Assignment_3

Abhinav Garg

2/12/2017

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.

Ans: These responses are numerical and the parameter of interest is a mean

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?”

Ans: The parameter of interest would be a mean. Since we’re loooking at individual numerical responses of over 100 individuals.

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

Ans: These responses would be categorical as they would be answered in yes / no / no response. This parameter would be proportion

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

Ans: The parameter of interest here is mean. Since we’re loooking at individual numerical responses of over 100 individuals.

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.

Ans: The parameter of interest here is proportion since 85 percent of the population expect to get a job and the remaining precent don’t.

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.

```
u <- 45
se <- 1.2
left <- 45 - 1.96*se
right <- 45 + 1.96*se
left
```

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

```
right
```

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

So as we can see the confidence interval is between (42.648, 47.352) With 95% confidence we can say that the average number of adults that live with one or more chronic conditions lies between 42.648 and 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. Write down the null and alternative hypotheses for a two-sided test of whether the nutrition label is lying.

Null Hypothesis : The mean of a one ounce serving of potato chips is 130 calories

Alternate Hypothesis : The mean of a one ounce seving of potato chips is greater or lesser than 130 calories

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

Assuming the null hypothesis to be true : u = 130 x_bar = 136 n = 35 sd = 17

Z = x_bar - u / (sd / sqrt(n))

```
u  = 130
x_bar = 136
n = 35
sd = 17
num <- x_bar - u
den <- sd / sqrt(n)
Z <- num /den
R <- 2*pnorm(-abs(Z))
```

Z-statistic = 2.0880282

P-value for two-tailed = 0.0367953

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?

Since the p-value is 0.036, if the significance level is 0.05, the p-value is lesser than the significance level, therefore we would have to reject the null hypothesis that the calories in one serving is 130 calories and accept that it is more than 130 calories. However if the significance level is 0.025 we would be within the confidence interval and accept the null hypothesis. Thus we should use the 0.025 significance level. But that being said, we cannot assume the significance level after calculating or taking the samples as this is ethically not correct.

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:

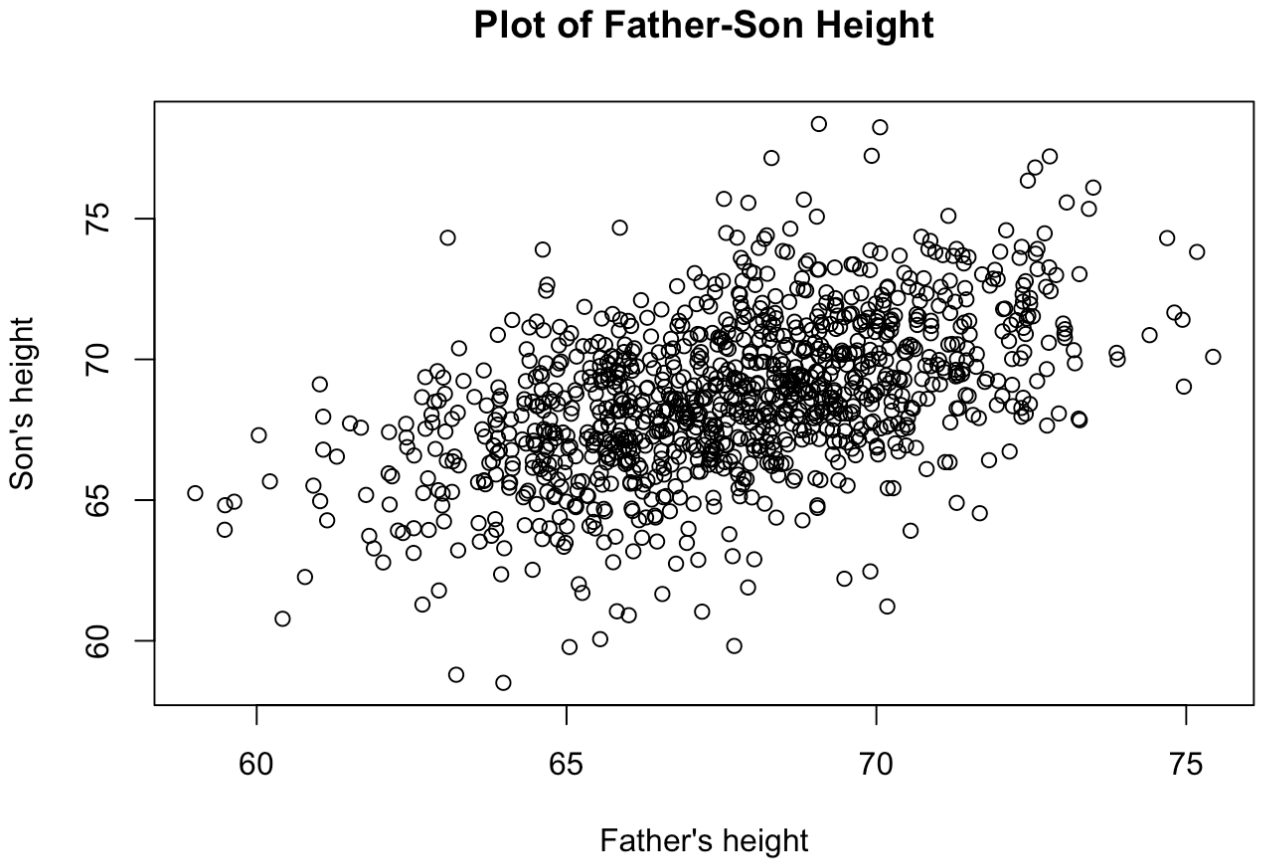
a. (5 pt) Perform an exploratory analysis of the father and son heights. What does the relationship look like? Would a linear model be appropriate here?

```
summary(height)
```

```
##     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
```

We see a summary of the dataset initially

```
plot(sheight ~ fheight, data = height, xlab = "Father's height", ylab="Son's heigh
t" )
title("Plot of Father-Son Height")
```

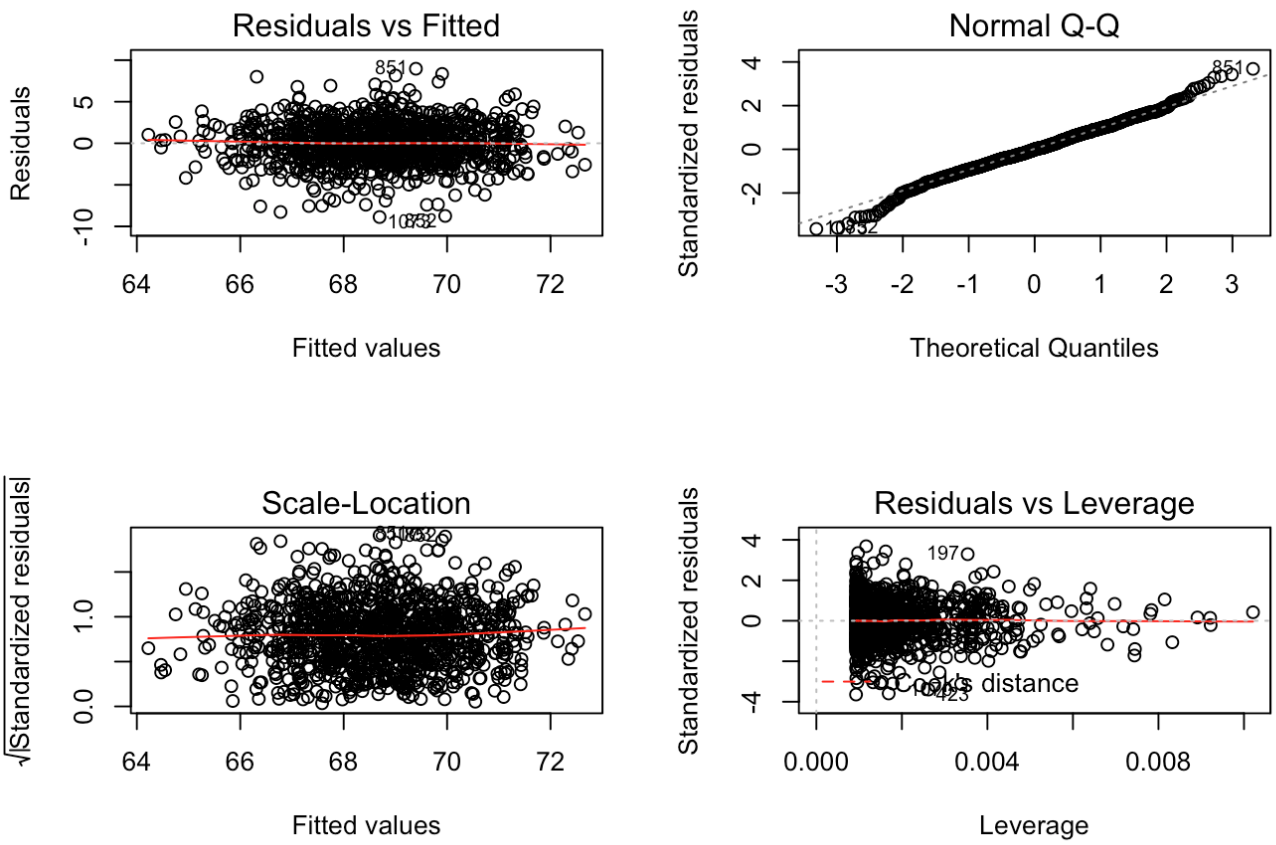



```
cor(height$sheight, height$fheight)
```

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

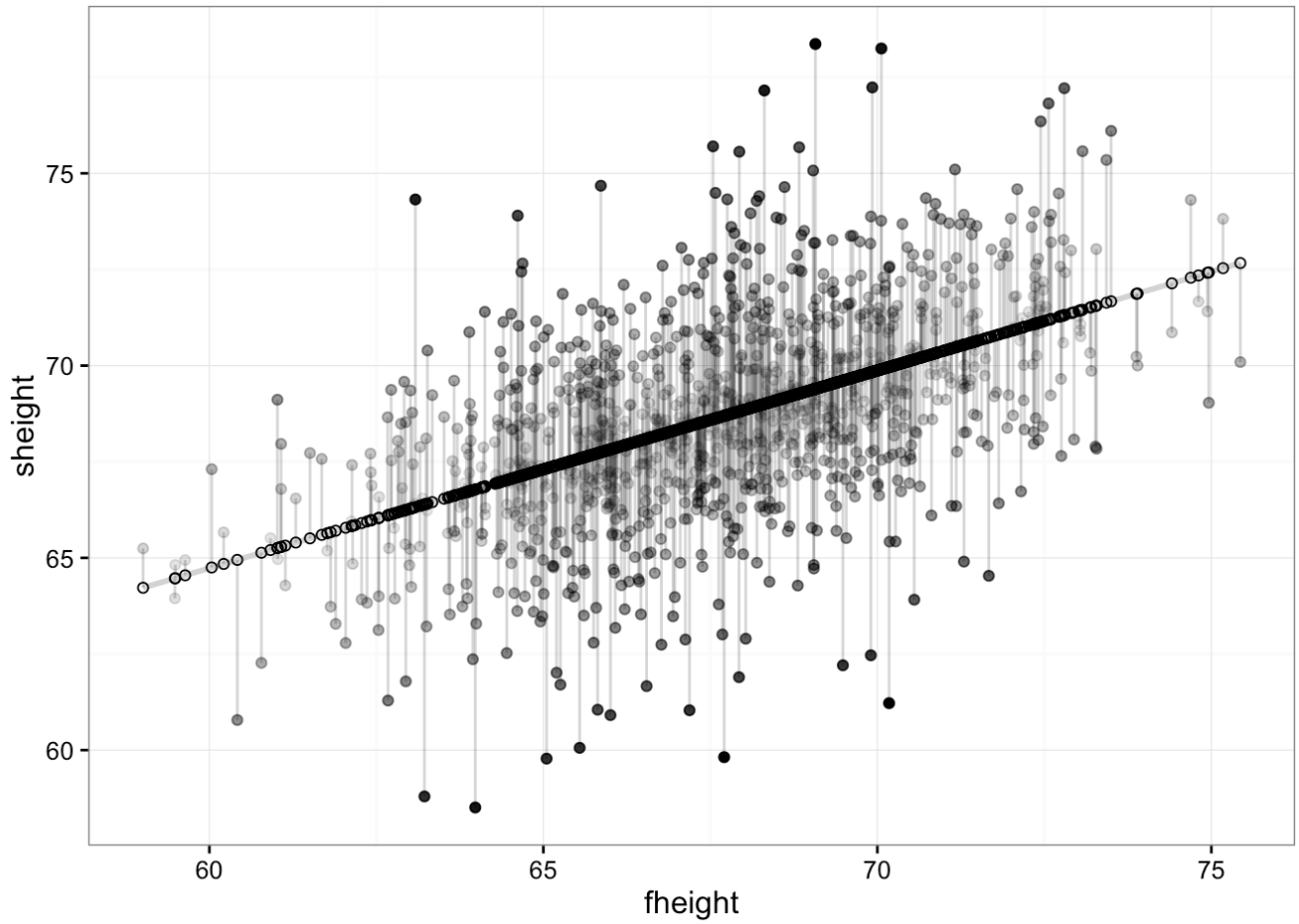
We can see that there is a positive correlation between the son and father's height, therefore a linear model could be utilized here. Let us also observe the residuals just to ensure there isn't a pattern.

```
fit <- lm(sheight ~ fheight, data = height)
par(mfrow = c(2, 2))
plot(fit)
```



We see different kinds of plots to observe this data. Let us focus on the residual plot to the top left. From this plot we can observe that the variables are randomly distrbuted and that there is no pattern between the residuals and the independent variable. Thus we can say that a linear model is appropriate to use. If there was a visible pattern (like u-shaped or inverted-u) in this plot we would have to utilize non-linear models.

```
predicted <- predict(fit)
residuals <- resid(fit)
ggplot(height, aes(x = fheight, y = sheight)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  geom_segment(aes(xend = fheight, yend = predicted), alpha = .2) +
  geom_point(aes(alpha = abs(residuals))) +
  guides(alpha = FALSE) +
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```



We can see how the further the further the actual data point is from the regression line, the darker is the point

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 = β0 + β1 x fheight

filling in estimated coefficient values and interpret the coefficient estimates.

```
mod3 <- lm(height$sheight ~ height$fheight)
summary(mod3)
```

```
##
## Call:
## lm(formula = height$sheight ~ height$fheight)
##
## 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 ***
## height$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
```

Therefore β0 = 33.886 β1 = 0.5149

The estimates of the slope means that for an increase in every inch of the father's height, the son's height increase by 0.51409 inches. The estimate of the intercept means that if the father's height is zero the son's height is 33.86 inches which doesn't make sense. There would've been no data where the father's height is zero. However it is acceptable to interpret this as a coefficient in the model.

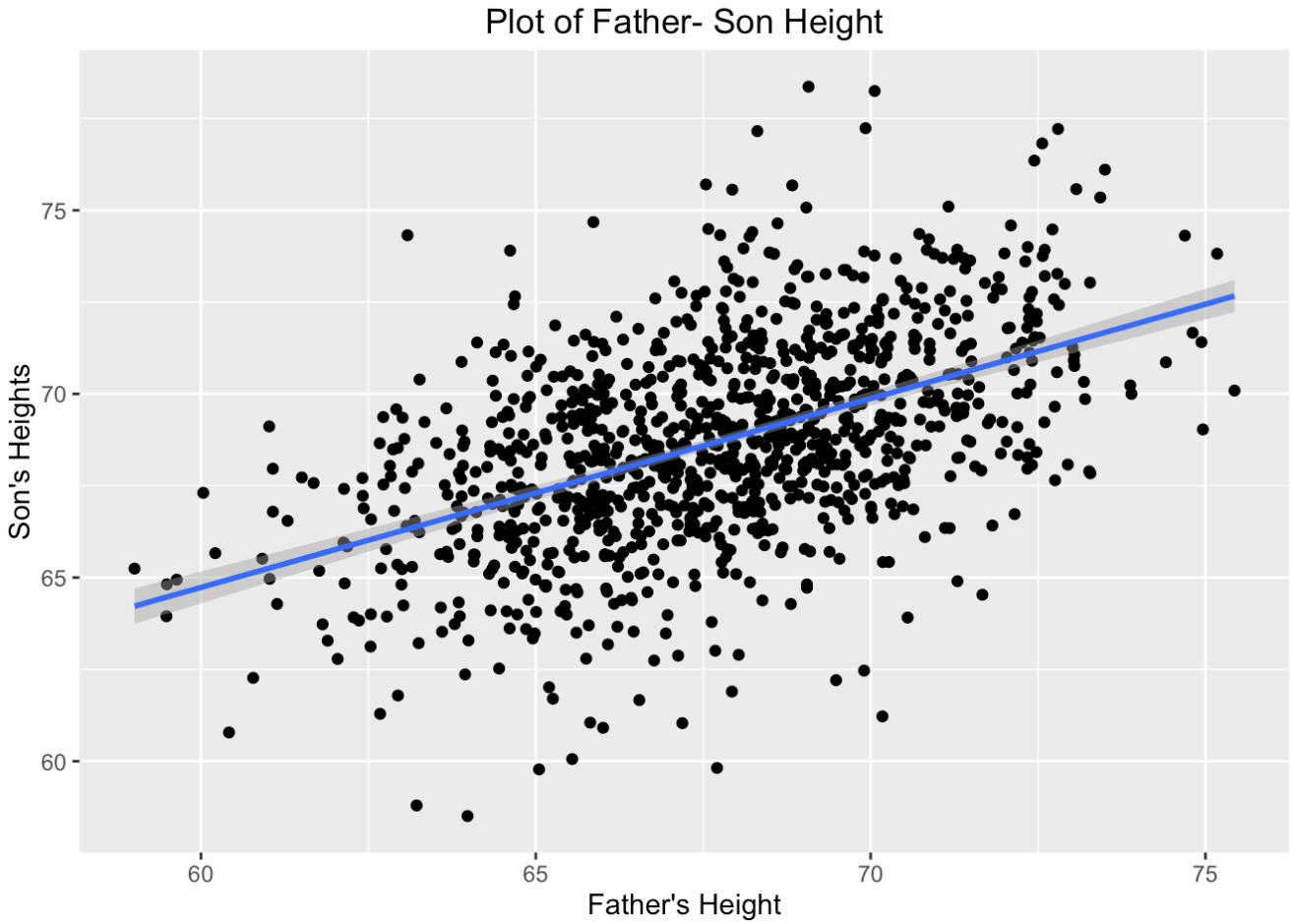
c. (5 pt) Find the 95% confidence intervals for the estimates. You may find the confint() command useful.

```
confint(mod3,level=0.95)
```

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

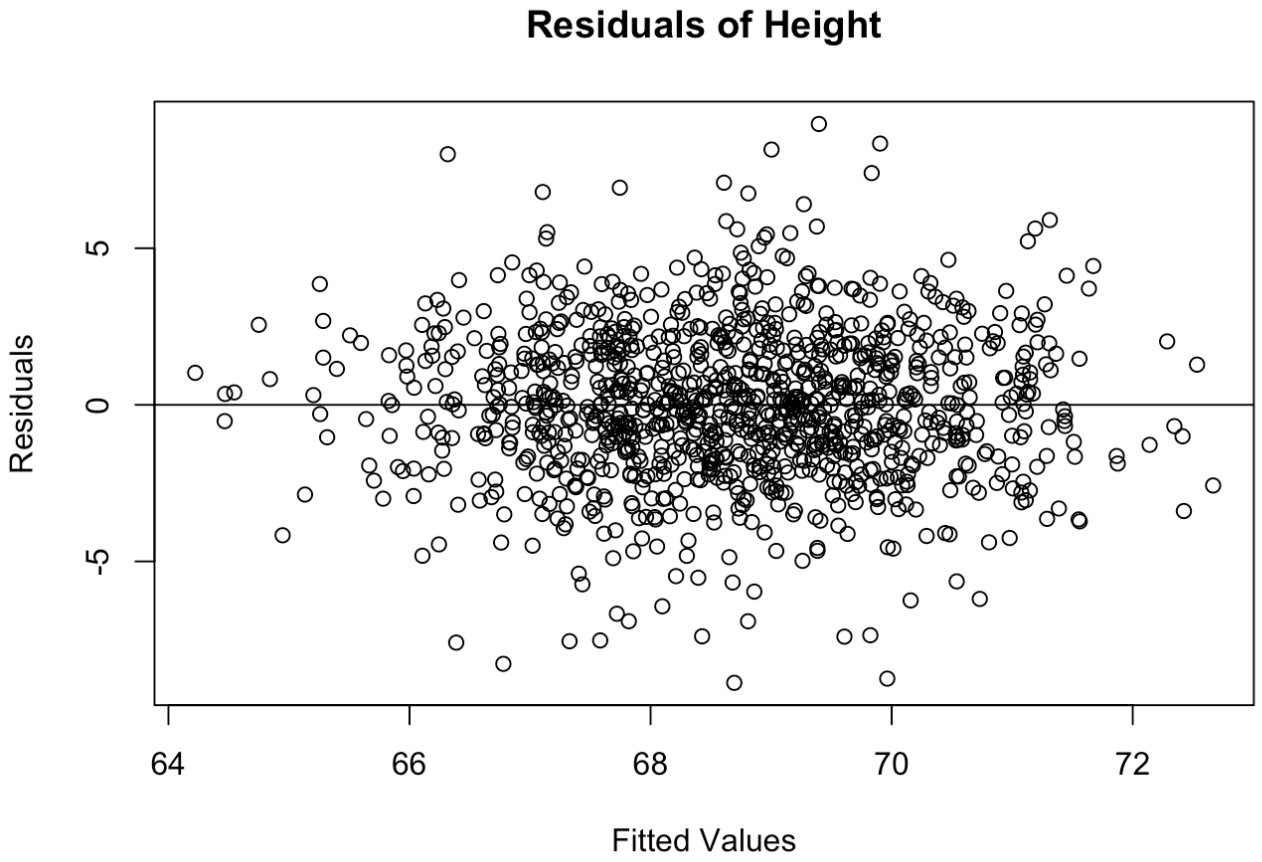
d. (5 pt) Produce a visualization of the data and the least squares regression line.

```
library(ggplot2)
g<- ggplot(data = height,mapping = aes(x= fheight, y = sheight))
g <- g + geom_point() + xlab("Father's Height") + ylab("Son's Heights")
g <- g + labs(title = "Plot of Father- Son Height")
g <- g + geom_smooth(method = 'lm', se = TRUE)
g
```



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?

```
fit <- lm(sheight ~ fheight, data = height)
res <- resid(fit)
fitted <- fit$fitted.values
par(mfrow = c(1,1))
plot(fitted, res,
     ylab = "Residuals",
     xlab = "Fitted Values",
     main = "Residuals of Height")
abline(0, 0)
```



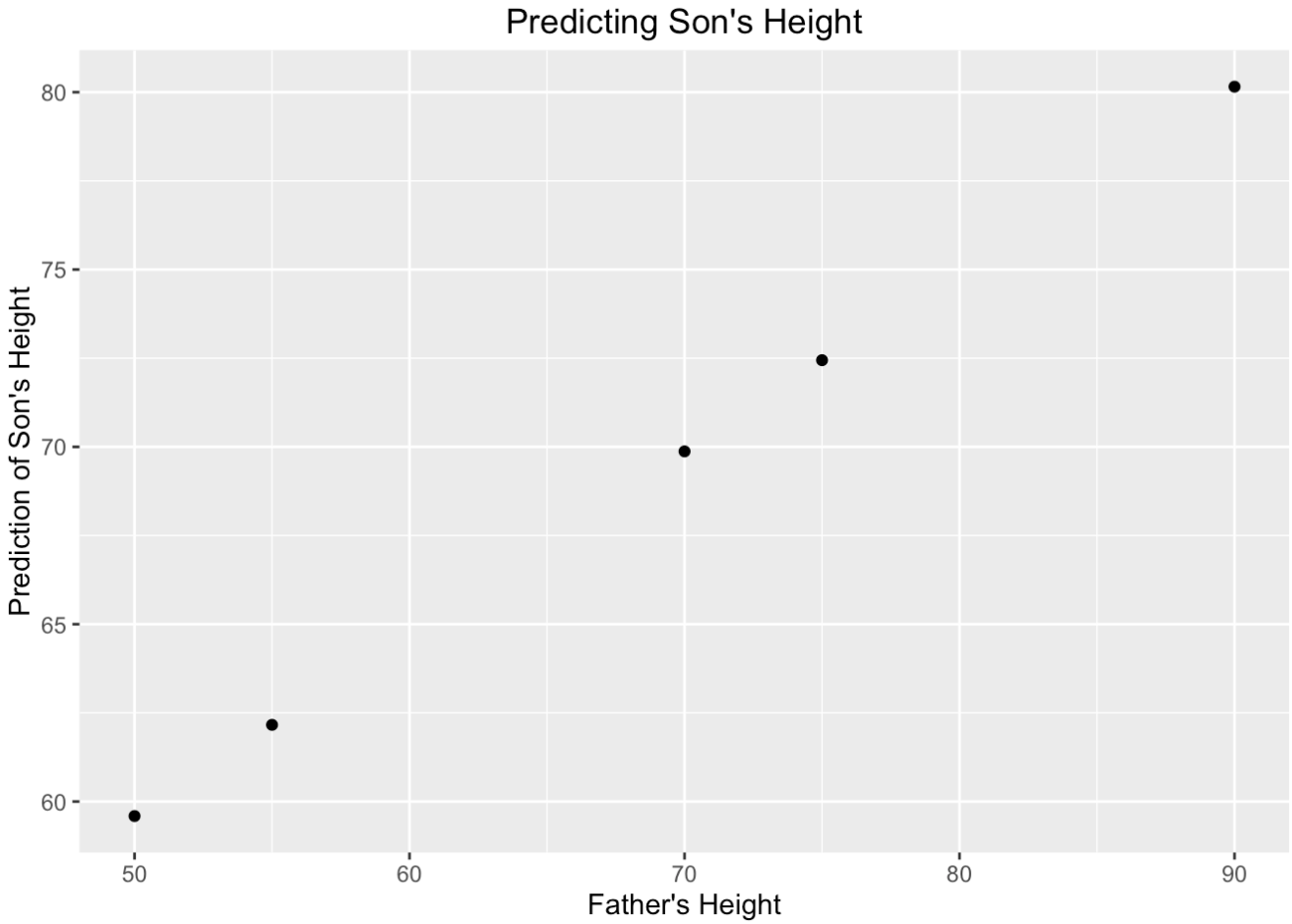
As mentioned in the exploratory analysis section, since the plot of fitted values vs the residuals is random and scattered around zero, a linear model is appropriate to use.

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

```
values <- data.frame(fheight=c(50,55,70,75,90))
predictions <-  predict(fit, newdata = values)
predictions
```

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

```
val <- as.numeric(c(50,55,70,75,90))
m <- ggplot(mapping = aes(x = val, y =predictions))
m <- m + geom_point() + labs(title = "Predicting Son's Height") + xlab("Father's H
eight")+ ylab("Prediction of Son's Height")
m
```



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

```
summary(fit)
```

```
##
## 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
```

The estimates of the slope and height are statistically significant as the p-value for both (< 2e-16) is very small. Indicating that the possibility of observing this by chance is very small. The estimates of the slope means that for an increase in every inch of the father's height, the son's height increase by 0.51409 inches. The estimate of the intercept means that if the father's height is zero the son's height is 33.86 inches which doesn't make sense. There would've been no data where the father's height is zero hence we are getting this as an anomaly. However it is acceptable to interpret this as a coefficient in the model. The estimates are practically significant as an increase in 1 inch in the father's height means an increase in the son's height by 0.5 inches which could be applicable in a practical setting, although there could be some exceptions to this trend

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:

```
mod1 <- lm(score ~ fatheriq, data = gifted)
summary(mod1)
```

```
##
## 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
```

```
mod2 <- lm(score ~ motheriq, data = gifted)
summary(mod2)
```

```
##
## 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)

```
est_conf <- confint(mod1)
est_conf
```

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

Father Intercept Estimate: 130.429 (95% CI : 78.154, 182.704) Score Estimate: 0.2501 (95% CI : -0.205, 0.705)

```
est_conf_2 <- confint(mod2)
est_conf_2
```

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

Mother Intercept Estimate: 111.093 (95% CI : 86.997, 135.188) Score Estimate: 0.4066 (95% CI : 0.202, 0.610)

c. (5 pt) How are these interpreted? Assuming the null hypothesis is true, we can say with 95% confidence that the true intercept for the model of analytical skills of a young child and his/her father will lie between 78.154 and 182.704 and the true slope would lie between -0.205 and 0.705

Similarly, Assuming the null hypothesis is true, we can say with 95% confidence that the true intercept for the model of analytical skills of a young child and his/her mother will lie between 86.997 and 135.188 and the true slope would lie between 0.202 and 0.610

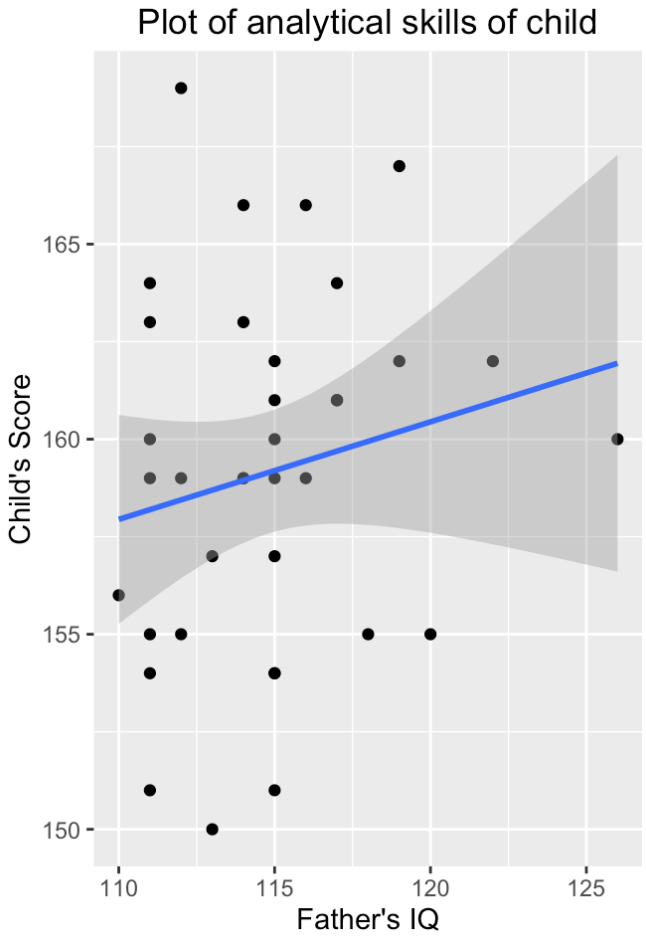
d. (5 pt) What conclusions can you draw about the association between the child's score and the mother and father's IQ?

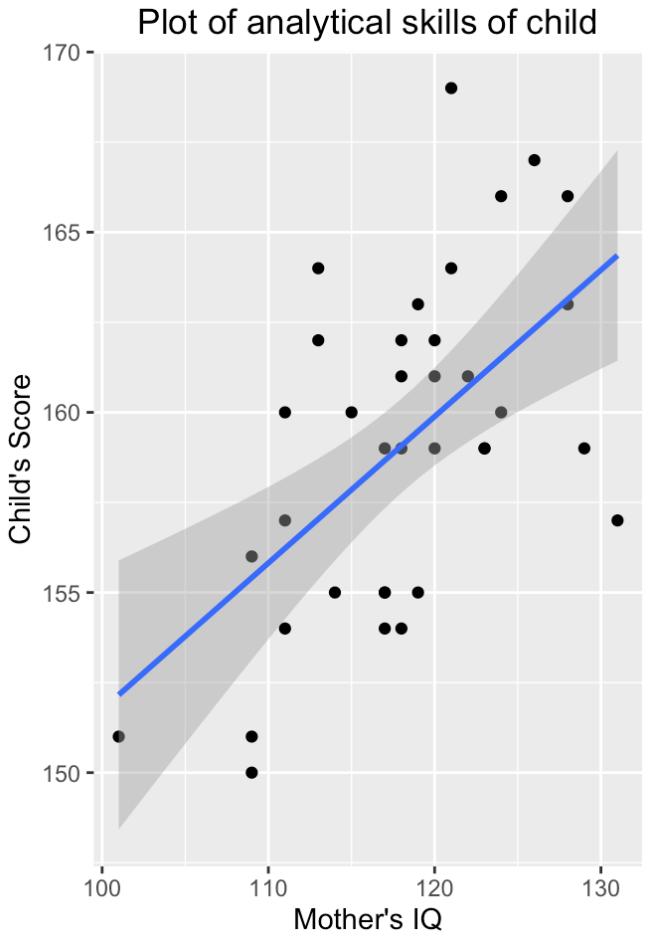
Looking at the Adjusted R-squared in both models we can see that the Adjusted R-squared for the model fitting the association between the father's IQ and the child's score is 0.007 whereas the Adjusted R-squared for the model fitting the association between the mother's IQ and the child's score is 0.3065. Thus the adjusted r-squared for the model fitting the association betweeen the mother's IQ and the child's score is greater and statistically more significant than that between the father's. Thus we can interpret that the mother's IQ is a better predictor of the analytical skills of a young child

```
g1<- ggplot(data = gifted,mapping = aes(x= fatheriq, y = score))
g1 <- g1 + geom_point() + xlab("Father's IQ") + ylab("Child's Score")
g1 <- g1 + labs(title = "Plot of analytical skills of child")
g1 <- g1 + geom_smooth(method = 'lm', se = TRUE)


g2<- ggplot(data = gifted,mapping = aes(x= motheriq, y = score))
g2 <- g2 + geom_point() + xlab("Mother's IQ") + ylab("Child's Score")
g2<- g2 + labs(title = "Plot of analytical skills of child")
g2 <- g2 + geom_smooth(method = 'lm', se = TRUE)


grid.arrange(g1, g2, ncol = 2)
```


Plot of analytical skills of child
Child's Score
Father's IQ
165
160
155
150
110
115
120
125


Plot of analytical skills of child
Child's Score
Mother's IQ
170
165
160
155
150
100
110
120
130