A5 - Ram Yoogesh Gopu (20867060)

```
knitr::opts_chunk$set(echo = TRUE,
                      warning = FALSE,
                      message = FALSE,
                      fig.align = "center",
                      fig.width = 6,
                      fig.height = 7,
                      out.height = "40%")
set.seed(12314159)
library(loon.data)
library(loon)
```

```
## Loading required package: tcltk
```

```
library(gridExtra)

imageDirectory <- "./img"
dataDirectory <- "./"
path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)

source("graphicalTests.R")
source("numericalTests.R")
source("generateData.R")
```

(1)

Reading the data

```
labData <- read.csv("labData.csv")
```

(A)

```
results <- data.frame((labData[(labData$type == "randomized" | labData$type == "observational") & labDa
```

(2)

(B)

(i)

```
Bmod <- lm(y ~ x, results)
summary(Bmod)
```

```
##
## 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)

From the p-value (which is 0.3111) greater than 0.05, we can conclude that it is not strong to oppose the hypothesis.

(iii)

```
numericalTest(results, discrepancyFn = slopeDiscrepancy, generateFn = mixCoords)
```

```
## [1] 0.32
```

The p-value (0.32, which is greater than 0.05) indicates that its not strong to oppose the hypothesis that X,Y are independent.

(iv)

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

```
## [1] 0.3145
```

Similarly, change in X doesnt result in change in Y, because the p-value (0.3145, which is greater than 0.05) indicates that its not strong to oppose the hypothesis.

(V)

From the above tests, we can conclude that there is no casual effect. We can conclude that change in X doesnt result in change in Y. This is because of the fact that the P-value is larger in all cases , hence we coudnt oppose the hypothesis.

(C)

(i)

```
Cmod <- lm (y ~ x, data = results[results$type == "observational", ])
summary(Cmod)
```

```
##
## Call:
## lm(formula = y ~ x, data = results[results$type == "observational",
##     ])
##
## 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)

since the p-value is very small (less than 0.05), it is a strong proof against the hypothesis. Also another way to look at is that the beta_1 is negative, which results in decrease of value y with increase in X.

(iii)

```
resval <- results[results$type=="observational", ]
samobs <- data.frame(x = resval$x, y = resval$y)
numericalTest(samobs, discrepancyFn = slopeDiscrepancy, generateFn = mixCoords)
```

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

It is a strong evidence against the null hypothesis, since p-value is 0. It also gives a strong evidence that X and Y are not independent of each other.

(iv)

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

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

It is a strong evidence against the null hypothesis, since p-value is 0. It also gives a strong evidence that X and Y are not independent of each other and their correlation cofficient is non-zero between X and Y.

(V)

From all the tests we can conclude that there is a casual relation between X and Y. Also a strong evidence that X and Y are not independent of each other from p-value which is 0 as inferred from the above tests. Also from the summary beta_1 is negative, which is a strong evidence that states that increase in value of X results in decrease of value Y.

(D)

(i)

```
Dmod <- lm(y~x, data = results[results$type == "randomized", ])
summary(Dmod)
```

```
##
## Call:
## lm(formula = y ~ x, data = results[results$type == "randomized",
##     ])
##
## 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)

Since the p-value is ver less (than 0.05), there is a strong evidence against the hypothesis beta_1 = 0.

(iii)

```
resval_2 <- results[results$type == "randomized", ]
samdobs <- data.frame(x = resval_2$x, y = resval_2$y)
numericalTest(samdobs, discrepancyFn = slopeDiscrepancy, generateFn = mixCoords)
```

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

Since the p-value is 0, this is a strong evidence against the NUll hypothesis. This also means that X and Y won't be independent.

(iv)

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

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

Since the p-value is again 0, this is a strong evidence against the null hypothesis. Also the correlation cofficient will be non-zero, which infers that X and Y will be correlated.

(V)

From the above tests we can conclude that there is a casual relation between X and Y. Also from the summary beta_1 is postive, which is also a evidence suggesting the increase in value of X results in increase in value of Y.

(3)

(E)

A lurking variable is a variable on which we don't have contol over, which can affect the observed relationships between measured variables and may also cause bias in our results. In the above experiment, Z is a lurking variable.The equation of the plane has 3 variables, and its difficult to specifiy the relationship between the two variables if the other one is kept constant.

(F)

We cannot come up with a conclusion about the relationship between X and Y when all the data is combines. But the relationship from "observational" and "randomized" data are producing different effects. So, for the observational data, the casual effect is negative, which states that the increase in value of X decreases the value of Y. For the randomized data, the casual effect is positive, which states that the increase in value of X increases the value of Y.

(G)

From the experiments we can conclude that Correlation isn't enough. This is because, when we used the whole data, it explained that there was no relationship between X and Y. On the contrary, when we performed the experiments on the seperated datasets, (observational and randomized) gave us the conclusion that there was relationship between X and Y (Positive and Negative). Hence we can conclude that if we just use the result of an experiment combining large volumes of data, we may lose some important relationships in the data.

(H)

We will end up in wrong conclusions if we use the whole data and we wont be able to indetify the relationships in the subgroups of data which could be dangerous. So we can conclude that , one must focus on meaningful insights rather than more data.