Assignment 3

Question 1

1 a. Mean

Since the individual responses are numerical, the parameter of interest is Mean.

1 b. Mean

Since the individual responses are numerical, the parameter of interest is Mean.

1 c. Proportion

Since the response is categorical – cite information or do not cite information, the parameter of interest is Proportion.

1 d. Mean

Since the individual responses are numerical, the parameter of interest is Mean.

1 e. Proportion

Since the response is categorical – yes or no, the parameter of interest is Proportion.

Question 2

Proportion, p =0.45

Standard error (SE) = 0.012

Since a normal model may be used and confidence interval is 95%, z = 1.96

Confidence Interval = p +(-) z * SE

= 0.45 +(-) 1.96 * 0.012

= (0.4265, 0.4735)

= 42.65% to 47.35%

We are 95% confident that 42.65% to 47.35% US adults live with one or more chronic conditions.

Question 3

a. n = 35

Sample mean = 136 cal

Standard deviation = 17 cal

Since, sample size > 30, we can assume a normal distribution.

$\boldsymbol{H_0}$ = Nutrition label is not lying i.e. one ounce (28 gm) of potato chips has 130 calories. $\mu = 130$

$\boldsymbol{H_A}$ = Nutrition label is lying i.e. one ounce (28 gm) of potato chips does not have 130 calories. $\mu \neq 130$

b. Test Statistic = (sample mean - μ)/SE

SE = standard deviation / $\sqrt{n}$

= 17/$\sqrt{35}$

Test Statistic = 136 – 130 /(17$\sqrt{35}$ /)
= 2.088

p-value

pnorm(-2.088, lower.tail = T) + pnorm(2.088, lower.tail = F)
= 0.0036

3. c. We get type 1 error when we reject the null and null hypothesis is true. We get type 2 error when we fail to reject the null and null hypothesis is to be rejected. If the potato chip company is doing the test and I am the potato chip company, then type 2 error is more dangerous and costly for me; if the label is lying and we identify the label as not lying. This would give a wrong picture to the customers and the purpose of performing this test would be defeated. Hence, I would try to minimize type 2 error and trade off type 1 error by taking a slightly higher value of alpha, i.e. 0.05.

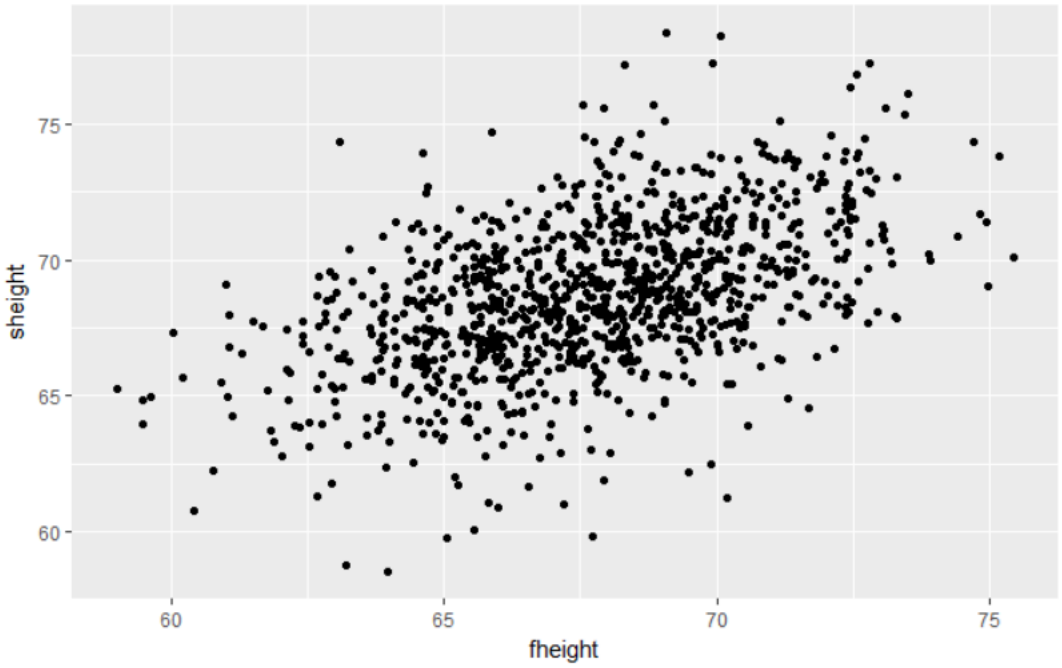
 # Question 4

```
install.packages("ggplot2")
install.packages("UsingR")
library(ggplot2)
library(UsingR)

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

 # Question 4(a)

```
ggplot(height, aes(fheight, sheight)) + geom_point(na.rm = TRUE)
cor(height$fheight, height$sheight)
```



With increase in father's height, son's height also increases. Also there is a medium positive correlation between both the attributes.
#There is a strong linear relationship as seen in the graph
A linear model would definitely be appropriate here

 # Question 4(b)

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

Intercept: 33.8866, slope: 0.51409
y = 33.8866 + 0.51409*(fheight)

#If Father's height is zero, son's height is 33.8860.
For every one unit change in Father's height, son's height increases by 0.51409.

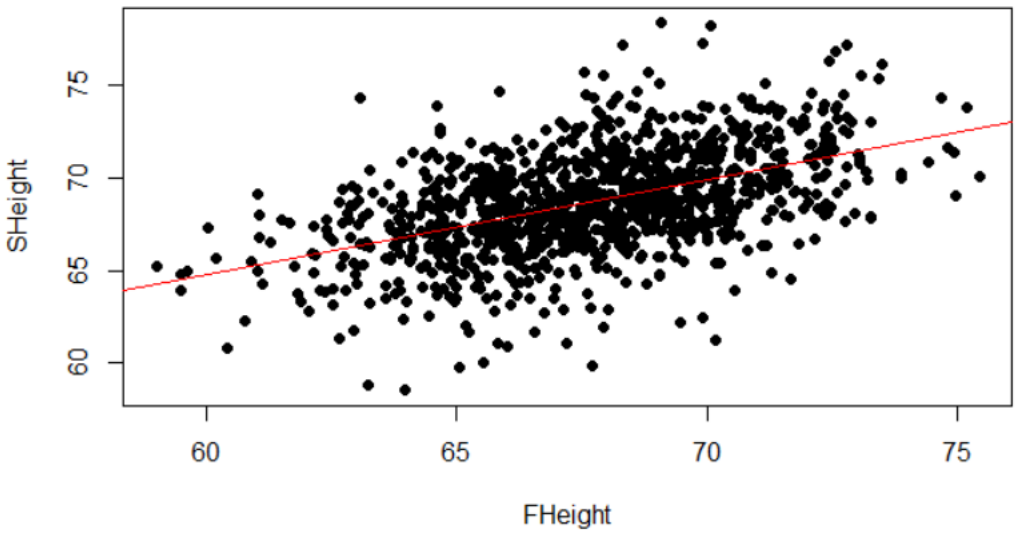
 # Question 4(c)

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

CI for intercept: 30.291 to 37.481
CI for fheight: 0.461 to 0.567

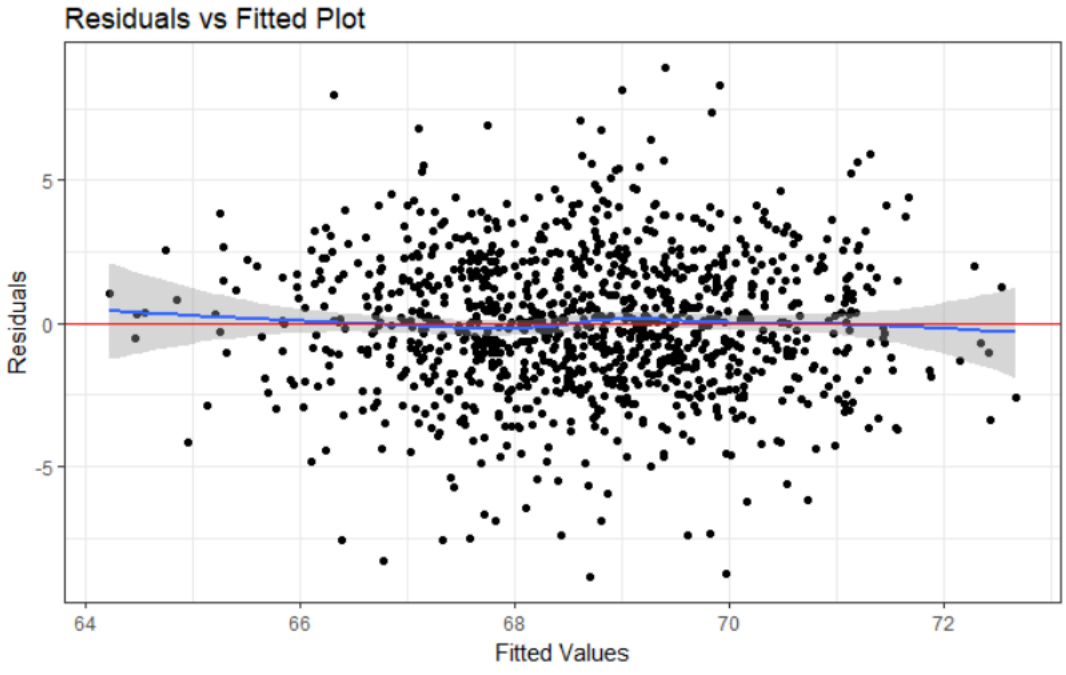
 # Question 4(d)

```
plot(height$fheight, height$sheight, pch=16, xlab = 'FHeight', ylab='SHeight')
abline(fsmodel, col=2)
```



 # Question 4(e)

```
p2 <- ggplot(fsmodel, aes(.fitted, .resid)) + geom_point()
p2 <- p2 + stat_smooth(method = "loess")
p2 <- p2 + geom_hline(yintercept = 0, col="Red")
p2 <- p2 + xlab("Fitted Values") + ylab("Residuals")
p2 <- p2 + ggtitle("Residuals vs Fitted Plot") + theme_bw()
p2
```



The plot is pretty symmetrically distributed and tendng to cluster towards the middle.

Since, the variation is uniform for all values, we can say that the linear model is a good fit for our data.

 # Question 4(f)

```
predict(fsmodel, newdata = data.frame(fheight = c(50, 55, 70, 75, 90)))
```

Predicted Heights are: 9.59126, 62.16172, 69.87312, 72.44358, 80.15498

 # Question 4(g)

For everyone unit change in Father's height, son's height increases by 0.51409. If Father's height is zero, son's height is 33.8860.
Father's height is statistically significant as p-value is very small. However, practically, it doesn't make sense; If father's height is zero then son's height is 33.8860 but father's height can never be zero. Hence, it is not practically significant.

 # Question 5

```
install.packages("openintro")
library(openintro)
data(gifted)
View(gifted)
```

 # Question 5(a)

```
model1 <- lm(score ~ fatheriq, data = gifted)
summary(model1)

model2 <- lm(score ~ motheriq, data = gifted)
summary(model2)
```

#Question 5(b)

#Slope for model1(Father) = 0.2501
Slope for model2(Mother)= 0.4066

```
confint(model1, level=0.95)
```

Range of slope for 95% confidence interval: -0.205 to 0.705

```
confint(model2, level=0.95)
```

Range of slope for 95% confidence interval: 0.202 to 0.610

#Question 5(c)

For one unit change in father's IQ, child's test score increases by 0.2501
For one unit change in mother's IQ, child's test score increases by 0.4066

#Question 5(d)

#Since the p-value is high for model1(FatherIQ) and not significant, we can say that child's score is not very much related to Father's IQ
Since p-value is very low for model2(MotherIQ) and very much significant, we can say that child's score is strongly associated with Mother's IQ.