PPHA 34600 Program Evaluation

Problem Set 1

4/14/2022

(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.

Using the potential outcomes framework, we would theoretically measure the impact of disconnecting a household's electricity by taking the difference of the potential outcome on electricity payments for a given household under treatment(disconnecting its electricity) minus the outcome for that same household not being treated(not having its electricity disconnected).

Mathematically, this would be represented as follows:

$$\tau_{householdi} = Y_{householdi}(D = 1) - Y_{householdi}(D = 0)$$

$$\tau_{householdi} = Y_{householdi}(1) - Y_{householdi}(0)$$

Where: $D_i = 1$ for treated households and $D_i = 0$ for untreated households (ie. control group). In this case treatment refers to disconnecting a household's electricity.

$Y_i(D_i)$ is the outcome in electricity payments (measured in rupees) for household i with treatment status D_i and τ_i is the treatment effect for an individual household i.

(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.

The treatment effect for each individual household is only a theoretical estimation. In reality, it is impossible to calculate τ_i because we need to observe the two potential outcomes for each individual household under both states of the world: treatment and no treatment(ie. control). That is, observing the potential outcome for a treated household and at the same time, what would have been its potential outcome had the household not been treated. This is known as the fundamental problem of causal inference, since we cannot ever observe both $Y_{householdi}(1)$ and $Y_{householdi}(0)$ simultaneously for the same household.

(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.

This early data on payments does not correspond to a process of randomized assignment to treatment and control groups, it simply is a registry of household payments and whether they were disconnected or not. In the absence of random assignment, this is not the same as the ATE and instead is a naive estimator based on **observed outcomes** instead of **potential outcomes**.

What this is measuring is:

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

Which is NOT the same as ATE:

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

For example, if the households that were behind on their payments and paid after their electricity was disconnected were households in a coal mine neighborhood where workers hadn't been paid and once they were paid, they paid for their electricity bill, then what caused them to settle their bill was receiving their overdue wages and not the fact that their electricity was cut-off. This would be a case of selection bias, where the households that pay their bills are systematically different from the households that don't pay their electricity bills for reasons **other** than treatment. In this example, we would be incorrectly overestimating the effect of disconnecting a household's electricity on their payments when there are economic and labor factors behind that are responsible for that causality.

(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.

In this context, the average treatment effect on the treated is:

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

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

From the data, we cannot observe the second term $E[Y_i(0)|D_i = 1]$, that is what would have been the potential outcome (endline_payments) for households whose electricity actually got disconnected ($D_i = 1$) had their electricity not been disconnected $Y_i(0)$. Similar as before, the fundamental problem of causal inference remains and we cannot actually calculate ATT from the given data.

(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?

```
# First we segment the data into treatment and control groups:
treatment <- filter(keller, keller_trt == 0)
control <- filter(keller, keller_trt == 1)

# Then, we segment the treatment and control groups by each baseline characteristic
# and run a t-test to check if they are balanced:

# (1) Pre-treatment payments (baseline_payments)
payments_t <- treatment %>% pull(baseline_payments)
payments_c <- control %>% pull(baseline_payments)
mt1 = mean(treatment$baseline_payments)
mc1 = mean(control$baseline_payments)
sdt1 = sd(treatment$baseline_payments)
sdc1 = sd(control$baseline_payments)
t_test <-t.test(payments_t, payments_c)
```

```
p1 = t_test$p.value

# (2) Electricity usage (baseline_elec_use)
elec_t <- treatment %>% pull(baseline_elec_use)
elec_c <- control %>% pull(baseline_elec_use)
mt2 = mean(treatment$baseline_elec_use)
mc2 = mean(control$baseline_elec_use)
sdt2 = sd(treatment$baseline_elec_use)
sdc2 = sd(control$baseline_elec_use)
t_test <-t.test(elec_t, elec_c)
p2 = t_test$p.value

# (3) Household size (baseline_hhsize)
hsize_t <- treatment %>% pull(baseline_hhsize)
hsize_c <- control %>% pull(baseline_hhsize)
mt3 = mean(treatment$baseline_hhsize)
mc3 = mean(control$baseline_hhsize)
sdt3 = sd(treatment$baseline_hhsize)
sdc3 = sd(control$baseline_hhsize)
t_test <-t.test(hsize_t, hsize_c)
p3 = t_test$p.value

# (4) Household head age (baseline_hh_head_age)
hage_t <- treatment %>% pull(baseline_hh_head_age)
hage_c <- control %>% pull(baseline_hh_head_age)
mt4 = mean(treatment$baseline_hh_head_age)
mc4 = mean(control$baseline_hh_head_age)
sdt4 = sd(treatment$baseline_hh_head_age)
sdc4 = sd(control$baseline_hh_head_age)
t_test <-t.test(hage_t, hage_c)
p4 = t_test$p.value

baseline = c("Payments", "Electricity Use", "Household Size", "Household Head Age")
p_value = c(p1, p2, p3, p4)
mean_treatment = c(mt1, mt2, mt3, mt4)
sd_treatment = c(sdt1, sdt2, sdt3, sdt4)
mean_control = c(mc1, mc2, mc3, mc4)
sd_control = c(sdc1, sdc2, sdc3, sdc4)
diff_means = c(mt1-mc1,mt2-mc2, mt3-mc3, mt4-mc4)
df = data.frame(baseline, mean_treatment, sd_treatment, mean_control,sd_control,diff_means, p_value)

kbl(df, caption = "Balance Table", booktabs = T, digits = 3,col.names =
      c("Baseline Characteristic", "Mean Treated", "Std Dev Treated",
      "Mean Control", "Std Dev Control", "Diff Means", "p-value"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

From the table, we can observe that after running a Welch Two-Sample t-test, the p-values of all four baseline characteristics are not statistically significantly different than 0 at any of the conventional levels of statistical significance (1%, 5% and 10%).

$$H_0 : meantreatment - meancontrol = 0$$

$$H_A : meantreatment - meancontrol \neq 0$$

Table 1: Balance Table

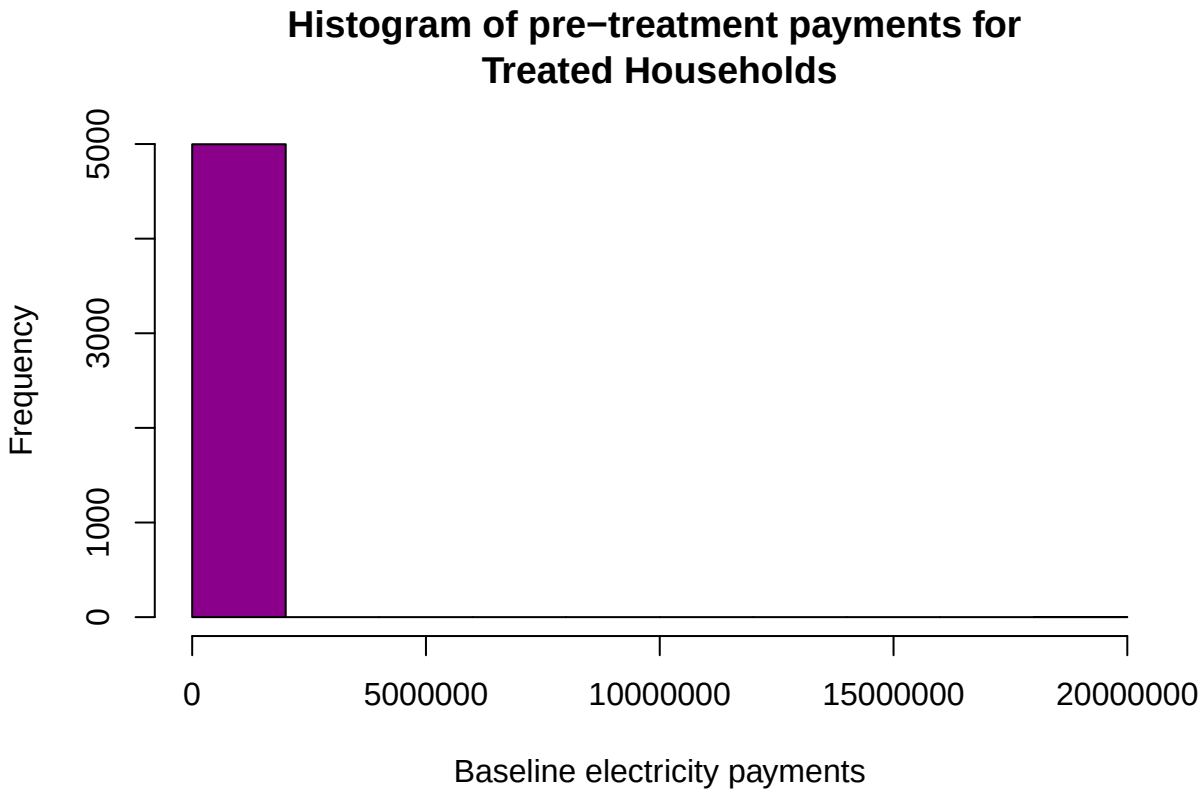
Baseline Characteristic	Mean Treated	Std Dev Treated	Mean Control	Std Dev Control	Diff Means	p-value
Payments	4130.085	282869.182	130.358	74.530	3999.728	0.317
Electricity Use	388.297	390.538	399.325	400.609	-11.028	0.163
Household Size	7.995	2.019	7.968	2.028	0.027	0.505
Household Head Age	35.067	4.974	35.005	4.972	0.063	0.529

This means that for all cases, we fail to reject the null hypothesis that the true difference in means between the the treatment and control groups is equal to 0, therefore the t-statistic results indicate that the treatment and control groups are statistically balanced in all the baseline characteristics.

It's noticeable from the table that the standard deviation for the treatment group baseline payments is much larger than the respective standard deviation for the control group (ie. $282869.182 > 74.530$). This suggests that there may be an issue with the distribution of the data for this baseline characteristic within the treatment group.

(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?

```
hist(treatment$baseline_payments,
     main = "Histogram of pre-treatment payments for \nTreated Households",
     xlab="Baseline electricity payments",
     col="darkmagenta")
```



```
min(treatment$baseline_payments)
```

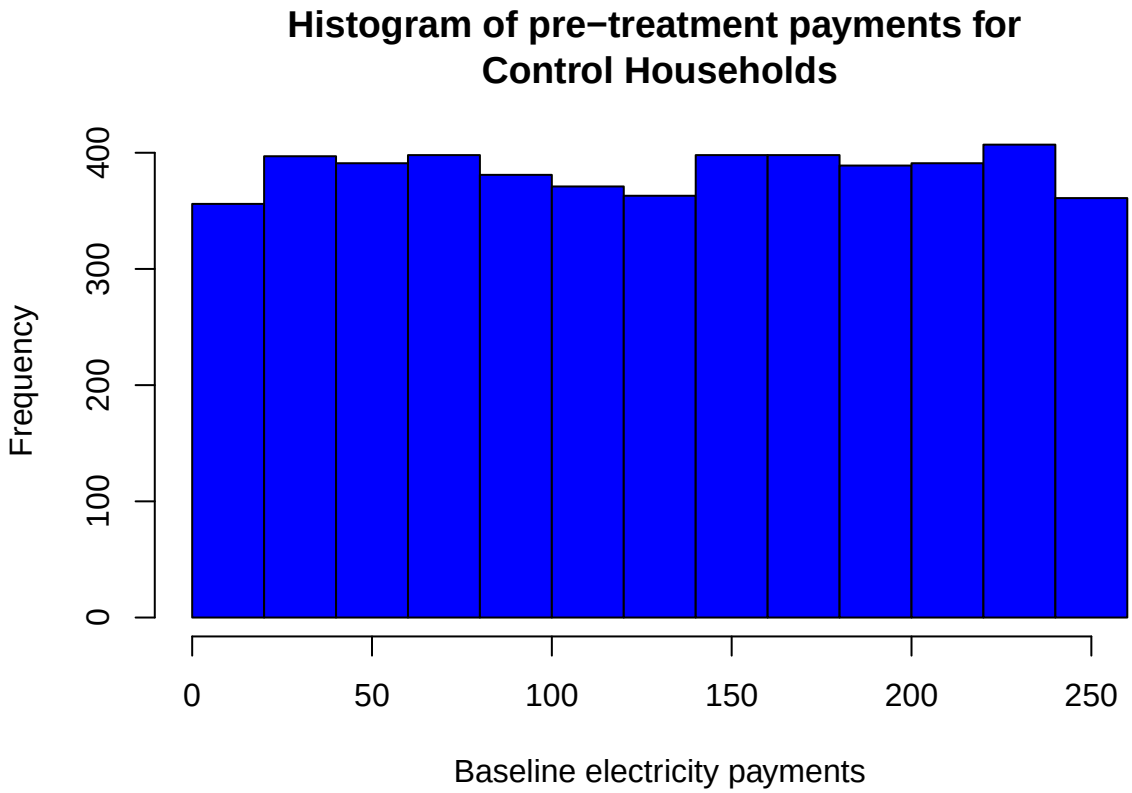
```
## [1] 0.0278301
```

```
max(treatment$baseline_payments)
```

```
## [1] 20000000
```

In the histogram of electricity payments for the treatment group, the graph looks very distorted due to an outlier that is very large = $20.000.000

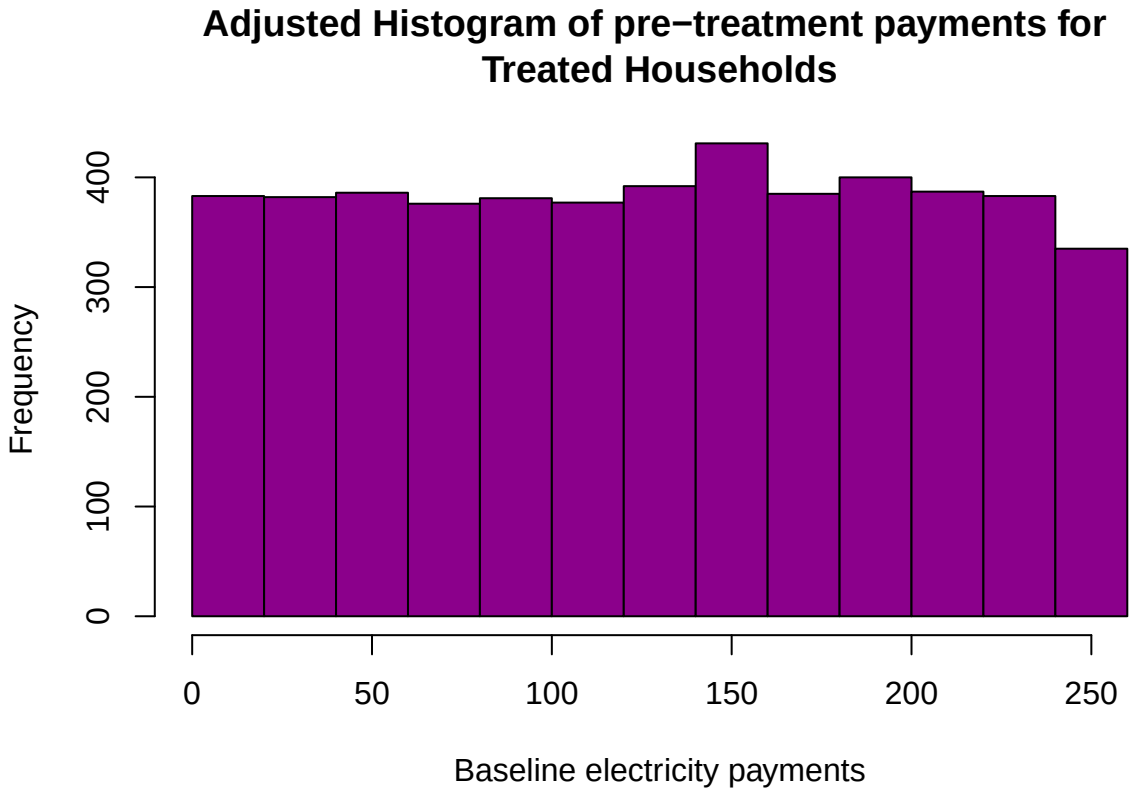
```
hist(control$baseline_payments,
     main = "Histogram of pre-treatment payments for \nControl Households",
     xlab="Baseline electricity payments",
     col="blue")
```



The histogram of electricity payments for the control group does not show any extreme outliers like in the case of the treatment group. It shows that the data for electricity payments at the baseline is evenly distributed.

Since the outlier in the previous case distorts the data, it is important to drop this observation:

```
keller_filtered <- keller[keller$baseline_payments <20000000,]
treatment <- filter(keller_filtered, keller_trt == 0)
hist(treatment$baseline_payments,
     main = "Adjusted Histogram of pre-treatment payments for \nTreated Households",
     xlab="Baseline electricity payments",
     col="darkmagenta")
```



After dropping the extreme outlier, we can observe that the data for baseline payments is evenly distributed for the treatment group. We need to recalculate the balance table with this adjustment:

```
# (1) Pre-treatment payments (baseline_payments)
payments_t <- treatment %>% pull(baseline_payments)
payments_c <- control %>% pull(baseline_payments)
mt1 = mean(treatment$baseline_payments)
mc1 = mean(control$baseline_payments)
sdt1 = sd(treatment$baseline_payments)
sdc1 = sd(control$baseline_payments)
t_test <-t.test(payments_t, payments_c)
p1 = t_test$p.value

# (2) Electricity usage (baseline_elec_use)
elec_t <- treatment %>% pull(baseline_elec_use)
elec_c <- control %>% pull(baseline_elec_use)
mt2 = mean(treatment$baseline_elec_use)
mc2 = mean(control$baseline_elec_use)
sdt2 = sd(treatment$baseline_elec_use)
sdc2 = sd(control$baseline_elec_use)
t_test <-t.test(elec_t, elec_c)
p2 = t_test$p.value

# (3) Household size (baseline_hhsize)
hsize_t <- treatment %>% pull(baseline_hhsize)
hsize_c <- control %>% pull(baseline_hhsize)
mt3 = mean(treatment$baseline_hhsize)
mc3 = mean(control$baseline_hhsize)
sdt3 = sd(treatment$baseline_hhsize)
sdc3 = sd(control$baseline_hhsize)
```

```
t_test <-t.test(hsize_t, hsize_c)
p3 = t_test$p.value

# (4) Household head age (baseline_hh_head_age)
hage_t <- treatment %>% pull(baseline_hh_head_age)
hage_c <- control %>% pull(baseline_hh_head_age)
mt4 = mean(treatment$baseline_hh_head_age)
mc4 = mean(control$baseline_hh_head_age)
sdt4 = sd(treatment$baseline_hh_head_age)
sdc4 = sd(control$baseline_hh_head_age)
t_test <-t.test(hage_t, hage_c)
p4 = t_test$p.value

baseline = c("Payments", "Electricity Use", "Household Size", "Household Head Age")
p_value = c(p1, p2, p3, p4)
mean_treatment = c(mt1, mt2, mt3, mt4)
sd_treatment = c(sdt1, sdt2, sdt3, sdt4)
mean_control = c(mc1, mc2, mc3, mc4)
sd_control = c(sdc1, sdc2, sdc3, sdc4)
diff_means = c(mt1-mc1,mt2-mc2, mt3-mc3, mt4-mc4)
df = data.frame(baseline, mean_treatment, sd_treatment, mean_control,sd_control,diff_means, p_value)

kbl(df, caption = "Adjusted Balance Table", booktabs = T, digits = 3,col.names =
      c("Baseline Characteristic", "Mean Treated", "Std Dev Treated",
      "Mean Control", "Std Dev Control", "Diff Means","p-value"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

Table 2: Adjusted Balance Table

Baseline Characteristic	Mean Treated	Std Dev Treated	Mean Control	Std Dev Control	Diff Means	p-value
Payments	129.311	74.012	130.358	74.530	-1.047	0.481
Electricity Use	388.348	390.560	399.325	400.609	-10.977	0.165
Household Size	7.995	2.019	7.968	2.028	0.027	0.502
Household Head Age	35.066	4.974	35.005	4.972	0.061	0.537

From the adjusted balance table, we can observe that the treatment and control groups remain balanced in all four baseline characteristics at all conventional levels of statistical significance. The p-values increased in size in every case, showing that the degree of balance is higher than previously anticipated. With the outlier removed, the difference in means for baseline payments is not as drastic as before and the standard deviation is 74 both for the treatment and control groups.

This tells us that KELLER'S randomization worked and that treatment and control groups are statistically similar, on key observables, at the baseline. In order to estimate the causal effect of keller_trt, we need to assume:

$$E[\varepsilon_i|D_i] = 0$$

Which means that the expectation of the error term, conditional on treatment, is zero. This is to say that there is no selection bias because we observe everything that is correlated with treatment and affects the outcome Y_i and those observable characteristics are balanced at the baseline. With this assumption in place, if we find a treatment effect, we can attribute it solely to the treatment and not to other unobservable characteristics that might be correlated with the treatment and the output.

(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?

Assuming that keller_trt is randomly assigned, we can use it to estimate ATE as follows:

$$\tau^{ATE} = \overline{Y_i}(D = 1) - \overline{Y_i}(D = 0)$$

This difference in means is essentially what a regression does:

$$Y_i = \beta + \tau D_i + \varepsilon_i$$

Where τ captures the ATE coefficient. Since the treatment and control groups are balanced (statistically equal) at the baseline, we can run this regression to recover the ATE.

```
reg <- lm(endline_payments ~ keller_trt, data = keller_filtered)
summary(reg)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt, data = keller_filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48651 -34705 -15078  18125 541990
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  48673.3      696.8  69.856 < 0.0000000000000002 ***
## keller_trt   -4679.1      985.2  -4.749          0.00000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49260 on 9997 degrees of freedom
## Multiple R-squared:  0.002251,   Adjusted R-squared:  0.002151
## F-statistic: 22.55 on 1 and 9997 DF,  p-value: 0.000002071
```

From the regression, the difference in endline payments between a household that receives treatment (gets electricity disconnected) and a household that doesn't (ie. control), is estimated to be approximately -4.679,1 rupees. The coefficient is statistically significant at all conventional levels of significance since the p-value of $0.000 < 0.05$ and 0.01 and 0.1.

Relative to the mean endline payments for the control group (given by the intercept), which is equivalent to 48.673,3 rupees, the estimated ATE is economically significant since the percentage change in endline payments for a treated household is -9.61%, which is a considerable amount.

```
(((48673.3 -4679.1) - 48673.3)/48673.3)*100
```

```
## [1] -9.613279
```

(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?

It is a bad idea to include the endline amount of power consumed in the regression because this post-treatment variable could potentially be correlated with our outcome of interest (electricity payments) and therefore, it could likely have been impacted by the treatment. That is, a decrease in the amount of power consumed could be a direct effect of the households that were treated by getting their electricity cut off.

```
reg2 <- lm(endline_payments ~ keller_trt + endline_elec_use, data = keller_filtered)
summary(reg2)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt + endline_elec_use,
##     data = keller_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135.24  -63.95    0.40   63.97  128.97
##
## Coefficients:
##                     Estimate Std. Error   t value            Pr(>|t|)
## (Intercept)      128.767262   1.280390   100.569 <0.0000000000000002 ***
## keller_trt         1.146583   1.487091     0.771               0.441
## endline_elec_use 125.002689   0.001885 66316.896 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.27 on 9996 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.204e+09 on 2 and 9996 DF,  p-value: < 0.00000000000000022
```

By running this regression, we obtain that, ceteris paribus, our ATE estimate is not statistically significant at any of the conventional levels of significance. (p-value: $0.441 > 0.05$). On the contrary, the coefficient on endline electricity use is highly statistically significant, because as expected, this variable was directly affected by the treatment.

This regression provides misleading results that the intervention had no statistically significant ATE due to bias introduced by including a post-treatment outcome control.

```
reg3 <- lm(endline_payments ~ keller_trt + baseline_elec_use, data = keller_filtered)
summary(reg3)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt + baseline_elec_use,
##     data = keller_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3685.0 -3029.8  -140.7   279.6  9037.7
##
```

```
## Coefficients:
##                       Estimate  Std. Error  t value             Pr(>|t|)
## (Intercept)         410.93161    52.14240    7.881  0.00000000000000359 ***
## keller_trt        -6043.18701    60.57536  -99.763 < 0.0000000000000002 ***
## baseline_elec_use   124.27611     0.07656 1623.290 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3028 on 9996 degrees of freedom
## Multiple R-squared:  0.9962, Adjusted R-squared:  0.9962
## F-statistic: 1.321e+06 on 2 and 9996 DF,  p-value: < 0.00000000000000022
```

When we control for baseline electricity consumption, we find that holding everything else constant, the difference in endline payments between a household that receives treatment (gets electricity disconnected) and a household that doesn't (ie. control), is estimated to be approximately -6.043,18 rupees. The coefficient is highly statistically significant at all conventional levels of significance.

Compared to the regression in (7), we find that the size of the coefficient increases from -4.679,1 rupees in (7) to -6.043,18 rupees, a difference of -1.0722.28 rupees. In both cases the coefficient on ATE is statistically significant at all conventional levels of significance. In terms of economic significance, the estimated percentage change in endline payments between a treated and control household is equivalent to -12.42%, a considerable increase from -9.61% in (7).

```
(((48673.3 -6043.18) - 48673.3)/48673.3)*100
```

```
## [1] -12.4158
```

Also, the standard errors decrease, which is what drives the differences in the size of the coefficients. For example, the standard error for the ATE coefficient in (7) was 985.2 compared to 60.57 in this regression, which shows more precision in our estimate.

Since baseline electricity consumption was balanced at the baseline, it is not required to include it in order to recover the ATE, however, the fact that the size of the ATE estimate increases in magnitude, suggests that baseline electricity consumption might be correlated with the outcome and the error term such that excluding it, lead to omitted variable bias and therefore led us to underestimate the size of the ATE.

(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?

Since not every household that was assigned to treatment was treated, this tells us that there was non-compliance in the RCT, either from outside the experiment (eg. because KELLER implementers failed to disconnect all the households it had intented to) or from inside the experiment (eg. households that reconnect themselves illegally).

In light of this new information, what I actually estimated in (7) was the Intent to Treat (ITT), that is, the impact of offering the program. Mathematically:

$$ITT = \bar{Y_i}(R = 1) - \bar{Y_i}(R = 0)$$

Where $R = assignmentstatus$ and $D = actualtreatmentindicator$ This is **different** from the ATE for the treated:

$$ATE = \overline{Y_i}(D = 1) - \overline{Y_i}(D = 0)$$

Instead of recovering the ATE of the treatment (disconnecting households' electricity), we are recovering the ATE of the assignment of treatment, known as ITT.

(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)?

Using the two treatment variables we observe, we can estimate the Local Average Treatment Effect (LATE), which is the effect of the treatment for the compliers. We can classify four types of households:

- **Never takers**: Never treated, regardless of (R) status, $D = 0$
- **Always takers**: Always treated, regardless of (R) status, $D = 1$
- **Compliers**: Treated in the treatment group, untreated in the control group. If $(R = 1)$, then $(D = 1)$. If $(R = 0)$, then $(D = 0)$
- **Defiers**: Untreated in the treatment group, treated in the control group.If $(R = 1)$, then $(D = 0)$. If $(R = 0)$, then $(D = 1)$

Assumptions:

- Typically there are no defiers in the sample
- Never takers and always takets are equally distributed in the treatment and control groups

With these assumption, LATE is essentially a weighted average of the ITT with respect to the share of compliers.

Mathematically, LATE is estimated as follows:

$$\tau^{LATE} = \frac{\overline{Y}(R_i = 1) - \overline{Y}(R_i = 0)}{\hat{\pi}^c}$$

$$\tau^{LATE} = \frac{\hat{\tau}^{ITT}}{\hat{\pi}^c}$$

To estimate LATE, we need to follow three steps:

1. Regress Y_i on R_i to recover τ^{ITT}
2. Regress D_i on R_i to recover π^c
3. Calculate τ^{LATE}

```
# Step (1):
itt <- lm(endline_payments ~ keller_trt, data = keller_filtered)
summary(itt)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt, data = keller_filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48651 -34705 -15078  18125 541990
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  48673.3      696.8  69.856 < 0.0000000000000002 ***
## keller_trt   -4679.1      985.2  -4.749          0.00000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49260 on 9997 degrees of freedom
## Multiple R-squared:  0.002251,   Adjusted R-squared:  0.002151
```

```
## F-statistic: 22.55 on 1 and 9997 DF,  p-value: 0.000002071
```

```
# Step (2):
pi_c <- lm(keller_trt_yes ~ keller_trt, data = keller_filtered)
summary(pi_c)
```

```
##
## Call:
## lm(formula = keller_trt_yes ~ keller_trt, data = keller_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7033  0.0000  0.0000  0.2967  0.2967
##
## Coefficients:
##                          Estimate           Std. Error t value
## (Intercept) -0.0000000000001577  0.00457026871149776     0.0
## keller_trt   0.70325934813033908  0.00646236661625688   108.8
##                         Pr(>|t|)
## (Intercept)                    1
## keller_trt  <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3231 on 9997 degrees of freedom
## Multiple R-squared:  0.5423, Adjusted R-squared:  0.5422
## F-statistic: 1.184e+04 on 1 and 9997 DF,  p-value: < 0.00000000000000022
```

```
# Step (3):
late = itt$coefficients[2] / pi_c$coefficients[2]
late
```

```
## keller_trt
##    -6653.4
```

$$\tau^{LATE} = \frac{-4,679.066}{0.7032593}$$
$$\tau^{LATE} = -6,653.4$$

As a result, we obtain that the estimated LATE is -6,653.4, which means that on average, electricity disconnections are associated with a decrease of -6,653.4 rupees in electricity payments for complying households. Compared to (7), this is larger than the estimated ITT of -4,669.6, which tells us that the effect of disconnecting electricity on endline payments is higher for complying households than an average household assigned to treatment.