Environmental Econometrics PS 1

Gabrial Erismann

2022-10-22

Exercise 1

Question i

The average treatment effect ATE, average treatment of controls ATC and average treatment of treated ATT are given by the following:

(1)

$$ATE = \mathbb{E}[y_{1i} - y_{0i}] = 1/N * \sum_{i}^{N} y_{1i} - y_{0i}$$

(2)

$$ATT = \mathbb{E}[y_{1i} - y_{0i}|D_i = 1] = 1/N * \sum_{i}^{N} y_{1i} - y_{0i} \; \forall \; i \; where \; D_i = 1$$

(3)

$$ATC = \mathbb{E}[y_{1i} - y_{0i}|D_i = 0]1/N * \sum_{i}^{N} y_{1i} - y_{0i} \; \forall \; i \; where \; D_i = 0$$

The ATE effect is the average effect on the outcome of interest of assigning individual i to treatment or to control. This is an exceptional situation where we see the realizations of the individuals with and without treatment. As such, we can calculate the average treatment effect directly using (1) to yield $ATE = 0.5$.

Note that since we have a linear operation we can move the subtraction out of the expectation and sum over all elements in each column instead of doing the element-wise operations.

The ATT is the effect of treament on the invididuals who were treated. Again, this is a unique (theoretical) situation where we can see the outcome of all individual $i „ $ so we can calcuate it directly using (2) to yield $ATT = 0.\bar{6}\bar{6}$.

The ATC is the effect of treatment on the invididuals who were not treated. Again, clearly this is not observed in the real world as each individual only gets one outcome and if they were treated they wouldn't be in the control group (ideally!). This can be calcuated directly using (3) to yield $ATC = 0.\bar{3}\bar{3}$.

The naive estimator of the average ATE is $\hat{\beta} = \mathbb{E}[y_{1i}|D_i = 1] - \mathbb{E}[y_{0i}|D_i = 0]$ or

(4)

$$\hat{\beta} = \mathbb{E}[y_{1i}|D_i = 1] = 1/N * \sum_{i}^{N} y_{1i} \; \forall \; i \; where \; D_i = 1 - \mathbb{E}[y_{0i}|D_i = 0]1/N * \sum_{i}^{N} y_{0i} \; \forall \; i \; where \; D_i = 0$$

It can be calculated using (4): to yield $\hat{\beta} = 1$.

These theoretical values should be equal given a large enough sample, but since we only have 6 observations in this small sample it is likely the randomness inherent in the data should result in the estimated values not being the same. In this case we can see that $\hat{\beta}$ is not equal to ATE

Note that the code in the annex reads in the table and performs the calculations.

Question ii

The ATT, ATC and ATE give different information. If we want to know what the effect of heat is on the population as a whole we use the ATE. If we want to know what the effect of heat is on people who are exposed we look at ATT. If we want to know the effect of exposing people to heat who have not been exposed we use the ATC.

Question iii

At first glance this appears to be quite an extreme value for $\hat{\beta}$ and I would expect $\beta < \hat{\beta}$. It does not seem likely that a heatwave would be so lethal for the 70+ population. If the data is drawn from a registry with some bias towards elderly people with other conditions (such as hospital admittance) then that could explain the high value. Otherwise, simply the small sampling size could have resulted in an overrepresentation of elderly people at higher risk.

Here we calculated already that $\hat{\beta} = 1$, $ATT = 0.6\bar{6}.$, $ATC = 0.3\bar{3}..$ So we can get the difference $\hat{\beta} - ATT = 0.3\bar{3}$ & $\hat{\beta} - ATC = 0.6\bar{6}$

Question iv

If the lowered health outcomes due to decreased healthcare availability *only* extend to the control group we get the following result of $\hat{\beta} = ATE = 0$. The effect of the distribution of healthcare resources would mask the effect of the heatwave. This would mean that an effective experiment would need to be conducted in a geographically sepatate enough area that healthcare resources are not rushed over to deal with the heatwave.

Code

```
# load in data
d = c(1,1,1,0,0,0)
y_1 = c(1,1,1,0,0,1)
y_2 = c(1,0,0,0,0,0)
df <- as.data.frame(cbind(d,y_1,y_2))

# calculate ATE using (1)
ate = (sum(df$y_1)-sum(df$y_2))/6

# calculate att by subtracting as inin

trues <- df[df$d == 1,]
falses <- df[df$d == 0,]
att =  (sum(trues$y_1)-sum(trues$y_2))/3
atc =  (sum(falses$y_1)-sum(falses$y_2))/3

ate_est = sum(trues$y_1-falses$y_2)/6

# load in data
d = c(1,1,1,0,0,0)
y_1 = c(1,1,1,0,0,1)
y_2 = c(1,0,0,1,1,1)
df <- as.data.frame(cbind(d,y_1,y_2))

# calculate ATE using (1)
ate_ = (sum(df$y_1)-sum(df$y_2))/6

# calculate att and atc
```

```r
trues <- df[df$d == 1,]
falses <- df[df$d == 0,]
att_ =  (sum(trues$y_1)-sum(trues$y_2))/3
atc_ =  (sum(falses$y_1)-sum(falses$y_2))/3

# estimate ATE
ate_est_ = sum(trues$y_1-falses$y_2)/6
```

Exercise 2

Question i

a Yes, I expect attenuation bias.

b There is always some amount of measurement and classification error, this is the nature of measuring and classifying. At the same time, these seem like highquality datasets so I would expect the classification scheme to be quite good. Nonetheless, there could be some "measurement error" in the sense that the temperature in urban areas is often higher at night than the temperature at the Automatic Weather Stations where official tempereature is recorded due to the urban heat island effect. Thus perhaps higher effective temperatures are not being consistently measured for some groups of people (in urban areas).

Another way in which attenuation bias can occur is model misspecification. In this case, the relationship between heat and mortality is far from linear, yet it is estimated as a linear regression parameter. This means that there could be significant divergences between the modeled value and the actual value which would increase the error in the estimation.

c This implies that there will be some amount of bias. Given the OLS regression below where H represents the number of heat days, D the per capita number of doctors in the commune, and y be the mortality rate of people of over 70, for every individual in the data set i:

$$Y_i = \alpha + \beta_1 H_i + \beta_2 D_i + \epsilon_i$$

Suppose the urban heat island is active and it has not been corrected for in the data set. In this case we have an additive error and can define our new $\bar{H}_i = H_i + u_i$ where u represents some additional number of heat days due to the urban heat island effect. Let $u_i \geq 0$ where $u_1 > 0$ for i in urban aras and $u_i = 0$ for i in rural areas. Suppose further that there is no measurement error in the doctors per commune D_i or in the mortality rate of over 70s Y_i.

$$Y_i = \alpha + \beta_1 \bar{H}_i + \beta_2 D_i + \epsilon_i$$

Using the formula for β_1 in a multivariate regression, we have our new estimator of the heat effect $\bar{\beta_1}$

$$\bar{\beta_1} = \frac{var(D)cov(Y, \bar{H}) - cov(D, \bar{H})cov(Y, D)}{var(\bar{H})var(D) - cov(D, \bar{H})}$$

Under usual assumptions, $\bar{\beta_1}$ converges to the following:

$$\bar{\beta_1} \longrightarrow \frac{\sigma_D^2(\beta_1\sigma_H^2 + \beta_1\sigma_{HD}) - \sigma_{\bar{H}D}(\beta_2\sigma_D^2 + \beta_1\sigma_{HD})}{\sigma_D^2(\sigma_H^2 + \sigma_u^2) - (\sigma_{\bar{H}D})^2}$$

While it is possible to simplify a bit further by pulling out β_1 and β_2, we can see that we will not have an intuitive result from this equation. In this case, it is quite plausible that the number of doctors D is correlated

with the measurement error u as there are likely more doctors in urban areas than in rural areas. Thus, we might have D correlated with the error term. This is indeed a noted problem in many OECD countries as one report noted: *"There are large differences in the density of doctors between predominantly urban and rural regions in France..." (OECD 2015)*

In this case we cannot reduce our equation further. It is important to note that in addition to bias in β_1 in general we have bias in β_2. This holds even if D is not correlated with u and would only be unbiased if D is not correlated with H.

However, if we assume that there is a governmental program that ensures doctors are well spread throughout the country in both urban and rural regions and in regions that do and don't experience heatwaves such that D,H and u are uncorrelated, then our formula reduces down to the following:

$$\bar{\beta_1} \longrightarrow \beta \frac{(\sigma_D^2 \sigma_H^2)}{\sigma_D^2(\sigma_H^2 + \sigma_u^2)} = \beta\lambda$$

In this case, as the size of σ_u increases, then λ decreases and so does our estimate of β_1. Intuitively, these deaths go unaccounted for and are attributed to the error term rather than being attributed correctly to β_1

*Derived from some old lecture notes.

d In order to fix the data, I would try to compensate by adjusting the heatdays variable by modelling the urban heat island effect. Basically, I would want to estimate $u_i \forall\, i$ It has a fairly consistent effect and there is data available on many cities, thus if geographic data is available for the individuals it could be possible to make adjustments for this for affected inviduals living in cities.

Question ii

a Yes, I expect omitted variable bias.

b In general, it is always possible that there are some ommitted variables that could have a larger or smaller effect on the analysis. In this case, I think there are a couple of potential ommitted variables that come to mind such as: 1) socioeconomic status of the indivudual 2) normal temperatures of the area

The formula for ommitted variable bias for the estimator of β_1, assuming that $cov(D, Z) = 0$ is given by:

$$\hat{\beta}_1 = \beta_1 + \frac{cov(H, Z}{Var(H)}\beta_1$$

Since the $Var(H) > 0$ for all values, the direction of bias is dependent on the $Cov(H, Z)$.

In the case of the socioeconomic status of the individual reflected by income S_i, I expect $cov(Y, S) < 0$, that is to say those who are less well off will have higher mortality, while those who are wealthy will have less mortality. In an area such as the south of France, I expect wealthier people to live close to the coast and to thus be exposed to fewer heatdays thanks to the cooling effect of the ocean. Therefore, we have $cov(H, S) < 0$. As such, I expect a negative bias term and therefore that $\frac{cov(H,Z}{Var(H)} < 0$ and thus $\hat{\beta}_1 < \beta_1$

In the case of the normal temperatures of the area, I expect those who are less routinely exposed to heatdays would be more likely to be affected by a significant heatwave. This could be because people in hot regions have access to infrastucture designed to relieve heat that may not be in clase in typically cooler regions. Thus, the same number of heatdays in a given year could lead to different outcome due to the different levels of preparedness within the population.

Let Z_i represent the median number of heatdays the individual is exposed to in a year. For any i, the number of heat days i is exposed to in a given year is going to be strongly correlated with the number of heatdays i has been exposed to in past years. Therefore $cov(H, Z) > 0$ Following the reasoning above, we have that

$cov(Y, Z) < 0$. Therefore, following the same formula, at first look I would expect $\frac{cov(H,Z)}{Var(H)} > 0$ and thus $\hat{\beta}_1 > \beta_1$

Question iii

The external validity concerns whether the results of this model are relevant to related literature and whether the results can be assumed to hold to a greater or lesser degree in another region. The answer to these questions can differ based on the context we want to use the study results to examine. If we are looking heat-realted mortality in different regions and want to use this study we would need to ask *Is the relationship between heatdays and mortality in Southern France representative of my region of interest?* The validity of the study would then depend on the region being compared. If we try to use this study to inform the effects of heatdays on mortality in the North of France, this might be quite different than using the study to look at mortality rates on the Adriatic coast of Italy. Additionally, if we were the authors of the study, we would want to additionally refer to literature in the subject area from similar regions and justify our estimated value for $\beta_h eat$ in the context of other, comparable estimates.

There are geographical, climatic, cultural, economic and social differences between regions that should be thought about and accounted for in some way before results are simply applied cross-context.

Another question of external validity concerns the choice of regressors and model: if other studies examining the same phenomenon are using a different specification for their regression model, it would be worth at least running the same specification for intercomparability between results in addition to our own specification.

Question iv

Using a simple regression model, we estimate that exposure to one additional day of heat above 35°C increases the mortality rate for an average 70+ year old by .03.

References

OECD. 2015. *Geographic Distribution of Doctors*. https://doi.org/https://doi.org/https://doi.org/10.1787/health_glance-2015-42-en.