Understanding and Applying Instrunmental Variables, PS3

Andrew Caughey

Questions:

The state of California is increasingly worried about wildfire season. In particular, sparking electricity transmission lines have been identified as a cause of large fires in 2018 and 2019. In response to this, the state has required that its electric utilities shut off power during times of high fire risk, as part of a program of public safety power shutoffs, which they are calling. However, as part of a cost-benefit analysis to inform future policy, the state needs an estimate of the impact of blackouts on household adoption of solar PV (since, with the right inverter, a solar panel can continue to power a house even when grid power is off). The state have hired an analysis team, called the CALifornia Blackouts and Electricity Analysis Research Service (CALBEARS). However, these analysts are in a bit over their heads, so they've called you in to help (go Maroons!).

1.

CALBEARS are interested in answering the following question: What is the effect of hours of electricity outages experienced by a household on the kW of solar PV they install? To make sure everybody is on the same page, explain to them what the ideal experiment would be for answering this question. Describe the dataset that you'd like to have to carry out this ideal experiment, and use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (ie, what is "i" here?).

A randomized control trial is our ideal experiment, but it may not be possible because it would costly and ethically challenging to implelment. In a perfect world, we'd like to randomly assign half of the population of households to a treatment group which would experience mandatory electricity outages and a control group which would not. Our unit of anaalysis, here, is the household i. In the real world, the best we can hope for is a large, well-balanced random sample of households. To ensure balance, we'd like to know a laundry list of pre-treatment observed characteristics; these might include family member demographics, employment status, income, type of housing, age, etc. This helps us address the fundamental problem of causal inference: though we never see household i at $D_i(1)$ and $D_i(0)$ at the exact same moment in time, n math, if we let Y be the kW hours of solar PV installed for an indivdual household i, and D be an indicator where 1 means an indiivudal was assigned to experience blackouts and 0 otherwise, we'd like to estimate the causal impact of outages with

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

However, since hours of electricity outages is a continuous variable, though, we need think a bit about how to implement outages for the treated group. The expected hours of a blackout should probably be normally distributed and centered around the average historical length of previous mandatory outages, but there should still be some variation in treatment status. In this case, treatment is no longer $D_i(1)$ or $D_i(0)$, but instead:

$$S_i \in (0, 1, ..., \bar{S})$$

Here, we'd like to estimate an average causal response, which will be a weighted average of those who received treatment over the share of the total set of people.

Further, we should think about wheter or not the policy views spillover effects as endogenous or not. Given shut-offs happen in response to an emergency, it seems reasonable to think that the state is interested in understanding what happens when many households experience outages simultaneously for roughly the same amount of time. In our ideal experiment, then, treated groups should all experience blackouts at the same time. This is important, because we might imagine households could be less likely to install panels if they are more likely to be the only person on the block experiencing a blackout of totally unknowable duration; they might prefer to just wait it out at a neighbors house or go out. If many people experience blackouts simultaneously for roughly the same amount of time, though, their treatment status might impact their other treated neighbors decision - that could be the intended effect of the policy.

Finally, we might think that the number of outages is an important factor, alongside total outage hours. In the same vein, we might guess that the total number of outage hours required to have an impact on solar panel might be very high, or the impact on installation could behave differently over a range of outage hour values. If that's the case, ideally, we'd like panel data that tracks the same households over time for multiple blackouts. Though household i will change over time, there are many time-invariant unobserved characteristics that we can difference out if we use panel data. We can use this to make across-household, within time, comparisions as well as within-household, across time, comparisons.

2.

CALBEARS are on board with your explanation, but, as they've discussed with you, they won't be able to implement your preferred solution. They don't think that a selection-on-observables approach will work (they're very sophisticated). They're also limited by state privacy laws: they will only be able to give you one wave of data (no repeated observations). Given these limitations, describe the type of research design you would try to use to answer their question of interest. Be explicit about the assumptions required for this design to work, describing them in both math and words.

Without repeated observations, we won't be able to get the panel data we'd like. I'm assuming here that, by "no repeated observations", this also rules out repeated cross-section data and other time series data. In that case, we're stuck in familiar terrority: cross-sectional data from a single point in time. I'm not clear from the question if this means they won't be able to implement a randomized control trial, but I'm assuming that's the case since that follows the pattern of previous psets.

If we can't implement and RCT and we're tepid about selection-on-observables strategies, we might be able to explore causal relationships using a natural experiment. In this approach, we're looking for naturally occuring or policy induced variation that impacts treatment status as-if by random. In essence, we're hoping that even while we can't manually assign households to treatment and control, there's some Other Thing out in the wild we'll call Z that is at least quasi-randomly impacting households hours of power outages. If that's the case we can use this random variation as a lever (or insturnment) to make causal inferences about the impact of outage hours of kW of solar installed. In essence, we'll use the quasi-random variation and chuck out the non-random variation - in theory, this means breaking our assignment to treatment vector into two parts $B_i\epsilon_i$ is the non-random, selected junk we can't work with, but C_i is random variation we can exploit. In math, this means we're working with:

$$Y - i = \alpha + \gamma D_i + \beta X_I + \epsilon_i \text{ , where } D_i = B_i\epsilon_i + C_i$$

Since we can't just intuitively split up the variation when collecting our data, ultimately, we'll need to use two-stage OLS regression analysis to implement this approach. We'll run a regression with our insturnment to create predicted values of our treatment indicator, and then use those fitted values along with a running variable of observables to predict outcomes. In math, our first stage is

$$D - i = \alpha + \gamma Z_i + \beta X_i + \nu_i$$

Armed with our predicted treatment indicators $\hat{D}_i$, generated from the random variation in our insturnment Z_i, we run another regression to predict outcomes:

$$Y_i = \alpha + \gamma \hat{D}_i + \delta X_i + \epsilon_i$$

Importantly, we need to remember that, because we're using fitted values and have introduced more noise, our standard errors here won't be correct unless we account for this

We make some assumptions for this to fly. First, we're asumming our insturnment Z's variation is at least quasi-random, and that it has a statistically significant covariance with power outages. Hopefully, this relationship isn't just statistically significant, but also has some magnitude in it, too - that'll make the insturnment more useful in the long run. Luckily, we can test this - in math, we're looking for

$$Cov(Z_i, D_i) \neq 0$$

Our more challenging assumption is meeting the fundamentally untestable exlusion restriction. We need to assume that our insturnment Z only impacts our outcome - household purchases of solar kW hours - through its impact on treatment assignment - household power outages. If Z impacts something else, say household wages for example, Z might impact installation of solar panels through multiple causal pathways. That's bad - we need the covariance of our instrunment to be uncorrelated with our error term, $Cov(Z_i, \epsilon_i) = 0$. However, we can never test this, because our error term is, by definition, unobserved.

We also require that we have no non-linear estimators (though different function forms are OK, for example $\beta_1 age + \beta_2 age^2$. We also need to make a decision about how we thinkg our IV impacts different households. If we think it has the same impact for every household, the first-stage and exclusion restrictions suffice. However, if we think the IV impacts treatment differently for households, we need to assume monotonacity $D_i(Z_i = 1) - D_i(Z_i = 0) \geq 0$ for all i - that the insturnment impacts treatment in the same *direction* for all households, though not the same magnitues - also independence $Y_i(D_i, Z_i), D_i(1), D_i(0) \perp Z_i$, which is an expansion of the exclusion restriction. This implies that potential outcomes are the same when the insturnment is a 1 or a 0. With heterogeenous treatment effects, we can estimate $\hat{\tau}^{IV} = \tau^{LATE}$ for those who are impacted by the insturnmental variable.

3.

CALBEARS are interested in this research design. It sounds promising. They'd like you to propose a specific approach. Please describe a plausible instrumental variable you could use to evaluate the effect of power outage hours on kW of solar PV installed. Why is your proposed instrument a good one? Do you have any concerns about your ability to estimate the treatment effect using your instrument? If yes, why? If no, why not?

A plauisble intsurnment needs to impact a household's purchase of solar PV kW hours *only* by having a quasi-random impact on that households total hours of power outages. Some of the most common causes of power-outages are storms, trees, and lightning; I'm concerned, though, that these variables might make panels more viable in some areas than others (stormier areas with high winds might be less inclined to buy panels because they get less sun or could be damaged by wind).

Otherwise, power outages are often caused by excavation digging, animals destroying equipment, and vehicles colliding with utility polls. These events aren't totally random, but are probably quasi-random, and are likeley to play essentially no role in households decisions to buy a panel, other than the frequency of outages. For this reason, I'd suggest we use outages caused by vehicle excavations, collisions with utility poles, and animals destorying equipment, excluding outages caused by weather and high-demand for electricity.This possible insturnment has plausible quasi-variation and seems to meet the exclusion restriction more than the other common, weather-based causes of outages.

This isn't a perfect insturnment by any means. It could be the case that animal-incuded power outages correlates highly with trees or stormy weather conditions, and if forested and stormy areas have a negative impact on household's propensity to buy panels, then our insturnment might be compromised. More concretely, these are probably relatively rare events, so its not clear that we would actually have enough observations to make this a viable insturnment. I'm also not sure if we would have data on the concrete cause of outages in the real world.

common outage causes from: https://energized.edison.com/stories/8-common-causes-of-outages

4

CALBEARS is intrigued by your approach. After an internal discussion, they've come back to you with great news! It turns out that one of the California utilities ran a small pilot program where they randomly cut power for different lengths of time to different households as part of an equipment test. With this new information, please describe to CALBEARS how you would estimate the impacts of power outages on solar PV adoption? Use both words and math.

It's encouraging to hear that people randomly got their power cut off, at least for our research design. Assuming we have data about how many kW of solar PV they installed after experiencing some total hours of otages, we can exploit this policy-induced random variation to estimate the impacts outages had on installing panels.

First, as mentioned above, wed'd like to conduct a balance test just to make sure the randomization process worked, and that households who had power cut and not cut are similar in observable characteristisc.

Importantly, because power was cut for different lengths of time, we need to recognize we are working with a continuous treatment variable, not a binary one or zero. As I discussed in question 1, when treatment is no longer $D_i(1)$ or $D_i(0)$, but instead:

$$S_i \in (0, 1, ..., \bar{S})$$

we'd like to estimate an average causal response, which will be a weighted average of those who received treatment over the share of the total set of people. In math, we're working with:

$$\hat{\tau}^{IV} = \sum_{s=1}^{\bar{S}} w_s E[Y_i(s) - Y_i(s-1) | S_i(1) \geq s \geq S_i(0)]$$

This looks grosser than it is, really we are just getting a weighted sum of the difference between oncomes under s and s-1 conditional on the treatment value being greater than 0 i.e. getting some kind of treatment. This us an estimate of the average causal response. If we have perfect compliance (which seems likeley since these outages were forced on households), and everyone was effected equally by the blackouts, this is the average treatment effect meaning those who actually were impacted by outages.

5.

CALBEARS agree that your approach is a good one. So good, in fact, that they'd like to see it in action! They are willing to share some data with you, in the form of ps3_data.csv. Please report the results of an analysis of the impacts of power outage hours on kW of PV adoption, using utility_outage_hours as your treatment variable and installed_pv_contractors as your outcome variable.

First, it's important we inspect the data. Here's the code I used to do that with no output, just for readability.

Ideally, we'd like to do a balance check to make sure the randomization here actually worked (i.e. created statistically similar control and treatment groups). Assuming this worked, we can exploit the efficacy of randomizaton estimate the average treatment effect with:

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

Because of randomization, this actually simplifies even further to something that looks an awful lot like the naieve estimator. We simply difference the means between the treated and control group. Remember, though, we can only do this because randomization has eliminated the selection problem.

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

In regression form, this is just

$$Y_i = \alpha + \beta_1 D_i$$

This regression's output leads us to think that there's actually not a statistically significant relationship between the total hours of outages and purchasing more kW of solar panels, and theres probably not a causal effect.

```
# outcome: installed_pv_contractors
# treatment: utility_outage_hours
# remmeber, this variable has some fat negative values, hard to interpret

lm(installed_pv_contractors ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                 estimate std.error statistic p.value
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          3.54      0.0645      55.0     0
## 2 utility_outage_hours 0.000111  0.000298     0.374   0.708
```

```
# 1315 NA obs
gobears %>%
  filter(is.na(installed_pv_backchecks)) %>%
  nrow()

# lets plot some vars, get a sesne of data

# excatly even
gobears %>%
  ggplot(aes(x = iou)) +
  geom_bar()

# pretty normal, but one extreme outliers at -1500?
# remeber this and check it out later
```

```r
gobears %>%
  ggplot(aes(x = utility_outage_hours)) +
  geom_histogram(bins = 50)

# lots of 0s, and skewed to the left, spike center at 5
gobears %>%
  ggplot(aes(x = installed_pv_contractors)) +
  geom_histogram(bins = 30)

# spike at 0, kind of spikey normal with long tai.
# not smooth - but plateau
gobears %>%
  ggplot(aes(x = installed_pv_backchecks)) +
  geom_histogram(bins = 60)

# spike right before 5, normal-ish
gobears %>%
  ggplot(aes(x = installed_pv_contractors_v2)) +
  geom_histogram(bins = 60)

# some center at just above zero, another around 80
# normalish
# negative values
gobears %>%
  ggplot(aes(x = survey_outage_hours)) +
  geom_histogram(bins = 60)
```

6

CALBEARS like your analysis, but they're a bit worried about the quality of their data on PV adoption. The way they normally collect these data is by collecting details on solar installations from contractors. However, they went and did some back-checks in a subsample of data that they gave you, and noticed that the contractor reports seem to be off. They would like you to make a graph showing the relationship between their back-checks (installed_pv_backchecks) and the contractors' estimates (installed_pv_contractors). Describe to them what you find. Is this likely to be a problem for your analysis? Why or why not? Next, estimate the impacts of power outages on PV adoption using the backcheck data and using the contractor estimates. Report what you find. Do your estimates differ? If no, explain why not. If yes, explain why.

When we compare the two measurements of outcomes, there's a real disparity. As the plot below shows, if the to measures were exactly the same we'd expect a straight line with a slope of one. In reality, though, we aren't even close to that - for example, when our backchecks told us no solar panels had been installed, our contractrs often told us there were 5kw hours of Solar panels! The correlation sign between the two measurements isn't even consistent across different values, and is really only positive between about 4 kW hours and 15 kW hours.

Why is this the case? It could be that the subsample selected for backchecks wasn't a properly randomized sample, so our backchecked measurements suffer from unobserved selection bias, which is why the measures are so off when we compare them to each other. Checking the balance between the subsample and everyone else would be a good place to start here.

It could also be that one or both of our measurements is impacted by measurement error. Instead of observing Y_i, we observe $Y_i + \gamma_i$, where gamma is some error that pushes us farther from the truth. This sounds bad, but because this measurement error is in our outcome variable, if certain assumptions are met we don't need to be too worried. This is because, after a lot of math, outcome variable measurement error expressed in

$$\tau = \frac{Cov(Y_i + \gamma_i, D_i)}{Var(D_i)}$$

simplifies down to

$$\tau = \frac{\tau Cov(D_i, D_i)}{Var(D_i)}$$

This is the same as

$$\tau * \frac{Var(D_i)}{Var(D_i)} = \tau$$

Here, we see that even when we measure our outcomes with error, we still can get to the same estimate we would have arrived at if we measured without error. Importantly, though, this is only true if we assume that the noise in our measurement isn't actually correlated with any of the unobserved characteristics in our error term $Cov(\gamma_i, \epsilon_i) = 0$. Further, we're assuming that the covariance of the noise in outcomes isn't correlated with treatment status, either $Cov(\gamma_i, D_i) = 0$. If either of these assumptions are violated, we have a pretty serious problem.

It seems reasonable to be to think that at least one of these assumptions are likely met in this case - if randomization was done properly and the samples are balanced, the reason our contractors might mismeasure solar kW hours should be independent of wheter or not a house was randomly assigned to treatment.

Perhaps, though measurement error is correlated with an unobserved variable in our error term: for example, maybe contractors are lazy and don't actually check and measure installed panels when they would have to use a ladder to climb multiple floors onto a steep roof to do so. But suppose our backchecks climbed roofs and found these installed panels were actually broken. This might explain why contracts often put "5" when our backchecks said 0. In this case, perhaps wealthier, multi-story households are more likeley to have a systematic measurement error correlated with an unobserved variable indicating steep roofs or multi-story

homes floating around in our error term. This would violate our assumption and could be a problem for our analysis. In this case, we might prefer to lean on our sample of backchecks more heavily.

When we use the outcome measurements from our backchecked sample and compare this to the regression from our contractor's measurement of outcomes, though, the relationship between power outage hours and solar panel installation still isn't statistically significant and the effect is close to zero. We are much closer to statistical significance: instead of there being a 70% chance the minimal effect we observed was due to random chance, now, there's only a 10% probability we are observing this relationship due to random chance. However, because any causal impact from blackouts is essentially zero, this isn't all too important a distinction. If the backcheck measure is more reliable, this may suggest that the $Cov(\gamma_i, \epsilon_i) = 0$ and $Cov(\gamma_i, D_i) = 0$ assumptions are more or less holding.

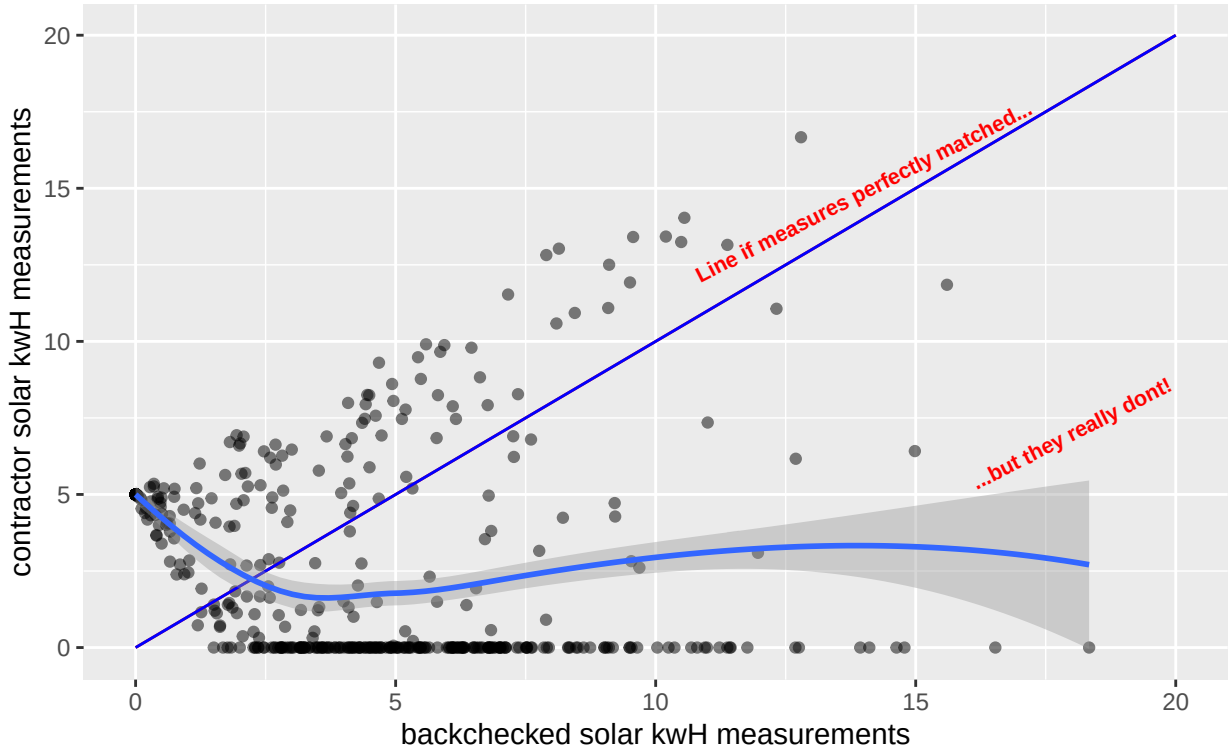
```
# plot relationship between back-checks and contractor estimates

# whats a good way to represent this? scatter?

same_measurement_annotation <- data.frame(
  x = c(18, 14),
  y = c(7, 14.9),
  label = c("...but they really dont!", "Line if measures perfectly matched...")
)

# only comparing sampled obs with backchecks
gobears %>%
  filter(!is.na(installed_pv_backchecks) & !is.na(installed_pv_contractors)) %>%
  # plot
  ggplot(aes(x = installed_pv_backchecks, y = installed_pv_contractors)) +
  geom_point(alpha = .5) +
  annotate("segment",
    x = 0, xend = 20, y = 0, yend = 20,
    color = c("red", "blue"), fill = c("red", "blue")
  ) +
  geom_text(
    data = same_measurement_annotation, aes(x = x, y = y, label = label),
    color = "red",
    size = 2.9, angle = 27, fontface = "bold"
  ) +
  geom_smooth() +
  labs(title = "How much do our measurements disagree?", subtitle = "A lot!") +
  xlab("backchecked solar kwH measurements") +
  ylab("contractor solar kwH measurements")
```

```
## Warning: Ignoring unknown parameters: fill
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

How much do our measurements disagree?

A lot!




```
# outcome: installed_pv_contractors
# treatment: utility_outage_hours
# remmeber, this variable has some fat negative values, hard to interpret

lm(installed_pv_contractors ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                 estimate std.error statistic p.value
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          3.54      0.0645      55.0     0
## 2 utility_outage_hours 0.000111  0.000298     0.374   0.708
```

```
# outcome: installed_pv_backchecks
# treatment: utility_outage_hours

lm(installed_pv_backchecks ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                  estimate std.error statistic  p.value
##   <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)          2.85       0.138       20.6   1.19e-73
## 2 utility_outage_hours 0.0000690  0.000630     0.110 9.13e- 1
```

7.

The challenge with back-checks is that they're very expensive to do. Fortunately, CALBEARS realized that there's another field on the contractors' reports that seems to match the back-checks much better. They'd like you to make a graph showing the relationship between their back-checks and this new measurement (installed_pv_contractors_v2). Describe to them what you find. Is this likely to be a problem for your analysis? Why or why not? Next, estimate the impacts of power outages on PV adoption using the backcheck data and using the (new) contractor estimates. Report what you find. Do your estimates differ? If no, explain why not. If yes, explain why.

This alternate measure from contractors fits the backchecks much more clearly. It's nearly a perfect match, but its consistently about 3 kwH too high.

For the same reasons as above, this is unlikely to be a problem for our analysis. Our outcome variable measurement error expressed in

$$\tau = \frac{Cov(Y_i + \gamma_i, D_i)}{Var(D_i)}$$

ultimately simplifies to

$$\tau * \frac{Var(D_i)}{Var(D_i)} = \tau$$

assuming our measurement error isn't correlated with a variable in our error term or with our randomized teratment assignment mechanism.

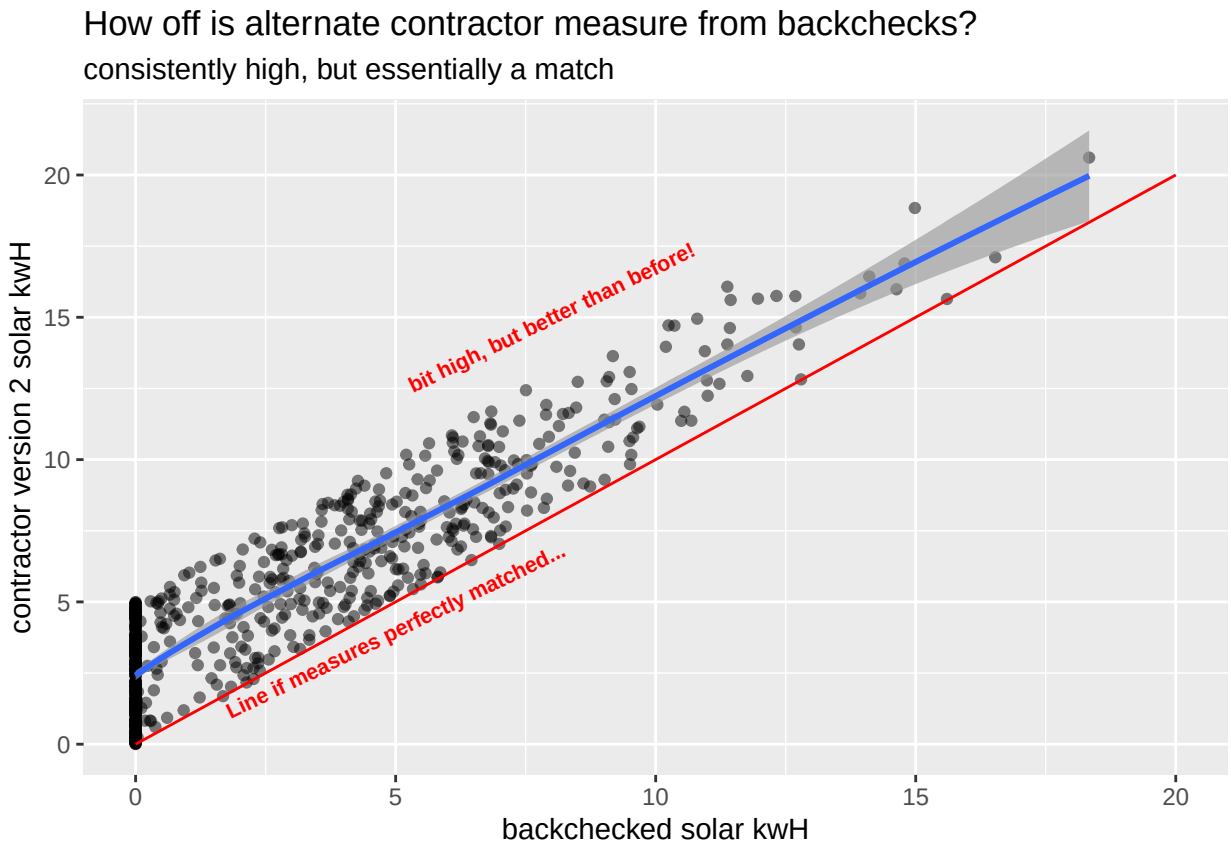
Indeed, when we use the alternate contractor measure that more closely mirrors the backchecks, we still have a statistically unsignificant result with a very, very small coefficient. It's still really unlikely that there is much of a causal relationship between outages and solar panel purchases, and the size of any causal impact would be very close to 0. Even though *installed_pv_contractors_v2* more closely mirrors or backchecks and may have less measurement error, because were grappling wtih measurement error in the outcome Y_i variable, our estimates don't really change too much regardless.

```
# plot relationship between back-checks and contractor estimates

# plot annotation
v2_annotation <- data.frame(
  x = c(8, 5),
  y = c(15, 4),
  label = c("bit high, but better than before!", "Line if measures perfectly matched...")
)

# only comparing sampled obs with backchecks
gobears %>%
  filter(!is.na(installed_pv_backchecks) & !is.na(installed_pv_contractors_v2)) %>%
  # plot
  ggplot(aes(x = installed_pv_backchecks, y = installed_pv_contractors_v2)) +
  geom_point(alpha = .5) +
  geom_smooth() +
  annotate("segment",
    x = 0, xend = 20, y = 0, yend = 20,
    color = "red", fill = "red"
  ) +
  geom_text(
    data = v2_annotation, aes(x = x, y = y, label = label),
    color = "red",
    size = 2.9, angle = 26, fontface = "bold"
```

```
  ) +
  geom_smooth() +
  labs(title = "How off is alternate contractor measure from backchecks?", subtitle = "consistently high
  xlab("backchecked solar kwH") +
  ylab("contractor version 2 solar kwH")
```

```
## Warning: Ignoring unknown parameters: fill
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# checking with all our different outcome measures

lm(installed_pv_contractors_v2 ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                  estimate std.error statistic p.value
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          5.15       0.0851      60.6     0
## 2 utility_outage_hours 0.0000684  0.000393     0.174   0.862
```

```
lm(installed_pv_contractors ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                  estimate std.error statistic p.value
```

```
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          3.54      0.0645      55.0     0
## 2 utility_outage_hours 0.000111  0.000298     0.374   0.708
```

```
lm(installed_pv_backchecks ~ utility_outage_hours, data = gobears) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)          2.85       0.138       20.6   1.19e-73
## 2 utility_outage_hours 0.0000690  0.000630     0.110 9.13e- 1
```

8.

CALBEARS comes back to you again with yet another data problem. This time, they're worried that the utilities aren't measuring the hours of power outages very well. CALBEARS explain to you that, in one utility (labeled iou == 1 in the data, because #privacy), something was going wrong with the power grid meters they were using. Houses that only had an hour or two of power outages appear to be reported accurately, but the higher the hours of outage, the more inflated utility 1's measurements are. In the other utility (labeled iou == 2), there are still imperfect measurements, but CALBEARS is convinced that the measurement problems are random. Explain the implications of these data issues in each utility to CALBEARS. Are these measurement issues going to be a problem for your analysis? Use words and math to explain why or why not. Despite any misgivings you might have, run your analysis anyway, separately for each utility this time (using your preferred PV variable from the three described above), and report your findings.

There are more serious issues. The total hours of power outages are the treament variable in our research design. While measurement error in our outcome variable ended up not being a serious problem, measurement error in our treatment indicators will be.

Now, instead of measuring D_i, we're getting $\tilde{D}_i = D_i + \gamma_i$. This is a problem, because simplifying down, we're left with

$$\hat{\tau} = \tau * (\frac{Var(D_i)}{Var(D_i) + Var(\gamma_i)})$$

Because we don't know the variance of γ_i, we can't estimate this thing away *and* our estimate will always be smaller than the truth.

Even if the measurement error in utility 2's data is random, we're still going to have attenuation bias. This is classical measurement error, which is very common. Even random noise will "stretch" our distribution out and make our data's line of best fit slope seem flatter than the true slope really is.

In contrast, the measurement error in utility 1's data isn't random - it's correlated with an observation's distance from the distribution's center. Still, this is effectively stretching the data out too, just more dramatically - points that were close to the distribution's true center may not be move much, but points farther away from the distribution's true center are pushed and pulled along the x axis by *a lot*. This means that our line of best fit's slope will, again, really flatten out. That causes us to underestimate the true coefficient for the variable, or, in other words, we understate the size of any causal impact outages has on buying solar panels. In this way, even non-classical measurement error in utility one is essentially "stretching" the data horizontally and attenuating our estimate - points that were close to the distribution's true center may not be move much, but points farther away from the distribution's true center are pushed and pulled along the x axis by *a lot*. This means that our line of best fit's slope will really flatten out.

Further, we're also introducing a bunch of non-sensical negative values into the also get a shift into the data that don't have any natural interpretation. It's not clear, at least to me, how a household can have negative hours of outages - even if they are pumping electricity back into the grid. This noise is especially confusing.

Together, all of this causes us to underestimate the true coefficient for the variable, or, in other words, we understate the size of any causal impact outages has on buying solar panels.

These measurement issues could be a problem for our analysis if we use them as-is. If we proceed, we need to keep in mind that we're underestimating the true effect. Further, I'd be hesitant about usng an IV approach to handle the measurement error in these two variables, because we don't really know that utility 1 and utility 2 have the same kind of customers, so their measurements might not really be of the same kind of household unit. Without conducting a balance check, we might just be introducing more and more selection bias, and failing to solve the problem.

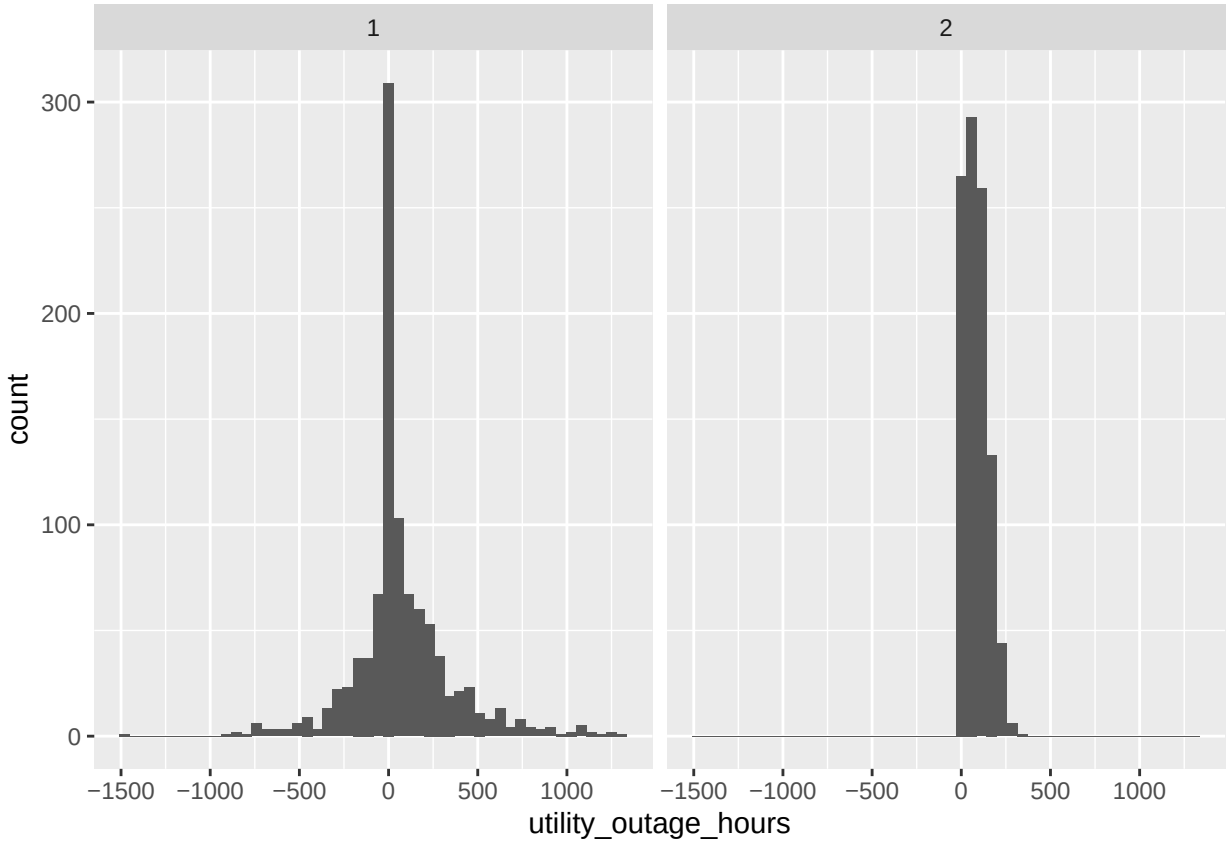
Throwing caution to the winds, we run the estimate anyway.

When we regress utility 1's outage hours on solar kW hours purchased as measured in installed_pv_contractors_v2, the results are hard to interpret. It seems like power outages have an attenuated, minisucle, and *negative* effect on consumer's installation of solar kW hours. We also end up with a near-zero but negative p value,

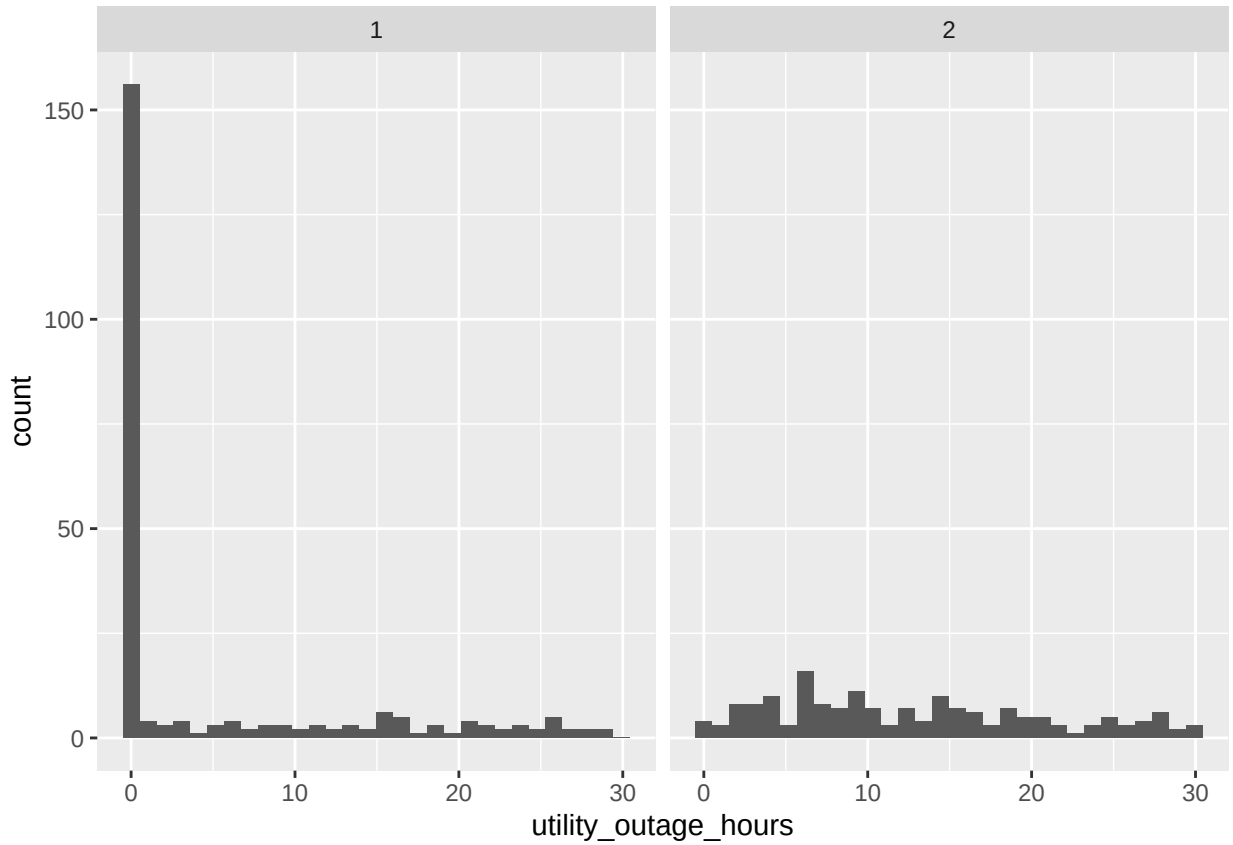
which has no natural interpretation. This is, perhaps, because our measurment error introduced a ton of negative values that also don't have any natural interpretation. Without correcting for these negative values, we're going to have a hard time making use of this data.

Using our alternate contractor measurement as the outcome variable, when we regress utility 2's outage hours on solar kW hours purchased, we still don't find a statistically significant relationship. With a p value of about 11%, though, there is a decent chance the relationship we've observed isn't just due to random change, but this is still outside our standard 5% confidence interval. Our estimated effect is, again, really small and close to zero. Still, we need to remember that, because of attenuation bias, the actual effect be much larger.

```
# first lets take a look at these two variabeles
# iou 2 is super tighly clustered arond normal
gobears %>%
  ggplot(aes(x = utility_outage_hours)) +
  geom_histogram(bins = 50) +
  facet_grid(~iou)
```




```
# shrink scale closer to distribution centers
# iou 1 is a plateau, but iou 2 has a downward slope and is closer to normal
# utility 1 has clear center at 0, but 2 does not - its ceneterd closer to 6
gobears %>%
  filter(utility_outage_hours >= 0 & utility_outage_hours < 30) %>%
  ggplot(aes(x = utility_outage_hours)) +
  geom_histogram(bins = 30) +
  facet_grid(~iou)
```




```
# all negative values are in utility 1
# did some homes really go without power for 41 days? over what period of time?
# probably not - question tells us higher values are inflated to be even higher


### running regression, using v2 PV outcome

# utility one
lm(installed_pv_contractors_v2 ~ utility_outage_hours,
  data = filter(gobears, iou == 1)
) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                  estimate std.error statistic   p.value
##   <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           5.26       0.117       45.1   5.91e-243
## 2 utility_outage_hours -0.0000653  0.000405    -0.161 8.72e-  1
```

```
filter(gobears, iou == 1) %>%
  nrow()
```

```
## [1] 999
```

```
# utility two
lm(installed_pv_contractors_v2 ~ utility_outage_hours,
  data = filter(gobears, iou == 2)
) %>%
```

```
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                 estimate std.error statistic   p.value
##   <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           4.84      0.179       27.1  1.05e-121
## 2 utility_outage_hours  0.00274   0.00172      1.59 1.11e-  1
```

```
filter(gobears, iou == 2) %>%
  nrow()
```

```
## [1] 1001
```

9.

CALBEARS conducted a survey of households to understand their power outage experiences, and asked households to report the number of hours of outages they experienced (survey_outage_hours). Describe how you could use these data to correct any issues you reported in (8). What conditions need to be satisfied in order for this to work? Carry out your proposed analysis. Report your results. What do you find? How do your results compare to your estimates in (8)? Which estimates would you send to CALBEARS as your final results?

Now that we have two measures of essentailly the same units and variable, we can use an insturnmental variables approach to start to tackle some of the measurment error introduced in the utility companies outage estimates. Essentailly, we'll use the two stage OLS regression described above. first, we leverage the non-noise variation in one measurement Z, which contains both information about D_i and noise ζ_i , or $\dot{D}_i + \zeta_i$. In this way, our insturnment contains some truth about $\tilde{D}_i$. In the first stage, we use this to our advantage and regress to generate predicated values:

$$D_i = \gamma Z_i + \nu_i$$

With our predicted values in hand, we plop them into a new regression to estimate our outcome without measurement error:

$$Y_i = \tau \hat{D}_i + \epsilon_i$$

In this way, we use random variation in one of our measures to handle classical measurement error (and nonclassical measurement error in rare casees, but the conditions aren't met here as the TA announcement pointed out).

This requies some assumptions, though. This only works on continuous variables, since binary variables won't mathematically produce the variation we need to exploit. We also need to make sure that Our insturnment actually correlates with the measured-with-error D_i variable: $Cov(Z_i, \tilde{D}_i \neq 0$. We again need to meet the exclusion restriction $Cov(Z_i, \epsilon_i) = 0$, meaning our insturnment isn't correlated with some other unobserved variable floating around in the error term. We also assume that the random noise in D_i, ζ_i isn't correlated with our measurement error - if its truly random thats reasonable. As long our our two measure's have some information about each other, and their measurement errors aren't correlated with each other, we can focus on the piece of variation that's common between the two and get closer to the truth.

We also need to account for the fact that we're using fitted values and adjust our standard errors.

We can apply these tools to the classical measurement error from utility 2's data. When we do so, we finally find that there still isn't a statsitically significant relationship between total hours of power outages and kW h of solar installed. The p value here is .20, so there's still a large probability the relationship we're observing is just due to random chance. Further, the magnitude of our estiamte is still basically zero, though it is slightly smaller than our attenuated measurements from section 8. This is actually pretty surprising to me, because we know our estimate from 8 suffers from attenuation bias due to measurement error - now that the error is out of the way we'd expect this estimate to be larger.

Ultimately, none of our estimats ended up being statisticlly significant, though our estimate of utility 2's consumers in section 8 without IV was closer to being statistially signifiant at a .11 p value. The magnitudes are also essentially the same, and are both nearly zero. My weasel answer is we should send both results and point out just how similar they are to show that, even when we account for measurement error, the results are similar. If pressed, though, I'd send over the estimate from part 8 because we're not just notifying them that measurement error exists, but we're doing our best to grapple with it.

```
###only using utility 2 because it has classical measurement error
gobears_u2 <- gobears %>%
  filter(iou == 2)
```

```r
## lets use survey as an insturnment - they strongly covary
cov(gobears_u2$utility_outage_hours, gobears_u2$survey_outage_hours)
```

```
## [1] 4158.76
```

```r
instrument_survey_model_2 <- lm(utility_outage_hours ~ survey_outage_hours,
  data = gobears_u2)

fitted_outage_hours_2 <- fitted(instrument_survey_model_2)

# outcome ~ fitted values
lm(installed_pv_contractors_v2 ~ fitted_outage_hours_2,
  data = gobears_u2) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term                  estimate std.error statistic   p.value
##   <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)            4.88      0.182       26.8  7.04e-120
## 2 fitted_outage_hours_2  0.00223   0.00177      1.26 2.07e-  1
```

```r
#robust standard errors
iv_robust(installed_pv_contractors_v2 ~ utility_outage_hours | survey_outage_hours,
          data = gobears_u2)
```

```
##                         Estimate Std. Error  t value      Pr(>|t|)     CI Lower
## (Intercept)          4.880876388  0.1772564 27.53568 1.194128e-124  4.533038800
## utility_outage_hours 0.002232089  0.0017785  1.25504  2.097576e-01 -0.001257934
##                         CI Upper  DF
## (Intercept)          5.228713975 999
## utility_outage_hours 0.005722112 999
```