Regression Analysis

Akolade Sofiyyah Iwalewa, 22201441

2022-10-24

Introduction to dataset:

The Health Behavior in School-aged Children (HBSC) survey is a WHO collaborative cross-national study that monitors the health behaviors, health outcomes and social environments of school-aged children every four years. The HBSC Ireland team, based at the Health Promotion Research Center, University of Gal way conducted the nationally representative survey of Irish school children in 2006, 2010, 2014 and 2018. This data set contains: • P_Content The percentage of children that reported being happy with they way they are, • Age the Age of the child, • Sex the Sex of the child Male or Female and • Year the information was collected

Objective

The objectives are:

- establish a relationship between P_Content and Year, so I can use it to predict P_Content when only the Year of the information is known.
- fit a simple linear regression model to the data with P_Content as the response variable and Year as a numeric predictor variable for females.

Comment: I first need to carry out exploratory data analysis on the dataset before fitting a model.

```
# Loading the data set into R
df <-read.csv("Assignment.csv", header = TRUE)
```

```
# Taking a look at the data set
head(df)
```

```
##   Year                Age    Sex P_Content
## 1 2006 10 years and under   Male     67.72
## 2 2006 10 years and under Female     69.39
## 3 2010 10 years and under   Male     78.41
## 4 2010 10 years and under Female     71.77
## 5 2014 10 years and under   Male     76.70
## 6 2014 10 years and under Female     77.64
```

```
# Checking the dimension of the data set ie. the number of rows and columns in the data set
dim(df)
```

```
## [1] 64  4
```

comment: There are 64 rows, 4 columns in the data set

```
# Checking the summary of the data set in order to have a better understanding of it.
summary(df)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2006   Length:64          Length:64          Min.   :28.70
##  1st Qu.:2009   Class :character   Class :character   1st Qu.:47.56
##  Median :2012   Mode  :character   Mode  :character   Median :57.28
##  Mean   :2012                                         Mean   :57.15
##  3rd Qu.:2015                                         3rd Qu.:69.75
##  Max.   :2018                                         Max.   :78.41
```

```
# Sub-setting the data set so it contains only the percentage of school aged children that re
ported being happy with they way they are in 2006 with respect to the female gender.
df_2006_female <- df[df$Year == 2006 & df$Sex == 'Female', ]
```

```
# Checking if the data was properly filtered
head(df_2006_female)
```

```
##    Year                Age    Sex P_Content
## 2  2006 10 years and under Female     69.39
## 10 2006           11 years Female     74.69
## 18 2006           12 years Female     73.16
## 26 2006           13 years Female     63.68
## 34 2006           14 years Female     61.91
## 42 2006           15 years Female     57.25
```

```
# Sub-setting the data set so it contains only the percentage of school aged children that re
ported being happy with they way they are in 2006 with respect to the male gender.
df_2006_male <- df[df$Year == 2006 & df$Sex == 'Male', ]
```

```
# Checking if the data was properly filtered
head(df_2006_male)
```

```
##    Year                Age  Sex P_Content
## 1  2006 10 years and under Male     67.72
## 9  2006           11 years Male     74.37
## 17 2006           12 years Male     66.25
## 25 2006           13 years Male     55.66
## 33 2006           14 years Male     48.20
## 41 2006           15 years Male     41.95
```

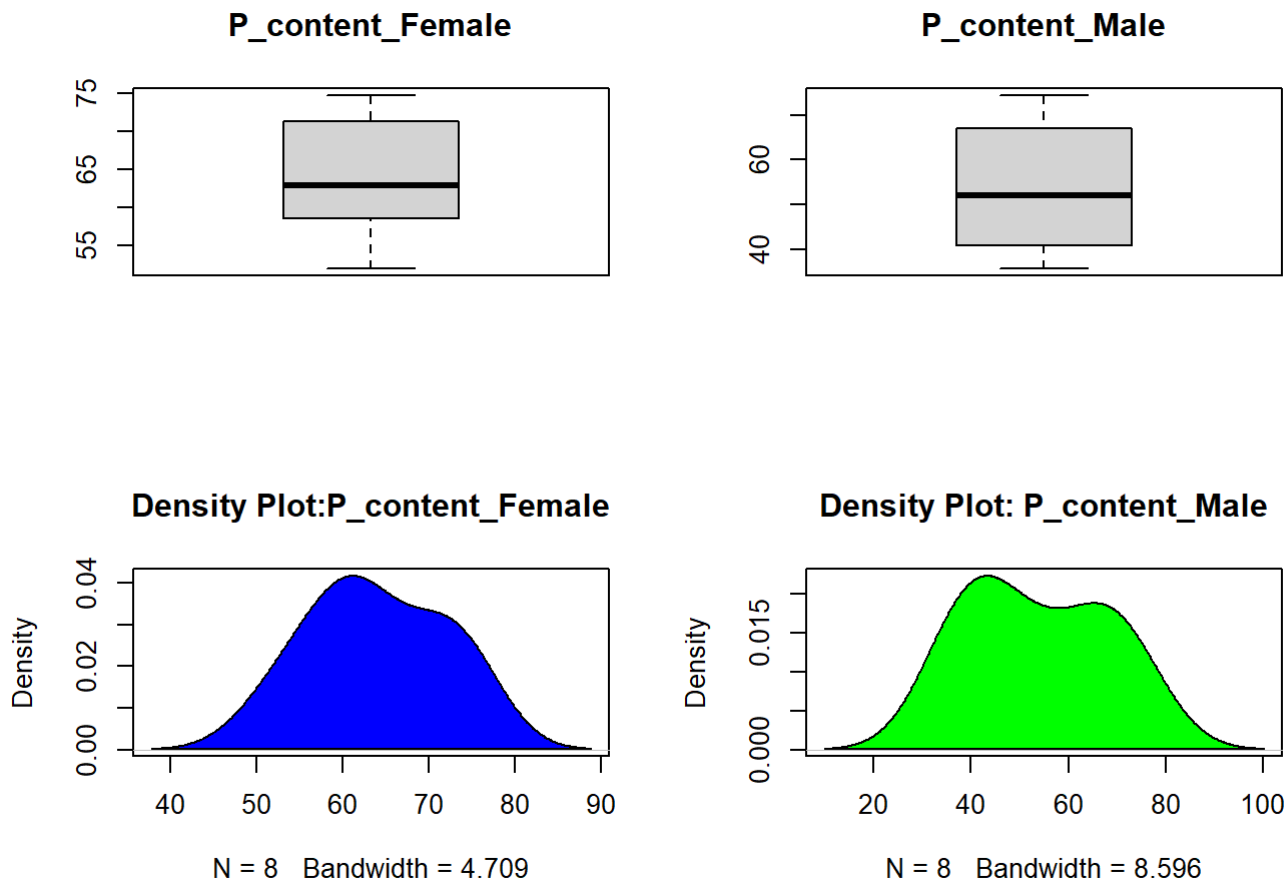
```
# Getting the summary of both filtered data
summary(df_2006_female)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2006   Length:8           Length:8           Min.   :51.93
##  1st Qu.:2006   Class :character   Class :character   1st Qu.:59.19
##  Median :2006   Mode  :character   Mode  :character   Median :62.80
##  Mean   :2006                                         Mean   :63.98
##  3rd Qu.:2006                                         3rd Qu.:70.33
##  Max.   :2006                                         Max.   :74.69
```

```
summary(df_2006_male)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2006   Length:8           Length:8           Min.   :35.71
##  1st Qu.:2006   Class :character   Class :character   1st Qu.:41.44
##  Median :2006   Mode  :character   Mode  :character   Median :51.93
##  Mean   :2006                                         Mean   :53.72
##  3rd Qu.:2006                                         3rd Qu.:66.62
##  Max.   :2006                                         Max.   :74.37
```

```
# Plotting a boxplot and a density plot
par(mfrow=c(2, 2))
boxplot(df_2006_female$P_Content, main="P_content_Female")
boxplot(df_2006_male$P_Content, main="P_content_Male")
plot(density(df_2006_female$P_Content), main="Density Plot:P_content_Female")
polygon(density(df_2006_female$P_Content), col="blue")
plot(density(df_2006_male$P_Content), main="Density Plot: P_content_Male")
polygon(density(df_2006_male$P_Content), col="green")
```



Description of the distributions and the difference in the distributions for the percentage of school aged children that reported being happy with they way they are in 2006 with respect to sex:

- With respect to the female gender:

a. The minimum P_Content for the females as at 2006 is 51.93 while the maximum P_Content is 74.69
b. 25% of the females have their P_Content lower than 59.19
c. 50% of the females have their P_Content lower than 62.80
d. 75% of the females have their P_Content lower than 70.33
e. The average P_Content for females is 63.98
f. The female P_Content is positively skewed

- With respect to the male gender

a. The minimum P_Content for the males as at 2006 is 35.71 while the maximum P_Content is 74.37
b. 25% of the males have their P_Content lower than 41.44
c. 50% of the males have their P_Content lower than 51.93
d. 75% of the males have their P_Content lower than 66.62
e. The average P_Content for males is 53.72
f. The male P_Content is positively skewed

- Difference in distributions for both females and males in 2016

a. The male's P-content are mostly similar compared to the female's P_Content
b. The females maximum P_content is slightly higher than the males maximum P_Content
c. The males minimum P_Content is much lower than the females minimum P_Content, which is implying that the males has the lowest P_Content in 2016.

```
# Sub-setting the data set so it contains only the percentage of female school aged children
 that reported being happy with they way they are with respect to each year (2006; 2010; 201
4; 2018)
df_2010_female <- df[df$Year == 2010 & df$Sex == 'Female', ]
df_2014_female <- df[df$Year == 2014 & df$Sex == 'Female', ]
df_2018_female <- df[df$Year == 2018 & df$Sex == 'Female', ]
```

```
# Checking if the data was properly filtered
head(df_2010_female)
```

```
##    Year                Age    Sex P_Content
## 4  2010 10 years and under Female     71.77
## 12 2010           11 years Female     71.16
## 20 2010           12 years Female     71.47
## 28 2010           13 years Female     64.82
## 36 2010           14 years Female     57.32
## 44 2010           15 years Female     54.77
```

```
head(df_2014_female)
```

```
##    Year                Age    Sex P_Content
## 6  2014 10 years and under Female     77.64
## 14 2014           11 years Female     75.03
## 22 2014           12 years Female     71.34
## 30 2014           13 years Female     63.38
## 38 2014           14 years Female     64.33
## 46 2014           15 years Female     54.75
```

```
head(df_2018_female)
```

```
##    Year                Age    Sex P_Content
## 8  2018 10 years and under Female     71.25
## 16 2018           11 years Female     72.81
## 24 2018           12 years Female     68.75
## 32 2018           13 years Female     60.19
## 40 2018           14 years Female     56.55
## 48 2018           15 years Female     51.54
```

```
# Getting the summary of all the filtered data
summary(df_2006_female)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2006   Length:8           Length:8           Min.   :51.93
##  1st Qu.:2006   Class :character   Class :character   1st Qu.:59.19
##  Median :2006   Mode  :character   Mode  :character   Median :62.80
##  Mean   :2006                                         Mean   :63.98
##  3rd Qu.:2006                                         3rd Qu.:70.33
##  Max.   :2006                                         Max.   :74.69
```

```
summary(df_2010_female)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2010   Length:8           Length:8           Min.   :54.77
##  1st Qu.:2010   Class :character   Class :character   1st Qu.:56.90
##  Median :2010   Mode  :character   Mode  :character   Median :61.07
##  Mean   :2010                                         Mean   :63.06
##  3rd Qu.:2010                                         3rd Qu.:71.24
##  Max.   :2010                                         Max.   :71.77
```

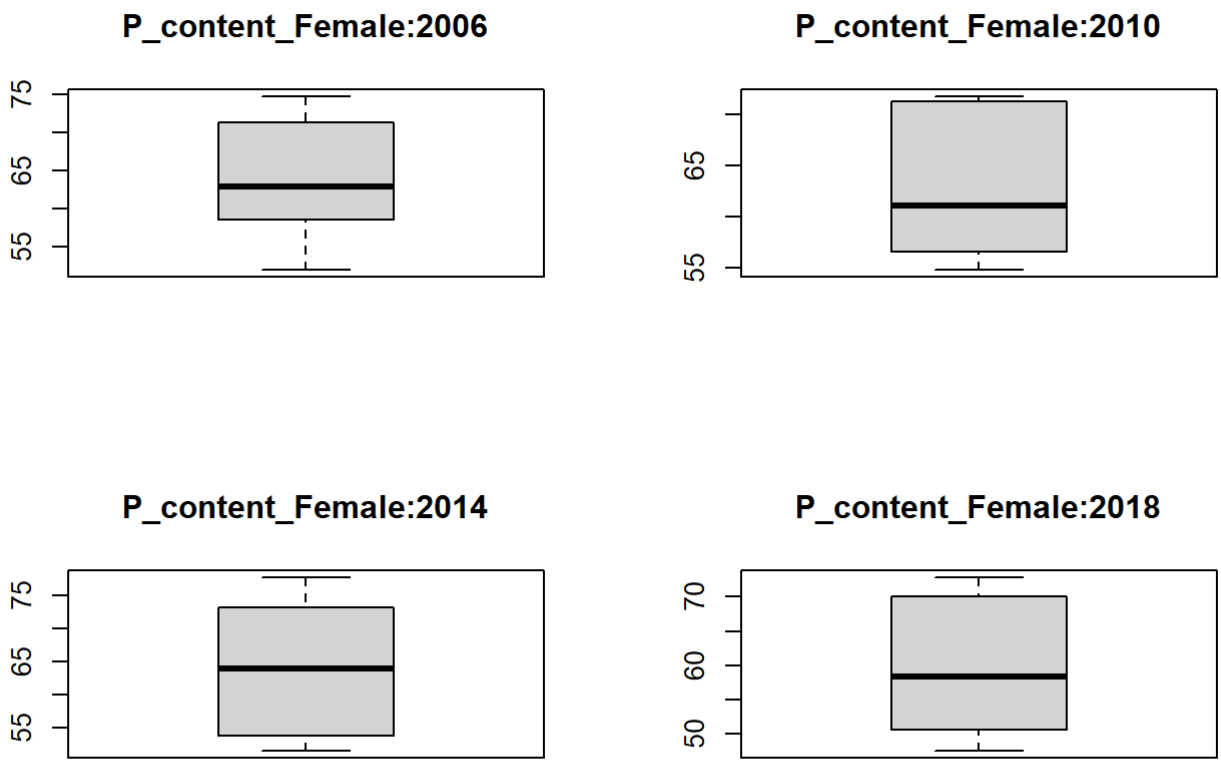
```
summary(df_2014_female)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2014   Length:8           Length:8           Min.   :51.54
##  1st Qu.:2014   Class :character   Class :character   1st Qu.:54.26
##  Median :2014   Mode  :character   Mode  :character   Median :63.85
##  Mean   :2014                                         Mean   :63.85
##  3rd Qu.:2014                                         3rd Qu.:72.26
##  Max.   :2014                                         Max.   :77.64
```

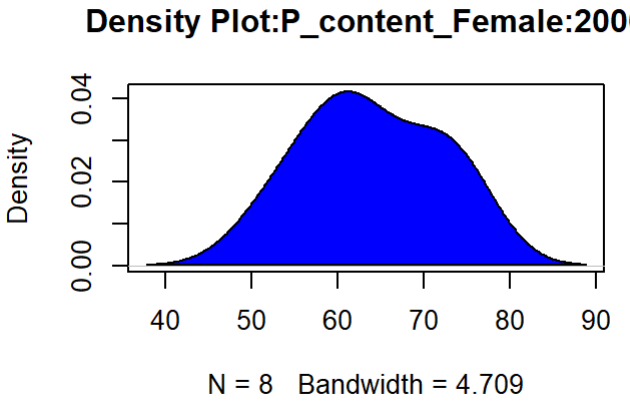
```
summary(df_2018_female)
```

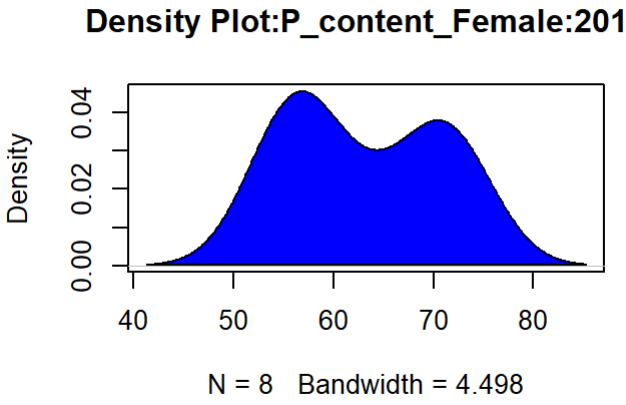
```
##       Year          Age                Sex              P_Content
##  Min.   :2018   Length:8           Length:8           Min.   :47.58
##  1st Qu.:2018   Class :character   Class :character   1st Qu.:51.07
##  Median :2018   Mode  :character   Mode  :character   Median :58.37
##  Mean   :2018                                         Mean   :59.79
##  3rd Qu.:2018                                         3rd Qu.:69.38
##  Max.   :2018                                         Max.   :72.81
```

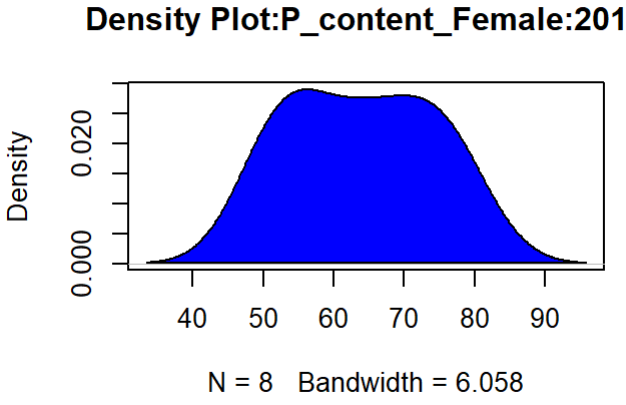
```
# Plotting a box plot and a density plot for the data
par(mfrow=c(2, 2))
boxplot(df_2006_female$P_Content, main="P_content_Female:2006")
boxplot(df_2010_female$P_Content, main="P_content_Female:2010")
boxplot(df_2014_female$P_Content, main="P_content_Female:2014")
boxplot(df_2018_female$P_Content, main="P_content_Female:2018")
```

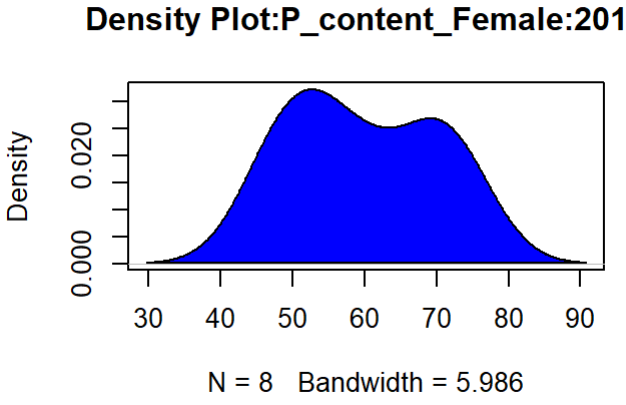



```
plot(density(df_2006_female$P_Content), main="Density Plot:P_content_Female:2006")
polygon(density(df_2006_female$P_Content), col="blue")
plot(density(df_2010_female$P_Content), main="Density Plot:P_content_Female:2010")
polygon(density(df_2010_female$P_Content), col="blue")
plot(density(df_2014_female$P_Content), main="Density Plot:P_content_Female:2014")
polygon(density(df_2014_female$P_Content), col="blue")
plot(density(df_2018_female$P_Content), main="Density Plot:P_content_Female:2018")
polygon(density(df_2018_female$P_Content), col="blue")
```









Description of the difference in the distributions for the percentage of female school aged children that reported being happy with they way they are with respect to year (2006; 2010; 2014; 2018)

a. The minimum P_Content for each of the year is slightly different from one another, the minimum P_Content for each year is given below:

- 2006: 51.93
- 2010: 54.77
- 2014: 51.54
- 2018: 47.58

b. The distribution for the year 2014 is not skewed. i.e the mean = median , while the year 2006, 20011 and 2018 has a positive skewed distribution.

c. The female's P-content are mostly similar in the following order respectively: 2010, 2018, 2014, 2006

```
# Sub-setting the data set so it contains only r the percentage of male school aged children
 that reported being happy with they way they are with respect to year (2006; 2010; 2014; 201
8)
df_2010_male <- df[df$Year == 2010 & df$Sex == 'Male', ]
df_2014_male <- df[df$Year == 2014 & df$Sex == 'Male', ]
df_2018_male <- df[df$Year == 2018 & df$Sex == 'Male', ]
```

```
# Checking if the data was properly filtered
head(df_2010_male)
```

```
##    Year                Age  Sex P_Content
## 3  2010 10 years and under Male     78.41
## 11 2010           11 years Male     70.00
## 19 2010           12 years Male     63.97
## 27 2010           13 years Male     54.42
## 35 2010           14 years Male     46.26
## 43 2010           15 years Male     42.04
```

```
head(df_2014_male)
```

```
##    Year                Age  Sex P_Content
## 5  2014 10 years and under Male     76.70
## 13 2014           11 years Male     74.41
## 21 2014           12 years Male     66.77
## 29 2014           13 years Male     48.93
## 37 2014           14 years Male     41.18
## 45 2014           15 years Male     30.23
```

```
head(df_2018_male)
```

```
##    Year                Age  Sex P_Content
## 7  2018 10 years and under Male     75.82
## 15 2018           11 years Male     69.67
## 23 2018           12 years Male     60.76
## 31 2018           13 years Male     47.51
## 39 2018           14 years Male     38.17
## 47 2018           15 years Male     32.52
```

```
# Getting the summary of all the filtered data
summary(df_2006_male)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2006   Length:8           Length:8           Min.   :35.71
##  1st Qu.:2006   Class :character   Class :character   1st Qu.:41.44
##  Median :2006   Mode  :character   Mode  :character   Median :51.93
##  Mean   :2006                                         Mean   :53.72
##  3rd Qu.:2006                                         3rd Qu.:66.62
##  Max.   :2006                                         Max.   :74.37
```

```
summary(df_2010_male)
```

```
##       Year          Age                Sex              P_Content
##  Min.   :2010   Length:8           Length:8           Min.   :39.07
##  1st Qu.:2010   Class :character   Class :character   1st Qu.:41.37
##  Median :2010   Mode  :character   Mode  :character   Median :50.34
##  Mean   :2010                                         Mean   :54.19
##  3rd Qu.:2010                                         3rd Qu.:65.48
##  Max.   :2010                                         Max.   :78.41
```

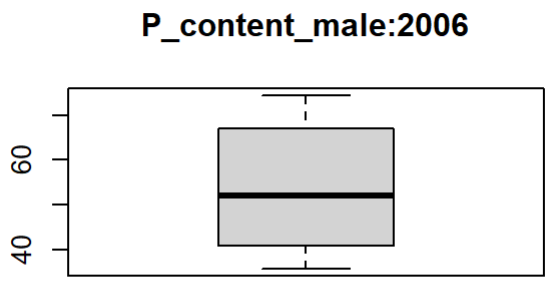
```
summary(df_2014_male)
```

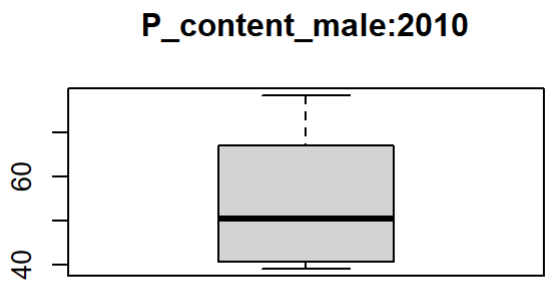
```
##       Year          Age                Sex              P_Content
##  Min.   :2014   Length:8           Length:8           Min.   :30.23
##  1st Qu.:2014   Class :character   Class :character   1st Qu.:32.60
##  Median :2014   Mode  :character   Mode  :character   Median :45.05
##  Mean   :2014                                         Mean   :50.40
##  3rd Qu.:2014                                         3rd Qu.:68.68
##  Max.   :2014                                         Max.   :76.70
```

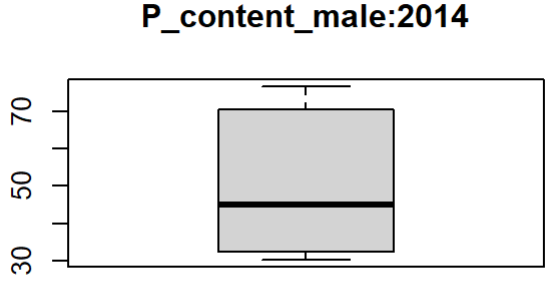
```
summary(df_2018_male)
```

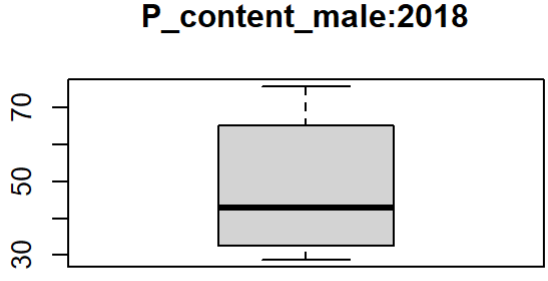
```
##       Year          Age                Sex              P_Content
##  Min.   :2018   Length:8           Length:8           Min.   :28.70
##  1st Qu.:2018   Class :character   Class :character   1st Qu.:32.62
##  Median :2018   Mode  :character   Mode  :character   Median :42.84
##  Mean   :2018                                         Mean   :48.23
##  3rd Qu.:2018                                         3rd Qu.:62.99
##  Max.   :2018                                         Max.   :75.82
```

```
# Plotting a box plot and a density plot for the data
par(mfrow=c(2, 2))
boxplot(df_2006_male$P_Content, main="P_content_male:2006")
boxplot(df_2010_male$P_Content, main="P_content_male:2010")
boxplot(df_2014_male$P_Content, main="P_content_male:2014")
boxplot(df_2018_male$P_Content, main="P_content_male:2018")
```

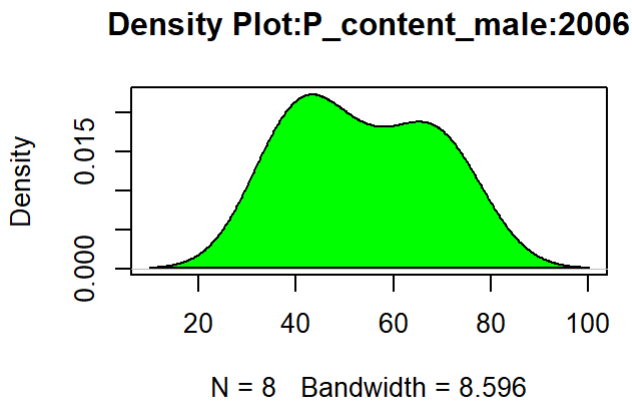


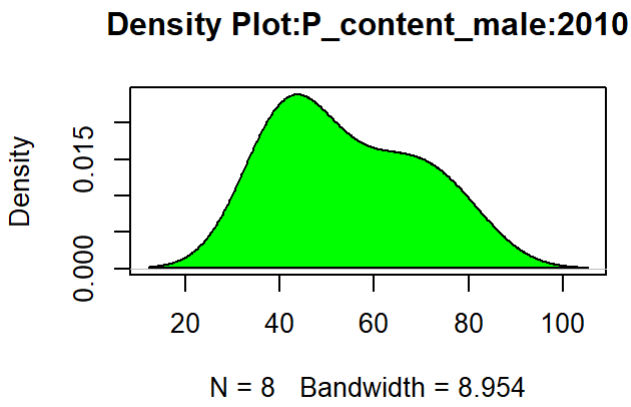


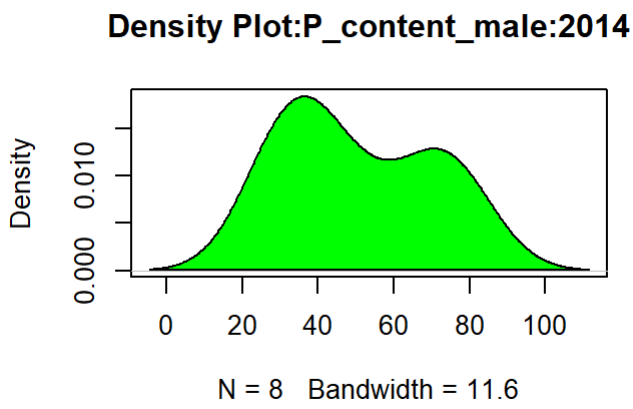


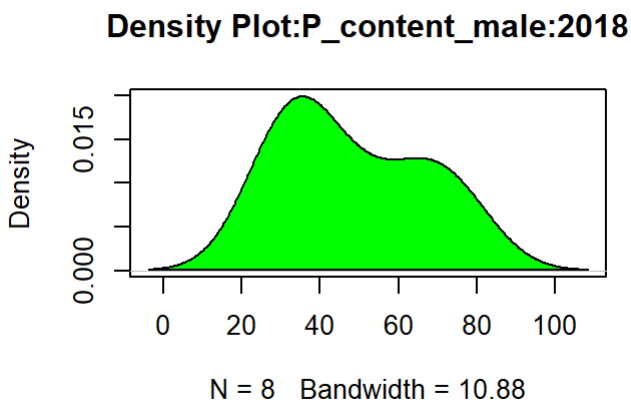


```
plot(density(df_2006_male$P_Content), main="Density Plot:P_content_male:2006")
polygon(density(df_2006_male$P_Content), col="green")
plot(density(df_2010_male$P_Content), main="Density Plot:P_content_male:2010")
polygon(density(df_2010_male$P_Content), col="green")
plot(density(df_2014_male$P_Content), main="Density Plot:P_content_male:2014")
polygon(density(df_2014_male$P_Content), col="green")
plot(density(df_2018_male$P_Content), main="Density Plot:P_content_male:2018")
polygon(density(df_2018_male$P_Content), col="green")
```









Description of the difference in the distributions for the percentage of male school aged children that reported being happy with they way they are with respect to year (2006; 2010; 2014; 2018)

a. The male's P-content are mostly similar in the following order respectively: 2014, 2018, 2006, 2010

```
# Sub-setting the data so I have separate data frame for each of the year
df_2006 <- df[df$Year == 2006, ]
df_2010 <- df[df$Year == 2010, ]
df_2014 <- df[df$Year == 2014, ]
df_2018 <- df[df$Year == 2018, ]
```

```
# Checking if the data was properly filtered
head(df_2006)
```

```
##    Year                Age    Sex P_Content
## 1  2006 10 years and under   Male     67.72
## 2  2006 10 years and under Female     69.39
## 9  2006           11 years   Male     74.37
## 10 2006           11 years Female     74.69
## 17 2006           12 years   Male     66.25
## 18 2006           12 years Female     73.16
```

```
head(df_2010)
```

```
##    Year                Age    Sex P_Content
## 3  2010 10 years and under   Male     78.41
## 4  2010 10 years and under Female     71.77
## 11 2010           11 years   Male     70.00
## 12 2010           11 years Female     71.16
## 19 2010           12 years   Male     63.97
## 20 2010           12 years Female     71.47
```

```
head(df_2014)
```

```
##    Year                Age    Sex P_Content
## 5  2014 10 years and under   Male     76.70
## 6  2014 10 years and under Female     77.64
## 13 2014           11 years   Male     74.41
## 14 2014           11 years Female     75.03
## 21 2014           12 years   Male     66.77
## 22 2014           12 years Female     71.34
```

```
head(df_2018)
```

```
##    Year                Age    Sex P_Content
## 7  2018 10 years and under   Male     75.82
## 8  2018 10 years and under Female     71.25
## 15 2018           11 years   Male     69.67
## 16 2018           11 years Female     72.81
## 23 2018           12 years   Male     60.76
## 24 2018           12 years Female     68.75
```

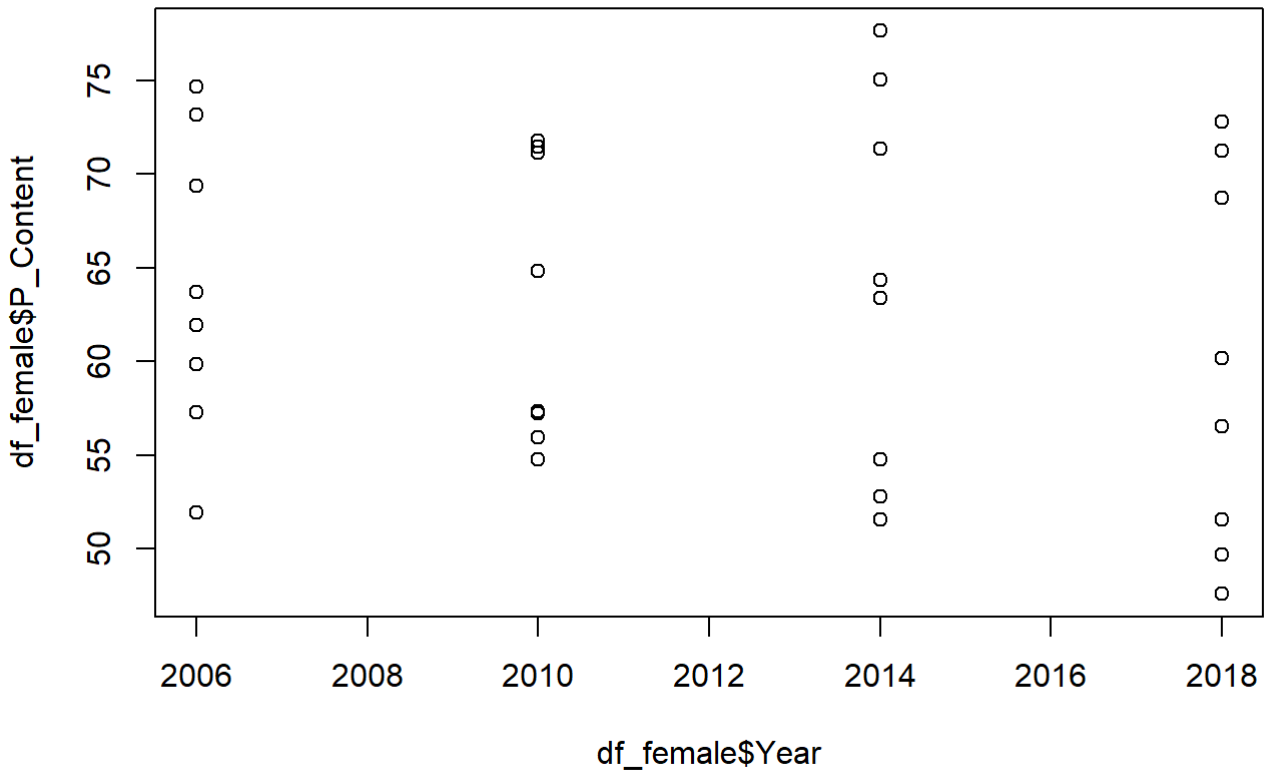
```
# Converting the categorical variable Sex to a factor.
df_2006$Sex <-as.factor(df_2006$Sex)
df_2010$Sex <-as.factor(df_2010$Sex)
df_2014$Sex <-as.factor(df_2014$Sex)
df_2018$Sex <-as.factor(df_2018$Sex)
```

```
# Plotting a bar chart to compare the gender for each of the year
par(mfrow=c(2, 2))
barplot(table(df_2006$Sex))
barplot(table(df_2010$Sex))
barplot(table(df_2014$Sex))
barplot(table(df_2018$Sex))
```

Describing the frequency of the categorical variable Sex with respect to year (2006; 2010; 2014; 2018)
The female and the male sex has the same frequency for each of the year. i.e. No of female = No of male = 8 for the year 2006, 2010, 2014, 2018.

```
# Filtering the data based on sex
df_female <- df[df$Sex == 'Female', ]
df_male <- df[df$Sex == 'Male', ]
```

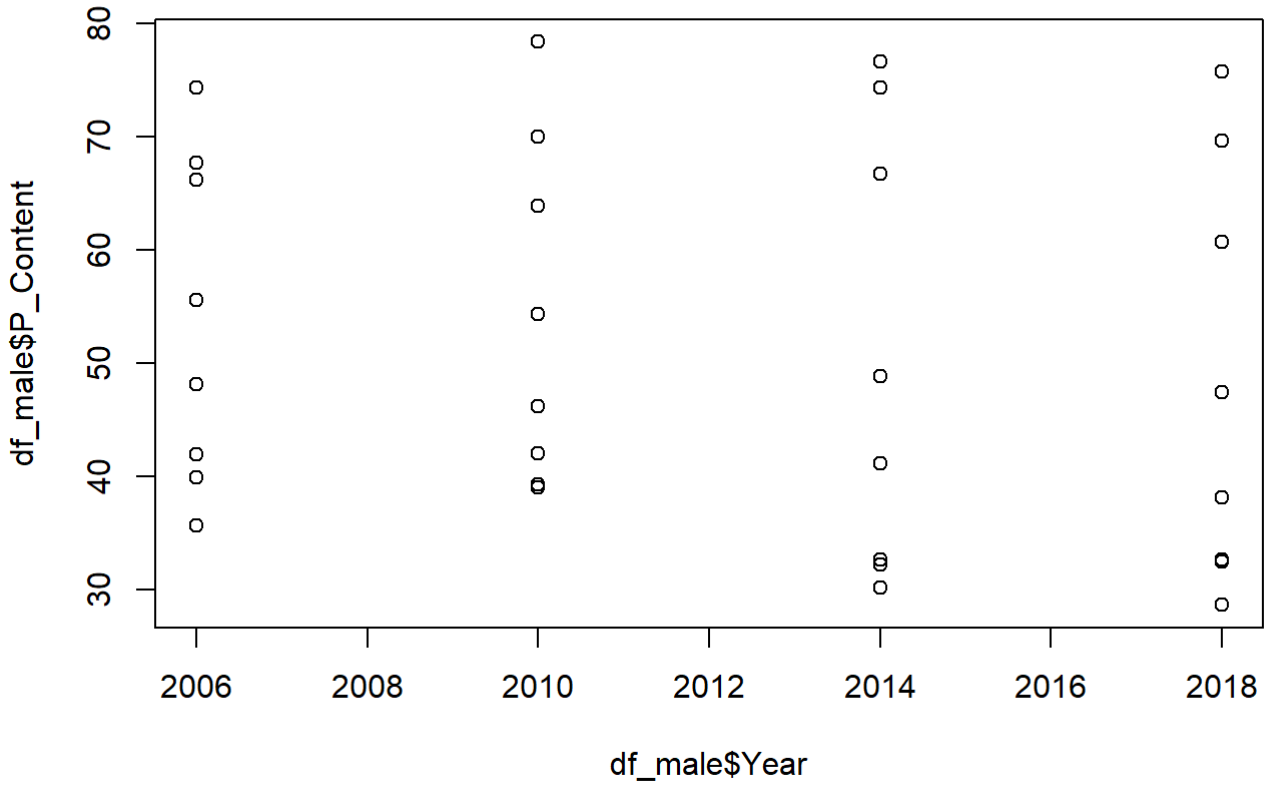
```
# Making a scatter plot to determine the relationship between P_Content and Year for the fema
les
plot(x = df_female$Year, y = df_female$P_Content)
```




```
# Determining the correlation between P_Content and Year for the females.
cor(x = df_female$Year, y = df_female$P_Content)
```

```
## [1] -0.1528523
```

```
# Making a scatter plot to determine the relationship between P_Content and Year for the male
s
plot(x = df_male$Year, y = df_male$P_Content)
```




```
# Making a scatter plot to determine the relationship between P_Content and Year for the male
s
cor(x = df_male$Year, y = df_male$P_Content)
```

```
## [1] -0.141013
```

Discussing the relationship between P_Content and Year the females

-0.1528523 indicates a very week negative linear relationship between P_Content and Year

Discussing the relationship between P_Content and Year the males

-0.141013 indicates a very week negative linear relationship between P_Content and Year

Comment: I am done with the exploratory data analysis, I will move to the regression analysis

Fitting a linear regression model

I want to fit a simple linear regression model to the data with P_Content as the response variable and Year as a numeric predictor variable for females.

$$PContent = \beta_0 + \beta_1 Year + E_i$$

where: - P_Content is the percentage of children that reported being happy with they way they are -

$$\beta_0$$

is the intercept, that is the value of the P_Content when no year is involved. -

$$\beta_1$$

is the slope or regression co–efficient, that is the change in P_Content when there is a one unit change in Year. - Year is the Year the information was collected. -

$$E_i$$

is the Residual (error)

```
linearMod <- lm(P_Content ~Year, data = df_female)
linearMod
```

```
##
## Call:
## lm(formula = P_Content ~ Year, data = df_female)
##
## Coefficients:
## (Intercept)         Year
##    655.0157      -0.2944
```

Interpretation the estimate of the intercept term

The intercept has a value of 655.0157, which implies that is the value of the P_Content when no year is involved.

Interpretation of the estimate of the slope

This model estimates that increasing the Year by an additional year when the information was collectecd it will result in a -0.2944 decrease of the P_Content.

Explanation of standard error of a parameter

The standard error of a parameter is a way to measure the uncertainty in the estimate of that parameter.

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = P_Content ~ Year, data = df_female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 655.0157   699.2138   0.937    0.356
## Year         -0.2944     0.3475  -0.847    0.404
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Calculating the standard error of the intercept, that is β0
N = length(df_female$Year)
MSE = sum(linearMod$residuals^2/(N-2))
SXX = sum((df_female$Year-mean(df_female$Year))^2)
VARB0 = MSE*(1/ N + (mean(df_female$Year)^{2}/SXX))
standard_error= sqrt(VARB0)
standard_error
```

```
## [1] 699.2138
```

```
N = length(df_female$Year)
SSE = sum(linearMod$residuals^2)
MSE = SSE/(N-2)
SXX = sum((df_female$Year - mean(df_female$Year))^2)
VARB1 = MSE/SXX
standard_error2 = sqrt( VARB1)
standard_error2
```

```
## [1] 0.3475209
```

comment on the standard error of the estimate of the intercept and slope term.

- The standard error of the regression intercept(699.2138) was large relative to the coefficient estimate(655.0157) of the regression intercept, the predictor variable was not statistically significant.
- The standard error of the regression slope(0.3475209) was large relative to the coefficient estimate(-0.2944) of the regression slope, the predictor variable was not statistically significant.

```
# Calculating confidence interval for 60
N = length(df_female$Year)
MSE = sum(linearMod$residuals^2/(N-2))
SXX = sum((df_female$Year-mean(df_female$Year))^2)
VARB0 = MSE*(1/ N + (mean(df_female$Year)^{2}/SXX))
alpha=0.05
beta0 = linearMod$coefficients[1]
c(beta0 - qt(1-alpha/2,N-2)*sqrt(VARB0),
beta0 + qt(1-alpha/2,N-2)*sqrt(VARB0))
```

```
## (Intercept) (Intercept)
##   -772.9695   2083.0008
```

comment on the confidence interval for β0: I am 95% confident that β0 lies between -772.9695 < β0 < 2083.0008

```
# Calculating confidence interval for B1
N = length(df_female$Year)
SSE = sum(linearMod$residuals^2)
MSE = SSE/(N-2)
SXX = sum((df_female$Year - mean(df_female$Year))^2)
VARB1 = MSE/SXX
beta1= linearMod$coefficients[2]
alpha=0.05
c(beta1 - qt(1-alpha/2,N-2)*sqrt( VARB1),
beta1 + qt(1-alpha/2,N-2)*sqrt( VARB1))
```

```
##       Year       Year
## -1.0041387  0.4153262
```

comment on the confidence interval for B1: I am 95% confident that B1 lies between -1.0041387 < B1 < 0.4153262

Explanation of What the confidence interval of a parameter measure

The confidence interval of a parameter measures the probability that a parameter will be between a pair of values.

Does a 95% confidence interval always contain the population parameter?

No, the interval can only either contain the population parameter or not.This is because A 95% confidence interval means that there is a 95% probability that the population parameter falls within the given interval, However the true population parameter remains the same regardless of the sample.

Computing an hypothesis test

$$H_o : \beta_0 = 0$$

Versus

$$H_a : \beta_0 \neq 0$$

The test statistic is:

$$T = \beta_0(estimated) - \beta_0 / \sqrt{Var(\beta_0)}$$

```
# Computing the t statistic
N = length(df_female$Year)
MSE = sum(linearMod$residuals^2/(N-2))
SXX = sum((df_female$Year - mean(df_female$Year))^2)
VARB0 = MSE*(1/ N + (mean(df_female$Year)^{2}/SXX))
T = (linearMod$coefficients[1]-0)/sqrt(VARB0)
T
```

```
## (Intercept)
##   0.9367888
```

```
# Computing the P-value
alpha = 0.05
TDIST = qt(1-alpha/2, N-2)
TDIST
```

```
## [1] 2.042272
```

```
PVALUE = 2 *( 1- pt(T, df = N- 2))
PVALUE
```

```
## (Intercept)
##   0.3563479
```

Commenting on the result of the hypothesis testing

Comparing |T| = 0.9367888 with t distribution = 2.042272 , |0.9367888| < 2.042272 , I will accept the null hypothesis. At the 5% level of significance, the evidence is strong enough to indicate that β0 = 0. Indicating that when the P_Content is at the mean the year is zero.

Computing the second hypothesis test

$$H_o : \beta_1 = 0$$

Versus

$$H_a : \beta_1 \neq 0$$

The test statistic is:

$$T = \beta_1(estimated) - \beta_0 / \sqrt{Var(\beta_1)}$$

```
# Computing the t statistic
N = length(df_female$Year)
MSE = sum(linearMod$residuals^2/(N-2))
SXX = sum((df_female$Year - mean(df_female$Year))^2)
VARB1 = MSE/SXX
T = (linearMod$coefficients[2]-0)/sqrt(VARB1)
T
```

```
##       Year
## -0.8471612
```

```
# Computing the P-value
alpha = 0.05
TDIST = qt(1-alpha/2, N-2)
TDIST
```

```
## [1] 2.042272
```

```
PVALUE = 2*(1-pt(T, df = N - 2))
PVALUE
```

```
##     Year
## 1.596387
```

Commenting on the result of the hypothesis testing

Comparing |T| = -0.8471612 with t distribution = 2.042272 , |-0.8471612| < 2.042272 , I will accept the null hypothesis. At the 5% level of significance, the evidence is strong enough to indicate that β1 = 0. Indicating that no relation exists between P-Content and year.

Computing an hypothesis test for F-statistic

$$H_o : \beta_1 = 0$$

Versus

$$H_a : \beta_1 \neq 0$$

```
# Computing the F-statistic and the P-value
MSR = sum((fitted(linearMod) - mean(df_female$P_Content))^2) / 1
MSE = sum(linearMod$residuals^2/(N-2))
F = MSR/MSE
F
```

```
## [1] 0.7176821
```

```
alpha = 0.05
FDIST = qf(1-alpha,1,N-2)
FDIST
```

```
## [1] 4.170877
```

```
PVALUE = pf(1-F, 1, N - 2)
PVALUE
```

```
## [1] 0.4009016
```

```
# Alternatively I could use the summary of the regression model to interpret the F-statistic
 and the P-value\
summary(linearMod)
```

```
##
## Call:
## lm(formula = P_Content ~ Year, data = df_female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 655.0157   699.2138   0.937    0.356
## Year         -0.2944     0.3475  -0.847    0.404
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

Commenting on the result of the hypothesis testing (F-statistics)

Comparing F = 0.7176821 with f distribution = 4.170877, 0.7176821 < 4.170877 so I will accept the null hypothesis. At the 5% level of significance, the evidence is strong enough to indicate that β1 = 0. Indicating that no relation exists between speed and stopping distance.

```
# Calculating the coefficient of determination, R squared
SST = sum((df_female$P_Content-mean(df_female$P_Content))^2)
SSE = sum(linearMod$residuals^2)
R2 <- (SST - SSE) /SST
R2
```

```
## [1] 0.02336381
```

Commenting on the result of value of R squared

Approximately 2.3% of the observed variation in the P_Content can be explained by the year of information collected.

```
# Calculating the residual Standard Error
N = length(df_female$Year)
RMSE = sqrt(SSE/(N-2))
RMSE
```

```
## [1] 8.791661
```

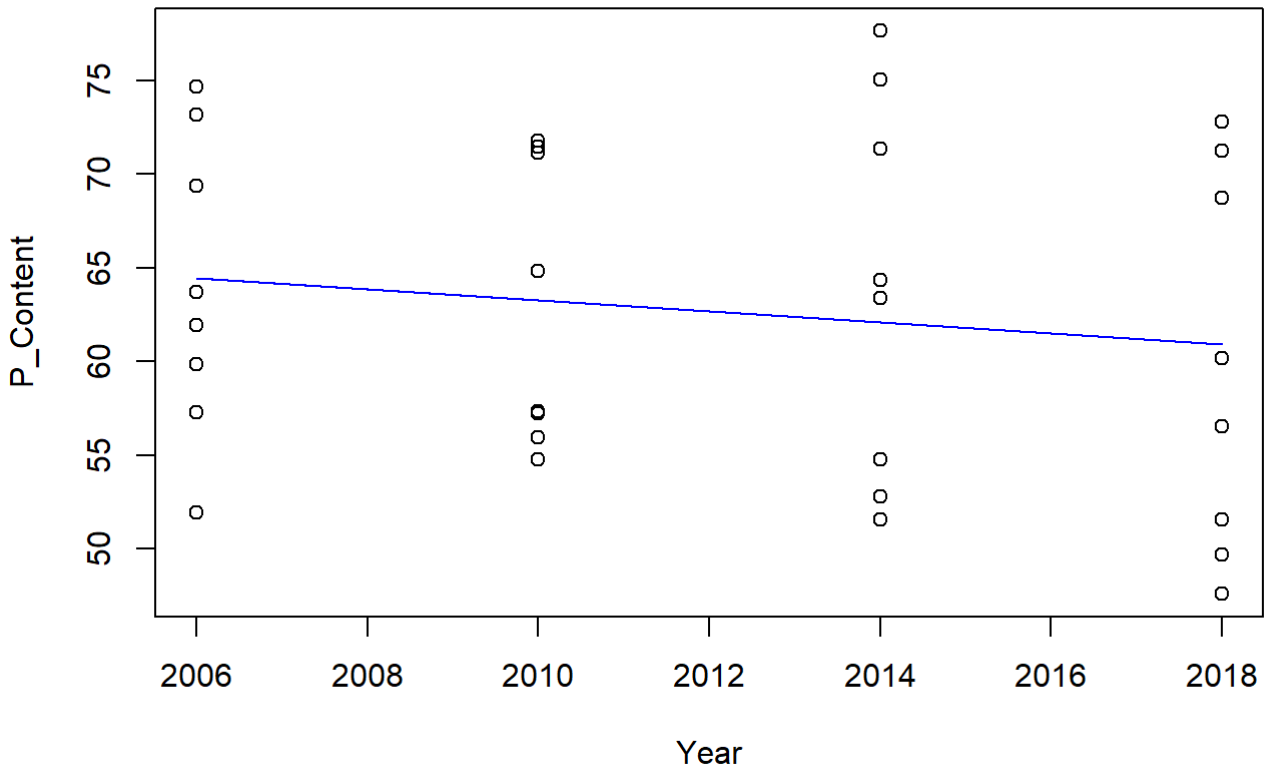
Commenting on the result of value residual Standard Error

I can say that the speed accurately predicts distance with about 8.791661 P_Content error on average.

```
# plotting the shape of the prediction intervals for the estimated values of
N = length(df_female$Year)
SXX =  sum((df_female$Year - mean(df_female$Year))^2)
MSE = SSE/(N-2)
Var_E = MSE*(1 + 1/N + (df_female$Year-mean(df_female$Year))^2/SXX)
Yhat = fitted(linearMod)
alpha = 0.05
cbind(Yhat- qt(1-alpha/2,N-2)*sqrt( Var_E),
Yhat + qt(1-alpha/2,N-2)*sqrt( Var_E))
```

```
##        [,1]     [,2]
## 2  45.71272 83.16078
## 4  44.97060 81.54765
## 6  43.79297 80.37003
## 8  42.17985 79.62790
## 10 45.71272 83.16078
## 12 44.97060 81.54765
## 14 43.79297 80.37003
## 16 42.17985 79.62790
## 18 45.71272 83.16078
## 20 44.97060 81.54765
## 22 43.79297 80.37003
## 24 42.17985 79.62790
## 26 45.71272 83.16078
## 28 44.97060 81.54765
## 30 43.79297 80.37003
## 32 42.17985 79.62790
## 34 45.71272 83.16078
## 36 44.97060 81.54765
## 38 43.79297 80.37003
## 40 42.17985 79.62790
## 42 45.71272 83.16078
## 44 44.97060 81.54765
## 46 43.79297 80.37003
## 48 42.17985 79.62790
## 50 45.71272 83.16078
## 52 44.97060 81.54765
## 54 43.79297 80.37003
## 56 42.17985 79.62790
## 58 45.71272 83.16078
## 60 44.97060 81.54765
## 62 43.79297 80.37003
## 64 42.17985 79.62790
```

```
plot(df_female$Year,df_female$P_Content,xlab="Year",ylab="P_Content")
lines(df_female$Year,Yhat,col="blue")
lines(df_female$Year,Yhat +qt(1-alpha/2,N-2)*sqrt(Var_E),col="red")
lines(df_female$Year,Yhat -qt(1-alpha/2,N-2)*sqrt(Var_E),col="red")
```



I, Akolade Sofiyyah Iwalewa confirm that this assignment is my own work. I have not copied in part or whole or otherwise plagiarised the work of other students and/or persons. I confirm that I have read and understood the UCD School of Mathematics and Statistics regulations on plagiarism in the Week 5 folder on bright space.