# Assignment 1

Randomized Controlled Trials

4/11/2022

## Contents

# Background

You have been asked by a well-meaning NGO, Kenya Electricity Loss Lessening Experts, Registered (KELLER) to help them learn about the impacts of their flagship program. KELLER works with the Kenyan government to disconnect electricity users who do not pay their bills, and hypothesizes that these disconnections increase payment.

# Question 1:

KELLER would like to know about the payment impacts of their disconnections program. They say they're interested in measuring the impact of their disconnections, but don't exactly know what that means. Use the potential outcomes framework to describe the impact of treatment (defined as "disconnecting a household's electricity") for household $i$ on electricity payments (measured in rupees) formally (in math) and in words.

## Answer

To evaluate the impact of the disconnections program, we would ideally want to observe a household $i$ in both states: in the state of disconnection and non-disconnection and then get the difference between these two states of the same household to come up with the causal impact of the disconnections program. This would correspond to estimating:

$$\tau_i = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

Where $\mathbb{E}[Y_i(1)]$ is the electricity payments of the household $i$ when their electricity is disconnected and $\mathbb{E}[Y_i(0)]$ is the electricity payments of the same household when their electricity is not disconnected. Therefore, the difference between the two potential outcomes, $\tau_i$, would give us the impact of the program.

# Question 2

KELLER are extremely impressed. They want to know how they can go about measuring $\tau_i$. Let them down gently, but explain to them why estimating $\tau_i$ is impossible.

## Answer

As discussed above, estimating $\tau_i$ assumes that we have both $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$. In other words, to estimate $\tau_i$, we would need to observe the same individual in two mutually exclusive states, which is impossible since we cannot both disconnect and not disconnect the electricty of the same househol simultaneously. This is the *fundamental problem of causal inference*, we cannot observe both $Y_i(1)$ and $Y_i(0)$.

# Question 3

KELLER are on board with the idea that they can't estimate individual-specific treatment effects. They suggest estimating the average treatment effect instead. They are willing to give you some of their early data on payments. They have data on households who did and didn't get disconnected, and want you to compare the average payments across the two sets of households. Describe what this is actually measuring, and provide an example of why this may differ from the average treatment effect.

### Answer

If we compare the payments of households who have been disconnected and households who have not, without information on which households were part of the disconnections program, we would not be able to estimate the average treatment effect of interest. Firstly, individuals who have been disconnected in the past would skew towards individuals who are more likely to fall behind on payments and this would compromise the results by introducing a significant selection bias. Secondly, if only a part of the population was included in the disconnections program, without knowing the non-disconnected individuals' status of program participation, we cannot tell whether the payments made by them were due to the disconnections program or not. Without random assignment to treatment and control groups, we cannot estimate the average treatment effect (ATE) in this context.

## Question 4

KELLER have realized the error of their ways. Their CEO tells you, "Okay, we understand that our data won't let us estimate the average treatment effect. But can't we estimate the average treatment effect on the treated?" First formally (in math) define the ATT in this context, and then explain whether or not the KELLER data will allow you to estimate it. If so, describe how what you see in the data corresponds to the necessary components of the ATT. If not, explain why not, and describe what you can't see in the data that you'd need to observe.

### Answer

The ATT in this context would be:

$$\tau^{ATT} = \mathbb{E}[\tau_i | D_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$$

To estimate ATT we would need to have information on which individuals were randomly assigned to the disconnections program, and then we would need to measure the impact of the disconnections program on households who were assigned to it. The KELLER data does not give us random assignment to the program and corresponding outcomes as we would see in a randomized controlled trial output. Therefore, it would not allow us to estimate $\tau^{ATT}$. And even if we had RCT results, we would face the problem that we can never observe $\mathbb{E}[Y_i(0)|D_i = 1]$, i.e., the potential outcome of non-participation for individuals who participated in the disconnections program. But, random assignment would give us sufficient basis that it would be equal (or very close) to $\tau^{ATE}$ by eliminating selection bias.

## Question 5

KELLER forgot to tell you that they ran a randomized pilot study to estimate the effects of disconnections on payments. They're happy to share those data with you: find it in `ps1_data_22.csv`. This experience has made you a little bit skeptical of KELLER's skills, so start by checking (with a proper statistical test) that the treatment group and control group are balanced in pre-treatment payments, electricity usage, household size, and household head age. Use `keller_trt` as your treatment variable. Report your results. What do you find?

### Answer

```r
# load data
keller_df <- read_csv("ps1_data_22.csv")
```

```r
balance_payments <- lm(baseline_payments ~ keller_trt, keller_df)
balance_elec_use <- lm(baseline_elec_use ~ keller_trt, keller_df)
balance_hhsize <- lm(baseline_hhsize ~ keller_trt, keller_df)
balance_hh_head_age <- lm(baseline_hh_head_age ~ keller_trt,
    keller_df)

stargazer(balance_payments, balance_elec_use, balance_hhsize,
    balance_hh_head_age, type = "latex", header = FALSE, no.space = TRUE,
    column.sep.width = "3pt", font.size = "small")
```

Table 1:

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | baseline_payments | baseline_elec_use | baseline_hhsize | baseline_hh_head_age |
| | (1) | (2) | (3) | (4) |
| keller_trt | −3,999.728 | 11.028 | −0.027 | −0.063 |
| | (3,999.974) | (7.912) | (0.040) | (0.099) |
| Constant | 4,130.085 | $388.297^{***}$ | $7.995^{***}$ | $35.067^{***}$ |
| | (2,828.692) | (5.595) | (0.029) | (0.070) |
| Observations | 10,000 | 10,000 | 10,000 | 10,000 |
| R$^2$ | 0.0001 | 0.0002 | 0.00004 | 0.00004 |
| Adjusted R$^2$ | −0.000 | 0.0001 | −0.0001 | −0.0001 |
| Residual Std. Error (df = 9998) | 199,998.700 | 395.607 | 2.024 | 4.973 |
| F Statistic (df = 1; 9998) | 1.000 | 1.943 | 0.445 | 0.396 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

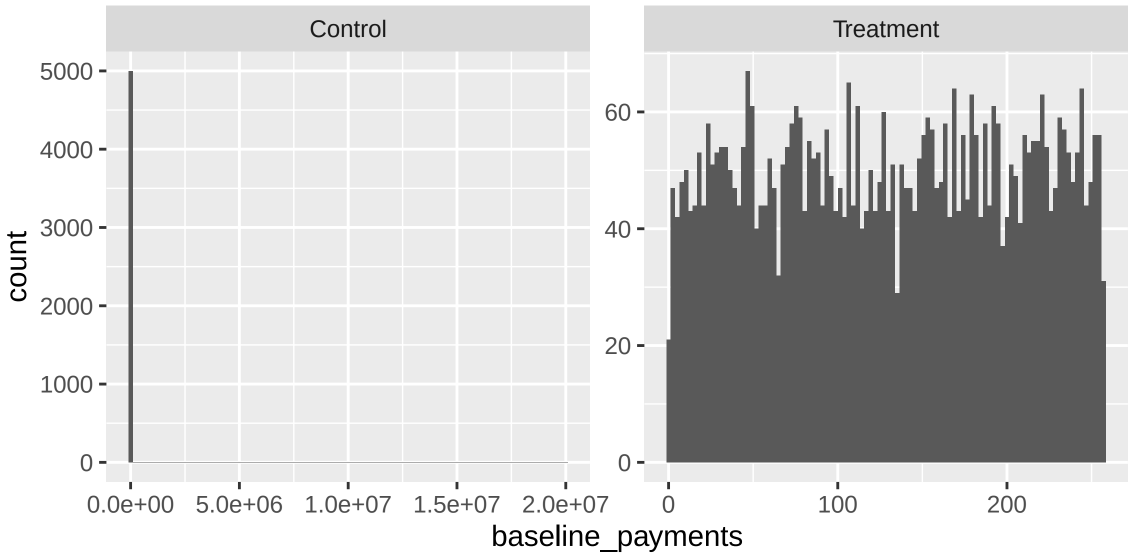
Based on the results above, the treatment and control groups seem to be balanced. We can infer this from the fact that none of the coefficients in the four regression models are significantly different than 0. The balance tests *do not* indicate any selection bias. That being said, the coefficient for baseline payments seem to be unusually high, which might warrant further inspection.

# Question 6

Plot a histogram of pre-treatment payments for treated farms and control households. What do you see? Re-do your balance table to reflect any necessary adjustments. What does this table tell you about whether or not KELLER's randomization worked? What assumption do we need to make on unobserved characteristics in order to be able to estimate the causal effect of keller_trt?

## Answer

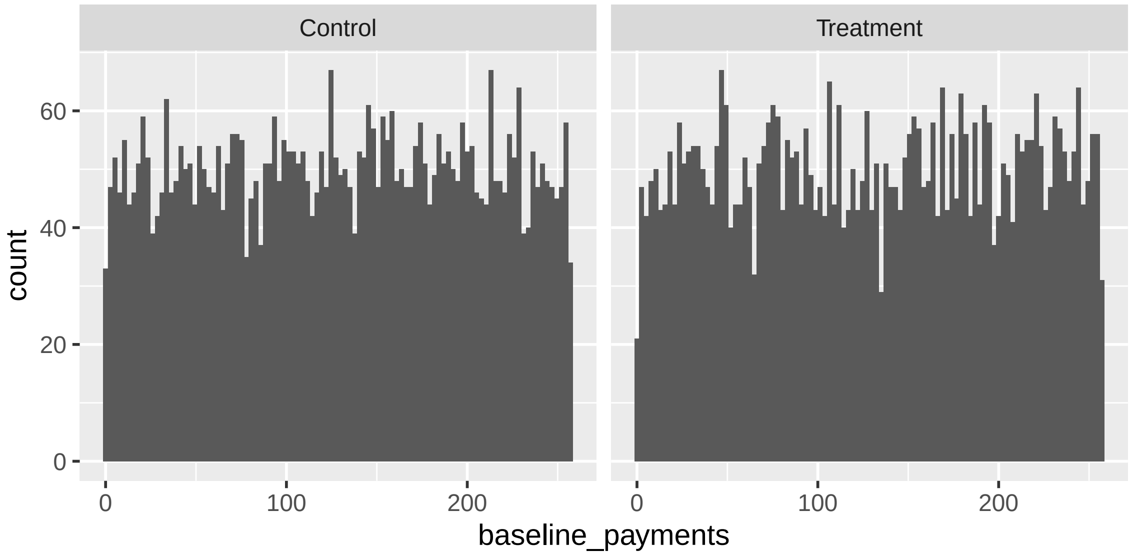
```r
keller_df %>%
    mutate(keller_trt = ifelse(keller_trt == 0, "Control", "Treatment")) %>%
    ggplot() + geom_histogram(aes(baseline_payments), bins = 100) +
    facet_wrap(~keller_trt, scales = "free")
```



There seems to be an outlier in the untreated group that is skewing the baseline payment distribution. By the extremity of the value, we can assume that this is a data error and drop that observation from the data.

```r
keller_df <- keller_df %>%
    filter(baseline_payments < 2000)

keller_df %>%
    mutate(keller_trt = ifelse(keller_trt == 0, "Control", "Treatment")) %>%
    ggplot() + geom_histogram(aes(baseline_payments), bins = 100) +
    facet_wrap(~keller_trt, scales = "free_x")
```



```r
balance_payments <- lm(baseline_payments ~ keller_trt, keller_df)
balance_elec_use <- lm(baseline_elec_use ~ keller_trt, keller_df)
balance_hhsize <- lm(baseline_hhsize ~ keller_trt, keller_df)
balance_hh_head_age <- lm(baseline_hh_head_age ~ keller_trt,
    keller_df)

stargazer(balance_payments, balance_elec_use, balance_hhsize,
    balance_hh_head_age, type = "latex", header = FALSE, no.space = TRUE,
    column.sep.width = "3pt", font.size = "small")
```

Table 2:

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | baseline_payments | baseline_elec_use | baseline_hhsize | baseline_hh_head_age |
| | (1) | (2) | (3) | (4) |
| keller_trt | 1.047 | 10.977 | −0.027 | −0.061 |
| | (1.486) | (7.913) | (0.040) | (0.099) |
| Constant | 129.311$^{***}$ | 388.348$^{***}$ | 7.995$^{***}$ | 35.066$^{***}$ |
| | (1.051) | (5.596) | (0.029) | (0.070) |
| Observations | 9,999 | 9,999 | 9,999 | 9,999 |
| R$^2$ | 0.00005 | 0.0002 | 0.00005 | 0.00004 |
| Adjusted R$^2$ | −0.0001 | 0.0001 | −0.0001 | −0.0001 |
| Residual Std. Error (df = 9997) | 74.272 | 395.618 | 2.024 | 4.973 |
| F Statistic (df = 1; 9997) | 0.496 | 1.924 | 0.451 | 0.381 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

After we filter drop the outlier from data, histograms for both groups approximate a uniform random distribution. And the coefficient for the `baseline_payments` variable became more reasonable with a significantly lower standard error. This strengthens the assumption of random assignment.

To make causal claims on the effect of the program, we need to assume that the unobserved characteristics of these households are, on average, not significantly different from each other. The information we have on the observed characteristics are supporting this assumption so far.

# Question 7

Assuming that `keller_trt` is indeed randomly assigned, describe how to use it to estimate the average treatment effect, and then do so. Please describe your estimate: what is the interpretation of your coefficient (be clear about your units)? Is your result statistically significant? Is the effect you find large or small, relative to the mean in the control group?

## Answer

Now that we can make the assumption that the treatment is randomly assigned, we can use it to estimate $\tau^{ATE}$ as:

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

Where $\overline{Y(1)}$ and $\overline{Y(0)}$ is the average payments after the experiment (endline) for the treatment and control groups respectively.

```
ate_payments <- lm(endline_payments ~ keller_trt, keller_df)
ate_elec_use <- lm(endline_elec_use ~ keller_trt, keller_df)
```

```r
stargazer(ate_payments, ate_elec_use, type = "latex", header = FALSE,
    no.space = TRUE, column.sep.width = "3pt", font.size = "small")
```

Table 3:

| | *Dependent variable:* | |
|---|---|---|
| | endline_payments | endline_elec_use |
| | (1) | (2) |
| keller_trt | $-4{,}679.066^{***}$ | $-37.441^{***}$ |
| | (985.230) | (7.882) |
| Constant | $48{,}673.340^{***}$ | $388.348^{***}$ |
| | (696.767) | (5.574) |
| Observations | 9,999 | 9,999 |
| $R^2$ | 0.002 | 0.002 |
| Adjusted $R^2$ | 0.002 | 0.002 |
| Residual Std. Error (df = 9997) | 49,259.050 | 394.063 |
| F Statistic (df = 1; 9997) | $22.555^{***}$ | $22.566^{***}$ |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

On average, disconnected households seem to make less payments than households who were not disconnected. The coefficient indicates that, on average, treated group paid 4,679.066 rupees less than the control group. The result is significant at 0.01 level.

```r
control_mean <- keller_df %>%
    filter(keller_trt == 0) %>%
    pull(endline_payments) %>%
    mean()

print(paste("Average payment amount in the control group:", round(control_mean,
    2)))
```

```
[1] "Average payment amount in the control group: 48673.34"
```

```r
print(paste("Percent impact: ", round(ate_payments$coefficients["keller_trt"]/control_mean *
    100, 2), "%", sep = ""))
```

```
[1] "Percent impact: -9.61%"
```

This effect seems to be fairly large, considering that the average payments in the control group is 48,673.34 rupees. The program seems to decrease the payments by 9.61%.

## Question 8

KELLER is convinced that the reason their disconnections are effective is because they are getting households to use less electricity. They want you to estimate the effects of the disconnections, but controlling for the endline amount of power consumed. Is this a good idea? Why or why not? Run this regression and describe your estimates. How do they differ from your results in (7)? What about controlling for baseline electricity consumption? Run this regression and describe your estimates. How do they differ from your results in (7)? How do the two estimates differ? What is driving any differences between them?

## Answer

In RCTs, controlling for post-treatment outcomes (e.g. endline electricity consumption) are not a good idea. It might lead to some portion of the treatment effect being captured by the control coefficient $\gamma$ rather than the ATE parameter of interest $\tau$. This can lead to bias as our treatment can have an impact on electricity consumption as well, which will likely lead us to underestimate the ATE.

Controlling for baseline consumption might be a good idea. Even though we made sure that there were no statistically significant differences in baseline measures between groups, it can help us shrink the standard errors of the coefficient of interest and increase the statistical precision of the model.

We estimate these two models anyway:

$$Y_i = \alpha + \tau D_i + \gamma X_i^{endline} + \epsilon_i$$

$$Y_i = \alpha + \tau D_i + \gamma X_i^{baseline} + \epsilon_i$$

Where $X_i^{endline}$ is the electricity consumption by the household $i$ after the experiment, $X_i^{baseline}$ is the electricity consumption by the household $i$ before the experiment, $D_i$ is the dummy treatment variable, $\alpha$ is the constant, $\epsilon$ is the error term and $Y_i$ is the electricity payments by the household $i$ after the experiment.

```r
ate_payments_control_end <- lm(endline_payments ~ keller_trt +
    endline_elec_use, keller_df)
ate_payments_control_base <- lm(endline_payments ~ keller_trt +
    baseline_elec_use, keller_df)

stargazer(ate_payments_control_end, ate_payments_control_base,
    type = "latex", header = FALSE, no.space = TRUE, column.sep.width = "3pt",
    font.size = "small")
```

Table 4:

| | *Dependent variable:* | |
|---|---|---|
| | endline_payments | |
| | (1) | (2) |
| keller_trt | 1.147 | $-6{,}043.187^{***}$ |
| | (1.487) | (60.575) |
| endline_elec_use | $125.003^{***}$ | |
| | (0.002) | |
| baseline_elec_use | | $124.276^{***}$ |
| | | (0.077) |
| Constant | $128.767^{***}$ | $410.932^{***}$ |
| | (1.280) | (52.142) |
| Observations | 9,999 | 9,999 |
| $R^2$ | 1.000 | 0.996 |
| Adjusted $R^2$ | 1.000 | 0.996 |
| Residual Std. Error (df = 9996) | 74.267 | 3,028.325 |
| F Statistic (df = 2; 9996) | $2{,}203{,}926{,}641.000^{***}$ | $1{,}320{,}520.000^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

When we control for `endline_elec_use`, we see that the average treatment effect is no longer significant. However, interpreting this as treatment did not effect the outcome would be incorrect because we controlled for a post-treatment outcome. And almost all of our treatment effects seems to be captured by this outcome variable, leading to a biased causal estimator.

When we control for `baseline_elec_use`, ATE remains significant and it's magnitude increases further. This means that the change in payments cannot be explained just by the baseline amount of electricity consumed by households. This was already hinted when we looked at the balance tests and saw that two groups were balanced in terms of their baseline payments. Nevertheless, adding baseline electricity usage into our model yields significantly lower standard errors and higher F Statics, indicating a more precise causal model. Finally, both models have drastically higher $R^2$ values, revealing that most of the variation in the payments come from electricity usage, as we would expect.

# Question 9

One of the KELLER RAs (the real workforce!) informs you that not everybody who was supposed to be disconnected – (keller_trt = 1) actually got disconnected. She tells you that the actual treatment indicator is keller_trt_yes. (Since disconnections are expensive, KELLER assures you that nobody in the control group got disconnected). In light of this new information, what did you actually estimate in question (7)? How does this differ from what you thought you were estimating?

## Answer

In this case, it looks like we have some non-compliance in the form of *never-takers* (never treated regardless of treatment assignment). And what we ended up estimating happens to be intent to treat (ITT) estimate (or the effect of assignment to treatment), which can also be helpful in a policy setting, since in real life policy implementation, the rate of non-compliance will likely be similar to the one we observe in the experiment.

Nonetheless, this is different than what we think we were estimating, which was ATE. The actual ATE could be different than what we estimated above, as not all the units that were assigned to the treatment group received the treatment.

# Question 10

KELLER aren't actually interested in the effect of assignment to treatment - they want to know about the actual effects of their disconnections. Describe (in math, and then in words) what you can estimate using the two treatment variables we observe, keller_trt and keller_trt_yes. Estimate this object (you can ignore standard errors just for this once). Interpret your findings. How does this compare to what you estimated in (7)?

## Answer

Using the two treatment variables, we can approximate the Local Average Treatment Effect (LATE) as:

$$\hat{\tau}^{LATE} = \frac{\overline{Y}(R=1) - \overline{Y}(R=0)}{Pr(D_i = 1|R_i = 1)}$$

Where the denominator of the above expression corresponds to the fraction of treatment group units receiving treatment. And the numerator is what we estimated with the models above using the `keller_trt` variable, since we found out that `killer_trt` gives us only the $R_i$ (assignment to treatment) and not $D_i$ (the actual treatment).

```r
numerator <- ate_payments_control_base$coefficients["keller_trt"]
denominator <- lm(keller_trt_yes ~ keller_trt, data = keller_df)$coefficients["keller_trt"]

att <- numerator/denominator
print(round(att, 2))
```

```
keller_trt
  -8593.11
```

This approximately gives us the Local Average Treatment Effect (LATE), and since we assume there was no selection bias in treatment assignment and there is treatment effect homogeneity, ATT and ATE as well. Therefore, we can conclude by saying that the average treatment effect of disconnection is a reduction in electricity payments by 8,593.11 rupees. Noncompliance appears to have led to an underestimation of the ATE by around 2,550 rupees.