IMT 573: Problem Set 5 - Learning from Data

Naga Soundari Balamurugan

Due: Tuesday, November 6, 2018

Collaborators: Jayashree Raman

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset5.Rmd` file from Canvas. Open `problemset5.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset5.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:
4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_ps5.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages. If you have not already installed them, do so in your console before loading the library here.

```
# Load standard libraries
library(tidyverse)
library(Sleuth3) # Contains data for problemset
library(UsingR) # Contains data for problemset
library(MASS) # Modern applied statistics functions
library(kableExtra)
```

Problem 1:

Davis et al. (1998) collected data on the proportion of births that were male in Denmark, the Netherlands, Canada, and the United States for selected years. Davis et al. argue that the proportion of male births is declining in these countries. We will explore this hypothesis. You can obtain this data as follows:

```
# Import male births data
malebirths <- Sleuth3::ex0724
```

1a.

Use the `lm` function in **R** to fit four (one per country) simple linear regression models of the yearly proportion of males births as a function of the year and obtain the least squares fits. Write down the estimated linear model for each country.

```
#Linear model for male births of Denmark
lm_Denmark <- lm(Denmark ~ Year, data = malebirths)
summary(lm_Denmark)
```

```
##
## Call:
## lm(formula = Denmark ~ Year, data = malebirths)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003225 -0.001339  0.000089  0.001119  0.003790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
## Year        -4.289e-05  2.069e-05  -2.073   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001803 on 43 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.06968
## F-statistic: 4.296 on 1 and 43 DF,  p-value: 0.04424
```

```
#Linear model for male births of Netherlands
lm_Netherlands <- lm(Netherlands ~ Year, data = malebirths)
summary(lm_Netherlands)
```

```
##
## Call:
## lm(formula = Netherlands ~ Year, data = malebirths)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031437 -0.0008246  0.0002819  0.0009287  0.0021478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.724e-01  2.792e-02   24.08  < 2e-16 ***
## Year        -8.084e-05  1.416e-05   -5.71 9.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001233 on 43 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.418
## F-statistic: 32.61 on 1 and 43 DF,  p-value: 9.637e-07
```

```
#Linear model for male births of Canada
lm_Canada <- lm(Canada ~ Year, data = malebirths)
summary(lm_Canada)
```

```
##
## Call:
## lm(formula = Canada ~ Year, data = malebirths)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.494e-03 -6.161e-04 -8.312e-05  4.951e-04  1.284e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.338e-01  5.480e-02  13.390 3.98e-11 ***
## Year        -1.112e-04  2.768e-05  -4.017 0.000738 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000768 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.4592, Adjusted R-squared:  0.4307
## F-statistic: 16.13 on 1 and 19 DF,  p-value: 0.0007376
```

```
#Linear model for male births of USA
lm_USA <- lm(USA ~ Year, data = malebirths)
summary(lm_USA)
```

```
##
## Call:
## lm(formula = USA ~ Year, data = malebirths)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
## Year        -5.429e-05  9.393e-06  -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

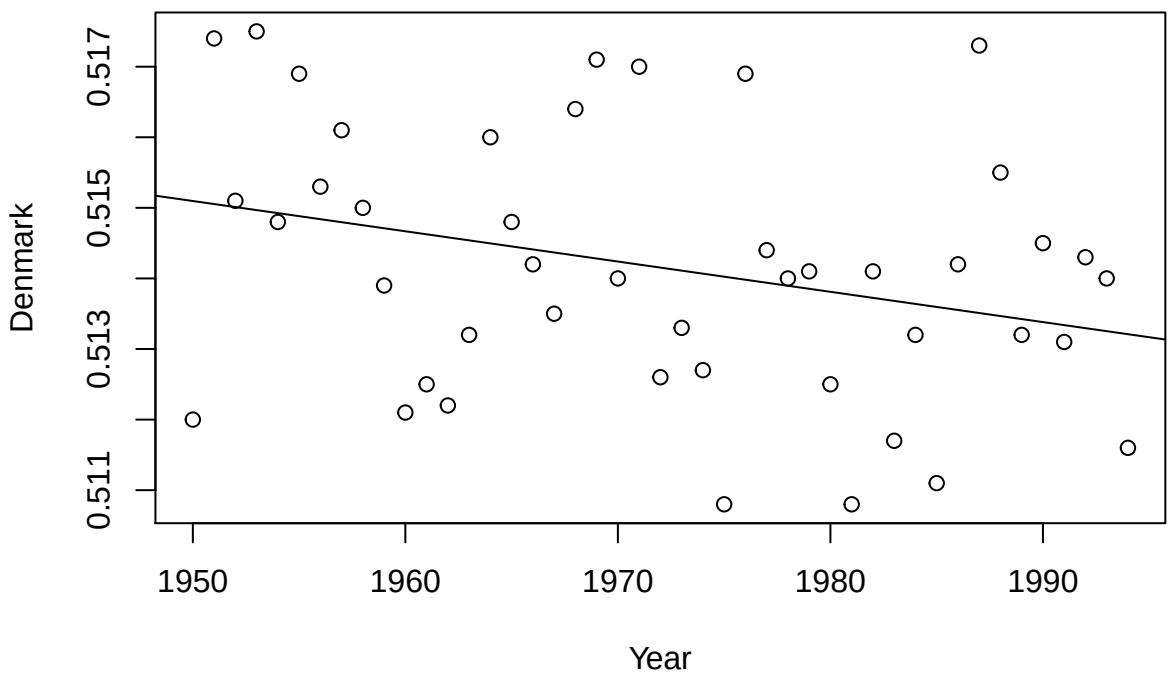
birthrate_Denmark = 5.987e-01 + ((-4.289e-05) * year) birthrateNetherlands = 6.724e-01 + ((-8.084e-05) * year) birthrateCanada = 0.7338 + ((-1.112e-04) * year) birthrateUSA = 6.201e-01 + ((-5.429e-05) * year)

1b.

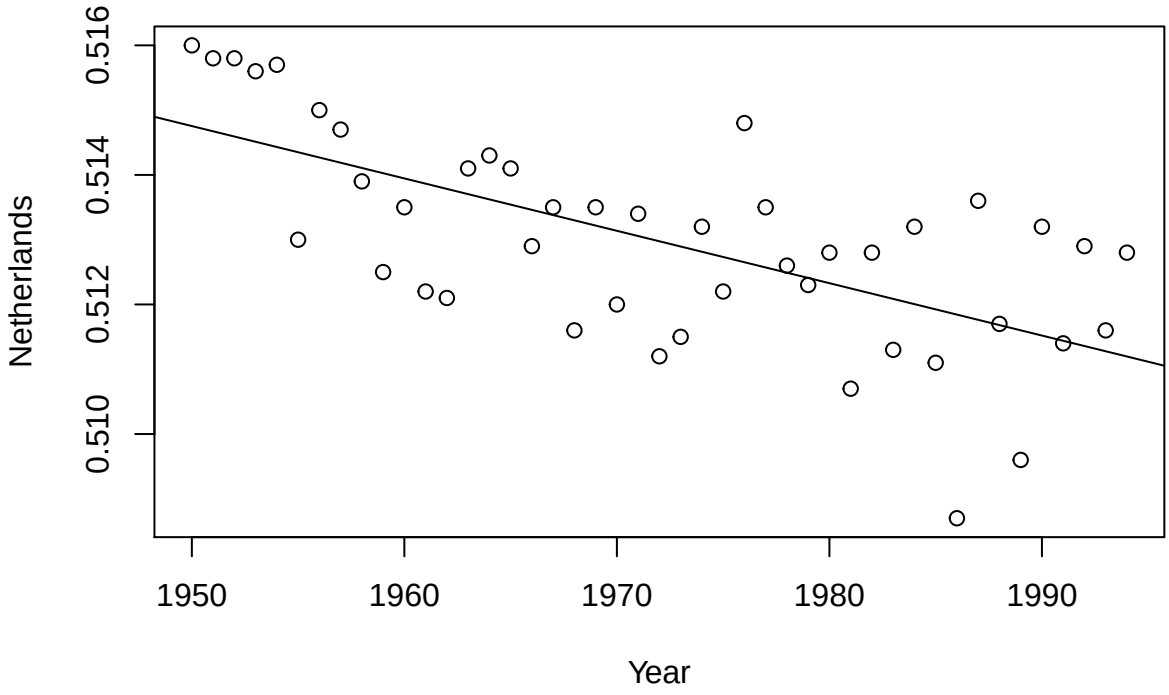
Obtain the t-statistic for the test that the slopes of the regression lines are zero, for each of the four countries. Is there evidence that the proportion of births that are male is truly declining over this period?

```
#Plots with Regression lines
with(malebirths,plot(Year, Denmark))
```

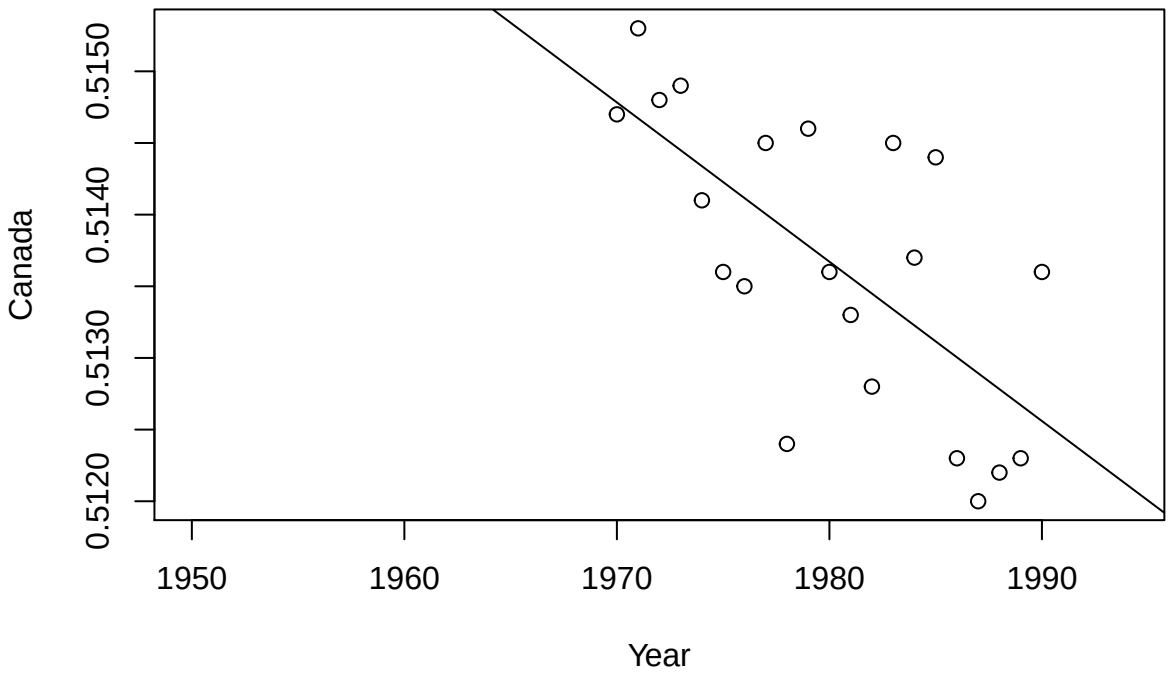
```
    abline(lm_Denmark)
```



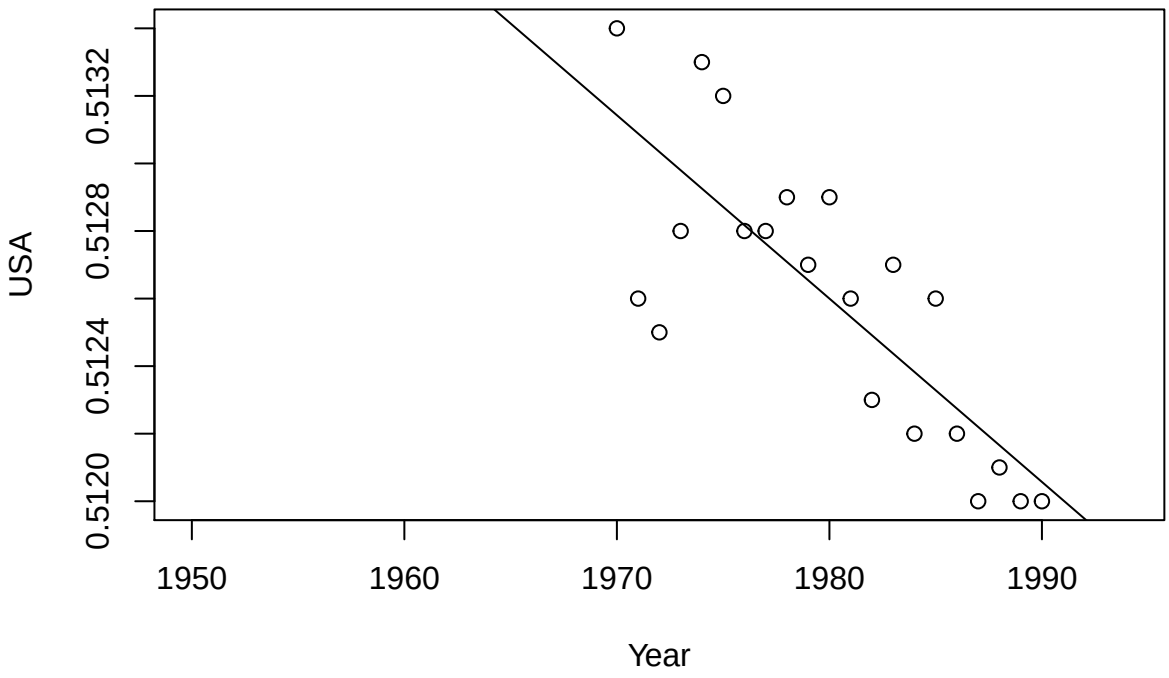

```
with(malebirths,plot(Year, Netherlands))
    abline(lm_Netherlands)
```




```
with(malebirths,plot(Year, Canada))
    abline(lm_Canada)
```




```
with(malebirths,plot(Year, USA))
    abline(lm_USA)
```



From the summary of the linear models, the t statistics and p-values are as follows. * Denmark: t-value -2.073, p-value 0.04424 * Netherlands: t-value -5.71, p-value 9.637e-07 * Canada: t-value -4.017, p-value 0.0007376 * USA: t-value -5.779, p-value 1.44e-05

In the above plots, we can see the regression lines for each country. The slope of the regression line is almost zero for the country Denmark. That means there is no significant relationship the dependent(male proportion) and independent variables(year). The slopes of the regression lines are non-zero for all the other countries. The countries Canada and USA shows a clear decline in the male proprtion whereas the decline is not much in Netherlands and Denmark.

Also, the p-values of all four countries are less that 0.05 and is against the null hypothesis. Hence we can say that there is evidence that the proportion of births that are male is truly declining over this period.

Problem 2:

Regression was originally used by Francis Galton to study the relationship between parents and children. One relationship he considered was height. Can we predict a man's height based on the height of his father? This is the question we will explore in this problem. You can obtain data similar to that used by Galton as follows:

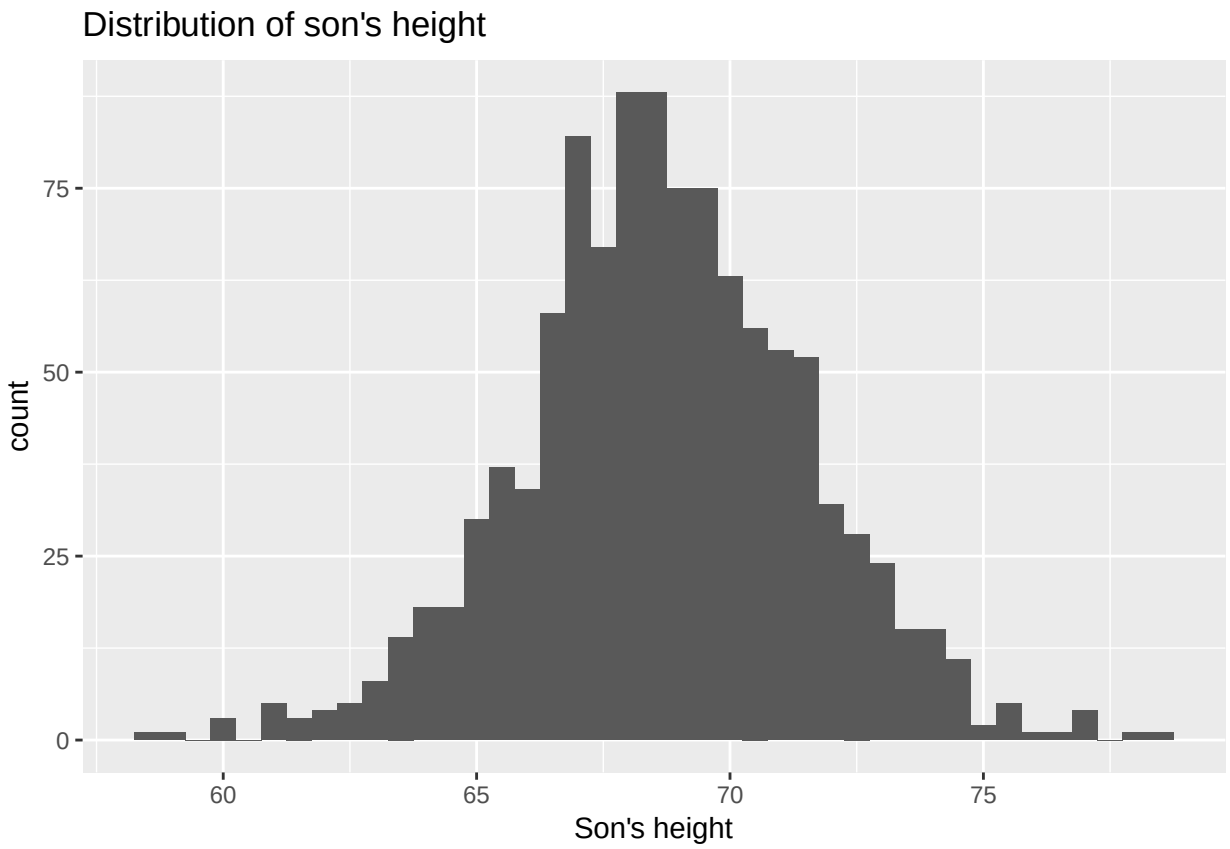
```
# Import and look at the height data
heightData <- tbl_df(get("father.son"))
```

2a.

Perform an exploratory analysis of the dataset. Describe what you find. At a minimum you should produce statistical summaries of the variables, a visualization of the relationship of interest in this problem, and a

statistical summary of that relationship.

```
#Plot the distribution of son's height as a histogram
ggplot(heightData, aes(x = heightData$sheight)) + geom_histogram(binwidth = 0.5) +
    xlab("Son's height") + ggtitle("Distribution of son's height")
```




```
#Correlation between Father's and son's height
correlation <- cor(heightData$fheight, heightData$sheight)
correlation
```

```
## [1] 0.5013383
```

```
corr_matrix <- cor(heightData, use="complete.obs")

kable(corr_matrix, "latex") %>% kable_styling(bootstrap_options = c("striped", "hover"))
```

	fheight	sheight
fheight	1.0000000	0.5013383
sheight	0.5013383	1.0000000

From the plot, we see that most of the height falls in the range of 67 - 72. The curve almost looks like a normal distribution. Also the correlarion value is **0.501** which is a positive correlation but is not a strong one.

2b.

Use the `lm` function in R to fit a simple linear regression model to predict son's height as a function of father's height. Write down the model,

$$\hat{y}_{\texttt{sheight}} = \hat{\beta}_0 + \hat{\beta}_i \times \texttt{fheight}$$

filling in estimated coefficient values and interpret the coefficient estimates.

```
#Build a liner model to that predicts son's height as a function of father's height
lm_height <- lm(sheight ~ fheight, data = heightData)
summary(lm_height)
```

```
##
## Call:
## lm(formula = sheight ~ fheight, data = heightData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8772 -1.5144 -0.0079  1.6285  8.9685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.88660    1.83235   18.49   <2e-16 ***
## fheight      0.51409    0.02705   19.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.437 on 1076 degrees of freedom
## Multiple R-squared:  0.2513, Adjusted R-squared:  0.2506
## F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

> From the linear model summary, the model can be represented as, **Son's height(y) = 33.8866(Intercept) + 0.51409 * Father's Height**. Thus in order to find the son's height the father's height is to be multiplied by 0.514 and a constant value of 33.886 to be added. As the p-values are less than 0.05, the model is considered statistically significant.

2c.

Find the 95% confidence intervals for the estimates. You may find the `confint()` command useful.

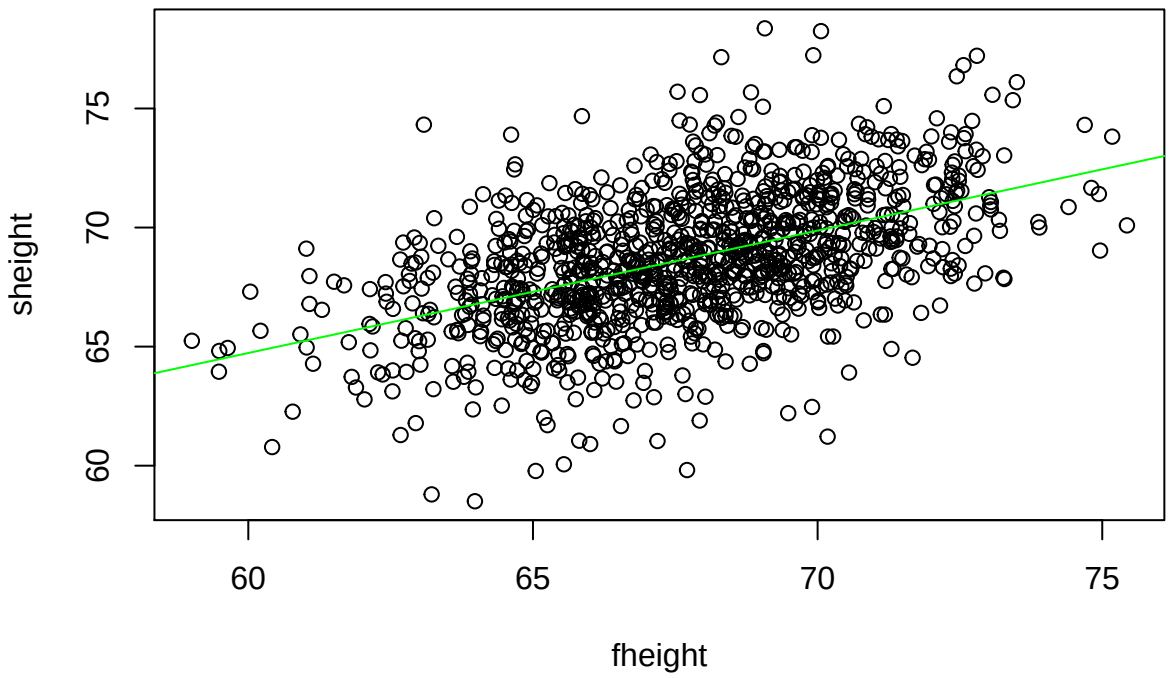
```
#Estimate the confidence intervals
conf_intervals <- confint(lm_height, level = 0.95)
kable(conf_intervals, "latex") %>% kable_styling(bootstrap_options = c("striped", "hover"))
```

	2.5 %	97.5 %
(Intercept)	30.2912126	37.4819961
fheight	0.4610188	0.5671673

2d.

Produce a visualization of the data and the least squares regression line.

```
with(heightData, plot(fheight, sheight))
    abline(lm_height, col = 'green')
```

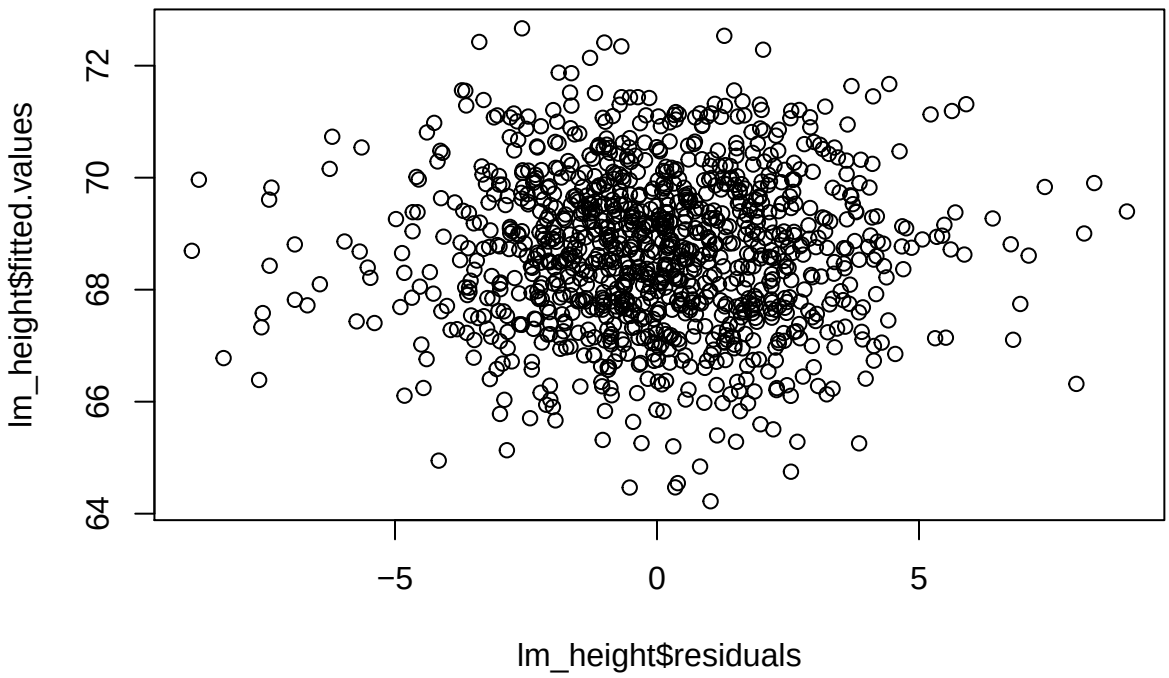


The regression line has a positive slope.

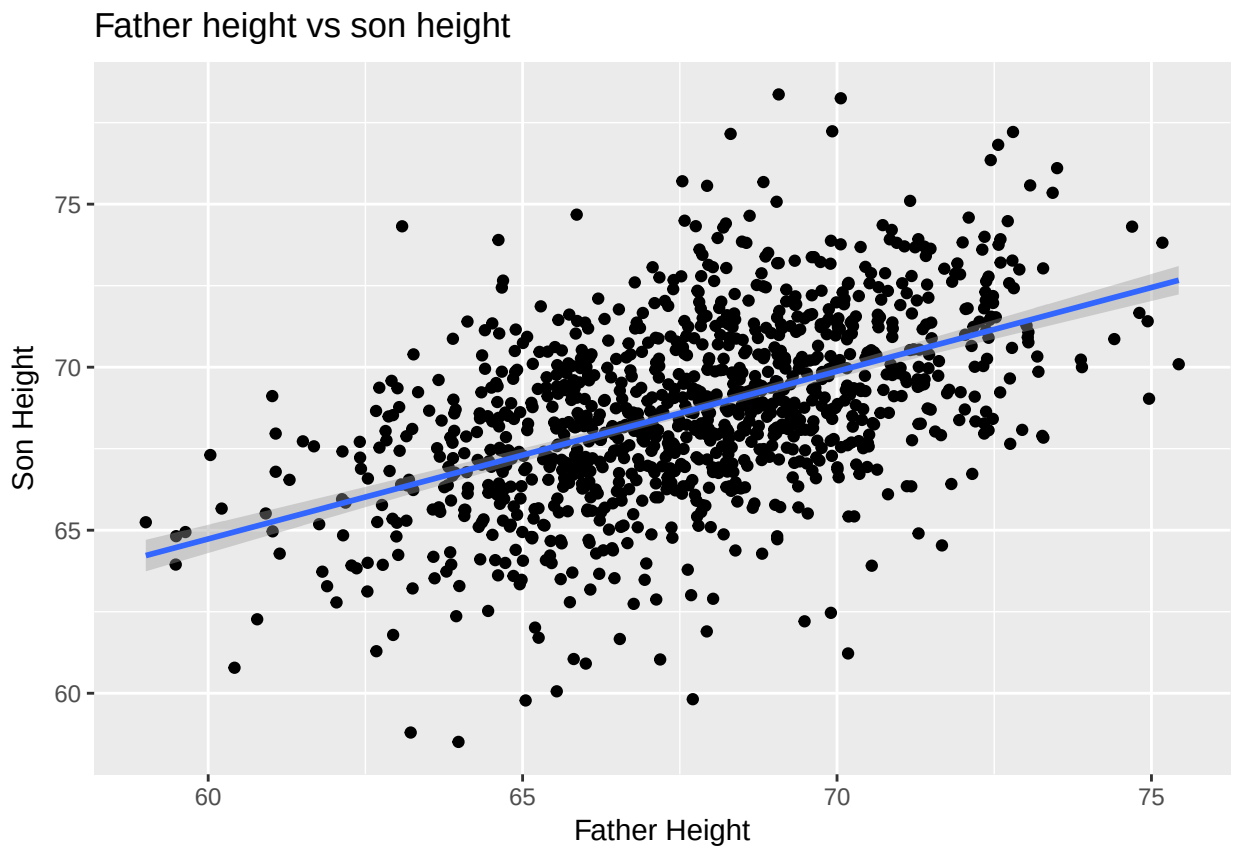
2e.

Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using `names()`). Discuss what you see. Do you have any concerns about the linear model?

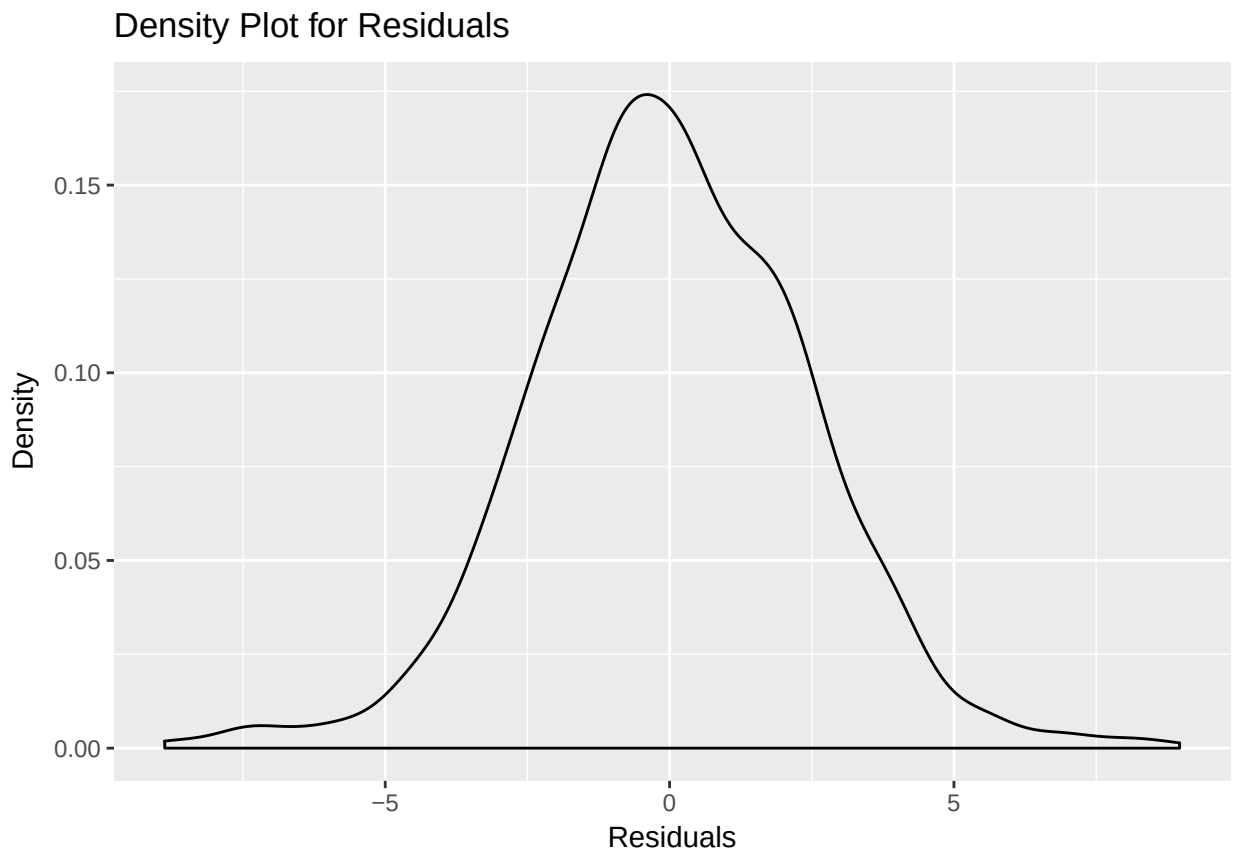
```
#Plot of residuals against fitted values
plot(lm_height$residuals, lm_height$fitted.values)
```




```
#Fit the plot with linear model
ggplot(lm_height, aes(x = fheight, y = sheight)) + geom_point() +
  xlab('Father Height') + ylab('Son Height') + ggtitle('Father height vs son height') + geom_smooth(metl
```




```
#Density plot of residuals
ggplot(lm_height, aes(x = lm_height$residuals)) + geom_density() +
  xlab('Residuals') + ylab('Density') + ggtitle('Density Plot for Residuals')
```



The density plot curve is more like a normal distribution. Hence using a linear model is more appropriate.

2f.

Using the model you fit in part (b) predict the height was 5 males whose father are 50, 55, 70, 75, and 90 inches respectively. You may find the `predict()` function helpful.

```
new <- data.frame(fheight = c(50, 55, 70, 75, 90))
predictedVals <- predict(lm_height, new, se.fit = TRUE)
predictedVals$fit
```

```
##        1        2        3        4        5
## 59.59126 62.16172 69.87312 72.44358 80.15498
```

Extra Credit:

EC(a).

What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model?

i) For each value of the explanatory variable, the distribution of possible response variables values is normal.

ii) The normal distribution for response variable(Y) values corresponding to a particular value of explanatory variable(X) has a mean µ{Y|X} that lies on the straight line µ{Y|X} = ??0 + ??1X. This line is called the population regression line. The parameter ??0 is the intercept of the population regression line. It represents the mean of the Y values when X = 0. The parameter ??1 is the slope of the population regression line. It represents the change in the mean of Y per unit increase in X.

iii) The normal distribution of response variable(Y) values corresponding to a particular value of explanatory variable(X) has standard deviation ??{Y|X}. That standard deviation is usually assumed to be the same for all values of X so that we may write ??{Y |X} = ??.

Reference: http://www.public.iastate.edu/~dnett/S401/wreginf.pdf

EC(b).

Why can an R^2 close to one not be used as evidence that the simple linear regression model is appropriate?

The R^2 value of 0.99 indicates that 99 times out of a 100, the plot will tell more of the story than a simple summary. Though the R^2 value is close to one, the regression line can over and under-predict the data (bias) at different points along the curve. This indicates a bad fit, and serves as a reminder to always check the residual plots. These biases can occur when the linear model is missing important predictors, polynomial terms, and interaction terms. This is called specification bias, and it is caused by an underspecified model. For this type of bias, the residuals can be fixed by adding the proper terms to the model. Thus, if an R-squared value is close to one, then the model is over-fitting the data, which means it cannot be used as evidence that the simple linear regression model is appropriate.

Reference: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-asses

EC(c).

Consider a regression of weight on height for a sample of adult males. Suppose the intercept is 5 kg. Does this imply that males of height 0 weigh 5 kg, on average? Would this imply that the simple linear regression model is meaningless?

We cannot imply that males of height 0 weigh 5kg, as we would not actually shift up or down the fitted line by a full meter. A regression of weight on height for a sample just explains the relationship between weight and height. Thus it is meaningful only within the range of normal weight and height of adult males. For this same reason, we cannot imply that the simple linear regression modle is meaningless.

EC(d).

Suppose you had data on pairs (X, Y) which gave the scatterplot been below. How would you approach the analysis?

As a first step, a closer look at the original data is to be made. This would help me to understand if there is any causal relationship between the explanatory and response variables. Also, once after I know the data, I would swap the x and y axis variables to check if there is any other distinct pattern. The correlation betweem both the variables could be checked to understand the relationship between them. Any outliers could be removed from the data and fitted against linear model. As we could see two groups in the plot, I would check if there is any grouping among the explanatory variables. In that case, the variables could be splitted into two different groups to check for the other patterns/further analysis.

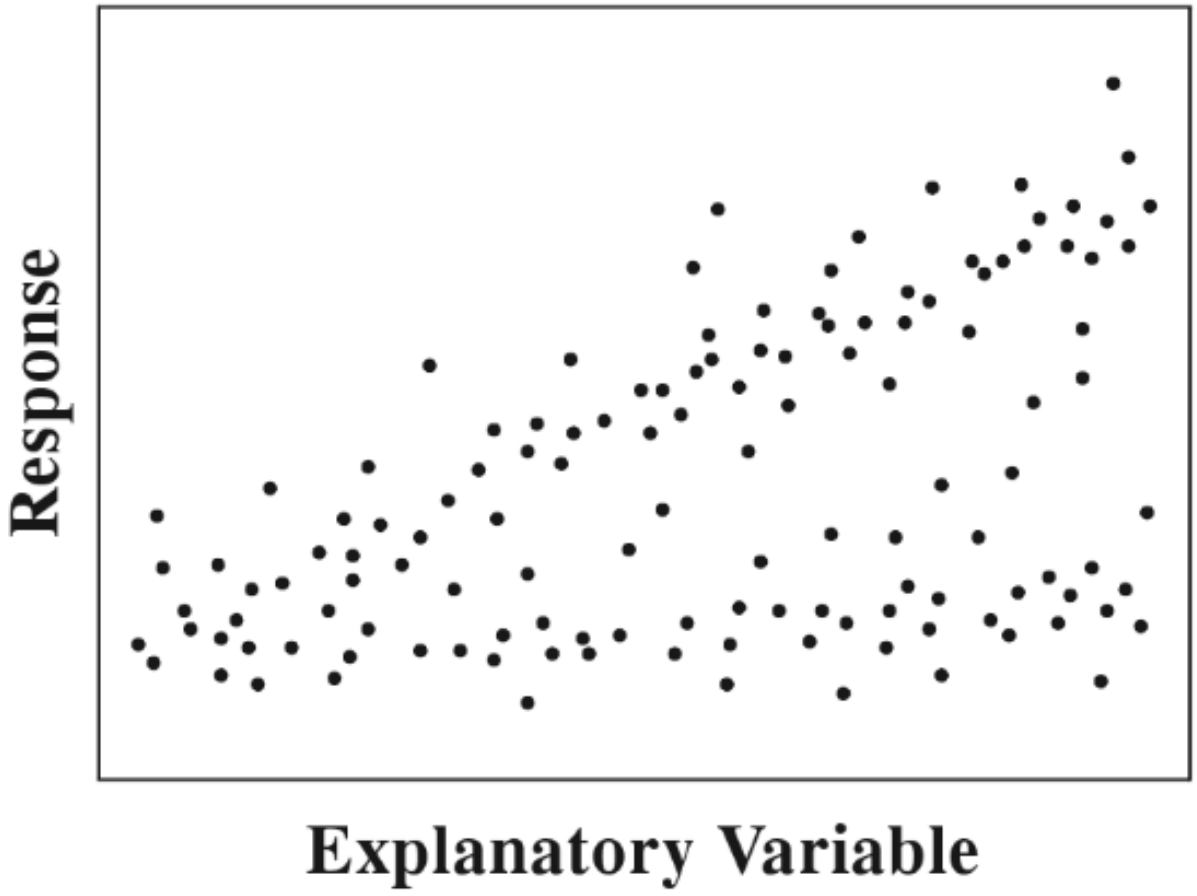


Figure 1: Scatterplot