition of words, works and thoughts includes such representations as diagrams, drawings, sketches, pictures, objects, text, lecture hand-outs, artistic works and other such expressions of ideas, but hereafter the term 'work' is used to embrace all of these. Plagiarism comprises not only direct copying of aspects of another person's work but also the reproduction, even if slightly rewritten or adapted, of someone else's ideas. In both cases, someone else's work is presented as the student's own. Under the Australian *Copyright Act 1968* a copyright owner can take legal action in the courts against a party who has infringed their copyright.

Self Plagiarism: the use of one's own previously assessed material being resubmitted without acknowledgement or citing of the original.

Student Signature	Amills	Submission Date: 19-Aug-2020

Does human development really mean more consumption?

Statistical Methods for Data Scientists: MA5820

ASSESSMENT TASK 3: [CAPSTONE REPORT] by Mills, A

Executive Summary

The development and growth of populations is promoted by various charities and organisations around the world in the hope of eliminating poverty and increasing the quality of peoples lives. This report outlines the investigation of three important questions. 1.Does water consumption keep increasing once a country “is developed”? 2. Do “developed” countries remove more wood from land over time? 3. Do countries with more forest area harvest more wood? These questions are answered using three difference statistical analysis techniques (independent t-test, chi-squared independence test and regression analysis) on data sourced from gapminder.com. It was observed from this testing and analysis that 1. No, water consumption does not keep increasing once a country’s human development index (HDI) goes above 0.66. 2. Yes, higher HDI (developed) countries do remove more wood from their land over time than less developed countries. 3. Yes, there is a significant relationship (positive correlation) between the amount of forest land a country has and the amount of wood removed from that country. At the end of the day the answer to the title of this study is: it depends. It is hoped that this study highlights the complexity of the consumption versus development conundrum and encourages more people to do their own investigations. Education can solve a lot of the world’s problems.

Introduction

Consumption typically conjures up images of shopping centres and fast-food restaurants, but consumption is also occurring at a more fundamental level of human existence. At the bottom of Maslow's Hierarchy of Needs there is a list of physiological needs that includes water to drink and shelter to sleep and survive the elements. There are several important factors in consumption such as. As one progresses up the Hierarchy of Needs the relative importance of the lower tiers decreases, this inherently suggests that the availability of water and shelter is adequate to require less attention. Following this pretence it’s reasonable to assume for an individual person that their consumption of water and resources would plateau at a certain level.

Defined by the united nations, the human development index (HDI) uses a range of measures including life expectancy, education, and income. This report is an attempt to determine if municipal water consumption and wood removal has an identifiable relationship with the HDI. Houses are predominantly made from wood (at least structurally). The three objectives of this investigation are as follows:

1. Examine if there was a statistically significant increase in the mean water municipal water withdrawal(m3/person) from a simple random sample of HDI category 3 countries between the year 2000 and 2010.
2. Examine if there was a statistically significant relationship between HDI and the removal of wood from a county's land between the year 2000 and 2010.
3. Determine if there was a statistically significant relationship between the forest area a country had and the amount of wood that it removed during the calendar year 2000.

Data

The specified constraints limited the data to that available under the resources folder of the gap minder website and the inclusion of the human development index.

Raw data sourced included the following:

- hdi_human_development_index.csv
- municipal_water_withdrawal_cu_meters_per_person.csv
- wood_removal_cubic_meters.csv
- forest_land_total_area_ha.csv

The raw data files were in a wide format with countries in column 1 and yearly measurements in the remaining columns. Not all files had measurements for identical years therefore missing values were dealt with during analysis.

Preprocessing consisted of importing each file into R-studio and converting it from a wide format to a long format resulting in the table structure of:

<country> <year> <measure>.

Each of these dataframes were then joined together resulting in a single data frame called “main” that had the structure of:

<country><year><measure><measure><measure>

It was decided to follow the guidance of the assessment example and create one categorical variable for HDI that had three possible factors HDI-1, HDI-2, HDI-3 where HDI-1 would represent the lowest HDI observations and HDI-3 would represent the highest HDI measurements.
A numerical summary of the main dataset grouped by the categorical variable can be seen in Table 1.

Table 1: Summary of data grouped by HDI Catagory

HDI-1									
	n	mean	sd	median	min	max	range	skew	kurtosis
country*	133	6.56	3.32	7.00E+00	1.00E+00	1.20E+01	1.10E+01	-0.13	-1.31
year*	133	7.93	4.83	7.00E+00	1.00E+00	2.00E+01	1.90E+01	0.38	-0.81
waterusage	11	11.44	6.8	9.62E+00	3.73E+00	2.16E+01	1.79E+01	0.44	-1.58
hdi	133	0.28	0.03	2.80E-01	1.90E-01	3.20E-01	1.30E-01	-0.72	-0.48
wood	124	9831250	16369464	6.08E+06	2.24E+05	9.45E+07	9.43E+07	4.21	17.84
forestlandarea	133	9320451	12706062	4.40E+06	1.81E+05	4.34E+07	4.32E+07	1.66	1.54
hdicat*	133	1	0	1.00E+00	1.00E+00	1.00E+00	0.00E+00	NaN	NaN
HDI-2									
	n	mean	sd	median	min	max	range	skew	kurtosis
country*	1779	55.75	29.83	5.60E+01	1	1.05E+02	104	-0.09	-1.14
year*	1779	13.32	7.46	1.30E+01	1	2.60E+01	25	0.01	-1.19
waterusage	181	39.75	37.03	2.79E+01	3.5	2.76E+02	272.5	2.53	9.96
hdi	1779	0.51	0.1	5.10E-01	0.33	6.60E-01	0.33	-0.1	-1.26
wood	1376	16319096	44960525	5.50E+06	12400	4.35E+08	4.35E+08	6.04	40.37
forestlandarea	1765	18830978	44679214	5.01E+06	1000	5.47E+08	5.47E+08	7.23	73.02
hdicat*	1779	2	0	2.00E+00	2	2.00E+00	0	NaN	NaN
HDI-3									
	n	mean	sd	median	min	max	range	skew	kurtosis
country*	2202	56.8	32.13	5.60E+01	1	1.13E+02	1.12E+02	0.01	-1.19
year*	2202	15.25	7.3	1.60E+01	1	2.60E+01	2.50E+01	-0.3	-1.06
waterusage	281	110.49	58.6	9.23E+01	25.3	3.74E+02	3.49E+02	1.57	3.23
hdi	2202	0.78	0.07	7.70E-01	0.66	9.50E-01	2.90E-01	0.25	-1.04
wood	1582	21951790	62705196	4.20E+06	0	5.09E+08	5.09E+08	5.2	30.29
forestlandarea	2157	30838228	1.09E+08	2.24E+06	0	8.15E+08	8.15E+08	5.36	31.13
hdicat*	2202	3	0	3.00E+00	3	3.00E+00	0.00E+00	NaN	NaN
hdicat*	2202	3	0	3.00E+00	3	3.00E+00	0.00E+00	NaN	NaN

Null value handling.
Missing values were common in the main data frame so when samples were extracted the author used the "dplyr" package filter function with the !is.na() function conditions incorporated to ensure that only rows with measurements were placed in the random samples. In addition, where is was appropriate do to so values equal to 0 were also filtered out.

Data subset for objective 1.
Two simple random samples were extracted from the main data frame for the year 2000 and the year 2010. During exploratory data analysis it was observed that the water removal variable was not normally distributed therefore a new variable was generated for each subset for the log transformation. Following the extraction of both samples normality evaluation was carried out. An example of this can be seen in figure 1.

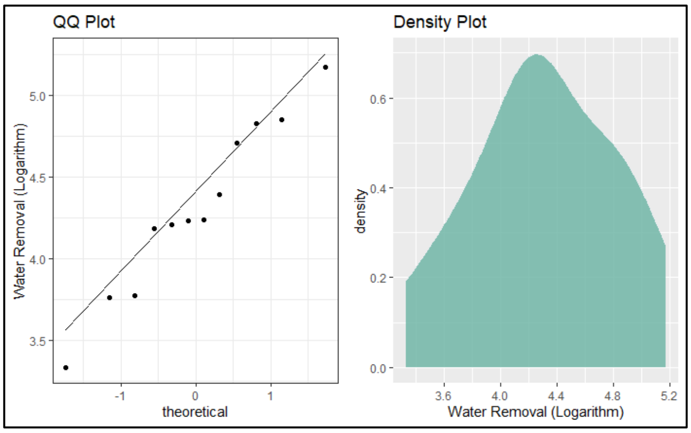


Figure 1: SRS normality evaluation plots for year 2000 in objective 1.

Data subset for objective 2.
Considerable data processing was conducted to generate a test samples for this objective. First the difference was calculated between wood removal measurement between 2000 and 2010. Second, the quantity of the positive values and the quantity of the negative values were counted for each HDI category. This resulted in two separate count lists that were then combined, and row names labelled. An example of this is figure 2.

	increase_count	decrease_count
hdi-1	99	36
hdi-2	890	480
hdi-3	922	627

Figure 2:Counts of all positive and negitive values from the differences in wood removal between 2000 and 2010

Data subset for objective 3.
A simple random sample of 100 cases was extracted from the main data frame using the filter() and sample_() functions from the dplyr R package. New logs were created for the logarithmic transforms for both wood removal and forest land area. Both new logs were evaluated in QQ plots and Density plots.

Table 2: Example first 6 rows of Ojb-3 SRS using head(y2k)

hdicat	forestlandarea	wood	logofwood	logofforestlandarea
3	2950000	7790000	15.86835	14.89731573
2	872000	13500	9.510445	13.6785447
2	51900000	23100000	16.95534	17.76482935
2	109000	2200000	14.60397	11.59910316
3	2190000	6580000	15.69955	14.5994121
3	172000	20600	9.933046	12.05524976

Table 2 shows the first 6 rows of the SRS. Figure 3 displayed is the first of these plots. The sample data frame was named y2k for reference when referring to R code appendix. Image displayed may not be exactly reproducible due to the random sampling of the data that occurs every time the sample extraction code is run.

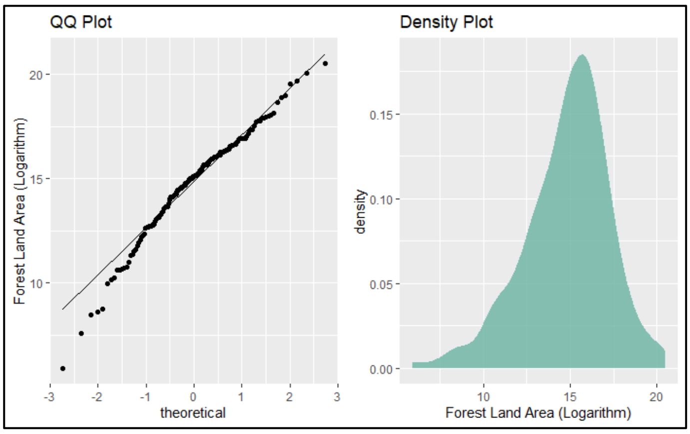


Figure 3: 1st Pair of normality evaluation plots for objective 3 data subset

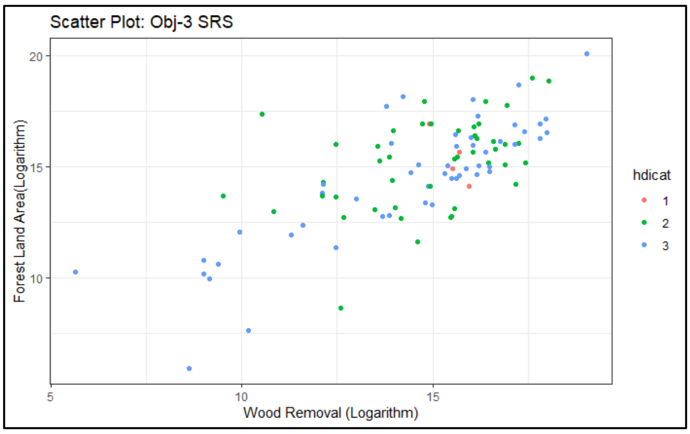


Figure 4: Exploratory Data Analysis (EDA) Scatter Plot for objective-3

Methods

Objective 1
As described in the data section of the report two simple random samples were taken. The two independent samples t-test was determined to be suitable for meeting the objective to examine if there was a statistically significant increase in the mean water municipal water withdrawal(m3/person) from two simple random samples of HDI category 3 countries from the year 2000 and 2010.

Test Type	**Independent t-test**
Null Hypothesis (Ho)	Difference in means is 0
Alternative Hypothesis (HA)	Difference in means is greater than 0
Alpha	0.05

Assumptions:
1. The data are simple random samples.
2. Water removed from the municipal grid is a representative sample of the entire population of countries of the same human development index.
3. Observations come from a population that is normally distributed. The Q-Q plot of the sample data (Figure 1), indicates that this is a reasonable assumption.

The t.test function with the following arguments was used to produce t-value and p-value test results:
conf.level = 0.95
var.equal=TRUE
alternative="greater"

Objective 2
Examine if there was a statistically significant relationship between HDI and the removal of wood from a county's land between the year 2000 and 2010.

Test Type	Chi squared Test of independence
Null Hypothesis (Ho)	HDI does not influence removal of wood
Alternative Hypothesis (HA)	HDI does influence removal of wood
Alpha	0.05

Assumptions:
1. All counts are greater than 5, and none less than 1.
2. The data represents actual counts.
3. Observations occur in 1 and only 1 of several distinct categories.

The chi-sqared test for independence was performed with the chisq.test function in R with the data outlined in the previous section and the default arguments.

Objective 3
Determine if there was a statistically significant relationship between the forest area a country had and the amount of wood that it removed during the calendar year 2000.

To investigate this relationship, it was decided that a correlation test and a regression analysis would be proformed.

Analysis	Linear Regression Analysis
Test 1 Type	Pearson Correlation
Test 2 Type	Durbin-Watson Test
Alpha	0.05

Following the production and examination of the EDA scatterplot (Figure 4) a simple Pearson correlation coefficient was produced using the base R function cor with the Pearson method selected because no significant outliers had been identified.

With the Pearson correlation coefficient being reasonable value to allow regression analysis to proceed the cor.test function was run with method = "pearson" and alternative = "greater" arguments in order to inspect the p-value.

The fit linear model function lm was used with default arguments to produce a simple linear model for logofforestlandarea ~ logofwood.

Predicted values and residual values were extracted from the model and saved into the sample data frame using the predict() and residuals() functions. Upper and lower prediction interval were also produced from the model using the predict() function with interval = “prediction” augment.

Results and Discussion

Objective 1 tested two independent and random samples of countries from the year 2000 and 2010 with the the t-test function in R. The results of this test are shown below and visualised in Figure 5.

```
        Two Sample t-test

data:  rs2000$logofwaterusage and
rs2010$logofwaterusage
t = -0.90117, df = 22, p-value = 0.8114
alternative hypothesis: true difference in means is
greater than 0
95 percent confidence interval:
 -0.661137       Inf
sample estimates:
mean of x mean of y
 4.449563  4.677114
```

Interpretation: The difference in means was not within the statistical significance condition set. P-value = 0.05 therefore the author rejects the alternative hypothesis. Evidence does not exist to suggest that removal of municipal water (water consumption) increased amongst HDI category 3 countries between the year 2000 and 2010.

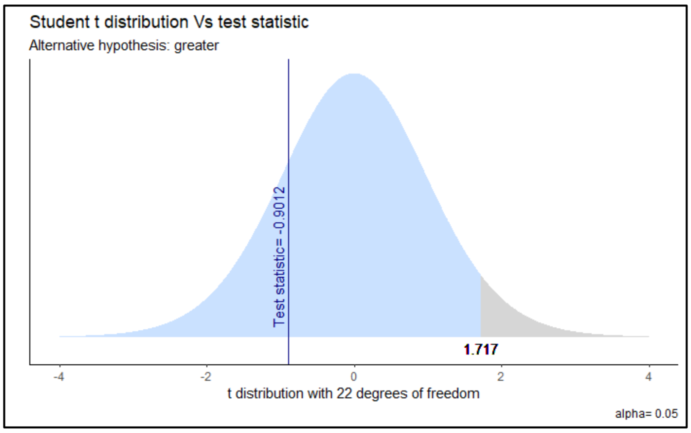


Figure 5:Visualisation of two sample t-test result. The alternative hypothesis was rejected.

Objective 2 tested if there was a statistically significant relationship between HDI category and the removal of wood from a county's land between the year 2000 and 2010. The results of the $\chi 2$ test of independence are shown below:

```
        Pearson's Chi-squared test

data:  hdi_counts
X-squared = 16.173, df = 2, p-value = 0.0003076
```

Table 3: Expected values from Chi-squared test function.

	increase_count	decrease_count
hdi-1	84.47446	50.52554
hdi-2	857.25933	512.74067
hdi-3	969.26621	579.73379

Interpretation: There is strong evidence to support a relationship between Human Development Index (HDI) and the removal of wood.
The data suggests that countries with a higher HDI will removal more wood over time rather than less.

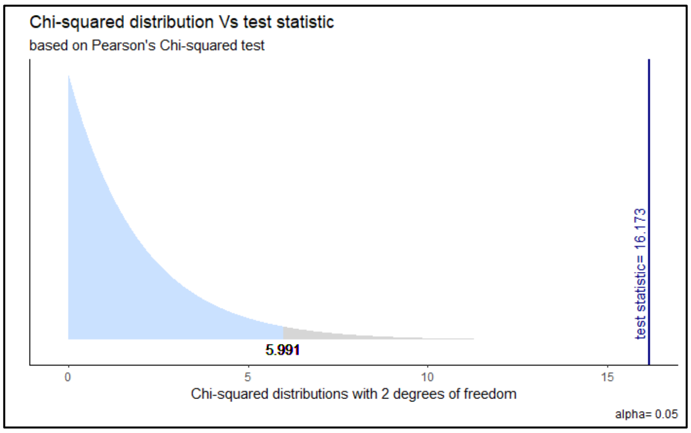


Figure 6: Visualisation of Chi-squared test for independence. The alternative hypothesis was accepted.

Objective 3 conducted a correlation test and a regression analysis to assess if there was statistically significant relationship between the forest area a country had and the amount of wood that it removed during the calendar year 2000.

```
        Pearson's product-moment correlation

data:  y2k$logofwood and y2k$logofforestlandarea
t = 12.151, df = 142, p-value < 2.2e-16
alternative hypothesis: true correlation is greater
than 0
95 percent confidence interval:
 0.6391188 1.0000000
sample estimates:
      cor
0.7139548
```

Correlation test indicated reasonable relationship with p-value below alpha and correlation coefficient of 0.71.

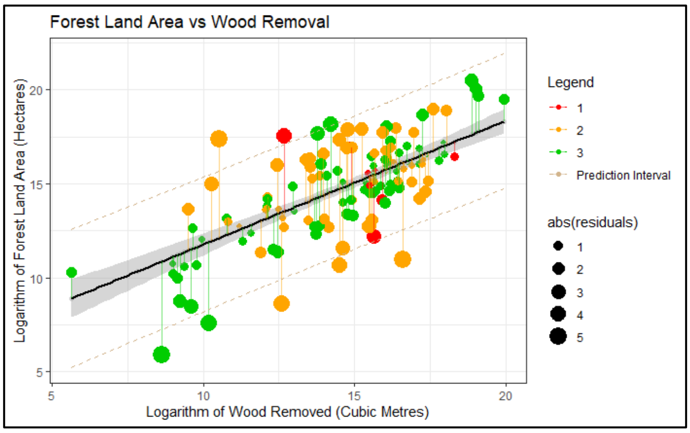


Figure 7: Regression analysis on the relationship of forest land area and wood removal.

Summary of linear model:

```
Call:
lm(formula = logofforestlandarea ~ logofwood, data = y2k)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1995 -0.9444 -0.1079  1.0983  5.4394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.57696    0.85436   5.357 3.33e-07 ***
logofwood    0.69873    0.05751  12.151  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.783 on 142 degrees of freedom
Multiple R-squared:  0.5097,    Adjusted R-squared:
0.5063
F-statistic: 147.6 on 1 and 142 DF,  p-value: < 2.2e-16
```

The linear regression analysis shown in figure 7 visualises the convincing relationship between the about of forest land a country has and the amount of wood that is removed from the country. The significance of the slope p-value of < 2.2e-16 and the R-squared value of 0.51 is evidence in favour of regression model being accepted. The following figures 8 to 10 show checks made to the assumptions of residuals.

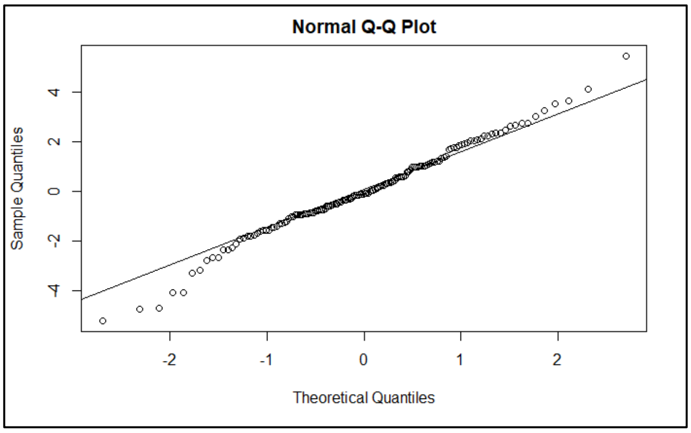


Figure 8: Q-Q Plot to check normality assumption. The residuals appear normally distributed.

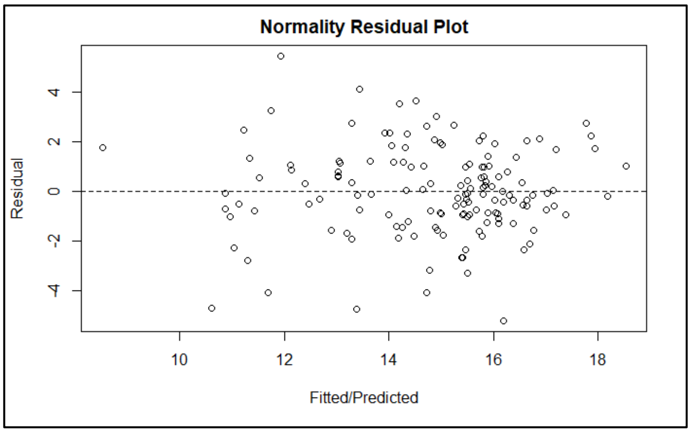


Figure 9: Plot of residuals to check assumption of variance normality. Good +/- spread is observed.

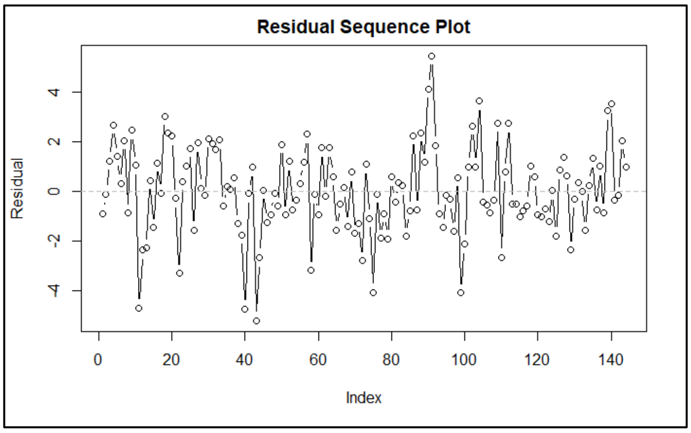


Figure 10: Plot of residuals in sequence to ensure approximately independent

```
        Durbin-Watson test

data:  fit
DW = 1.5708, p-value = 0.004772
alternative hypothesis: true autocorrelation is greater than 0
```

The null hypothesis is that there is no autocorrelation. Since this p-value is less than 0.05, we can reject the null hypothesis and conclude that the residuals in this regression model are autocorrelated (that there is some association among the residuals). This is concerning and does bring into question the validity of the regression analysis. Perhaps the log transformation was not the best choice in this case. In future the author would experiment with different transformations before proceeding to reporting.

Concluding Remarks

This report successfully proposed, analysed and answered the three objectives outlined in the introduction with minimal uncertainty.
Among category 3 countries it was presented that there was no statistically significant increase in mean water consumption over a ten-year time period. There is a statistically significant relationship between the human development index of a country and the removal of wood from a county's land between the year 2000 and 2010.
There is a statistically significant relationship between the forest area a country had and the amount of wood that it removed during the calendar year 2000. A limitation to the above tests was that only one or two groups of data were taken from different points in time. On reflection more thorough tests would have taken more sample points in the time domain. Another note worthy limitation was that the author was constrained in where data could be sourced. If gap minder had “gaps” in the data, the author was unable to look elsewhere.

References

JCU Online. Learn.jcu.edu.au. (2020). Retrieved 19 August 2020, from https://learn.jcu.edu.au/ultra/stream.

S. Mangiafico, S. (2020). R Handbook. Rcompanion.org. Retrieved 15 August 2020, from https://rcompanion.org/handbook/.

Verzani, J. (2015). Using R for Introductory Statistics, Second Edition. CRC Press.

Grolemund, G. & Wickham, H. (2017). R for Data Science

Wickham, H (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1

Wickham, H (2019). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.3.0

Bratsas, C (2020). gginference: Visualise the Results of Inferential Statistics using 'ggplot2'. R package version 0.1.1

Hothorn, T (2019). lmtest: Testing Linear Regression Models. R package version 0.9-37

CODE APPENDIX FOLLOWS…

Assessment-3

Adam Mills

15/08/2020

Data Import

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyv
erse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_c
onflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
# Import Data files
hdi <- read_csv("hdi_human_development_index.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   country = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
waterwithdrawal <- read_csv("municipal_water_withdrawal_cu_meters_per_person.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   country = col_character(),
##   `1966` = col_logical(),
##   `1967` = col_logical(),
##   `1968` = col_logical(),
##   `1969` = col_logical(),
##   `1971` = col_logical(),
##   `1972` = col_logical(),
##   `1973` = col_logical(),
##   `1976` = col_logical(),
##   `1977` = col_logical(),
##   `1978` = col_logical(),
##   `1979` = col_logical(),
##   `1981` = col_logical(),
##   `1983` = col_logical()
## )
## See spec(...) for full column specifications.
```

```
wood <- read_csv("wood_removal_cubic_meters.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   country = col_character()
## )
## See spec(...) for full column specifications.
```

```
forest_land <- read_csv("forest_land_total_area_ha.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   country = col_character()
## )
## See spec(...) for full column specifications.
```

Data Processing

Combine data sets

```
hdi_water <- left_join(tidy_waterwithdrawl, tidy_hdi)
```

```
## Joining, by = c("country", "year")
```

```
hdi_water_wood <- left_join(hdi_water, tidy_wood)
```

```
## Joining, by = c("country", "year")
```

```
main <- left_join(hdi_water_wood, tidy_forestland)
```

```
## Joining, by = c("country", "year")
```

Create Catagorical Variable from numeric HDI

```
# Create human development index catagorical variable
main$hdicat[main$hdi <= 0.32] = 1
```

```
## Warning: Unknown or uninitialised column: `hdicat`.
```

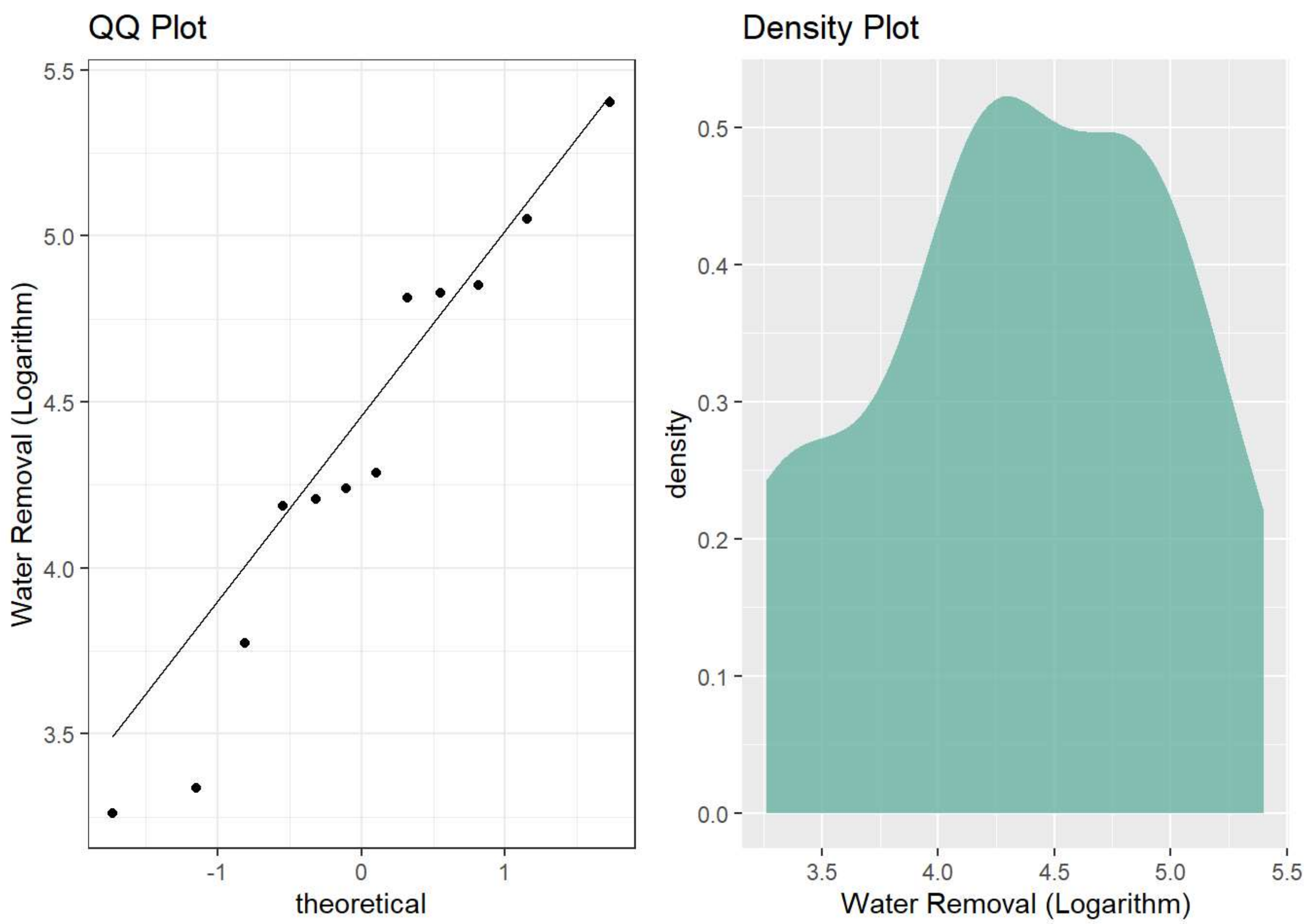
```
main$hdicat[main$hdi >= 0.33 & main$hdi < 0.66] = 2
main$hdicat[main$hdi >= 0.66 & main$hdi <= 1] = 3

# Convert to factor
main$hdicat <- as.factor(main$hdicat)

# Output numerical summery by HDI group
describeBy(main, group = main$hdicat)
```

```
##
##  Descriptive statistics by group
## group: 1
##                 vars   n       mean          sd   median    trimmed        mad
## country*           1 133       6.56        3.32 7.00e+00       6.57       4.45
## year*              2 133       7.93        4.83 7.00e+00       7.66       5.93
## waterusage         3  11      11.44        6.80 9.62e+00      11.16       6.79
## hdi                4 133       0.28        0.03 2.80e-01       0.28       0.03
## wood               5 124 9831250.00 16369464.26 6.08e+06 6056400.00 2920722.00
## forestlandarea     6 133 9320451.13 12706062.31 4.40e+06 6450018.69 4610886.00
## hdicat*            7 133       1.00        0.00 1.00e+00       1.00       0.00
##                      min      max      range  skew kurtosis         se
## country*        1.00e+00 1.20e+01 1.1000e+01 -0.13    -1.31       0.29
## year*           1.00e+00 2.00e+01 1.9000e+01  0.38    -0.81       0.42
## waterusage      3.73e+00 2.16e+01 1.7870e+01  0.44    -1.58       2.05
## hdi             1.90e-01 3.20e-01 1.3000e-01 -0.72    -0.48       0.00
## wood            2.24e+05 9.45e+07 9.4276e+07  4.21    17.84 1470021.29
## forestlandarea  1.81e+05 4.34e+07 4.3219e+07  1.66     1.54 1101755.33
## hdicat*         1.00e+00 1.00e+00 0.0000e+00   NaN      NaN       0.00
## ------------------------------------------------------------
## group: 2
##                 vars    n        mean          sd    median    trimmed
## country*           1 1779       55.75       29.83 5.600e+01      56.17
## year*              2 1779       13.32        7.46 1.300e+01      13.30
## waterusage         3  181       39.75       37.03 2.790e+01      33.67
## hdi                4 1779        0.51        0.10 5.100e-01       0.51
## wood               5 1376 16319096.44 44960525.14 5.495e+06 7448818.51
## forestlandarea     6 1765 18830978.39 44679214.16 5.010e+06 9705610.76
## hdicat*            7 1779        2.00        0.00 2.000e+00       2.00
##                       mad      min      max         range  skew kurtosis
## country*            38.55     1.00 1.05e+02        104.00 -0.09    -1.14
## year*                8.90     1.00 2.60e+01         25.00  0.01    -1.19
## waterusage          24.61     3.50 2.76e+02        272.50  2.53     9.96
## hdi                  0.13     0.33 6.60e-01          0.33 -0.10    -1.26
## wood           7134271.20 12400.00 4.35e+08 434987600.00  6.04    40.37
## forestlandarea 6895572.60  1000.00 5.47e+08 546999000.00  7.23    73.02
## hdicat*              0.00     2.00 2.00e+00          0.00   NaN      NaN
##                        se
## country*             0.71
## year*                0.18
## waterusage           2.75
## hdi                  0.00
## wood           1212054.53
## forestlandarea 1063489.41
## hdicat*              0.00
## ------------------------------------------------------------
## group: 3
##                 vars    n        mean           sd    median    trimmed
## country*           1 2202       56.80        32.13 5.600e+01      56.76
## year*              2 2202       15.25         7.30 1.600e+01      15.56
## waterusage         3  281      110.49        58.60 9.230e+01     102.77
## hdi                4 2202        0.78         0.07 7.700e-01       0.78
## wood               5 1582 21951790.42  62705196.30 4.195e+06 7478889.18
## forestlandarea     6 2157 30838227.64 108757053.46 2.240e+06 5733939.84
## hdicat*            7 2202        3.00         0.00 3.000e+00       3.00
##                       mad  min      max     range  skew kurtosis         se
## country*            41.51 1.00 1.13e+02 1.120e+02  0.01    -1.19       0.68
```

```
## year*                 8.90  1.00 2.60e+01 2.500e+01 -0.30    -1.06       0.16
## waterusage           41.51 25.30 3.74e+02 3.487e+02  1.57     3.23       3.50
## hdi                   0.09  0.66 9.50e-01 2.900e-01  0.25    -1.04       0.00
## wood            6192301.29  0.00 5.09e+08 5.090e+08  5.20    30.29 1576522.93
## forestlandarea 3260682.18  0.00 8.15e+08 8.150e+08  5.36    31.13 2341705.95
## hdicat*               0.00  3.00 3.00e+00 0.000e+00   NaN      NaN       0.00
```

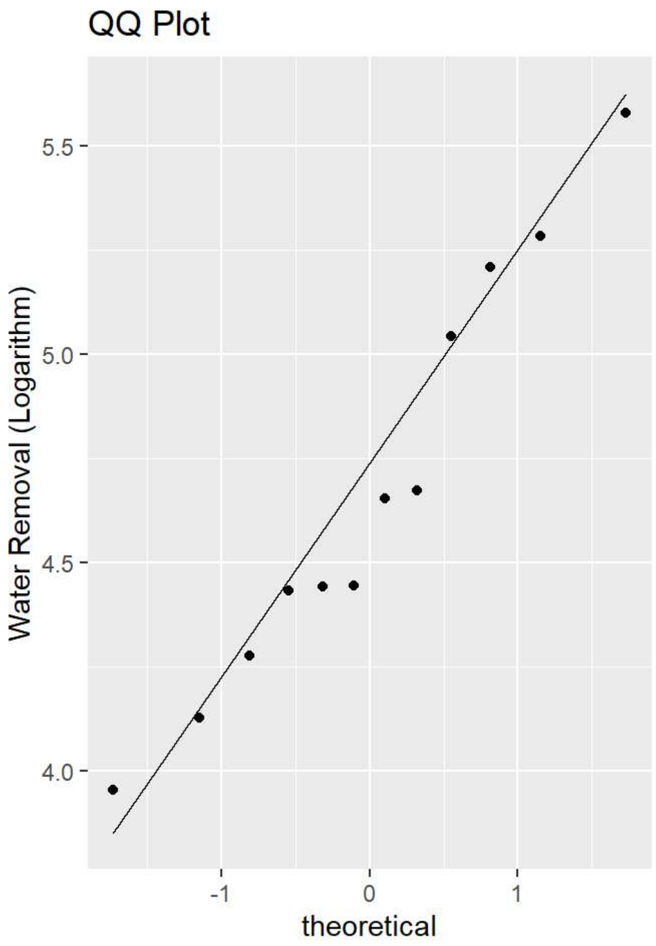
Random sample of non-null hdi cat 3 from 2000

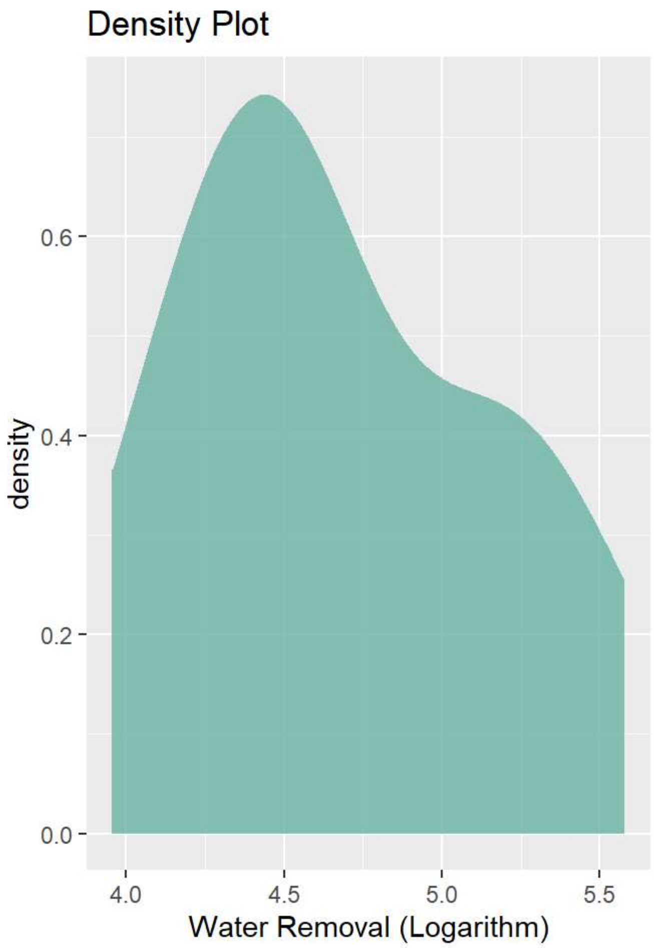
```
# Generate Random stample of 12 from year 2000 where HDI is 3. Generate log transform of wate
r usage.
rs2000 <- filter(main, hdicat == 3 & !is.na(waterusage) & year == 2000) %>%
  sample_n(12) %>%
  select(waterusage) %>%
  mutate(logofwaterusage = log(waterusage))

# Plot qq plot - assess normality
aplot <- ggplot(data = rs2000,aes(sample=logofwaterusage)) +
  stat_qq() + stat_qq_line() +
  theme_bw() +
  ggtitle("QQ Plot") +
  labs(y="Water Removal (Logarithm)")

# Plot density of data - assess normality
bplot <- ggplot(data = rs2000, aes(x=logofwaterusage)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) +
  ggtitle("Density Plot")+ labs(x="Water Removal (Logarithm)")

grid.arrange(aplot, bplot, ncol=2)
```



Random sample of non-null hdi cat 3 from 2010

```
# Generate Random stample of 12 from year 2010 where HDI is 3. Generate log transform of wate
r usage.
rs2010 <- filter(main, hdicat == 3 & !is.na(waterusage) & year == 2010) %>%
  sample_n(12) %>%
  select(waterusage) %>%
  mutate(logofwaterusage = log(waterusage))
# Plot qq plot - assess normality
aplot <- ggplot(data = rs2010,aes(sample=logofwaterusage)) +
  stat_qq() + stat_qq_line() +
  ggtitle("QQ Plot") +
  labs(y="Water Removal (Logarithm)")
# Plot density of data - assess normality
bplot <- ggplot(data = rs2010, aes(x=logofwaterusage)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+
  ggtitle("Density Plot")+ labs(x="Water Removal (Logarithm)")

grid.arrange(aplot, bplot, ncol=2)
```





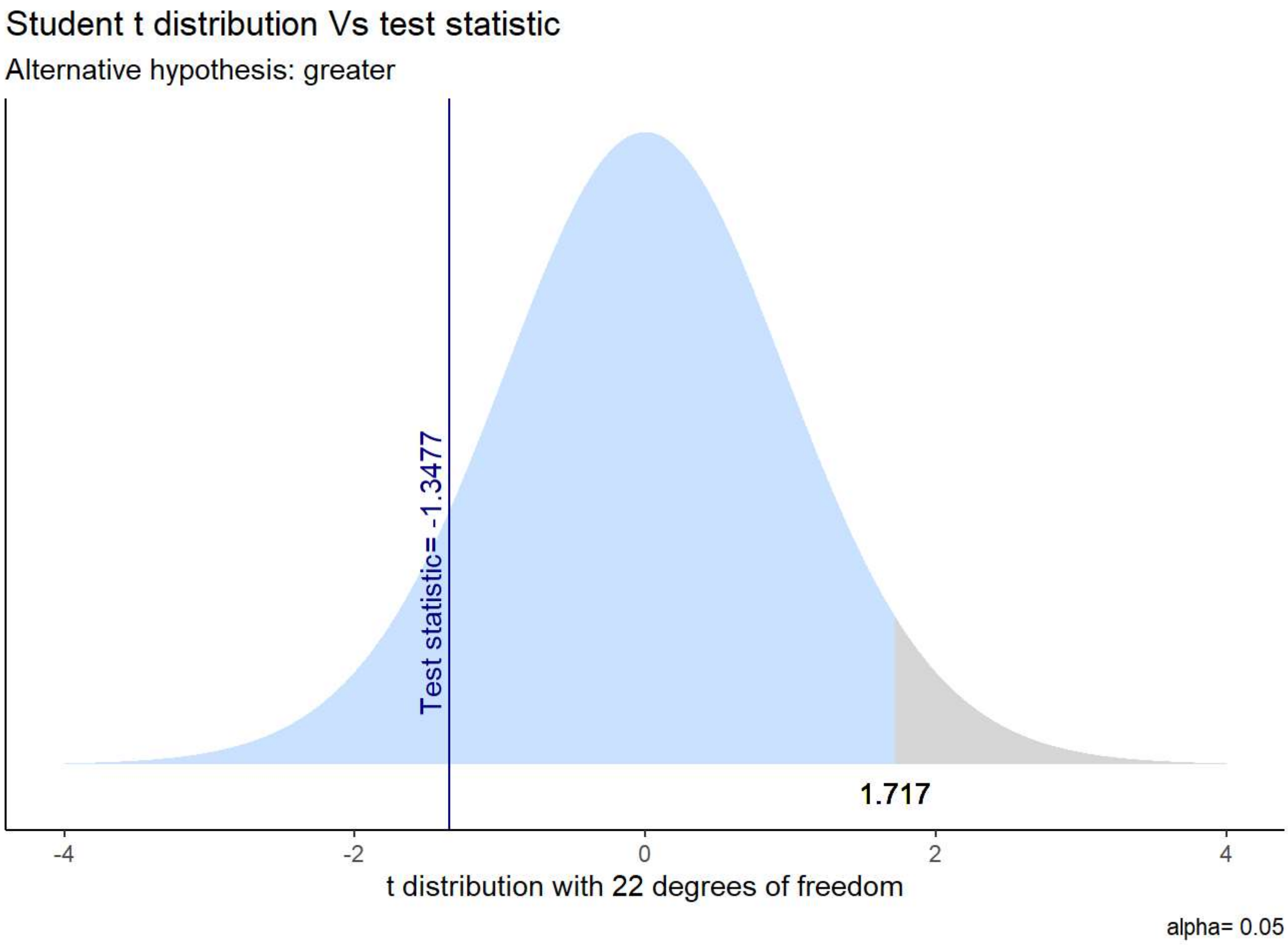
Perform independent t-test Objective: Examine if there was a statistically significant increase in the mean water municipal water withdrawal(m3/person) from a simple random sample of HDI category 3 countries between the year 2000 and 2010.

Ho: Difference in means is 0 HA: Difference in means is greater than 0 Alpha of 0.05

```
df <- t.test(rs2000$logofwaterusage, rs2010$logofwaterusage, conf.level = 0.95, var.equal=TRU
E, alternative="greater")
df
```

```
##
##  Two Sample t-test
##
## data:  rs2000$logofwaterusage and rs2010$logofwaterusage
## t = -1.3477, df = 22, p-value = 0.9043
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.7379132        Inf
## sample estimates:
## mean of x mean of y
##  4.352639  4.677114
```

```
library(gginference)
ggttest(df)
```



Interpretation The difference in means was not within the statistical significance condition set. P-vaule = 0.05 therefore the author rejects the alternative hypothesis.

```
# calculate the difference in wood removal between 2010 and 2000 (negative values indicate a
 decrease in wood removal)
wood$diff <- wood$"2010" - wood$"2000"

# Merge new wood dataframe with data that contains hdi categories
tidy_wood2 <- wood %>% pivot_longer(cols = 2:23, names_to = "year", values_to = "wood")
main2 <- left_join(main, tidy_wood2)
```

```
## Joining, by = c("country", "year", "wood")
```

```r
# Count all the distint postive & negative difference values from category 1
decrease_count <- c(n_distinct(main2[main2$diff < 0 & main2$hdicat == "1",]))
increase_count <- c(n_distinct(main2[main2$diff > 0 & main2$hdicat == "1",]))

# Count all the distint postive & negative difference values from category 2
decrease_count <- c(decrease_count,n_distinct(main2[main2$diff < 0 & main2$hdicat == "2",]))
increase_count <- c(increase_count,n_distinct(main2[main2$diff > 0 & main2$hdicat == "2",]))

# Count all the distint postive & negative difference values from category 3
decrease_count <- c(decrease_count,n_distinct(main2[main2$diff < 0 & main2$hdicat == "3",]))
increase_count <- c(increase_count,n_distinct(main2[main2$diff > 0 & main2$hdicat == "3",]))

# Bind the two lists together
hdi_counts = cbind(increase_count, decrease_count)

# Name the rows
rownames(hdi_counts) = c("hdi-1","hdi-2","hdi-3")

# Check output
hdi_counts
```

```
##       increase_count decrease_count
## hdi-1             99             36
## hdi-2            890            480
## hdi-3            922            627
```

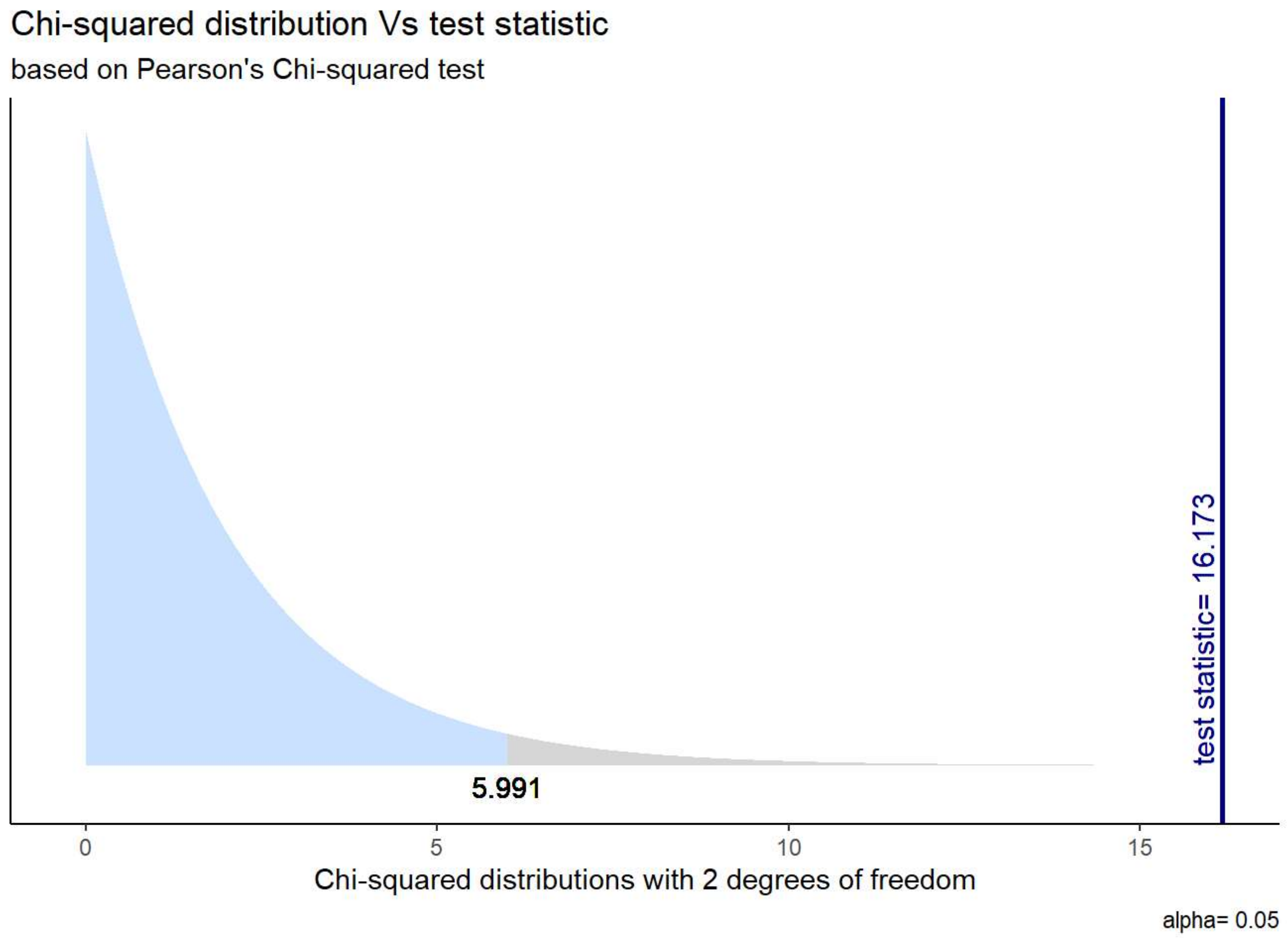
Objective: Examine if there was a statistically significant relationship between HDI and the removal of wood from a county’s land between the year 2000 and 2010.

The null hypothesis is that Human Development Index (HDI) does not influence the removal of wood. The alternative hypothesis is that Human Development Index (HDI) influences the amount of wood removal for a country.

```r
chitest = chisq.test(hdi_counts)
chitest
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  hdi_counts
## X-squared = 16.173, df = 2, p-value = 0.0003076
```

```r
ggchisqtest(chitest)
```




```
hdi_counts
```

```
##       increase_count decrease_count
## hdi-1             99             36
## hdi-2            890            480
## hdi-3            922            627
```

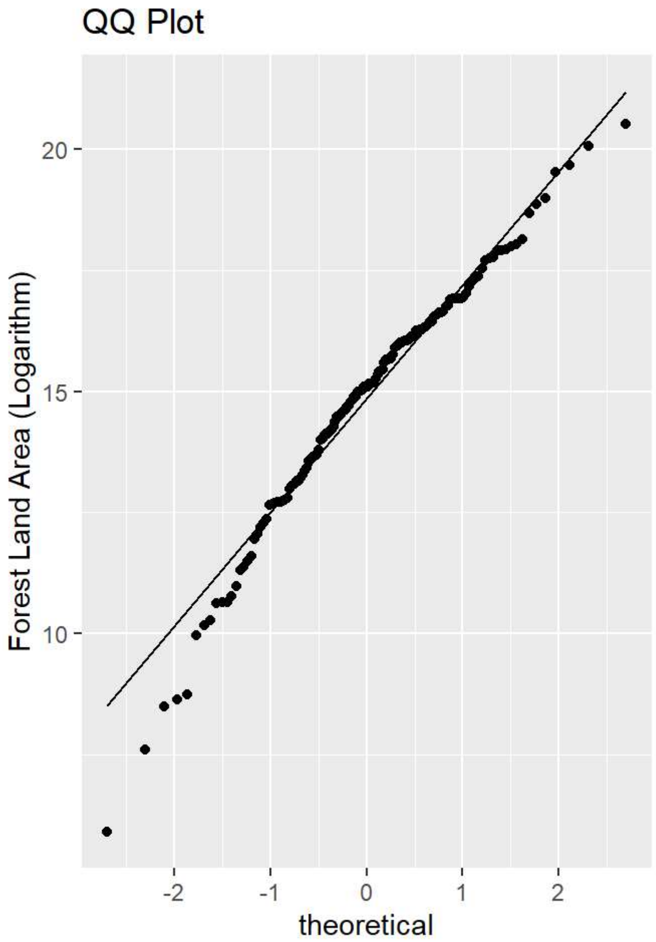
```
hdi_counts_expected <- chitest$expected
```

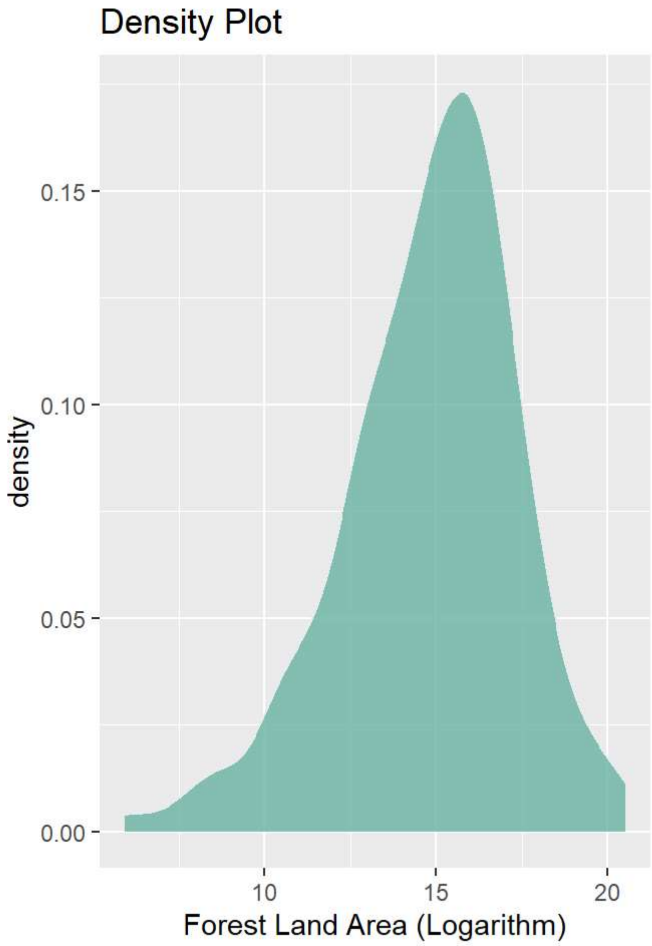
There is strong evidence to support a relationship between Human Development Index (HDI) and the removal of wood. The data suggests that countries with higher a higher HDI will removal more wood over time rather than less.

```
y2k <- filter(main, year==2000 & !is.na(forestlandarea)
              & forestlandarea>0
              & !is.na(wood)
              & wood>0 & !is.na(hdicat)) %>%
        select(hdicat, forestlandarea, wood)

y2k$logofwood <- log(y2k$wood)
y2k$logofforestlandarea <- log(y2k$forestlandarea)

head(y2k)
```

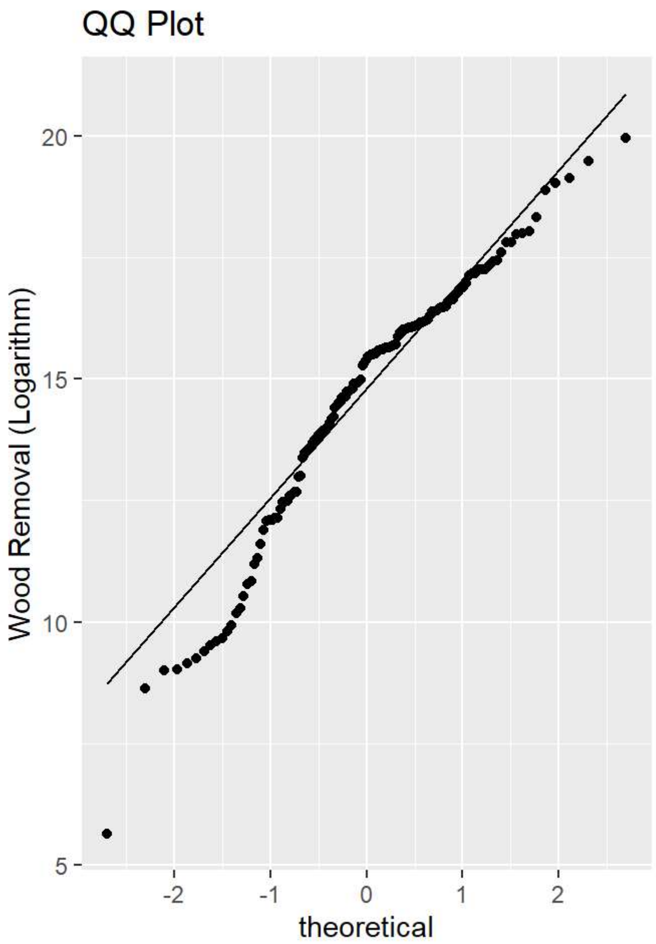
```
## # A tibble: 6 x 5
##   hdicat forestlandarea     wood logofwood logofforestlandarea
##   <fct>           <dbl>    <dbl>     <dbl>               <dbl>
## 1 2             1350000  3040000      14.9                14.1
## 2 3              769000   447000      13.0                13.6
## 3 2             1580000   186000      12.1                14.3
## 4 2            59700000  4260000      15.3                17.9
## 5 3            31900000 10700000      16.2                17.3
## 6 2              333000    72200      11.2                12.7
```

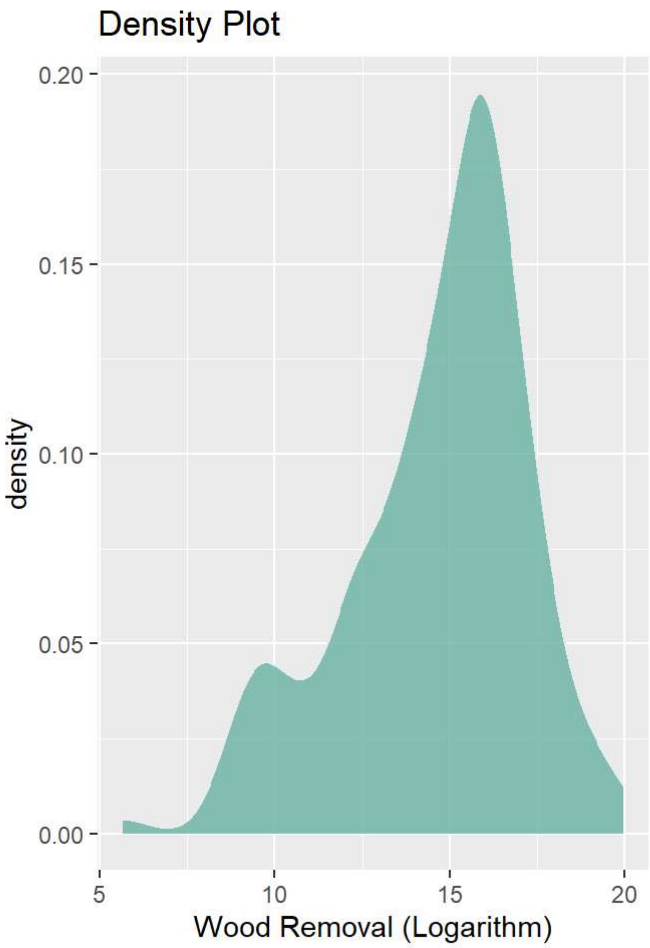
```
write.csv(head(y2k),"y2khead.csv")

# Plot qq plot - assess normality
aplot <- ggplot(data = y2k,aes(sample=logofforestlandarea)) +
  stat_qq() + stat_qq_line() +
  ggtitle("QQ Plot") +
  labs(y="Forest Land Area (Logarithm)")
# Plot density of data - assess normality
bplot <- ggplot(data = y2k, aes(x=logofforestlandarea)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+
  ggtitle("Density Plot")+ labs(x="Forest Land Area (Logarithm)")
grid.arrange(aplot, bplot, ncol=2)
```



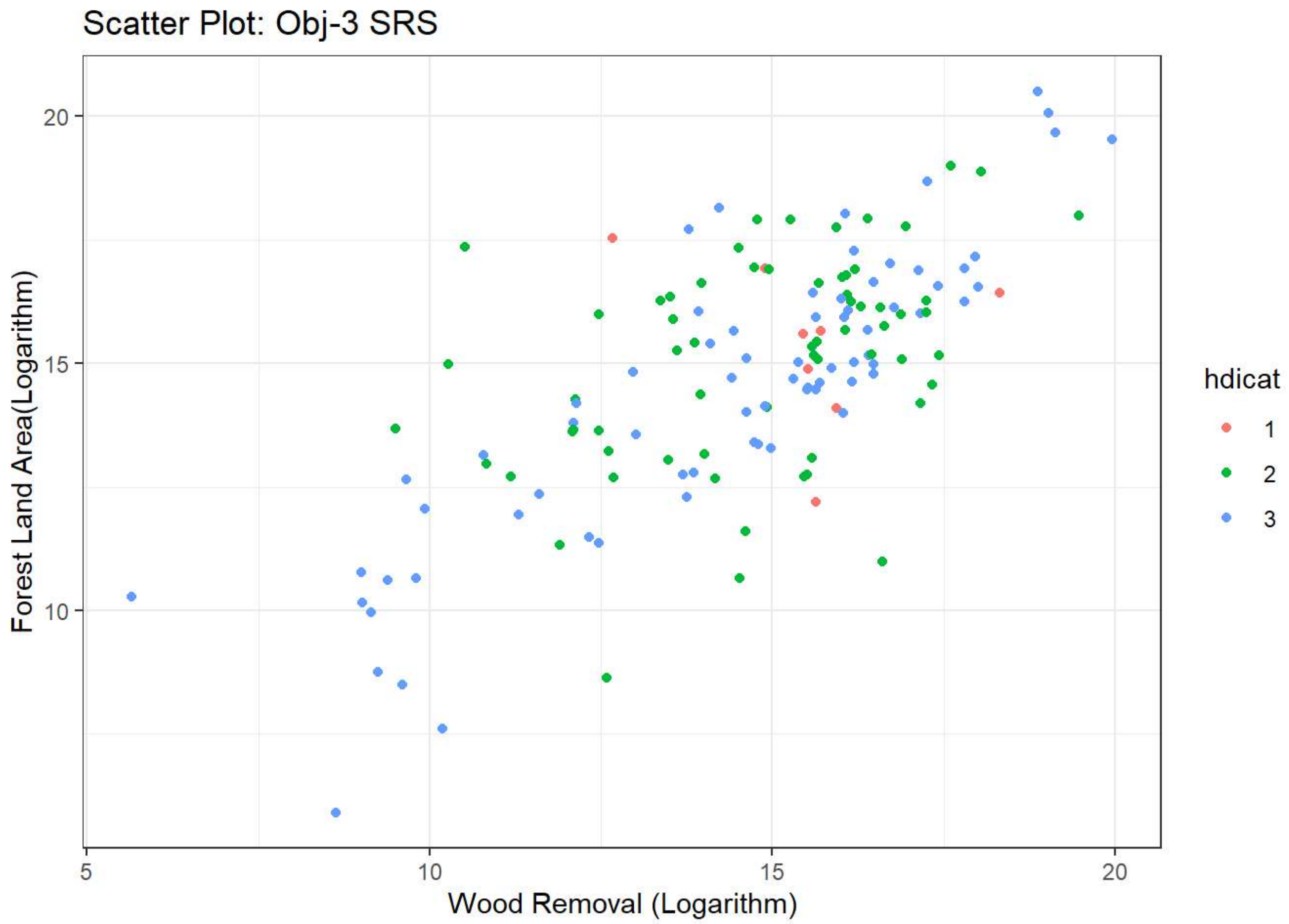


```r
# Plot qq plot - assess normality
aplot <- ggplot(data = y2k,aes(sample=logofwood)) +
  stat_qq() + stat_qq_line() +
  ggtitle("QQ Plot") +
  labs(y="Wood Removal (Logarithm)")
# Plot density of data - assess normality
bplot <- ggplot(data = y2k, aes(x=logofwood)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+
  ggtitle("Density Plot")+ labs(x="Wood Removal (Logarithm)")
grid.arrange(aplot, bplot, ncol=2)
```






```r
y2k %>% select(hdicat, logofforestlandarea, logofwood) %>%
  ggplot(aes(x = logofwood, y = logofforestlandarea, color = hdicat)) +
          geom_point() +
          theme_bw() +
          ggtitle("Scatter Plot: Obj-3 SRS") +
          labs(x="Wood Removal (Logarithm)", y="Forest Land Area(Logarithm)")
```




```
cor(y2k$logofwood, y2k$logofforestlandarea, method = "pearson")
```

```
## [1] 0.7139548
```

```
cor.test(y2k$logofwood, y2k$logofforestlandarea, method = "pearson", alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  y2k$logofwood and y2k$logofforestlandarea
## t = 12.151, df = 142, p-value < 2.2e-16
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.6391188 1.0000000
## sample estimates:
##       cor
## 0.7139548
```

```
fit <- lm(logofforestlandarea ~ logofwood, data = y2k)
summary(fit)
```

```
##
## Call:
## lm(formula = logofforestlandarea ~ logofwood, data = y2k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1995 -0.9444 -0.1079  1.0983  5.4394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.57696    0.85436   5.357 3.33e-07 ***
## logofwood    0.69873    0.05751  12.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.783 on 142 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.5063
## F-statistic: 147.6 on 1 and 142 DF,  p-value: < 2.2e-16
```

```
y2k$predicted <- predict(fit)   # Save the predicted values
y2k$residuals <- residuals(fit) # Save the residual values
pred.int <- predict(fit, interval = "prediction")
```

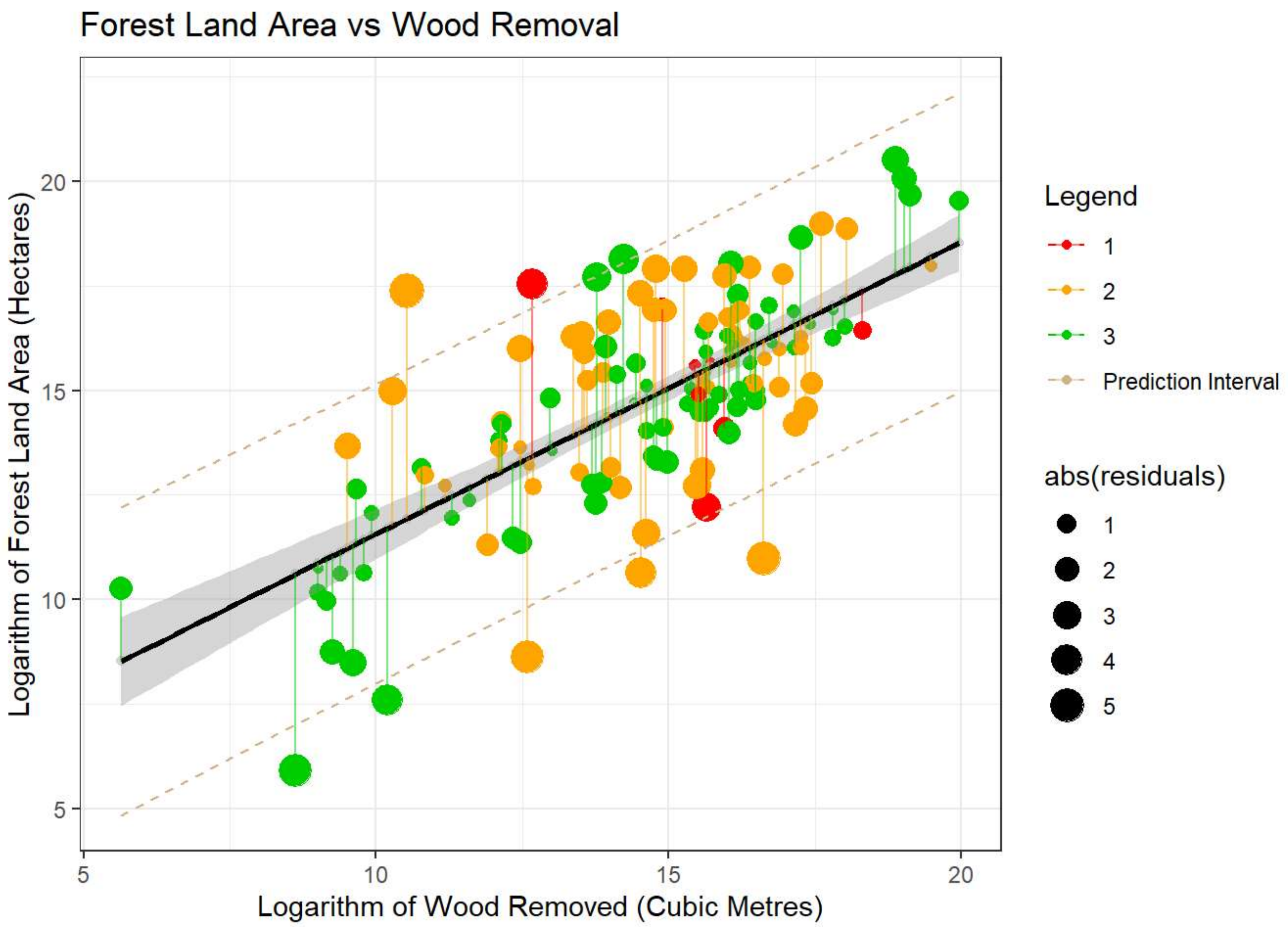
```
## Warning in predict.lm(fit, interval = "prediction"): predictions on current data refer to
_future_ responses
```

```
y2k <- cbind(y2k, pred.int)
y2k %>% select(logofforestlandarea,logofwood, predicted, residuals, lwr, fit, upr) %>% head()
```

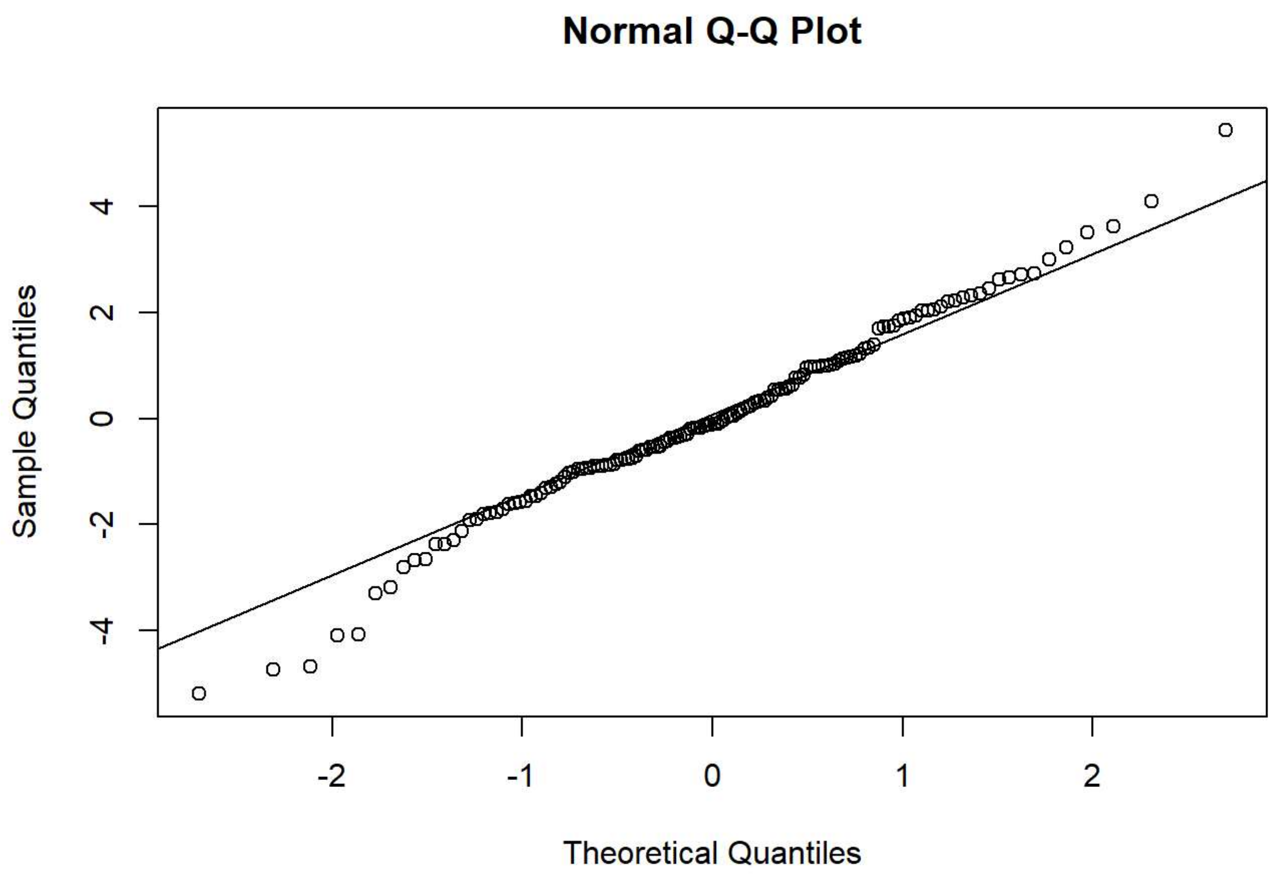
```
##   logofforestlandarea logofwood predicted  residuals       lwr      fit
## 1            14.11562  14.92737  15.00717 -0.8915545 11.469275 15.00717
## 2            13.55285  13.01031  13.66767 -0.1148190 10.125141 13.66767
## 3            14.27294  12.13350  13.05501  1.2179255  9.505908 13.05501
## 4            17.90484  15.26478  15.24293  2.6619130 11.704461 15.24293
## 5            17.27812  16.18575  15.88644  1.3916739 12.344294 15.88644
## 6            12.71590  11.18720  12.39380  0.3221013  8.834475 12.39380
##        upr
## 1 18.54506
## 2 17.21019
## 3 16.60411
## 4 18.78140
## 5 19.42859
## 6 15.95312
```

```
colors <- c("1" = "red", "2" = "orange", "3" = "green3", "Prediction Interval" = "tan")

ggplot(data = y2k, aes(x = logofwood, y = logofforestlandarea, color = hdicat)) +
  geom_point(aes(size = abs(residuals))) +
  geom_point(aes(y = predicted), color="lightgrey") +  # Add the predicted values
  geom_smooth(method = "lm",aes(group=1), color = "black") +
  geom_segment(aes(xend = logofwood, yend = predicted), alpha=0.5) +
  geom_line(aes(y = lwr, color = "Prediction Interval"), linetype = "dashed")+
  geom_line(aes(y = upr, color = "Prediction Interval"), linetype = "dashed")+
  theme_bw() +
  ggtitle("Forest Land Area vs Wood Removal") +
  labs(y="Logarithm of Forest Land Area (Hectares)", x = "Logarithm of Wood Removed (Cubic Me
tres)", color = "Legend") +
  scale_color_manual(values = colors)
```

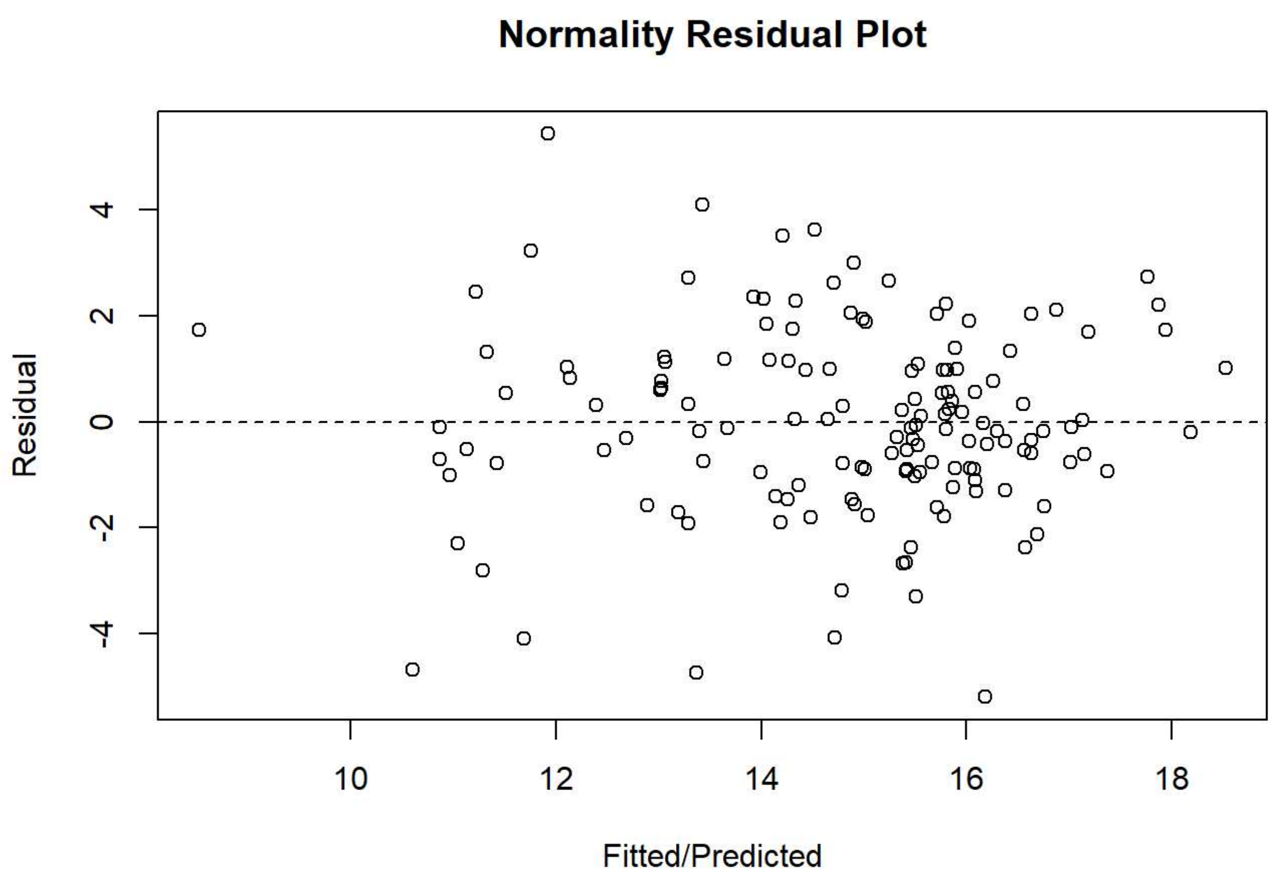
```
## `geom_smooth()` using formula 'y ~ x'
```




```
qqnorm(y2k$residuals)
qqline(y2k$residuals)
```

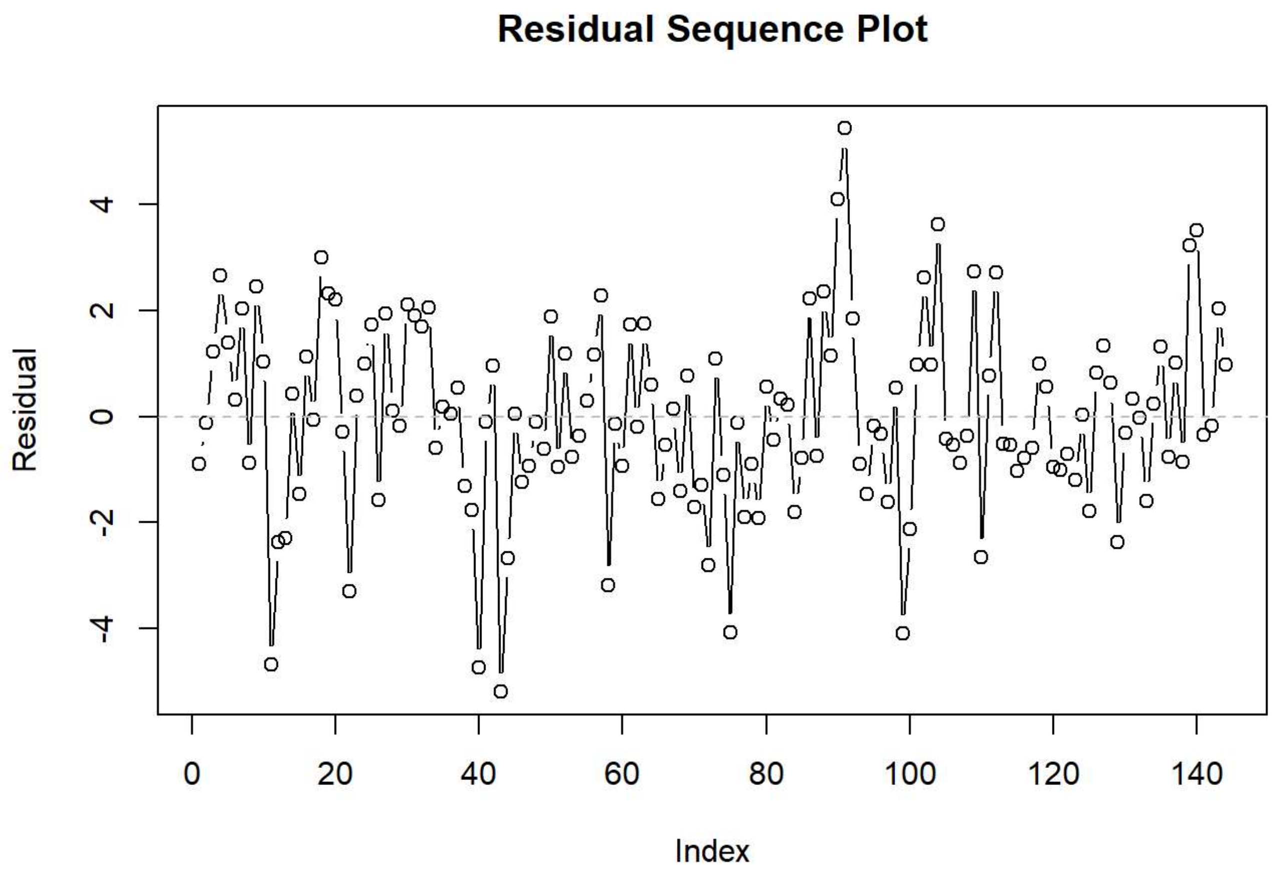



```
plot(y2k$predicted, y2k$residuals, xlab="Fitted/Predicted", ylab="Residual", main="Normality
 Residual Plot") + abline(h=0, lty=2)
```



```
## integer(0)
```

```
plot(y2k$residuals, type='b', ylab="Residual", main = "Residual Sequence Plot") + abline(h=0,
lty=2, col='grey')
```




```
## integer(0)
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
dwtest(fit)
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.5708, p-value = 0.004772
## alternative hypothesis: true autocorrelation is greater than 0
```