November 25th:

I think we’re getting to the end of the simulations paper/simulations portion of the Medicare paper. I wanted to write up what we’ve done so far.

* We started doing simulations for this paper because we wanted to know how to define power outage exposure, as it’s a new thing and there are a few definitions out there (mostly based on ‘when more than p% of people are out for t length of time or longer, that’s an outage), but also because our data are really messy and have a lot of missingness and we wanted to know what bias that mess will introduce.
* So in summary: 2 sources of bias to quantify. Exposure misclassification bias that’s coming from a definition of power outage that may not be accurately representing when people have health effects from an outage, and bias from missing data.
* There are two types of missing data in the dataset. The data come in units of ‘city-utility’, which we have aggregated up to the county level.
  + Some time series at the city-utility level that show customers out are missing long chunks of time
  + Some counties are missing coverage of part of their population
* We want to test how these potential sources of bias will influence effect estimates with two types of simulation.
* We created fake data to represent city-utilities and counties. We made 100 datasets containing 100 counties each, and each county contained city-utility units. We drew the number of city-utilities in a county from the empirical distribution of city-utilities in counties in our data. (Mean, SD?)
* We assigned a number of customers to each city-utility, again drawn from the empirical distribution of customers at the city-utility in the real data.
* In each city-utility, we created hourly power outage data for a year, which is what we have for years 2018 and 2019 in our real dataset.
* We drew customers out in each hour from the empirical distribution of customers out each hour in the real outage data. (This was more complicated and I did something fancy to try to introduce some autocorrelation to make it look like the real data, but idk if that’s interesting.)
* We then cleaned the data as we did the real data – we aggregated the number of customers out to the hourly level at the city-utility level, and then aggregated those hours to the county level.
* We then identified ‘power outage days’.
  + We think that 8 hours might be the clinically relevant threshold at which power outages start to have health effects. Partly, we have chosen 8 hours because that’s how long a DME battery lasts.
  + We made a binary indicator that was 1 if the number of customers out in a county was above the x percentile for more than 8 consecutive hours that ended on a given day, and 0 otherwise
* We simulated outcome data.
  + We chose a baseline daily hospitalization rate of 0.001, 0.1% of people. (When we have Medicare data we can get the empirical rate, and base it off that.)
  + We hypothesized that this rate could increase on days when a power outage was affecting health by 0.0005, or 0.05%.
* First we just simulated data to validate a model and the procedure, and have a comparison to draw from. We drew from a Poisson distribution hospitalization counts for each day of the simulated year, based on the baseline rate, the number of people in a county, and whether or not that day was ‘exposed’ to power outage according to our above definition and data.
* lambda <- baseline\_hosp\_rate + effect\_of\_outage \* exp1$exposed, where exp1 is a vector of 1s and 0s representing whether there was an outage on a day or not.
* We set up a simulated case-control analysis. For every exposed day in our simulated outcome dataset, we chose two control days, one 2 weeks before the exposure, and one 2 weeks after the exposure, if there were such days available in the county. We created a simulated outcome dataset of just the cases and controls with the simulated exposure data (power outage y/n based on 8 hours), the simulated hospitalization counts, and the number of people in the county to use as an offset. We put the simulated hospitalization counts into a Poisson model, where the outcome was hospitalization counts, the exposure was power outage y/n, and the offset was the number of people in the county.
* We ran this model on 100 datasets of 100 counties, and plotted the mean bias in estimates of the hospitalization rate during a power outage, as well as confidence interval coverage.

I have done everything up to this point and I’m working on the rest.

* Then we moved on to models that actually tested some of the decisions we made in data cleaning and exposure definition.
* First we wanted to test the impact of missing data on the estimates of hospitalization rate due to power outages.
* We did this by generating power outage exposures based on the 8 hours x percent definition (x percent backed up by literature), and generated outcome data based on this. But then, we randomly removed 10%, 20%, and 30% of these binary indicators that were 1 showing power outage, and replaced them with 0s, and used the exposure data with missingness to run the same model on 100 datasets of 100 counties each.
  + This was meant to represent cases where there is missingness in a time series, like, half of the time series observations in a city-utility time series are 0 because they were not recorded, which is a thing that happens a lot in our data (could put an empirical number here of how often and how much that happens).
  + It also could represent cases where a chunk of the population in a county is missing – also common in our dataset.
  + We can’t really tell if, say, we are missing coverage for 20% of the population of a county, this will result in 20% of the power outages in this county being cut out of the dataset. That would assume a uniform distribution of power outages among people, which we know isn’t true. So we can’t really draw a straight line from person-coverage in a county or time-coverage in a city-utility to how much missingness there will be in the final cleaned data.
  + The same is true for time series – we do have decisions to make about what time series to include in the dataset. Should we drop a time series if it’s missing more than half of the observations? We also can’t know exactly how this will affect final missingness, but maybe the least bias would come from including everything that we have.
  + Hopefully these simulations will show that the bias in these situations is minimal.
* Second, we wanted to assess the impact of exposure misclassification/defining the exposure wrong.
* To do this, we simulated datasets that had indicators for if days were exposed to an 4 hour, 8 hour, or 12 hour outage as defined above as p% of people out for n hours or more ending on a day.
* We generated outcome data based on each of these new thresholds – so we generated some outcome data assuming that 4 hour outages would be clinically relevant, and again assuming 12 hour outages would be clinically relevant.
* Then, we analyzed the data where the outcome was generated with 4 hour outages, using the exposure data for 8 hour outages instead of the indicators for 4 hour exposure. We repeated this for 12 hour outages. This was meant to represent the scenario where we cleaned our data assuming that 8 hour outages are the relevant length, but actually 4 or 12 hour outages are.

What do we think?

Critiques?

Is this enough for a paper by itself? (Hopefully? I think it would be nice to write up these results that we’ve been working on for so long).

November 24th:

Justification for simulations:

* We decided on a baseline hospitalization rate of 0.001 – 0.1% of people.
* We thought that the effect of an outage might raise that rate by half. That’s maybe a bit much. We can do a sensitivity simulation on the hypothesized rate
* We calculated lambda for each exposed day, unexposed day, and by county. The hospitalization rate should depend on both the county, exposure status, and baseline rate
* This is lambda
* We then generated hospitalization counts from a Poisson distribution with rate lambda
* We then did a case-control setup. We got controls by looking two weeks prior and two weeks after cases. In this case we know there are no lagged effects of power outages, so this is just fine. In the real analysis we’ll need to be more careful about control selection.
* By calculations (in my notebook), we can see what the beta that we should get out of the simulation is, for the effect of an outage. It’s 0.405.
* We built a model and so far it returned 0.404, despite really high variance in the number of people by county, and power outages being a really rare event. Amazing.

July 18th:

Next steps:

* Make model of exposure and outcome
* Run that 100 times or whatever
* According to the case control thing
* For guessing the rate, Dan suggests that we can try a couple of different numbers, and then we can do sensitivity checks based on the rate
* We should also hypothesize about whether we think that the effect size is small or large compared to the baseline rate
* For county-level covariates, we could add season but it’s not a big deal/we don’t need to do it right away
* We also need to simulate actual counts rather than the rate, which means that we need to multiple the lambda by each population of each county in the outcome generating set
* Ok, then for our actual model.
* What we’re going to do is generate outcome datasets based on different scenarios – for example, we’ll do it with power outages that are only 4 hours creating differences in hospitalizations, or maybe 12 hours, and then we’ll run the models using the 8 hour power outage data
* That will show us how biased we would be if we were guessing the power outage length that’s relevant wrong
* For missing data, we should come up with a hypothesis about whether or not the missingness we’ve got is going to lead to us missing power outages, or over counting them
* We can then remove some, generate outcomes based on the data that has none removed, and run the model with the removed data
* We need to write about the outcome generating assumptions
* And then make histograms of the effect estimates

Copy paste of unedited notes:

**Dan:**

* Can try a couple different numbers
* And can do sensitivity checks
* Should be small relative to the baseline rate
* First: want to simulate counts rather than the rate; need actual counts
* To actually test the model
* glm(outcome ~ additive + equation, data = dataset, family = poisson(link=“log”))
* lm(outcome ~ additive + equation, data = dataset)
* glm(outcome ~ additive + equation + offset(N), data = dataset, family = poisson(link=“log”))
* glm(outcome ~ additive + equation + offset(log(N)), data = dataset, family = poisson(link=“log”))
* Going to match things

log(E[outcome])=additive+outcome+log(N)

log(E[outcome])-log(N)

log(E[outcome]/N)=log(rate)

glm(outcome ~ year \* power\_outage + offset(log(N)), data = dataset, family = poisson(link=“log”))

* You clean the data one way, for example 8 hr outage, and then produce an outcome
* Then you clean it another way and produce a different outcome
* Then you run the model for the first way, and you see how biased your results are
* Create outcome with 4, model with 8
* For missingness, create another dataset with the number of power outages and randomly remove 10%
* The data you actually have always produces the outcome

Other point:

* Figure out a way to scale it up
* Generate outcome data 100 times
* Look at the distribution of the effect
* Outcome on true data should be tight
* Should see a wider range or bias on simulated effects
* Come up with outcome generating assumptions
* Histograms of effect estimates
* We can discuss what types of changes in the data might the model be sensitive to

­July 11th:

* Ok so, I’m thinking that we need to meet with Dan to get more details on what the simulations should be
  + What is the baseline/real scenario we’re comparing this to? What is the baseline/real scenario that we’re comparing model estimates to? Like, if there’s an outage, we need to guess how many people would end up being hospitalized due to this outage, right?
  + Need to build the model to evaluate this, right?
    - How do we build this model? Matched pairs? Is there example code somewhere? I think that I saw that.
  + Need to also get realistic estimates of the number of hospitalizations and perhaps build in seasonality and another variable to the mix
* Also need to reread the model document
* Need to get the fake or real medicare data
  + Meet with Vivian to do this?
* Ok.

July 10th:

Ok to relate power outages to a health outcome

Next steps:

* Tidy scripts
* Push to github
* Apply scripts to generated data
* Generate outcome dataset

Want to rewrite scripts in a new format, where we can just lift them into the other data.

b01\_read and clean has got us to a place where we have the data recording changes at the city-utility level with names and stuff all reconciled.

b02\_expand I think we need to look at. Half of it probably needs to go in a new script?

Recall that we’re always going to be matching on county fips once we agg to the county level.

Ok so from that we need to pull the aggregation strategy. Makes sense to have new code because it’s chunked.

Meeting with Dan after:

* First we discussed git.ignore – Dan said just to save the data in a parent folder then back-propagate to that folder, that way you never accidentally push the data. I think this is a great idea and I’m reorganizing my stuff this way.
* We discussed explaining how I created the dataset. This is what we came up with:

1. I got the distribution of lengths of power outages from the data. For this, I just got the distribution of the lengths of time that customers out wasn’t 0.
2. I then selected n of these lengths such that the sum of those lengths was 6% of the total time in the simulated dataset. This is how much time in the original dataset customers out isn’t 0.
3. I then sampled JUST from the distribution of non-zero customers out values to fill in these ‘outages’ with non-zero customers out values.
4. I checked to make sure there was around 6% out.

We discussed next steps. Dan thinks the next step is to create a health outcome dataset. He thinks we should model hospitalizations with a Poisson model.

* First want to model the overall hospitalization rate – this is a number between 0 and 1 and should be small. Have something simple like baseline\_rate + B\*power outage + B2\*county-specific differences, with some added risk that we decide on from literature with a power outage.
* That will give the hospitalization rate.
* Then, multiply that county-specific rate by the number of people in the county, then sample from rpoisson(1, county-rate)
* That should give the number of hospitalizations per day.

Meeting with Dan Feb 23rd:

* Simulated counties
* Found out how long the whole time series spent out
* Added outages with lengths sampled from the EDF until it matched the time spent out
* Sampled customers out from the distribution of customers out after filtering out all observations that were 0.
* Going to make a github.
* Wondering about making a model.
* Wondering about making a git repo with everything but want to ignore data because it’s all way too big

Meeting with Marianthi Feb 22nd:

Main takeaways:

* For the simulation, it’s probably good to start with 5-10 datasets and then go from there
* The data processing pipeline is slow on 3000 observations (it takes a couple days) so it might be better to rewrite it in data.table so that it’s lightning fast.
* However, each of our simulations has only 100 counties, and I know the pipeline is actually pretty fast on 100 counties since I tested it on 100 counties for speed.
* So maybe the pipeline doesn’t need to be rewritten, since that’s probably a week of work? I’m going to ask Dan about this since I didn’t realize it during this meeting.
* May be possible to also just improve some of the slow code.

Meeting with Dan February 3rd:

* Data should resemble the dataset that you have both in the way that it will go through the data cleaning pipeline
* It should also resemble the data you have in that it should have similar distributions in it (like, if there are mostly counties with few subcounty areas, then the data should be similar). Distributions should match
* Also should think about what problems you have in your dataset that you’re simulating for.
* So to create the outage series, we could do 3 steps:

1. we want to find out how often an outage affects 0-10% of people, 10%-20% of people, etc., and then randomly input those outages
2. Then, we want to ask how long an outage lasts, find deciles, and randomly assign a length to the outage
3. Don’t need to worry about the correlation between those two variables.

* Can focus on using data.table, and using the code that Dan sent.
* Just realized I deleted the code Dan sent in the zoom. Great! Lol.
* Read the data.table vignette.

W­­­hat I’ve done so far:

* I simulated 100 counties.
* Our data’s geographic unit is this ‘sub-county area’, which is made up of a utility + a city where that utility is operating. These units might overlap in actual space, because neighbours might have different utility providers, but that’s our unit. I got the empirical distribution of the number of sub-county areas in a county, and then I drew from that distribution to assign a number of sub-county areas to each county.
* I got the empirical distribution from the data of the number of customers per sub-county area, and drew from that distribution to assign a number of customers to each sub-county area.
* I made a ten-minute interval time series.
* Going to find the mean and the variance of the percentage of customers out by sub-county area, and use a Weibull distribution maybe? to assign customers out at ten-minute intervals to the fake counties. I didn’t add any autocorrelation.

Questions:

* Should I be simulating like, a number of customers out at a ten-minute interval? How should I deal with the idea of a power outage? Should I be simulating individual customers, and then they’ll be some probability that if someone is exposed they get sick, or can I just have count data?
* Do I need to add autocorrelation? Marianthi is thinking that my simulation is too complicated. What should we improve upon? What should we leave simple?

02/01/2023

Goals of simulation:

1. Want to know how to define a power outage. As Marianthi says, this is a *measurement error* issue. We want to know what threshold of people in a county should be out and for how long in order for us to call it a power outage and see health effects.

2. Want to deal with missing data/person-coverage. There are two types of missingness – there is a problem where sometimes counties are missing entire sub-county areas. This is an issue of person-coverage – that means everyone in the county is not represented by the data. The other issue is regular old missingness where the time series has gaps/NAs in it. We want to know when we should completely discard a time series, and how much person-coverage a county should have in order to be included in the analysis.

Language for this dataset:

The power outages US dataset has information by state-county-cityname-utility. Our smallest spatial unit is the city-utility. We have customer counts at the city-utility, and these are nested within counties. So that’s what we’re going to simulate at.