PS-11

Dilip Nikhil Francies

2023-11-28

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
getwd()
```

```
## [1] "C:/Users/dilip/OneDrive - Indiana University/Stats/PS11"
```

Question 1: Unusual

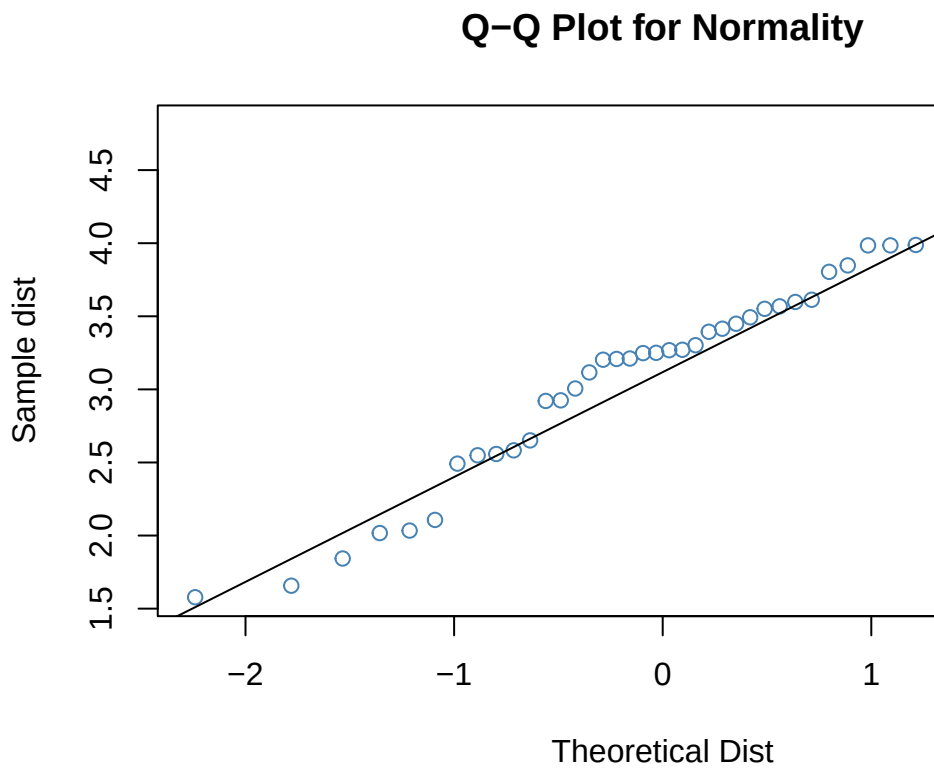
Let' read the given data from the given link:

```
unusual <- matrix(scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/unusual.dat"),ncol = 2, byrow = TRU

#convert this matrix into a data frame
unusual_df <- as.data.frame(unusual)
unusual_df$V1
```

```
##  [1] 4.813 3.208 2.034 3.804 3.449 4.025 4.189 1.578 2.558 3.302 3.493 3.848
## [13] 1.657 2.017 3.268 3.211 3.988 3.394 3.985 2.925 3.250 4.482 2.492 2.549
## [25] 3.568 2.583 3.006 3.598 3.248 3.271 3.613 3.985 2.921 2.107 3.203 1.843
## [37] 3.116 3.551 3.415 2.651
```

```
qqnorm(unusual_df$V1,main = 'Q-Q Plot for Normality', xlab = 'Theoretical Dist',
       ylab = 'Sample dist', col = 'steelblue')
qqline(unusual_df$V1)
```



Question 1a: Normal probability plot

```
summary(unusual_df$V1)
```

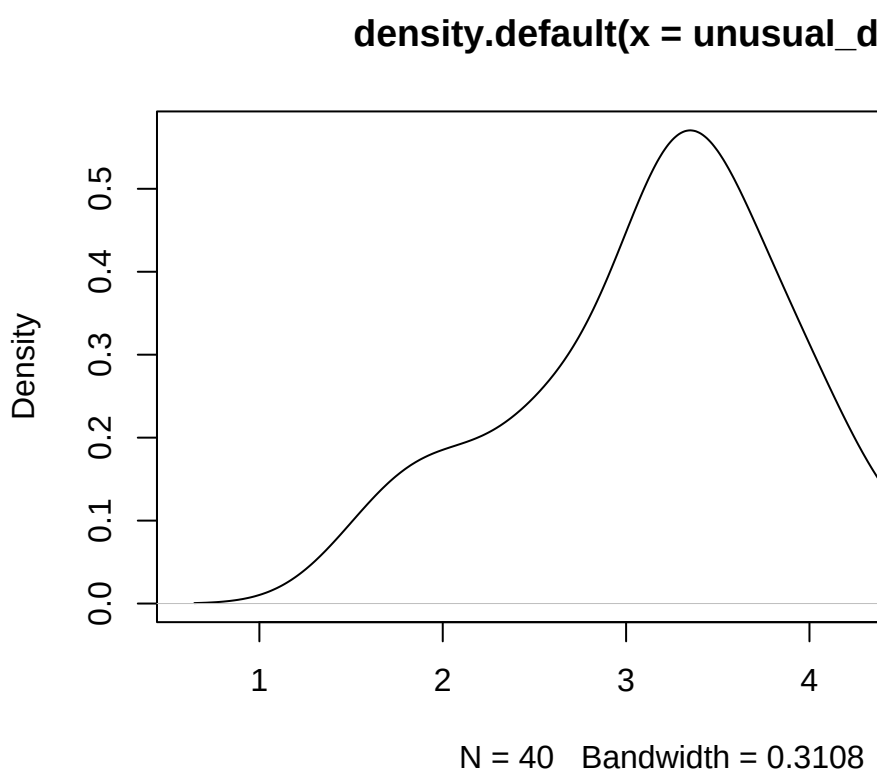
Lets quickly see the summary of the data:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.578   2.634   3.259   3.180   3.602   4.813
```

```
sd(unusual_df$V1)
```

```
## [1] 0.7577768
```

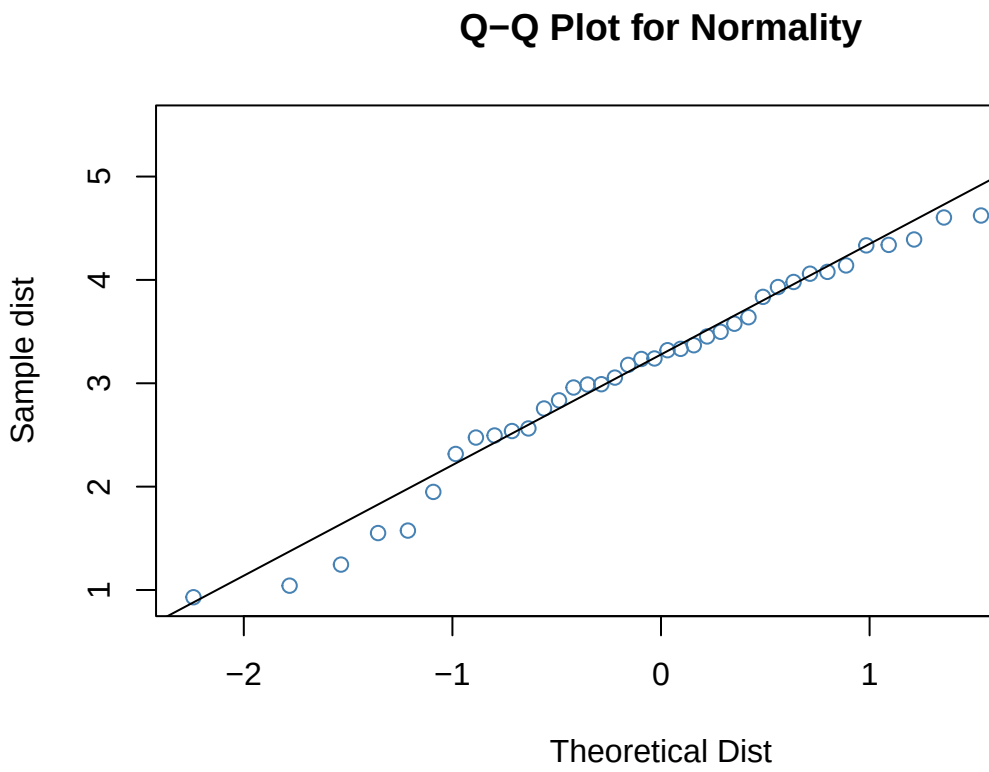
```
density <- density(unusual_df$V1)
plot(density)
```



Plot a density plot to see how the curve is:

Does the X values appear to have been drawn from a normal distribution? Well, to a certain extent yes!! Sure, there is deviation in the bottom left of the graph in QQplot, but overall, it does look like it was drawn from an approximately normal distribution. Additionally, the IQR is 1.277 times the standard deviation, which is slightly less than the usual 1.35.

```
qqnorm(unusual_df$V2,main = 'Q-Q Plot for Normality', xlab = 'Theoretical Dist',
       ylab = 'Sample dist', col = 'steelblue')
qqline(unusual_df$V2)
```



Question 1B: Normal probability

```
summary(unusual_df$V1)
```

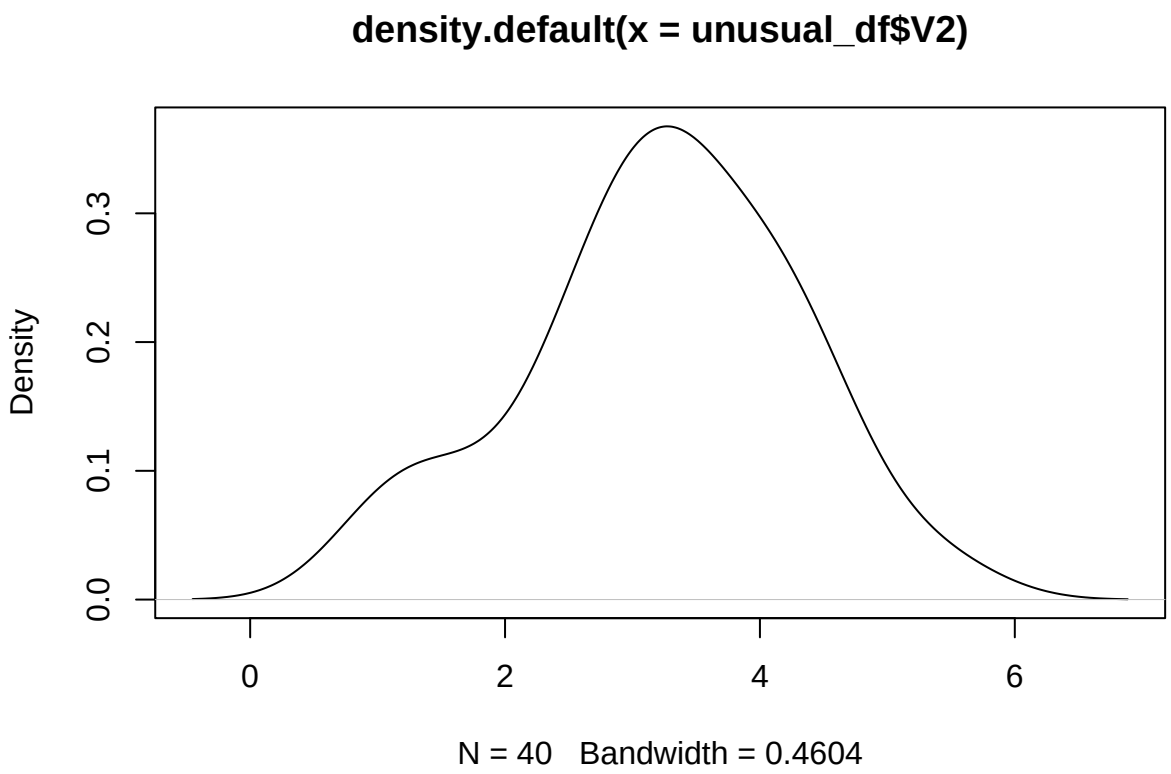
Summary of the dataset:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.578   2.634   3.259   3.180   3.602   4.813
```

```
sd(unusual_df$V1)
```

```
## [1] 0.7577768
```

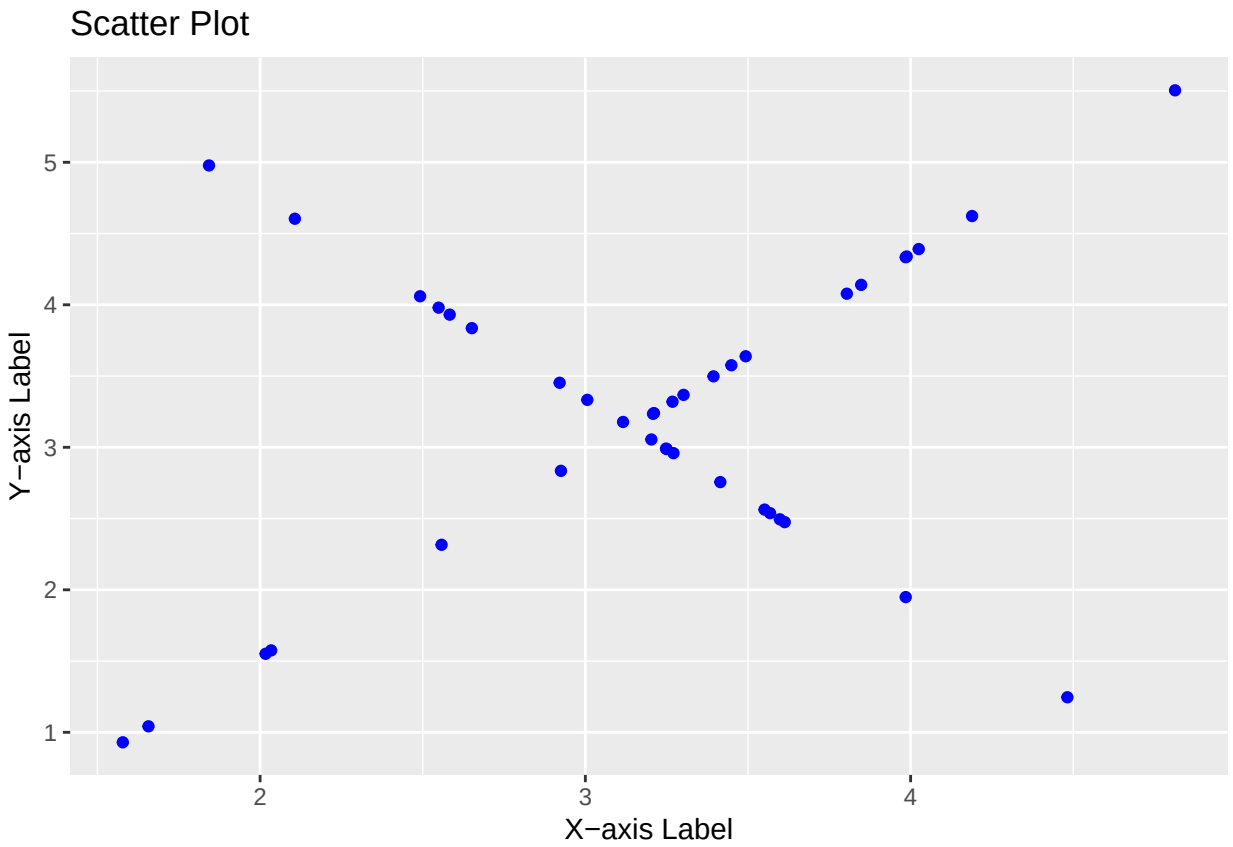
```
density_y <- density(unusual_df$V2)
plot(density_y)
```



Desnity Plot

Overall, the values of Y looks likes they are drawn from a normal distribution as the QQ plot looks right, though at the tails, the distribution does look deviated a little bit from the qqline. Moreover, the IQR is 1.32 times the standard deviation, very slightly less than the usual 1.35 times.Overall, the data does look like it was drawn from a normal distribution. perhaps, if we had more observations, we could have been more sure.

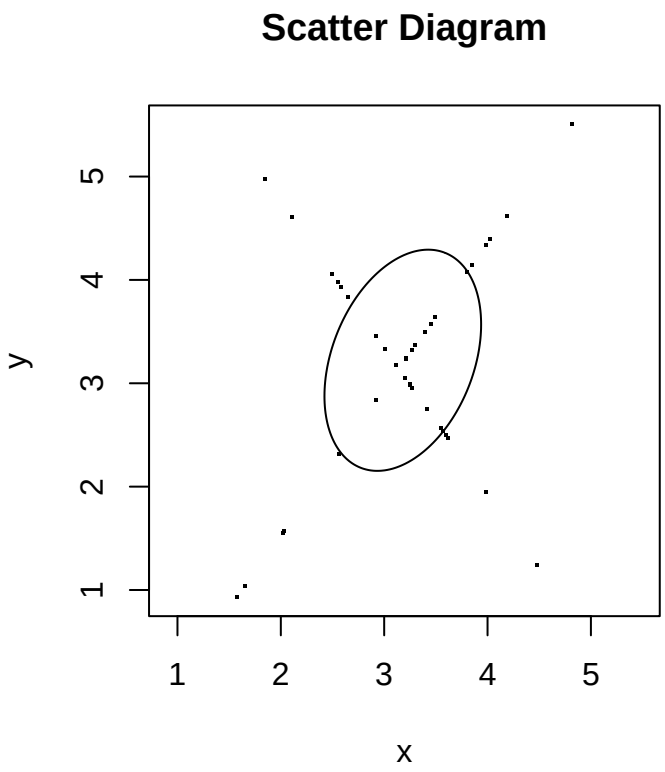
```
#scatter plot
library(ggplot2)
scatter_plot <- ggplot(unusual_df, aes(x = V1, y = V2)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot", x = "X-axis Label", y = "Y-axis Label")
# Display the scatter plot
print(scatter_plot)
```



Question 1 C:

Let's see how the ellipse depicts the data cloud.

```
source("http://mtrosset.pages.iu.edu/StatInfeR/binorm.R")
binorm.scatter(cbind(unusual_df$V1,unusual_df$V2))
```



Correlation between these two columns:

```
cor(unusual_df$V1,unusual_df$V2)
```

```
## [1] 0.3244167
```

Explanation : Well, just because two samples are drawn from a normal population, and we have Pearson's correlation of 0.32 does not mean they have to be a bivariate distribution. There is no clear trend in the scatter plot, and the ellipse does not do a great job in indicating the data cloud, if not all, very very few data points are inside the ellipse. Hence, its not drawn from a bi-variate normal distribution.

Question 2:

Answer : Yes, there is a simpler explanation for this phenomena and it is called regression towards the mean. That is, when something is exceptionally high in their first trial, the next trial would probably move towards the average performance. The reason behind the first one to be exceptionally high is may be because of a combination of luck, skill,focus etc. One may perform really well once, and may not be able to replicate it irrespective of how high the praise is. Performance on the first and second trial are not always perfectly correlated. Hence, if the first trial is exceptionally high, the second trial has very high chances of moving closer towards the average performances. Just because the performance in the second trial decreased, does not mean that the praises has a negative impact on the persons ability. There's is no causal relationship between praise and decreased in performance. Its just the phenomena of regression towards the mean.

Question 3: Missing Test Marks

```
#Given:

n <- 33
```

```r
mu_A <- 75
sd_A <- 10
mu_B <- 64
sd_B <- 12
#ellipsoidal meaning they are bi-variate variables
ro <- 0.5
```

Question 3A: Lets try and predict the marks in Test 2 given the marks in test A and its correlation.

The regression line is given by:

y = mx + c, where y = predicted values, m is the slope, x is the independent value, and c is the y-intercept.

Hence,

```r
m <- ro * (sd_B/sd_A)
c <- mu_B - (m * mu_A)

#prediction of marks in Second Test
marks_B <- m * 80 + c
marks_B
```

```
## [1] 67
```

Hence, our best prediction for Jill's test 2 would be 67, as giving the same score as test1 would likely be too high because of difference in the score distribution, and yes, these are not perfectly correlated, its 0.5.

Question 3B: Given that Jack score 76, one standard deviation above the test 2 mean, we can calculate marks in test A as:

```r
# correlation *  standard deviation units above the average
# 0.5 * 1 = 0.5

marks_A <- mu_A + 0.5 * sd_A
marks_A
```

```
## [1] 80
```

Alternatively, lets find it using the regression line:

```r
m <- ro * (sd_A/sd_B)
c <- mu_A - (m * mu_B)
#prediction y

marks_A <- m * 76  + c
marks_A
```

```
## [1] 80
```

Hence, the best prediction for Jack for test 1 would be 80, and not 85. I beleive 85 would be an likely too high considering the distribution of scores in Test 1.

Question 4: Baseball

Question 4a: This is because of the phenomena called "regression towards the mean or just regression". Given the fact that two variables are positively correlated, that is Team's win in one season (X) and team's win in next season "(Y) is 0.54, the best prediction of Y given X through regression would be above the average, but not as extreme as Y. In this case, the value of X is high (98), which is close to 1.5 standard

deviation away from the mean. Because the the number of wins predicted for next year is equal to the number of wins this year (98), from the statistical phenomena of regression towards the mean, the prediction is definitely too high, and is an over-estimation.

The best estimation would be slightly above the average number of wins of 81 games in the season, perhaps 86,87,etc something closer towards the mean may be reasonable. But purely based on the fact that the team will win the same number of games the next season where the distribution has correlation of 0.54 is likely too high. The team may end up winning all the games, or even losing all the games, but the "best prediction" will definitely be number of games closer to the mean, and not 98.

Question 4B: we know that the regression line is given by the formula:

y = mx + c where m is the slope c is the y intercept. m = correlation * sigma(y)/sigma(x)

Given,

```
ro <- 0.54
sigma_x <- 11.7       #given
sigma_y <- 11.7        # given
m <- ro * (sigma_y/sigma_x)
mu_y <- 81
mu_x <- 81
m
```

```
## [1] 0.54
```

```
c <- mu_y - (m * mu_x)
c
```

```
## [1] 37.26
```

#Therefore our best estimation would be :

```
prediction_wins <- m * 98 + c
prediction_wins
```

```
## [1] 90.18
```

Therefore, our best prediction for the number of games the team might win would be 90.18 games. [90 or 91]

Question 4C:

At least one team has won 96 games clearly means that this may be an extremely high wins in a distribution where mean is 81 and the standard deviation is 11.7. The regression prediction does not take into consideration such outliers, as the predictions are made based on the historical data and the overall trend captured from that data. Moreover, given that the correlation between a team's wins one season and their wins the next season is 0.54 does not signify a very high correlation. So, any kind of extreme values(such as this 96,95,94) wins the previous season will be predicted to more than average in the previous season, but not necessarily as high as other 98. Again, these are only the statistical predictions that are being made, the best possible prediction considering all the data points, trends etc, Hence, with regression, the extremely high values would be predicted closer towards the average of the distribution. Hence the executive's suspicions are definitely misplaced.

Question 5:adults.txt

Read the given data:

```
adults <- read.table("adults.txt",header=TRUE)
```

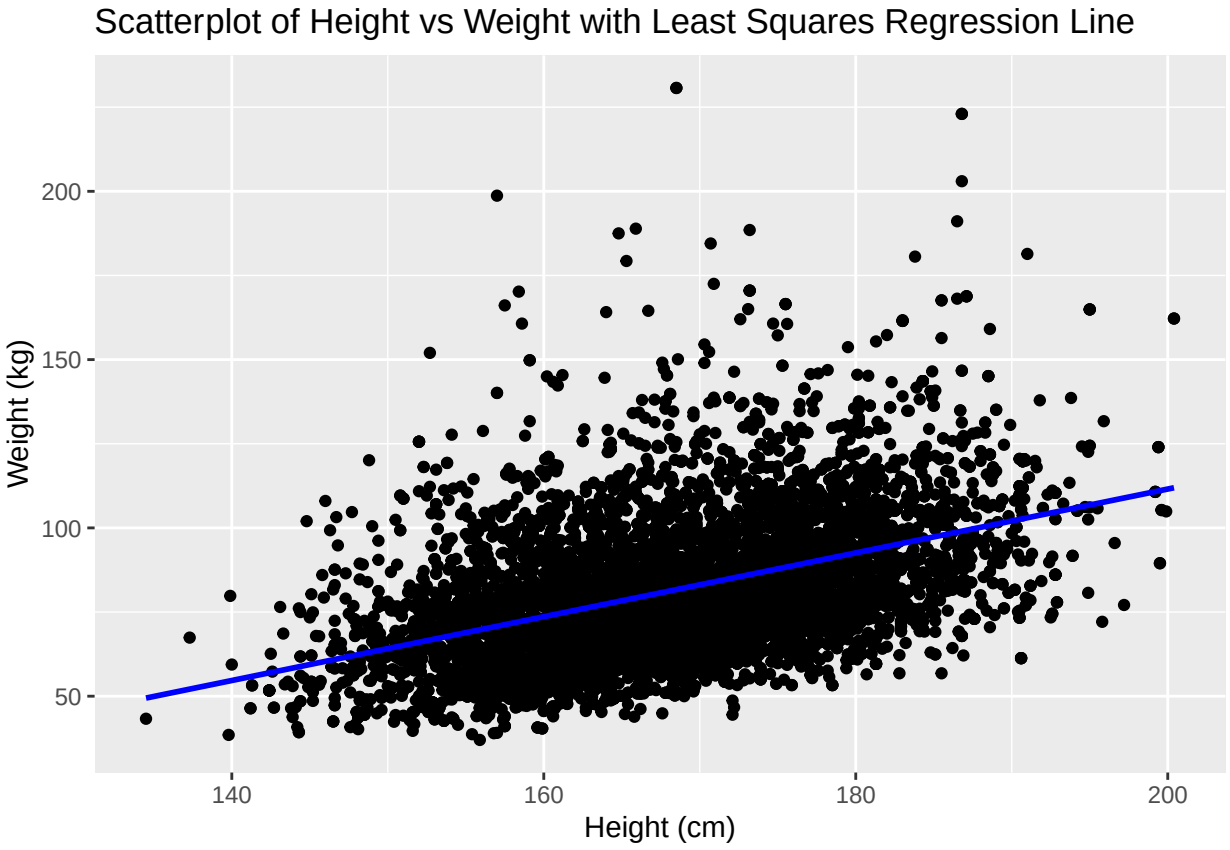
Question 5a: Using ggplot:

```
library(ggplot2)
```

```
scatterplot <- ggplot(adults, aes(x = Height, y = Weight)) +
  geom_point() +  # Scatterplot
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  labs(x = "Height (cm)", y = "Weight (kg)", title = "Scatterplot of Height vs Weight with Least Squares

# Print the plot
print(scatterplot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
scatterplot <- ggplot(adults, aes(x = Weight , y = Height)) +
  geom_point() +  # Scatterplot
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  labs(x = "Weignt (cm)", y = "Height (kg)", title = "Scatterplot of Weight vs Height with Least Squares

# Print the plot
print(scatterplot)
```

Question 5b:

```
## `geom_smooth()` using formula = 'y ~ x'
```

Question 5c: Lets fit a linear regression model to the given data and predict the weight for the given height of 180cm.

```
model <- lm(Weight~Height , data = adults)
w <- predict(model, data.frame((Height = c(180))))
# Display the predictions
print(w)
```

```
##        1
## 92.58016
```

Hence the best possible prediction using linear regression for the weight of the adult who is 180cm tall is 92.58016 Kgs.

```
model <- lm(Height ~ Weight , data = adults)
h <- predict(model, data.frame((Weight = c(92.58016 ))))
# Display the predictions
print(h)
```

Question 5d:

```
##        1
## 171.1131
```

Hence, the best possible prediction for the height of an adult who;s weight is 92.58016kgs is 171.1131 cms.

```
cor(adults$Height,adults$Weight)
```

Question 5e: Is your answer to (d) 180 cm? If not, explain why not.

```
## [1] 0.4499657
```

```
mean(adults$Height)
```

```
## [1] 168.857
```

```
sd(adults$Height)
```

```
## [1] 10.10038
```

```
mean(adults$Weight)
```

```
## [1] 82.0162
```

```
sd(adults$Weight)
```

```
## [1] 21.28064
```

Well, its not 180cm. Because the linear regression model takes into account the trends, patterns, information, variance in the entire data set and models an equation to predict the target variable. Yes, it might miss some of the information which is accounted as residual errors, and moreover the correlation between height and weight is not perfect either, its only 0.45. The model inherently regresses extreme values to towards the mean, in this case, from 180cm to 171cm where the mean is 169cm for the data point of 92.58 kgs which is only 0.5 standard deviation away from the mean of 82kg.