Assignment 1

Manyi Luo - 1003799419

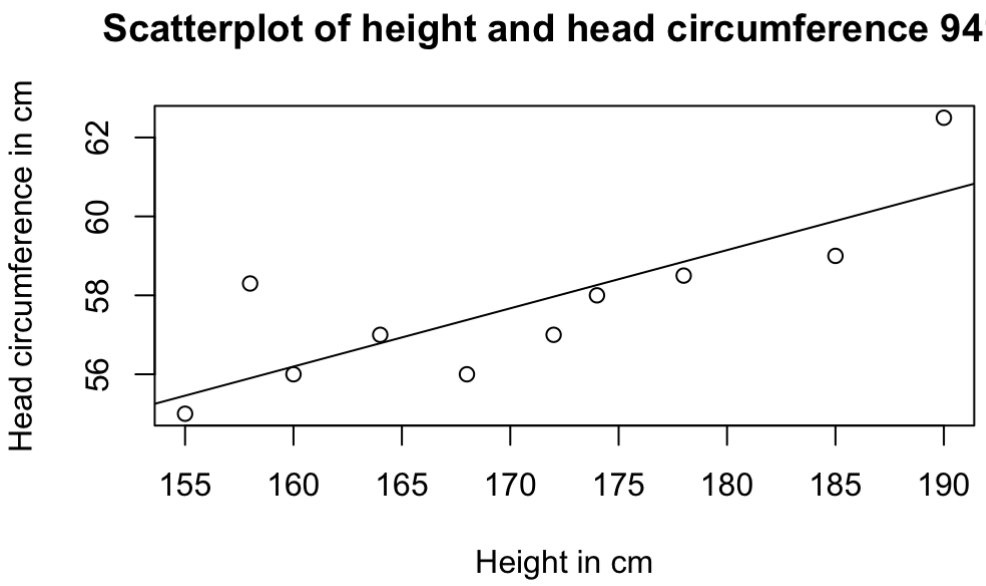
9/21/2019

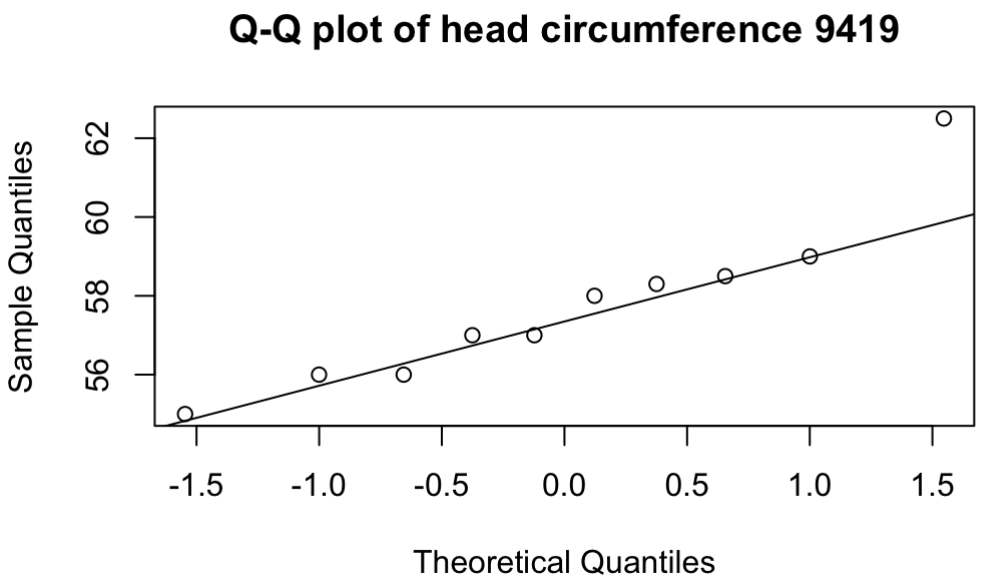
Solution

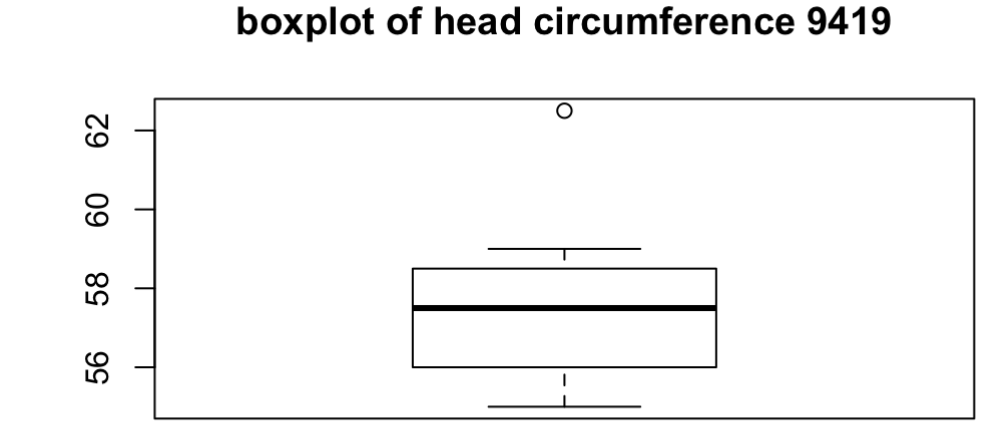
1. Using words, give a title to summarise what your assignment #1 is about. Specify your explanatory variable and response variable and, give a brief explanation for your choice.

My assignment is about finding the relationship between height and head circumference using simple linear regression. So my title is: determining the relationship between height and head circumference. My explanatory variable (x) is height and the response variable (y) is head circumference. I choose height as explanatory variable (x) because in daily observations, taller people may have greater stature, so I am wondering whether higher people may have greater head circumference.

2. Draw at most 3 plots to visually describe your data. Is your response variable approximately Normal?


Scatterplot of height and head circumference 9419
Head circumference in cm
56
58
60
62
155
160
165
170
175
180
185
190
Height in cm


Q-Q plot of head circumference 9419
Sample Quantiles
56
58
60
62
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles


boxplot of head circumference 9419
56
58
60
62

Through the scatterplot, we can see that ten data points are located around the fitted line, eventhough some of them are above and some of them are below. There are two outliers identified, and the heights are 158cm and 190cm (head circumference are 58.3cm and 62.5cm) respectively. And based on the trend line, we can see height and head circumference follows a positive linear relationship: as height increases, head circumference increases.

We can see the response variable (head circumference) is approximately normal based on the Q-Q plot. This is because, as data points evenly distributed through out the theoretical quantiles (x axis), most of them lies around the fitted line, and there is only one outstanding outlier located at theoretical quantile 1.5.

From the boxplot, we can clearly see that the distance between the 25th percentile and the median and the distance between the median and the 75th percentile are approximately the same, eventhough the distance between the median and the 75th percentile is slightly larger (since most of the data gather between 56-58cm), making it slightly left-skewed. There is only one outstanding outlier at 62cm, which doesn't really affect the distribution. Above all, the response variable is approximately Normal with a weak tendency of left-skewed.

3. Numerically describe the centre, spread and any unusual points of your variables/data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   56.25   57.50   57.73   58.45   62.50
```

```
## [1] 2.2
```

```
## [1] 2.104519
```

For variable headcircumference:

We can clearly see from the first chart that mean is 57.73cm and median is 57.5cm. And the interquartile range is calculated as 2.2.

Spread(standard deviation) of data is calculated as 2.104519.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.0   161.0   170.0   170.4   177.0   190.0
```

```
## [1] 16
```

```
## [1] 11.62564
```

For variable height:

We can clearly see from the first chart that mean is 170.4cm and median is 170cm. And the interquartile range is calculated as 16.

Spread(standard deviation) of data is calculated as 11.62564.

Also referring to the scatterplot from part (3), we can identify two unusual points of data, which are (158cm, 58.3cm) and (190cm, 62.5cm) respectively (height, headcircumference). We can clearly see that (190cm, 62.5cm) is an outlier because its the maximum, and it deviates a lot from the other data. Also, we might deduce that (158cm, 58.3cm) is an outlier from the data of variable headcircumference. We can see that the mean is slightly greater than the median, meaning that the existence of (158cm, 58.3cm) may drag the mean up. This also corresponds to our conclusion that the response variable is approximately Normal with a weak tendency of left-skewed. If data are sampled perfectly (eg: large enough sample size to dilute the influence of outliers), we might observe that the response variable becomes even closer to Normal.

4. Fit and describe a simple linear regression model between head circumference and height.

```
##
## Call:
## lm(formula = headcircumference ~ height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3759 -0.7776 -0.3063  0.1119  2.3996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.58745    6.33142   5.147 0.000878 ***
## height       0.14755    0.03708   3.979 0.004065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.293 on 8 degrees of freedom
## Multiple R-squared:  0.6644, Adjusted R-squared:  0.6224
## F-statistic: 15.84 on 1 and 8 DF,  p-value: 0.004065
```

We set up the linear regression model to describe the relationship between height and head circumference, as height is the explanatory variable and head circumference is the response variable. Therefore, our simple linear regression model can be expressed as $Y_i = \beta_0 + \beta_1 X_i + e_i$, where Y_i is each individual's head circumference, β_0 is the true intercept of linear regression line, β_1 is the true slope of linear regression line, X_i is each individual's height and last but not least, the e_i is the random error associated with each data.

Using our ten sampling data points to find the fitted regression line, we can calculate the estimated β_0 (the intercept) is 32.58745, while the estimated β_1 equals 0.14755. Estimated β_0 represents that the head circumference is at least 32.58745cm (when height is 0cm, head circumference is 32.58745cm). The estimated β_1 represents the slope parameter, meaning that 1cm increase in height results 0.14755cm increase in head circumference. So the simple linear regression model can be represented as $\hat{Y}_i = 32.58745 + 0.14755\hat{x}_i$.

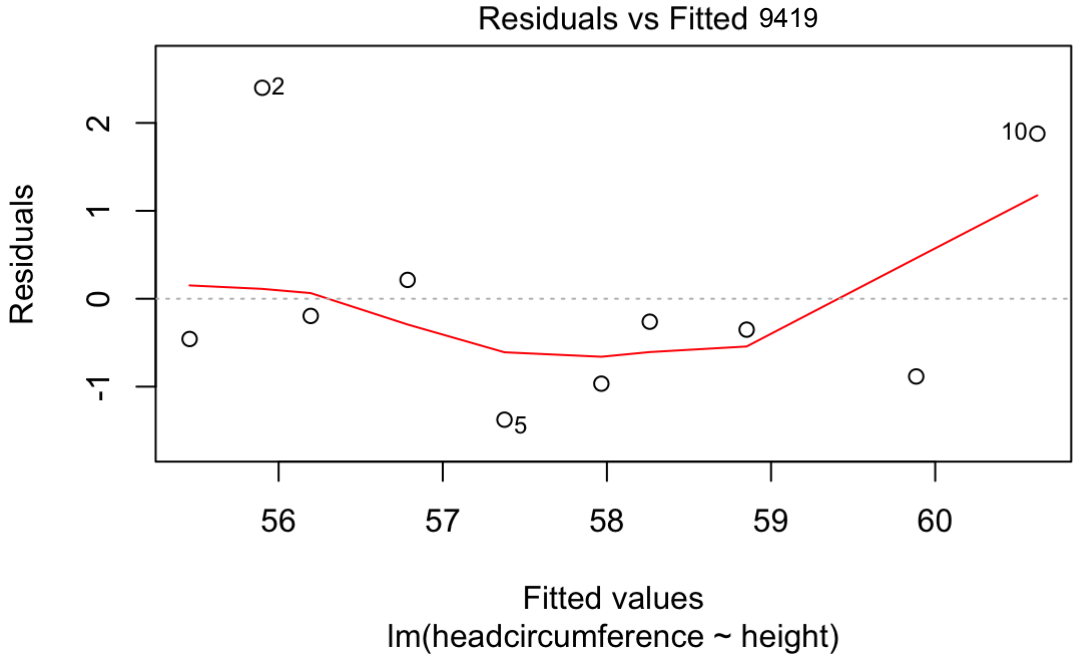
5. Are the regression parameters zero? Interpret the estimates of the regression

parameters (p-value is held for four decimal places).

The p-value of estimated β_0 is 0.0009. This comes from setting H_0 as $\beta_0 = 0$ and H_1 as $\beta_0 \neq 0$, and with a significance level of 5%, we have strong evidence to reject H_0 because 0.0009 is much more smaller than 0.05, meaning that we support H_1. β_0 is not likely to be zero.

The p-value of estimated β_1 is 0.0005. This is resulted from setting H_0 as $\beta_1 = 0$ and H_1 as $\beta_1 \neq 0$, and with a significance level of 5%, we have strong evidence to reject H_0 because 0.0005 is much more smaller than 0.05, meaning that we support H_1. β_1 is not likely to be zero either.

6. Visually and numerically describe the residuals of the fit in part (4) above. Use at most one plot (p-value is held for four decimal places.



From the graph above, we can see that only 5 out of 10 data points lie around the trend line. We cannot identify a pattern in residuals' variation, as they all randomly spread out around the line. Therefore, the assumption of constant variance is satisfied.

And analyzing numeric data generated from part (4), the residual standard error is 1.293 on 8 degrees of freedom. The p-value of estimated residual standard error is 0.0041. This comes from setting H_0 as $residual\ error = 0$ and H_1 as $residual\ error \neq 0$, and with a significance level of 5%, we have evidence to reject H_0 because 0.0041 is smaller than 0.05, meaning that we support H_1. Residual error is not likely to be zero, which coincides with our graph.

What's more, the multiple R-squared is 0.6644, meaning that 66.44% of change in head circumference can be explained by the change in height. This also means that the linear regression model we set up previously is well fitting.

7. Identify a potential lurking variable. Describe at least one other issue or limitation of your fit.

A potential lurking variable is gender, which can influence both height and head circumference, distorting the linear relationship. For example, female might have lower average height and bigger head circumference comparing to male.

One of the limitation of my fit is that, because the sample size in this case is small (only 10 data randomly selected from 60 data), this might result rather high linearity or poor linearity. Statistically, if we want to get a representative result, we need to make our sample size greater or equal to 30.

8. Identify another pair of variables to explore a simple linear regression model. Specify the response and explanatory variable.

Another pair of variables to explore might be weight and head circumference. But in order to make our sampling data more specific, we need to elimiate the effect of lurking variable, for exmaple: we need to specify sex before collecting data of weight and head circumference. In this case, the explanatory variable is weight and the response variable is head circumference. And the simple linear regression model can also be used to find the relationship between weight and head circumference, for example: as weight increases, head circumference also increases.

Appendix

```
Q2
#The result of doing a random sampling of 10 data from a1_60_data.csv (60 data in total)
without replacement.
height = c(155,158,160,164,168,172,174,178,185,190)
headcircumference = c(55,58.3,56,57,56,57,58,58.5,59,62.5)

#The first plot
h_h_manyi = lm(headcircumference~height)
plot(height, headcircumference, xlab="Height in cm", ylab="Head circumference in cm",mai
n="Scatterplot of height and head circumference 9419")
abline(h_h_manyi)

#The second plot
qqnorm(headcircumference, main="Q-Q plot of head circumference 9419")
qqline(headcircumference, main="Q-Q plot of head circumference 9419")

#The third plot
boxplot(headcircumference, main="boxplot of head circumference 9419")

Q3
summary(headcircumference)
sd(headcircumference)
IQR(headcircumference)
summary(height)
sd(height)
IQR(height)

Q4
h_h_manyi = lm(headcircumference~height)
summary(h_h_manyi)

Q6
h_h_manyi = lm(headcircumference~height)
plot(h_h_manyi, which = 1)
```