EDA A5 Q2

Data

Set up the following:

```
## Set this up for your own directory
#imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
#dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
#path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
source(file = "/Users/rudranibhadra/Downloads/graphicalTests.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/generateData.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/numericalTests.R", echo = FALSE)
```

and then read in the data as:

```
labData <- read.csv("/Users/rudranibhadra/Downloads/labData.csv")
```

Interest here lies **only** in that subset of the data where:

- `type` is either `"observational"` **or** `"randomized"`, **and**
- `rep` is 1

The result should be a `data.frame` having five variables and 216 rows (108 for each `type`). The value of `rep` should be 1 for all rows and the value of `type` should only be one of the two mentioned above.

a. *(2 marks)* Construct the `data.frame` described above. Assign it to the variable `results`. Show your code.

```
results<-labData[labData$rep==1 & (labData$type=='observational' | labData$type=='randomized'),]
```

Analysis

The dataset `results` from part (a) contains 216 (x, y) pairs, half of which are of `type "observational"` and half of which are of `type "randomized"`. To address the problem of whether changes in x cause changes in y you will pursue fitting a straight line model to the data of the form

$$y = \beta_0 + \beta_1 x + r$$

where r is the residual representing the fact that y and x need not lie exactly on a line. All fitting will be done by least-squares using the `lm()` function from R.

Of interest will be the hypothesis $H_0 : \beta = 0$. As a surrogate, we will use the model to test $H_0 : \beta_1 = 0$ and to provide estimates for β_1 in place of estimates of β.

Note that parts (b), (c), and (d) are identical, the difference lies in which data are used.

b. Following Cukier's and Mayer-Schoenberger's advice, here you will use **all** of the data in `results`.

c. *(2 marks)* Fit a straight line model of y to x and print the `summary()` of the fitted model. Show your code.

```
l1<-lm(y~x, data=results)
summary(l1)
```

```
##
## Call:
## lm(formula = y ~ x, data = results)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5398  -3.7322   0.7447   4.5601  15.5062
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.30128    0.88146  20.762   <2e-16 ***
## x            0.04770    0.04698   1.015    0.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.477 on 214 degrees of freedom
## Multiple R-squared:  0.004795,   Adjusted R-squared:  0.0001441
## F-statistic: 1.031 on 1 and 214 DF,  p-value: 0.3111
```

ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis H_0 : $\beta_1 = 0$? Justify your answer.

As the p-value is much greater than 0.05, there is less evidence against the null hypothesis that $\beta_1 = 0$. Hence, x and y are most likely independent.

iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer.

```
d<-results[,c('x','y')]
numericalTest(d,discrepancyFn=slopeDiscrepancy,generateFn =mixCoords )
```

```
## [1] 0.311
```

Since the p value is greater than 0.05, it indicates weak evidence against the null hypothesis and hence x and y are most likely independent.

iv. *(2 marks)* Repeat part (iii) above but now use the absolute value of the sample correlation as the discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer(s).

```
numericalTest(d,discrepancyFn =correlationDiscrepancy,generateFn =mixCoords)
```

```
## [1] 0.3005
```

Since the p value is greater than 0.05, it indicates weak evidence against the null hypothesis and hence x and y are most likely independent.

v. *(2 marks)* What do the above empirical tests, based on **all** the data here, suggest one might conclude about the causal relation between x and y? If you have found a causal effect, in what direction would y be expected to change if a value of x were increased? Justify your answers.

There does not seem to be a causal relationship between x and y since they appear to be independent with respect to each other because the p values from the above tests are all greater than 0.05.

c. Now restrict the analysis to only that subset of `results` which has `type "observational"`. This analysis will be based on only 108 (x, y) pairs.

d. *(2 marks)* Fit a straight line model of y to x and print the `summary()` of the fitted model. Show your code.

```
r1<-results[results$type=='observational',]
l2<-lm(y~x, data=r1)
summary(l2)
```

```
##
## Call:
## lm(formula = y ~ x, data = r1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4802 -0.6540  0.1250  0.7446  3.4619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.82778    0.27584   97.26   <2e-16 ***
## x           -0.41159    0.01417  -29.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.257 on 106 degrees of freedom
## Multiple R-squared:  0.8884, Adjusted R-squared:  0.8874
## F-statistic: 844.2 on 1 and 106 DF,  p-value: < 2.2e-16
```

ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis H_0 : $\beta_1 = 0$? Justify your answer.

As the p-value is much lesser than 0.05, there is strong evidence against the null hypothesis that $\beta_1 = 0$. Hence, x and y are not likely to be independent.

iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer.

```
d1<-r1[,c('x','y')]
numericalTest(d1,discrepancyFn=slopeDiscrepancy,generateFn =mixCoords)
```

```
## [1] 0
```

Since the p value is 0 (less than 0.05), it indicates strong evidence against the null hypothesis and hence x and y are not likely to be independent.

iv. *(2 marks)* Repeat part (iii) above but now ise the absolute value of the sample correlation as the discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer.

```
numericalTest(d1,discrepancyFn =correlationDiscrepancy,generateFn =mixCoords)
```

```
## [1] 0
```

Since the p value is 0 (less than 0.05), it indicates strong evidence against the null hypothesis and hence x and y are not likely to be independent.

v. *(2 marks)* What do the above empirical tests, based on **only** the `"observational"` data here, suggest one might conclude about the causal relation between x and y? If you have found a causal effect, in what direction would y be expected to change if a value of x were increased? Justify your answers.

Based on the tests above, there seems to be a causal relationship between x and y because they appear to be dependent on each other since there is strong evidence against $H_0 : \beta_1 = 0$. Since here the slope estimate is less than 0 (-0.41159), an increase in x causes y to decrease.

d. Now restrict the analysis to only that subset of `results` which has `type "randomized"`. This analysis will be based on only 108 (x, y) pairs.

e. *(2 marks)* Fit a straight line model of y to x and print the `summary()` of the fitted model. Show your code.

```
r2<-results[results$type=='randomized',]
l3<-lm(y~x, data=r2)
summary(l3)
```

```
##
## Call:
## lm(formula = y ~ x, data = r2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8241  -6.2315   0.1759   6.4259  12.7685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.97222    1.30229   9.961  < 2e-16 ***
## x            0.37037    0.07224   5.127 1.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.507 on 106 degrees of freedom
## Multiple R-squared:  0.1987, Adjusted R-squared:  0.1912
## F-statistic: 26.29 on 1 and 106 DF,  p-value: 1.334e-06
```

ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis $H_0 : \beta_1 = 0$? Justify your answer.

As the p-value is much lesser than 0.05, there is strong evidence against the null hypothesis that $\beta_1 = 0$. Hence, x and y are not likely to be independent.

iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer.

```
d2<-r2[,c('x','y')]
numericalTest(d2,discrepancyFn=slopeDiscrepancy,generateFn =mixCoords )
```

```
## [1] 0
```

Since the p value is 0 (less than 0.05), it indicates strong evidence against the null hypothesis and hence x and y are not likely to be independent.

iv. *(2 marks)* Repeat part (iii) above but now ise the absolute value of the sample correlation as the discrepancy measure. Show your code.

What do you conclude about the hypothesis? Justify your answer.

```
numericalTest(d2,discrepancyFn =correlationDiscrepancy,generateFn =mixCoords)
```

```
## [1] 0
```

Since the p value is 0 (less than 0.05), it indicates strong evidence against the null hypothesis and hence x and y are not likely to be independent.

v. *(2 marks)* What do the above empirical tests, based on **only** the `"randomized"` data here, suggest one might conclude about the causal relation between x and y? If you have found a causal effect, in what direction would y be expected to change if a value of x were increased? Justify your answers.

Based on the tests above, there seems to be a causal relationship between x and y because they appear to be dependent on each other since there is strong evidence against $H_0 : \beta_1 = 0$. Since here the slope estimate is more than 0 (0.37), an increase in x causes y to increase.

Conclusions

You have now conducted three separate analyses of the data. Based on the results in these analyses, address the following.

e. *(4 marks)* What is meant by a "lurking" variable and how might it have shown up in this data?

A lurking variable is one which is not included as an explanatory or response variable in analysis but it can affect the interpretation of relationship between variables. Here the lurking variable might have affected the relationship between x and y.

f. *(3 marks)* What conclusion do you draw about the causal relation, if any, between changes in x and changes in y? Justify your answer.

There is no causal relationship observed between x and y when taking the dataset when taking both observational and randomized based on the analysis above. If the type is observational or randomized, only then there is there is strong evidence against the null hypothesis that x and y are independent, suggesting that x and y are related to each other. When the type is observational, an increase in x causes y to decrease as slope estimate < 0. When the type is randomized, an increase in x causes y to increase as slope estimate > 0.

g. *(2 marks)* In 2008, then editor of *Wired* magazine, Chris Anderson wrote in an article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete":

There is now a better way. Petabytes allow us to say: "Correlation is enough."

Drawing on your analyses above, give an illustration as to why Chris Anderson might be mistaken, in spite of the volume of data observed.

From the first analysis, it seemed like x and y are most likely independent until we took the subsets of observational and randomized seperately and analysed x and y. Based on the values of the slope estimate, x and y have a positive or negative relationship. This shows that even a large amount of data is not enough to show correlation.

h. *(2 marks)* Why might Cukier's and Mayer-Schoenberger's advice (quoted above) actually be dangerous?

Correlation does not imply causation. No conclusions should be drawn simply on the basis of correlation between two variables X and Y. Instead, we must understand the underlying mechanisms that connect the two variables.