Applied Microeconometrics II , Lecture 2

Ciprian Domnisoru
Aalto University

Validity typology (Shadish, Cook and Campbell)

Validity typology (Shadish, Cook and Campbell)

- Validity: A judgment about the extent to which evidence supports an inference as being true or correct.
- Statistical conclusion validity: the appropriate use of statistics. Is there covariance - how large/ how statistically significant?
- Internal validity: does the covariance result from a causal relationship; are your estimates biased ?
- Construct validity: Are we measuring what we say we are measuring?
- External validity: Can we generalize findings to other samples, populations, treatments, measurements, and settings ?

Statistical conclusion validity

- Small sample sizes mean low power- you fail to detect an effect. Solutions: increase sample size, try to limit the number of subgroups.
- Clustering- violation of test assumptions.
- Restrictions of range in the independent variable; Heterogeneity in dependent variable
- Report effect sizes. Statistically significant but economically unimportant effects.
- Fishing (or "p hacking"). Correct for repeated testing (Bonferonni)

Fishing or p-hacking

Replicability crisis; Publication bias- "Precise Nulls" team; Pre-registering experiments;

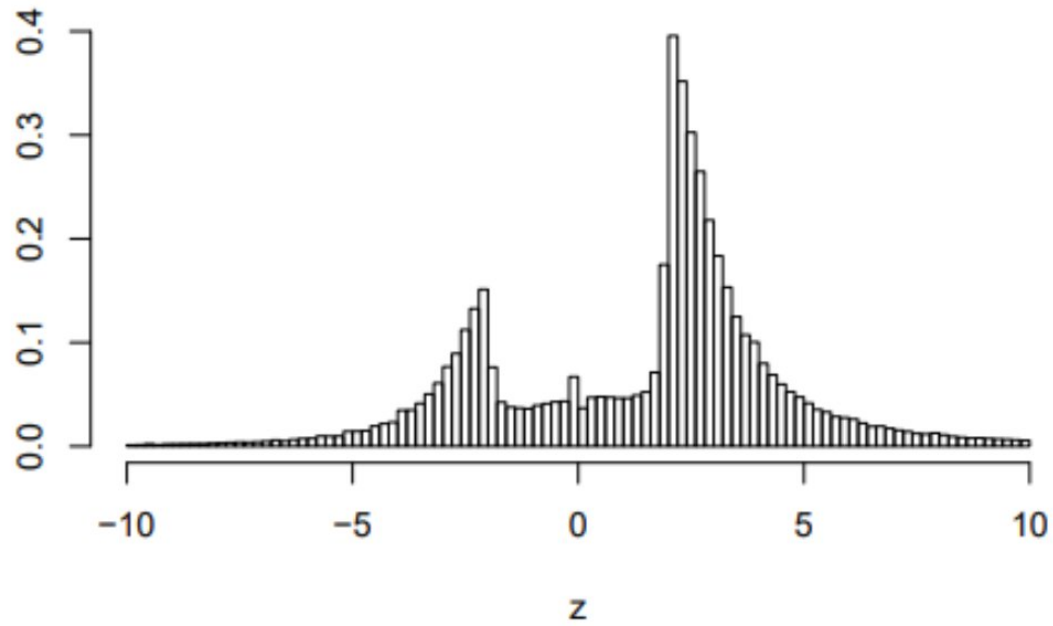
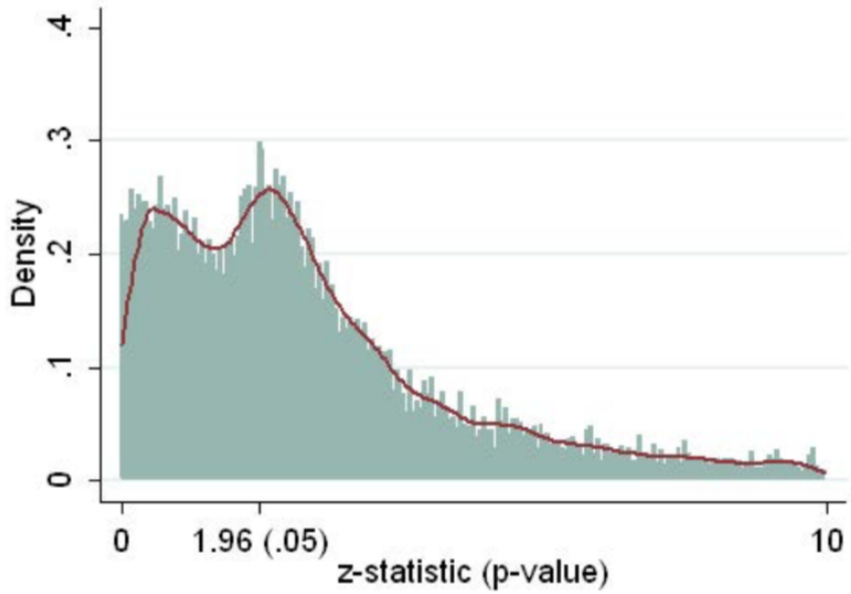


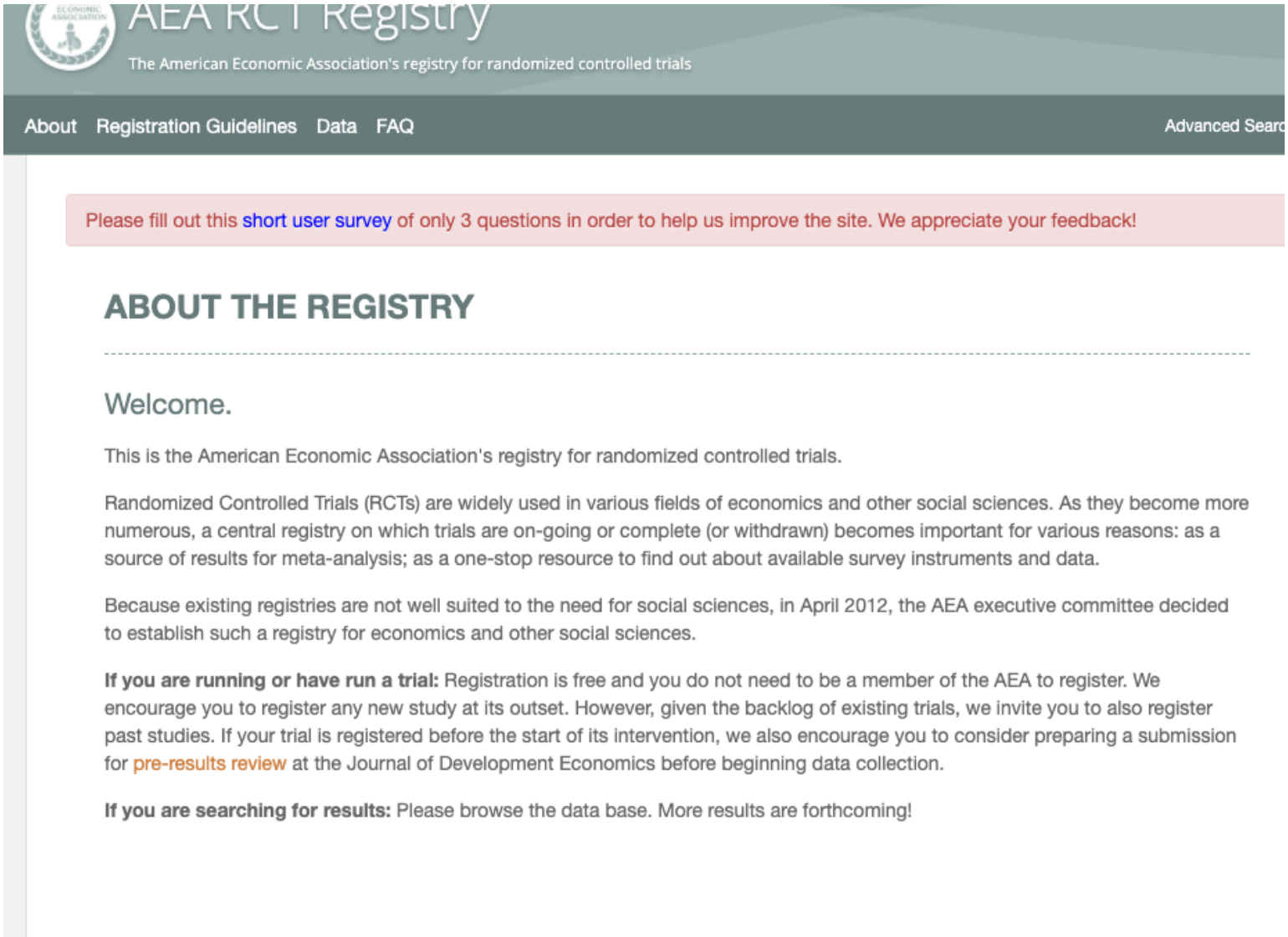
Figure 1: The distribution of more than one million z-values from Medline (1976–2019).

Statistical conclusion validity



- Brodeur, Abel ,Lé, Mathias , Sangnier, Marc , Zylberberg, Yanos, 2013. "Star Wars: The Empirics Strike Back," IZA Discussion Papers 7268, Institute of Labor Economics (IZA).
- Andrew Gelman refers to the "The Garden of Forking Paths" to describe the near infinite number of choices facing researchers in cleaning and analyzing data.
- Gelman argues that scientists can make false discoveries when they do not pre-specify a data analysis plan and instead choose 'one analysis for the particular data they saw.

Pre-registration of experiments



Fishing, or p-hacking

> J Clin Epidemiol. 2006 Sep;59(9):964-9. doi: 10.1016/j.jclinepi.2006.01.012. Epub 2006 Jul 11.

Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health

Peter C Austin [1], Muhammad M Mamdani, David N Juurlink, Janet E Hux

Affiliations + expand

PMID: 16895820 DOI: 10.1016/j.jclinepi.2006.01.012

Abstract

Objectives: To illustrate how multiple hypotheses testing can produce associations with no clinical plausibility.

Study design and setting: We conducted a study of all 10,674,945 residents of Ontario aged between 18 and 100 years in 2000. Residents were randomly assigned to equally sized derivation and validation cohorts and classified according to their astrological sign. Using the derivation cohort, we searched through 223 of the most common diagnoses for hospitalization until we identified two for which subjects born under one astrological sign had a significantly higher probability of hospitalization compared to subjects born under the remaining signs combined ($P<0.05$).

Results: We tested these 24 associations in the independent validation cohort. Residents born under Leo had a higher probability of gastrointestinal hemorrhage ($P=0.0447$), while Sagittarians had a higher probability of humerus fracture ($P=0.0123$) compared to all other signs combined. After adjusting the significance level to account for multiple comparisons, none of the identified associations remained significant in either the derivation or validation cohort.

Conclusions: Our analyses illustrate how the testing of multiple, non-prespecified hypotheses increases the likelihood of detecting implausible associations. Our findings have important

Bonferroni correction

- Suppose you have five outcomes and four treatments.
- What's the probability of observing at least one significant result just due to chance?
  P(at least one significant result) $= 1 -$ P(no significant results)=1-$(1 - 0.05)^{20}$=0.64
- The Bonferroni correction sets the significance cut-off at α/n. Reject if p-value is less than 0.0025. P(at least one significant result) $= 1 -$ P(no significant results) $= 1 - (1 - 0.0025)^{20} \approx 0.0488$
- Great, close to the desired 0.05 level, but all tests are not necessarily independent of each other. The Bonferroni correction could be too conservative, leading to a high rate of false negatives.

Alternatives to Bonferroni correction

Example of Five Outcomes and Four Treatments, with p-values in parentheses

	Y1	Y2	Y3	Y4	Y5
Treat 1	0.022	0.043	0.083**	0.079***	0.032
	(0.516)	(0.258)	(0.031)	(0.001)	(0.178)
Treat 2	0.043	0.060	0.099***	0.083***	0.046**
	(0.168)	(0.109)	(0.006)	(0.001)	(0.048)
Treat 3	0.030	-0.006	-0.016	0.008	0.009
	(0.356)	(0.877)	(0.665)	(0.726)	(0.691)
Treat 4	0.042	0.093**	0.070*	0.044*	0.052**
	(0.179)	(0.014)	(0.052)	(0.064)	(0.024)
Sample Size	726	678	678	678	678

"An overview of multiple hypothesis testing commands in Stata" David McKenzie, Lead Economist, Development Research Group, https://blogs.worldbank.org/impactevaluations/overview-multiple-hypothesis-testing-commands-stata

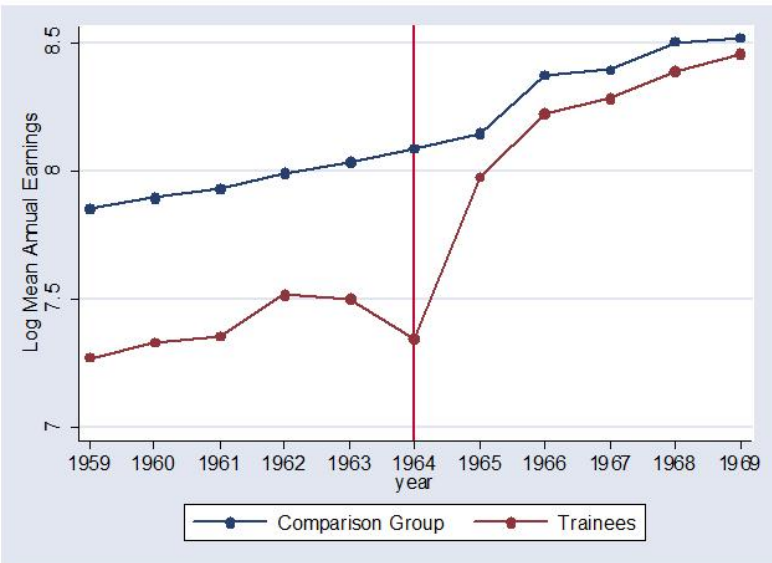
Reducing bias: where do we want more variance?

- **Restriction of range in independent variable** weakens relationship with dependent variable, increases bias. Solution? Pilot testing.
- **Heterogeneity of dependent variable** increases error variance and bias. Solution? Pilot testing; Blocking: male/female (usually sources of variability which are not of interest to the experimenter.
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)} =$

  $\beta_1 + \frac{Cov(u,X)}{Var(X)} = \beta_1 + \frac{\rho X,u * \sigma_u * \sigma_x}{\sigma_x^2} = \beta_1 + \frac{\rho X,u * \sigma_u}{\sigma_x}$

Threats to internal validity

- Selection; Non response bias
- History (events occuring concurrently)
- Maturation: naturally occuring changes are confused for treatment
- Regression (to the mean): in initial stages, the treated are observed in extreme circumstances that dissapear and make it seem like the treatment was (more) effective than it really was.
- "Ashenfelter Dip"



- Testing effects (practice, familiarity)

How do randomized experiments help?

- If you randomize, you should get rid of selection bias, history, maturation, regression and testing problems, as they are equally distributed across randomized groups.
- Problems: Failure to randomize
- Failure to follow treatment protocol. Mix up of Treatment and Controls.
- Attrition

Duflo, Hanna, Ryan."Incentives Work: Getting Teachers to Come to School"

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- Regular teachers in India have few incentives, often don't show up to work.
- Can an incentive program for para-teachers increase teacher presence? Student outcomes?
- Paper combines:
  1. A randomized experiment in teacher incentives
  2. A regression discontinuity design that tests how teachers respond to financial incentives : change in teacher behavior just before and after the end of he month
  3. Structural model estimated using the treatment group: simple dynamic labor supply model, teachers choose each day whether to go and teach or not.

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- Seva Mandir, an NGO in rural Rajasthan, who runs 150 "non-formal education center" (NFE): single teacher school for students who do not attend regular school.
- Students are 7-14 year old, illiterate when they join.
- Teacher absence rate 35%
- Schools teach basic hindi and math skills and prepare students to "graduate" to primary school.
- In 1997, 20 million children were served by such NFEs

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- Teachers in intervention schools received a camera with non-temperable time and date stamp.
- Instructed to take two pictures of themselves and the children every day (pictures separated by at least 5 hours, at least 8 children per picture).
- Payment is calculated each month and is a non-linear function of attendance: • Up to 10 days: Rs 500. • Each day above 10 days: Rs 50. • In non-intervention schools, teachers receive Rs 1000, and are reminded by attending at least 20 days is compulsory.

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- "We originally picked 120 schools, out of which 7 closed immediately after they were picked to be in the study (unrelated to the study)".
- 57 treatment schools, the rest control. Data collection: • Teacher and student attendance: Monthly random checks. • In treatment schools: Camera data • Students learning: tests in September 03-April 04-Oct 04 • Long term impact: a new sets of random checks was done in 2006-2007, and a new set of test scores were done in 2007
- Findings: "Over the 30 months in which attendance was tracked, teachers at program schools had an absence rate of 21 percent, compared to 44 percent at baseline and the 42 percent in the comparison schools."

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- 1. What was the power of the Experiment? At what level was the experiment randomized? "Need to take into account clustering at that level in computing our standard error"
- 2. What the randomization successful (was there balance between treatment and control group in covariates)
  - Ways to enforce balance: Stratifying (randomization within groups)
  - Ways to check balance: Compare covariates
- 3. Did we have attrition (lost observations)? • If so, how did we deal with it?
- 4 Did we have non-compliance? • If so, how did we deal with it?
- 5. Did we have contagion (externalities) between treatment and control group?

Duflo, Hanna, Ryan: Table 2

TABLE 2—TEACHER ATTENDANCE

September 2003–February 2006			Difference between treatment and control schools		
Treatment (1)	Control (2)	Diff (3)	Until mid-test (4)	Mid- to post-test (5)	After post-test (6)
Panel A. All teachers					
0.79	0.58	0.21	0.20	0.17	0.23
		(0.03)	(0.04)	(0.04)	(0.04)
1,575	1,496	3,071	882	660	1,529
Panel B. Teachers with above median test scores					
0.78	0.63	0.15	0.15	0.15	0.14
		(0.04)	(0.05)	(0.05)	(0.06)
843	702	1,545	423	327	795
Panel C. Teachers with below median test scores					
0.78	0.53	0.24	0.21	0.14	0.32
		(0.04)	(0.05)	(0.06)	(0.06)
625	757	1,382	412	300	670

Notes: Child learning levels were assessed in a mid-test (April 2004) and a post-test (November 2004). After the post-test, the "official" evaluation period was ended. Random checks continued in both the treatment and control schools. Standard errors are clustered by school. Panels B and C only include the 109 schools where teacher tests were available.

Duflo, Hanna, Ryan: Table 2

```
. regress open treat, cluster(schid)

Linear regression                               Number of obs     =      3,071
                                                F(1, 112)         =      49.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0497
                                                Root MSE          =     .45266

                                  (Std. Err. adjusted for 113 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
        open |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .2071212    .029586     7.00   0.000     .1485003    .2657422
       _cons |   .5795455    .0226667   25.57   0.000     .5346344    .6244565
------------------------------------------------------------------------------
```

Duflo, Hanna, Ryan: Table 2 with robust errors

```
. regress open treat, robust

Linear regression                               Number of obs     =      3,071
                                                F(1, 3069)        =     159.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0497
                                                Root MSE          =     .45266

------------------------------------------------------------------------------
             |               Robust
        open |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .2071212   .0164199    12.61   0.000     .1749262    .2393163
       _cons |   .5795455   .0127667    45.40   0.000     .5545133    .6045776
------------------------------------------------------------------------------
```

OLS essentials: homoskedasticity (univariate case)

- $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Var $(\hat{\beta}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}$
- If $\sigma_i = \sigma$ for all i, expression above simplifies to $\frac{\sigma^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]}$.

OLS essentials: robust standard errors (univariate case)

- If $\sigma_i = \sigma$ for all i, expression above simplifies to $\frac{\sigma^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]}$.
- Heteroskedasticity: Var(u_i) may vary with i, the above expression no longer simplifies, we are left with Var $(\hat{\beta}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}$.
- `,robust` command in Stata produces corrected standard errors.

Robust standard errors

- Var $(\hat{\beta}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}$ where we can estimate σ_i^2 by $\hat{u_i^2}$. Why can we estimate the variance of the residual , σ_i^2, by just $\hat{u_i^2}$? Think about the definition of variance and OLS properties. $Var[X] = E[X^2] - (E[X])^2$

In the multivariate case, $\widehat{Avar(\hat{\beta})} = (X'X)^{-1}\left[\sum_{i=1}^{n} \hat{u_i^2} x_i x_i'\right](X'X)^{-1}$.

Robust standard errors

The variance covariance matrix of the error terms in the multivariate robust case is

$$\begin{pmatrix} \sigma_1^2 & 0 & . & . & 0 & 0 \\ 0 & \sigma_2^2. & . & . & 0 & 0 \\ . & . & . & . & 0 & . \\ . & . & . & . & \sigma_{n-1}^2 & . \\ 0 & 0 & . & . & . & \sigma_n^2 \end{pmatrix}$$

, which can be estimated by

$$\begin{pmatrix} \hat{u}_1^2 & 0 & . & . & 0 & 0 \\ 0 & \hat{u}_2^2. & . & . & 0 & 0 \\ . & . & . & . & 0 & . \\ . & . & . & . & \hat{u}_{n-1}^2 & . \\ 0 & 0 & . & . & . & \hat{u}_n^2 \end{pmatrix}$$

Correlated errors

- Even if we acknowledge heteroskedasticity, we still assume the errors are independently distributed (how can you see the independence assumption in the matrices above?)
- Correlations in error terms are violations of the assumption of independently distributed errors.
- Correlated errors can introduce bias in the estimation of standard errors: errors are too low, you risk concluding there is a "significant" treatment difference more often than you should.

OLS essentials: clustering

- Suppose you create a sample of high school students by first drawing a sample of schools and then randomly selecting some number of students per school. This is called cluster sampling, where each school is a cluster.
- Clusters also arise naturally, without any deliberate sample design. Depending on the situation, you may consider individuals from the same families, neighborhoods, cities or even states to be a cluster.

Clustering

Suppose we drop the assumption of no serial correlation within classrooms. Then the variance-covariance matrix of the vector of error terms changes from

$$\begin{pmatrix} \sigma^2_{1,C1} & 0 & . & . & 0 & 0 \\ 0 & \sigma^2_{2,C1} & . & . & 0 & 0 \\ . & . & . & . & 0 & . \\ . & . & . & . & \sigma^2_{n-1,Cn} & . \\ 0 & 0 & . & . & . & \sigma^2_{n,Cn} \end{pmatrix}$$ to

$$\begin{pmatrix} \sigma^2\Omega_{C1} & 0 & . & . & 0 & 0 \\ 0 & \sigma^2\Omega_{C2} & . & . & 0 & 0 \\ . & .. & . & . & . & \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sigma^2\Omega_{Cn} \end{pmatrix}$$, where C_1, C_2 and C_n indicate classrooms.

Clustering

In the matrix $\begin{pmatrix} \sigma^2\Omega_{C1} & 0 & . & . & 0 & 0 \\ 0 & \sigma^2\Omega_{C2}. & . & . & 0 & 0 \\ . & .. & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sigma^2\Omega_{Cn} \end{pmatrix}$,Ω_{C1} is

$\begin{pmatrix} \sigma^2_{1,C1} & \sigma_{1,2} & . & . & \sigma_{1,k-1} & \sigma_{1,k} \\ \sigma_{2,1} & \sigma^2_{2,C1}. & . & . & \sigma_{2,k-1} & ....\sigma_{2,k}. \\ . & . & . & . & \sigma^2_{k-1,C1} & . \\ \sigma_{k,1} & \sigma_{k,2} & . & . & \sigma^2_{k-1,k} & \sigma^2_{k,C1} \end{pmatrix}$,

where k is the number of students in classroom 1 and $\sigma_{1,2}$ is the covariance between the error terms between student 1 and student 2, $\sigma_{1,2}$=σ_1*σ_2*$\rho_{1,2}$

Clustering: estimation

- For robust standard errors, $\widehat{Avar(\hat{\beta})} = (X'X)^{-1}\left[\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i'\right](X'X)^{-1}$
- For cluster robust standard errors, it can be shown that $\widehat{Avar(\hat{\beta})} = (X'X)^{-1}\left[\sum_{g=1}^{G} X_g' \hat{u}_g' \hat{u}_g X_g\right](X'X)^{-1}$

OLS essentials: clustering

- If we don't account for clustering, the standard errors will be incorrectly too small. The reason is that the observations within a cluster are not completely independent of each other, so the individual errors do not average out as fast.
- Moulton (1990) quantified the differences between the variance that accounts for clustering and the variance that doesn't (V_{REG})

  $V_{TRUE}(\hat{\beta}_1) = [1 + (N_g - 1)\rho_x\rho_\epsilon]\ \widehat{V_{REG}(\hat{\beta}_1)}$, clusters are all the same size, N_g, $\rho_x$and ρ_ϵ are the within cluster correlation of x and ϵ.

Construct validity

Start with a clear explanation of constructs for the the person, setting, treatment, and outcome of interest.

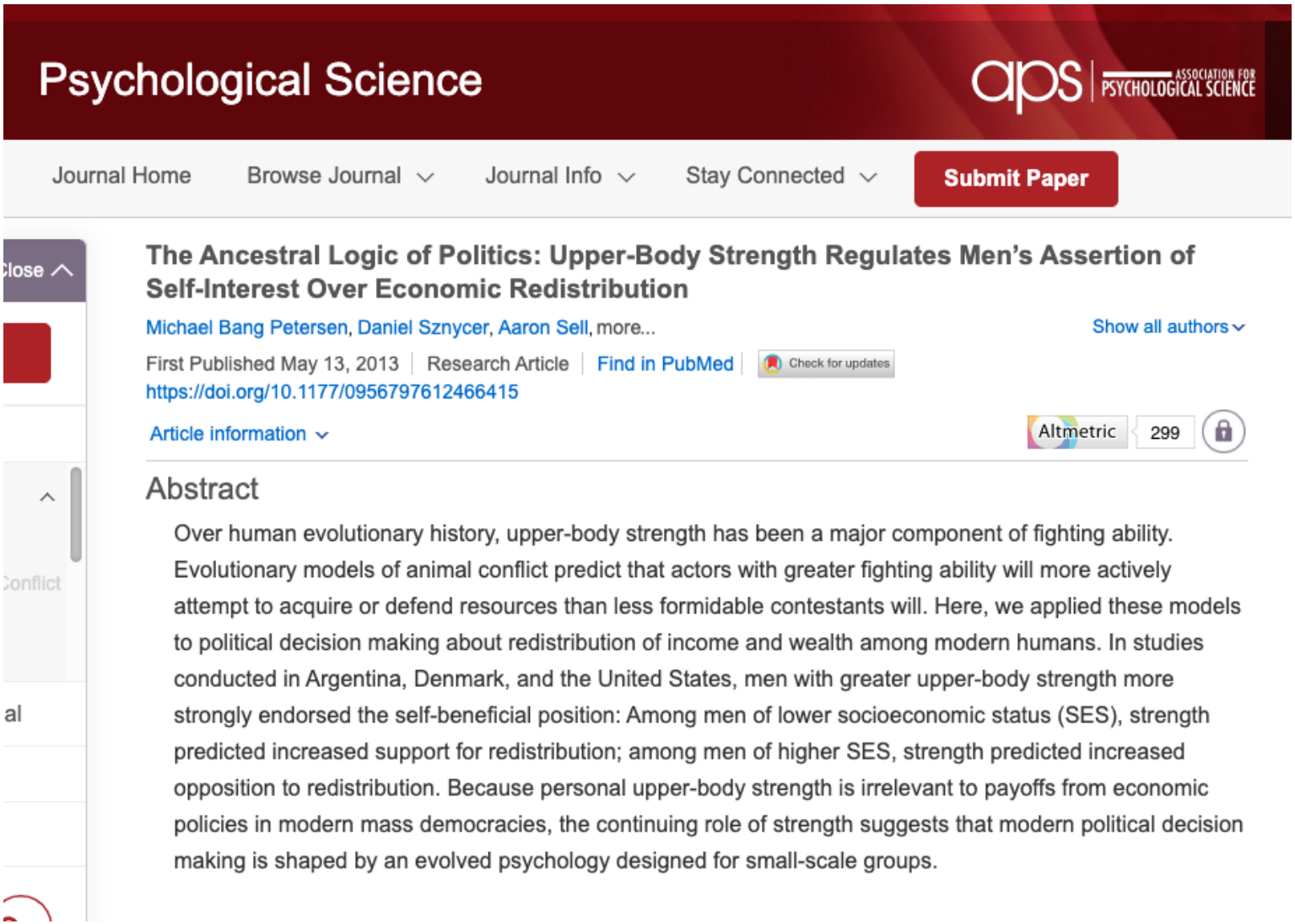
- Experimenter expectancies: subjects try to guess them, or they are somehow conveyed to subjects.
- Compensatory equalization
- Resentful demoralization
- Hawthorne and John Henry (compensatory equalization) effects.

Construct validity:John Henry (compensatory equalization)

Duflo, Hanna, Ryan: Incentives Work: Getting Teachers to Come to School

- "While the reduced form results inform us that this program was effective in reducing absenteeism, [...] they do not allow us to identify the response to the financial incentive separately from a possible independent effect of collecting daily data on absence."

Construct validity



Psychological Science

aps | ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Journal Home | Browse Journal | Journal Info | Stay Connected | Submit Paper

The Ancestral Logic of Politics: Upper-Body Strength Regulates Men's Assertion of Self-Interest Over Economic Redistribution

Michael Bang Petersen, Daniel Sznycer, Aaron Sell, more... Show all authors

First Published May 13, 2013 | Research Article | Find in PubMed | Check for updates

https://doi.org/10.1177/0956797612466415

Article information | Altmetric 299

Abstract

Over human evolutionary history, upper-body strength has been a major component of fighting ability. Evolutionary models of animal conflict predict that actors with greater fighting ability will more actively attempt to acquire or defend resources than less formidable contestants will. Here, we applied these models to political decision making about redistribution of income and wealth among modern humans. In studies conducted in Argentina, Denmark, and the United States, men with greater upper-body strength more strongly endorsed the self-beneficial position: Among men of lower socioeconomic status (SES), strength predicted increased support for redistribution; among men of higher SES, strength predicted increased opposition to redistribution. Because personal upper-body strength is irrelevant to payoffs from economic policies in modern mass democracies, the continuing role of strength suggests that modern political decision making is shaped by an evolved psychology designed for small-scale groups.

External validity

- Nonrepresentative sample
- Limited duration. People may react differently to a temporary program than to a permanent program.
- Experiment specificity: geographic, small scale tightly controlled.
- General equilibrium effects