Regression Adjustment in Experiments with Heterogeneous Treatment Effects

Jeff Wooldridge
Michigan State University
(Based on Joint Work with Akanksha Negi)

Camp Resources XXV
August 14, 2018

1. Introduction and Motivation
2. Overview of Linear Regression Adjustment
3. Potential Outcomes, Random Assignment, Random Sampling
4. Linear Regression Adjustment
5. Nonlinear Regression Adjustment
6. More than Two Treatment Levels
7. Difference-in-Differences
8. Estimating Lower Bound WTP
9. Summary and Additional Comments

1. Introduction and Motivation

- Simplest program evaluation: control group and treated group.
- With randomized intervention, the difference-in-means (SDM) estimator is unbiased and consistent.
- Regression Adjustment (RA): Possibly improve precision by regressing on covariates that predict the outcome, Y.

- RA is consistent but, without strong functional form assumptions, generally biased.
- Rely on asymptotic variance to argue for improvements.
- Types of Regression Adjustment:
  - Linear versus Nonlinear
  - Pooled Regression Adjustment versus Full (Separate) Regression Adjustment
- Not all combinations are (theoretically) desirable.

2. Overview of Linear Regression Adjustment

- It matters whether treatment effect is constant or heterogeneous.
- The literature has focused on linear RA.
- Nonlinear methods, suitably chosen, are just as robust and can be much more efficient.

Constant Treatment Effect

- Pooled RA is consistent, has a smaller asymptotic variance than SDM.
- No improvement only when covariates are useless for predicting outcome.
- Holds for different sampling schemes.
  - ▸ Finite Population [Lin (2013, Annals of Applied Statistics)]
  - ▸ Random Sampling [Imbens and Rubin (2015)].

Heterogeneous Treatment Effects

- Finite Population, No Sampling Error: Lin (2013)
  - PRA may not improve over SDM.
  - FRA improves over PRA and SDM.

- Random Sampling from Large Population: Imbens and Rubin (2015).
  - ▸ PRA may not improve over SDM
  - ▸ FRA improves over PRA and SDM
  - ▸ Imbens and Rubin ignore sampling variation in $\bar{\mathbf{X}}$.
  - ▸ Negi and Wooldridge (2018): Account for sampling variation in $\bar{\mathbf{X}}$.
    - ▸ Full RA is still most efficient (asymptotically).

3. Potential Outcomes, Random Assignment, Random Sampling

- Binary treatment, $W \in \{0, 1\}$.
- Potential outcomes:

$$[Y(0), Y(1)]$$

- Observed outcome:

$$Y = (1 - W)Y(0) + WY(1) = Y(0) + W[Y(1) - Y(0)]$$

- Observed covariates:

$$\mathbf{X} = (X_1, X_2, \ldots, X_K).$$

- Random Assignment:

$$W \text{ is independent of } [Y(0), Y(1), \mathbf{X}].$$

- Assume $\{(Y_i, W_i, \mathbf{X}_i) : i = 1, 2, \ldots, N\}$ are independent and identically distributed.

4. Linear Regression Adjustment

- Tempting to say "linear model," but we do not assume a model.
- Let

$$\mu_0 = E[Y(0)],\ \mu_1 = E[Y(1)]$$

- Interested in the average treatment (causal) effect,

$$\tau = \mu_1 - \mu_0.$$

Constant Treatment Effect

- In the population,

$$Y(1) = \tau + Y(0)$$

- For any draw i from the population,

$$Y_i(1) = \tau + Y_i(0)$$

- Linearly project $Y(0)$ onto $\mathbf{X}$:

$$L[Y(0)|1, \mathbf{X}] = \alpha_0 + \mathbf{X}\boldsymbol{\beta}_0$$

- By construction,

$$Y(0) = \alpha_0 + \mathbf{X}\boldsymbol{\beta}_0 + U(0)$$

$$E[U(0)] = 0$$
$$E[\mathbf{X}'U(0)] = \mathbf{0}$$

- By random assignment,

$$W \text{ independent of } [U(0), \mathbf{X}]$$

- We can write

$$Y = \alpha_0 + \tau W + \mathbf{X}\boldsymbol{\beta}_0 + U(0)$$

$$E[U(0)] = 0$$
$$E[WU(0)] = 0$$
$$E[\mathbf{X}'U(0)] = \mathbf{0}$$

- Not a "true" model.

- Two estimators:

1. Difference in means:

$$\hat{\tau}_{SDM} = \bar{Y}_1 - \bar{Y}_0$$

- $\hat{\tau}_{SDM}$ is unbiased (conditional on $N_0, N_1 > 0$), consistent, $\sqrt{N}$-asymptotically normal.
- Inference from

$$Y_i \text{ on } 1, W_i, i = 1, \ldots, N.$$

2. Pooled RA: $\hat{\tau}_{PRA}$ from the regression

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, i = 1, \ldots, N.$$

- $\hat{\tau}_{PRA}$ is unbiased conditional on

$\{(W_i, \mathbf{X}_i) : i = 1, \ldots, N\}$ if

$$E[Y(0)|\mathbf{X}] = \alpha_0 + \mathbf{X}\boldsymbol{\beta}_0$$

- Otherwise, $\hat{\tau}_{PRA}$ is only consistent and

$\sqrt{N}$-asymptotically normal.

- In the constant treatment effect case,

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{SDM}-\tau)\Big] = \frac{\sigma^2_{Y(0)}}{Var(W)} = \frac{\sigma^2_{Y(0)}}{\rho(1-\rho)}$$

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{PRA}-\tau)\Big] = \frac{\sigma^2_{U(0)}}{Var(W)} = \frac{\sigma^2_{U(0)}}{\rho(1-\rho)}$$

$$\sigma^2_{Y(0)} = Var(\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2_{U(0)} \geq \sigma^2_{U(0)}$$

- Put good predictors of $Y(0)$ in $\mathbf{X}$.
- Past outcomes on Y?

Heterogeneous Treatment Effects

- Now write

$$Y(0) = \mu_0 + V(0)$$
$$Y(1) = \mu_1 + V(1)$$

- Treatment effect for unit i is

$$Y_i(1) - Y_i(0) = (\mu_1 - \mu_0) + [V_i(1) - V_i(0)]$$

- Still interested in

$$\tau = \mu_1 - \mu_0$$

- Population demeaned covariates:

$$\dot{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$$

- Linearly project each of $V(0)$ and $V(1)$ onto $\dot{\mathbf{X}}$:

$$V(0) = \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0)$$
$$V(1) = \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1)$$

- Then

$$Y(0) = \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0)$$
$$Y(1) = \mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1)$$

- The observed Y is

$$Y = (1 - W)[\mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0)] + W[\mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1)]$$
$$= \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) + \tau W + W\dot{\mathbf{X}}\boldsymbol{\delta} + W[U(1) - U(0)]$$

where

$$\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$$

- We do not want to assume anything about

$$E[Y(0)|\mathbf{X}],\ E[Y(1)|\mathbf{X}]$$

- Four different estimators.

1. Difference in means:

$$\hat{\tau}_{SDM} = \bar{Y}_1 - \bar{Y}_0$$

2. Pooled RA, $\hat{\tau}_{PRA}$:

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i$$

3. Full RA from separate regressions: $(\hat{\alpha}_w, \hat{\boldsymbol{\beta}}_w')$ from

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 0$$

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 1$$

and then

$$\hat{\tau}_{FRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{X}}\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)$$

- Same as coefficient on W_i in

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, W_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}), i = 1, 2, \ldots, N$$

4. Regression adjustment with known $\boldsymbol{\mu}_{\mathbf{X}}$:

$$\hat{\tau}_{IRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \boldsymbol{\mu}_{\mathbf{X}}\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right).$$

- Not an estimator if $\boldsymbol{\mu}_{\mathbf{X}}$ is unknown.
- "Infeasible" regression adjustment.

THEOREM (Negi and Wooldridge, 2018): Let

$$\sigma_w^2 = Var[U(w)], w = 0, 1$$

$$\mathbf{\Omega_X} = Var(\mathbf{X}) \ \ (K \times K)$$

• Under random assignment, finite moment conditions, and random sampling:

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{SDM} - \tau)\Big] = \boldsymbol{\beta}_1'\mathbf{\Omega_X}\boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0'\mathbf{\Omega_X}\boldsymbol{\beta}_0/(1-\rho) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho)$$

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{PRA}-\tau)\Big] = \left[\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)}\right] \cdot (\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_{\mathbf{X}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho).$$

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{FRA}-\tau)\Big] = (\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_{\mathbf{X}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho).$$

$$AVar\Big[\sqrt{N}\,(\hat{\tau}_{IRA}-\tau)\Big] = \sigma_1^2/\rho + \sigma_0^2/(1-\rho).$$

Implications

1. $\hat{\tau}_{IRA}$ is asymptotically most efficient.
2. No gain from knowing $\boldsymbol{\mu}_{\mathbf{X}}$ if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$.
   - Treatment effects could still be heterogenous.

3. The function

$$f(\rho) = \frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)}$$

is minimized at $\rho = 1/2$, and $f(1/2) = 1$.

▸ Therefore, PRA is asymptotically as efficient as FRA

if at least one of two conditions holds:

(i) $\beta_1 = \beta_0$

(ii) $\rho = 1/2$ (50-50 assignment)

4. FRA is always asymptotically more efficient than SDM:

$$\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0/(1-\rho) - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\boldsymbol{\Omega}_{\mathbf{X}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

can be written as

$$\boldsymbol{\lambda}'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\lambda}$$

where

$$\boldsymbol{\lambda} = \sqrt{\left(\frac{1-\rho}{\rho}\right)}\,\boldsymbol{\beta}_1 + \sqrt{\left(\frac{\rho}{1-\rho}\right)}\,\boldsymbol{\beta}_0.$$

- $\boldsymbol{\lambda} = \mathbf{0}$ if $\rho = 1/2$ *and* $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$.
- Latter condition seems unrealistic unless

$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \mathbf{0}$.

- Some intuition. The FRA estimator of μ_1 is

$$\hat{\mu}_{1,FRA} = \hat{\alpha}_1 + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1$$

- By OLS mechanics, SDM is based on

$$\hat{\mu}_{1,SDM} = \bar{Y}_1 = \hat{\alpha}_1 + \bar{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1$$

- $\hat{\mu}_{1,FRA}$ uses a more efficient estimator of $\boldsymbol{\mu}_{\mathbf{X}}$.

5. PRA and SDM cannot always be ranked.

▸ But PRA is more efficient if at least one of the conditions

(i) $\beta_1 = \beta_0$

(ii) $\rho = 1/2$

holds.

▸ If we know $\rho = 1/2$, PRA is attractive for small sample sizes: it conserves on parameters compared with FRA.

- Use Stata `teffects` to get proper standard errors.

`teffects ra y x1 ... xK`

- Accounts for sampling variation in $\bar{\mathbf{X}}$.
- Or use `vce(unconditional)` option with `margins`.

Simulations

- Broadly, two designs.

1. $E[Y(g)|\mathbf{X}]$ a quadratic in X_1, X_2: X_1, X_2, X_1^2, X_2^2, X_1X_2.

2. $E[Y(g)|\mathbf{X}]$ is probit with the index a quadratic in X_1, X_2.

- "Mild" versus "strong" heterogeneity in the covariates.
- RA uses only linear function of X_1, X_2.
  - ▸ Linear model is "misspecified."

- DGP1 = linear mean, mild heterogeneity
- DGP2 = linear mean, strong heterogeneity
- DGP3 = probit mean, mild heterogeneity
- DGP4 = probit mean, strong heterogeneity

DGP1	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	0.004	1.248	–0.037	1.107	–0.027	1.205	–0.017	1.455
PRA	0.033	1.599	–0.041	1.004	–0.050	0.963	–0.017	1.217
FRA	–0.038	1.128	–0.044	0.934	–0.047	0.962	–0.002	1.191
IRA	–0.059	1.056	–0.044	0.836	–0.057	0.878	–0.012	1.109

DGP2	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	–0.019	1.292	–0.019	0.900	–0.008	0.879	–0.029	1.055
PRA	0.012	1.958	–0.013	1.014	–0.004	0.860	–0.016	1.112
FRA	0.039	1.089	–0.009	0.871	–0.001	0.855	–0.002	0.997
IRA	0.028	0.967	–0.005	0.681	0.009	0.655	0.001	0.818

DGP3	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	–0.003	0.080	0.000	0.059	0.000	0.060	0.000	0.070
PRA	0.002	0.084	0.000	0.052	0.000	0.047	0.001	0.058
FRA	0.006	0.074	0.001	0.049	0.000	0.047	0.000	0.056
IRA	0.006	0.073	0.003	0.047	0.000	0.044	0.001	0.053

DGP4	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	–0.002	0.104	0.004	0.066	0.000	0.066	–0.001	0.075
PRA	–0.001	0.140	0.005	0.070	0.000	0.058	0.000	0.073
FRA	0.011	0.092	0.006	0.061	0.000	0.058	0.000	0.062
IRA	0.011	0.082	0.005	0.049	0.002	0.045	0.000	0.048

DGP1	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 1,000$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	–0.016	0.569	0.009	0.490	–0.018	0.519	0.003	0.661
PRA	0.009	0.717	–0.002	0.433	0.002	0.433	–0.007	0.540
FRA	–0.014	0.509	0.000	0.412	0.002	0.433	–0.012	0.515
IRA	–0.013	0.476	–0.005	0.375	–0.007	0.388	–0.017	0.482

DGP2	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 1,000$	Bias	SD	Bias	SD	Bias	SD	Bias	SD
SDM	0.005	0.580	–0.001	0.391	–0.001	0.395	–0.004	0.470
PRA	0.022	0.851	–0.002	0.426	0.007	0.390	0.003	0.488
FRA	0.011	0.501	0.005	0.380	0.007	0.390	–0.006	0.433
IRA	0.014	0.430	0.003	0.298	–0.002	0.297	–0.008	0.340

5. Nonlinear Regression Adjustment

- Can use linear models and OLS for any kind of outcome Y: linear projection.
- Might we do better if Y has special features?
- Imbens and Rubin (2015, p. 128) warn against nonlinear regression adjustment, but argument is incomplete.

- There are a few nonlinear models/estimation methods leads to a consistent estimator of τ.
  - $Y \in \{0, 1\}$
  - $Y \in [0, 1]$
  - $Y \geq 0$ (discrete, continuous, or mixed)
- Consistency only guaranteed using full (separate) RA.

• Two combinations of models/estimation methods consistently estimate the ATE.

1. If $0 \leq Y \leq 1$ (binary, fractional) use

(a) Model: A logistic conditional mean function,

$$E(Y|\mathbf{X}, W = w) = \frac{\exp(\alpha_w + \mathbf{X}\boldsymbol{\beta}_w)}{1 + \exp(\alpha_w + \mathbf{X}\boldsymbol{\beta}_w)}$$

(b) Estimator: The Bernoulli quasi MLE.

- With this combination, μ_0 and μ_1 are consistently estimated even if the means are misspecified.

$$\hat{\mu}_{0,FRA} = N^{-1} \sum_{i=1}^{N} \frac{\exp(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0)}{1 + \exp(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0)}$$

$$\hat{\mu}_{1,FRA} = N^{-1} \sum_{i=1}^{N} \frac{\exp(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1)}{1 + \exp(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1)}$$

$$\hat{\tau}_{FRA} = \hat{\mu}_{1,FRA} - \hat{\mu}_{0,FRA}$$

- Get the proper standard error using the Stata `teffects` command with the `ra` option:

```
teffects ra (y x1 ... xK, logit) (w)
teffects ra (y x1 ... xK, flogit) (w)
```

- "By hand" (but incorrect standard error):

```
glm y x1 x2 ... xK if w == 0,
  fam(bin) link(logit) robust
predict y0hat
glm y x1 x2 ... xK if w == 1,
  fam(bin) link(logit) robust
predict y1hat
gen tehat = y1hat - y0hat
mean tehat
```

2. If $Y \geq 0$, use

(a) Model: An exponential mean,

$$E(Y|\mathbf{X}, W = w) = \exp(\alpha_w + \mathbf{X}\boldsymbol{\beta}_w)$$

(b) Estimator: Poisson quasi-log likelihood.

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \left[\exp(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0) \right]$$

- Just as robust as linear FRA.

• Robustness comes because both methods use the canonical link function in the linear exponential family.

• If $\alpha_1^*,\ \boldsymbol{\beta}_1^*$ are the plims from the QMLE for

$$E(Y|\mathbf{X}, W = 1)$$

can show

$$\mu_1 = E[m(\mathbf{X}, \alpha_1^*, \boldsymbol{\beta}_1^*)]$$

even though $m(\cdot)$ is misspecified for the conditional mean.

- Only holds for special choices of model $m(\cdot)$ and the quasi-log likelihood.
  - Linear-Homoskedastic Normal (OLS)
  - Logistic-Bernoulli
  - Exponential-Poisson
- All have in common that the residuals always add to zero when an intercept is included in the estimation.

• Consistency follows from Wooldridge (2007, Journal of Econometrics) on doubly robust IPWRA estimators.

• In observational studies, weight the quasi-log likelihood by the inverse of the propensity scores:

$$p(\mathbf{X}_i) = P(W_i = 1|\mathbf{X}_i)$$
$$1 - p(\mathbf{X}_i) = P(W_i = 0|\mathbf{X}_i)$$

• The resulting estimates of the ATEs are "doubly robust."

1. Estimate a binary response model (flexible logit, probit, hetprobit, and so on) and obtain the propensity scores, and $p(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})$.

2. Using $W_i = 0$ and $W_i = 1$ separately, use a weighted estimation method, with weights $1/[1 - p(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})]$ and $1/p(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})$, respectively.

- In the linear case,

$$\min_{\alpha_0,\boldsymbol{\beta}_0} \sum_{i=1}^{n}(1-W_i)(Y_i-\alpha_0-\mathbf{X}_i\boldsymbol{\beta}_0)^2/[1-p(\mathbf{X}_i,\hat{\boldsymbol{\gamma}})]$$

$$\min_{\alpha_1,\boldsymbol{\beta}_1} \sum_{i=1}^{n}(Y_i-\alpha_1-\mathbf{X}_i\boldsymbol{\beta}_1)^2/p(\mathbf{X}_i,\hat{\boldsymbol{\gamma}})$$

$$\hat{\tau}_{IPWRA} = \left(\hat{\alpha}_1+\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1\right)-\left(\hat{\alpha}_0+\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\right)$$

- In randomized trials,

$$P(W_i = 1|\mathbf{X}_i) = P(W_i = 1) = \rho$$

so IPW weighting has no effect.

- So far, no proof that nonlinear regression adjustment results in asymptotic efficiency gains over SDM or linear FRA.
- Simulations show the gains can be nontrivial.

Simulations

- Data generated to be fractional probit outcome, index quadratic in X_1, X_2.
- True ATE is 0.075.
- Estimation is logit with only linear function of X_1, X_2.
- The logit model is (badly) "misspecified."

	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD
SDM	.0008	0.064	.0002	0.058	0.0017	0.063
Linear-PRA	0.0007	0.041	–0.00005	0.035	–0.0004	0.037
Linear-FRA	0.0019	0.041	–0.000003	0.035	–0.0014	0.037
Logit-FRA	–0.0009	0.028	–0.0008	0.026	–0.0005	0.030

- Data generated to be product of lognormal and Poisson, exponential mean, index quadratic in X_1 and X_2.
- Y is actually continuous.
- Linear regression adjustment can be used to improve efficiency.
- Separate Poisson RA using index linear in X_1, X_2.

- True ATE is about 2.2.

	$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
$N = 200$	Bias	SD	Bias	SD	Bias	SD
SDM	–0.081	1.428	–0.098	1.199	–0.112	1.226
Linear-PRA	–0.092	1.166	–0.098	0.989	–0.124	1.071
Linear-FRA	–0.149	1.133	–0.098	0.986	–0.086	1.057
Poisson-FRA	–0.025	0.706	–0.006	0.599	–0.022	0.597

6. More than Two Treatment Levels

- Let $Y(g), g = 1, \ldots, G$ be the potential outcomes.
- Means of potential outcomes:

$$\mu_g = E[Y(g)], g = 1, \ldots, G$$

- Treatment indicators

$$\mathbf{W} = (W_1, \ldots, W_G).$$

- Interested in linear combinations of the μ_g.
- For example, difference-in-differences.
- Estimating lower bound willingness-to-pay using randomized bid values.
- Let

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_G)'$$

be the vector of means.

• If $\hat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}$ are consistent and $\sqrt{N}$-asymptotically normal, would like to rank the $G \times G$ matrices

$$AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\right]$$

$$AVar\left[\sqrt{N}\left(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\right]$$

• If $AVar\left[\sqrt{N}\left(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\right] - AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)\right]$ is PSD, all linear combinations are estimated more efficiently using $\hat{\boldsymbol{\mu}}$.

- Still assume random assignment:

$$\mathbf{W} \text{ is independent of } [Y(1), Y(2), \ldots, Y(G), \mathbf{X}].$$

- Still assume random sampling.
- The previous estimators all have extensions to this case.
- Within-group sample averages:

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} W_{ig} Y_i, g = 1, \ldots, G$$

- Let

$$\hat{\boldsymbol{\mu}}_{SM} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_G)'$$

- Pooled RA: $\hat{\boldsymbol{\mu}}_{PRA}$ from

$$Y_i \text{ on } W_{i1}, W_{i2}, \ldots, W_{iG}, \mathbf{X}_i - \bar{\mathbf{X}}, i = 1, \ldots, N$$

- Full RA: Obtain $\hat{\alpha}_{g,FRA}$ and $\hat{\boldsymbol{\beta}}_{g,FRA}$ from

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_{ig} = 1$$

$$\hat{\mu}_g = \hat{\alpha}_g + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_g$$

- Negi and Wooldridge (in progress):

$$AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM}-\boldsymbol{\mu}\right)\right]-AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]$$

is PSD.

$$AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{PRA}-\boldsymbol{\mu}\right)\right]-AVar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]$$

is PSD.

- Pooled RA is efficient if

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_G$$

- The equal assignment case

$$\rho_g = P(W_g = 1) = 1/G, g = 1, \ldots, G$$

no longer ensures Pooled RA is efficient.

- Any linear combination of $\{\mu_1, \mu_2, \ldots, \mu_G\}$ is estimated most efficiently (asymptotically) using full (separate) regression adjustment.
- Moreover, we can use the nonlinear methods (logit, Poisson) on each of the treatment groups of Y is and LDV.
- In observational studies, IPWRA extends immediately: Estimate the $p_g(\mathbf{X}_i)$ by, say, multinomial logit.

7. Difference-in-Differences

- C is the control group, T is the treatment group.
- B is the before period, A is the after period.
- The standard DID treatment effect is

$$\tau = (\mu_{TA} - \mu_{TB}) - (\mu_{CA} - \mu_{CB})$$

- Estimating the means by separate regression adjustment is generally better than not controlling for covariates, or putting them in additively.

- Equivalently, supplement the usual DID regression by interacting covariates with dummies.
- Demean covariates before interacting with

$$D_T \cdot D_A$$

- Application to monitoring attendance in principles of microeconomics.
- Lecture 2 is the treated group.
- 2005: Before period, no monitoring.

```
. reg score d06 dlec2 d06_dlec2, robust

Linear regression                               Number of obs     =      1,338
                                                F(3, 1334)        =      11.
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0231
                                                Root MSE          =     6.0907

------------------------------------------------------------------------------
             |               Robust
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |  -.2538108   .4788325    -0.53   0.596    -1.193158    .6855359
       dlec2 |  -1.667177   .4529949    -3.68   0.000    -2.555837   -.7785167
   d06_dlec2 |   2.885365   .6663384     4.33   0.000     1.578179    4.19255
       _cons |   79.50235   .3299327   240.97   0.000     78.85511    80.1496
------------------------------------------------------------------------------
```

```
. reg score d06 dlec2 d06_dlec2 priGPA ACT, robust

Linear regression                               Number of obs     =      1,338
                                                F(5, 1332)        =     135.
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3345
                                                Root MSE          =     5.0307

------------------------------------------------------------------------------
             |               Robust
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |  -.3312425   .3975366    -0.83   0.405    -1.111109    .4486236
       dlec2 |  -1.743972   .3760746    -4.64   0.000    -2.481735   -1.006209
   d06_dlec2 |   3.031274     .54996     5.51   0.000     1.952392    4.110157
      priGPA |   3.777703   .2620455    14.42   0.000     3.263636     4.29177
         ACT |   .6063293   .0444399    13.64   0.000     .5191494    .6935092
       _cons |   56.12003   1.017628    55.15   0.000      54.1237    58.11636
------------------------------------------------------------------------------
```

```
. sum priGPA

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      priGPA |      1,338    2.587402    .5436816        .857       3.93

. gen priGPA_dm = priGPA - r(mean)

. sum ACT

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         ACT |      1,338    22.51121     3.49894         13         32

. gen ACT_dm = ACT - r(mean)
. gen d06_dlec2_priGPA_dm = d06_dlec2*priGPA_dm
. gen d06_dlec2_ACT_dm = d06_dlec2*ACT_dm
```

```
. reg score d06 dlec2 d06_dlec2 priGPA ACT ///
>         c.d06#c.priGPA_dm c.dlec2#c.priGPA_dm c.d06#c.ACT_dm c.dlec2#c.ACT
>         d06_dlec2*priGPA_dm d06_dlec2_ACT_dm, robust

Linear regression                               Number of obs     =      1,338
                                                F(11, 1326)       =       61.
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3357
                                                Root MSE          =     5.0377

------------------------------------------------------------------------------
             |               Robust
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf.
-------------+----------------------------------------------------------------
         d06 |  -.3568272     .40052    -0.89   0.373    -1.142549
       dlec2 |  -1.743008   .3771367    -4.62   0.000    -2.482858
  d06_dlec2 |   3.034592   .5525984     5.49   0.000     1.950529
     priGPA |     3.3895   .5335311     6.35   0.000     2.342842
         ACT |   .6518106   .0873472     7.46   0.000     .4804569
```

```
  c.d06#c.priGPA_dm |   .7347998   .7726823     0.95   0.342    -.7810133
                    |
c.dlec2#c.priGPA_dm |   .4326894   .7022318     0.62   0.538     -.944917
                    |
      c.d06#c.ACT_dm |  -.0457111   .1258555    -0.36   0.717    -.2926088
                    |
   c.dlec2#c.ACT_dm |  -.1305484   .1227289    -1.06   0.288    -.3713124
                    |
d06_dlec2_priGPA_dm |  -.8012812   1.053115    -0.76   0.447    -2.867235
   d06_dlec2_ACT_dm |   .1704278   .1775869     0.96   0.337    -.1779541
              _cons |   56.11503   2.013184    27.87   0.000     52.16566
------------------------------------------------------------------------------
```

```
. test c.d06#c.priGPA_dm c.dlec2#c.priGPA_dm c.d06#c.ACT_dm ///
>         c.dlec2#c.ACT_dm d06_dlec2_priGPA_dm  d06_dlec2_ACT_dm

 ( 1)  c.d06#c.priGPA_dm = 0
 ( 2)  c.dlec2#c.priGPA_dm = 0
 ( 3)  c.d06#c.ACT_dm = 0
 ( 4)  c.dlec2#c.ACT_dm = 0
 ( 5)  d06_dlec2_priGPA_dm = 0
 ( 6)  d06_dlec2_ACT_dm = 0

       F(  6,  1326) =    0.39
            Prob > F =    0.8865
.
. tab g

          g |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        340       25.41       25.41
          2 |        343       25.64       51.05
          3 |        340       25.41       76.46
          4 |        315       23.54      100.00
------------+-----------------------------------
      Total |      1,338      100.00
```

8. Estimating Lower Bound WTP

- A common lower bound WTP estimator is based on the area under the WTP survival function:

$$E(WTP) = \int_0^\infty S(a)da$$

- Let $b_1, b_2, ..., b_G$ be G bid values.

$$Y(g) = 1[WTP > b_g], g = 1, \ldots, G.$$

$$E[Y(g)] = P(WTP > b_g) = S(b_g)$$

- The ABERS (1955) estimator without monotonicity imposed ($b_0 = 0$):

$$\hat{\mu}_{ABERS} = \sum_{g=1}^{G} (b_g - b_{g-1}) \bar{Y}_g$$

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} Y_i 1[B_i = b_g]$$

is the fraction of yes votes at bid value b_g.

- Lewbel (2000, Journal of Econometrics) and

Watanabe (2010, AJAE) allow for covariates.

- Point: $\hat{\mu}_{LW}$ is a linear combination of means at different treatment levels.
- The simple means, $\bar{Y}_g$, are obtained from

$$Y_i \text{ on } Bid1_i, Bid2_i, ..., BidG_i, i = 1, \ldots, N$$

- Full regression adjustment:

1. Use the linear regressions

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ if } Bid_i = g.$$

2. Or use logistic regression for each g.

- Application to Cal Oil Data (Carson et al., 2004)

```
. reg vote bid5 bid25 bid65 bid120 bid220, nocons robust

Linear regression                               Number of obs     =      1,085
                                                F(5, 1080)        =     241.
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5274
                                                Root MSE          =      .4816

------------------------------------------------------------------------------
             |               Robust
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bid5 |   .6894977   .0313386    22.00   0.000     .6280063    .7509892
       bid25 |   .5694444   .0337689    16.86   0.000     .5031844    .6357044
       bid65 |   .4854772   .0322687    15.04   0.000     .4221608    .5487936
      bid120 |   .4033149   .0365476    11.04   0.000     .3316026    .4750272
      bid220 |   .2894737   .0301044     9.62   0.000     .2304039    .3485435
------------------------------------------------------------------------------

. lincom bid5*5 + bid25*(25 - 5) + bid65*(65 - 25) + bid120*(120 - 65) ///
>         + bid220*(220 - 120)

 ( 1)  5*bid5 + 20*bid25 + 40*bid65 + 55*bid120 + 100*bid220 = 0
```

```
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
         (1) |   85.38515    3.90513    21.86   0.000     77.72265    93.04765
------------------------------------------------------------------------------
```

```
. teffects ra (vote c.linc c.linc#c.linc inc_miss i.notax i.lowspend ccoast
>         i.notax#c.linc i.lowspend#c.linc coastip wildip envist) (bid), pomeans

Iteration 0:   EE criterion =  3.217e-24
Iteration 1:   EE criterion =  6.808e-31

Treatment-effects estimation                    Number of obs     =      1,085
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
POmeans      |
         bid |
           5 |   .6846359   .0287713    23.80   0.000     .6282452    .7410266
          25 |   .5968978   .0306565    19.47   0.000     .5368121    .6569835
          65 |   .4890953   .0293534    16.66   0.000     .4315638    .5466268
         120 |   .3780313    .033163    11.40   0.000      .313033    .4430297
         220 |   .2895015   .0285561    10.14   0.000     .2335326    .3454703
------------------------------------------------------------------------------

.
. lincom _b[POmeans:5bn.bid]*5 + _b[POmeans:25.bid]*(25 - 5)
    + _b[POmeans:65.bid]*(65 - 25) + _b[POmeans:120.bid]*(120 - 65)
    + _b[POmeans:220.bid]*(220 - 120)
```

```
 ( 1)  5*[POmeans]5bn.bid + 20*[POmeans]25.bid + 40*[POmeans]65.bid
    + 55*[POmeans]120.bid + 100*[POmeans]220.bid = 0

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval
-------------+----------------------------------------------------------------
         (1) |   84.66682   3.792276    22.33   0.000     77.23409    92.09954
------------------------------------------------------------------------------
```

9. Summary and Further Comments

- In the linear case, need to use full RA to ensure (asymptotic) efficiency gains.
  - ▸ Except in special cases.
- In nonlinear case:
  - ▸ Need to use full RA to ensure consistency.
  - ▸ Only certain combinations of mean functions/QMLEs ensure consistency.

- In observational studies, full RA also has important benefits.
- Słoczyński (2017, "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands")
  - Shows theat pooled RA estimates

$$P(W = 1) \cdot \tau_{atu} + P(W = 0)\tau_{att}$$

  - The weights are reversed.