INFX 573: Problem Set 5 - Learning from Data

Elton Sequeira

Due: Tuesday, November 8, 2016

Collaborators: Rajat Sethi, Sanath Kumar, Akshay Singh

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.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her 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`, rename the R Markdown file to `YourLastName_YourFirstName_ps5.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

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

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:

```
#Add data into a data frame

birthTable_df <- tbl_df(ex0724)
```

   (a) 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.

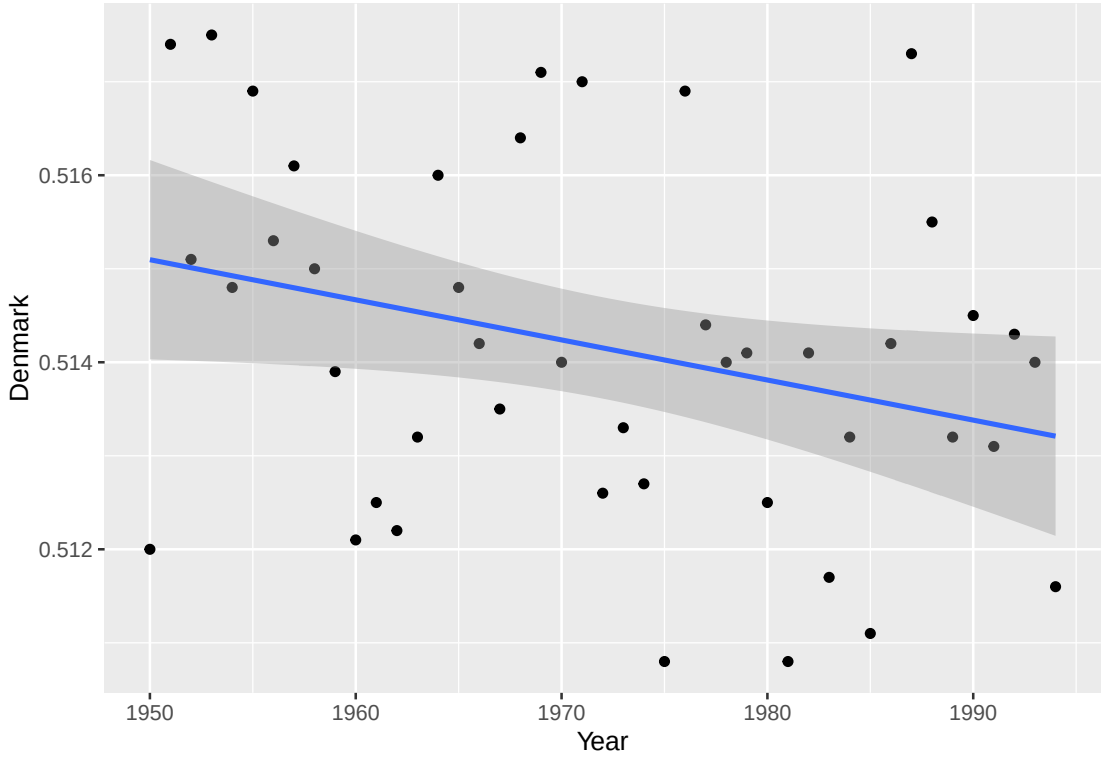
```r
attach(birthTable_df)

#lm model for Denmark
denmarkProportion <- lm(Denmark ~ Year)

denmarkProportion
```

```
##
## Call:
## lm(formula = Denmark ~ Year)
##
## Coefficients:
## (Intercept)         Year
##   5.987e-01   -4.289e-05
```

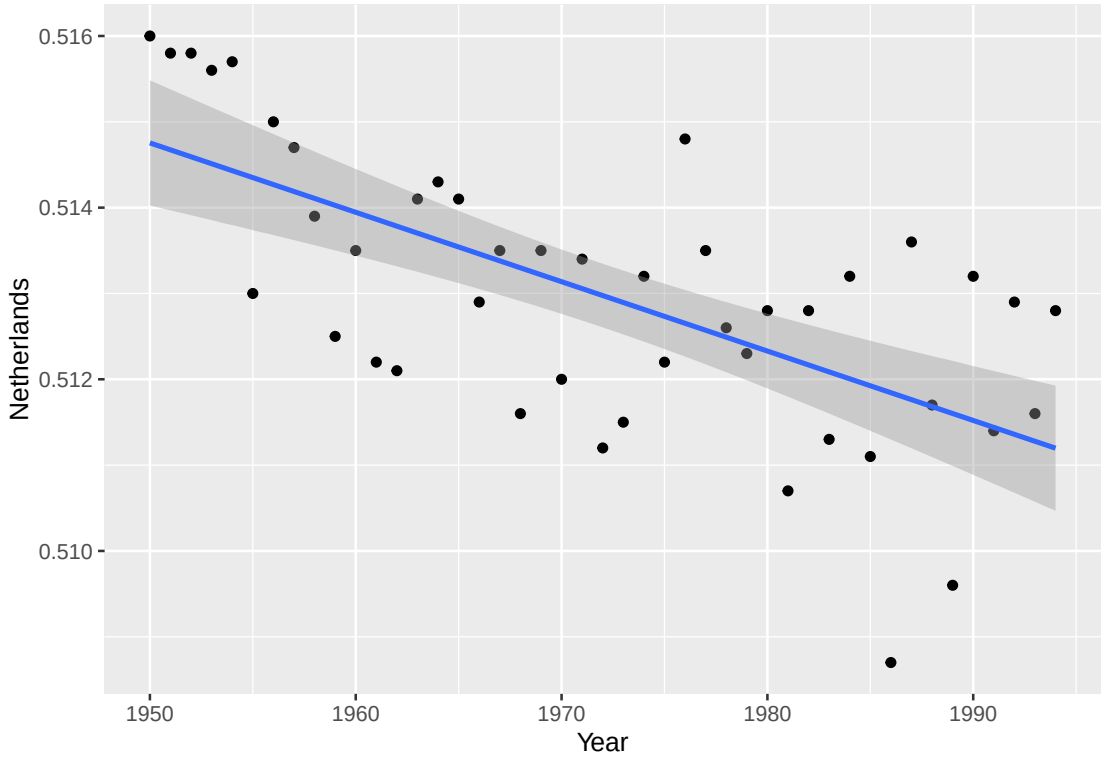
```r
ggplot(denmarkProportion, aes(x=Year, y = Denmark)) + geom_point() + geom_smooth(method = lm)
```




```r
#lm model for Netherlands
nethlandsProportion <- lm(Netherlands ~ Year)
nethlandsProportion
```

```
##
## Call:
## lm(formula = Netherlands ~ Year)
##
## Coefficients:
## (Intercept)         Year
##   6.724e-01   -8.084e-05
```

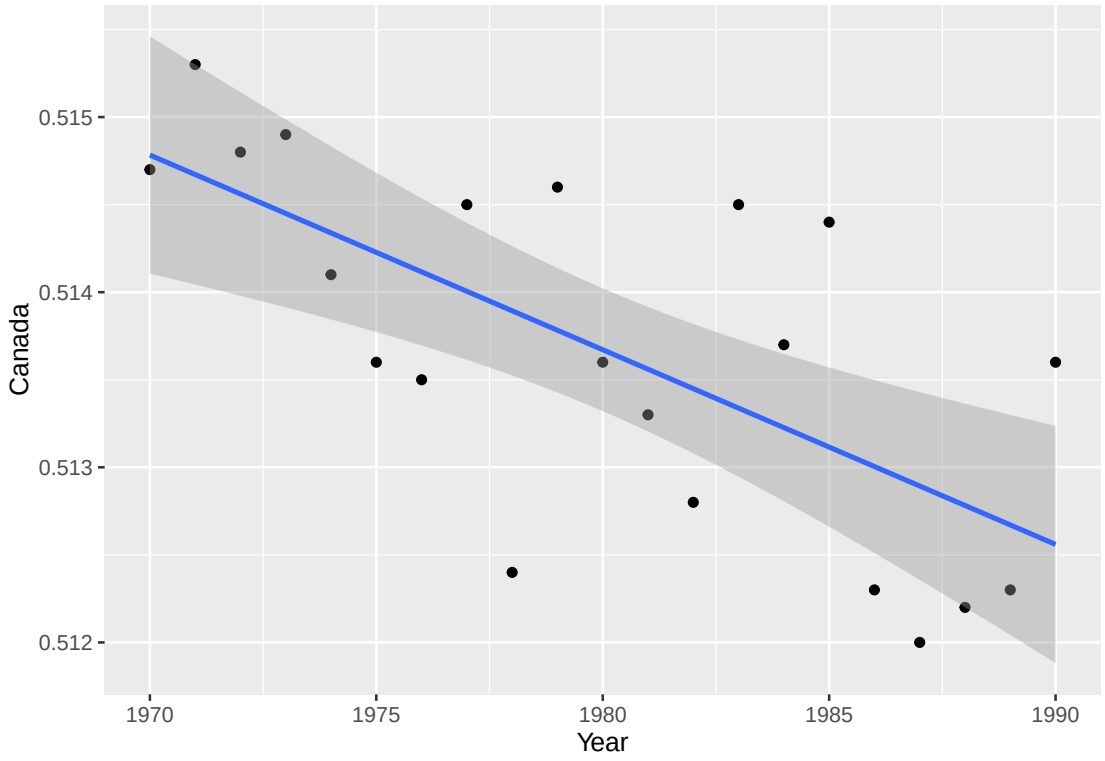
```
ggplot(nethlandsProportion, aes(x=Year, y = Netherlands)) + geom_point() + geom_smooth(method =
```




```
#lm model for Canada
canadaProportion <- lm(Canada ~ Year)
canadaProportion
```

```
##
## Call:
## lm(formula = Canada ~ Year)
##
## Coefficients:
## (Intercept)         Year
##   0.7337857   -0.0001112
```

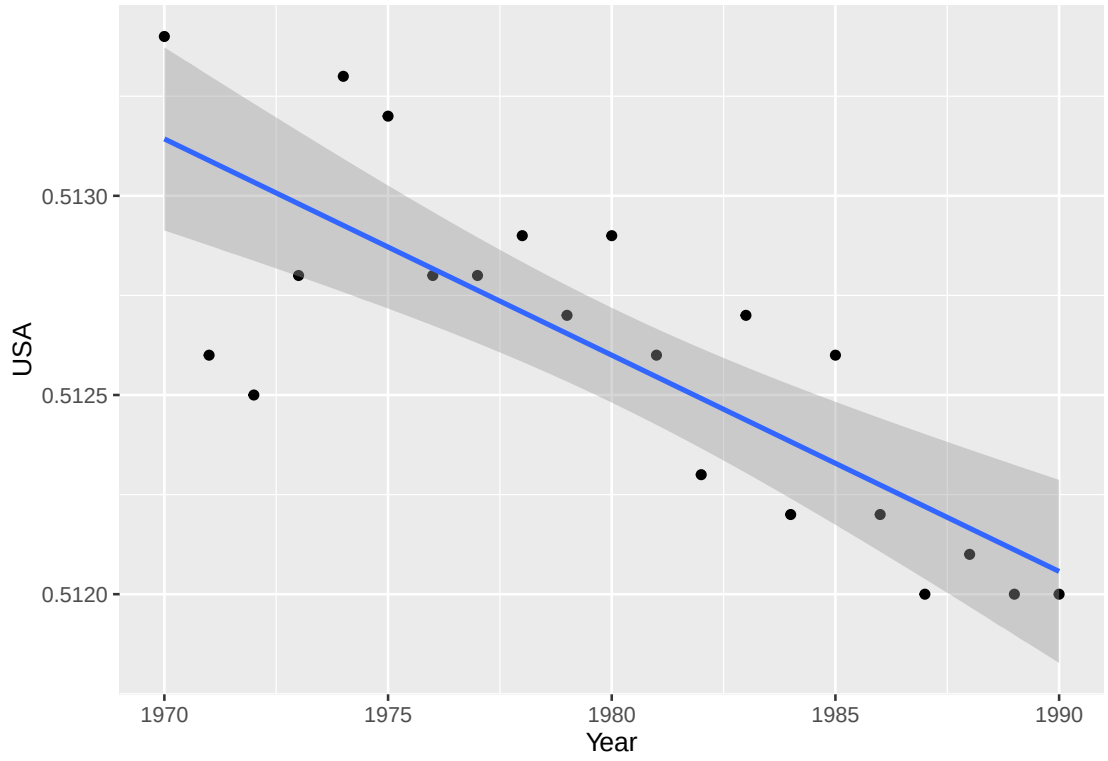
```
ggplot(canadaProportion, aes(x=Year, y = Canada)) + geom_point() + geom_smooth(method = lm)
```




```
#lm model for USA
usaProportion <- lm(USA ~ Year)
usaProportion
```

```
##
## Call:
## lm(formula = USA ~ Year)
##
## Coefficients:
## (Intercept)         Year
##   6.201e-01   -5.429e-05
```

```
ggplot(usaProportion, aes(x=Year, y = USA)) + geom_point() + geom_smooth(method = lm)
```



Estimated linear model for each of the countires
birthProportionDenmark = (-4.289e-05) Year + 5.987e-01 birthProportionNethrlands = (-8.084e-05) Year + 6.724e-01 birthProportionCanada = (-0.000111)Year + 0.7337857 birthProportionUSA = (-5.429e-05)Year + 6.201e-01

(b) 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?

```
summary(denmarkProportion)
```

```
##
## Call:
## lm(formula = Denmark ~ Year)
##
## 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
```

```r
summary(nethlandsProportion)
```

```
##
## Call:
## lm(formula = Netherlands ~ Year)
##
## 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
```

```r
summary(canadaProportion)
```

```
##
## Call:
## lm(formula = Canada ~ Year)
##
## 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
```

```r
summary(usaProportion)
```

```
##
## Call:
## lm(formula = USA ~ Year)
##
## 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
```

1.439e-05 P-values denmark: 0.0442 netherlands: 9.637e-07 Canada:0.0007376 USA: 1.439e-05
For the null hypothesis to be true (for the birthrate to stay constant), the p value is a lot less that 0.05 which indicates that the birth rates are declining in each of the 4 countries.

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: xx

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

(a) 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.

```
str(heightData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1078 obs. of  2 variables:
##  $ fheight: num  65 63.3 65 65.8 61.1 ...
##  $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
```

```
dim(heightData)
```

```
## [1] 1078    2
```

There are 1078 rows of height data with two columns that represent the heights of father and son. The data type for all columns in this data set is num.
Lets view the distribution and the min and max values

```
summary(heightData)
```

```
##     fheight         sheight
##  Min.   :59.01   Min.   :58.51
##  1st Qu.:65.79   1st Qu.:66.93
##  Median :67.77   Median :68.62
##  Mean   :67.69   Mean   :68.68
##  3rd Qu.:69.60   3rd Qu.:70.47
##  Max.   :75.43   Max.   :78.36
```

```
sd(heightData$fheight)
```

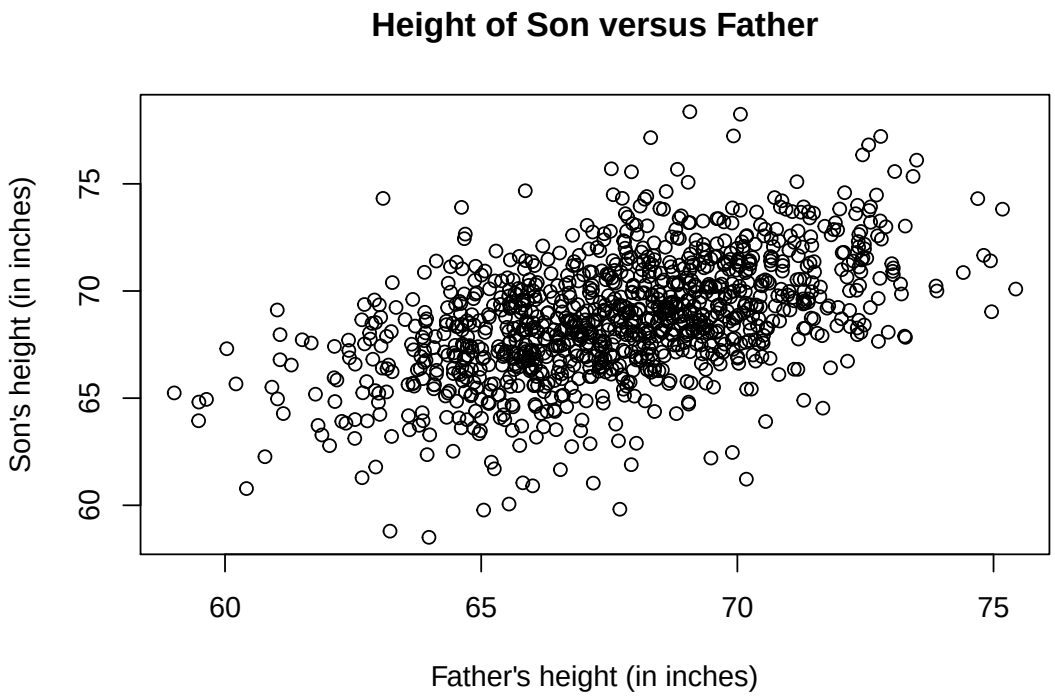
```
## [1] 2.744868
```

```r
sd(heightData$sheight)
```

```
## [1] 2.814702
```

From the results we observe the following. THe mean of father heigh is 67.69 and the mean of son is 68.68. The standard deviation of son data is greater than father data.

```r
#Plotting sons' heights against fathers' heights
plot(heightData$sheight ~ heightData$fheight
     , main = "Height of Son versus Father"
     , ylab = "Son's height (in inches)"
     , xlab = "Father's height (in inches)")
```



The visualization suggests that there is an increase in son height with an incrase in father height. We shall calculated the correlation between the father and son variables.

```r
cor.test(heightData$fheight, heightData$sheight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  heightData$fheight and heightData$sheight
## t = 19.006, df = 1076, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4552586 0.5447396
## sample estimates:
##       cor
## 0.5013383
```

From the correlation test we can see that the value is 0.5013383 which indicates that there is a correlation between fathers height and sons height.

(b) 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.

```
#Fit a linear regression model for the height dataset

attach(heightData)
lmHeight <- lm(sheight ~ fheight)
lmHeight
```

```
##
## Call:
## lm(formula = sheight ~ fheight)
##
## Coefficients:
## (Intercept)      fheight
##     33.8866       0.5141
```

The coeffecient is : 0.5141 Intercept is : 33.8866

The linear regression model is: sheight = (0.5141)fheight + 33.8866.

The coeffcients suggest that for every unit increase of father height the son height increases by a factor of 0.5141 of the father height and an addtion of 33.8866.

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

```
#Find the confidence intervals for the estimates
confint(lmHeight,level = 0.95)
```

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

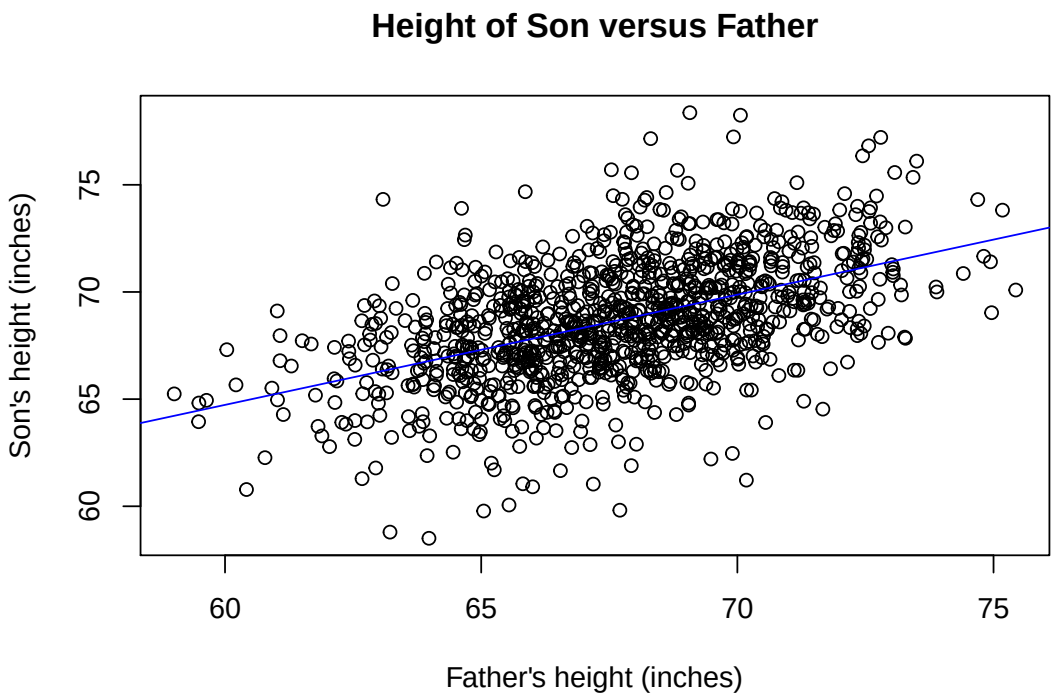
Confidence intervals of 95 % is the intervals between which 95% of the obsevations lie within.

From the confidence intervals estimate we can say that 95% of observations lie between.

sheight = (0.4610188)fheight + 30.2912126 and sheight = (0.5671673)fheight + 37.4819961

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

```
#Plotting sons vs father heights
plot(heightData$sheight ~ heightData$fheight
     , main = "Height of Son versus Father"
     , ylab = "Son's height (inches)"
     , xlab = "Father's height (inches)")
#Drawing the least squares regression line
abline(lmHeight, col = "blue")
```



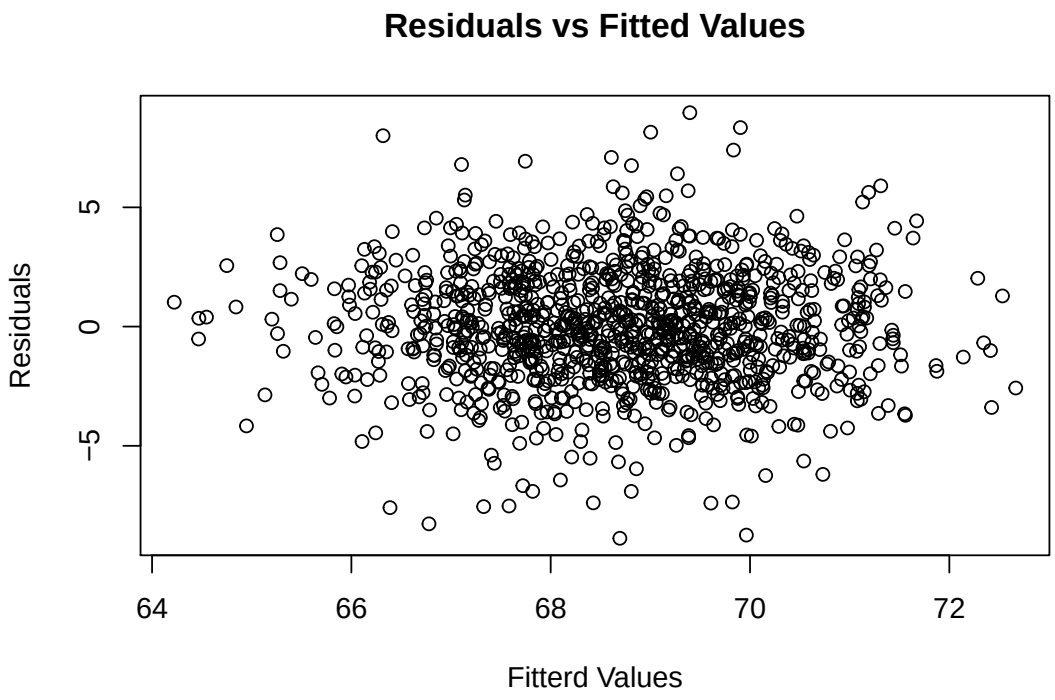
(e) Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using `names()`). Discuss what you see. Do you have any concerns about the linear model?

```
#Inspect tnames of the elements of the linear model
names(lmHeight)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Use the residuals and fitter.values to produce a visualization.

```
plot(lmHeight$residuals ~ lmHeight$fitted.values,
     main="Residuals vs Fitted Values",ylab ="Residuals",xlab="Fitterd Values")
```



The plot shows that the residuals are symmetrically distributed around 0 but there is no noticeable relation between the fitted values and the residuals. Let's take a corrlation test to obsever the relation between fitted values and residuals.

```
 #Correlation test for fitted variables and residuals
cor.test(lmHeight$residuals,lmHeight$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  lmHeight$residuals and lmHeight$fitted.values
## t = 5.4346e-14, df = 1076, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05970724  0.05970724
## sample estimates:
##          cor
## 1.656776e-15
```

THe p value is 1 which indicates that the null hypothesis is true for all cases and there is absolutely no relation between residuals and fitted values.

(f) 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.

```
#Add the data set to aa vector
fatherHeight <- data.frame(fheight <-c(50,55,70,75,90))
```

Now predict values according to the linear model by using the predict function

```r
predict(lmHeight,fatherHeight,interval = "predict")
```

```
##        fit      lwr      upr
## 1 59.59126 54.71685 64.46566
## 2 62.16172 57.33140 66.99204
## 3 69.87312 65.08839 74.65785
## 4 72.44358 67.64470 77.24246
## 5 80.15498 75.22740 85.08255
```

The result is a table which shows the fitted values for sons height correspoinding to the fathers heights. It also shows the upper and lower values which falls within the confidence level of our linear model. For example for father with height 50, the sons height will be 59.59126 as fitted by the value and the confidence levels for sons heights will lie between 54.71685 and 64.46566.

3. **Extra Credit:**

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

Looking at the distribution we cannot make any assumptions of the values the explanatory variable can take. It can be any numberical value. The explanatory variable can vary vastly with even a small change in response variable and that makes it difficult to make any assumptions about the distribution.

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

A R^2 value close to one makes a better fit for a linear model. However if there are certain biases in the dataset it can cause the regression line to be curved. but the linear regression will not give the same picture as this specific dataset and it will either over estimate or under estimated the model in order to maintain the linearity. Hence for this case even a R^2 value close to 1 cannot be use as evidence that the simple linear regression model is appropriate.

(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?

The regression line extends in both direction infinitely and is a apt fit for predictinig values. The line will have a 0 intercept but in a practical sense there would not be a value of 0. For example in the example we worked on there would never be a value of 0 as height. But for a s pecific range of values the linear regression model would definitley be useful in predicting outcomes and this means that the model is not meaningless.

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

I would first begin ananlysis by observation. From observation is seems that there is no deductible relationship between explanatory variable and response variable and hence there would need to calculate correlation and create a linear regression model.

I would then calculate the correlatin between the two variables to see the strength of the relationship. Depending on the correlation we can either accept or reject the null hypothesis which means there is no relation between the two variables.

In addtion I would try and fit a linear model to see and recognize any pattern in the values between the two variables. Over this I will use a plot between the residuals and fitted values of the model to deduct if the linear model is a good fit.

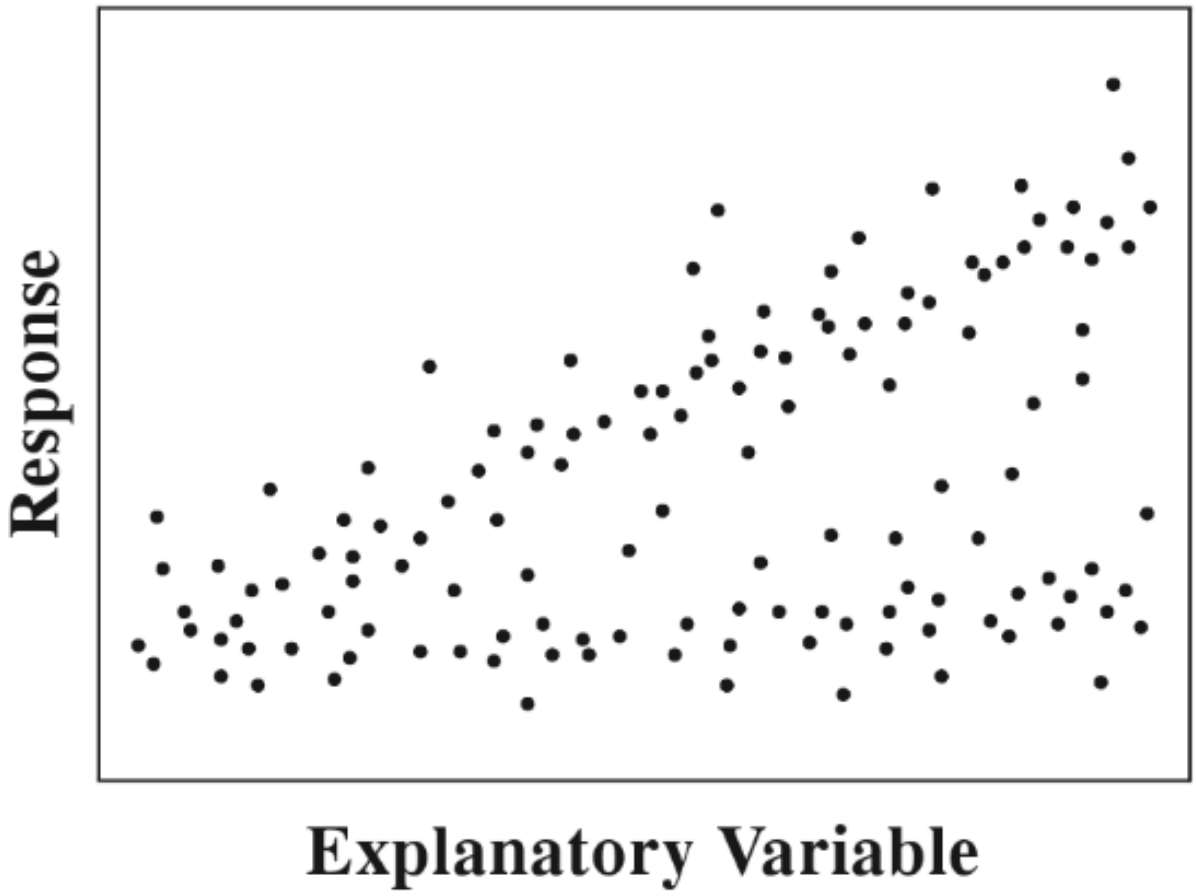


Figure 1: Scatterplot for Extra Credit (d).