INFX 573: Problem Set 5 - Learning from Data

Aakash Bang

Due: Tuesday, November 8, 2016

Collaborators:

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
```

```
Male_Births <- Sleuth3::ex0724
```

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:

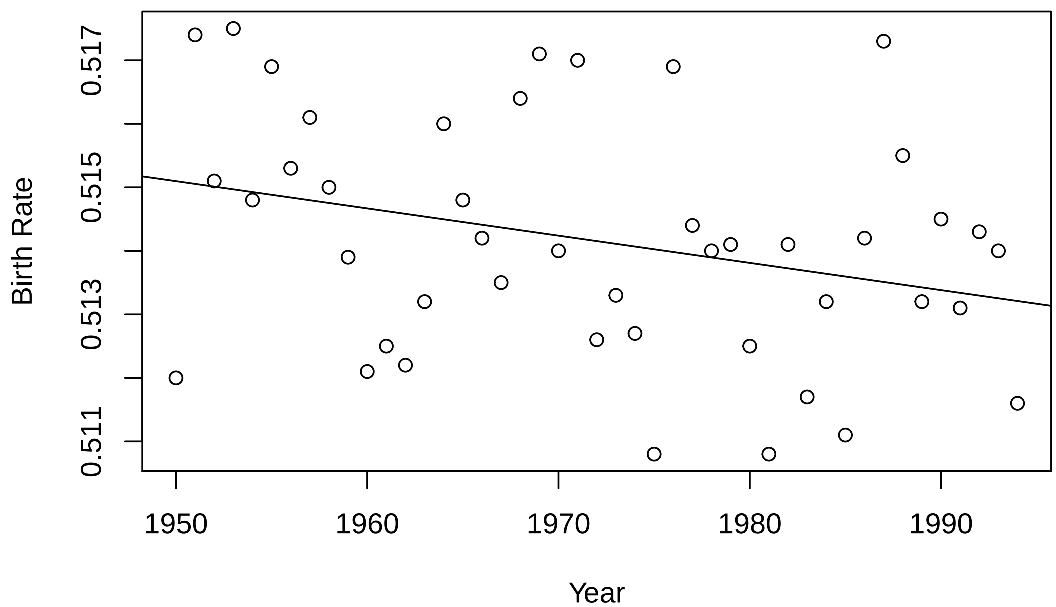
   (a) Use the `lm` function in **R** to fit four (one per country) simple linear regression models of the yearly proportion of males births as a function of the year and obtain the least squares fits. Write down the estimated linear model for each country.

   ```
   # Regression Line for Denmark
   fit_denmark <- lm(formula = Denmark ~ Year, data = Male_Births)
   fit_denmark
   ```

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

```
## Coefficients:
## (Intercept)         Year
##   5.987e-01   -4.289e-05
```

```r
plot(Male_Births$Denmark ~ Male_Births$Year, xlab = "Year", ylab = "Birth Rate")
abline(fit_denmark)
```

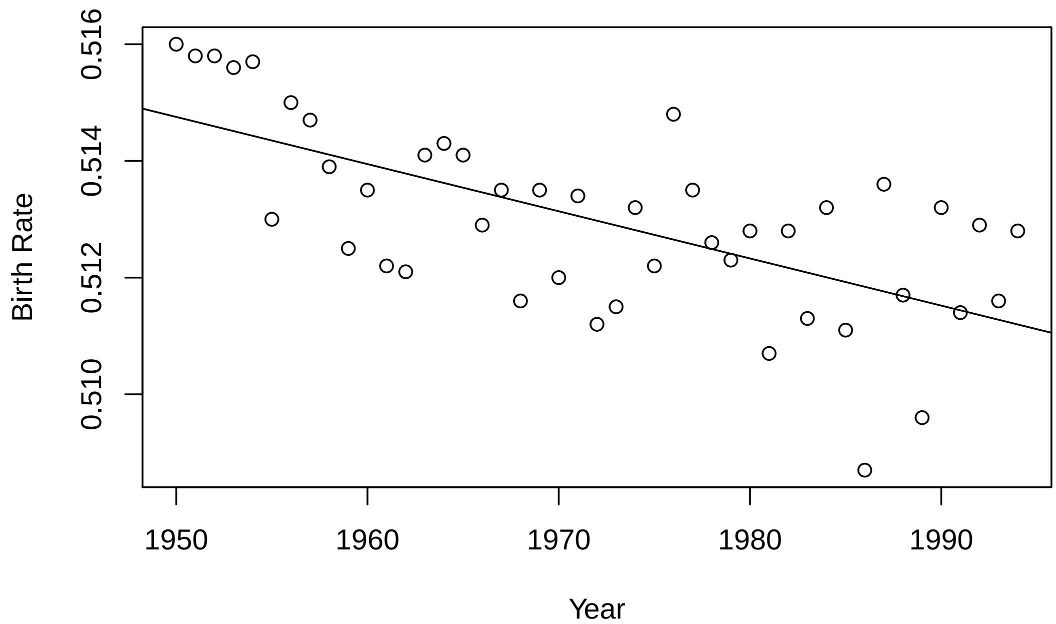


birthrateDenmark = -4.289e-05*year + 5.987e-01

```r
# Regression Line for Netherlands
fit_Netherlands <- lm(formula = Netherlands ~ Year, data = Male_Births)
fit_Netherlands
```

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

```r
plot(Male_Births$Netherlands ~ Male_Births$Year, xlab = "Year", ylab = "Birth Rate")
abline(fit_Netherlands)
```

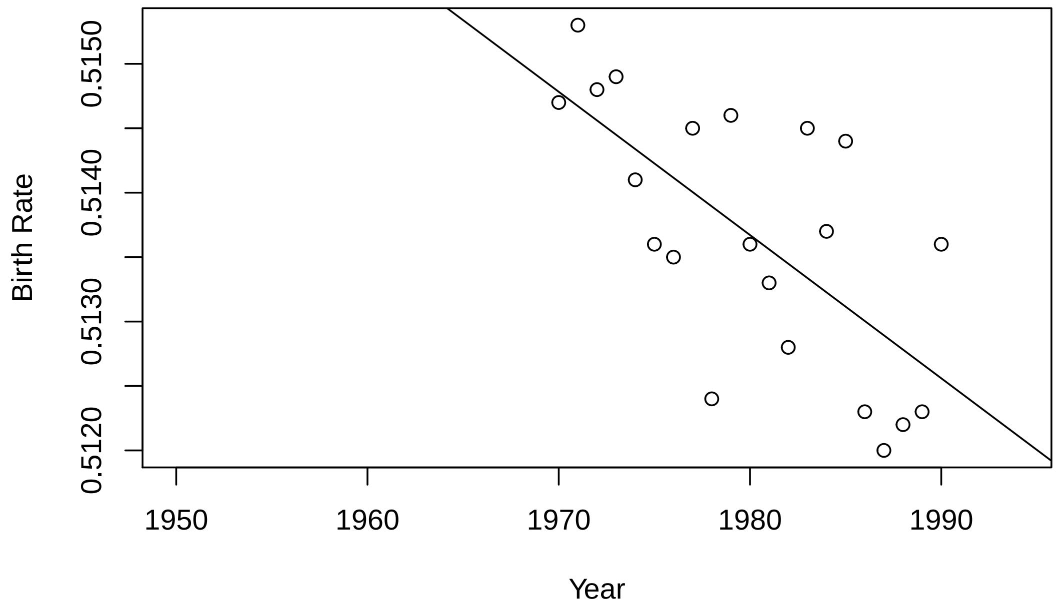


birthrateNetherlands = -8.084e-05*year + 6.724e-01

```r
# Regression Line for Canada
fit_Canada <- lm(formula = Canada ~ Year, data = Male_Births)
fit_Canada
```

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

```r
plot(Male_Births$Canada ~ Male_Births$Year, xlab = "Year", ylab = "Birth Rate")
abline(fit_Canada)
```

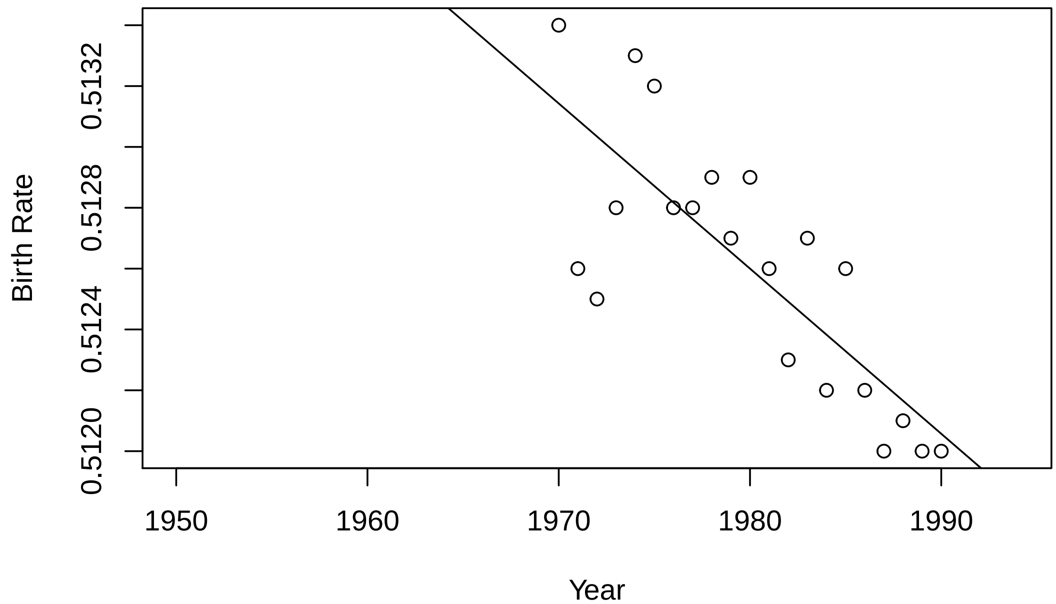


birthrateCanada = -0.0001112*year + 0.7337857

```r
# Regression Line for USA
fit_USA <- lm(formula = USA ~ Year, data = Male_Births)
fit_USA
```

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

```r
plot(Male_Births$USA ~ Male_Births$Year, xlab = "Year", ylab = "Birth Rate")
abline(fit_USA)
```



birthrateUSA = -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?

```r
#t-statistic for Denmark
summary(fit_denmark)
```

```
##
## Call:
## lm(formula = Denmark ~ Year, data = Male_Births)
##
## 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
#t-value = -2.073, p-value = 0.0442
```

```r
#t-statistic for Netherlands
summary(fit_Netherlands)
```

```
##
## Call:
## lm(formula = Netherlands ~ Year, data = Male_Births)
##
## 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
#t-value = -5.71, p-value = 9.64e-07
```

```r
#t-statistic for Canada
summary(fit_Canada)
```

```
##
## Call:
## lm(formula = Canada ~ Year, data = Male_Births)
##
## 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
#t-value = -4.017, p-value = 0.000738
```

```r
#t-statistic for USA
summary(fit_USA)
```

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

```
## 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
```

```r
#t-value = -5.779, p-value = 1.44e-05
```

The t-statistic are more extreme than -2 and +2, and p-values are less than 0.05, suggesting that we can reject the null hypothesis (male births are not declining). In other words, male birth is declining over this period.

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:

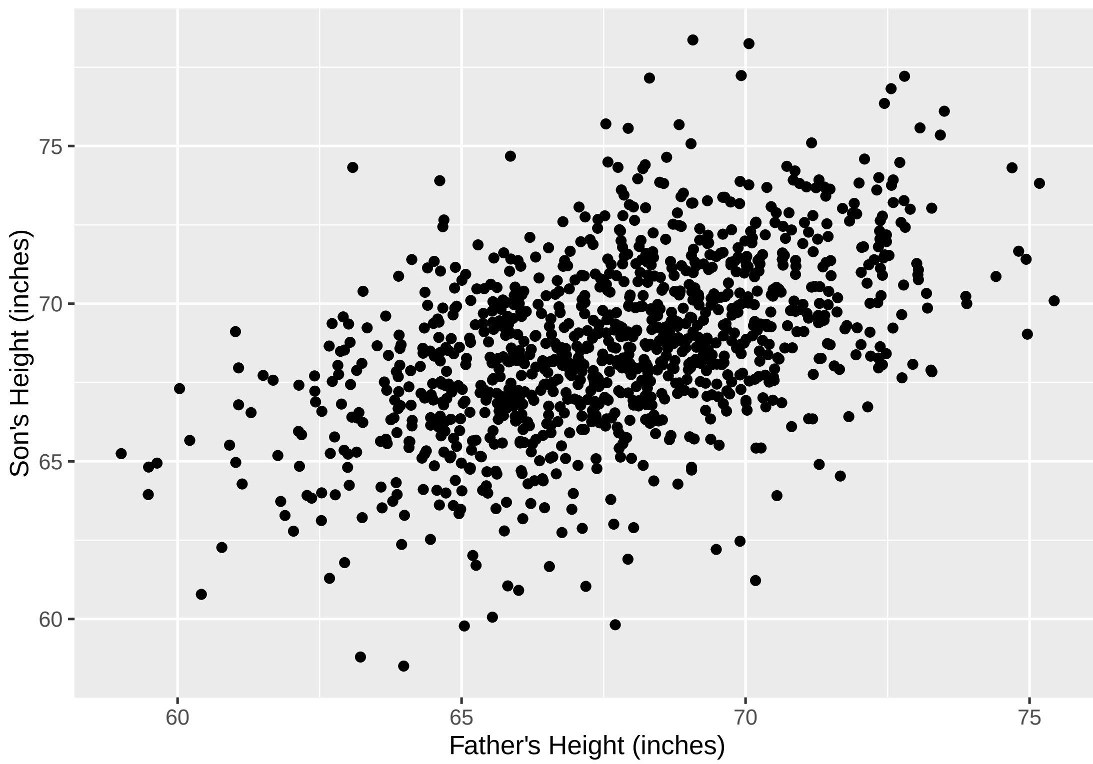
```r
# Import and look at the height data
heightData <- tbl_df(get("father.son"))
```

   (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.

```r
#statistical summary of the variables
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
```

```r
#visualization
ggplot(heightData, aes(fheight, sheight))+geom_point() +
labs(x = "Father's Height (inches)", y = "Son's Height (inches)")
```



It is interesting to see that as the father's height goes up, the son's height has a higher tendency to go up too.

(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 simple linear regression model
fit_sonHeight = lm(formula = sheight ~ fheight, data = heightData)
summary(fit_sonHeight)
```

```
##
## 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
```

```
fit_sonHeight
```

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

As per the summary, we can write the equation for the linear model as: sheight = 0.5141*fheight + 33.8866

For this model, the regression coefficient beta_i indicates the change in the response variable (sheight) for one unit change in the predictor variable (fheight). Thus, for an increase of 1 inch to the father's height we can expect an increase of 0.5141 inch in the son's height. It is important to note that this equation holds for only the range of values of the data has been collected.

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

```
#Find the 95% confidence interval

confint(fit_sonHeight, level = 0.95)
```

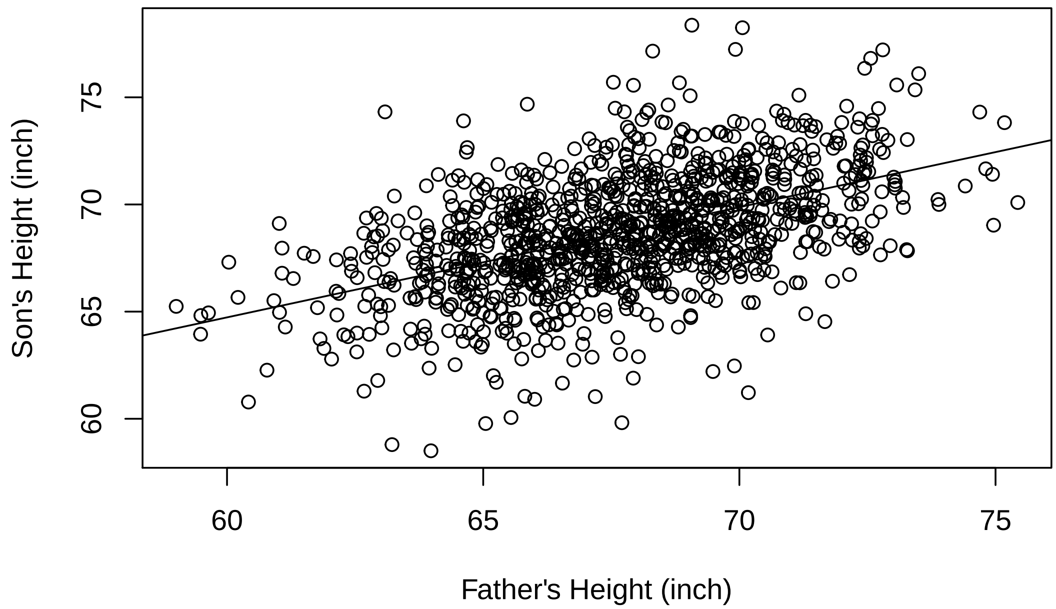
```
##                  2.5 %     97.5 %
## (Intercept) 30.2912126 37.4819961
## fheight      0.4610188  0.5671673
```

We can say with 95% confidence that beta_a lies between 0.4610188 and 0.5671673

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

```
# visualize relationship between fheight and sheight
plot(heightData$fheight, heightData$sheight,
xlab = "Father's Height (inch)", ylab = "Son's Height (inch)" )

# draw the least squares regression line
abline(fit_sonHeight)
```



(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?

```r
names(fit_sonHeight)
```

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

```r
residuals(fit_sonHeight)
```

```
##            1            2            3            4            5
## -7.549320518 -3.189432294 -3.937262188 -4.897126876 -1.035698698
##            6            7            8            9           10
## -2.043843444 -3.410828758 -3.165011904 -3.236827219 -4.334628227
##           11           12           13           14           15
##  1.022303623 -0.888502884 -0.939312002 -1.389889738 -1.848607024
##           16           17           18           19           20
## -2.591991494 -2.989964076 -2.052904112 -3.438394247 -2.717010607
##           21           22           23           24           25
## -3.605699113 -4.121340976  0.546305037 -0.312543404 -0.875422298
##           26           27           28           29           30
## -1.312839748 -1.192957249 -1.230181614 -1.812453255 -1.615901663
##           31           32           33           34           35
## -2.375460072 -2.204042616 -1.619202118 -3.046443067 -2.648375866
##           36           37           38           39           40
## -2.626129125 -2.805758011 -3.212520458 -4.390581934  1.251267433
##           41           42           43           44           45
##  0.423048888 -0.045610592  0.017480915 -0.497696232 -0.474796414
```

```
##          46           47           48           49           50 
##  1.651610444 -0.093493379 -6.433844280 -1.501642356 -0.690917366 
##          51           52           53           54           55 
##  0.167196487 -0.296185711 -0.663097904 -0.437084264  0.035412008 
##          56           57           58           59           60 
## -0.270065897 -0.778626280 -0.794462663 -0.920354365 -1.725711657 
##          61           62           63           64           65 
## -1.230643968 -0.620670541 -1.364184860 -1.572385589 -1.470870305 
##          66           67           68           69           70 
## -1.513833405 -1.745407337 -2.069684088 -2.846581324 -3.329580662 
##          71           72           73           74           75 
## -2.946462676 -3.640587809  1.860225243  2.281611743  1.304137765 
##          76           77           78           79           80 
##  1.278954303  1.282886041  0.483236276  0.464206224  0.684499731 
##          81           82           83           84           85 
##  0.541204255  0.635911455 -0.167036297  0.447075292 -0.375922595 
##          86           87           88           89           90 
## -0.216290319  0.116000909 -0.540760234 -0.514137299 -0.965763441 
##          91           92           93           94           95 
## -0.408536148 -1.177500783 -0.045185671 -1.231809237 -1.256862315 
##          96           97           98           99          100 
## -1.283594057 -0.807166782 -1.300360878 -0.583115367 -1.592844342 
##         101          102          103          104          105 
## -1.694391204 -2.726917798 -1.696968809 -2.808427293 -3.721415968 
##         106          107          108          109          110 
##  2.481860301  2.991852489  2.191823951  2.316841460  1.663529726 
##         111          112          113          114          115 
##  1.509323610  1.810568992  1.750030102  1.182320365  1.412419651 
##         116          117          118          119          120 
##  0.593792097  0.609404304  0.130282046  0.484382391  0.873493002 
##         121          122          123          124          125 
##  0.234369977 -0.507444670  0.256542544  0.113457481  0.050642153 
##         126          127          128          129          130 
## -0.456481594 -0.714303903 -1.006836579 -0.642387828 -0.821788545 
##         131          132          133          134          135 
## -0.904210026 -1.227948506 -1.903205886 -3.391252138  2.614688577 
##         136          137          138          139          140 
##  2.476442442  2.256267373  2.481986131  2.739712775  1.972221011 
##         141          142          143          144          145 
##  1.206614827  1.720636825  0.898841785  0.733899166  1.800290795 
##         146          147          148          149          150 
##  0.561337916  0.719309646  0.393519322  0.726053610  0.182004139 
##         151          152          153          154          155 
##  0.015308736  0.133562580 -0.366483521 -0.228146526 -0.531668437 
##         156          157          158          159          160 
## -0.322913060 -0.986682154 -2.575892643  4.288392671  2.811879391 
##         161          162          163          164          165 
##  3.854350538  2.964102656  3.377198232  2.462113473  2.147711164 
##         166          167          168          169          170 
##  2.500735498  1.712579696  1.766511791  1.668353743  1.576386544 
##         171          172          173          174          175 
##  1.530073205  1.281262567  1.039887763  1.046845499  1.162054402 
##         176          177          178          179          180 
##  0.026946632 -0.515032586 -0.680480615  3.857967348  3.386384510 
```

```
##         181         182         183         184         185 
##  3.613236253  2.639723729  3.303625096  2.120430141  2.598257226 
##         186         187         188         189         190 
##  2.241349711  1.826296299  0.819806704  1.033570797  4.187133519 
##         191         192         193         194         195 
##  4.073441414  3.622528817  3.353553128  2.995469909  2.616144304 
##         196         197         198         199         200 
##  2.324543094  8.003255538  5.442858656  3.865042378  3.616255354 
##         201         202         203         204         205 
##  2.560981725  6.751380784  7.093602167  5.901547869 -4.818473013 
##         206         207         208         209         210 
##  5.339813720  3.591276323  3.711810584 -4.456682821 -2.036181060 
##         211         212         213         214         215 
## -7.357182705  3.280337809 -4.094563645 -7.402425332 -4.546391439 
##         216         217         218         219         220 
## -3.547572138 -2.996731758 -3.058385489 -4.114599078 -5.963614403 
##         221         222         223         224         225 
## -1.937790543 -2.947703222 -2.850429872 -4.008435512 -3.586611265 
##         226         227         228         229         230 
## -4.981214007 -0.455548844 -1.843159299 -1.066749004 -2.356374230 
##         231         232         233         234         235 
## -2.339795194 -2.037973366 -2.089951647 -2.198188552 -3.484180192 
##         236         237         238         239         240 
## -3.311773147 -4.661443598 -0.009016003 -0.379773844 -0.055167765 
##         241         242         243         244         245 
## -0.966452762 -0.676800413 -1.302263927 -1.650971009 -1.416779793 
##         246         247         248         249         250 
## -1.019166136 -2.330428929 -1.957194664 -1.842430538 -3.005425792 
##         251         252         253         254         255 
## -2.799590649 -2.524552853 -3.378735919 -4.134563822  0.902462882 
##         256         257         258         259         260 
##  0.600731530  0.536295550  0.953143203 -0.014966588 -0.237475177 
##         261         262         263         264         265 
##  0.265131054 -0.228149267 -5.391354890 -2.849681116 -1.850511525 
##         266         267         268         269         270 
## -3.589695641 -0.102142686 -1.043803584 -0.666873710 -0.613530122 
##         271         272         273         274         275 
## -0.782423016 -0.765505802 -0.981290464 -1.070444318 -0.568468247 
##         276         277         278         279         280 
## -1.610169354 -1.765660462 -1.642620503 -1.486344337 -2.148516404 
##         281         282         283         284         285 
## -2.177641887 -1.785823380 -2.458158060 -2.459333375 -3.180891716 
##         286         287         288         289         290 
## -2.358938099 -3.344303343  1.737870221  1.529971642  2.126584122 
##         291         292         293         294         295 
##  1.837056693  1.339263165  1.180360089  0.674466681  0.463427763 
##         296         297         298         299         300 
## -0.123244601  0.742131310 -0.293302886 -0.188462701  0.335081797 
##         301         302         303         304         305 
##  0.145177756 -0.264314883 -0.427049864  0.084303897 -0.927216569 
##         306         307         308         309         310 
## -0.791968391 -1.187685475 -0.234938274 -1.272797267 -1.642547450 
##         311         312         313         314         315 
## -0.822138191 -0.732354563 -1.497519308 -1.531592328 -2.319124873 
```

```
##         316         317         318         319         320 
## -1.420630068 -1.632832247 -1.879364631 -3.047175858 -2.731998400 
##         321         322         323         324         325 
##  3.238853634  2.267149069  2.325138632  1.685806236  1.877961690 
##         326         327         328         329         330 
##  1.616795594  0.998786670  1.879586855  1.013560311  1.266195138 
##         331         332         333         334         335 
##  0.417645158  0.461415500  0.255616549  0.157141033  0.998830954 
##         336         337         338         339         340 
## -0.265556588  0.105104884 -0.294442609 -0.151975195 -0.122545542 
##         341         342         343         344         345 
##  0.047138344 -0.187101821 -0.329330033 -1.375334272 -1.103788678 
##         346         347         348         349         350 
## -0.655542638 -1.121097939 -1.017182163 -1.979390319  2.955116189 
##         351         352         353         354         355 
##  2.662128458  1.985806920  1.903415364  2.163431715  1.717846883 
##         356         357         358         359         360 
##  2.148443709  0.821629353  1.132662390  1.446649298  1.690406308 
##         361         362         363         364         365 
##  1.158488113  0.988784179  1.243774578  0.607180922  0.101055811 
##         366         367         368         369         370 
##  0.542196241  0.177075898  0.517863104 -0.564554608  0.047002321 
##         371         372         373         374         375 
## -0.653364255 -0.963612058 -1.659336497  4.138422805  3.552991910 
##         376         377         378         379         380 
##  3.349257552  2.174165598  2.560942560  2.337009107  2.182338515 
##         381         382         383         384         385 
##  1.917710543  1.828863893  1.772559456  1.904477732  1.488699510 
##         386         387         388         389         390 
##  1.486941263  1.954373022  1.111390949  0.859528685  1.197521578 
##         391         392         393         394         395 
##  0.358195739  0.240338195 -0.148113758  4.182020693  3.565863898 
##         396         397         398         399         400 
##  3.250366427  2.324158514  2.389633338  2.537895513  2.374891897 
##         401         402         403         404         405 
##  2.617766937  1.520946125  0.864434305  0.729090171  4.699917114 
##         406         407         408         409         410 
##  4.672887051  3.795810853  3.707780237  3.487485465  2.732503400 
##         411         412         413         414         415 
##  1.666374075  1.963024008  5.064569422  5.484583810  3.387810643 
##         416         417         418         419         420 
##  3.637810778  2.024863332  4.629300349  8.151747800 -4.165330595 
##         421         422         423         424         425 
## -4.397042931  0.092183980 -8.271382039 -1.875112825  0.074072255 
##         426         427         428         429         430 
## -0.294812949  2.968543040 -1.649808898 -5.637762696  1.390831221 
##         431         432         433         434         435 
##  0.549083220 -4.497157752 -3.497778992 -4.525255176 -5.520827464 
##         436         437         438         439         440 
## -1.986475485 -2.387577959 -3.302808121 -3.025704983 -3.618434519 
##         441         442         443         444         445 
## -4.864695754  0.821873405 -1.462888129 -0.933447502 -2.631248222 
##         446         447         448         449         450 
## -2.581049069 -2.088575488 -2.392377869 -2.632835083 -2.985479647 
```

```
##         451         452         453         454         455
## -3.746318585 -4.570049223  0.123878967  0.017222335 -0.636215736
##         456         457         458         459         460
## -0.343478636 -1.308177593 -0.996858657 -1.356835908 -1.482939709
##         461         462         463         464         465
## -1.045068074 -2.380873782 -2.035909345 -2.405678827 -2.577727027
##         466         467         468         469         470
## -2.282927149 -2.269134787 -2.804976822 -3.263977852  1.507905714
##         471         472         473         474         475
##  1.401090723  0.220064646 -0.064974348  0.629756131 -0.350102478
##         476         477         478         479         480
## -0.248751352 -0.711485837 -3.486091137 -4.676935637 -1.760437753
##         481         482         483         484         485
## -2.470469682 -0.068363565 -0.003059954 -0.667482049 -0.834641792
##         486         487         488         489         490
## -1.015693781 -0.740362645 -0.875387261 -1.605599758 -1.349426932
##         491         492         493         494         495
## -0.864106255 -0.919440855 -0.822352004 -1.658792381 -1.976064268
##         496         497         498         499         500
## -1.658146023 -1.891878029 -2.229115654 -1.760182271 -2.455244350
##         501         502         503         504         505
## -3.012410151 -2.582018386  2.218731817  2.286712411  1.032497054
##         506         507         508         509         510
##  1.134678004  0.918782672  1.073320623  0.796418419 -0.170741632
##         511         512         513         514         515
##  0.287647361 -0.118998035  0.014531984 -0.341609600 -0.512497595
##         516         517         518         519         520
## -0.166179153  0.058683229  0.157404465 -0.070758990 -0.434102977
##         521         522         523         524         525
## -0.193019062 -0.485859072 -0.864082612 -1.463250029 -0.804839027
##         526         527         528         529         530
## -1.232652814 -1.401964163 -1.186450226 -1.064415076 -1.499661599
##         531         532         533         534         535
## -1.203505578 -2.081020787 -1.941188075 -2.456274391 -2.494140757
##         536         537         538         539         540
##  3.857451958  3.073832008  2.427127077  1.603469118  2.417727796
##         541         542         543         544         545
##  1.823501510  1.593041702  0.940236448  0.577220003  0.910136172
##         546         547         548         549         550
##  0.515233980  0.703753170 -0.233272042  0.196287715  0.100233396
##         551         552         553         554         555
##  0.476631073 -0.467795173 -0.484116104 -0.394556034  0.026199662
##         556         557         558         559         560
##  0.121192064 -0.615982756 -0.349044921 -0.111379734 -0.520127446
##         561         562         563         564         565
## -0.956518584 -0.900222426 -1.879284146 -1.571540721  3.390140321
##         566         567         568         569         570
##  2.714972594  2.529337171  2.331660578  2.185673182  2.205944965
##         571         572         573         574         575
##  1.617506217  1.841601958  0.994719840  1.925553736  1.311455510
##         576         577         578         579         580
##  0.206828320  1.042816425  0.505300846  0.523866950  0.632331719
##         581         582         583         584         585
##  0.122432249  0.605779235  0.420151602 -0.498952678 -0.565600690
```

```
##          586          587          588          589          590 
## -0.006776941 -1.034715921 -0.712179528  4.139316630  3.906986317 
##          591          592          593          594          595 
##  3.293088672  2.426219526  2.224091815  2.731735177  2.764149030 
##          596          597          598          599          600 
##  1.733662782  2.318791395  2.003560969  1.544138142  1.888508198 
##          601          602          603          604          605 
##  1.384333399  1.614683392  1.222460389  1.958252376  0.723632483 
##          606          607          608          609          610 
##  0.246105987 -0.211022623  0.359560060  5.307896020  4.380018759 
##          611          612          613          614          615 
##  3.032073323  3.081352240  2.559210884  3.158959859  1.998324951 
##          616          617          618          619          620 
##  2.155955791  2.167794730  1.924072275  0.860063659  1.295236203 
##          621          622          623          624          625 
##  4.209837674  3.804104926  4.104828835  3.690333035  3.165760595 
##          626          627          628          629          630 
##  1.968800825  1.632180526  6.933304176  4.754464919  4.110180686 
##          631          632          633          634          635 
##  2.919973210  2.715892034  6.406283771  4.432464139  8.343688769 
##          636          637          638          639          640 
## -6.912291784  2.729159120 -2.864912195 -1.004644842  1.402358037 
##          641          642          643          644          645 
##  0.310794534  4.414340456  0.593138224 -6.197892353 -0.863888343 
##          646          647          648          649          650 
## -0.960672325 -3.003959566 -7.392165742 -2.912550245 -3.631370065 
##          651          652          653          654          655 
## -4.822340049 -2.113980471 -2.389505185 -2.863227011 -3.241792826 
##          656          657          658          659          660 
## -3.660770957 -3.557003416  0.347453597 -1.046381008 -1.210875925 
##          661          662          663          664          665 
## -1.199720032 -2.198016722 -1.753317097 -2.283677937 -2.250404110 
##          666          667          668          669          670 
## -3.154485655 -4.073990691 -3.496240977  1.147346305 -0.176037520 
##          671          672          673          674          675 
## -0.807078631 -0.557475725 -0.869483985 -0.886144096 -1.789389167 
##          676          677          678          679          680 
## -1.879991423 -2.216156355 -1.873581809 -2.065795216 -2.224427754 
##          681          682          683          684          685 
## -2.153025312 -2.756852481 -2.821668994 -2.891438726 -3.107287648 
##          686          687          688          689          690 
##  2.556559092  1.585146235  0.912853044  0.243258742  0.084565160 
##          691          692          693          694          695 
## -0.176879207 -0.582579831  0.728382928 -2.031688861 -5.673090644 
##          696          697          698          699          700 
##  0.128901089 -0.848534235 -0.425037232 -0.026043450 -0.461419334 
##          701          702          703          704          705 
## -0.666059743 -0.583233298 -0.576804909 -1.025709317 -1.727829408 
##          706          707          708          709          710 
## -0.656146609 -1.173511313 -1.483330169 -1.102655255 -1.984596098 
##          711          712          713          714          715 
## -1.925230237 -2.075224673 -2.241378133 -2.105931208 -2.162257874 
##          716          717          718          719          720 
## -2.301354229 -2.591485584 -2.352828690  2.679176774  2.552197385 
```

```
##         721         722         723         724         725 
##  1.475345605  1.081885037  1.336060999  1.320918248 -0.163937907 
##         726         727         728         729         730 
##  0.226689115  0.155647222  0.615597172  0.841266298  0.306471871 
##         731         732         733         734         735 
## -0.177464407  0.347112755  0.422040392 -0.615294736 -0.348105883 
##         736         737         738         739         740 
## -0.791977664 -0.518978094 -0.636249443 -0.960427270 -0.107662718 
##         741         742         743         744         745 
## -1.446767273 -0.847153161 -0.853009530 -1.087654005 -1.037981022 
##         746         747         748         749         750 
## -1.341329846 -1.545713529 -2.050437241 -1.931290112 -3.112732942 
##         751         752         753         754         755 
## -2.649142210 -3.309949996  2.785279007  2.263076122  1.531179258 
##         756         757         758         759         760 
##  1.905201983  1.764683451  1.213794428  1.666860943  1.826353542 
##         761         762         763         764         765 
##  1.094292047  0.642080206  0.786003233  0.293665575  0.561125983 
##         766         767         768         769         770 
##  0.183822293 -0.296645354  0.459255598  0.257705551 -0.463951074 
##         771         772         773         774         775 
## -0.687939067  0.165302297 -0.485975879 -0.702747144 -0.516705767 
##         776         777         778         779         780 
## -0.798849741 -1.189519749 -1.284811307 -1.137687029 -2.215477427 
##         781         782         783         784         785 
##  3.981979061  3.257037431  2.441659396  1.865190889  2.184233319 
##         786         787         788         789         790 
##  2.438748847  1.523302204  1.744129527  1.792821336  0.977441966 
##         791         792         793         794         795 
##  0.761255989  0.987603262  0.845677064  0.751871029  1.311344445 
##         796         797         798         799         800 
##  1.121127694  0.939099389 -0.076258444  0.328537243  0.054671476 
##         801         802         803         804         805 
##  0.280665660  0.367567953 -0.843409666 -1.183476486  4.547311529 
##         806         807         808         809         810 
##  3.923177995  3.928424900  2.511007480  3.132594693  2.421288524 
##         811         812         813         814         815 
##  2.210515640  2.619886458  2.615086752  2.144988651  1.925027600 
##         816         817         818         819         820 
##  2.140899245  0.995929277  1.676481318  0.899878283  1.421671863 
##         821         822         823         824         825 
##  0.706562463  1.010632168  0.330221258 -0.676492952  5.512939661 
##         826         827         828         829         830 
##  3.582462901  3.565717639  3.207917457  2.921717005  2.891910573 
##         831         832         833         834         835 
##  1.970314406  2.583362960  2.434336198  1.514810192  1.805870562 
##         836         837         838         839         840 
##  1.103459451  4.316685112  4.858847022  3.737191773  2.971788177 
##         841         842         843         844         845 
##  3.110645457  2.544724923  2.029458434  6.796174721  5.605938444 
##         846         847         848         849         850 
##  4.054330034  3.889141502  3.209280692  5.693750433  5.221715677 
##         851         852         853         854         855 
##  8.968478813 -8.743284614  1.540916577  0.396943250  1.471849840 
```

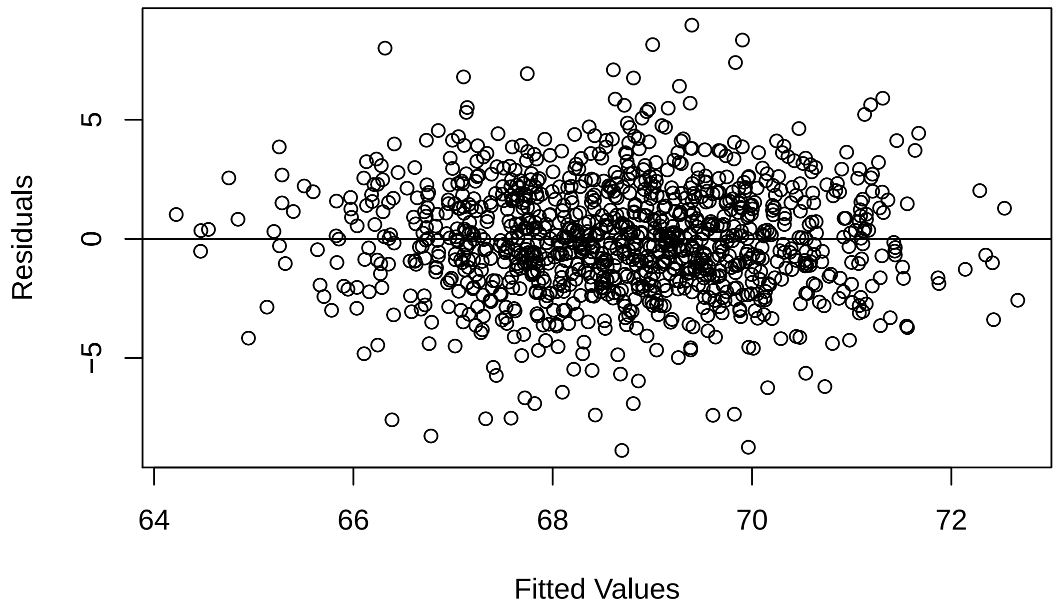
```
##          856          857          858          859          860 
##  2.905875337 -1.061552721 -6.910148818  1.510890334  2.068747073 
##          861          862          863          864          865 
## -5.731984395 -2.572537507 -4.664720036 -2.422006258 -3.815460108 
##          866          867          868          869          870 
## -4.266914447 -0.520498395 -2.221230289 -2.771972252 -2.997079420 
##          871          872          873          874          875 
## -2.925450034 -3.579032884 -6.245864307 -0.992296736 -1.651270933 
##          876          877          878          879          880 
## -1.731973307 -1.784466334 -2.309649546 -2.035579488 -2.274169372 
##          881          882          883          884          885 
## -3.139890119 -3.757720851 -3.203750953 -4.590130726  0.172744471 
##          886          887          888          889          890 
## -0.916724729 -0.730549555 -0.887759564 -0.822638186 -1.364119639 
##          891          892          893          894          895 
## -1.144799831 -1.514609136 -2.067856275 -1.915276036 -1.980686728 
##          896          897          898          899          900 
## -1.697328359 -2.642619860 -2.485659262 -3.179389457 -3.698535260 
##          901          902          903          904          905 
## -4.187737492  1.975566233  1.138536012  0.039401362  0.476744056 
##          906          907          908          909          910 
## -0.358221327  0.233489628  0.326412660 -0.385024862 -5.470119405 
##          911          912          913          914          915 
## -0.943585721 -3.858235770  1.721985741  0.411202107 -0.570615118 
##          916          917          918          919          920 
## -0.584183699 -0.618681331 -1.021739968  0.031520090 -1.290869634 
##          921          922          923          924          925 
## -1.464087582 -1.339519526 -0.922204781 -1.427366675 -1.125437078 
##          926          927          928          929          930 
## -1.689962615 -1.566518148 -1.994245576 -2.162774545 -1.965427571 
##          931          932          933          934          935 
## -1.932382933 -2.225195673 -2.552031012 -2.562924943  1.567945618 
##          936          937          938          939          940 
##  1.705210854  1.478785397  1.632961910  0.689264831  1.447444579 
##          941          942          943          944          945 
##  0.386428459  0.627950374  0.422352636  0.435988172  0.030293202 
##          946          947          948          949          950 
##  0.403164614  0.131251766  0.216064953  0.141792061 -0.575767023 
##          951          952          953          954          955 
## -0.894873844 -1.086366203 -0.876240723 -0.332976163 -0.722298637 
##          956          957          958          959          960 
## -0.940250308 -0.954278596 -1.291021065 -1.371546810 -1.005424910 
##          961          962          963          964          965 
## -1.193308727 -1.825408209 -1.966702622 -2.205402332 -2.280651354 
##          966          967          968          969          970 
## -2.668852466 -2.801679762  3.350174082  1.944998698  1.834501349 
##          971          972          973          974          975 
##  1.240292894  1.318071750  1.018360789  1.293638770  0.673796039 
##          976          977          978          979          980 
##  0.978309682  0.547953795  0.711440073  0.545759462  0.662990964 
##          981          982          983          984          985 
##  0.273431755  0.180373017  0.208792946 -0.016661794  0.164318768 
##          986          987          988          989          990 
##  0.179359096  0.199053067  0.037893666 -0.341811880 -1.120156569 
```

```
##         991         992         993         994         995
## -0.837729072 -1.109617521 -0.933549710 -1.109290052 -0.735894733
##         996         997         998         999        1000
## -1.633203894  3.028308450  2.181361796  3.050111495  2.889867743
##        1001        1002        1003        1004        1005
##  2.195399495  1.535078644  2.191354836  1.104142120  1.014290782
##        1006        1007        1008        1009        1010
##  1.997966028  1.384245909  1.004036442  0.598681741  1.378897978
##        1011        1012        1013        1014        1015
##  1.086436763  0.119568553  0.163910582  0.322355679  0.554619095
##        1016        1017        1018        1019        1020
## -0.414866878  0.171365277 -0.236351870 -0.520593564 -1.634873942
##        1021        1022        1023        1024        1025
##  3.436372382  3.506197296  3.668358411  2.939776083  2.466993091
##        1026        1027        1028        1029        1030
##  2.385146036  2.790695191  2.584203554  1.583647664  1.912701017
##        1031        1032        1033        1034        1035
##  1.586134683  1.580070675  1.332671761  1.946137415  1.254843310
##        1036        1037        1038        1039        1040
##  1.309027961  0.602998826  0.414716046 -0.364875489 -1.278200379
##        1041        1042        1043        1044        1045
##  4.136000253  3.682608593  3.774236726  3.211133930  2.378770554
##        1046        1047        1048        1049        1050
##  2.131856040  2.588984847  2.049679070  1.836506945  1.220304149
##        1051        1052        1053        1054        1055
##  0.962337404  4.331060985  4.028872564  4.195720592  3.619775234
##        1056        1057        1058        1059        1060
##  2.313494450  2.811195962  2.270234930  1.997450895  5.865147140
##        1061        1062        1063        1064        1065
##  4.680763131  3.452156878  2.926232447  1.284000836  4.123250188
##        1066        1067        1068        1069        1070
##  7.400658027 -7.523332626  5.628732981 -4.251839278 -7.593037101
##        1071        1072        1073        1074        1075
## -3.658603782 -6.671128941 -8.877150660  2.423122015 -2.290051307
##        1076        1077        1078
## -1.483926919 -0.950717935 -3.015475796
```

```r
ft = lm(fit_sonHeight$residuals ~ fit_sonHeight$fitted.values)

#Visualization of the residuals
plot(fit_sonHeight$fitted.values, fit_sonHeight$residuals,
xlab = "Fitted Values", ylab = "Residuals")

abline(ft)
```



From the plot we can observe the plot of the fitted values and the residuals is similar to the plot between father's height and son's height. The outliers in the father's height vs son's height plot (eg. at point 68) have corresponding residual values in the above plot. Also, there is no pattern in the fitted values vs residuals plot, they roughly form a horizontal line around the 0 line which suggests the variances of the error terms are equal and no one residual stands out. Hence we can say this linear regression has highly appropriate.

My concern would be how this model will behave when the height values are outside the range of the data which was used to build this model. For example - what if the father's height is 55 inches. How accurately can we predict the son's height in this case? This model may not be ideal for this dataset

(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.

```
#Create data frame with father's heights
predict_son_height = data.frame(fheight = c(50, 55, 70, 75, 90))

#Predict son's heights
predict(fit_sonHeight, predict_son_height, 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
```

From the fitted values we can see that the son's heights are 59.59126, 62.16172, 69.87312, 72.44358, 80.15498 when their father's heights are 50, 55, 70, 75, 90 respectively.

3. **Extra Credit:**

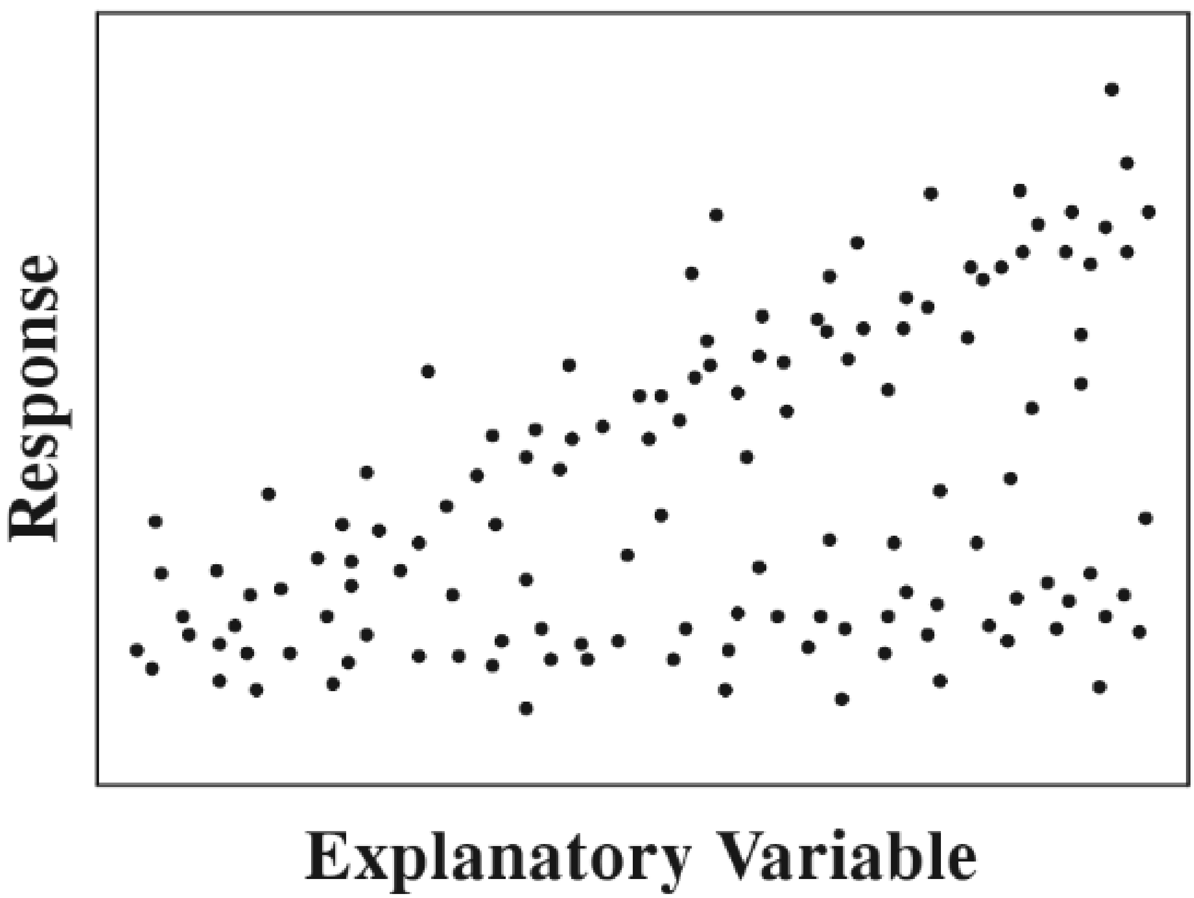


Figure 1: Scatterplot for Extra Credit (d).

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

There are no assumptions made about the distribution of the explanatory variable in the normal simple linear regression model. The assumptions made in linear regression are as follows:

(a) All values of the dependent variable (y variable) are independent of each other.
(b) For each value of X, the distribution of possible Y values is normal.
(c) Linear regression needs the relationship between the independent and dependent variables to be linear.
(d) Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are not independent from each other. Also the error of the mean has to be independent from the independent variables.
(e) Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other.
(f) Homoscedasticity - The error is uniform across all values of the independent variables.

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

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. R squared explains what proportion of variability in the response has been explained by the regression. R squared close to 1 may indicate most of the variability in the regression has been explained whereas we might expect it to be otherwise i.e. in cases where residual errors might be large due to unmeasured factors, a R squared value closer to 0 might be closer to the truth.

(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?

A regression of weight and height for a sample is for us to understand the approximate relationship between weight and height for adult males. But it only makes sense within the range of normal weight and height of adult males. Linear regression is simply a modeling frame-work. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave. By applying linear regression outside of the realm of the original data is extrapolation and if we extrapolate it to a male with height 0, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed. However, this doesn't make the linear regression model meaningless. It holds meaning with the range of data for which it was designed but may not be extrapolated outside of that range.

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

I would take a look at the original data and try to make sense of the variables. I will try to identify what the predictor and response variables should be and whether they have been plotted accordingly. It may happen that reversing the axes might make a significant difference. Since we can observe 2 different fitted lines, I will try to find a grouping of the explanatory variable and consider splitting it into two separate groups to make the statistical inference more accurate.