PS1_PEval

12265092

11/04/2022

R Markdown

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 Question 1

Let i be household units $i \in \{1, 2, ...N\}$

Treatment Indicator D_i: $D_i \in \{0, 1\}$

Treated: $D_i = 1$: disconnecting a household's electricity
Untreated: $D_i = 0$: not disconnecting a household's electricity

Outcome treated: $Y_i(D_i = 1)$: Electricity payments of household i in case household's electricity is disconnected
Outcome untreated: $Y_i(D_i = 0)$: Electricity payments of household i in case household's electricity is not disconnected

We get the impact of treatment(i.e disconnecting household's electricity) τ_i from the difference between the above outcomes.

$\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

Question 2: KELLER are extremely impressed. They want to know how they can go about measuring τ_i. Let them down gently, but explain to them why estimating τ_i is impossible

The impact of treatment τ_i is the difference between the two outcomes i.e difference between the electricity payments if the household's electricity is disconected vs the electricity payments of the household's if electricity is not disconnected.

From above: $\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While we need both the outcomes at a given time to compute the impact of treatment, the problem is that at a given time, we cannot observe both the outcomes $ or $ we can only observe either $Y_i(D_i = 1) $ or $ Y_i(D_i = 0)$ at a given time.

In case a household is treated (i.e disconnecting the electricity), then the observed outcome would be $Y_i(D_i = 1) $(Electricity payments of household i when household's electricity is disconnected), and $ Y_i(D_i = 0)$ (Electricity payments of household i when household's electricity is not disconnected) would become an unobserved outcome.

Due to the un-observable outcome *or* not being able to observe both the outcomes at a given time, measuring τ_i is impossible.

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.

Average Treatment Effect τ^{ATE}

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

ATE measures the average effect of treatment across a population of households.

Calculating the average outcomes(Electricity payments) of both sets (the ones who got disconnected and the ones who did not get disconnected) and subtracting them, we are determining a Naive Estimator τ_N

$\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

where $\bar{Y}(D = 1)$ is the average outcome (Electricity payments) for households with treatment status 1 i.e disconnecting the electricity and $\bar{Y}(D = 0)$ is the average (Electricity payments) for households with treatement status 0 i.e not disconnecting the electricity. The Naive estimator τ_N(a sample average) is calculated based on observed outcomes where as the ATE(Average of population E[]) is calculated on potential outcomes.

Here we are observing $Y_i(D_i = 1)$ and $Y_j(D_j = 0)$, knowing that i is not equal to j.

This brings us to the assumptions that the expectation of Y is same as (conditional expectation of Y that D_i is 1)and same as the (conditional expectation of Y given D_i is 0). We are assuming that the average of Y given $D_i = 1$ is a good counterfactual for when $D_i = 0$. Below in mathematical expression form:

$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
and
$E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$

There can be a problem when i and j significantly differ from each other. Which is, the households that receive treatment differ a lot from the households that donot receive any treatment. This leads to a bias called the Selection bias. This bias can be explained through the following example:

Consider two types of households , one which are poor and other which are rich. Households that are treated i.e whose electricity got disconnected i.e treated are economically backward households and ones that are not treated are rich households. Consider a program running parallely, a financial benefit program for economically backward communities. This could potentially increase the electricity payments of the poor households who majorly conform to the treated group. If this happens, it would be impossible to isolate the effect on Electricity payments caused by the treatment from the total effect on Electricity Payments caused by both the treatment and the financial benefit program.

Due to the above stated bias and assumptions, the Naive estimator τ^N suggested by Keller might not be a good estimator of the τ^{ATE} i.e the Average Treatment Effect.

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 to Question 4

Average Treatment Effect on the Treated (ATT) :

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

$E[Y_i(1)|D_i = 1]$ is the average effect of the treatment (disconnecting the electricity) for households in the treatment group (actually got disconnected), which can be observed in the data $E[Y_i(0)|D_i = 1]$ is the average effect of the treatment (disconnecting the electricity) for households in the control group (who actually didnt get disconnected), which is a counterfactual that cannot be observed in the data

Since we have unobservable in the calculation of τ^{ATT}, it is not possible to calculate the Average treatment effect of the treated (τ^{ATT}) from the KELLER data.

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 for Question 5

```
data <- read_csv('ps1_data_22.csv')
```

```
## Rows: 10000 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): keller_trt, keller_trt_yes, baseline_hhsize, baseline_payments, ba...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
reg1 <- data %>%
  select("baseline_hhsize",
         "baseline_payments",
         "baseline_elec_use",
         "baseline_hh_head_age") %>%
  lapply(function(x) lm(x ~ keller_trt, data = data))

balance_test_summary1 <- reg1 %>%
  sapply(function(x) coef(summary(x))[c(2,8)]) %>%
  t()
balance_test_summary1 %>%
  kable(col.names = c("Difference in means", "P-value")) %>%
  kable_styling(position = "center", font_size = 11, latex_options = "hold_position")
```

	Difference in means	P-value
baseline_hhsize	-0.0269925	0.5048567
baseline_payments	-3999.7275772	0.3173646
baseline_elec_use	11.0276798	0.1634185
baseline_hh_head_age	-0.0626144	0.5290177

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

P value for baseline_hhsize = 0.5048, Difference in Mean = -0.026 P value for baseline_payments = 0.3173 Difference in Means = -3999.72 (~4000) P value for baseline_elec_use = 0.1634 Difference in Means = 11.02 P value for baseline_hh_head_age = 0.529 Difference in means = -0.062

Null Hypothesis H0: Difference in Means = 0

From the P- values observed for the pre-payment variables baseline_hhsize(household size), baseline_payments(Payments), baseline_elec_use(household electricity usage), baseline_hh_head_age(household head age) using the keller_trt as the treatment variable, we cannot reject the null hypothesis. This says that the randomization is fine and the treatment group and the control group are balanced in pre-treatment payments, electricity usage, household size, and household head age. Thus they are not statistically significant.

From the Difference in Means observed, we find an odd value for the variable baseline_payments(Payments).

```
summary(data)
```

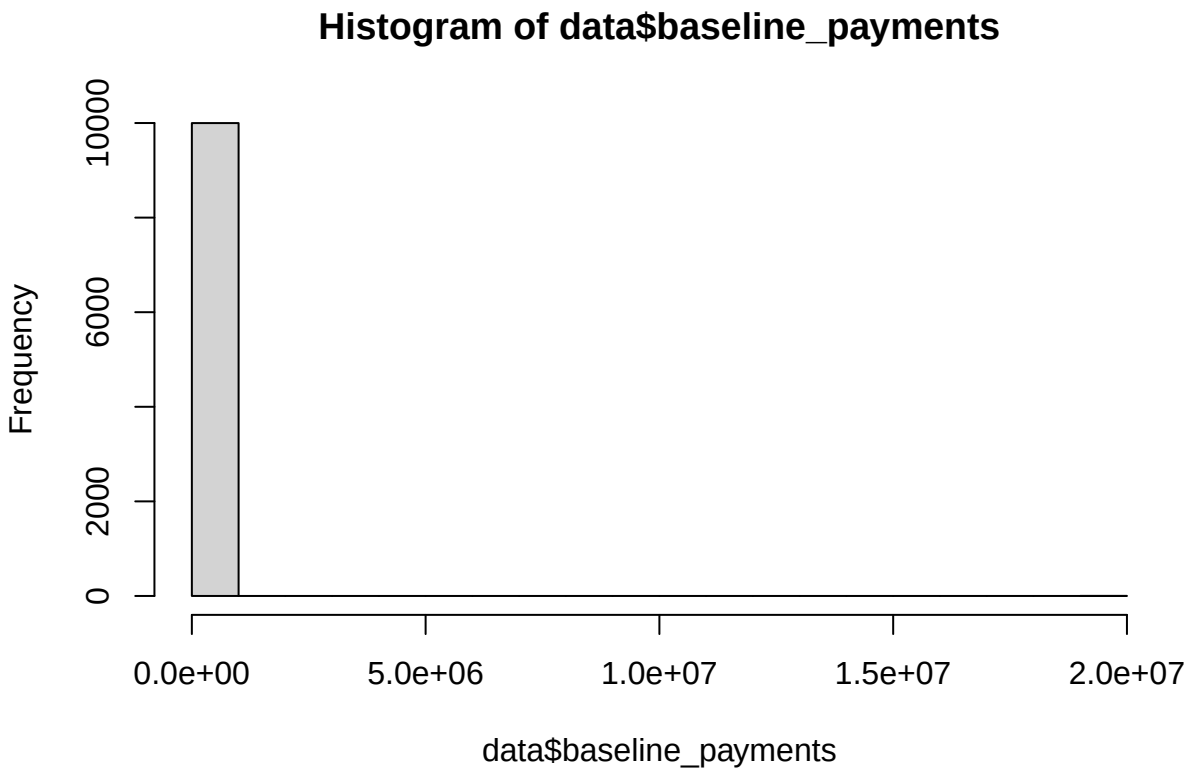
```
##    keller_trt     keller_trt_yes   baseline_hhsize  baseline_payments
##  Min.   :0.0000   Min.   :0.0000   Min.   : 1.000   Min.   :       0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7.000   1st Qu.:      66
##  Median :1.0000   Median :0.0000   Median : 8.000   Median :     130
##  Mean   :0.5001   Mean   :0.3517   Mean   : 7.981   Mean   :    2130
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 9.000   3rd Qu.:     194
##  Max.   :1.0000   Max.   :1.0000   Max.   :15.000   Max.   :20000000
##  baseline_elec_use  baseline_hh_head_age endline_hhsize   endline_payments
##  Min.   :   0.005   Min.   :13.00        Min.   : 1.000   Min.   :     0.7
##  1st Qu.: 115.954   1st Qu.:32.00        1st Qu.: 7.000   1st Qu.: 11450.3
##  Median : 276.654   Median :35.00        Median : 8.000   Median : 31385.0
##  Mean   : 393.812   Mean   :35.04        Mean   : 7.981   Mean   : 46328.6
##  3rd Qu.: 539.248   3rd Qu.:38.00        3rd Qu.: 9.000   3rd Qu.: 64253.8
##  Max.   :4687.777   Max.   :52.00        Max.   :15.000   Max.   :585984.2
##  endline_elec_use  endline_hh_head_age
##  Min.   :   0.00   Min.   :14.00
##  1st Qu.:  90.57   1st Qu.:33.00
##  Median : 249.88   Median :36.00
##  Mean   : 369.60   Mean   :36.04
##  3rd Qu.: 513.10   3rd Qu.:39.00
##  Max.   :4687.78   Max.   :53.00
```

baseline_payments Min. : 0 1st Qu.: 66 Median : 130 Mean : 2130 3rd Qu.: 194 Max. :20000000

This can be due to the Max value present in the baseline_payments 20000000 which is odd despite the median of 130 and mean of 2130.

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?

```
hist(data$baseline_payments)
```



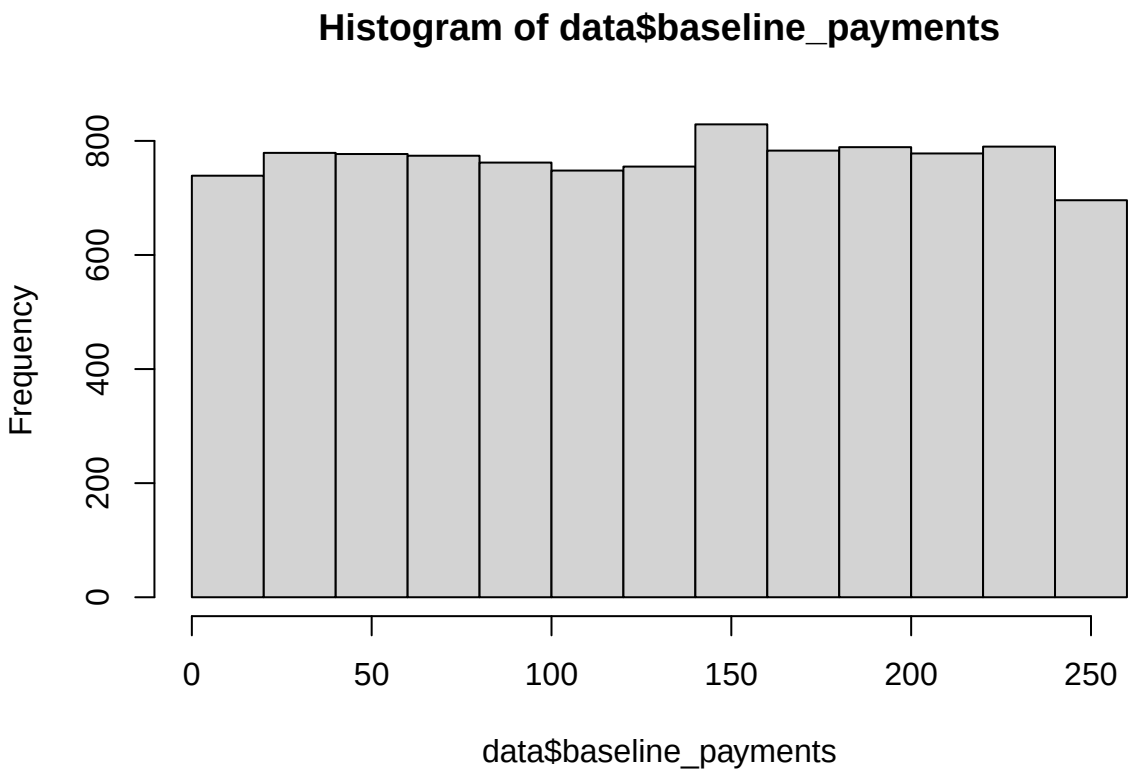
Lets filter the max values observed in the baseline_payments and then redo the test

```
data %>% filter(baseline_payments == 20000000)
```

```
## # A tibble: 1 x 10
##   keller_trt keller_trt_yes baseline_hhsize baseline_payments baseline_elec_use
##        <dbl>          <dbl>           <dbl>             <dbl>             <dbl>
## 1          0              0               7          20000000              133.
## # ... with 5 more variables: baseline_hh_head_age <dbl>, endline_hhsize <dbl>,
## #   endline_payments <dbl>, endline_elec_use <dbl>, endline_hh_head_age <dbl>
```

We see that there is only one row with baseline_payments value of 20000000. This is an outlier. Now lets remove this outlier and plot a histogram.

```
data <- data %>% filter(baseline_payments != 20000000)
hist(data$baseline_payments)
```



Our finding that only one outlier is correct. We now observe a reasonable with varied frequency across payments ranging 0 to 250 in the histogram.

Doing balance test again:

```
reg2 <- data %>%
  select("baseline_hhsize",
         "baseline_payments",
         "baseline_elec_use",
         "baseline_hh_head_age") %>%
  lapply(function(x) lm(x ~ keller_trt, data = data))

balance_test_summary2 <- reg2 %>%
  sapply(function(x) coef(summary(x))[c(2,8)]) %>%
  t()

balance_test_summary2 %>%
  kable(col.names = c("Difference in means", "P-value")) %>%
  kable_styling(position = "center", font_size = 11, latex_options = "hold_position")
```

	Difference in means	P-value
baseline_hhsize	-0.0271916	0.5017582
baseline_payments	1.0467155	0.4810651
baseline_elec_use	10.9765373	0.1654133
baseline_hh_head_age	-0.0614275	0.5368687

```
summary(data)
```

```
##    keller_trt     keller_trt_yes   baseline_hhsize  baseline_payments
##  Min.   :0.0000   Min.   :0.0000   Min.   : 1.000   Min.   :  0.02783
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7.000   1st Qu.: 65.99745
##  Median :1.0000   Median :0.0000   Median : 8.000   Median :130.25410
##  Mean   :0.5001   Mean   :0.3517   Mean   : 7.981   Mean   :129.83440
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 9.000   3rd Qu.:193.88830
##  Max.   :1.0000   Max.   :1.0000   Max.   :15.000   Max.   :257.27950
##  baseline_elec_use  baseline_hh_head_age endline_hhsize   endline_payments
##  Min.   :   0.005   Min.   :13.00        Min.   : 1.000   Min.   :     0.7
##  1st Qu.: 115.952   1st Qu.:32.00        1st Qu.: 7.000   1st Qu.: 11454.5
##  Median : 276.661   Median :35.00        Median : 8.000   Median : 31386.7
##  Mean   : 393.838   Mean   :35.04        Mean   : 7.981   Mean   : 46333.1
##  3rd Qu.: 539.275   3rd Qu.:38.00        3rd Qu.: 9.000   3rd Qu.: 64255.0
##  Max.   :4687.777   Max.   :52.00        Max.   :15.000   Max.   :585984.2
##  endline_elec_use  endline_hh_head_age
##  Min.   :   0.00   Min.   :14.00
##  1st Qu.:  90.57   1st Qu.:33.00
##  Median : 249.96   Median :36.00
##  Mean   : 369.62   Mean   :36.04
##  3rd Qu.: 513.10   3rd Qu.:39.00
##  Max.   :4687.78   Max.   :53.00
```

```
#summary(reg2)
```

Null Hypothesis H0: Difference in Means = 0

From the P- values observed for the pre-payment variables baseline_hhsize(household size), baseline_payments(Payments), baseline_elec_use(household electricity usage), baseline_hh_head_age(household head age) using the keller_trt as the treatment variable, we cannot reject the null hypothesis. This says that the randomization is fine and the treatment group and the control group are balanced in pre-treatment payments, electricity usage, household size, and household head age.

Assumption we need to make on unobserved characteristics in order to be able to estimate the causal effect of keller_trt? From the balance tests above, we found that the treatment is assigned randomly and randomization is fine. The distribution of observables and unobservables are same in both the treated and the non-treated groups. Hence with this, we can assume that there is no risk of selection bias in this randomization.

Which means D_i is exogenous, i.e the unobservable, conditioned on treatment is zero.

$E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

Substituting these into the expression for τ^{ATE}, we get

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

Initially we had a question of whether the i and j are significantly different. Here we confirmed the opposite with balance test and made sure that they are not significantly different from each other in terms of characteristics. Hence we can estimate ATE simply by taking difference in means of treatment group and control group.

The estimate for the same can be found by calculating the difference between mean of the treatment group and the mean of control group, as shown below:

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

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?

From the balance tests above, we found that the treatment is assigned randomly and randomization is fine. The distribution of observables and unobservables are same in both the treated and the non-treated groups. Hence with this, we can assume that there is no risk of selection bias in this randomization.

Which means D_i is exogenous, i.e the unobservable, conditioned on treatment is zero.

$E[Y_i(1)|D_i = 1] = E[Y_i(1)]$ $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

Substituting these into the expression for τ^{ATE}, we get

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

Initially we had a question of whether the i and j are significantly different. Here we confirmed the opposite with balance test and made sure that they are not significantly different from each other in terms of characteristics. Hence we can estimate ATE simply by taking difference in means of treatment group and control group.

The estimate for the same can be found by calculating the difference between mean of the treatment group and the mean of control group, as shown below:

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

The same can also be found using the following regression (Yi on Di)

```
endline_payments_rg <- lm(endline_payments ~ keller_trt, data = data)


stargazer(endline_payments_rg,
          type = "latex",
          header = FALSE,
          title = "Regression: Treatment Effect",
          column.labels = c("$Regression keller_trt, endline_payments$"),
          colnames = FALSE,
          model.numbers = FALSE,
          df = FALSE)
```

```
summary(endline_payments_rg)
```

Call: lm(formula = endline_payments ~ keller_trt, data = data)

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 < 2e-16 ***keller_trt -4679.1 985.2 -4.749 2.07e-06*** — 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: 2.071e-06

$\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$
$\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0) = -\ Rs.\ 4,679.066$
This means that the estimated average treatment effect is negative, i.e is a decrease in payments by 4,769.066 Rs. post treatment.

Table 1: Regression: Treatment Effect

	Dependent variable:
	endline_payments
	$Regressionkeller_trt, endline_payments$
keller_trt	−4,679.066***
	(985.230)
Constant	48,673.340***
	(696.767)
Observations	9,999
R^2	0.002
Adjusted R^2	0.002
Residual Std. Error	49,259.050
F Statistic	22.555***
Note:	*p<0.1; **p<0.05; ***p<0.01

```
control_data <- data %>% filter(keller_trt == 0)
summary(control_data$endline_payments)
```

Statistical significance of results: Is your result statistically significant? Is the effect you find large or small, relative to the mean in the control group?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      22   14637   34450   48673   66443  475379
```

```
summary(endline_payments_rg)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt, data = data)
##
## 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  < 2e-16 ***
## keller_trt   -4679.1      985.2  -4.749 2.07e-06 ***
## ---
## 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: 2.071e-06
```

We see a mean of endline payments for control group to be 48673. The mean and the median(34450) for endline payments for control group, and from the the p-value 2.071eˆ-06 , t value of -4.749, we can conclude that the estimated ATE is statistically significant.

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?

It is not a good idea to control for the endline amount of power consumed. As this is a post treatment variable which is endogenous, once conditioned on it will result in affecting our assumption that by random assignment households with disconnection and without disconnections would have similar endline payments.Thus this will result on an estimation bias.

```
endline_power_reg <- lm(endline_payments ~ keller_trt + endline_elec_use, data = data)

summary(endline_power_reg)
```

Conditioning on endline power use

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt + endline_elec_use,
##     data = data)
##
## 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)      1.288e+02  1.280e+00   100.569   <2e-16 ***
## keller_trt       1.147e+00  1.487e+00     0.771    0.441
## endline_elec_use 1.250e+02  1.885e-03 66316.896   <2e-16 ***
## ---
## 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: < 2.2e-16
```

```
summary(endline_payments_rg)
```

```
##
## Call:
## lm(formula = endline_payments ~ keller_trt, data = data)
```

```
##
## 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  < 2e-16 ***
## keller_trt   -4679.1      985.2  -4.749 2.07e-06 ***
## ---
## 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: 2.071e-06
```

```
stargazer(endline_payments_rg, endline_power_reg,
          type = "latex",
          header = FALSE,
          title = "Regression: Treatment Effect",
          column.labels = c("$Regression keller_trt, endline_elec_use, endline_payments$"),
          colnames = FALSE,
          model.numbers = FALSE,
          df = FALSE)
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Regression: Treatment Effect}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{\textit{Dependent variable:}} \\
## \cline{2-3}
## \\[-1.8ex] & \multicolumn{2}{c}{endline\_payments} \\
##  & $Regression keller_trt, endline_elec_use, endline_payments$ &  \\
## \hline \\[-1.8ex]
##  keller\_trt & $-$4,679.066$^{***}$ & 1.147 \\
##   & (985.230) & (1.487) \\
##   & & \\
##  endline\_elec\_use &  & 125.003$^{***}$ \\
##   &  & (0.002) \\
##   & & \\
##  Constant & 48,673.340$^{***}$ & 128.767$^{***}$ \\
##   & (696.767) & (1.280) \\
##   & & \\
## \hline \\[-1.8ex]
## Observations & 9,999 & 9,999 \\
## R$^{2}$ & 0.002 & 1.000 \\
## Adjusted R$^{2}$ & 0.002 & 1.000 \\
## Residual Std. Error & 49,259.050 & 74.267 \\
## F Statistic & 22.555$^{***}$ & 2,203,926,641.000$^{***}$ \\
## \hline
## \hline \\[-1.8ex]
```

```
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

Conditioning on the post-treatment variable endline power usage gives us $\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$ = $Rs.$ 1.147

P-value observed when conditioning with the post-treatment variable is 0.441. From the p value (>0.05) we can say that the difference in means between treatment group and control group is zero. And not statistically significant. Thus we can say that the estimated ATE on payments due to treatment when conditioned on the post-treatment variable endline power usage is zero.This means that the estimated average treatment effect on payments due to the treatment (disconnection) is zero

From Q7 with no conditioning we saw the value of $\hat{\tau}^{ATE}$ = -Rs. 4,679.066 This means that the estimated average treatment effect is negative, i.e is a decrease in payments by Rs. 4,769.066 post treatment.

Controlling on baseline power consumed The variable baseline power consumed does not create a significant difference in any observable feature between treatment and control groups as they are already balanced around this covariate. In simple words, the variable baseline power consumed in a balanced covariate. A balanced covariate can be used to reduce the standard errors.

```
baseline_power_reg <- lm(endline_payments ~ keller_trt + baseline_elec_use, data = data)


summary(endline_payments_rg)
```

Call: lm(formula = endline_payments ~ keller_trt, data = data)

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 < 2e-16 ***keller_trt -4679.1 985.2 -4.749 2.07e-06*** — 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: 2.071e-06

```
summary(endline_power_reg)
```

Call: lm(formula = endline_payments ~ keller_trt + endline_elec_use, data = data)

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) 1.288e+02 1.280e+00 100.569 <2e-16 ***keller_trt 1.147e+00 1.487e+00 0.771 0.441 endline_elec_use 1.250e+02 1.885e-03 66316.896 <2e-16*** — 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: < 2.2e-16

```
summary(baseline_power_reg)
```

Call: lm(formula = endline_payments ~ keller_trt + baseline_elec_use, data = data)

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) 4.109e+02 5.214e+01 7.881 3.59e-15 ***keller_trt -6.043e+03 6.058e+01 -99.763 < 2e-16***
baseline_elec_use 1.243e+02 7.656e-02 1623.290 < 2e-16 *** — 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: < 2.2e-16

```
stargazer(endline_payments_rg, baseline_power_reg, endline_power_reg,
          type = "latex",
          header = FALSE,
          title = "Regression : Treatment Effect",
          column.labels = c("$Regression keller_trt, baseline_elec_use, endline_payments$"),
          colnames = FALSE,
          model.numbers = FALSE,
          df = FALSE)
```

Table 2: Regression : Treatment Effect

	Dependent variable:		
	endline_payments		
	$Regression keller_t rt, baseline_e lec_u se, endline_p ayments$		
keller_trt	−4,679.066***	−6,043.187***	1.147
	(985.230)	(60.575)	(1.487)
baseline_elec_use		124.276***	
		(0.077)	
endline_elec_use			125.003***
			(0.002)
Constant	48,673.340***	410.932***	128.767***
	(696.767)	(52.142)	(1.280)
Observations	9,999	9,999	9,999
R^2	0.002	0.996	1.000
Adjusted R^2	0.002	0.996	1.000
Residual Std. Error	49,259.050	3,028.325	74.267
F Statistic	22.555***	1,320,520.000***	2,203,926,641.000***
Note:			*p<0.1; **p<0.05; ***p<0.01

When conditioning on pre-treatment variable baseline amount of power consumed $\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0) = -\ Rs.\ 6,043$
This means that the estimated ATE is a decrease in the post-treatment payments by $6,043 with a standard error of Rs. 60.58

From Question 7, no conditioning $\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0) = -\ Rs.\ 4,679.066$
This means that the estimated ATE is a decrease in post treatment payments by $4,769.066 with a standard error of Rs. 985.2

Standard error when no conditioning is Rs. 985.2 Standard error when conditioning on pre-treatment variable baseline amount of power consumed is Rs. 60.58

Thus, including the pre-treatment balanced covariate baseline amount of power consumed has reduced the standard error.

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?

Under the new information i.e actual treatment indicator is keller_trt_yes and not keller_trt, this means that the there is non comliance in the assigned treatment group. This means that the ATE calculated is of treatment assignment but not of actually treated and thus called as the "Intent to Treat" estimate.

In question 7, we thought we are estimating ATE
$\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$ $\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$
$D_i = 1$ signifies treatment i.e disconnecting household's electricity
$D_i = 0$ signifies no-treatment i.e not disconnecting household's electricity

But we are actually estimating $\tau^{experiment} = E[Y_i(R_i = 1)] - E[Y_i(R_i = 0)]$
$\hat{\tau}^{experiment} = \bar{Y}(R_i = 1) - \bar{Y}(R_i = 0)$
$R_i = 1$ signifies a household is assigned treatment i.e., selected for disconnecting household's electricity
$R_i = 0$ signifies a household not assigned treatment i.e not selected for disconnecting electricity

For both the above estimates to be same as we thought earlier, all households should comply to the assignment. I.e households selected for treatment undergoes treatment and households selected for non-treatment doesnt undergo treatment. Incase of non compliance both the above estimates will not be same and the calculated estimate in Question 7 would then become Intent to Treat estimate.

With the new information, we can say that $R_i $ is assignment/selected for treatment indicator, which we get from keller_trt and D_i is the actually treated indicator which we get from keller_trt_yes

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)?

Using keller_trt, we can estimate the average treatment effect on assigned to treatment units and using keller_trt_yes, we can find the average effect on actually treated. We thought we are estimating $\hat{\tau}^{ATE}$ in Question 7 but we are actually calculating $\hat{\tau}^{experiment}$
This is called the intent to treat (ITT) estimate

Using keller_trt $\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$ $\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$
$D_i = 1$ signifies treatment i.e disconnecting household's electricity
$D_i = 0$ signifies no-treatment i.e not disconnecting household's electricity

Using keller_trt_yes $\tau^{experiment} = E[Y_i(R_i = 1)] - E[Y_i(R_i = 0)]$
$\hat{\tau}^{experiment} = \bar{Y}(R_i = 1) - \bar{Y}(R_i = 0)$
$R_i = 1$ signifies a household is assigned treatment i.e., selected for disconnecting household's electricity
$R_i = 0$ signifies a household not assigned treatment i.e not selected for disconnecting electricity

ITT estimate can be found using regression

```
assigned_only_reg <- lm(endline_payments ~ keller_trt, data = data)

summary(assigned_only_reg)
```

Call: lm(formula = endline_payments ~ keller_trt, data = data)

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 < 2e-16 ***keller_trt -4679.1 985.2 -4.749 2.07e-06*** — 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: 2.071e-06

```
stargazer(assigned_only_reg,
          type = "latex",
          header = FALSE,
          title = "Regression: Treatment Effect",
          column.labels = c("$Regression: keller_trt, endline_payments$"),
          colnames = FALSE,
          model.numbers = FALSE,
          df = FALSE)
```

Table 3: Regression: Treatment Effect

	Dependent variable:
	endline_payments
	Regression : $keller_trt, endline_payments$
keller_trt	−4,679.066***
	(985.230)
Constant	48,673.340***
	(696.767)
Observations	9,999
R^2	0.002
Adjusted R^2	0.002
Residual Std. Error	49,259.050
F Statistic	22.555***
Note:	*p<0.1; **p<0.05; ***p<0.01

```
actually_treated_reg <- lm(endline_payments ~ keller_trt_yes, data = data)

summary(actually_treated_reg)
```

Call: lm(formula = endline_payments ~ keller_trt_yes, data = data)

Residuals: Min 1Q Median 3Q Max -48715 -34847 -15044 17893 537247

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 48737.2 611.2 79.743 < 2e-16 ***keller_trt_yes -6835.0 1030.5 -6.633 3.47e-11*** — Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49210 on 9997 degrees of freedom Multiple R-squared: 0.004381, Adjusted R-squared: 0.004281 F-statistic: 43.99 on 1 and 9997 DF, p-value: 3.471e-11

```
stargazer(actually_treated_reg,
          type = "latex",
          header = FALSE,
          title = "Regression: Treatment Effect",
          column.labels = c("$Regression keller_trt_yes, endline_payments$"),
          colnames = FALSE,
          model.numbers = FALSE,
          df = FALSE)
```

Table 4: Regression: Treatment Effect

	Dependent variable:
	endline_payments
	$Regression keller_t rt_y es, endline_p ayments$
keller_trt_yes	−6,834.980***
	(1,030.527)
Constant	48,737.210***
	(611.178)
Observations	9,999
R^2	0.004
Adjusted R^2	0.004
Residual Std. Error	49,206.440
F Statistic	43.990***
Note:	*p<0.1; **p<0.05; ***p<0.01

To see how many of the assigned treatment actually got treated, take ratio

```
ratio_assigned_actual = sum(data$keller_trt_yes) / sum(data$keller_trt)

print(ratio_assigned_actual)
```

```
## [1] 0.7032593
```

The above ratio says that 70.3%($Pr(D_i = 1|R_i = 1) = 0.703$) of the households assigned treatment actually got treated. There exists a non compliance.

We also know that households in the control group were not treated. But checking again

```
data %>% filter(keller_trt == 0 & keller_trt_yes == 1) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

i.e $Pr(D_i = 1|R_i = 0) = 0$, Hence τ^T is same as τ^{LATE}.

As Keller is only interested to know about the actual effects of their disconnections, we can calculate the treatment effect on actually treated by

$\tau^T = \frac{\hat{\tau}^{experiment}}{Pr(D_i=1|R_i=1)}$

τ^T = -4679.1/0.7032593 = - Rs. 6,653.45

This means that the estimated average effect on actually treated i.e τ^T is endline payments decreased by Rs. 6,653.45

How does this compare to what you estimated in (7)? $\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$
$\hat{\tau}^{ATE} = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$
$= -\ Rs.\ 4,679.066$
This means that the estimated average Intent to treatment effect τ^{ITT} is post-treatment payments decreased by Rs. 4,769.066

Comparing both the estimated average ITT and estimated average effect on actually treated, we see that the effect on post treatment payments (or) endline_payments is greater in the actual treated vs the effect on post treatment paymets (or) endline payments in the Intent to Treatment.