1)

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

**Mean**. Each student reports a numerical value: a number of hours.

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**. Each student reports a number, which is a percent-age, and we can average over these percentages.

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

**Proportion**. Each student reports Yes or No, so this is a categorical variable and we use a proportion.

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

**Mean**. Each student reports a number, which is a percentage

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**. Each student reports whether or not she expects to get a job, so this is a categorical variable and we use a proportion.

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.

Standard Error: 1.2%
95% Confidence Interval = point estimate (+ or -) Z * Standard Error
95% Confidence Interval = (0.45 (+ or -) 1.96 * 0.012) = ( 42.65% to 47.35% )
At a confidence interval of 95 %
42.65% to 47.35% of US Adults may report that they live with one or more chronic conditions
Since the point estimate reported is within the 95% confidence interval, the report is 95% of the times accurate.

95% confidence interval indicates a 95% precision of the reported value of point estimate. Thus, it covers the true mean value with 95% probability.

3)

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

H0 :- One ounce (28 gram) serving of potato chips is equal to 130 calories

Ha :- One ounce (28 gram) serving of potato chips is not equal to 130 calories

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

Z= (x-μ)/S.E

We have, x=136

μ=130

n=35

S.E= Standard deviation/√n

=17/√35

=17/5.9160797830996160425673282915616

=2.8735244660769563635327023130442

Z=136-130/ 2.8735244660769563635327023130442

= 2.0880281587410409562002335146688

P value:

2*(1-pnorm(2.0880281587410409562002335146688))= 0.03679529

OR

```
> pnorm(2.088,lower.tail=FALSE)+pnorm(-2.088,lower.tail=TRUE)
[1] 0.03679783
```

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?

If p value is greater than significance level we cannot reject null hypothesis. For alpha=0.05, P value is lower than the significance value. Thus, we reject the null hypothesis. For alpha=0.025, P value is higher than the significance value. Thus, we accept the null hypothesis.

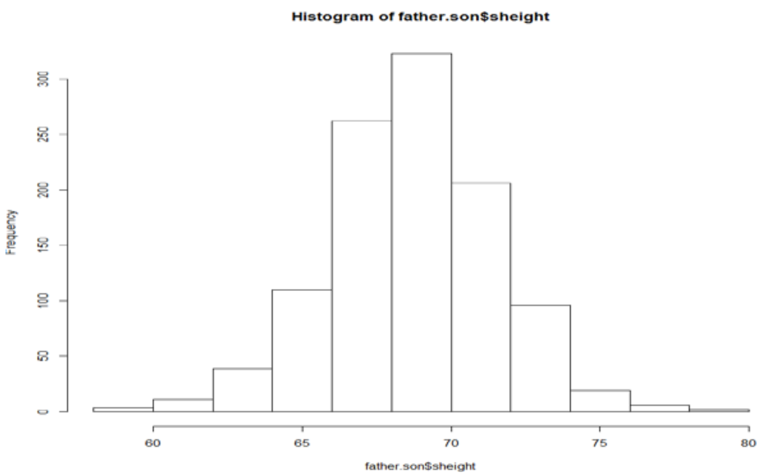
Current p value is greater than 0.025, thus, making the hypothesized mean correct and proving that the one ounce of potato chip has 130 calories based on hypothesized t test. I do not want to reject the null hypothesis, as the potato chip company, I want the number of calories to be 130. Therefore my p value should be greater than alpha. So alpha should be 0.025.

4)

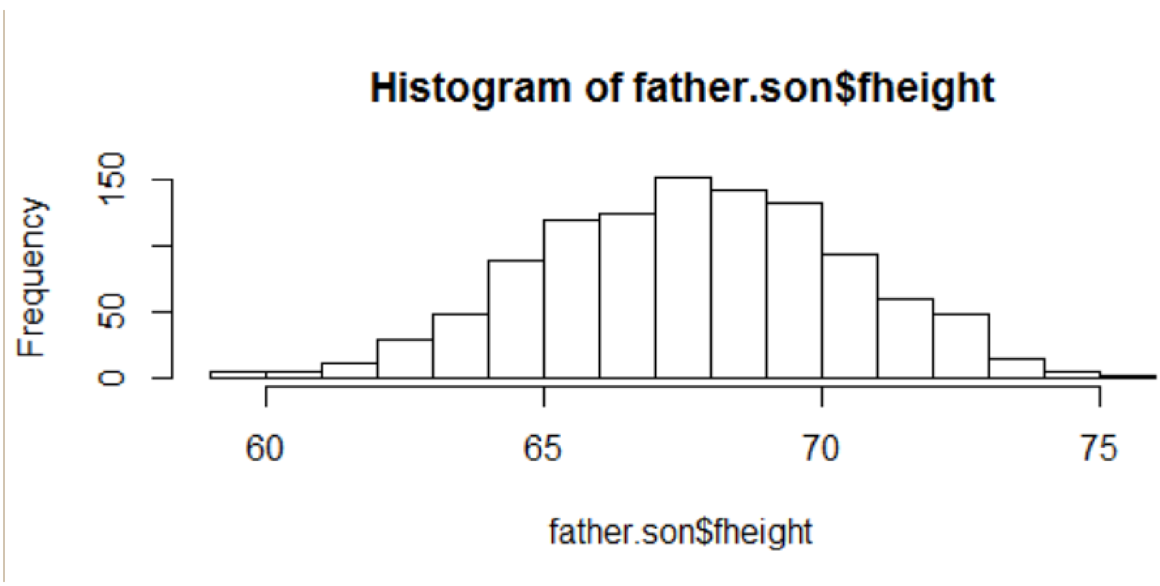
> height <- get("father.son")
> height

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?

plot(father.son$sheight)




hist(father.son$fheight)




> cor(father.son)

```
        fheight   sheight
fheight 1.0000000 0.5013383
sheight 0.5013383 1.0000000
```

Based on the above analysis, we can find that there is Medium positive correlation between Father and Son's height based on correlation. A linear model would be appropriate here since it looks like a normal distribution.

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.

```
> lm(formula = sheight ~ fheight, data = father.son)

Call:
lm(formula = sheight ~ fheight, data = father.son)

Coefficients:
(Intercept)     fheight
    33.8866      0.5141
```

**Sons Height = 33.8866 + 0.5141 x Father's Height**
When father's height=0, we get the son's height as 33.8866. 0.5141 is the slope which indicates change in son's height per unit change in father's height. For each unit increase in father's height, the son's height increases, on average by 0.5141 units.

c) (5 pt) Find the 95% confidence intervals for the estimates.

```
> mod<-lm(formula = sheight ~ fheight, data = father.son)

> summary(mod)

Call:
lm(formula = sheight ~ fheight, data = father.son)

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
```

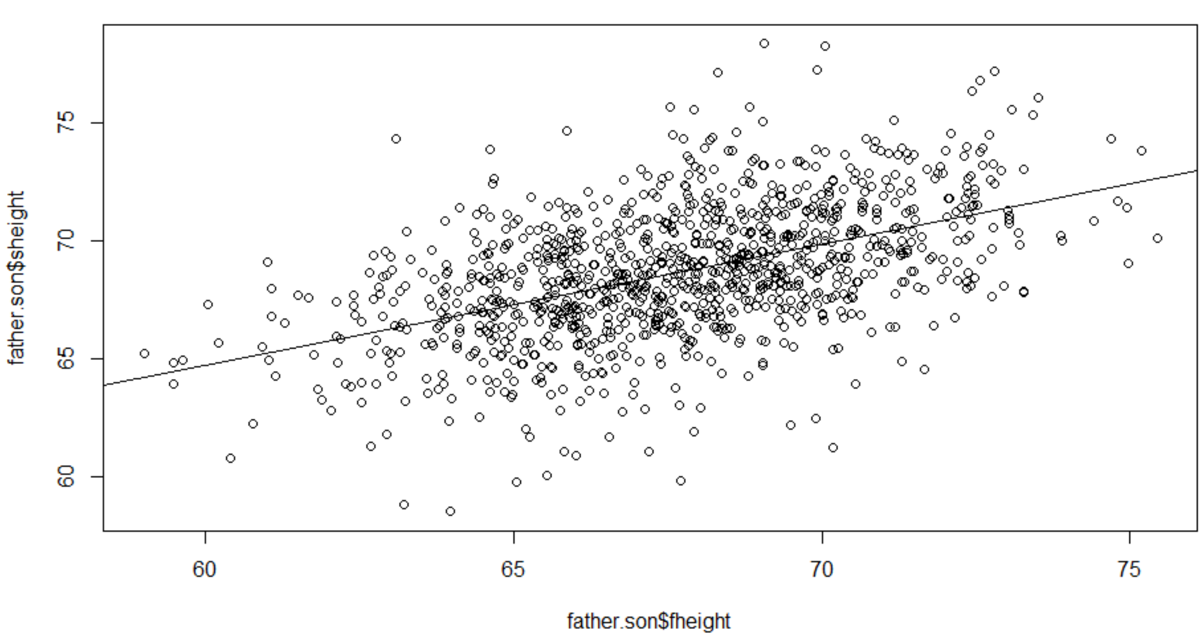
Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared:  0.2513,	Adjusted R-squared:  0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16

> confint(mod,level=0.95)

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

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

**plot(father.son$fheight,father.son$sheight)**
**mod<-lm(formula = sheight ~ fheight, data = father.son)**
**abline(mod)**



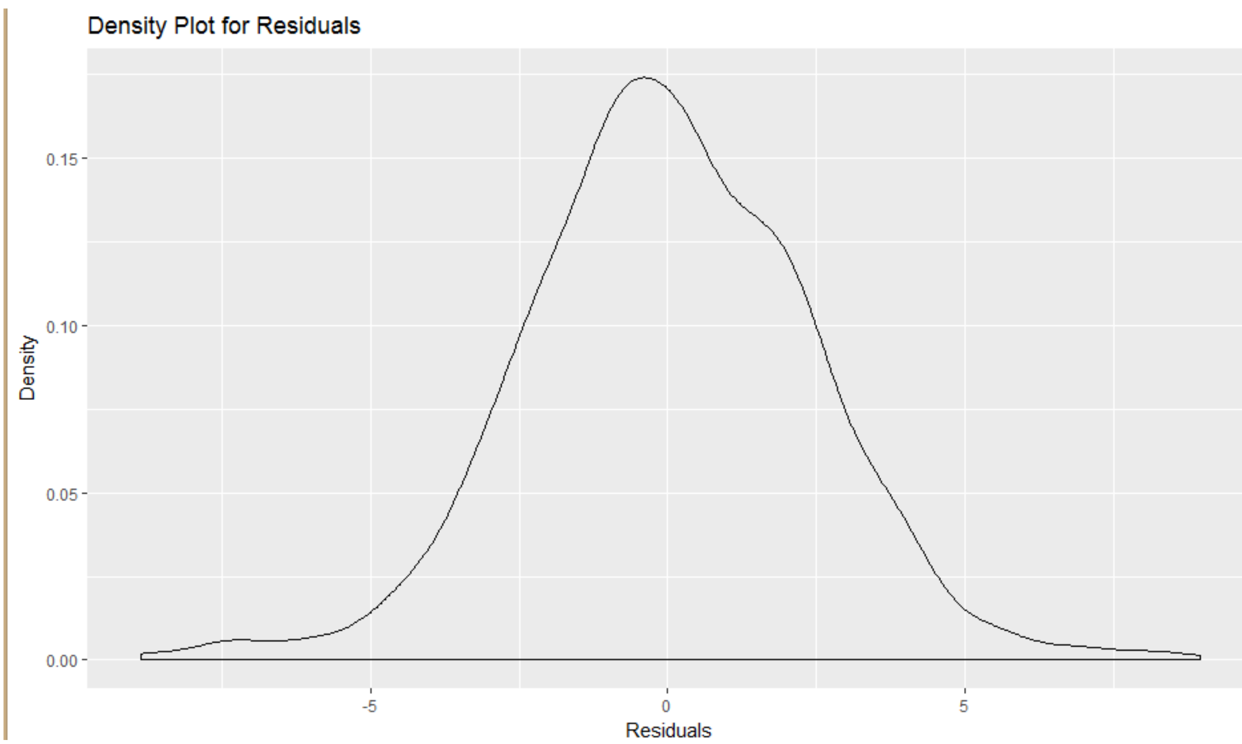

between the data and the least squares regression line looks positive.

e) (5 pt) Produce a visualization of the residuals versus the fitted values.
resid(mod)

```
> resid(mod)
           1            2            3            4            5            6            7            8            9           10           11           12 
-7.549320518 -3.189432294 -3.937262188 -4.897126876 -1.035698698 -2.043843444 -3.410828758 -3.165011904 -3.236827219 -4.334628227  1.022303623 -0.888502884 
          13           14           15           16           17           18           19           20           21           22           23           24 
-0.939312002 -1.389889738 -1.848607024 -2.591991494 -2.989964076 -2.052904112 -3.438394247 -2.717010607 -3.605699113 -4.121340976  0.546305037 -0.312543404 
          25           26           27           28           29           30           31           32           33           34           35           36 
-0.875422298 -1.312839748 -1.192957249 -1.230181614 -1.812453255 -1.615901663 -2.375460072 -2.204042616 -1.619202118 -3.046443067 -2.648375866 -2.626129125 
          37           38           39           40           41           42           43           44           45           46           47           48 
-2.805758011 -3.212520458 -4.390581934  1.251267433  0.423048888 -0.045610592  0.017480915 -0.497696232 -0.474796414  1.651610444 -0.093493379 -6.433844280 
          49           50           51           52           53           54           55           56           57           58           59           60 
-1.501642356 -0.690917366  0.167196487 -0.296185711 -0.663097904 -0.437084264  0.035412008 -0.270065897 -0.778626280 -0.794462663 -0.920354365 -1.725711657 
          61           62           63           64           65           66           67           68           69           70           71           72 
-1.230643968 -0.620670541 -1.364184860 -1.572385589 -1.470870305 -1.513833405 -1.745407337 -2.069684088 -2.846581324 -3.329580662 -2.946462676 -3.640587809 
          73           74           75           76           77           78           79           80           81           82           83           84 
 1.860225243  2.281611743  1.304137765  1.278954303  1.282886041  0.483236276  0.464206224  0.684499731  0.541204255  0.635911455 -0.167036297  0.447075292 
          85           86           87           88           89           90           91           92           93           94           95           96 
-0.375922595 -0.216290319  0.116000909 -0.540760234 -0.514137299 -0.965763441 -0.408536148 -1.177500783 -0.045185671 -1.231809237 -1.256862315 -1.283594057 
          97           98           99          100          101          102          103          104          105          106          107          108 
-0.807166782 -1.300360878 -0.583115367 -1.592844342 -1.694391204 -2.726917798 -1.696968809 -2.808427293 -3.721415968  2.481860301  2.991852489  2.191823951 
         109          110          111          112          113          114          115          116          117          118          119          120 
 2.316841460  1.663529726  1.509323610  1.810568992  1.750030102  1.182320365  1.412419651  0.593792097  0.609404304  0.130282046  0.484382391  0.873493002 
         121          122          123          124          125          126          127          128          129          130          131          132 
 0.234369977 -0.507444670  0.256542544  0.113457481  0.050642153 -0.456481594 -0.714303903 -1.006836579 -0.642387828 -0.821788545 -0.904210026 -1.227948506 
         133          134          135          136          137          138          139          140          141          142          143          144 
-1.903205886 -3.391252138  2.614688577  2.476442442  2.256267373  2.481986131  2.739712775  1.972221011  1.206614827  1.720636825  0.898841785  0.733899166 
         145          146          147          148          149          150          151          152          153          154          155          156 
 1.800290795  0.561337916  0.719309646  0.393519322  0.726053610  0.182004139  0.015308736  0.133562580 -0.366483521 -0.228146526 -0.531668437 -0.322913060 
         157          158          159          160          161          162          163          164          165          166          167          168 
```

**resid(mod)**
**plot(resid(mod))**
**hist(resid(mod))**

ggplot(mod, aes(x=residuals(mod))) + geom_density() + labs(x='Residuals',y='Density',title='Density Plot for Residuals')

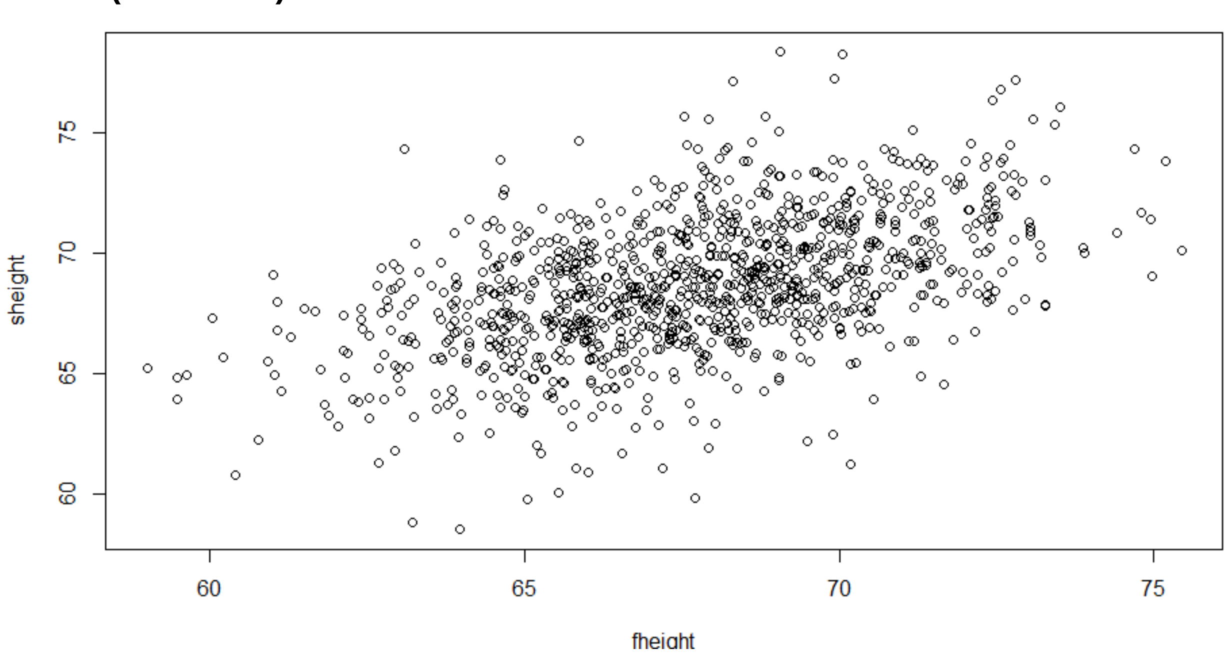



**The plot seems normally distributed. Therefore we can apply a linear model.**

```
> names(father.son)
[1] "fheight" "sheight"
```

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

**mod1<-(father.son)**




```
> new.df <- data.frame(fheight=c(50,55,70,75,90))
> predict(mod,newdata=new.df)
       1        2        3        4        5
59.59126 62.16172 69.87312 72.44358 80.15498
```

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

```
Coefficients:
  (Intercept)  height$fheight
     33.8866          0.5141
```

With 1 unit increase in father's height, the son's height increases approximately by 0.514 units. Therefore, as the slope is positive, statistically, we can say that the estimates are significant (small p-value: < 2.2e-16).

Practically, it is not significant as the intercept value of 33.886 means that when the father's height is zero, still the son's height is 33.8866. This is practically not possible.

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.

**install.packages("openintro")**
**library(openintro)**
**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").

**#score and father's IQ**

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

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
```

**#score and mother's IQ**

```
> mod3 <- lm(score ~ motheriq, data = gifted)
> summary(mod3)

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)

**Father’s IQ slope is 0.2501**
**Mother’s IQ slope is 0.4066**

```
> confint(mod2,level=0.95)
                 2.5 %      97.5 %
(Intercept) 78.1548748 182.7039518
fatheriq    -0.2051068   0.7053687
```

**(CIlower:** -0.2051068 **, CIupper:** 0.7053687**)**

```
> confint(mod3,level=0.95)
                 2.5 %      97.5 %
(Intercept) 86.9972563 135.1886542
motheriq     0.2029815   0.6102077
```

**(CIlower:** 0.2029815 **, CIupper:** 0.6102077**)**

c) (5 pt) How are these interpreted?

Regression 1: Child score vs father’s IQ
The 95 % confidence interval for $\beta 0$ is [78.1548748 182.7039518] and the 95 % confidence interval for $\beta 1$ is [-0.2051068, 0.7053687]. Therefore, we can conclude that when father’s IQ is zero, child’s score will, on average, fall somewhere between 78.1548748 and 182.7039518 units. Furthermore, for each 1 unit increase in father’s IQ, there will be an average increase in child’s score between -0.2051068 and 0.7053687

units. Here, for each unit increase in father's IQ, the child's score increases, on average by 0.2501 unit.
The lower end of the 95% confidence interval for father's IQ with child's score is negative. This indicates negative correlation. However, the higher end of the interval indicates positive correlation.

Regression 2: Child score vs mother's IQ
The 95 % confidence interval for $\beta 0$ is [86.9972563, 135.1886542] and the 95 % confidence interval for $\beta 1$ is [0.2029815   0.6102077]. Therefore, we can conclude that when mother's IQ is zero, child's score will, on average, fall somewhere between 86.9972563 and 135.1886542 units. Furthermore, for each 1 unit increase in mother's IQ, there will be an average increase in child's score between 0.2029815 and 0.6102077 units. Here, for each unit increase in mother's IQ, the child's score increases, on average by 0.4066 units. Both the ends of the 95% confidence interval point towards a positive correlation between child's score and mother's IQ.

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

There is less positive correlation between father's IQ and child's score and there is a littl e higher, yet less positive correlation between mother's IQ and child's score. For score v s father's IQ, the R squared value is close to 0 and f statistic value is close to 1. Therefo re, they are not significant.
For score vs mother's IQ,
Adjusted R-squared:  0.3065
F-statistic: 16.47 on 1.Therefore, these values are statistically significant.

Child's score increases by 0.2501 as father's IQ increases by 1 unit. Child's score Increases by 0.4066 as mother's IQ increases by 1 unit.