Regression adjustment in experiments: A menace to society?

Winston, not Peter :)

Peter's Class

March 15, 2021

David Freedman (1938–2008)

Freedman was a great statistician and probabilist,
but he argued for more humility
about what statistics can accomplish.

- Fancy statistics can't save a weak research design
- Prefer simple, transparent methods
  and lots of critical discussion
- Qualitative research is as important as quantitative

Two of his many excellent essays:

- "Statistical models and shoe leather" (with discussion)
  Sociological Methodology (1991)
- "On types of scientific inquiry:
  The role of qualitative reasoning"
  Statistical Models and Causal Inference: A Dialogue with the Social Sciences (2010)

Freedman on regression in the social sciences

from "Statistical models and shoe leather"

4-point scale:

1. Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong.
2. Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use.
3. Regression might work, but it hasn't yet.
4. Regression can't work.

"My own view is bracketed by categories 2 and 3, although good examples are quite hard to find."

Example: A welfare reform experiment

	Average Earnings ($)		
	Treatment group	Control group	Difference
Year 1	2,470	1,550	920***
Year 2	3,416	2,233	1,183***
Year 3	3,562	2,552	1,010***

Riccio et al. (MDRC, 1994)

- 5,508 welfare recipients and applicants (Riverside County, CA)
- Treatment group: Mandatory job search / basic education
- Control group: no mandate

Example: A welfare reform experiment

	Average Earnings ($)		
	Treatment group	Control group	Difference
Year 1	2,470	1,550	920***
Year 2	3,416	2,233	1,183***
Year 3	3,562	2,552	1,010***

- All the estimates are regression-adjusted.
- Adjustment is standard in the evaluation industry and common in academic publications.

The usual OLS adjustment

Y_i	Outcome (e.g.: earnings in year 1 after random assignment)
T_i	Treatment group dummy variable
X_i	Baseline covariate(s) (e.g.: previous earnings, education)
$\overline{X}$	Mean covariate value for subjects in the experiment

Ordinary least squares (OLS) regression:

$$Y_i \;=\; \widehat{\alpha} + \widehat{\beta} \cdot X_i + \widehat{\gamma} \cdot T_i + \widehat{\epsilon}_i$$

Regression-adjusted estimates:

Treatment group mean	$\widehat{\alpha} + \widehat{\beta} \cdot \overline{X} + \widehat{\gamma}$
Control group mean	$\widehat{\alpha} + \widehat{\beta} \cdot \overline{X}$
Difference	$\widehat{\gamma}$

The traditional case for adjustment

Random assignment reduces the need for dubious statistical models. So why use regression models?

Standard rationale: to improve precision

- Fisher (1932) example: Tea bushes
  - Y = Tea yield after random assignment
  - X = Tea yield before random assignment (explains 86% of the variation in Y)
  - Adjustment for X reduced the variance of the estimated treatment effect by a factor of 7
- In social experiments, gains are typically more modest
  - Example based on Angrist, Lang, & Oreopoulos (2009)
  - Y = First-year college GPA
  - X = High school GPA (explains 15% of the variation in Y)
  - Standard error of unadjusted estimate: 0.16
  - Standard error of adjusted estimate: 0.15

The traditional case for adjustment

Assumptions:

- Regression model is correct
- The covariates in the regression are:
  - Measured before random assignment
  - Pre-specified (no fishing)
  - Correlated with the outcome
- Sample size is much larger than no. of covariates

Results:

- The unadjusted and OLS regression-adjusted treatment effect estimates are both unbiased
- Adjustment tends to improve precision

Adjustment can hurt precision if:

- Too many covariates relative to sample size, or
- Covariates have very little correlation with outcome

More details on the traditional case

Cox & McCullagh 1982

If the classical regression model is correct,
OLS adjustment multiplies the true SE by about

$$\sqrt{(1-\rho^2)\,\frac{N-3}{N-3-K}}$$

ρ^2	Population R^2 (in regression of outcome on covariates)
N	Sample size
K	No. of covariates

- Adjustment tends to improve precision if $K/N \approx 0$ and $\rho^2 \gg 0$
- The gain is small if $\rho^2 < 0.1$
- Adjustment can hurt precision if K/N is too large or ρ^2 is too small

Freedman's critique

"On regression adjustments to experimental data" (2008a)

"On regression adjustments in experiments with several treatments" (2008b)

What are the properties of OLS adjustment when the regression model is false?

Freedman uses Neyman's model for randomization inference.

(Neyman 1923, 1990)

Completely randomized experiment:

"Population"	N subjects in the experiment
Treatment group	Simple random sample of fixed size N_T
Control group	The other $N - N_T$ subjects

Neyman's model (with covariate)

Observed data for person i

T_i	Treatment group dummy variable
X_i	Covariate (not affected by treatment)
Y_i	Outcome

Y_i is a composite of two potential outcomes (counterfactuals):

Y_{1i}	Outcome that would occur if $T_i = 1$
Y_{0i}	Outcome that would occur if $T_i = 0$

Individual treatment effect: $Y_{1i} - Y_{0i}$

Trying to estimate the average treatment effect (ATE)

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) = \overline{Y}_{1,\,pop} - \overline{Y}_{0,\,pop}$$

Key features of Neyman's model

T_i	Treatment	Observed	Random
Y_{1i}	Potential outcome	Observed if $T_i = 1$	Fixed
Y_{0i}	Potential outcome	Observed if $T_i = 0$	Fixed
Y_i	Outcome	Observed	Random
X_i	Covariate	Observed	Fixed

- Treatment effect $(Y_{1i} - Y_{0i})$ can vary with i
- No assumptions about relationship between Y_i and X_i
- No "error term" assumptions
- No imaginary superpopulation
- Random assignment is the source of randomness

- Does assume SUTVA
  (no interference; no hidden versions of treatments)

Freedman's conclusions

Freedman found 3 problems with OLS regression adjustment:

1. The precision problem:
   Adjustment can actually worsen asymptotic precision
   (even if there's only one covariate
   and even if the covariate is strongly correlated
   with the outcome).

2. The SE problem:
   The conventional OLS standard error estimator
   is inconsistent.

3. The small-sample bias problem:
   The adjusted estimator of average treatment effect
   has a bias of order $1 / N$.

"The reason for the breakdown is not hard to find:
randomization does not justify the assumptions
behind the OLS model."

Sidebar: On asymptotics

Both Freedman's analysis and mine rely on asymptotic theory.

Asymptotic approximations are often excellent in moderate or large samples and enable us to avoid strong parametric assumptions. But your mileage may vary.

"An important feature of large-sample theory is that it is nonparametric. Its limit theorems provide distribution-free approximations for statistical quantities such as significance levels, critical values, power, confidence coefficients, and so on. However, the accuracy of these approximations is not distribution-free but, instead, depends both on the sample size and on the underlying distribution. To obtain an idea of the accuracy, it is necessary to supplement the theoretical results with numerical work, much of it based on simulation."

Lehmann (1999), *Elements of Large-Sample Theory*

(Empirical example: Section 7 in paper)

Hidden bright side of Freedman's critique

Asymptotic precision:

Adjustment can't hurt when $\Pr(T_i = 1) = 0.5$

- Noted by Freedman (2008a), but not emphasized

Adjustment hurts only under "severe" conditions

- Suppose $\Pr(T_i = 1) \in [0.25, 0.75]$
- Then for adjustment to hurt, X_i must covary more with the treatment effect than with the expected outcome.
- Not noted by Freedman, but follows from his asymptotic variance formula.

SE estimation:

The conventional SE is consistent or asymptotically conservative when $\Pr(T_i = 1) = 0.5$

- Noted by Freedman (2008a), but not emphasized

The dark side stole the headlines

Freedman and many excellent applied statisticians summarized his papers in terms that emphasized the dangers of adjustment.

"Random assignment does not justify any form of regression with covariates.
If regression adjustments are introduced nevertheless, there is likely to be bias in any estimates of treatment effects and badly biased standard errors."

Richard Berk et al. (2010), *Journal of Experimental Criminology*

"Researchers should be extremely cautious about using multiple regression to adjust experimental data. Unfortunately, there is a tendency to use it freely."

Jas Sekhon (2009), "Opiates for the Matches"

What my 2013 paper is about

Revealing the hidden bright side (earlier slide)

Simple solutions to Freedman's major problems (precision, SE)

1. OLS adjustment with treatment $\times$ covariate interactions
   (see also Yang & Tsiatis 2001)
2. Huber–White sandwich SE estimator ("robust")

Perspective on Freedman's 3rd problem (small-sample bias)

- Literally a 2nd-order issue
- The bias can be estimated

Intuitions from parallels with other literatures

- Econometrics: "Agnostic" view of regression
  Goldberger 1991; Angrist & Pischke 2009
- Survey sampling: Regression estimators of pop. means

What my paper is not about

- "Three cheers for regression adjustment"
- Selling a new adjustment method
- Finding the optimal adjustment
- Specific practical guidance
- Nonlinear models (logit, probit, Cox, etc.)

But note that Freedman's and my results on OLS are applicable to binary outcomes

Regression estimators in survey sampling

Cochran (1977)
Sampling Techniques, 3rd ed.

Estimating the mean surface area $\overline{Y}_{pop}$ of leaves on a plant

Y_i	Surface area	Measured for random sample
X_i	Mass	Measured for all leaves

$\overline{Y}_{sample}$ is unbiased, but it ignores the auxiliary info (X).

If $\overline{X}_{pop} > \overline{X}_{sample}$, then we expect $\overline{Y}_{pop} > \overline{Y}_{sample}$.

OLS regression estimator of $\overline{Y}_{pop}$:

$$\widehat{\overline{Y}}_{OLS} \equiv \overline{Y}_{sample} + \widehat{\beta}_{OLS} \cdot (\overline{X}_{pop} - \overline{X}_{sample})$$

Consistency of regression estimators

Claim

Under simple random sampling,
$\widehat{\overline{Y}}_{OLS}$ is a consistent estimator of $\overline{Y}_{pop}$,
even if the regression model is horribly misspecified.

Informal argument (under suitable regularity conditions)

$$\widehat{\overline{Y}}_{OLS} - \overline{Y}_{sample} \;=\; \widehat{\beta}_{OLS} \cdot (\overline{X}_{pop} - \overline{X}_{sample})$$

1. $(\overline{X}_{pop} - \overline{X}_{sample}) \xrightarrow{p} 0.$
2. $\widehat{\beta}_{OLS} \xrightarrow{p}$ a finite limit ("population least squares").
3. Therefore, $(\widehat{\overline{Y}}_{OLS} - \overline{Y}_{sample}) \xrightarrow{p} 0.$
4. $\overline{Y}_{sample}$ is a consistent estimator of $\overline{Y}_{pop}$.
5. Therefore, $\widehat{\overline{Y}}_{OLS}$ is a consistent estimator of $\overline{Y}_{pop}$.

Adjustment improves asymptotic precision

Claim

Under simple random sampling,
even if the regression model is horribly misspecified,
$\widehat{\overline{Y}}_{OLS}$ is asymptotically more efficient than $\overline{Y}_{sample}$ unless X is uncorrelated with Y (in which case $\widehat{\overline{Y}}_{OLS}$ and $\overline{Y}_{sample}$ have equal asymptotic variance).

Informal argument (Cochran 1977)

First, imagine using a "fixed-slope regression estimator":

$$\widehat{\overline{Y}}_{fixed\,slope} \equiv \overline{Y}_{sample} + b \cdot (\overline{X}_{pop} - \overline{X}_{sample})$$

where b is a constant.

Note that $\overline{Y}_{sample}$ itself is a fixed-slope regression estimator (with $b = 0$).

Adjustment improves asymptotic precision

Informal argument (cont'd)

$$\widehat{\overline{Y}}_{fixedslope} = \overline{Y}_{sample} - b \cdot (\overline{X}_{sample} - \overline{X}_{pop})$$

is the sample mean of $Y_i - b \cdot (X_i - \overline{X}_{pop})$, so its variance is

$$\frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{i=1}^{N} \left[(Y_i - \overline{Y}_{pop}) - b \cdot (X_i - \overline{X}_{pop}) \right]^2 .$$

What choice of b minimizes this variance?
The "population least squares" slope, β_{PopLS}.
Call the resulting estimator $\widehat{\overline{Y}}_{PopLS}$.

$\widehat{\overline{Y}}_{PopLS}$ has lower variance than $\overline{Y}_{sample}$ if $\beta_{PopLS} \neq 0$
(i.e., if X is correlated with Y in the population).

Adjustment improves asymptotic precision

Informal argument (cont'd)

Asymptotically, $\widehat{\overline{Y}}_{OLS}$ is as efficient as $\widehat{\overline{Y}}_{PopLS}$:

$$\widehat{\overline{Y}}_{OLS} - \overline{Y}_{pop} = (\widehat{\overline{Y}}_{PopLS} - \overline{Y}_{pop}) + (\widehat{\beta}_{OLS} - \beta_{PopLS}) \cdot (\overline{X}_{pop} - \overline{X}_{sample})$$

- $(\widehat{\overline{Y}}_{PopLS} - \overline{Y}_{pop})$ is of order $1/\sqrt{n}$
- $(\widehat{\beta}_{OLS} - \beta_{PopLS}) \cdot (\overline{X}_{pop} - \overline{X}_{sample})$ is of order $1/n$

So we have

$$\text{Avar}(\widehat{\overline{Y}}_{OLS}) = \text{Avar}(\widehat{\overline{Y}}_{PopLS}) \leq \text{Avar}(\overline{Y}_{sample})$$

and the inequality is strict unless $\beta_{PopLS} = 0$.

Separated at birth?

	Survey sampling	Experiments
Estimand	$\overline{Y}_{pop}$	$\overline{Y}_{1,\,pop} - \overline{Y}_{0,\,pop}$
Outcome data	Y_i (sample)	Y_{1i} (treatment group) Y_{0i} (control group)
Auxiliary data	X_i (population)	X_i (population)
Cochran's classic paper	Sampling theory when the sampling-units are of unequal sizes *JASA* (1942)	Analysis of covariance: Its nature and uses *Biometrics* (1957)
Theory	Agnostic	Linear model

What the analogy suggests

Run separate OLS regressions
in the treatment group and the control group:

1. Estimate $\overline{Y}_{1,\,pop}$
   - Regress Y_i on X_i in treatment group $\longrightarrow$ $\widehat{\beta}_{treat}$
   - $\widehat{\overline{Y}}_{1,\,OLS} \;=\; \overline{Y}_{treat} + \widehat{\beta}_{treat} \cdot (\overline{X}_{pop} - \overline{X}_{treat})$
2. Estimate $\overline{Y}_{0,\,pop}$
   - Regress Y_i on X_i in control group $\longrightarrow$ $\widehat{\beta}_{control}$
   - $\widehat{\overline{Y}}_{0,\,OLS} \;=\; \overline{Y}_{control} + \widehat{\beta}_{control} \cdot (\overline{X}_{pop} - \overline{X}_{control})$
3. Take the difference: $\widehat{ATE}_{interact} \;=\; \widehat{\overline{Y}}_{1,\,OLS} - \widehat{\overline{Y}}_{0,\,OLS}$

Shortcut: OLS with "demeaning interactions"

- Regress Y_i on T_i, X_i, and $T_i \cdot (X_i - \overline{X}_{pop})$

Precision improvement in experiments: Preview of results

Comparing 3 estimators of average treatment effect:

1. Unadjusted: $\widehat{ATE}_{unadj} = \overline{Y}_{treat} - \overline{Y}_{control}$
2. Usual OLS adjustment: $\widehat{ATE}_{adj}$
   Regress Y_i on T_i and X_i
3. OLS with "demeaning interactions": $\widehat{ATE}_{interact}$
   Regress Y_i on T_i, X_i, and $T_i \cdot (X_i - \overline{X}_{pop})$

All 3 estimators are consistent and asymptotically normal, even if the regression models are horribly misspecified.

Asymptotically, $\widehat{ATE}_{interact}$ is always either the most efficient or tied for most efficient,
even if the regression models are horribly misspecified.

Finite-population asymptotics

Infinite-population asymptotics (more familiar):

- Imagine a sequence of samples of size $n = 1, 2, \ldots$
- Population and estimand (e.g., ATE) don't vary with n
- Example of a regularity condition:
  $$E(X_i^4) \ < \ \infty$$
- Example of an asymptotic result (consistency):
  $$\widehat{ATE}_n \ \xrightarrow{p} \ ATE$$

Finite-population asymptotics (used by Freedman and me):

- Imagine a sequence of populations of size $N = 1, 2, \ldots$
- Example of a regularity condition:
  $$\frac{1}{N}\sum_{i=1}^{N} X_{i,N}^4 \ < \ L \qquad \text{for all } N = 1, 2, \ldots$$
- Example of an asymptotic result (consistency):
  $$\widehat{ATE}_N - ATE_N \ \xrightarrow{p} \ 0$$
- Freedman and I drop the subscript N

Regularity conditions (same as Freedman's)

For simplicity, these slides assume a single covariate X_i.
(My paper allows multiple covariates but assumes the number of covariates is fixed as $N \to \infty$.)

Condition 1

Y_{1i}, Y_{0i}, and X_i have bounded fourth moments.
For example, there exists $L < \infty$ such that

$$\frac{1}{N} \sum_{i=1}^{N} X_i^4 \ < \ L \qquad \text{for all } N = 1, 2, \ldots$$

Regularity conditions (same as Freedman's)

Condition 2

The population means, variances, and covariances of Y_{1i}, Y_{0i}, and X_i converge to finite limits as $N \to \infty$. The limiting variances are positive.

For example:

- $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (Y_{0i} - \overline{Y}_{0,\,pop})(X_i - \overline{X}_{pop})$ exists.
- $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X}_{pop})^2$ exists and is positive.

Condition 3

Let N_T denote the treatment group size. Then

$$\lim_{N\to\infty} \frac{N_T}{N} = p \qquad \text{where } 0 < p < 1.$$

Additional notation

Limits of "population least squares" slopes:

$$\beta_1 \equiv \lim_{N\to\infty} \frac{\sum_{i=1}^{N} (X_i - \overline{X}_{pop})(Y_{1i} - \overline{Y}_{1,pop})}{\sum_{i=1}^{N} (X_i - \overline{X}_{pop})^2}$$

$$\beta_0 \equiv \lim_{N\to\infty} \frac{\sum_{i=1}^{N} (X_i - \overline{X}_{pop})(Y_{0i} - \overline{Y}_{0,pop})}{\sum_{i=1}^{N} (X_i - \overline{X}_{pop})^2}$$

Prediction errors:

$$U_{1i} \equiv (Y_{1i} - \overline{Y}_{1,pop}) - \beta_1 \cdot (X_i - \overline{X}_{pop})$$

$$U_{0i} \equiv (Y_{0i} - \overline{Y}_{0,pop}) - \beta_0 \cdot (X_i - \overline{X}_{pop})$$

Additional notation

Generic notation for population variances and covariances of any non-random variables Z_i and W_i:

$$\sigma_Z^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (Z_i - \overline{Z}_{pop})^2$$

$$\sigma_{Z,W} \equiv \frac{1}{N} \sum_{i=1}^{N} (Z_i - \overline{Z}_{pop}) (W_i - \overline{W}_{pop})$$

OLS with "demeaning interactions" is consistent and asymptotically normal

Theorem 1

Under Conditions 1–3,

$$\sqrt{N}\ \left(\widehat{ATE}_{interact} - ATE\right)\ \xrightarrow{d}\ \text{Normal}\ (0, V)$$

where

$$V \ = \ \frac{1-p}{p}\lim_{N\to\infty}\sigma^2_{U_1} + \frac{p}{1-p}\lim_{N\to\infty}\sigma^2_{U_0} + 2\lim_{N\to\infty}\sigma_{U_1,U_0}.$$

Note

$\widehat{ATE}_{unadj}$ and $\widehat{ATE}_{adj}$ are also consistent and asymptotically normal (as shown by Freedman).

Outline of proof of Theorem 1

1. Imagine a fixed-slope regression estimator $\widehat{ATE}_{ideal}$, using the population least squares slopes β_1 and β_0

2. $\widehat{ATE}_{ideal} - ATE \;=\; \overline{U}_{1,\,treat} - \overline{U}_{0,\,control}$

3. A finite-population multivariate CLT gives the asymptotic distribution of $\overline{U}_{1,\,treat} - \overline{U}_{0,\,control}$
   (Freedman 2008b; Li & Ding 2016)

4. Show that $\sqrt{N}\,(\widehat{ATE}_{interact} - \widehat{ATE}_{ideal}) \xrightarrow{p} 0$
   - This is $\sqrt{N}\,(\widehat{\beta}_{treat} - \beta_1) \cdot (\overline{X}_{pop} - \overline{X}_{treat})$ minus the analogous term for the control group
   - Show $\widehat{\beta}_{treat} \xrightarrow{p} \beta_1$
     (use Chebyshev and Cauchy-Schwarz to prove a WLLN)
   - By the Central Limit Theorem, $\sqrt{N}\,(\overline{X}_{pop} - \overline{X}_{treat})$ is stochastically bounded
   - So $\sqrt{N}\,(\widehat{\beta}_{treat} - \beta_1) \cdot (\overline{X}_{pop} - \overline{X}_{treat}) \xrightarrow{p} 0$

$\blacksquare$

OLS with "demeaning interactions" cannot hurt asymptotic precision

Corollary 1.1

Asymptotically, $\widehat{ATE}_{interact}$ is at least as efficient as $\widehat{ATE}_{unadj}$ (the difference in means), and strictly more efficient unless

$$(1-p)\beta_1 + p\beta_0 = 0.$$

Corollary 1.2

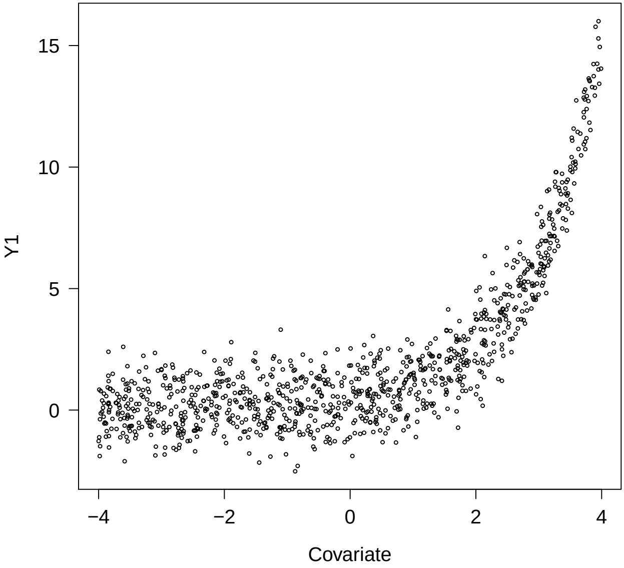
Asymptotically, $\widehat{ATE}_{interact}$ is at least as efficient as $\widehat{ATE}_{adj}$ (the usual OLS adjustment), and strictly more efficient unless either $\beta_1 = \beta_0$ or $p = 0.5$.

Simulation

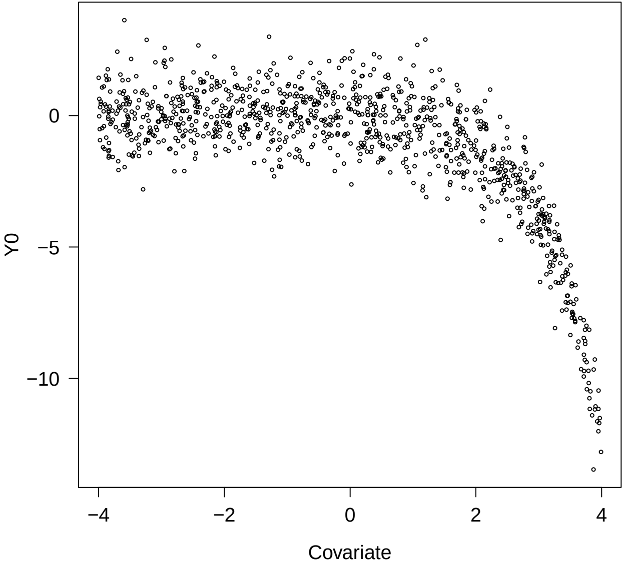
A road test of the 3 estimators
under "severe weather conditions":

1. Generate a finite population of 1,000 subjects
   - $X_i \sim$ Uniform $(-4, 4)$
   - $Y_{1i} = 0.25 \cdot e^{X_i} + 0.25 \cdot e^{X_i/2} + \nu_i$
   - $Y_{0i} = -0.25 \cdot e^{X_i} + 0.25 \cdot e^{X_i/2} + \epsilon_i$
   - ν_i and ϵ_i independent, standard Normal
2. Simulate a completely randomized experiment 40,000 times, assigning 75% of the subjects to treatment
3. Repeat step 2, assigning a different proportion to treatment (60%, 50%, 40%, or 25%)

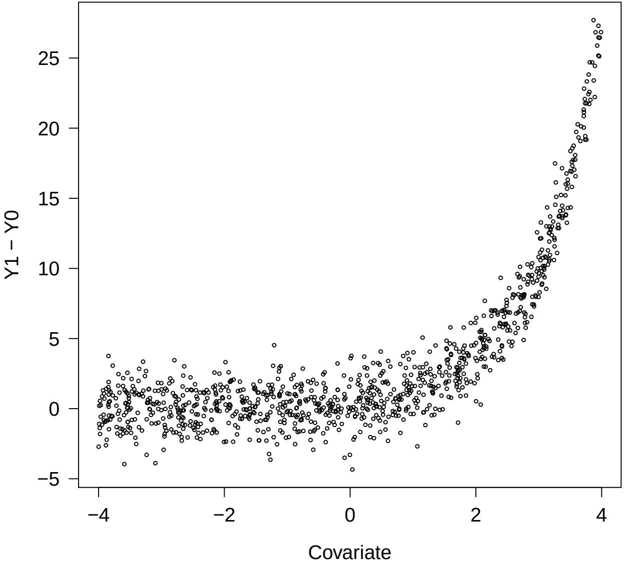
Potential outcome Y_{1i} vs. covariate X_i



Potential outcome Y_{0i} vs. covariate X_i



Treatment effect $Y_{1i} - Y_{0i}$ vs. covariate X_i



Simulation results

- All 3 estimators are approximately or exactly unbiased
- Empirical SDs ($\times$ 1,000) are shown below
  (and are very close to the asymptotic predictions)

	Proportion in treatment				
Estimator	75%	60%	50%	40%	25%
Unadjusted	93	49	53	78	142
Usual OLS-adjusted	171	73	47	80	180
"Demeaning interactions"	81	50	47	59	99

OLS with "demeaning interactions" solves Freedman's precision-worsening problem, even though "randomization does not justify the assumptions behind the OLS model."

Note that this simulation is an extreme scenario.
It was designed to illustrate that $\widehat{\mathrm{ACE}}_{interact}$ does well
even in extreme situations where $\widehat{\mathrm{ACE}}_{adj}$ hurts precision and
even when the outcome–covariate relationships are highly nonlinear.

The sandwich variance estimator

Conventional OLS variance estimator: $\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$

Sandwich estimator: $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}(\hat{\epsilon}_i^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

- Consistent under i.i.d. sampling
  - Don't need homoskedasticity or even linearity
    - *Mostly Harmless Econometrics*, section 3.1.3
    - Powell, "Asymptotics for least squares" (lecture notes)
- Downward bias and high variance in small samples
  - Bias corrections exist, but still have high variance
  - "HC2" (MacKinnon & White 1985) or "$J(1)$ jackknife" (Wu 1986) yields the familiar

$$\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_0^2}{n_0}$$

for the difference in means. In randomized experiments, this is unbiased or conservatively biased (Neyman 1923; Reichardt & Gollob 1999; Freedman, Pisani, & Purves 2007).

The sandwich estimator is consistent or asymptotically conservative in completely randomized experiments

Theorem 2

Let $\widehat{v}$ denote the sandwich variance estimator for $\widehat{ATE}_{interact}$. Under Conditions 1–3,

$$
\begin{aligned}
N\ \widehat{v} \quad &\xrightarrow{p} \quad \frac{1}{p}\lim_{N\to\infty}\sigma^2_{U_1} + \frac{1}{1-p}\lim_{N\to\infty}\sigma^2_{U_0} \\
&= \quad \text{Avar}\left(\sqrt{N}\left[\widehat{ATE}_{interact} - ATE\right]\right) + \\
&\qquad \lim_{N\to\infty}\sigma^2_{U_1-U_0}.
\end{aligned}
$$

A similar result holds for $\widehat{ATE}_{adj}$ and its sandwich variance estimator.

Comments on SE estimation

- Theorem 2 is analogous to Freedman's result for $\widehat{ATE}_{unadj}$ and the variance estimator

$$\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_0^2}{n_0}$$

(which, as we noted, is a sandwich-type estimator).

- Theorem 2 (together with the consistency and asymptotic normality of $\widehat{ATE}_{adj}$ and $\widehat{ATE}_{interact}$) implies that the sandwich SE estimator can be used to construct asymptotically valid confidence intervals for ATE.

- Freedman's critique conflates two independent questions:
    - To adjust, or not to adjust?
    - To use "robust," or not to use "robust"?

Practical takeaways

- Freedman's critique is a major theoretical contribution, but its practical implications have been exaggerated.
- Careful OLS adjustment is often a reasonable way to improve precision and power (but the gains are often modest).
- For careful OLS adjustment, the covariates should be:
  - Pre-specified (to prevent fishing)
  - Measured before random assignment
  - Much fewer than the sample size
- For transparency, reporting both unadjusted and adjusted estimates is a good idea.

Some things Freedman and I didn't cover

- Block-randomized experiments
- Cluster-randomized experiments
- Attrition
- High-dimensional covariate adjustment

  Bloniarz, Liu, Zhang, Sekhon, & Yu 2016
  Wager, Du, Taylor, & Tibshirani 2016
  Wu & Gagnon-Bartsch 2018
- Robust SEs in small samples

Cheat sheet on robust SEs

- Nice asymptotic properties
  - Consistent under i.i.d. sampling
  - Consistent or conservative under Neyman model (complete randomization of finite population)
- Not robust to outliers and heavy tails
- Downward bias and high variance in small samples
  - Bias corrections exist, but still have high variance
  - Therefore, the confidence interval $\widehat{\beta} \pm 1.96\,\widehat{\mathrm{SE}}$ may have coverage probability far below 95%
  - One possible remedy is Bell & McCaffrey's (2002) CI
    - Bias-reduced (HC2) version of robust SE
    - Instead of normal-distribution critical value, use t-distribution with Satterthwaite df
    - Similar to Welch's unequal-variances t-test
    - Discussions and extensions: Imbens & Kolesár 2016; Pustejovsky & Tipton 2016; Young 2016
    - clubSandwich R and Stata packages (among others)

Rosy vs. gloomy simulation scenarios

Asymptotic theory is only a rough guide.
Carefully designed simulations are valuable.

But simulations can be misleading
if they only examine rosy scenarios.

- In Neyman's finite-population framework, robust variance estimators are conservatively biased if treatment effects are heterogeneous.

  (Neyman 1923; Samii & Aronow 2012)

- Therefore, CI coverage probabilities look conservative in simulations with heterogeneous treatment effects and large enough N.
- For more stringent checks, we need to see gloomy scenarios:
  - Constant treatment effects, or
  - Infinite-population (i.i.d.) sampling

Simulation setup

- Lin (2013, section 7) used a subset ($N = 157$) of Angrist et al.'s (2009) data
  - Male first-year undergrads at a Canadian university
  - "Treatment": Peer advising and financial incentives
  - "Control": Peer advising only
  - Y = First-year college GPA
  - X = High school GPA
- Using a smaller subset here ($N = 30$ with lowest high-school GPAs)
  - 9 students assigned to "treatment"
  - 21 students assigned to "control"
- Treatment effect scenarios (both with $Y_{0i} = Y_i$)
  - Rosy: $Y_{1i} = \max(Y_{0i}, 2.0)$
  - Gloomy: $Y_{1i} = Y_{0i}$
- Simulate 250,000 replications of an RCT and estimate coverage probabilities of nominal 95% CIs

Rosy scenario (coverage-hacking)

Empirical coverage rates (%) of nominal 95% CIs

SE estimator	Critical value	ATE estimator: Unadjusted	ATE estimator: Usual OLS-adjusted
Huber–White	normal	98.1	97.3
HC1	normal	98.5	98.1
HC2	normal	98.5	98.1
HC2	Welch	99.0	
HC2	BM	99.2	99.0
HC3	normal	98.8	98.7

Gloomy scenario

Empirical coverage rates (%) of nominal 95% CIs

SE estimator	Critical value	ATE estimator: Unadjusted	ATE estimator: Usual OLS-adjusted
Huber–White	normal	92.0	91.3
HC1	normal	92.9	92.7
HC2	normal	93.2	93.1
HC2	Welch	94.9	
HC2	BM	95.1	95.1
HC3	normal	94.3	94.7

More takeaways

The methods literature emphasizes intellectual breakthroughs more than careful practical guidance.

We should try to accumulate a larger inventory of:

- Simulations with gloomy scenarios
- Bake-offs and road tests
- Reanalyses of previous RCTs