Program Evaluation Problem Set 3

Hye-Min Jung

5/21/2020

1.

- **The most ideal experiment would be to run randomized controlled trial, to answer the causal effect of hours of outages on kW amount of PV installed at the household level, because under randomization, we can easily estimate the impact of treatment, ATE simply from the difference in means between treated and control households.**
- Define: $D_i \in \{0, 1\}$ as the blakcout indicator for household i
  - when household i is treated, $D_i = 1$
  - when household i is not treated, $D_i = 0$
- Define: $Y_i(D_i)$ as the kW of solar PV installed as a function of D_i
  - when household i is treated, we observe kW of solar PV installed $Y(1)$
  - when household i is not treated, we observe kW of solar PV installed $Y(0)$
- Then, the impact of blackouts on kW of solar PV adopted by housdeholds is: $\tau_i = Y_i(1) - Y_i(0)$
- To put it simply in words, impact of treatment is a parameter that measures 'What is the outcome of power shut off, compared to the outcomes that we would have observed without the power shut off?'
  - Impacts is 'changes in outcomes caused by the policy' whereas, outcomes is 'things that we could "potentially" observe'.
  - We need to consider all possible outcomes we could have observed, spans both actual and alternative programs, to explain the impact of treatment.
- **However, since shutting off electricity for randomly selected households would cause lot of problems, a state won't be able to conduct RCT. Therefore, we would like to turn to alternative method, selection on observable design.**
  - For this design, we will assume that, conditional on observables and treatment assignment is independent of potential outcomes.
  - In other words, $(Y_i(1), Y_i(0)) \perp D_i | X_i$.
  - This means, once we control for covariates, we've eliminated selection and treatment is as good as random.
  - Also, we will have to assume common support, $0 < \Pr(D_i = 1 | X_i = x) < 1$.
  - This means, the probability that $D_i = 1$ for all levels of x_i is between 0 and 1.
  - There are both treated and untreated units for each level of X.
- And try to estimate the impact of treatment, ATE: $\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$.
  - However, all we can acutally see is $E[Y_i(1)|D_i = 1]$ and $E[Y_i(0)|D_i = 0]$.
  - But under conditional independence and common support, we can get from τ^{SOO} to τ^{ATE}
- To use selection-on-observables approach, we would need dataset at the household level, treatment status, pre-treatment observable chracteristics, post-treatment characteristics, post-treatment outcome.

2.

- Since we cannot observe everything we need, I would try to observe some (quasi) random variation in D_i, and turn to natural experiments research design instead.
  - In order to recover the causal effect of D_i, we should have nothing in the error term that is unobserved piece that is correlated with treatment or selection into treatment.
  - It is important to make sure that we have modeled everything in the model, so there is nothing left in the error term.
  - In other words, $E[\varepsilon_i|D_i] = 0 \Longleftrightarrow \text{Cov}(D_i, \varepsilon_i) = 0$

3.

- We could use "distance from the electric utilities" as an instrumental variable to evaluate the effect of power outage hours on kW of solar PV installed.
  - Proposed instrument should be a good one, because this affects outcome only through treatment. In other words, instrument cannot affect outcome through any other channel, and therefore can be "excluded" from a regression of Y_i on D_i.
- However, I have some concerns about the abiilty to estimate the treatment effect using propsed instrument.
  - Because IV is throwing out variation, the standard errors will be bigger than OLS standard erros.
  - Also, the exclusion restiction is fundamentally untestable so I will need to bear strong assumption to believe that proposed IV works properly.

4.

- Since California utilities randomly cut power for different lengths of time to different households in the small pilot program, at least for this small pilot program, treatment should being as good as random.
  - So we can believe that an additional hour of electricity outage has same effect on KW of PV adoption and we can estimate this by linear regression model.
  - Thanks to random treatment assignment, we can estimate the impacts as: $\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$.
  - Using linear regression model $Y_i = \alpha + \tau D_i + \varepsilon_i$ get $\hat{\tau}$ and it's same as $\hat{\tau}^{ATE}$.

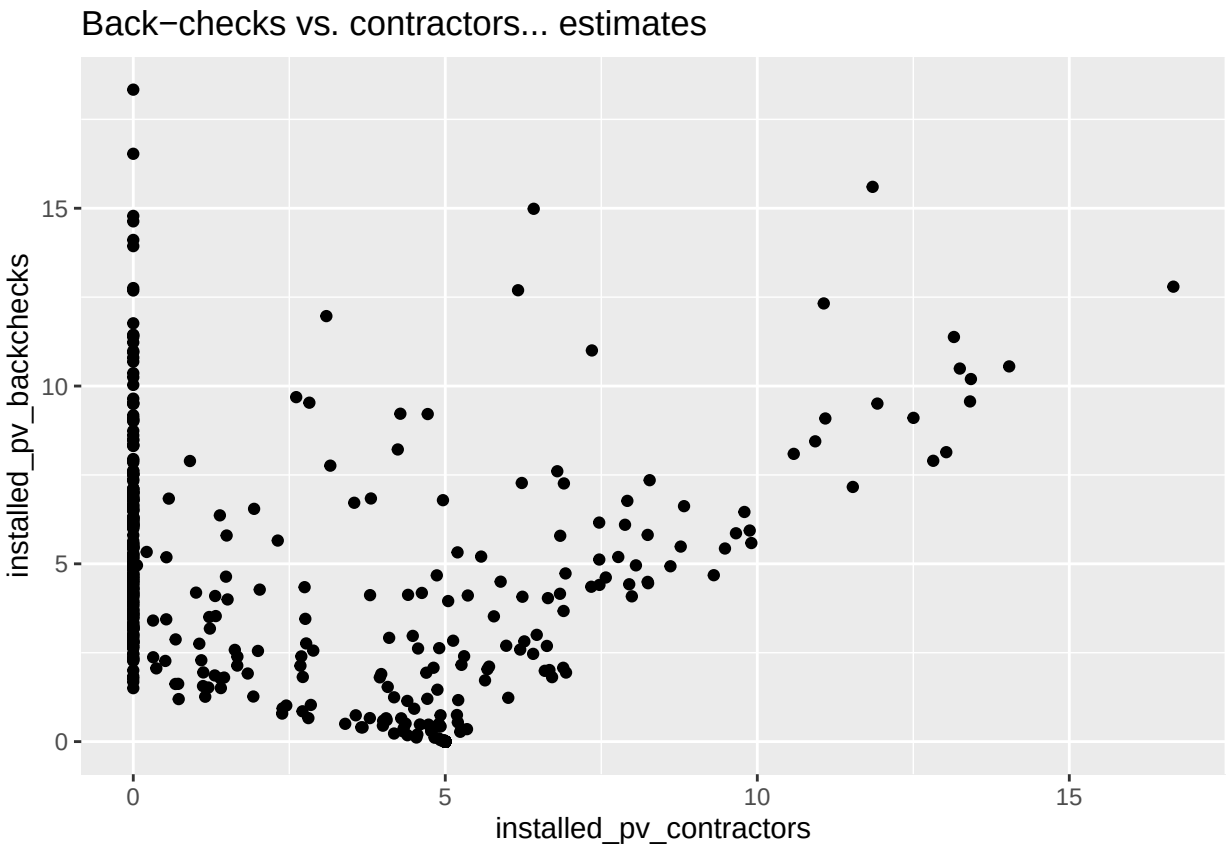
5.

- A unit increase in utility outage hours increases installed_pv_contractors by 0.0001115.
  - Please note that $\hat{\tau}$ is no statistically significant. However, I still report the coefficient here.
  - But it is still good practice to suspect any laten problems such as noise or small sample size.

```
summary(lm5 <- lm(installed_pv_contractors ~ utility_outage_hours, data = cal_raw))
```

```
##
## Call:
## lm(formula = installed_pv_contractors ~ utility_outage_hours,
##     data = cal_raw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.681 -3.525  1.430  1.456 13.113
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.5435687  0.0644675  54.967   <2e-16 ***
## utility_outage_hours 0.0001115  0.0002976   0.374    0.708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.7 on 1998 degrees of freedom
## Multiple R-squared:  7.019e-05,  Adjusted R-squared:  -0.0004303
## F-statistic: 0.1402 on 1 and 1998 DF,  p-value: 0.7081
```

6.



- 203 observations which were 0 for installed_pv_contractors were found to be positive values when backchecked.
  - Assuming backcheck data has no measurement error, installed_pv_contractor is likely to be cause a problem for the analysis.
  - This is measurement error is more like non-classical measurement error in Di, which is likely to cause attenuation bias.
  - Because, a second report as an instrument does not provide consistent estimates of beta under non-classical measurement error. Specifically, if the mismeasurement error covaries with D, in the denominator, Z and the mismeasurement will covary through D.
- Estimate is smaller using backcheck data.
  - A unit increase in utility outage hours increases installed_pv_backchecks by 6.902e-05.
  - However, still $\hat{\tau}$ is no statistically significant.
  - Estimate is smaller than the previous equation, because the measurement error correlated with treatment, resulted in OVB.

```
summary(lm5 <- lm(installed_pv_contractors ~ utility_outage_hours, data = cal_raw))
```

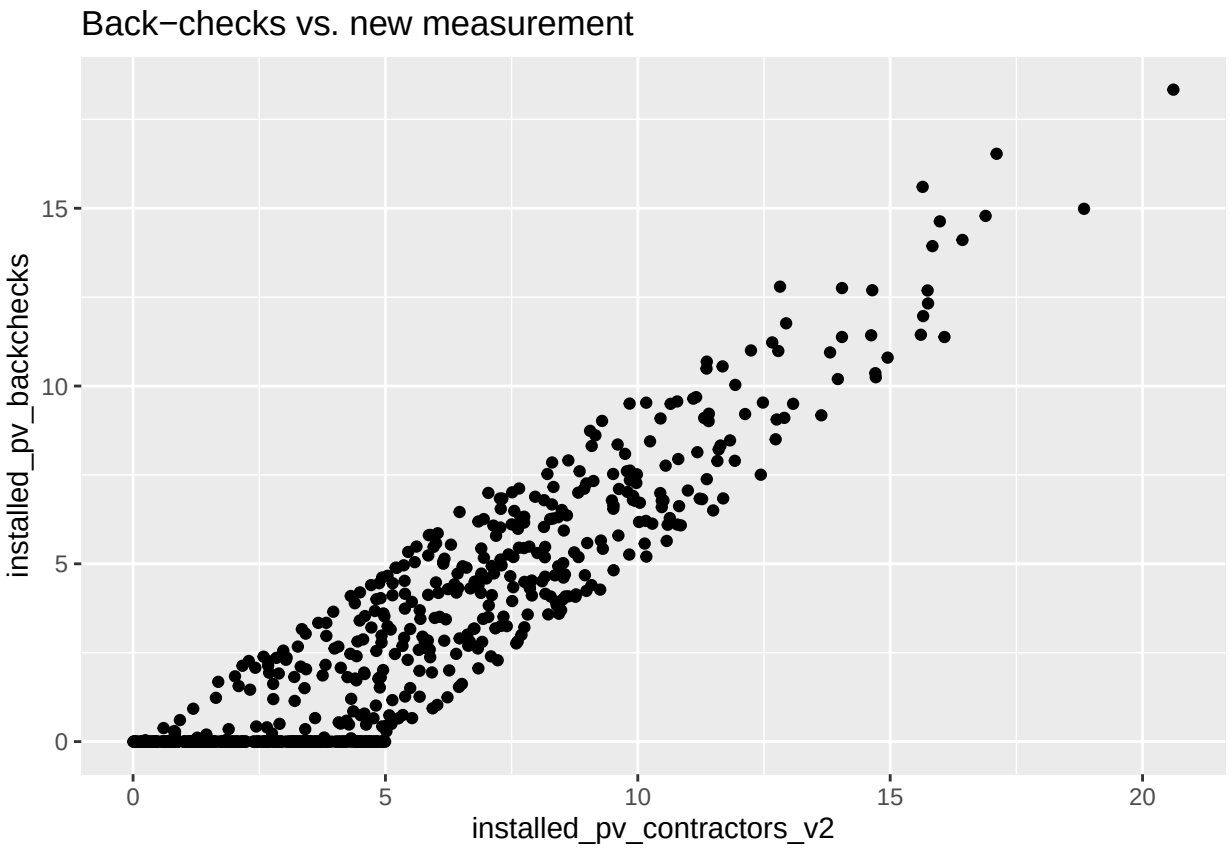
```
##
## Call:
## lm(formula = installed_pv_contractors ~ utility_outage_hours,
##     data = cal_raw)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.681 -3.525  1.430  1.456 13.113
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.5435687  0.0644675  54.967   <2e-16 ***
## utility_outage_hours 0.0001115  0.0002976   0.374    0.708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.7 on 1998 degrees of freedom
## Multiple R-squared:  7.019e-05,  Adjusted R-squared:  -0.0004303
## F-statistic: 0.1402 on 1 and 1998 DF,  p-value: 0.7081
```

```
summary(lm6 <- lm(installed_pv_backchecks ~ utility_outage_hours, data = cal_raw))
```

```
##
## Call:
## lm(formula = installed_pv_backchecks ~ utility_outage_hours,
##     data = cal_raw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.915 -2.849 -1.294  2.080 15.492
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.848e+00  1.383e-01   20.59   <2e-16 ***
## utility_outage_hours 6.902e-05  6.300e-04    0.11    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.47 on 683 degrees of freedom
##   (1315 observations deleted due to missingness)
## Multiple R-squared:  1.757e-05,  Adjusted R-squared:  -0.001447
## F-statistic: 0.012 on 1 and 683 DF,  p-value: 0.9128
```

7.



- Installed_pv_contracts_v2 is same or bigger than installed_pv_backchecks by constant differences.
  - Measurement error in this contractor reports is not likely to cause a problem, becuase measurement error in Y is fine.
  - Because measurement error in Y allow us this assumption: $\text{Cov}(\gamma_i, \varepsilon_i) = 0$ and $\text{Cov}(\gamma_i, D_i) = 0$
  - With the assumption stated above, even if we don't observe Y_i, but rather observe $\tilde{Y}_i = Y_i + \gamma_i$, we can recover true effect by estimating $\hat{\tau}$.
- Estimate is almost same(slightly small) using new measurement.
  - A unit increase in utility outage hours increases installed_pv_contractors_v2 by 6.843e-05.
  - However, still $\hat{\tau}$ is no statistically significant.
  - Estimate is almost same as using backcheck because like we are just moving observation little upward, this does should not change the real relationship.

```
summary(lm6 <- lm(installed_pv_backchecks ~ utility_outage_hours, data = cal_raw))
```

```
##
## Call:
## lm(formula = installed_pv_backchecks ~ utility_outage_hours,
##     data = cal_raw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -2.915 -2.849 -1.294  2.080 15.492
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.848e+00  1.383e-01   20.59   <2e-16 ***
## utility_outage_hours 6.902e-05  6.300e-04    0.11    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.47 on 683 degrees of freedom
##   (1315 observations deleted due to missingness)
## Multiple R-squared:  1.757e-05,  Adjusted R-squared:  -0.001447
## F-statistic: 0.012 on 1 and 683 DF,  p-value: 0.9128
```

```
summary(lm7 <- lm(installed_pv_contractors_v2 ~ utility_outage_hours, data = cal_raw))
```

```
##
## Call:
## lm(formula = installed_pv_contractors_v2 ~ utility_outage_hours,
##     data = cal_raw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1659 -2.5628 -0.7455  2.1454 15.4603
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.155e+00  8.511e-02  60.570   <2e-16 ***
## utility_outage_hours 6.843e-05  3.929e-04   0.174    0.862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.564 on 1998 degrees of freedom
## Multiple R-squared:  1.518e-05,  Adjusted R-squared:  -0.0004853
## F-statistic: 0.03033 on 1 and 1998 DF,  p-value: 0.8618
```

8.

- iou == 1, the measurement error is going to be a problem for the analysis. Because the measurement error is more severe for selected treatment units, in other words, the measurement error is not random and correlated with treatment status.
  - We want to estimate $Y_i = \alpha + \tau D_i + \varepsilon_i$.
  - We don't observe D_i, but rather $\tilde{D}_i = D_i + \gamma_i$
  - $\hat{\tau} = \tau \left(\frac{\text{Var}(D_i)}{\text{Var}(D_i)+\text{Var}(\gamma_i)} \right)$
  - We get attenuation biase.
- iou == 2, the measurement error is not going to be a problem for the analysis. Because the measurement error is random.
  - We don't observe Y_i, but rather $\tilde{Y}_i = Y_i + \gamma_i$
  - If we run, $\tilde{Y}_i = \alpha + \tau D_i + \varepsilon_i$, we get $\hat{\tau} = \tau$
  - With assumption: $\text{Cov}(\gamma_i, \varepsilon_i) = 0$ and $\text{Cov}(\gamma_i, D_i) = 0$
- iou == 1,
  - A unit increase in utility outage hours decreases installed_pv_contractors_v2 by 6.526e-05.
  - Coefficent is not statistically significant for utility_outage_hours.
- iou == 2,
  - A unit increase in utility outage hours increases installed_pv_contractors_v2 by 0.002736
  - Coefficent is not statistically significant for utility_outage_hours.

```
summary(lm8_iou_1 <- lm(installed_pv_contractors_v2 ~ utility_outage_hours, data = iou_1))
```

```
##
## Call:
## lm(formula = installed_pv_contractors_v2 ~ utility_outage_hours,
##     data = iou_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2765 -2.6283 -0.7435  2.1872 15.3433
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.264e+00  1.167e-01  45.099   <2e-16 ***
## utility_outage_hours -6.526e-05  4.049e-04  -0.161    0.872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.575 on 997 degrees of freedom
## Multiple R-squared:  2.606e-05,  Adjusted R-squared:  -0.0009769
## F-statistic: 0.02598 on 1 and 997 DF,  p-value: 0.872
```

```
summary(lm8_iou_2 <- lm(installed_pv_contractors_v2 ~ utility_outage_hours, data = iou_2))
```

```
##
## Call:
## lm(formula = installed_pv_contractors_v2 ~ utility_outage_hours,
##     data = iou_2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4784 -2.5841 -0.7658  2.0964 13.3147
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.840101   0.178586  27.102   <2e-16 ***
## utility_outage_hours 0.002736   0.001717   1.593    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.55 on 999 degrees of freedom
## Multiple R-squared:  0.002535,   Adjusted R-squared:  0.001537
## F-statistic: 2.539 on 1 and 999 DF,  p-value: 0.1114
```

9.

- I would use newly introduced data, survey_outage_hour to decompose into the true outage hours and an error term.
  - In order for this to work, (1) treatment should be as good as random and (2) measurement error is not in our original error term.

```
lm9 <- lm(utility_outage_hours ~ survey_outage_hours, data = cal_raw)
summary(lm(installed_pv_contractors_v2 ~ lm9$fitted.values, data = cal_raw))
```

```
##
## Call:
## lm(formula = installed_pv_contractors_v2 ~ lm9$fitted.values,
##     data = cal_raw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6938 -2.5909 -0.7353  2.1708 15.5033
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.888163   0.127903  38.218  < 2e-16 ***
## lm9$fitted.values 0.003579   0.001318   2.715  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.558 on 1998 degrees of freedom
## Multiple R-squared:  0.003676,   Adjusted R-squared:  0.003178
## F-statistic: 7.373 on 1 and 1998 DF,  p-value: 0.006679
```

- Compare to the estimate in (8), the effect is bigger and statistically significant.
  - A unit increase in utility outage hours increases installed_pv_contractors_v2 by 0.003579
  - And now our coefficent is statistically significant.
- This estimate I've just recovered from #9 would be the estimate that I would like to send to CAL-BEARS as final analysis result.