Answers to PS3 - Program Evaluation PPHA 34600

Diego Diaz

21-05-2020

Other group members: Piyush Tank, Matthew Mauer.

1 - An ideal experiment for this question requires an RCT experiment since we require random assignment of the treatment. Ideally we would like to create power outages on peoples' homes and make sure they are selected randomly across the population. In order to answer what is the effect of hours of electricity outages we would also have to choose the duration of the power outage randomly, ranging on the interval that we want to study. It is likely that the effect is non linear, for example, a household that is assigned a very long power outages (e.g. a year), will likely install solar PV to supply all his energy requirement. It is also important to consider the reason why households decide to invest in solar, and although there might be many, one of them is the unreliability of the power grid, and that is likely the factor they are considering when investing in solar PV after experiencing outages. This means that what households actually care about when deciding whether to invest in solar is their expectation of future power outages. With this consideration in mind, it is likely that only unexpected power outages and power outages that are related to problems on the grid will have an impact.

The dataset that the experiment would produce is a variable indicating the number of hours of power outage each household experienced during a given period of time, ranging from 0 to some arbitrary upper bound (let's call it T). The other variable would be their investment in solar PV capacity measured in kWs. Using the potentian outcomes framework we would then compare groups of households that experienced the same amount of power outages and that would give us the effect of the difference between the number of hours of outages in solar PV investment.

The impact of one hour of outage in solar PV investment for a household i is:

$$\tau_i = Y_i(h+1) - Y_i(h)$$

Where h is the number of hours of outage that household i would have experienced if he or she hadn't been treated.

Although we can't observe τ_i because we can't observe a household in both states, With our RCT we can estimate the Average Treatment Effect (ATE). This is:

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

Given that we randomly assign h to households, differences between average instalation of solar PV would give us an unbiased estimate of the (ATE).

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

Where $\overline{Y(h+1)}$ and $\overline{Y(h)}$ are simply the means between groups with $h+1$ and h hours of power outages respectively.

2 - Since we have a problem of endogeneity in the treatment variable, that is, number of hours of power outages, we can adopt an *instrumental variables* (IV) approach in order to isolate the variation in the treatment that is not correlated with the error term.

We need a variable that is correlated with the treatment assignment but that does not directly impact the outcome.

Without endogeneity concerns, we would estimate the impact of the treatment by OLS with the following regression model:

$$Y_i = \alpha + \tau h_i + \epsilon_i$$

Where h_i is the number of hours of outages and ϵ_i is the error for household i.

The problem is that $Cov(h_i, \epsilon_i) \neq 0$

We can get around this issue with instrumental variables by separating h_i as:

$$h_i = B_i \epsilon_i + C_i$$

With $Cov(h_i, \epsilon_i) = 0$

Then we can write:

$$Y_i = \alpha + \tau(B_i \epsilon_i + C_i) + \epsilon_i$$

We can't observe C_i, but we can use an instrument Z_i that is correlated with C_i and uncorrelated with the error term. By regressing the instrument on our treatment variable we isolate the variation in h that is not correlated with the error term. We are using the following regression model:

$$h_i = \alpha + \gamma Z_i + \eta_i$$

We are assuming that Z_i is uncorrelated with the error term, what is called the exclusion assumption. We also need the instrument to be correlated with the treatment, and we can test it with an F test from the previous regression.

3 -

We need a variable that is correlated with power outages. Since thunderstorms and extreme weather events like hurricanes and tornadoes are the most common cause of outages, we can use the presence of these events as an instrument for hours of power outages. We need to check that it is correlated with outages, and we can reason that it is unlikely that it directly affects PV adoption since people should not have an incentive to invest or not invest in solar power depending on the frequency of thunderstorms. It is likely that the frequency of cloudy days affects solar power PV adoption but it is also likely that it is unrelated to thunderstorms. IN case it turns out to be correlated with such frequency, we would have to control for the frequency of cloudy/sunny days and we would get the pure effect of the instrument in the regression.

We can tell the instrument is a good one by estimating the regression model from (2): D_i = alpha + gamma Z_i + eta_i , performing an F-test, and arguing that it does not effect the outcome as we did on the previous paragraph. Since the assumptions are likely to be fulfilled, we should be able to estimate the effect as long as we have data on the instrument.

4 - This is the best possible scenario since our ideal experiment from (1) was performed, we can now estimate the impact of outages on solar PV by following the same approach described (estimating $\hat{\tau}^{ATE} = \overline{Y(h+1)} - \overline{Y(h)}$) or by making a linear regression. This will be our approach since it will give more useful information for the policy maker. We will perform the regression mentioned in (2):

$$Y_i = \alpha + \tau h_i + \epsilon_i$$

Performing OLS will give us the coefficient that we are interested in τ, which can be interpreted as the effect in kW of installed capacity from a variation of 1 in the hours of power outage.

5 - We estimate the regression model from (4) using R with the giving data, the code an output are shown next:

```
ps3_data <- read.csv("C:/Google Drive/Program Eval/Assignments/data/ps3_data.csv")
linear = lm(installed_pv_contractors ~ utility_outage_hours, ps3_data)
summary(linear)
```

Call: lm(formula = installed_pv_contractors ~ utility_outage_hours, data = ps3_data)

Residuals: Min 1Q Median 3Q Max -3.681 -3.525 1.430 1.456 13.113

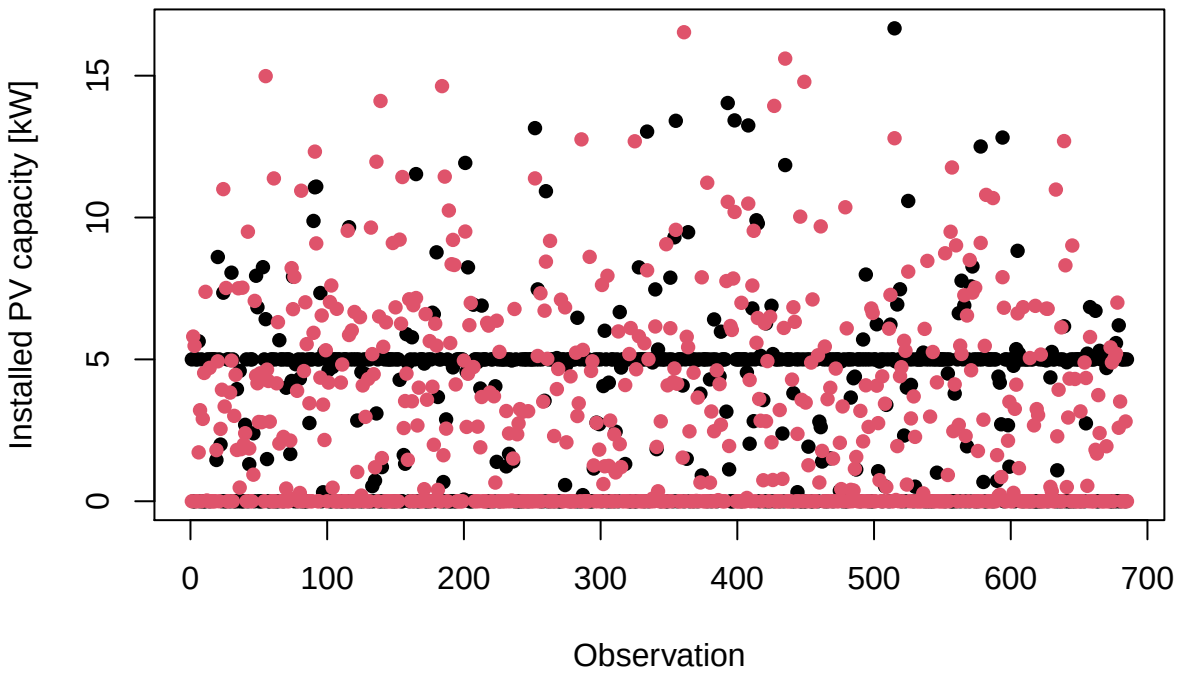
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5435687 0.0644675 54.967 <2e-16 *** utility_outage_hours 0.0001115 0.0002976 0.374 0.708
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

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

As we can see the coefficient on utility outage hours is $1.11 \cdot 10^{-4}$ and is not statistically significant.

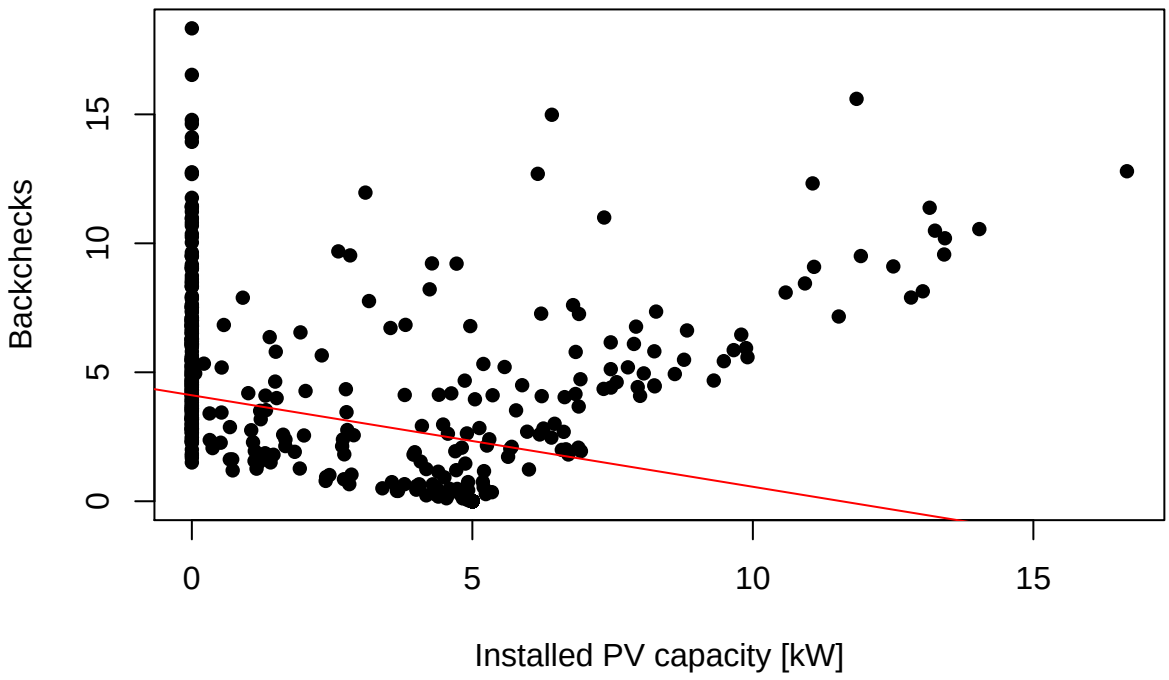
6 - Next we show a scatter plot of the contractors estimates (black) and the backchecks (red) on the same figure:

```
plot(ps3_data$installed_pv_contractors[! is.na(ps3_data$pv_adoption_backchecks)], pch = 16, ylab = "Ins
points(ps3_data$pv_adoption_backchecks[! is.na(ps3_data$pv_adoption_backchecks)], col=2, pch = 16)
```



There is an strange high concentration of observations arouns 5 kW for the contractor estimates. To be more certain about problems with the data, we show a different graph by putting the contractor estimates in the x-axis and the backchecks on the y-axis along with a regression line.

```
df = na.omit(ps3_data)
plot(df$installed_pv_contractors, df$pv_adoption_backchecks, pch = 16, xlab = "Installed PV capacity [k
abline(lm(df$pv_adoption_backchecks ~ df$installed_pv_contractors), col="red")
```



There seems to be underestimation of the real value the higher the measurement, which indicates the presence of non-classical errors. This will cause a downward bias in the estimate for the impact of hours of power outage on adoption of PV since the true value of the adoption tends to be higher for lower measurements. In practice this means that households that experienced non to few hours of outages actually experienced more relatively to the error to those that experienced longer hours of outages.

Next we show the estimates for both cases, first for contractor estimates and second for backchecks. We expect the second to be lower because of the downward bias.

```
linear = lm(installed_pv_contractors ~ utility_outage_hours, ps3_data)
summary(linear)
```

Call: lm(formula = installed_pv_contractors ~ utility_outage_hours, data = ps3_data)

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

```
linear = lm(pv_adoption_backchecks ~ utility_outage_hours, ps3_data)
summary(linear)
```

Call: lm(formula = pv_adoption_backchecks ~ utility_outage_hours, data = ps3_data)

Residuals: Min 1Q Median 3Q Max -2.915 -2.849 -1.294 2.080 15.492

Coefficients: Estimate Std. Error t value Pr(>|t|)

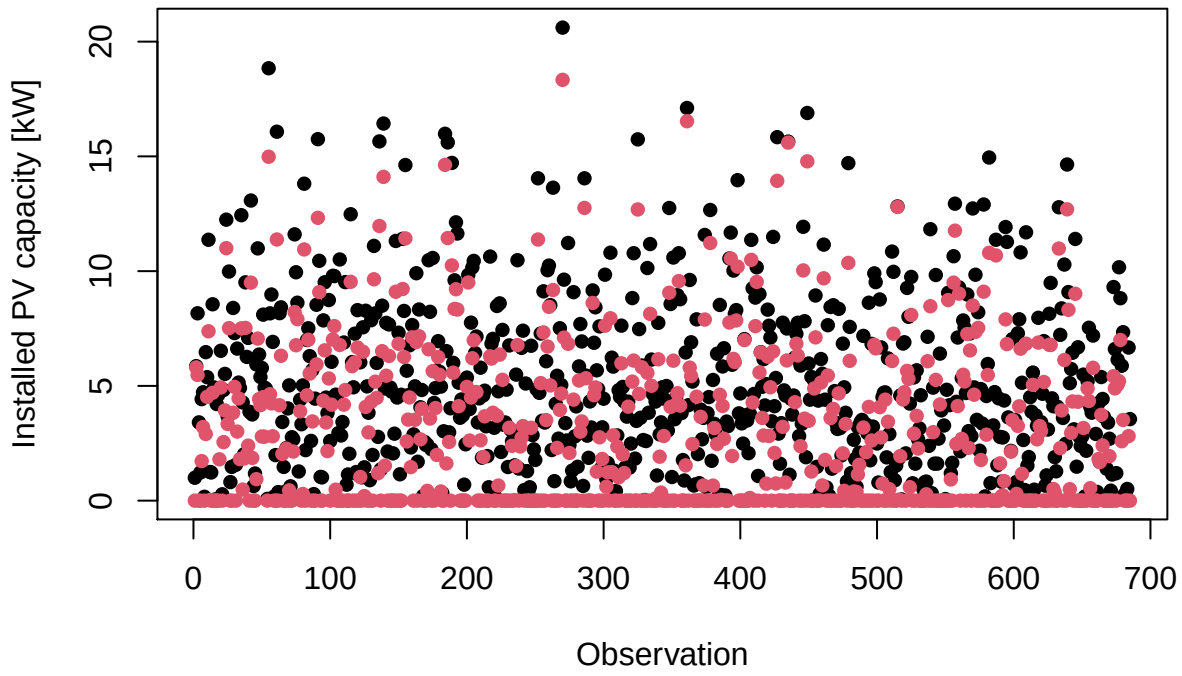
(Intercept) 2.848e+00 1.383e-01 20.59 <2e-16 *** utility_outage_hours 6.902e-05 6.300e-04 0.11 0.913
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

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

Although we can see that we have a decrease in the coefficient (as expected) from $1.11 \cdot 10^{-4}$ to $6.90 \cdot 10^{-5}$, we fail to reject the null hypothesis in both cases since the coefficient is not statistically significant.

7 - First we recreate the figures from (6) with the new variable for the contractor estimates.

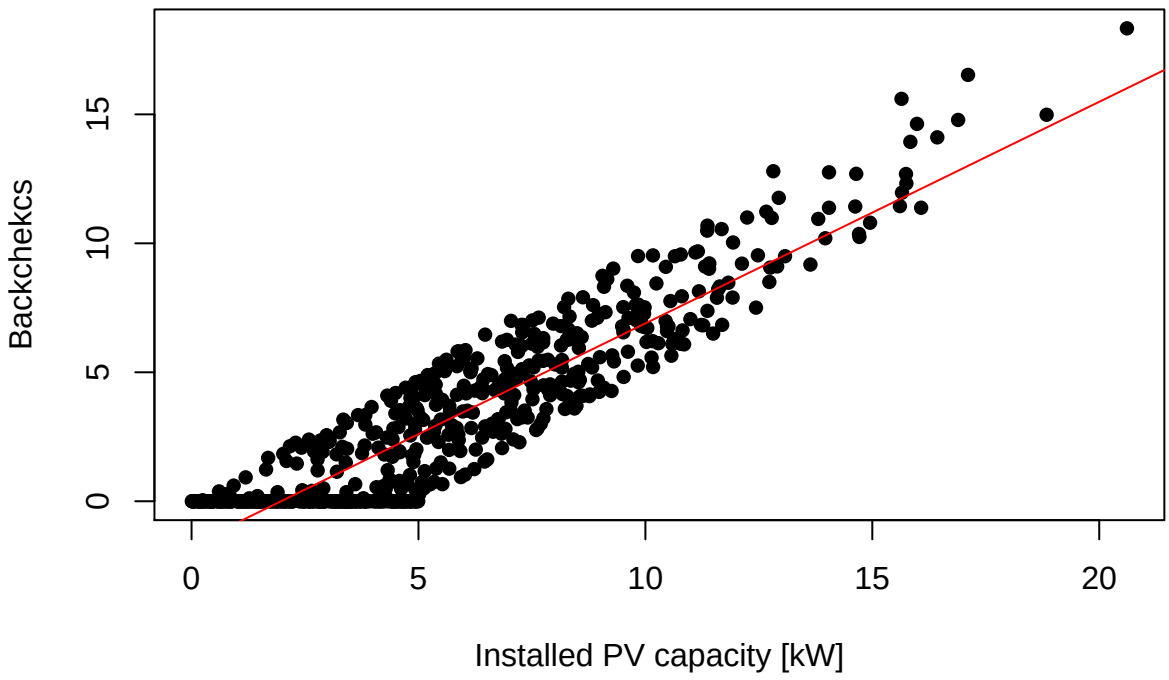
```
plot(ps3_data$installed_pv_contractors_v2[! is.na(ps3_data$pv_adoption_backchecks)], pch = 16, ylab = "
points(ps3_data$pv_adoption_backchecks[! is.na(ps3_data$pv_adoption_backchecks)], col=2, pch = 16)
```



The data for the contractor estimates seems cleaner since it doesn't have a higher concentration of observations at any given point, although it should have a higher concentration at 0 since it is likely more likely not to experience outages in California.

Next we put contractor estimates v2 in the x-axis and the backchecks on the y-axis along with a regression line.

```
df = na.omit(ps3_data)
plot(df$installed_pv_contractors_v2, df$pv_adoption_backchecks, pch = 16, xlab = "Installed PV capacity
abline(lm(df$pv_adoption_backchecks ~ df$installed_pv_contractors_v2), col="red")
```



In this case we clearly have classical errors in the sample, which will not cause any bias in the regression coefficients. However, it is still likely to be a problem since it will decrease the power of our statistical tests, potentially leading us to a type 2 error. This happens because the standard error of the coefficient increases.

Next we estimate the coefficient using the backchecks data and the new contractor estimates. We can expect the results to be similar (on average they will be) since estimating with classical errors does not bias our estimate, and we can expect the standard error to be lower with the true values (backchecks).

```
linear = lm(pv_adoption_backchecks ~ utility_outage_hours, ps3_data)
summary(linear)
```

Call: lm(formula = pv_adoption_backchecks ~ utility_outage_hours, data = ps3_data)

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

```
linear = lm(installed_pv_contractors_v2 ~ utility_outage_hours, ps3_data)
summary(linear)
```

Call: lm(formula = installed_pv_contractors_v2 ~ utility_outage_hours, data = ps3_data)

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

We can see the estimate staying almost the same $6.90 \cdot 10^{-5}$ with backchecks and $6.84 \cdot 10^{-5}$ with contractors v2, while the standard error slightly decreases. Neither estimate is statistically significant.

(as a note, the results shown use the sample of complete cases, eliminating nas so that the standard error does not change because of this)

8 -

In this case we learn that there is a problem with the treatment variable, which is called classical errors for $iou = 1$ and non-classical errors for $iou = 2$. Remembering our regression model for our problem, we have:

$$Y_i = \alpha + \tau h_i + \epsilon_i$$

The problem for data with classical errors in the treatment variable is that we don't observe h_i. Instead we see a noisy variable $\tilde{h}$ which we can describe as:

$$\tilde{h}_i = h_i + \gamma_i$$

The problem is that, for the case of classical errors, the effect we are interest in, τ, which we estimate by OLS as:

$$\hat{\tau} = \frac{\text{Cov}\left(Y_i, \tilde{h}_i\right)}{\text{Var}\left(\tilde{h}_i\right)}$$

$$= \frac{\text{Cov}\left(Y_i, h_i + \gamma_i\right)}{\text{Var}\left(h_i + \gamma_i\right)}$$

$$= \frac{\text{Cov}\left(\alpha + \tau h_i + \varepsilon_i, h_i + \gamma_i\right)}{\text{Var}\left(h_i + \gamma_i\right)}$$

Expanding the covariance term and the denominator we get to:

$$= \frac{\tau \,\text{Var}\left(h_i\right)}{\text{Var}\left(h_i\right) + \text{Var}\left(\gamma_i\right)} \neq \tau$$

Which is smaller than τ. This is called an attenuation bias. For the case of the utility 2 (when $iou = 2$), since the treatment variable has an extra term which correlates with the treatment:

$$\tilde{h}_i = \delta h_i + \gamma_i$$

We will also have a bias in the estimation of τ as before as we are only adding an extra term. τ becomes in this case:

$$= \frac{\tau \,\text{Var}\left(\delta h_i\right)}{\text{Var}\left(\delta h_i\right) + \text{Var}\left(\gamma_i\right)} \neq \tau$$

We next proceed to show our regression results for each utility. We use the backcheck data for being the true measurement of the installed PV capacity and estimate the effect of outages for both utilities separately. First for $iou == 1$ and second for $iou == 2$:

```r
util_1 = ps3_data[ps3_data$iou == 1,]
util_2 = ps3_data[ps3_data$iou == 2,]

linear = lm(pv_adoption_backchecks ~ utility_outage_hours, util_1)
summary(linear)
```

Call: lm(formula = pv_adoption_backchecks ~ utility_outage_hours, data = util_1)

Residuals: Min 1Q Median 3Q Max -3.103 -2.998 -1.001 2.186 15.343

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.9982165 0.1882659 15.925 <2e-16 *** utility_outage_hours 0.0001078 0.0006512 0.166 0.869
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 3.489 on 350 degrees of freedom (647 observations deleted due to missingness)
Multiple R-squared: 7.828e-05, Adjusted R-squared: -0.002779 F-statistic: 0.0274 on 1 and 350 DF, p-value: 0.8686

```r
linear = lm(pv_adoption_backchecks ~ utility_outage_hours, util_2)
summary(linear)
```

Call: lm(formula = pv_adoption_backchecks ~ utility_outage_hours, data = util_2)

Residuals: Min 1Q Median 3Q Max -2.829 -2.677 -1.959 2.001 13.796

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.6382565 0.3032259 8.701 <2e-16 *** utility_outage_hours 0.0006629 0.0029222 0.227 0.821
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 3.452 on 331 degrees of freedom (668 observations deleted due to missingness)
Multiple R-squared: 0.0001554, Adjusted R-squared: -0.002865 F-statistic: 0.05145 on 1 and 331 DF, p-value: 0.8207

For the first utility we find the coefficient of outages to be $\tau = 1.08 \cdot 10^{-4}$ and for the second utility $\tau = 6.63 \cdot 10^{-4}$, which expresses a change of $\tau[kW]$ of installed PV capacity per aditional hour of power outages. Neither is statistically significant so we fail to reject the null hyphotesis of no impact of outages on installed PV capacity.

9 - The survey is a variable that is highly correlated with our treatment (hours of power outages) and is therefore a candidate for a good instrument for an instrumental variables approach. An IV approach is a way of using this data to correct the measurement error when the treatment is non-binary as the issue reported in (8). For this method to produce an unbiased estimate we need the assumptions explained in (2) to be satisfied. These are $Cov(Z_i, h_i) \neq 0$ and the exclusion restriction $Cov(Z_i, \epsilon_i)$, which we can't be tested. We also need the measurement error in the instrument to be uncorrelated with the error in the treatment (γ_i), uncorrelated with the actual treatment h_i, and uncorrelated with the error in the first regression ϵ_i. This means we need to use the second utility data ($iou = 2$) since we can't have correlation with the treatment (Non-classical errors).

Using this approach we can perform a two stage linear regression by regressing Z_i on h_i first and use the predicted values as a regressor on the second stage, with installed PV capacity as the dependent variable. Starting with the first regression to check the first assumption, we obtain:

```r
tsls1 <- lm(utility_outage_hours~survey_outage_hours, util_2)
summary(tsls1)
```

```
##
## Call:
## lm(formula = utility_outage_hours ~ survey_outage_hours, data = util_2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -57.167  -9.980   0.144  10.855  65.020
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.091965   0.785343   2.664  0.00785 **
## survey_outage_hours 0.969402   0.007523 128.864  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.58 on 999 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9432
## F-statistic: 1.661e+04 on 1 and 999 DF,  p-value: < 2.2e-16
```

```
d.hat <- fitted.values(tsls1)
```

Since we find a very high correlation between the variables ($F - statistic = 1.66 \cdot 10^4$). We have tested the first assumption and it was positive. Now we can perform the second stage, regressing the fitted values in our outcome variable.

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
d.hat <- fitted.values(tsls1)
tsls2 <- lm(pv_adoption_backchecks~d.hat, util_2)
summary(tsls2)
```

```
##
## Call:
## lm(formula = pv_adoption_backchecks ~ d.hat, data = util_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.807 -2.679 -1.952  1.999 13.798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.6461075  0.3067931   8.625 2.71e-16 ***
## d.hat       0.0005724  0.0030115   0.190    0.849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.452 on 331 degrees of freedom
##   (668 observations deleted due to missingness)
## Multiple R-squared:  0.0001091,  Adjusted R-squared:  -0.002912
## F-statistic: 0.03612 on 1 and 331 DF,  p-value: 0.8494
```

We find the coefficient to be $\tau = 5.72 \cdot 10^{-4}$ and also not statistically significant. Since we have corrected the measurement error by using an instrumental variables approach, we would send this last coefficient to CALBEARS as final results. The coefficient compares to the ones from 8 since it is slightly smaller than the one calculated for the second utility.