R Notebook

```
load('2022essay_q2.Rda')
```

a) Assuming no attrition, is it likely that randomisation failed in this experiment? Provide evidence from the dataset for your answer.

```
#Using regression
summary(lm(d ~ x, data = e))
```

```
##
## Call:
## lm(formula = d ~ x, data = e)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6250 -0.6250  0.1875  0.3750  0.3750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6250     0.0489  12.780  < 2e-16 ***
## x            -0.6250     0.1094  -5.715 1.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4374 on 98 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:  0.2423
## F-statistic: 32.67 on 1 and 98 DF,  p-value: 1.18e-07
```

```
#t-test
t.test(e$x[e$d==1], e$x[e$d==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  e$x[e$d == 1] and e$x[e$d == 0]
## t = -5.7155, df = 49, p-value = 6.402e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.540641 -0.259359
## sample estimates:
## mean of x mean of y
##       0.0       0.4
```

b) Again assuming no attrition, calculate and report:

i. The true average treatment effect in terms of potential outcomes

ii. The observed average treatment effect from the experiment Then use your answers to explain whether selection bias is negative, zero, or positive in this experiment.

```
# i. true ATE in terms of potential outcome
mean(e$y1-e$y0)
```

```
## [1] -7.495
```

```
# ii. The observed average treatment effect from the experiment
mean(e$outcome[e$d==1]) - mean(e$outcome[e$d==0])
```

```
## [1] -9.4206
```

c) Explain the direction of the selection bias in (b): why is it negative, zero, or positive? Provide evidence from the dataset for your answer.

```
cor(e$outcome,e$x)
```

```
## [1] 0.4767462
```

```
cor(e$d,e$x)
```

```
## [1] -0.5
```

d) Still assuming no attrition, use an appropriate technique to adjust your calculation from (b) (ii) to come closer to recovering the true ATE from the experiment. How close is your new estimate to the true ATE?

```
summary(lm(outcome ~ d + factor(x), data = e))
```

```
##
## Call:
## lm(formula = outcome ~ d + factor(x), data = e)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1542  -3.1667  -0.1075   3.6908  15.3370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   27.413      1.150  23.846  < 2e-16 ***
## d             -7.499      1.454  -5.157 1.33e-06 ***
## factor(x)1     4.804      1.818   2.643  0.00958 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.297 on 97 degrees of freedom
## Multiple R-squared:  0.3935, Adjusted R-squared:  0.381
## F-statistic: 31.47 on 2 and 97 DF,  p-value: 2.921e-11
```

e) Now we'll examine attrition, assuming that it occurs as described by the variable R. Are units in this experiment missing at random? Provide evidence from the dataset for your answer.

```
#For control
mean(e$y0[e$R==1])
```

```
## [1] 26.37182
```

```
mean(e$y0[e$R==0])
```

```
## [1] 32.25917
```

```
#For treatment
mean(e$y1[e$R==1])
```

```
## [1] 19.85091
```

```r
mean(e$y1[e$R==0])
```

```
## [1] 17.62083
```

```r
t.test(e$y1[e$R==1], e$y1[e$R==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  e$y1[e$R == 1] and e$y1[e$R == 0]
## t = 2.9634, df = 97.145, p-value = 0.003827
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7365102 3.7236413
## sample estimates:
## mean of x mean of y
##  19.85091  17.62083
```

```r
t.test(e$y0[e$R==1], e$y0[e$R==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  e$y0[e$R == 1] and e$y0[e$R == 0]
## t = -8.0705, df = 97.542, p-value = 1.842e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.335087 -4.439610
## sample estimates:
## mean of x mean of y
##  26.37182  32.25917
```

f) Evaluate the following two statements about this dataset:

i. "You cannot estimate an ATE for always-reporters with this data"
ii. "Within values of x, units are missing-at-random". Are these statements correct? Provide evidence from the dataset for your answers.

```r
mean(e$x[e$R==1])
```

```
## [1] 0.09090909
```

```r
mean(e$x[e$R==0])
```

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

```r
t.test(e$x[e$R==1], e$x[e$R==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  e$x[e$R == 1] and e$x[e$R == 0]
## t = -29.496, df = 87, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9703511 -0.8478307
## sample estimates:
##  mean of x  mean of y
## 0.09090909 1.00000000
```