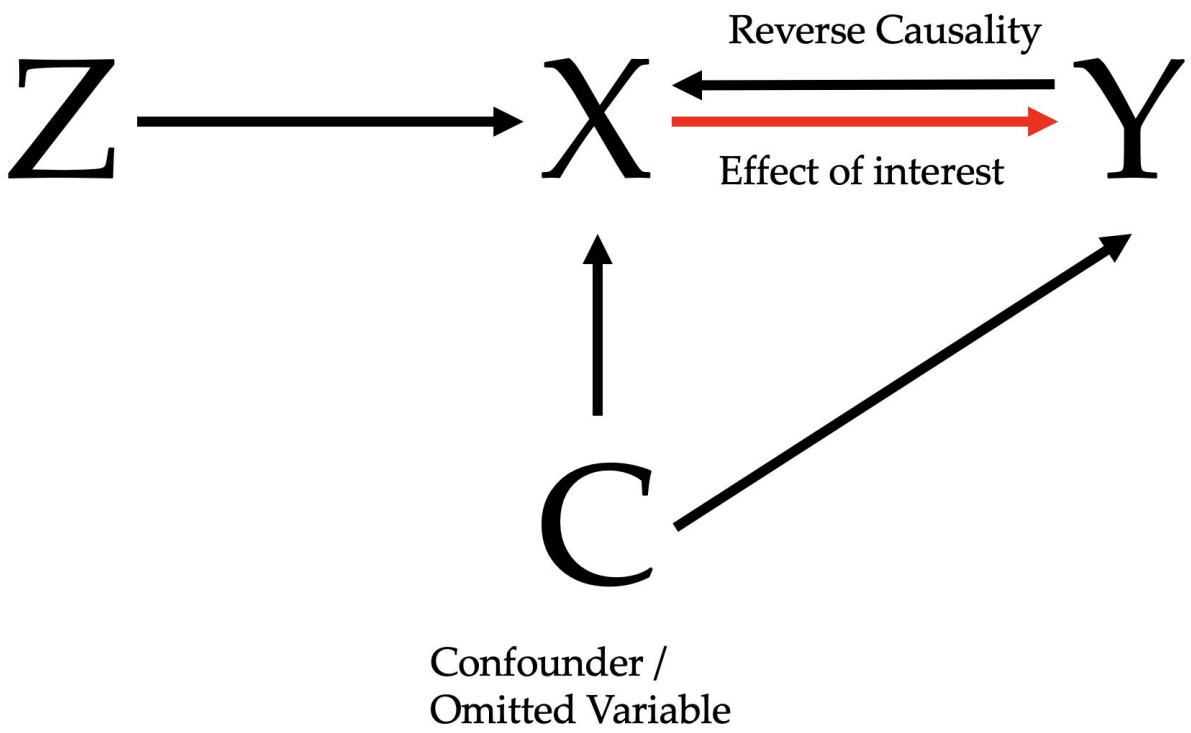
On Counterfactuals, Clones, and Cool (?) Regression Stuff

https://pollev.com/jonathanold608

Roadmap

- Your questions
- Recap: Selection bias and counterfactuals
- A closer look: how RCTs solve selection bias problem
- A small intro to regression and OLS

Recap - Selection Bias



Potential outcomes

Y_{0i} = Outcome of individual i, with "status" 0

Y_{1i} = Outcome of individual i, with "status" 1

Counterfactual outcomes!

- "Status" can be anything:
  - Treatment assignment
  - Actual treatment
  - Having 0 or 1 children
  - Drinking expensive whiskey or not

BUT: We **NEVER** observe an individual at more than one status at the same time! ➔ The fundamental problem of unobservability!

Does giving you an iPad improve your test scores? Potential outcomes

Difference in Potential Outcomes

Group

Potential outcome if i has iPad

$$E[Score_{1i} - Score_{0i} | iPad_i = 0]$$

$$= E[Score_{1i} | iPad_i = 0] - E[Score_{0i} | iPad_i = 0]$$

$$= E[Score_{1i} | iPad_i = 0] - E[Score_{0i} | iPad_i = 0]$$

$$+ E[Score_{1i} | iPad_i = 1] - E[Score_{1i} | iPad_i = 1]$$

$$= E[Score_{1i} | iPad_i = 1] - E[Score_{0i} | iPad_i = 0]$$

$$+ E[Score_{1i} | iPad_i = 0] - E[Score_{1i} | iPad_i = 1]$$

Treatment effect =

Observed difference +

Selection Bias

How RCTs solve selection bias

$$= E[\mathbf{Score_{1i}} | \mathbf{iPad_i=1}] - E[\mathrm{Score_{0i}} | \mathrm{iPad_i=0}]$$

$$\cancel{+ E[\mathrm{Score_{1i}} | \mathrm{iPad_i=0}] - E[\mathrm{Score_{0i}} | \mathrm{iPad_i=1}]}$$

Random assignment of X-variable!

How RCTs solve selection bias

$$E[IQ_{1i} - IQ_{0i} | \text{one more child}_i]$$

$$= E[IQ_{1i} | \text{one more child}_i] - \mathbf{E[IQ_{0i} | \text{same children}_i]}$$

$$\cancel{+ \mathbf{E[IQ_{0i} | \text{same children}_i]} - E[IQ_{0i} | \text{one more child}_i]}$$

Random assignment of X-variable!

Epic RCTs of History

On the 20th of May 1747, **I selected twelve patients in the scurvy**, on board the Salisbury at sea. **Their cases were as similar as I could have them**. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times light puddings, boiled biscuit with sugar, etc., and for supper, barley and raisins, rice and currants, sago and wine or the like.

Two were ordered each a quart of cyder a day. **Two others** took twenty-five drops of elixir vitriol three times a day … **Two others** took two spoonfuls of vinegar three times a day … **Two** of the worst patients were put on a course of sea-water … **Two others had each two oranges and one lemon given them every day** … **The two remaining patients,** took … an electary recommended by a hospital surgeon …

The consequence was, **that the most sudden and visible good effects were perceived from the use of oranges and lemons**; one of those who had taken them, being at the end of six days fit for duty … The other was the best recovered of any in his condition; and … was appointed to attend the rest of the sick. Next to the oranges, I thought the cyderhad the best effects …"

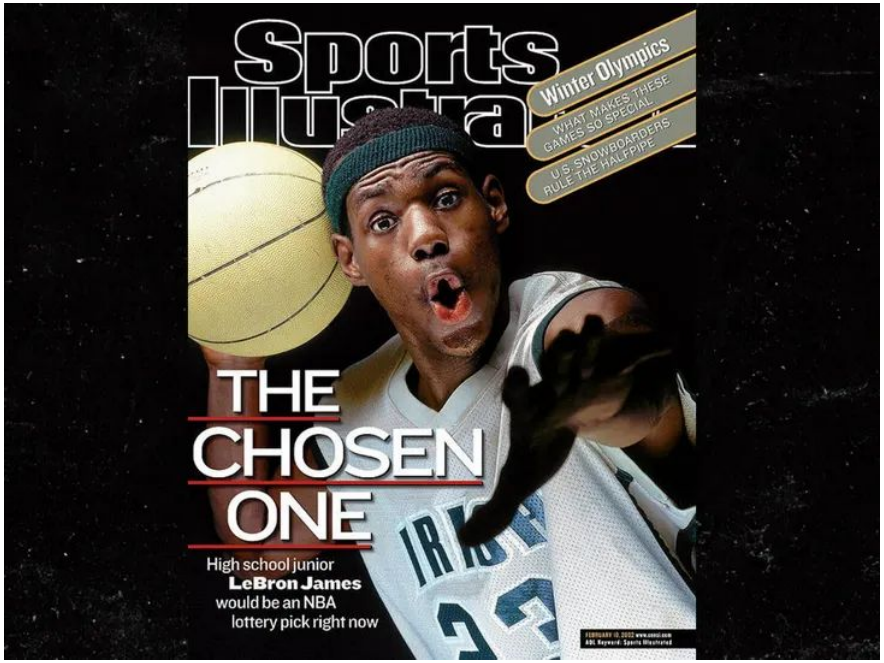
–Dr.James Lind "Treatise on Scurvy" (1753)

So... Do RCTs solve everything?

- Recall: We **NEVER** observe an observation unit at both status at the same time
- Weird things can happen:
  - Hawthorne effect (being aware of being in the study)
  - Pygmalion effect (self-fulfilling prophecies)
  - John Henry effect (control group works harder than treatment)
  - Spillover effects (e.g., from treatment to control)
- It all comes down to: Is your “control group” a **VALID COUNTERFACTUAL** for the “treatment group”

Now... What is a regression?

The name comes from: **regression to the mean**



What we mean by it

A specific **model of the world**

relating two (or more) variables

(in a **linear** fashion - for linear regressions)

$$\mathbf{y_i = a + bx_i + e_i}$$

Model of the world

Are models right?

Are models useful?

Linearity

Which relationships in nature are linear?
Is linearity useful?

We can be more flexible:

$$y_i = a + bx_i^2 + e_i$$

$$\log(y_i) = a + bx_i + e_i$$

$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3(x_{1i} * x_{2i}) + e_i$$

But linearity is a **strong** assumption:
https://www.nature.com/articles/431525a

Other assumptions

Does the model have to be complete?

Well… it depends if we want to interpret it causally

Maths? Estimating regressions with OLS

Published in Towards Data Science

You have **2** free member-only stories left this month. Sign up for Medium and get an extra one

Valentina Alto

Aug 17, 2019 · 4 min read · Member-only · Listen

Understanding the OLS method for Simple Linear Regression

Linear Regression is the family of algorithms employed in supervised machine learning tasks (to learn more about supervised learning, you can read my former article here). Knowing that supervised ML tasks are normally divided into classification and regression, we can collocate Linear Regression algorithms in the latter category. It differs from classification because of the nature of the target variable: in classification, the target is a categorical value ('yes/no', 'red/blue/green', 'spam/not spam'…); on the other hand, regression involves numerical, continuous values as target, hence the algorithm will be asked to

Maths? Estimating regressions with OLS

(Blackboard)