Assignment - 3

February 13, 2017

1) For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

a) (1 pt) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

Mean - Numerical data

b) (1 pt) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

Mean - Numerical data

c) (1 pt) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

Proportion - Categorical data (Yes/ No)

d) (1 pt) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

Mean - Numerical data

e) (1 pt) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

Proportion - Categorical data

2) (5 pt) In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

For 95% confidence interval, we chose the multiplier 1.96 as,

```
pnorm(1.96) - pnorm(-1.96)
```

```
## [1] 0.9500042
```

```
upr.int <- 45 + (1.96 * 1.2)
low.int <- 45 - (1.96 * 1.2)
upr.int
```

```
## [1] 47.352
```

```
low.int
```

```
## [1] 42.648
```

The 95% confidence interval is [42.648 47.352] Interpretation : The research is 95% confident that the true value of the percentage of U.S adults who report that they live with one or more chronic conditions lies within [42.648 47.352]

3)The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 136 calories with a standard deviation of 17 calories.

a) (4 pt) Write down the null and alternative hypotheses for a two-sided test of whether the nutrition label is lying.

Null Hypothesis : The mean of one ounce of potato chips is equal to 130 calories.

Alternative Hypothesis : The mean of one ounce of potato chips is not equal to 130 calories.

b) (4 pt) Calculate the test statistic and find the p value.

The test statistic:

```
sample_mean <- 136
population_mean <- 130
sample_size <- 35
sample_sd <- 17

test.s <-  (sample_mean - population_mean) / (sample_sd/sqrt(sample_size))
test.s
```

```
## [1] 2.088028
```

Considering a 95% confidence interval, the significant value (alpha value) is 1.96, then the p-value for the two sided test is:

```
a <-  pnorm(test.s, lower.tail = F)

b <- pnorm(-test.s, lower.tail = T)

a+b
```

```
## [1] 0.03679529
```

c) (2 pt) If you were the potato chip company would you rather have your alpha = 0.05 or 0.025 in this case? Why?

I would rather have the alpha value = 0.025 as I can accept the null hypothesis and thereby claim with a 97.5% confidence interval that the nutrition label on the bag of chips is true. By setting it at 0.05, the null hypothesis will be rejected which will indicate that my nutritional label is misleading.

4) Regression was originally used by Francis Galton to study the relationship between parents and children. He wondered if he could 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:
library(UsingR)
height <- get("father.son")

```
#install.packages("UsingR")
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

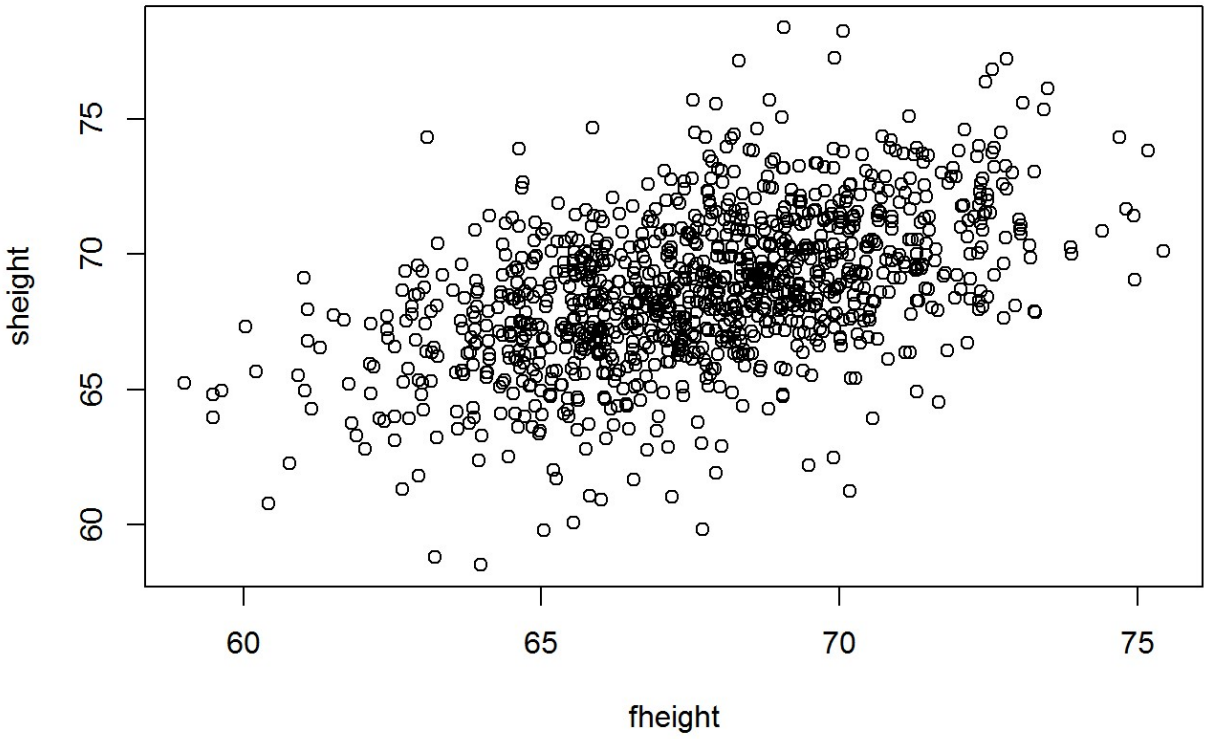
```
##
## Attaching package: 'UsingR'
```

```
## The following object is masked from 'package:survival':
##
##     cancer
```

```
height <- get("father.son")
```

a) (5 pt) Perform an exploratory analysis of the father and son heights. What does the relationship look like? Would a linear model be appropriate here?

```
attach(height)
plot(fheight,sheight)
```




```
cor(fheight,sheight, use = "everything", method = c("pearson") )
```

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

By plotting fheight vs. sheight, there appears to be a linear positive association between the two variables. The pearson correlation co-efficient is 0.501 which confirms a moderately positive correlation.

b) (5 pt) 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, ysheight = ??0 + ??1 x fheight filling in estimated coefficient values and interpret the coefficient estimates.

```
mod1 <- lm(sheight ~ fheight, data = height)
summary(mod1)
```

```
##
## Call:
## lm(formula = sheight ~ fheight, data = height)
##
## 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
```

As, B0 <- 33.88660 B1 <- 0.51409 ysheight = 33.88660 + (0.51409 * fheight)

The B0 coeffiecient indicates the value of the sheight will be 33.8866 when fheight = 0. Similarly, B1 coefficient indicates the slope of the linear model wherein, a 1 inch increase in the father's height predicts a 0.51409 inch increase in the son's height.

c) (5 pt) Find the 95% confidence intervals for the estimates. You may find the confint() command useful.

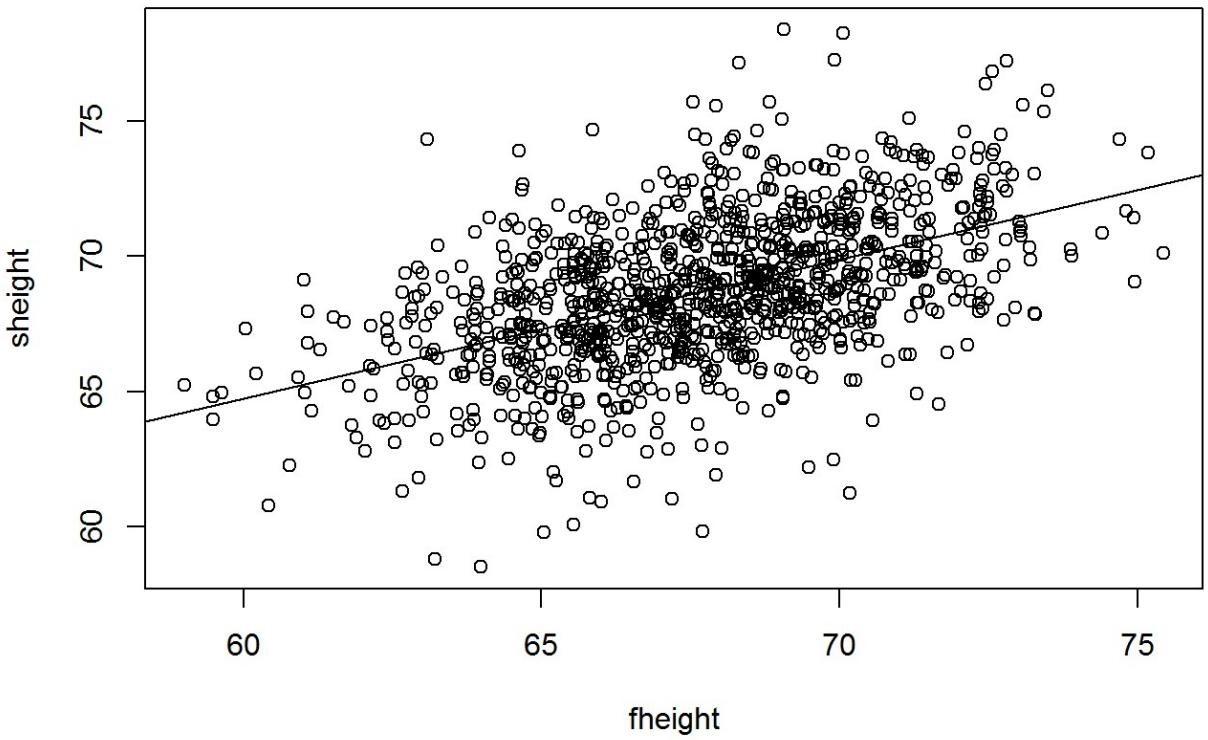
```
confint(mod1,level = 0.95)
```

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

The 95% confidence interval for the slope estimate is [0.4610785 0.5671076] while for the intercept estimate it is [30.2952569 37.4779519]

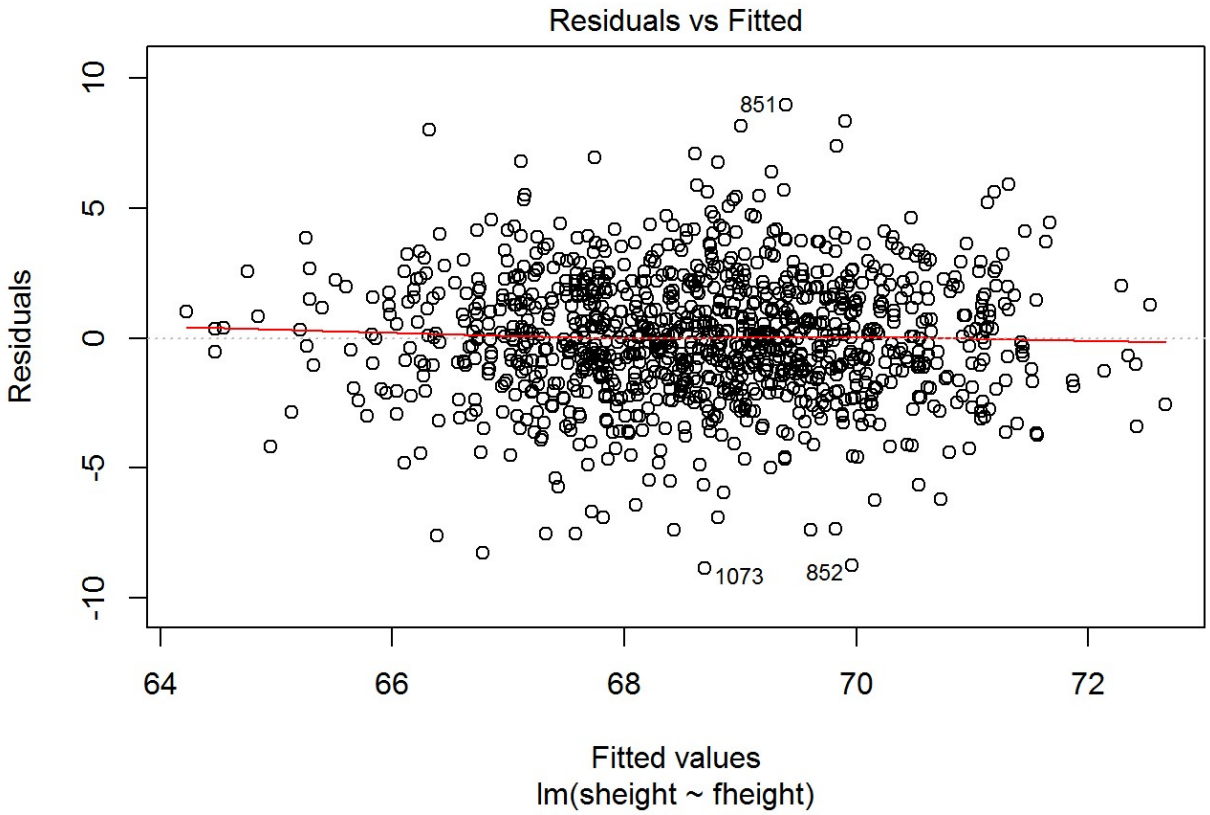
d) (5 pt) Produce a visualization of the data and the least squares regression line.

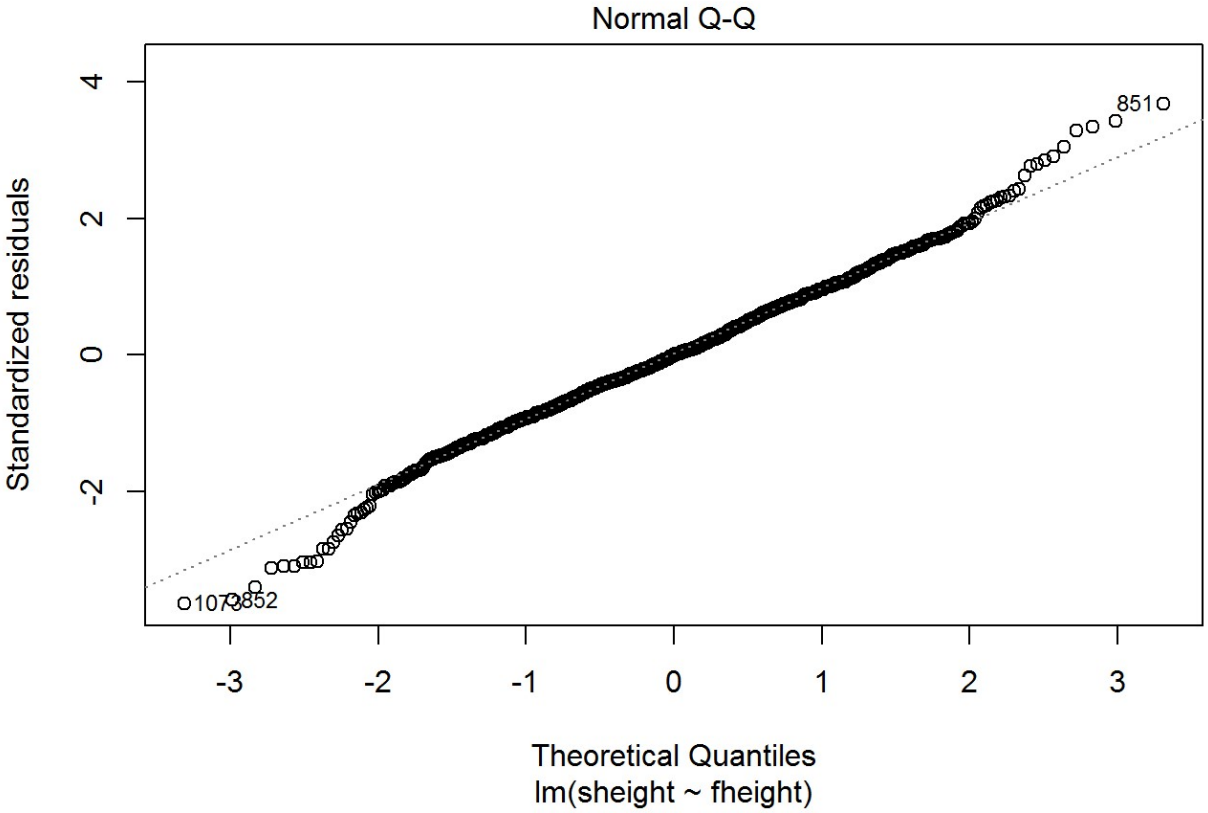
```
plot(fheight,sheight)
abline(mod1)
```

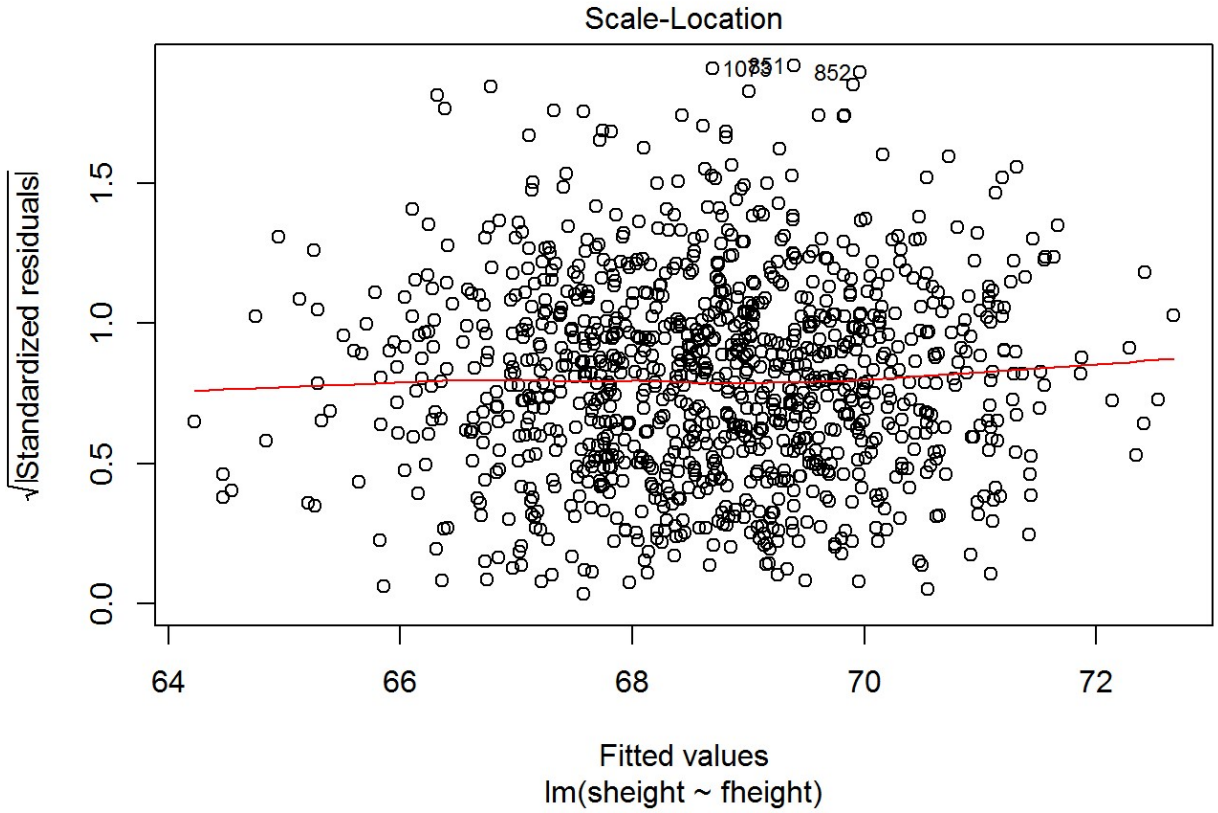


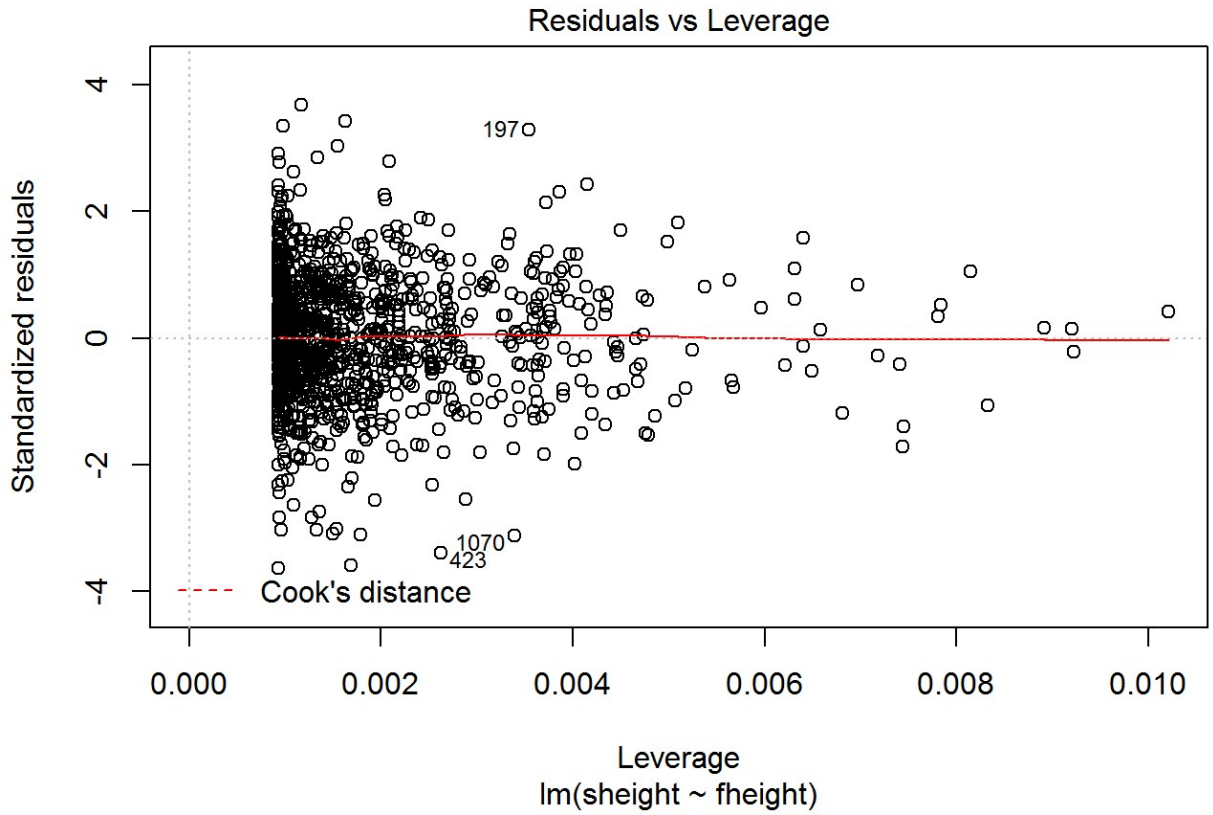
e) (5 pt) Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using names()). Discuss what you see. Do you have any concerns about the linear model?

```
plot(mod1)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
851
1073
852
64
66
68
70
72
Fitted values
lm(sheight ~ fheight)


Normal Q-Q
Standardized residuals
4
2
0
-2
851
1073
852
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(sheight ~ fheight)


Scale-Location
1078
851
852
√|Standardized residuals|
0.0
0.5
1.0
1.5
64
66
68
70
72
Fitted values
lm(sheight ~ fheight)



In the residuals vs fitted values plot, as the residuals vary around the horizontal line, the model is linear and is a moderate fit.

f) (5 pt) 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.

```
predict(mod1, newdata=data.frame("fheight" = 50),
                                        interval="prediction")
```

```
##        fit      lwr      upr
## 1 59.59126 54.71685 64.46566
```

```
predict(mod1, newdata=data.frame("fheight" = 55),
                                        interval="prediction")
```

```
##        fit     lwr      upr
## 1 62.16172 57.3314 66.99204
```

```
predict(mod1, newdata=data.frame("fheight" = 70),
                                        interval="prediction")
```

```
##        fit      lwr      upr
## 1 69.87312 65.08839 74.65785
```

```
predict(mod1, newdata=data.frame("fheight" = 75),
                                    interval="prediction")
```

```
##        fit     lwr      upr
## 1 72.44358 67.6447 77.24246
```

```
predict(mod1, newdata=data.frame("fheight" = 90),
                                    interval="prediction")
```

```
##        fit     lwr      upr
## 1 80.15498 75.2274 85.08255
```

g) (5 pt) What do the estimates of the slope and height mean? Are the results statistically significant? Are they practically significant?

The slope indicates that for a 1 inch increase in the father's height the son's height is expected to increase by 0.51409 inches.

The intercept indicates that for a case when the father's height is 0 inches, the expected height of the son is 33.8866 inches.

Yes the result is statistically significant, as the p-value is 2.2e-16 which is less than the significant alpha value of 0.05, we reject the null hypothesis.

However, the model is not practically significant as based on the Adjusted R-squared value of 0.2506, the model does not fit the data well as it does not seem to consider other possible predictors.

5) An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the father's IQ, the mother's IQ, and hours of educational TV. The data are here:library(openintro) data(gifted)

```
#install.packages("openintro")
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following object is masked from 'package:MASS':
##
##     mammals
```

```
## The following object is masked from 'package:datasets':
##
##     cars
```

```
data(gifted)
```

a) (5 pt) Run two regressions: one with the child's analytical skills test score ("score") and the father's IQ ("fatheriq") and the child's score and the mother's IQ score ("motheriq").

```
mod.f <- lm(score~fatheriq, data = gifted)
mod.m <- lm(score~motheriq, data = gifted)

summary(mod.f)
```

```
##
## Call:
## lm(formula = score ~ fatheriq, data = gifted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6942 -3.2565  0.3058  2.0559 10.5559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130.4294    25.7226   5.071 1.39e-05 ***
## fatheriq      0.2501     0.2240   1.117    0.272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.614 on 34 degrees of freedom
## Multiple R-squared:  0.03537,    Adjusted R-squared:  0.007003
## F-statistic: 1.247 on 1 and 34 DF,  p-value: 0.272
```

```
summary(mod.m)
```

```
##
## Call:
## lm(formula = score ~ motheriq, data = gifted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3569 -2.7497  0.1157  2.8794  8.7091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.0930    11.8567   9.370 6.02e-11 ***
## motheriq      0.4066     0.1002   4.058 0.000274 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.856 on 34 degrees of freedom
## Multiple R-squared:  0.3263, Adjusted R-squared:  0.3065
## F-statistic: 16.47 on 1 and 34 DF,  p-value: 0.000274
```

b) (5 pt) What are the estimates of the slopes for father and mother's IQ score with their 95% confidence intervals? (Note, estimates and confidence intervals are usually reported: Estimate (95% CI: CIlower, CIupper)

```
confint(mod.f,level = 0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 78.1548748 182.7039518
## fatheriq    -0.2051068   0.7053687
```

Estimate and confidence interval for slope for father's IQ is:

(0.2501: -0.2051068, 0.7053687)

```
confint(mod.m,level = 0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 86.9972563 135.1886542
## motheriq     0.2029815   0.6102077
```

Estimate and confidence interval for slope for mother's IQ is:

(0.4066: 0.2029815, 0.6102077)

c) (5 pt) How are these interpreted?

For the slope for the Father's IQ score: With an increase in 1 IQ of the father's IQ Score, the IQ of the child is expected to increase by 0.2501 points. There is 95% confidence that the true value of the slope estimate for the father's IQ lies within [-0.2051068 0.7053687]

Similarly, For the slope for the mother's IQ score: With an increase in 1 IQ of the mother's IQ Score, the IQ of the child is expected to increase by 0.4066 points There is 95% confidence that the true value of the slope estimate for the mother's IQ lies within [0.2029815, 0.6102077]

d) (5 pt) What conclusions can you draw about the association between the child's score and the mother and father's IQ?

Based on the single regression model of the child's score with that of the father's IQ, the p-value is higher than the significance level alpha = 0.05. We can accept the null hypothesis that there is no association between the child's IQ score and that of the father's, as assuming H0 is true, the probability of observing a more extreme test statistic in the direction of the alternative hypothesis than we did is 0.272

The adjusted r2 value is low at 0.007003 as well which indicates that the model does not fit the data well.

Similarly, Based on the single regression model of the child's score with that of the mother's IQ, the p-value is lower than the significance level alpha = 0.05. We can reject the null hypothesis to indicate there is an association between the child's IQ score and that of the mother's as assuming H0 is true, the probability of observing a more extreme test statistic in the direction of the alternative hypothesis than we did is 0.000274

The adjusted r2 value is comparatively high at 0.3065 which indicates that the model fits the data better than the linear model of the father's IQ although it does not completely fit the data.

Considering a multiple regression model confirms this interpretation. The adjusted r2 value is slightly higher at 0.3289 than the one obtained in the single regression model between the child's score and that of the mother's which indicates little to no impact of the father's iq on the child's score.

```
mod.both <- lm(score~fatheriq + motheriq, data = gifted)
summary(mod.both)
```

```
##
## Call:
## lm(formula = score ~ fatheriq + motheriq, data = gifted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9242 -2.7109  0.2544  2.5915  9.4466
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.77863   24.40057   3.270 0.002523 **
## fatheriq     0.26915    0.18421   1.461 0.153452
## motheriq     0.41017    0.09859   4.160 0.000213 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.793 on 33 degrees of freedom
## Multiple R-squared:  0.3672, Adjusted R-squared:  0.3289
## F-statistic: 9.577 on 2 and 33 DF,  p-value: 0.0005252
```