Assignment ON1

Elmer V Villanueva

Due at 3:55 PM on Monday 02 March 2020

Question 1

Consider the data shown in Table 1 from Anscombe [1]. Four pairs of data sets are given: Y_1 and X_1, Y_2 and X_2, Y_3 and X_3, and Y_4 and X_4.

Let's enter the data.

```
Y1 <- c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)
X1 <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
Y2 <- c(9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74)
X2 <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
Y3 <- c(7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73)
X3 <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
Y4 <- c(6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89)
X4 <- c(8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8)
```

EVV FEEDBACK: Accuracy and precision in data entry is very important here. A number of you have shown errors in data entry, causing your results to be wrong. Triple and quadruple check your data! There is no easy way around this.

1.1. (8 marks) For each pair, calculate the correlation coefficient.

```
cor(Y1, X1)
```

```
## [1] 0.8164205
```

```
cor(Y2, X2)
```

```
## [1] 0.8162365
```

```
cor(Y3, X3)
```

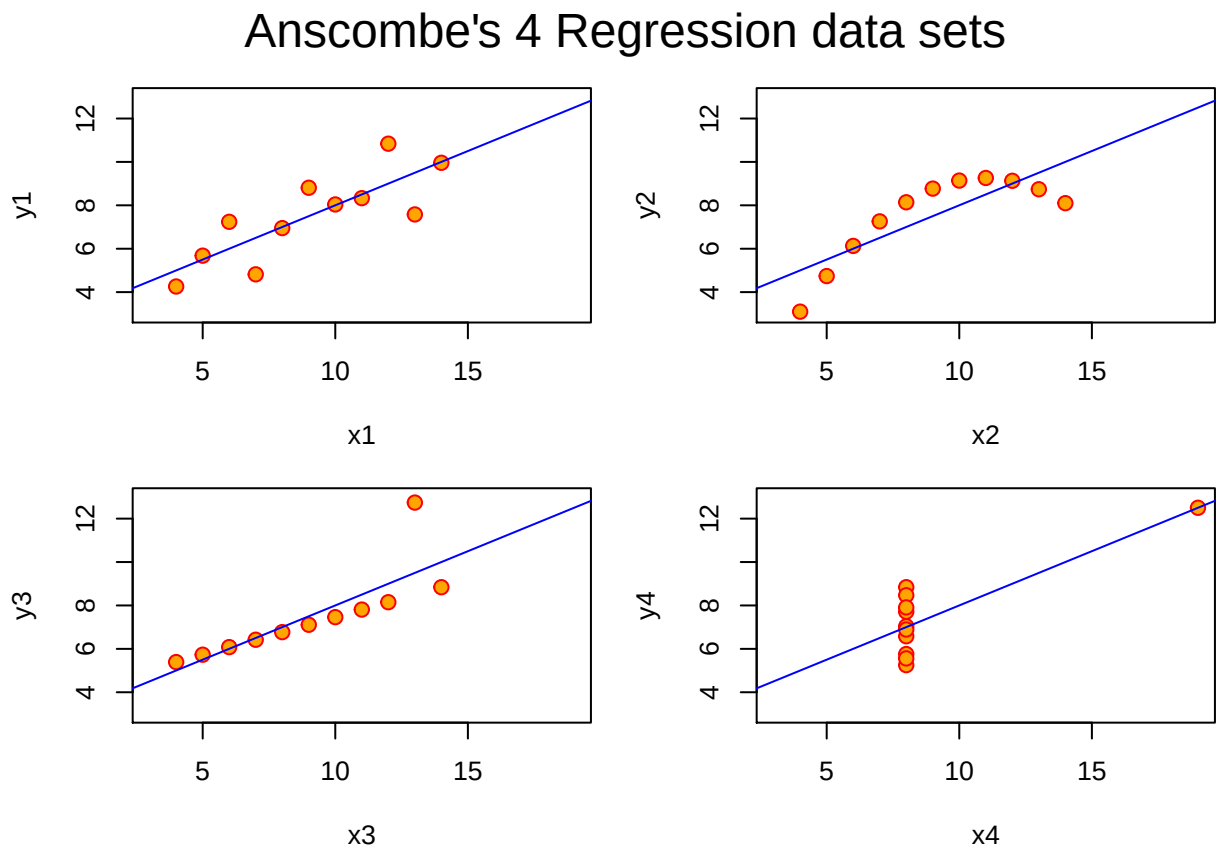
```
## [1] 0.8162867
```

```
cor(Y4, X4)
```

```
## [1] 0.8165214
```

1.2. (16 marks) For each pair, construct a separate scatterplot. The scatterplot should be properly formatted.

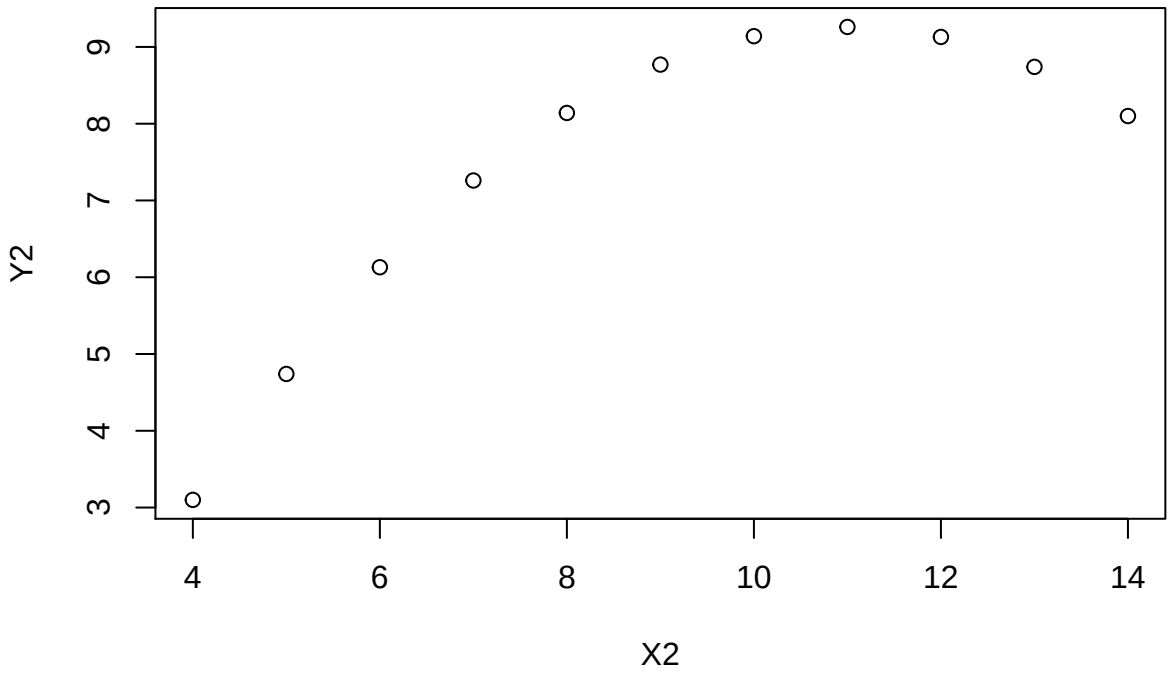
```r
ff <- y ~ x
mods <- setNames(as.list(1:4), paste0("lm", 1:4))
for(i in 1:4) {
  ff[2:3] <- lapply(paste0(c("y","x"), i), as.name)
  ## or   ff[[2]] <- as.name(paste0("y", i))
  ##      ff[[3]] <- as.name(paste0("x", i))
  mods[[i]] <- lmi <- lm(ff, data = anscombe)
}

op <- par(mfrow = c(2, 2), mar = 0.1+c(4,4,1,1), oma =  c(0, 0, 2, 0))
for(i in 1:4) {
  ff[2:3] <- lapply(paste0(c("y","x"), i), as.name)
  plot(ff, data = anscombe, col = "red", pch = 21, bg = "orange", cex = 1.2,
       xlim = c(3, 19), ylim = c(3, 13))
  abline(mods[[i]], col = "blue")
}
mtext("Anscombe's 4 Regression data sets", outer = TRUE, cex = 1.5)
```



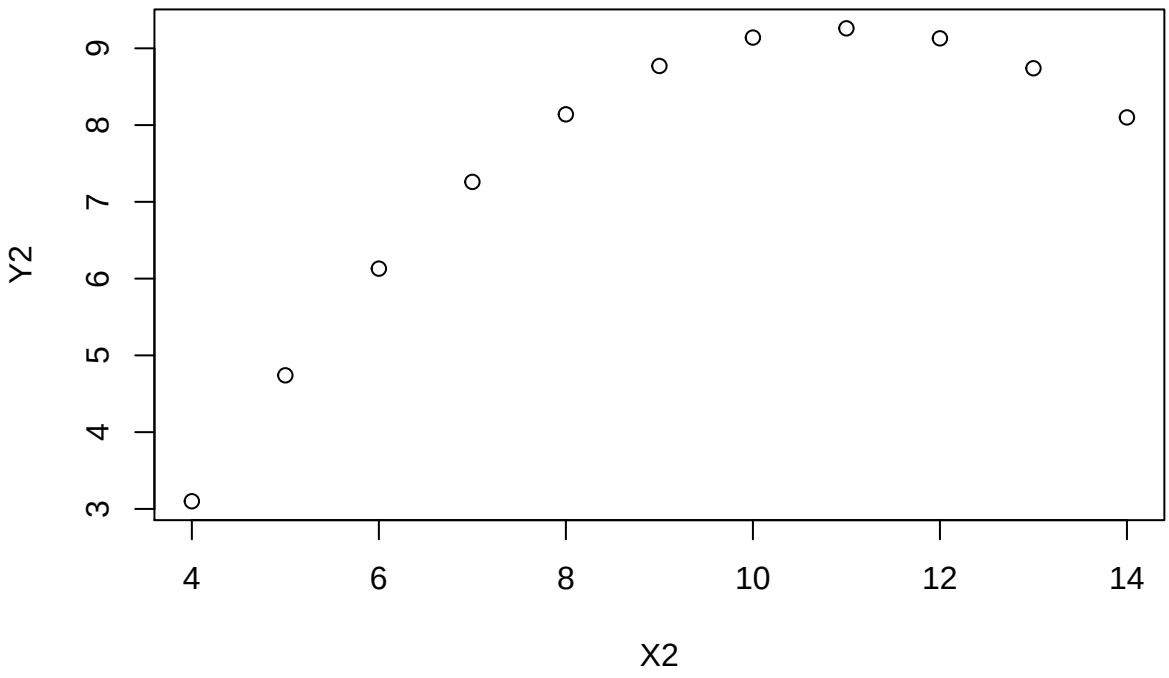

```r
par(op)
```

EVV FEEDBACK: Please be aware of the way the `plot()` function works. It is very quirky. The generic usage is `plot(x, y, ...)`. However, there is another format, as you know: `plot(y ~ x, ...)`. Note how the vertical and horizontal axes variables are switched in these two formats. If you make the mistake of failing to distinguish between the two, your graphs will be wrong. Let me give you an example:

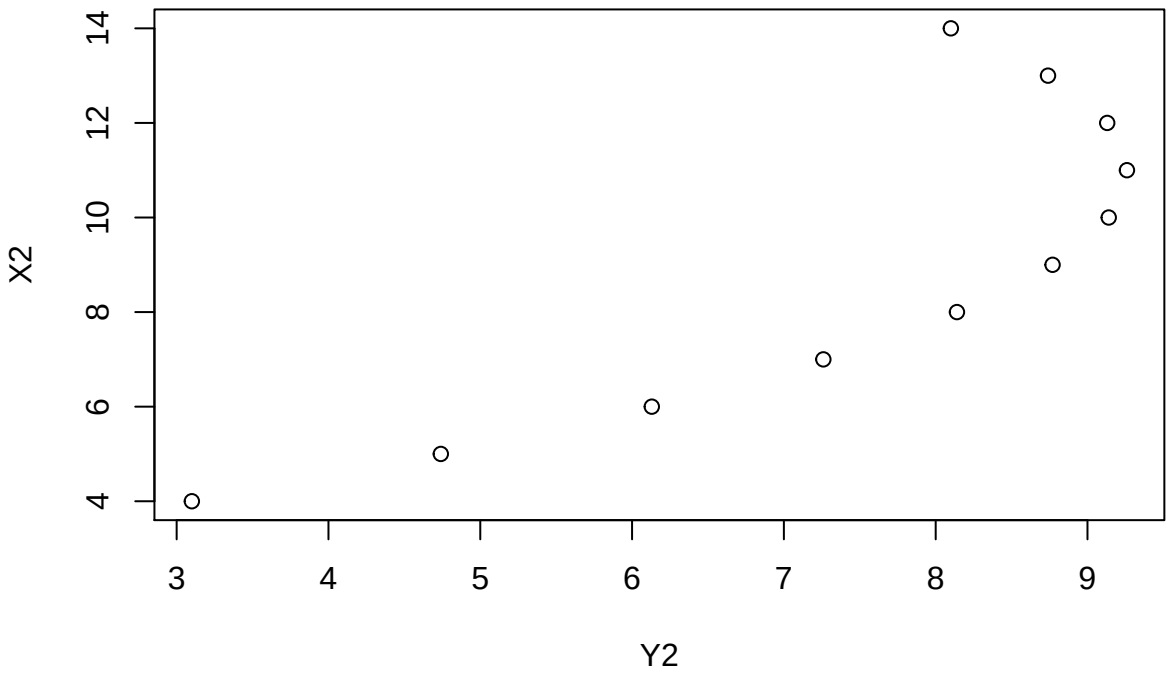
```r
plot(Y2 ~ X2)
```




```r
plot(X2, Y2)
```

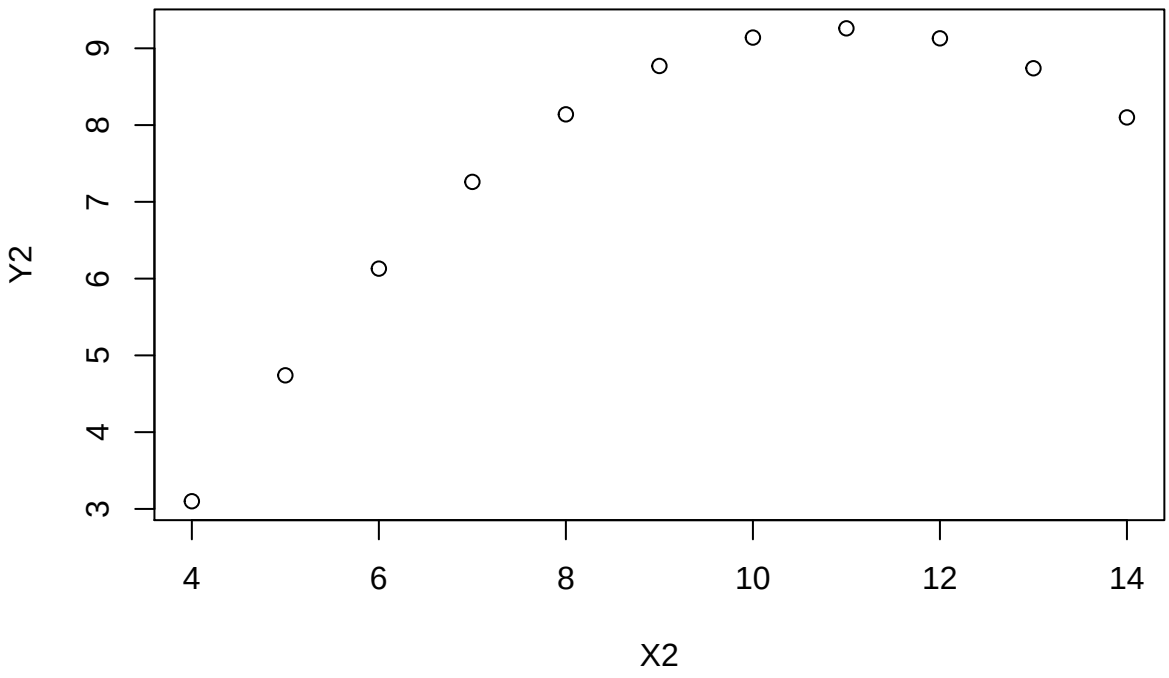



```
plot(Y2, X2)
```



The first two plots are equivalent. The third is erroneous. Some of you produced erroneous plots because you confused the two formats of the function. If you are confused, I would suggest that you modify the function a little bit:

```
plot(x=X2, y=Y2)
```



This will make you stop and think about your horizontal and vertical axes.

1.3. (4 marks) What do you notice about the results you presented in (1.1) and (1.2) above? Discuss your answer in less than 100 words.

> The correlation coefficients for all four datasetes are nearly equivalent at about 81%. However, the graphs show that these four datasets have quite different relationships.

Question 2

Do people tend to marry those of similar heights? A sample of 96 newly-married couples were selected. Shown in Table 2 are the heights of husbands (H) and wives (W), all measured to the nearest centimeter.

```
ID <- c(1:96)
H <- c(186,180,160,186,163,172,192,170,174,191,182,178,
       181,168,162,188,168,183,188,166,180,176,185,169,
       182,162,169,176,180,157,170,186,180,188,153,179,
       175,165,156,185,172,166,179,181,176,170,165,183,
       162,192,185,163,185,170,176,176,160,167,157,180,
       172,184,185,165,181,170,161,188,181,156,161,152,
       179,170,170,165,165,169,171,192,176,168,169,184,
       171,161,185,184,179,184,175,173,164,181,187,181)
W <- c(175,168,154,166,162,152,179,163,172,170,170,147,
       165,162,154,166,167,174,173,164,163,163,171,161,
       167,160,165,167,175,157,172,181,166,181,148,169,
```

```
       170,157,162,174,168,162,159,155,171,159,164,175,
       156,180,167,157,167,157,168,167,145,156,153,162,
       156,174,160,152,175,169,149,176,165,143,158,141,
       160,149,160,148,154,171,165,175,161,162,162,176,
       160,158,175,174,168,177,158,161,146,168,178,170)
MARRY <- data.frame(ID, H, W)
str(MARRY)
```

```
## 'data.frame':    96 obs. of  3 variables:
##  $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ H : num  186 180 160 186 163 172 192 170 174 191 ...
##  $ W : num  175 168 154 166 162 152 179 163 172 170 ...
```

```
head(MARRY)
```

```
##   ID   H   W
## 1  1 186 175
## 2  2 180 168
## 3  3 160 154
## 4  4 186 166
## 5  5 163 162
## 6  6 172 152
```

2.1. (4 marks) Produce a properly-formatted scatterplot of the data.

```
plot(W, H,
     main = "Heights of husbands and wives",
     ylab = "Husband's height (cm)",
     xlab = "Wife's height (cm)")
```

2.2. (2 marks) Calculate the covariance between the heights in centimeters of husbands and wives.

```
cov(H,W)
```

```
## [1] 69.41294
```

2.3. (4 marks) Transform the data from centimeters into inches. Then, calculate the covariance.

```
H.IN <- H/2.54
W.IN <- W/2.54
cov(H.IN, W.IN)
```

```
## [1] 10.75903
```

2.4. (4 marks) What do you notice about the results you reported in (2.2) and (2.3) above? Discuss your answer in less than 100 words.

The covariance changes when the units of measurement change.

2.5. (2 marks) Calculate the correlation between the heights in centimeters of husbands and wives.

```
cor(H, W)
```

```
## [1] 0.7633864
```

2.6. (2 marks) Calculate the correlation between the heights in inches of husbands and wives.

```
cor(H.IN, W.IN)
```

```
## [1] 0.7633864
```

2.7. (4 marks) What do you notice about the results you reported in (2.5) and (2.6) above? Discuss your answer in less than 100 words.

The correlation coefficient remains the same despite the change in the unit of measurement.

Assume that women in the population only marry husbands that are *exactly* five centimeters shorter than them.

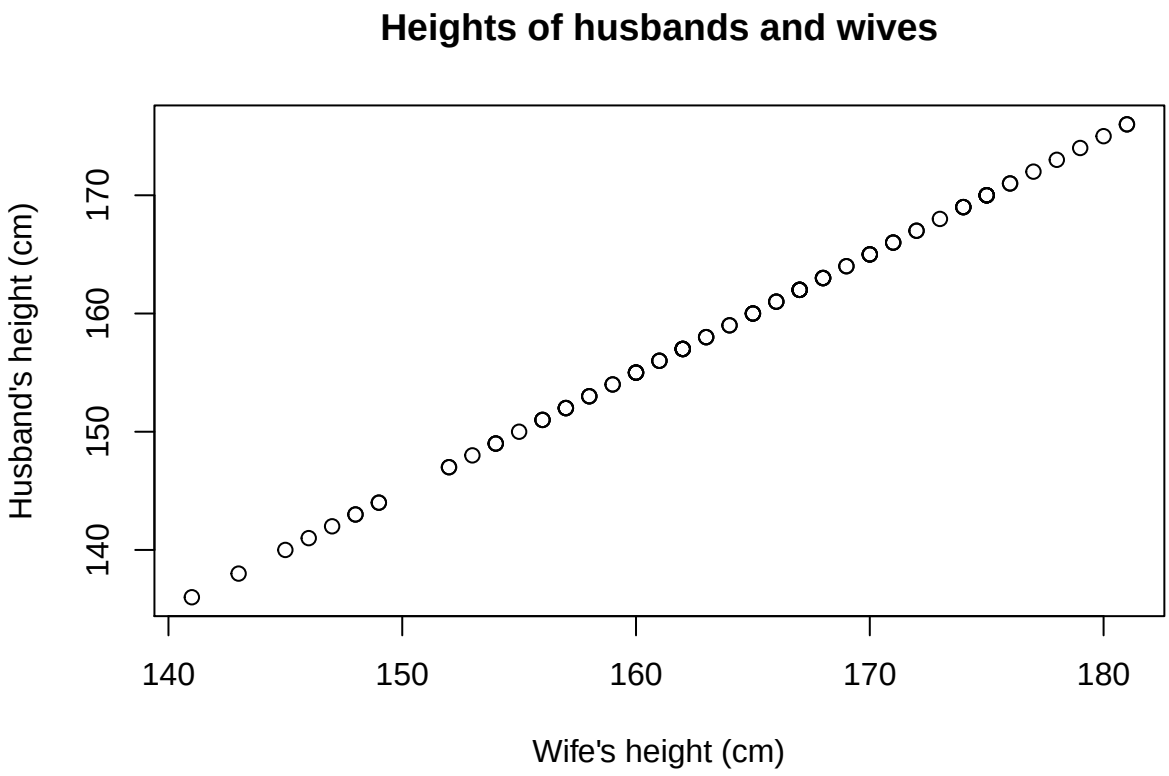
2.8. (4 marks) Produce the variable `H.NEW` that calculates new heights for husbands under this rule.

```
H.NEW <- W-5
summary(H.NEW)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   136.0   153.0   159.5   158.9   165.2   176.0
```

2.9. (4 marks) Produce a properly-formatted scatterplot of `W` and `H.NEW`.

```
plot(W, H.NEW,
     main = "Heights of husbands and wives",
     ylab = "Husband's height (cm)",
     xlab = "Wife's height (cm)")
```



2.10. (2 marks) Without any calculations, estimate the correlation between `W` and `H.NEW` using only the scatterplot.

The correlation coefficient will be 1.0