SciencesPo
DEPARTMENT OF ECONOMICS

# ScPoEconometrics: Advanced

## Instrumental Variables

Bluebery Planterose
SciencesPo Paris
2023-02-14

# Where Are We At?

## Today

- we will introduce *instrumental variables* (IV)
- To motivate IV, we will look back to London in 1850 and learn about John Snow.
- We will finally introduce the IV estimator formally.

# Setting the Scene

- In chapters 7, 8 and 9 of the book (and the intro course) we talk about the merits of *experimental methods*.
- Randomized Control Trials (RCTs) or *Quasiexperimental* (as good as random) settings allow us to estimate **causal** effects.
- In particular the RCT should be familiar to you.

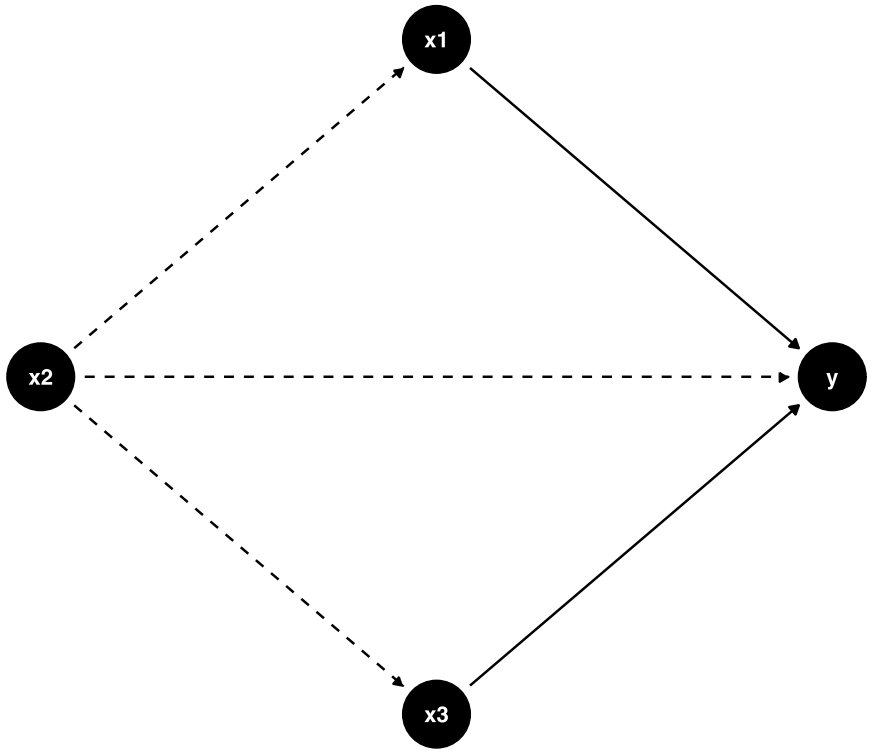
# Setting the Scene

- In chapters 7, 8 and 9 of the book (and the intro course) we talk about the merits of *experimental methods*.
- Randomized Control Trials (RCTs) or *Quasiexperimental* (as good as random) settings allow us to estimate **causal** effects.
- In particular the RCT should be familiar to you.

- If people have some sort of control about getting treatment, there will be *selection*.
- RCTs can break the self-selection of people into treatment by assigning randomly.
- So with experimental data, we have a good solution.
- What about non-experimental data?

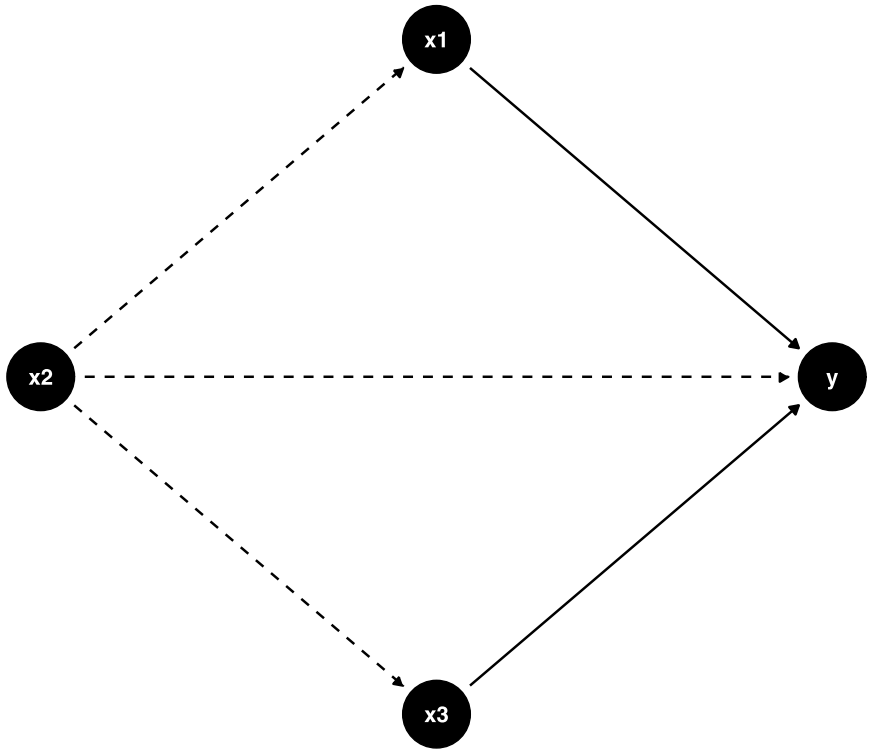
# Non-Experimental Data

- We talked about **omitted variable bias**.
- What if there is correlation between a variable in the error term $u$, $x_2$ say, and our explanatory variable $x_1$?
- We will obtain biased estimates because we cannot separate out what is what: effect of $x_1$, or of $x_2$?
- Remember that this can be so severe that we don't even get the correct sign of an effect.



# Non-Experimental Data

- We talked about **omitted variable bias**.
- What if there is correlation between a variable in the error term $u$, $x_2$ say, and our explanatory variable $x_1$?
- We will obtain biased estimates because we cannot separate out what is what: effect of $x_1$, or of $x_2$?
- Remember that this can be so severe that we don't even get the correct sign of an effect.




**IV** provides a solution to OVB.

# Welcome to London in 1850

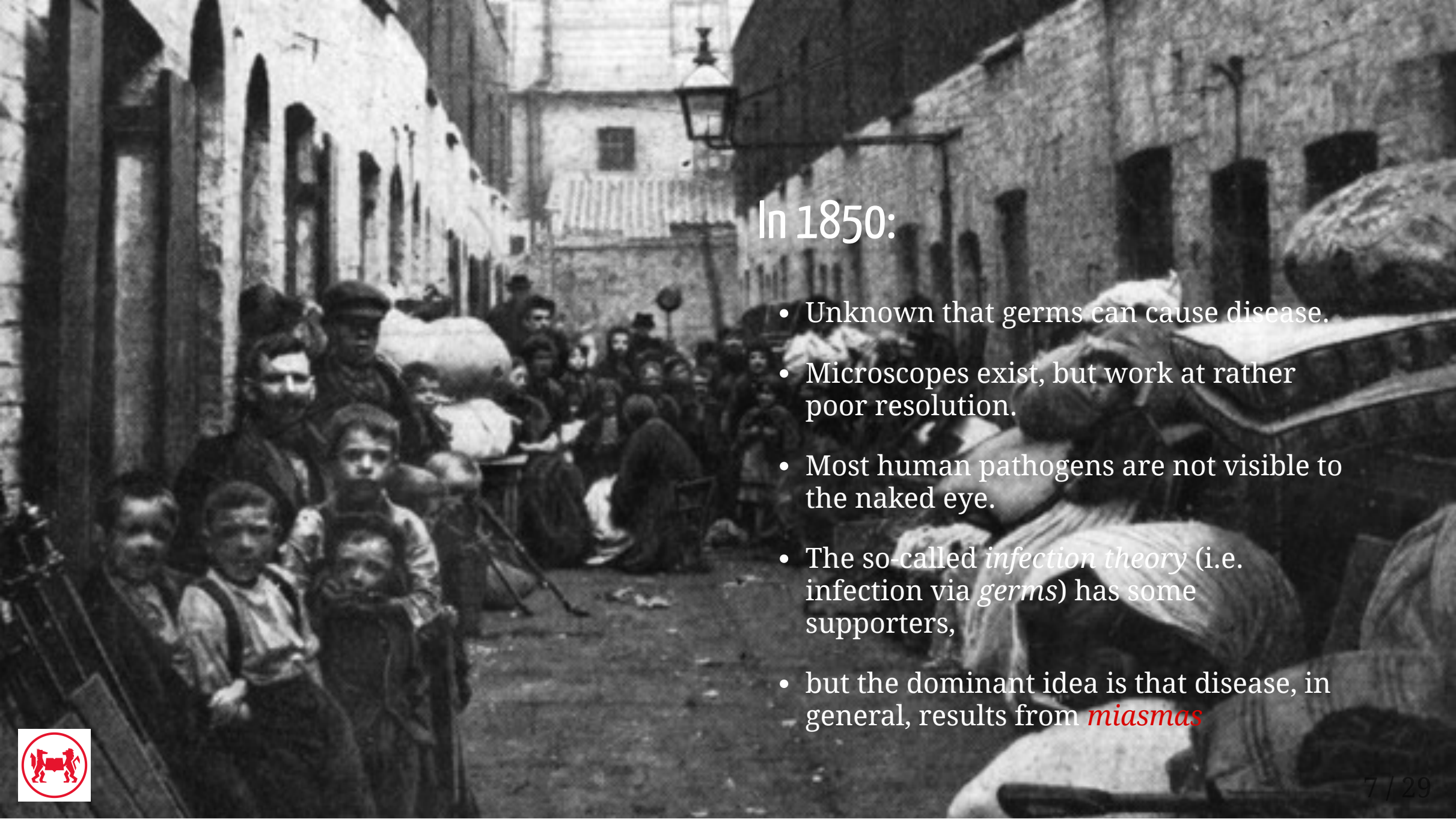
(Slum in Kensington)

# John Snow's (Non) Experiment: Cholera Hits the Town

- John Snow was a physician in London around 1850, when Cholera erupted several times in the City.
- There was a dispute at the time about how the disease is transmitted: via air or via water?


FATHER THAMES INTRODUCING HIS OFFSPRING TO THE FAIR CITY OF LONDON.

*(A Design for a Fresco in the New Houses of Parliament.)*



## In 1850:

- Unknown that germs can cause disease.
- Microscopes exist, but work at rather poor resolution.
- Most human pathogens are not visible to the naked eye.
- The so-called *infection theory* (i.e. infection via *germs*) has some supporters,
- but the dominant idea is that disease, in general, results from *miasmas*

# Let's Go Watch a Movie

# Let's Go Watch a Movie

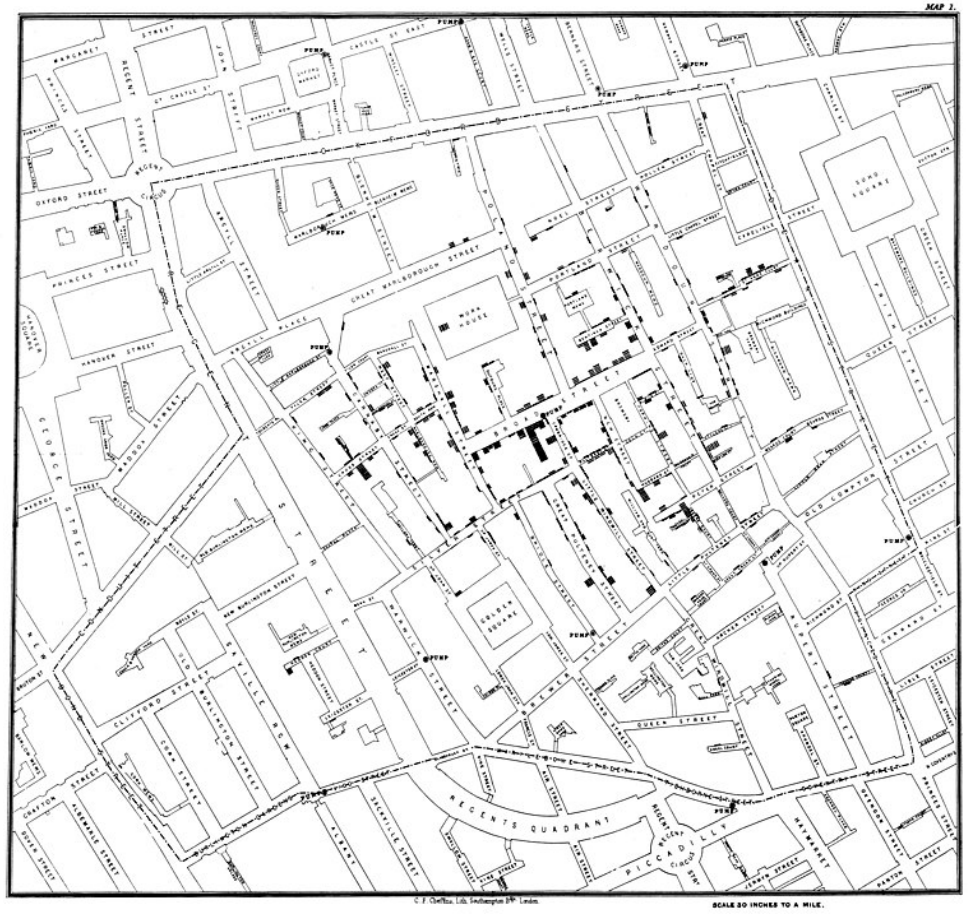
Click here!

# Snow's Detective Work

- Snow collected a lot of data.
- He first mapped the location of dead during the 1854 outbreak.
- This was the notorious *Broadstreet Pump Outbreak*

# Snow's Detective Work

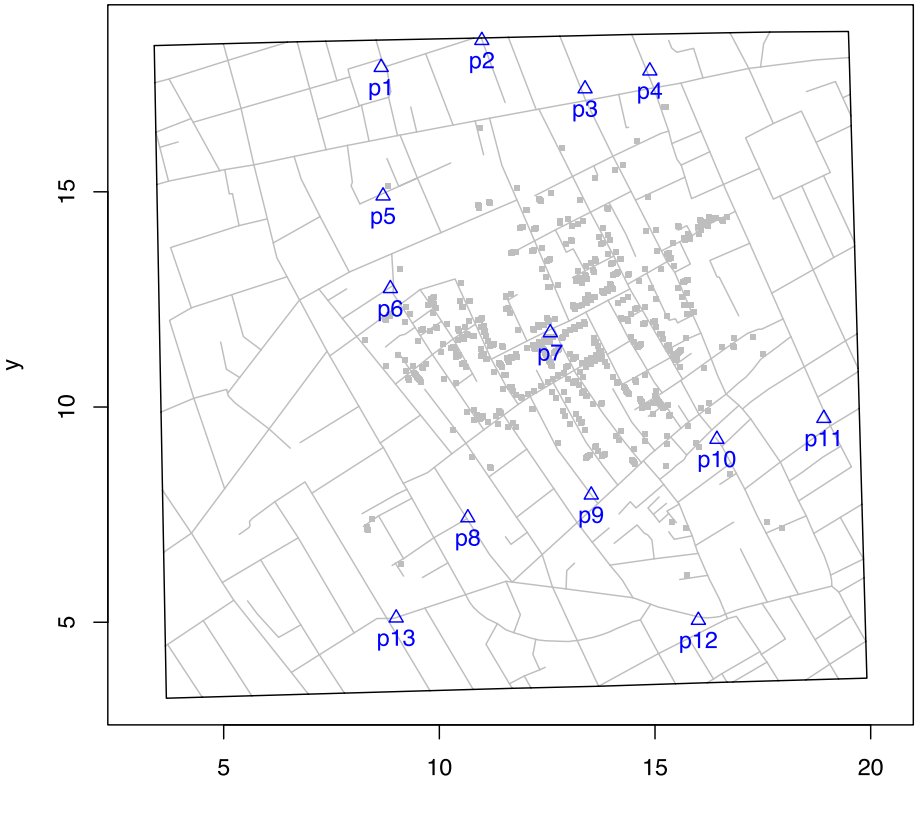
- Snow collected a lot of data.
- He first mapped the location of dead during the 1854 outbreak.
- This was the notorious *Broadstreet Pump Outbreak*



# The `cholera` package

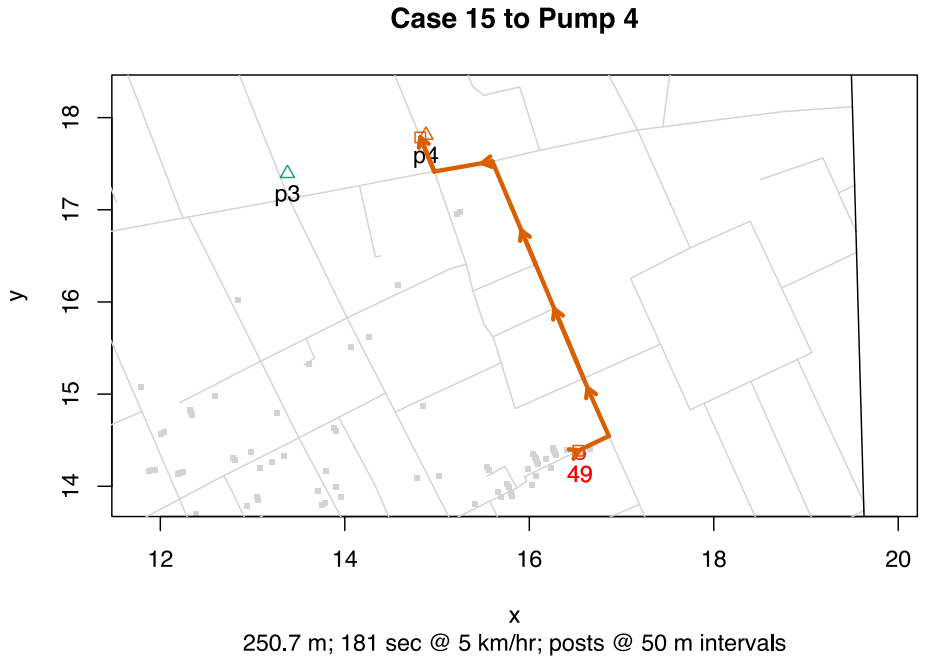
- The `cholera` package has some interesting features.
- For example an R version of Snow's map:

```
cholera::snowMap()
```



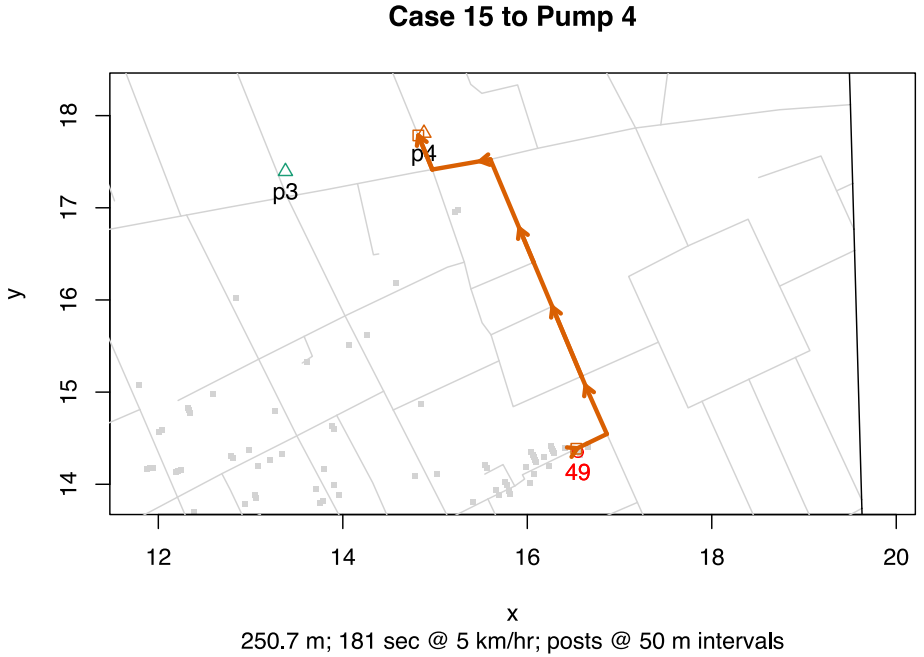
# cholera

...or the walking path of case number 15 in Snow's data:

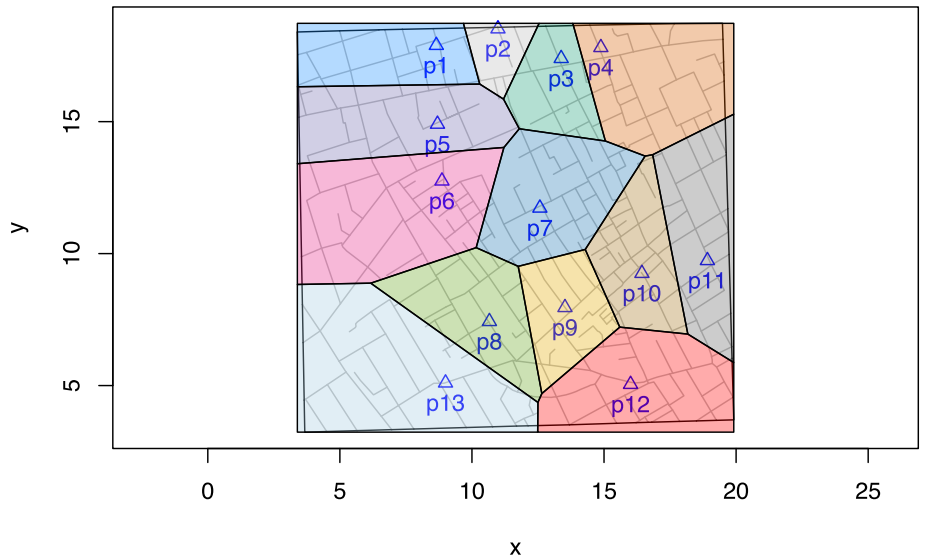


# cholera

...or the walking path of case number 15 in Snow's data:

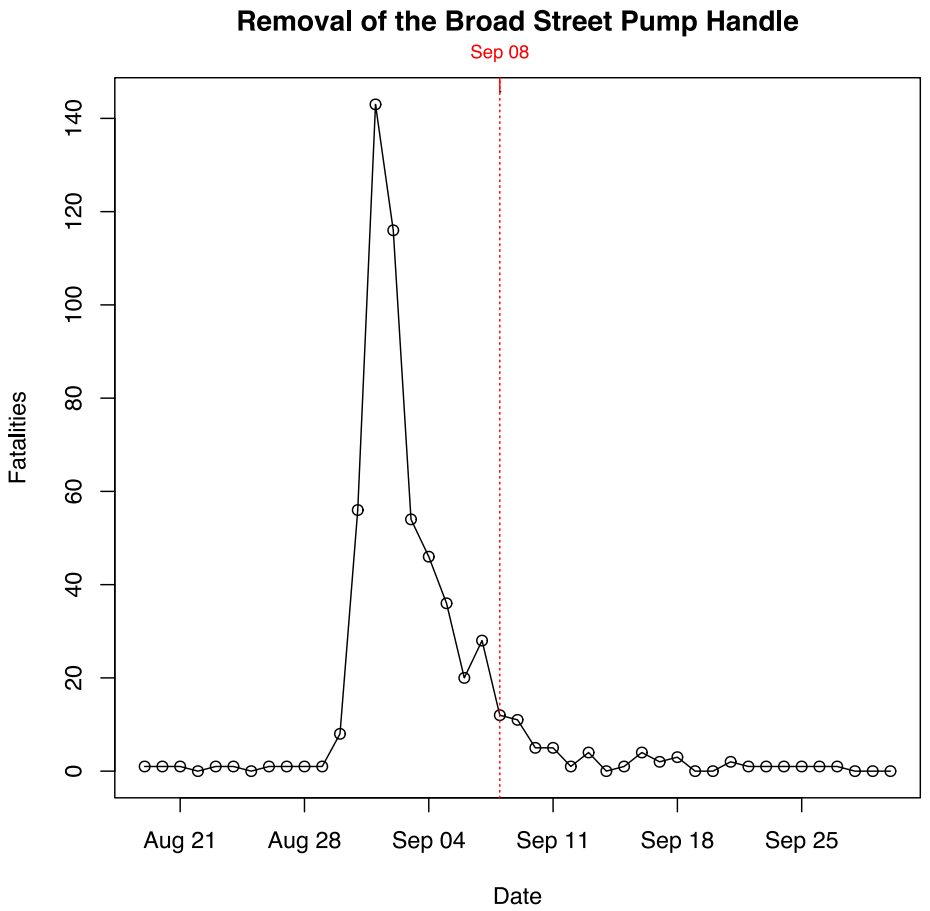



...or estimate Voronoi Polygons for pump neighborhoods:



# Removal of the Broad Street Pump?

- Snow identified the Broad Street Pump as culprit.
- He pleaded to have its handle removed.
- He was sceptical this was the reason the epidemic ended.



# Mapping London's Water Supply

- Water supply came from the River Thames
- Different supply companies had different intake points
- Southwark and Vauxhall water companies took in water beneath a major sewage discharge.
- Lambeth water did not.

# Snow's conclusion

- Snow collected the following data:

| area | numhouses | deaths | death1000 |
|---|---|---|---|
| Southwark and Vauxhall | 40046 | 1263 | 315 |
| Lambeth | 26107 | 98 | 37 |
| Rest of London | 256423 | 1422 | 59 |

- And concluded

> that if Southwark and Vauxhall water companies had moved their water intakes upstream to where Lambeth water was taking in their supply, roughly 1,000 lives could have been saved.

- For proponents of the miasma theory, this was still not evidence enough, because there were also many factors that led to poor air quality in those areas.

# We Need A Model.

Because: *It takes a model to beat a model*

# Snow's Model of Cholera Transmission

- Suppose that $c_i$ takes the value 1 if individual $i$ dies of cholera, 0 else.
- Let $w_i = 1$ mean that $i$'s water supply is impure and $w_i = 0$ vice versa. Water purity is assessed with a technology that cannot detect small microbes.
- Collect in $u_i$ all unobservable factors that impact $i$'s likelihood of dying from the disease: whether $i$ is poor, where exactly they reside, whether there is bad air quality in $i$'s surrounding, and other invidivual characteristics which impact the outcome (like genetic setup of $i$).

# Snow's Model of Cholera Transmission

- Suppose that $c_i$ takes the value 1 if individual $i$ dies of cholera, 0 else.
- Let $w_i = 1$ mean that $i$'s water supply is impure and $w_i = 0$ vice versa. Water purity is assessed with a technology that cannot detect small microbes.
- Collect in $u_i$ all unobservable factors that impact $i$'s likelihood of dying from the disease: whether $i$ is poor, where exactly they reside, whether there is bad air quality in $i$'s surrounding, and other invidivual characteristics which impact the outcome (like genetic setup of $i$).

We can write:

$$c_i = \alpha + \delta w_i + u_i$$

# Doing the Simple Thing is always right?

- John Snow could have used his data and assess the correlation between drinking pure water and cholera incidence.
- measure $Cor(c_i, w_i)$
- Suppose $Cor(c_i, w_i) \approx 0.5$. Does that prove the infection theory?

# Doing the Simple Thing is always right?

- John Snow could have used his data and assess the correlation between drinking pure water and cholera incidence.
- measure $Cor(c_i, w_i)$
- Suppose $Cor(c_i, w_i) \approx 0.5$. Does that prove the infection theory?

Note quite. Angus Deaton says:

> The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the 'poison miasmas' that were then thought to be the cause of cholera.

☹

# The Simple Thing

- It does not make sense to compare someone who drinks pure water with someone with impure water.
- because *all else is not equal*: pure water is correlated with being poor, living in bad area, bad air quality and so on - all factors that we encounter in $u_i$.
- This violates the crucial orthogonality assumption for valid OLS estimates, $E[u_i|w_i] = 0$ in this context.
- Another way to say this, is that $Cov(w_i, u_i) \neq 0$, implying that $w_i$ is *endogenous*.
- There are factors in $u_i$ that affect both $w_i$ and $c_i$

# Snow's Model and Some Algebra

Remember our simple model:

$$c_i = \alpha + \delta + u_i$$

Now let's condition on both values of $w$:

$$E[c_i|w_i = 1] = \alpha + \delta + E[u_i|w_i = 1]$$
$$E[c_i|w_i = 0] = \alpha + \quad + E[u_i|w_i = 0]$$

# Snow's Model and Some Algebra

Remember our simple model:

$$c_i = \alpha + \delta + u_i$$

Now let's condition on both values of $w$:

$$E[c_i|w_i = 1] = \alpha + \delta + E[u_i|w_i = 1]$$
$$E[c_i|w_i = 0] = \alpha + \quad + E[u_i|w_i = 0]$$

Now substract one line from the other:

$$E[c_i|w_i = 1] - E[c_i|w_i = 0] = \delta + \{E[u_i|w_i = 1] - E[u_i|w_i = 0]\}$$

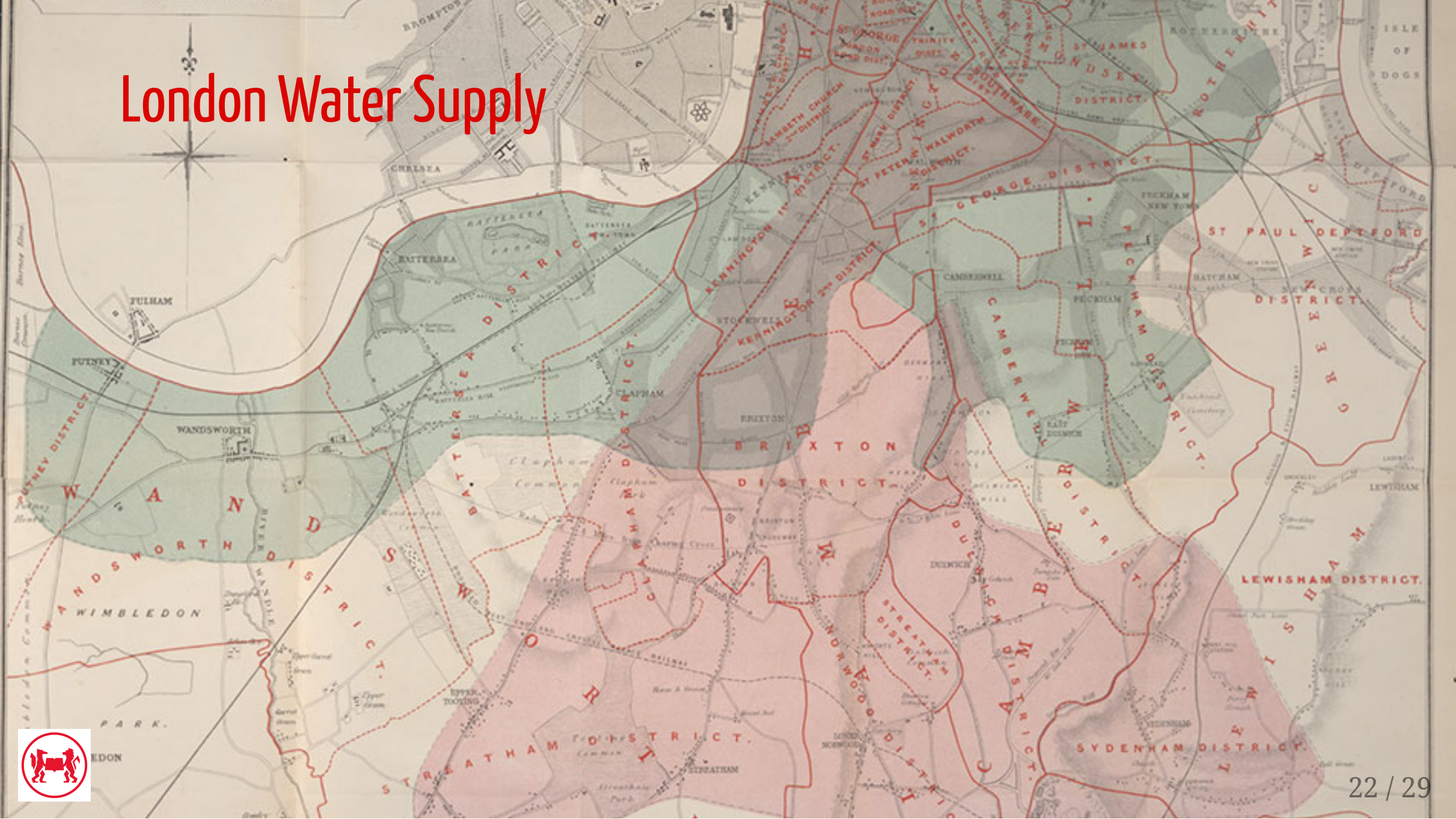
- The last term $\{E[u_i|w_i = 1] - E[u_i|w_i = 0]\}$ is not equal to zero (by what Deaton said!)
- A regression estimate for $\delta$ would be biased by that quantity.

# The IV Estimator

# John Snow Says

> […] the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. […] The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.

# London Water Supply



# Proposing an IV

- Snow is proposing an **instrumental variable** $z_i$, the *identity of the water supplying company* to household $i$:

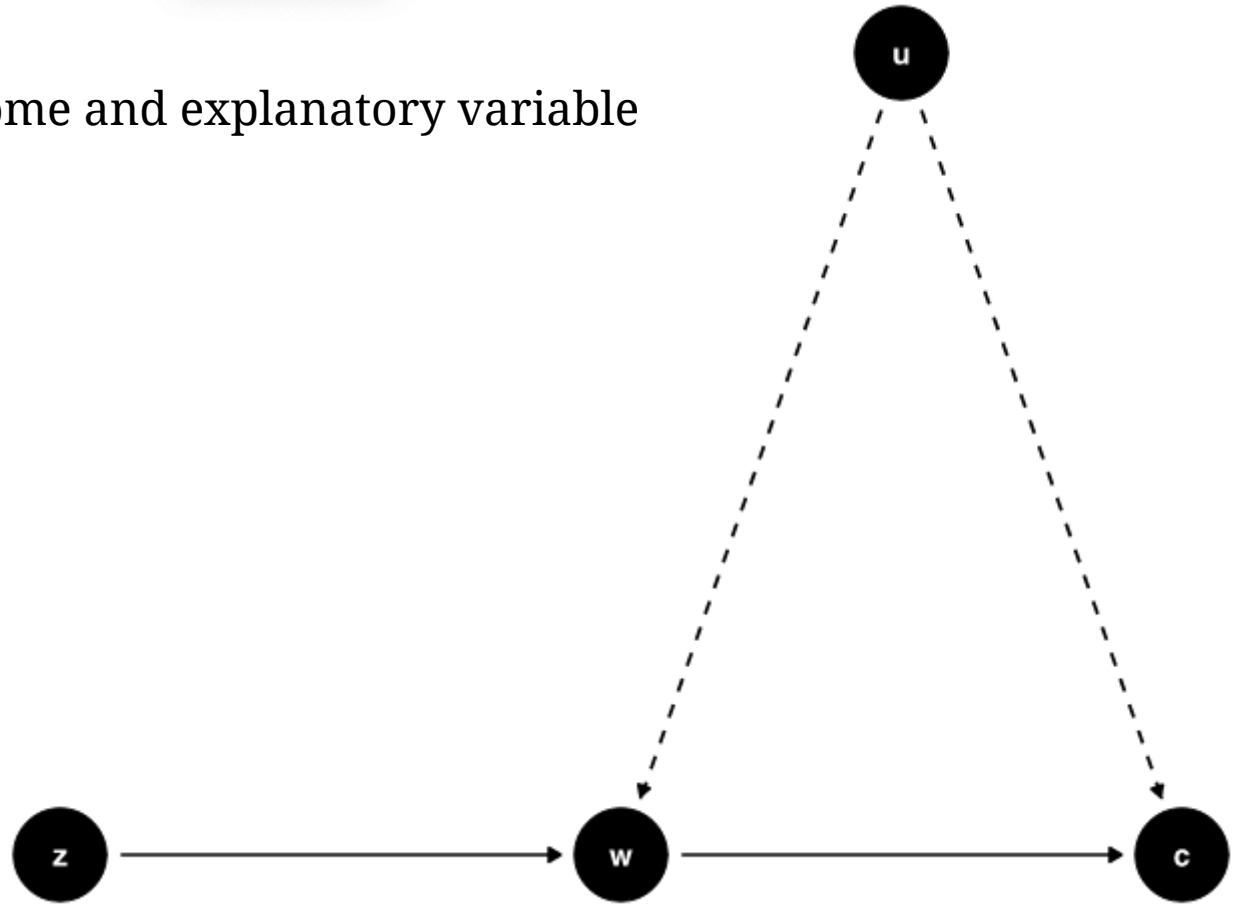
More formally, let's define the instrument as follows:

$$z_i = \begin{cases} 1 & \text{if water supplied by Lambeth} \\ 0 & \text{if water supplied by Southwark or Vauxhall.} \end{cases}$$

- $z_i$ is highly correlated with the water purity $w_i$.
- However, it seems to be uncorrelated with all the other factors in $u_i$, which worried us before: Water supply was decided years before, and now houses on the same street have different suppliers!

# Simple IV in a DAG

- $u$ affects both outcome and explanatory variable



# Defining Snow's IV Formally

# Defining Snow's IV Formally

Here are the conditions for a valid instrument:

1. **Relevance** or **First Stage**: Water purity is indeed a function of supplier identity. We want that

$$E[w_i|z_i = 1] \neq E[w_i|z_i = 0]$$

   i.e. the average water purity differs across suppliers. We can *verify* this condition with observational data. We want this effect to be reliably causal.

# Defining Snow's IV Formally

Here are the conditions for a valid instrument:

1. **Relevance** or **First Stage**: Water purity is indeed a function of supplier identity. We want that

$$E[w_i|z_i = 1] \neq E[w_i|z_i = 0]$$

   i.e. the average water purity differs across suppliers. We can *verify* this condition with observational data. We want this effect to be reliably causal.

2. **Independence**: Whether a household has $z_i = 1$ or $z_i = 0$ is unrelated to $u$, hence *as good as random*. Whether we condition $u$ on certain values of $z$ does not change the result - we want

$$E[u_i|z_i = 1] = E[u_i|z_i = 0].$$

# Defining Snow's IV Formally

Here are the conditions for a valid instrument:

1. **Relevance** or **First Stage**: Water purity is indeed a function of supplier identity. We want that

$$E[w_i|z_i = 1] \neq E[w_i|z_i = 0]$$

   i.e. the average water purity differs across suppliers. We can *verify* this condition with observational data. We want this effect to be reliably causal.

2. **Independence**: Whether a household has $z_i = 1$ or $z_i = 0$ is unrelated to $u$, hence *as good as random*. Whether we condition $u$ on certain values of $z$ does not change the result - we want

$$E[u_i|z_i = 1] = E[u_i|z_i = 0].$$

3. **Excludability** the instrument should affect the outcome $c$ *only* through the specified channel (i.e. via water purity $w$), and nothing else.

# Defining the IV Estimator

We are now ready to define a simple IV estimator. Like before, let's condition on the values of $z$:

$$E[c_i|z_i = 1] = \alpha + \delta E[w_i|z_i = 1] + E[u_i|z_i = 1]$$
$$E[c_i|z_i = 0] = \alpha + \delta E[w_i|z_i = 0] + E[u_i|z_i = 0]$$

which upon differencing both lines gives

$$E[c_i|z_i = 1] - E[c_i|z_i = 0] = \delta \{E[w_i|z_i = 1] - E[w_i|z_i = 0]\} + \underbrace{\{E[u_i|z_i = 1] - E[u_i|z_i = 0]\}}_{=0 \text{ by Exogeneity Assumption}}$$

# Defining the IV Estimator

We are now ready to define a simple IV estimator. Like before, let's condition on the values of $z$:

$$E[c_i|z_i = 1] = \alpha + \delta E[w_i|z_i = 1] + E[u_i|z_i = 1]$$
$$E[c_i|z_i = 0] = \alpha + \delta E[w_i|z_i = 0] + E[u_i|z_i = 0]$$

which upon differencing both lines gives

$$E[c_i|z_i = 1] - E[c_i|z_i = 0] = \delta \{E[w_i|z_i = 1] - E[w_i|z_i = 0]\} + \underbrace{\{E[u_i|z_i = 1] - E[u_i|z_i = 0]\}}_{=0 \text{ by Exogeneity Assumption}}$$

- Finally, if the IV is *relevant*, i.e. $E[w_i|z_i = 1] - E[w_i|z_i = 0] \neq 0$:

$$\delta = \frac{E[c_i|z_i = 1] - E[c_i|z_i = 0]}{E[w_i|z_i = 1] - E[w_i|z_i = 0]} (\#eq : IV)$$

# Special Case: Wald Estimator

Let's say that $x \mapsto y$ means that $x$ is an estimate for $y$:

1. $\overline{c}_1 \mapsto E[c_i | z_i = 1]$: the proportion of households supplied by Lambeth with cholera.
2. $\overline{w}_1 \mapsto E[w_i | z_i = 1]$: the proportion of households supplied by Lambeth with bad water.
3. $\overline{c}_0 \mapsto E[c_i | z_i = 0]$: the proportion of households not supplied by Lambeth with cholera.
4. $\overline{w}_0 \mapsto E[w_i | z_i = 0]$: the proportion of households not supplied by Lambeth with bad water.

The estimator would then be

$$\hat{\delta} = \frac{\overline{c}_1 - \overline{c}_0}{\overline{w}_1 - \overline{w}_0}$$

In this special case where all involved variables $c$, $w$, $z$ are binary, the estimator is called the *Wald estimator*.

**Summary**: IVs are a powerful tool to establish causality in contexts with observational data only and where we are concerned that the conditional mean assumption $E[u_i|x_i] = 0$ is violated, hence, we cannot say *all else equal, as $x$ changes, $y$ changes like this and that*. Then we say that $x$ is *endogenous*. The key features of IV $z$ are that

1. $z$ is *relevant* for $x$. For example, in a simple regression of $z$ on $x$, we want $z$ to have considerable predictive power. We can *test* this condition in data.
2. We need a theory according to which is *reasonable* to assume that $z$ is *unrelated* to other unobservable factors that might impact the outcome. Hence, $z$ is *exogenous* to $u$, or $E[u|z] = 0$. This is an **assumption** (i.e. we can not test this with data).

# END

| | |
|---|---|
| | bluebery.planterose@sciencespo.fr |
| | Original Slides from Florian Oswald |
| | Book |
| | @ScPoEcon |
| | @ScPoEcon |