Validity

Validity

The degree to which something measures what it purports to measure.

Where do challenges to validity come from?

1. P-hacking, dishonesty, and difficult research decisions

Where do challenges to validity come from?

1. P-hacking, dishonesty, and difficult research decisions
2. Reality is complicated and science is hard

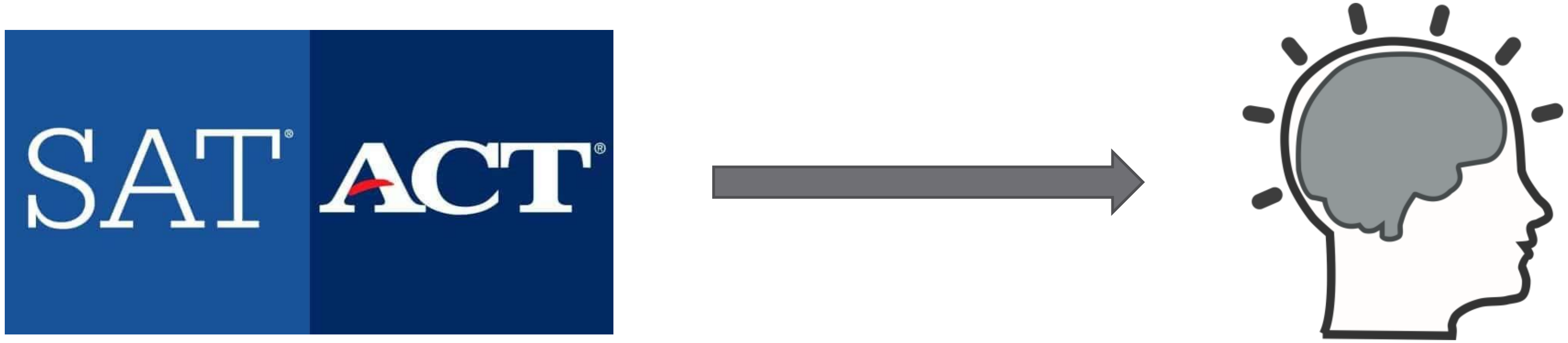
Types of Validity

Construct Validity

How well an operationalization represents a theory.

Construct Validity

How well an operationalization represents a theory.



Construct Validity

How well an operationalization represents a theory.

SAT® ACT®

QBR

Internal Validity

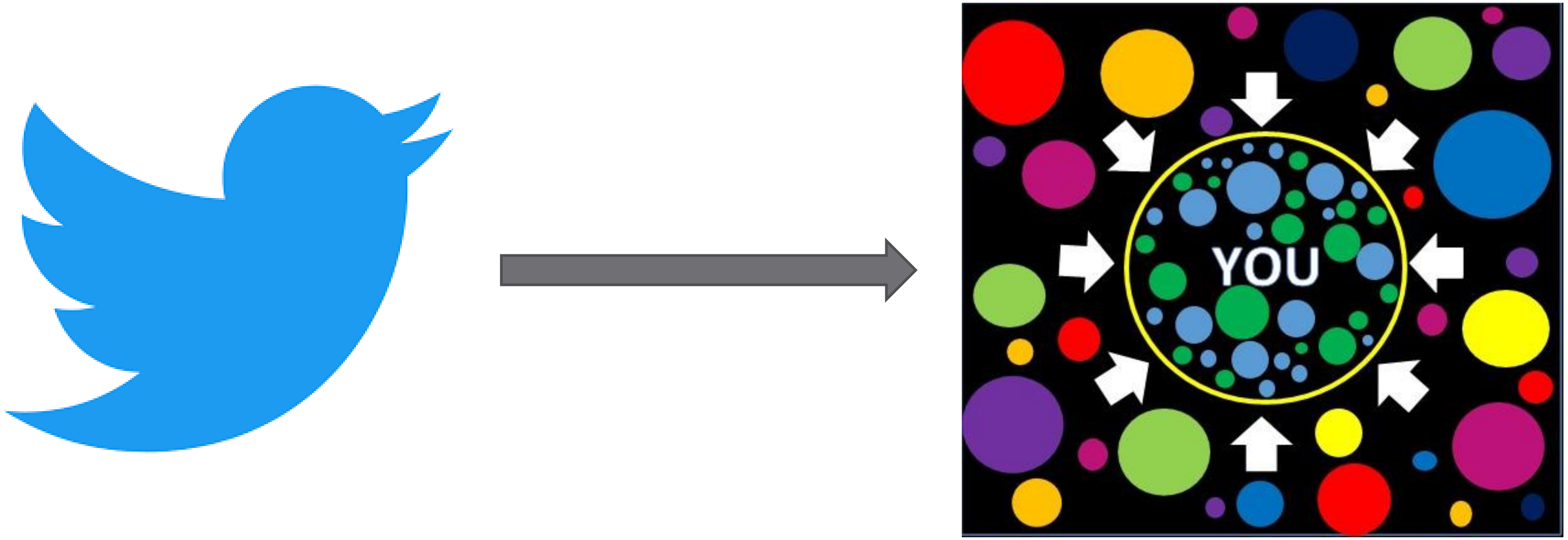
The validity of your research design.

- Are the conclusions of your study true?
- Can the effect on your dependent variable be correctly attributed to your independent variable?

Internal Validity

The validity of your research design.

- Are the conclusions of your study true?
- Can the effect on your dependent variable be correctly attributed to your independent variable?



External Validity

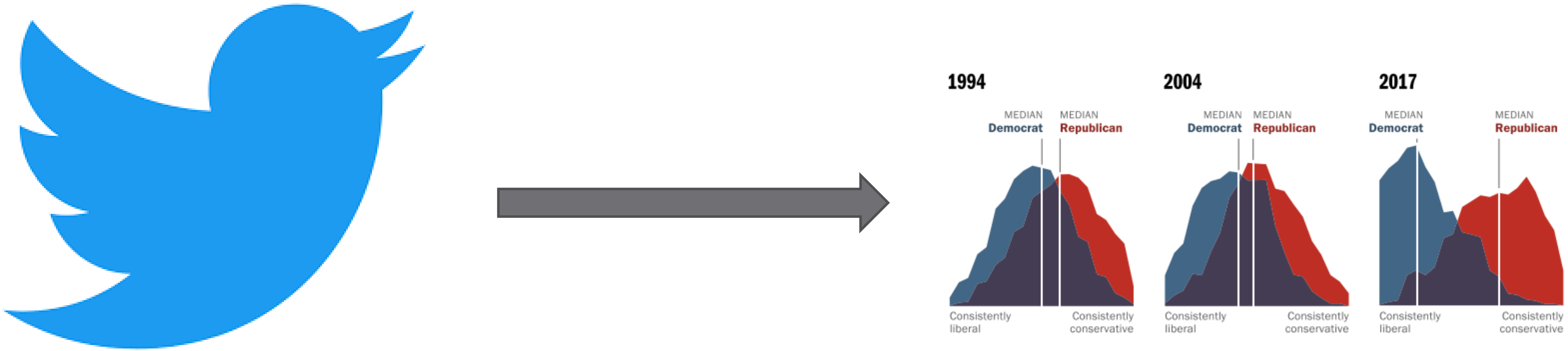
The extent to which your conclusions can be generalized.

- Population
- Setting
- Time

External Validity

The extent to which your conclusions can be generalized.

- Population
- Setting
- Time

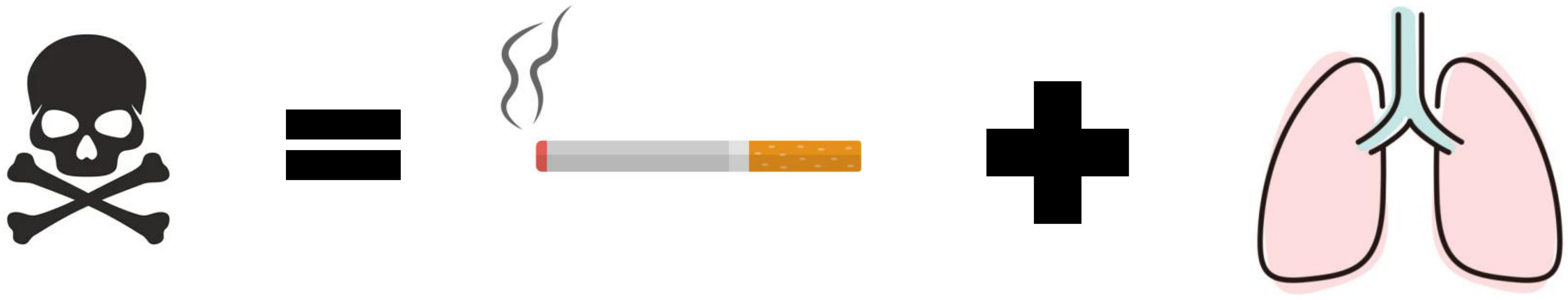


Threats to Validity

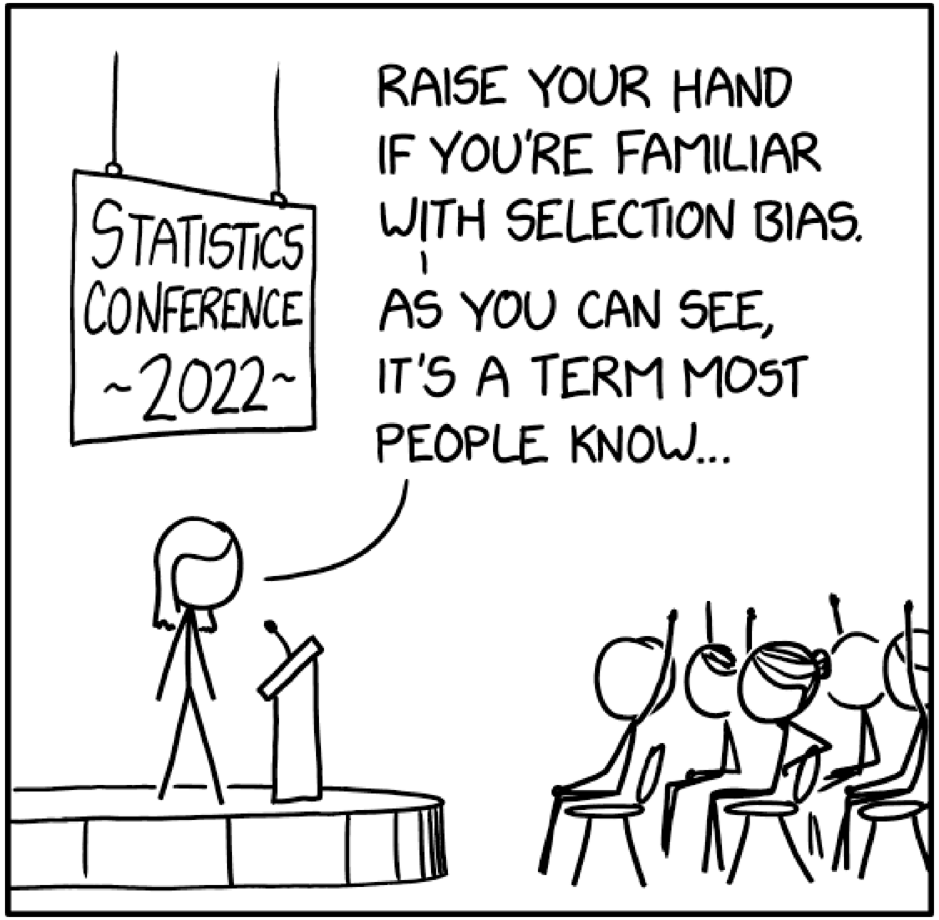
Bad Controls and Post-Treatment Bias

🚬 = ☠ ?

Bad Controls and Post-Treatment Bias

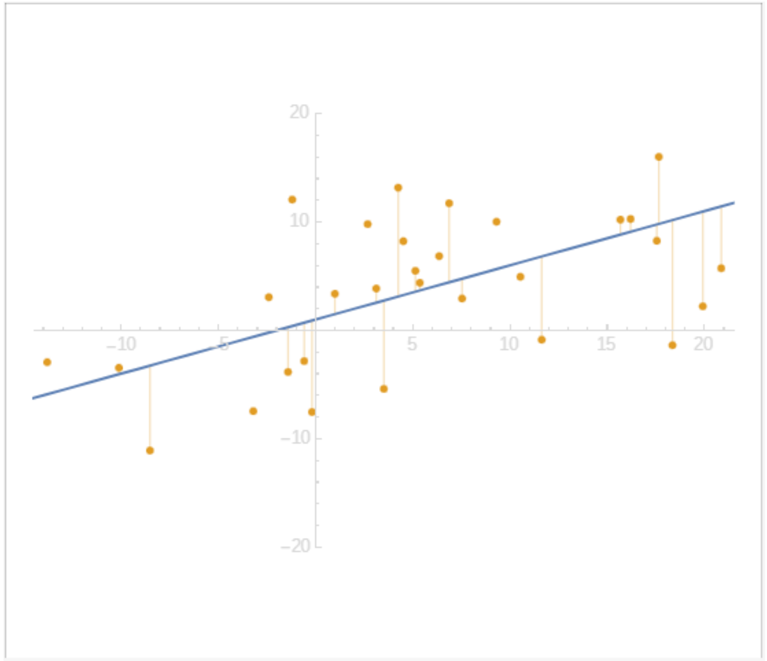


Selection Bias


STATISTICS CONFERENCE ~2022~
RAISE YOUR HAND IF YOU'RE FAMILIAR WITH SELECTION BIAS.
AS YOU CAN SEE, IT'S A TERM MOST PEOPLE KNOW...

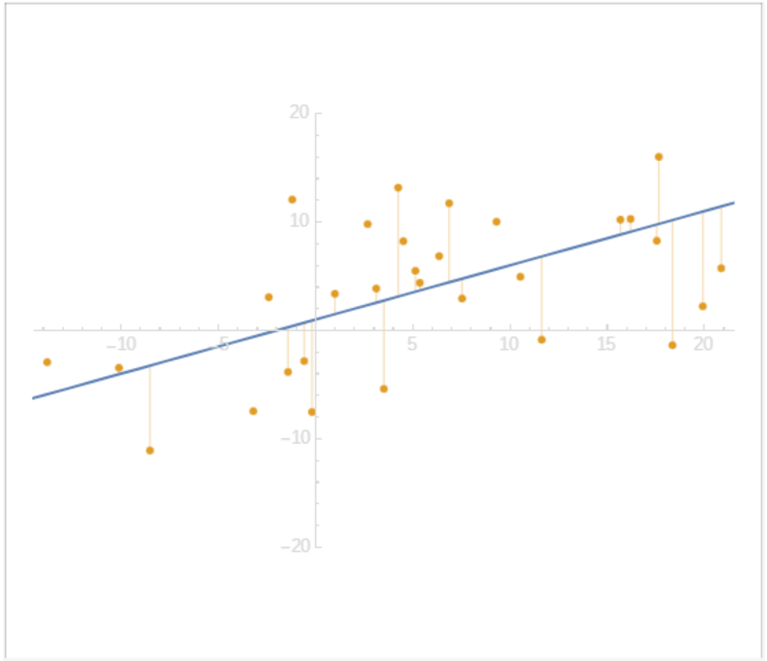
Endogeneity

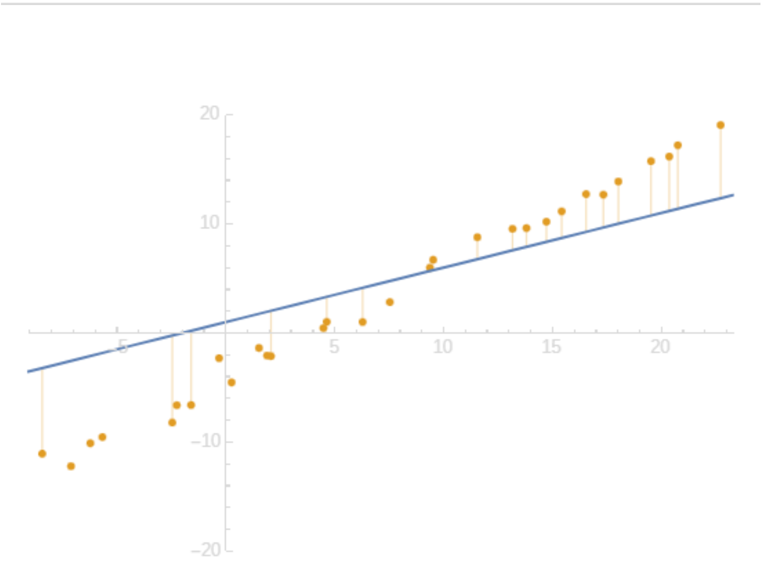
Correlation between the error term and your explanatory variables.



Endogeneity

Correlation between the error term and your explanatory variables.





A quick review

- You want to see if democracy increases or deceases the probability of civil war. Should you control for GDP?

A quick review

- You want to see if democracy increases or deceases the probability of civil war. Should you control for GDP?
- You want to see how someone's qualifications impacts their salary. Should you control for race?

A quick review

- You want to see if democracy increases or deceases the probability of civil war. Should you control for GDP?
- You want to see how someone’s qualifications impacts their salary. Should you control for race?
- You want to measure someone's approval of Trump so you use sentiment analysis on social media statements containing the word ‘Trump”. What are some validity issues with this?

Endogeneity: Omitted variables

True DGP:

$$y_i = \alpha + \beta x_i + \delta z_i + u_i$$

Endogeneity: Omitted variables

True DGP:

$$y_i = \alpha + \beta x_i + \delta z_i + u_i$$

Model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Endogeneity: Omitted variables

True DGP:

$$y_i = \alpha + \beta x_i + \delta z_i + u_i$$

Model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Because Z is not included in our model, it is absorbed by the error term:

$$\varepsilon_i = \delta z_i + u_i$$

Endogeneity: Omitted variables

True DGP:

$$y_i = \alpha + \beta x_i + \delta z_i + u_i$$

Model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Because Z is not included in our model, it is absorbed by the error term:

$$\varepsilon_i = \delta z_i + u_i$$

Thus, if $Corr(x, z) \neq 0$ and δ ≠ 0 then $Corr(x, \varepsilon) \neq 0$

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

The true model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

The true model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Substitute our true x_i for our observed value:

$$y_i = \alpha + \beta \hat{x}_i - \beta \epsilon_i + \varepsilon_i$$

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

The true model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Substitute our true x_i for our observed value:

$$y_i = \alpha + \beta \hat{x}_i - \beta \epsilon_i + \varepsilon_i$$

Collect our error terms into a single term:

$$u_i = \beta \epsilon_i + \varepsilon_i$$

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

The true model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Substitute our true x_i for our observed value:

$$y_i = \alpha + \beta \hat{x}_i - \beta \epsilon_i + \varepsilon_i$$

Collect our error terms into a single term:

$$u_i = \beta \epsilon_i + \varepsilon_i$$

Thus, your model is:

$$y_i = \alpha + \beta \hat{x}_i + u_i$$

Endogeneity: Measurement Error

You observe $\widehat{x_i} = x_i + \epsilon_i$, where x_i is the true value and ϵ_i is random noise.

The true model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Substitute our true x_i for our observed value:

$$y_i = \alpha + \beta \hat{x}_i - \beta \epsilon_i + \varepsilon_i$$

Collect our error terms into a single term:

$$u_i = \beta \epsilon_i + \varepsilon_i$$

Thus, your model is:

$$y_i = \alpha + \beta \hat{x}_i + u_i$$

Because both $\hat{x}_i$ and u_i are a function of ϵ_i, they are correlated

Endogeneity: Simultaneity

- $\mathrm{X} \rightarrow Y$
- $\mathrm{Y} \rightarrow X$

Endogeneity: Simultaneity

- $\mathrm{X} \rightarrow Y$
- $\mathrm{Y} \rightarrow X$

Assume:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Endogeneity: Simultaneity

- $\mathrm{X} \rightarrow Y$
- $\mathrm{Y} \rightarrow X$

Assume:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

If X is a function of Y, and Y is a function of ε_i, then X is necessarily correlated with ε_i

Some examples

- COVID restrictions and COVID death rates in a state
- Ideology measured via text analysis used as the independent variable to predict sentiment measured via text.
- Parenting, genetics, gender, sex, and environment

(Partial) Solutions

Random Assignment

Q: Imagine you are testing a vaccine. You take a completely random sample of 10,000 people across the world. Half are randomly assigned to the treatment group and half are randomly assigned to the placebo group. What other variables should you control for?

Random Assignment

Q: Imagine you are testing a vaccine. You take a completely random sample of 10,000 people across the world. Half are randomly assigned to the treatment group and half are randomly assigned to the placebo group. What other variables should you control for?

A: None.

Natural/Quasi Experiments

- Situations that arise naturally and create “as if random” assignment of a treatment.
  - Mariel Boatlift on the labor market
  - Media market borders
  - Close elections

Fixed and Random Effects

- Entity level dummy variables that allows the intercept to vary across observations.
- Reduces omitted variable bias.
- Often used with time and geolocation.

Instrumental Variables

- This is the most common solution you will find for endogeneity. We won’t be covering it because it’s a bit complicated and doesn’t actually work as well as we once believed.

See: Mellon, Jonathan. "Rain, Rain, Go Away: 192 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable." Available at SSRN 3715610 (2022).

Matching (but not really)

- Matching is often touted as a way to make causal claims with observational data. You need exact matching to do this though and this isn’t feasible in practice.
- Matching is more-or-less subject to the same challenges as regression.

Lowering Your Expectations

- There is usually no statistical fix for endogeneity and other challenges to validity. These are intrinsic to the data and research design.
- Research designs that fix these challenges are often impossible. E.g. random assignment of democracy.
- It is unlikely you will be able to make strong causal claims from your research. That's okay!

Vocabulary

- Construct validity
- Internal validity
- External validity
- Post-treatment bias
- Selection bias
- Endogeneity
- Omitted variable bias
- Measurement error
- Simultaneity
- Random assignment
- Fixed/Random effects