Project Causal Inference

Group 14 (Ariana, Jiraiyu, Ravi, Yashish)

4/17/2022

1 Executive Summary

AIDS infection in Africa is on the rise. One of the strategies to prevent this is to make sure the people getting tested are getting their results. In this document we evaluate an experiment to explore if giving incentives to people gets them to find out the results. We have done so using Propensity Score Matching technique and OLS regression. The results indicate that giving incentive indeed pushes people to find their results. Also, the distance to the VCT center is another factor which impacts decision to find the test result.

2 Background

Over the past two decades, the HIV/AIDS epidemic has afflicted millions of individuals in Africa. In the absence of significantly expanded prevention and treatment programs, the epidemic is expected to worsen in many other parts of the world. One intervention often suggested to alleviate the spread of the disease is HIV testing, and some have gone so far as to say that voluntary counseling and testing (VCT) is the "missing weapon in the battle against AIDS". Under the assumption that HIV testing is an effective prevention strategy, many international organizations and governments have called for increased investments in counseling and testing, requiring large amounts of monetary and human resources. For example, in South Africa, government expenditures on counseling and testing increased from $2.4 million in 2000 to $17.3 million in 2004, and in Mozambique, 55 percent of all HIV/AIDS program expenditures in 2000 were for HIV counseling and testing.

Underlying the emphasis on HIV testing for prevention—and the large-scale expenditures on testing—are two rarely challenged assumptions. First, many believe that knowledge of HIV status has positive effects on sexual behavior that prevent the spread of the disease. In particular, it is assumed that those diagnosed HIV-negative will protect themselves from infection and those diagnosed HIV-positive will take precautions to protect others. Second, many believe that it is difficult to get people to learn their HIV status, due primarily to psychological or social barriers, thus justifying expenditures on destigmatization and advertising campaigns. In this project, we evaluate a field experiment in rural Malawi designed to address these assumptions where in respondents were offered free door-to-door HIV test and were randomly assigned vouchers ($0-$3) redeemable upon obtaining the results at a nearby VCT center.

3 Data Overview

1. Overall data had 4820 observations and 9 features which are as follows:
    a. Gender (male)
    b. Unique Village Id (villnum)
    c. Got/Not got results (got)
    d. Received/Not Received incentives (any)

e. Total Incentives (tinc)
f. Distance to VCT center in Kilometers (distvct)
g. Married or Not (mar)
h. Age (age)
i. HIV Test Result (hiv2004)

2. Summary statistics (appendix 9.2):
Number of respondents: 4820
Number of Males: 2260
Age range (Males): 13 to 84
Number of Females: 2560
Age Range (Females): 11 to 81
Survey Response Rates: 67%
HIV Test Acceptance: 90.49%
Learned Result: 69.66%
HIV Prevalence: 6%
Incentives Received: 2016
No Incentives: 878 Average Age: 33

4 The Experiment

Free door-to-door tests were conducted to collect HIV samples. Men and women were randomly selected from approximately 140 villages located in the Northern, Central, and Southern regions of Malawi from, approximately, 25 percent of all households in each village. Between May and August of 2004, nurses from outside each area offered respondents free tests in their homes for HIV (samples were taken through oral swabs). Vouchers ranging from 0 to $3 were drawn randomly from the bag with an average amount of $1.01. For each subject, VCT center, to receive the results, were randomly selected.

5 Analysis

5.1 Randomization Test Randomization in an experiment refers to a random assignment of participants to the treatment in an experiment. It makes sure that the groups made for conducting an experiment are as similar as possible to each other so that the results come out as accurate as possible. Randomization prevents biases and eliminates any possibility of an unobserved confound and possibility of simultaneity. To check if the subjects selected for the experiments (appendix 9.3) were completely random, we ran 't.test' for each of the features between control and treatment. From the results of the experiment we could say that, except with respect to age (p-value = 0.005421), in all the features the subjects chosen were completely random. But since the age was not randomized in the data, we decided to match subjects in control and treatment using propensity score matching.

5.2 Propensity Score Matching (PSM) Propensity score matching is a statistical matching technique that attempts to estimate the effect of a treatment by accounting for the co-variates that predict receiving the treatment. PSM attempts to reduce the bias due to confounding variables that could be found in an estimate of the treatment effect obtained from simply comparing outcomes among units that received the treatment versus those that did not. In PSM, we find matched pairs that are as similar on observed variables, so that the treatment can be assumed to be random. If no match is found, throw away the data. In PSM we model treatment as a function of observed variables using logit or probit. We then get predicted probabilities of

treatment from the model and for each treated observation, find control observation with closest propensity score.

To simulate randomization and balance the number of subjects in treatment and control, PSM matches each individual in control and treatment based on given features (age, distance to VCT center, gender and if married or not in this case) using a propensity score (appendix 9.4). After propensity score matching we re-tested the randomization for each figure and there are no columns which fail randomization.

5.3 Power of the Experiment We needed to check whether our experiment was under-powered i.e. having a sample that is too small. If the sample lacks power, it means it cannot reliably detect effects and the chance of discovering effects that aren't there goes up. This could be because random noise in the data leads to more chance differences in small samples. To report a difference of 0.45 between control and treatment. with 95% level of significance and 80% power, we need at least 79 respondents in both treatment and control (appendix 9.5). In the matched data, we have 400+ rows of data both the groups. Thus the experiment is sufficiently powerful.

5.4 OLS Regressions Initially regressed age, marital status, gender, distance to VCT center, and incentives over if the subject proceeded to get the result. Then, regressed incentives over if the subject proceeded to get the result. Next, regressed distance to VCT center over if the subject proceeded to get the result. Lastly, regressed the incentive the subject got over if the subject proceeded to get the result.

We found that:

1. Any amount of incentive increases the odds of subjects finding out the results by almost 7 times.

2. Increasing incentive amount by $1 increases the odds of subject finding out the results by about 1.01 times.

3. Decreasing the distance to VCT center by 1 km, increases the odds of subject finding out the results by about 1.2 times.

Please refer to appendix 9.7 for all the results.

6 Assumptions

We have had below assumptions which have been tested in the Analysis section:

1. Randomization of the features in the experiment subjects
2. The experiments has enough power to provide the significant results.

7 Threats and Limitations to the Experiment

1. Omitted variable bias : There could be other factors, such as behavior of VCT center staff due to which a subject might be wary of getting the results
2. Selection Bias : The design of this experiment avoids selection bias because it randomized: Individual incentives to learn HIV status Location of VCT centers where HIV results were available
3. Measurement error: There is no measurement error as experiment also measured actual post-test attendance at centers to obtain results
4. Simultaneity: There can be no source of simultaneity

8 Conclusions and Recommendations

From the above experiments we have established causality between getting the getting incentives and if the subject proceeds to get the results. Since incentives increase the odds of a subject finding out their test results, long term programs should be developed to provide incentives to all patients who are getting test for HIV. And, since decreasing the distance to VCT increases the odds of subject finding out the results, infrastructure should be set up to increase the number of centers. This would make the centers more accessible for the subject to get the results. These two combined with each other could be an effective HIV prevention strategy in rural Malawi.

9 Appendix

9.1 Load data Column 'any' tells us if the subject is treatment or control.

```
### Count of all respondents
summary_data = all_data %>%
    mutate(villnum = as.factor(villnum),
           got = as.factor(got),
           any = as.factor(any),
           male = as.factor(male),
           mar = as.factor(mar))
summary(summary_data)
```

9.2 Summary statistics of the data

```
##     villnum       got          distvct         hiv2004            tinc
##  11     : 223   0   : 878   Min.   :0.000   Min.   :-1.0000   Min.   :  0.0
##  10     : 166   1   :2016   1st Qu.:1.030   1st Qu.: 0.0000   1st Qu.: 10.0
##  14     : 141   NA's:1926   Median :1.675   Median : 0.0000   Median :100.0
##  6      : 138               Mean   :2.003   Mean   : 0.0591   Mean   :104.5
##  140    : 122               3rd Qu.:2.743   3rd Qu.: 0.0000   3rd Qu.:200.0
##  (Other):4003               Max.   :5.192   Max.   : 1.0000   Max.   :300.0
##  NA's   :  27                               NA's   :1926      NA's   :1919
##    any            age        male       mar
##  0   : 679   Min.   :11.00   0:2560   0   : 963
##  1   :2222   1st Qu.:23.00   1:2260   1   :2544
##  NA's:1919   Median :32.00            NA's:1313
##              Mean   :33.65
##              3rd Qu.:43.00
##              Max.   :84.00
##              NA's   :441
```

```
treatment_data = final_data %>%
                 filter(any == 1)

control_data = final_data %>%
                 filter(any == 0)
```

```
print(t.test(treatment_data$age, control_data$age))
```

9.3 Randomization check

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$age and control_data$age
## t = 2.7866, df = 1059.1, p-value = 0.005421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4897056 2.8207641
## sample estimates:
## mean of x mean of y
##  33.76832  32.11309
```

```
print(t.test(treatment_data$distvct, control_data$distvct))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$distvct and control_data$distvct
## t = 1.2174, df = 1027.8, p-value = 0.2237
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04199419  0.17927316
## sample estimates:
## mean of x mean of y
##  2.029567  1.960928
```

```
print(t.test(treatment_data$male, control_data$male))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$male and control_data$male
## t = -0.4375, df = 991.56, p-value = 0.6618
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05452527  0.03464503
## sample estimates:
## mean of x mean of y
## 0.4601730 0.4701131
```

```
print(t.test(treatment_data$mar, control_data$mar))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$mar and control_data$mar
```

```
## t = 0.073219, df = 991.43, p-value = 0.9416
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03898788  0.04201007
## sample estimates:
## mean of x mean of y
## 0.7123350 0.7108239
```

```
match_output <- matchit(any ~ age + distvct + male + mar,
                        data = final_data,
                        method = 'nearest',
                        distance = "logit",
                        caliper = 0.001,
                        replace = FALSE,
                        ratio = 2)
```

9.4 Propensity score matching

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

```
print(summary(match_output))
```

```
##
## Call:
## matchit(formula = any ~ age + distvct + male + mar, data = final_data,
##     method = "nearest", distance = "logit", replace = FALSE,
##     caliper = 0.001, ratio = 2)
##
## Summary of Balance for All Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.7810        0.7773          0.1431     1.1778    0.0352
## age            33.7683       32.1131          0.1195     1.1673    0.0251
## distvct         2.0296        1.9609          0.0536     1.0880    0.0147
## male            0.4602        0.4701         -0.0199          .    0.0099
## mar             0.7123        0.7108          0.0033          .    0.0015
##          eCDF Max
## distance   0.0859
## age        0.0665
## distvct    0.0440
## male       0.0099
## mar        0.0015
##
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.7754        0.7754          0.0002     0.9999    0.0005
## age            30.9752       30.4367          0.0389     1.0390    0.0102
## distvct         1.9394        1.9578         -0.0144     1.0104    0.0149
## male            0.4566        0.4516          0.0100          .    0.0050
```

```
## mar             0.7097        0.6787          0.0685          .    0.0310
##          eCDF Max Std. Pair Dist.
## distance   0.0050          0.0005
## age        0.0347          0.6193
## distvct    0.0397          0.9393
## male       0.0050          0.9821
## mar        0.0310          0.8857
##
## Percent Balance Improvement:
##          Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance            99.8      100.0      98.7     94.2
## age                 67.5       75.3      59.4     47.8
## distvct             73.2       87.8      -1.6      9.8
## male                50.1          .      50.1     50.1
## mar              -1952.6          .   -1952.6  -1952.6
##
## Sample Sizes:
##               Control Treated
## All            619.      2197
## Matched (ESS)  416.43     403
## Matched        429.       403
## Unmatched      190.      1794
## Discarded        0.         0
```

```
data_match = match.data(match_output)
```

```
treatment_data = data_match %>%
                 filter(any == 1)

control_data = data_match %>%
                 filter(any == 0)

print(t.test(treatment_data$age, control_data$age))
```

9.5 Randomization re-check for matched data

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$age and control_data$age
## t = 0.58615, df = 823.67, p-value = 0.5579
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.107869  2.051248
## sample estimates:
## mean of x mean of y
##  30.97519  30.50350
```

```
print(t.test(treatment_data$distvct, control_data$distvct))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$distvct and control_data$distvct
## t = -0.30948, df = 825.51, p-value = 0.757
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1889351  0.1374710
## sample estimates:
## mean of x mean of y
##  1.939362  1.965094
```

```
print(t.test(treatment_data$male, control_data$male))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$male and control_data$male
## t = 0.26104, df = 826.57, p-value = 0.7941
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05882640  0.07687287
## sample estimates:
## mean of x mean of y
## 0.4565757 0.4475524
```

```
print(t.test(treatment_data$mar, control_data$mar))
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_data$mar and control_data$mar
## t = 0.83647, df = 828.81, p-value = 0.4031
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03594505  0.08933253
## sample estimates:
## mean of x mean of y
## 0.7096774 0.6829837
```

```
powerTestResults = power.t.test(type = 'two.sample',
                                power = .8,
                                sig = 0.05,
                                delta = .45,
                                sd = 1)
powerTestResults
```

9.6 Power Test

```
##
##      Two-sample t test power calculation
##
##               n = 78.49181
##           delta = 0.45
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
model = glm(got ~ distvct + any + age + male + mar,
           data = data_match,
           family = "binomial")
summary(model)
```

9.7 OLS Regressions

```
##
## Call:
## glm(formula = got ~ distvct + any + age + male + mar, family = "binomial",
##     data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9840  -0.9259   0.6149   0.7764   1.7833
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.525398   0.273300  -1.922  0.05455 .
## distvct     -0.212151   0.066430  -3.194  0.00141 **
## any          1.939661   0.159145  12.188  < 2e-16 ***
## age          0.008109   0.009018   0.899  0.36850
## male        -0.008069   0.161971  -0.050  0.96027
## mar          0.004866   0.224268   0.022  0.98269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1145.69  on 831  degrees of freedom
## Residual deviance:  965.62  on 826  degrees of freedom
## AIC: 977.62
##
## Number of Fisher Scoring iterations: 4
```

```
model = glm(got ~ any,
           data = data_match,
           family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = got ~ any, family = "binomial", data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7251  -0.9044   0.7155   0.7155   1.4776
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6827     0.1022  -6.677 2.44e-11 ***
## any           1.9148     0.1570  12.197  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1145.69  on 831  degrees of freedom
## Residual deviance:  978.04  on 830  degrees of freedom
## AIC: 982.04
##
## Number of Fisher Scoring iterations: 4
```

```
model = glm(got ~ any + distvct,
           data = data_match,
           family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = got ~ any + distvct, family = "binomial", data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8821  -0.9352   0.6268   0.7753   1.7970
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.26881    0.15997  -1.680  0.09288 .
## any          1.93909    0.15889  12.204  < 2e-16 ***
## distvct     -0.21649    0.06598  -3.281  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1145.69  on 831  degrees of freedom
## Residual deviance:  967.07  on 829  degrees of freedom
## AIC: 973.07
##
## Number of Fisher Scoring iterations: 4
```

```
model = glm(got ~ tinc + distvct,
            data = data_match,
            family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = got ~ tinc + distvct, family = "binomial", data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5307  -1.0293   0.4237   1.0921   1.6008
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.082298   0.153184  -0.537  0.59109
## tinc         0.011126   0.001126   9.884  < 2e-16 ***
## distvct     -0.168299   0.063935  -2.632  0.00848 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1145.7  on 831  degrees of freedom
## Residual deviance:  997.7  on 829  degrees of freedom
## AIC: 1003.7
##
## Number of Fisher Scoring iterations: 4
```

```
model = glm(got ~ tinc + distvct,
            data = data_match,
            family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = got ~ tinc + distvct, family = "binomial", data = data_match)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5307  -1.0293   0.4237   1.0921   1.6008
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.082298   0.153184  -0.537  0.59109
## tinc         0.011126   0.001126   9.884  < 2e-16 ***
## distvct     -0.168299   0.063935  -2.632  0.00848 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1145.7  on 831  degrees of freedom
## Residual deviance:  997.7  on 829  degrees of freedom
## AIC: 1003.7
##
## Number of Fisher Scoring iterations: 4
```

10 References

1. Data: https://github.com/NickCH-K/causaldata/blob/main/Python/causaldata/thornton_hiv/data.py
2. Actual experiment : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115776/