Two-stage Regression for Treatment Effect Estimation

A Dissertation Defense

Josh Errickson

August 22, 2016

1 Peters-Belson with Prognostic Heterogeneity

2 Further Extensions to PBPH

3 Linear Treatment Effect with Binary Response

Peters-Belson with Prognostic Heterogeneity

Motivation

- Treatment effect framework.
- Often post-hoc subgroup analysis is of interest.
- Subgroups can be based on “at risk”, defined by predicted risk (response).
    - We consider continuous predicted risk.
- Answered in current literature with Two-stage Peters-Belson method.
- Incorrectly applied
    - Underestimates S.E. of coefficient representing additional effect based on predicted risk.
    - High Type I error
- Approach based on estimating equations addresses the concern.
    - With a modest complication.

Examples in Literature

- Giné, Goldberg, and Yang (2012)
  - Fingerprint identification in Malawi increased repayment of loans, especially for those least likely to repay in absence of treatment.
- Dynarski, Hyman, and Schanzenbach (2011)
  - Small class size increased college enrollment only those in the lowest quintile of predicted college enrollment.
- Goldrick-Rab et al. (2011)
  - Financial aid for post-secondary education only beneficial for those in highest third of predicted drop-out.

Giné et al. Results

(Outcome is percentage repaid by due date)
Subgroup by Predicted Repayment:

Fingerprint:	Q1 (Lowest)	Q2	Q3	Q4	Q5 (Highest)
	.506 ***	.056	-.001	-.040	-.075 *

Continuous:

	Coef (SE)	
Fingerprint	0.794 (.045)	***
Fingerprint:Predicted Repayment	-0.896 (.043)	***

Giné et al. Simulation

- Using Giné et al.'s data, drop treatment group.
- Split control group into faux treatment/faux control groups.
  - On average, all treatment effects are 0.
- Confidence interval coverage of 0 only 72% (over 10,000 runs).

Causal Inference

- Goal is to estimate treatment effect.
- Z is indicator of treatment, let Y_c be the potential response that would be observed under control, and Y_t under treatment,

$$Y_{\text{observed}} = ZY_t + (1 - Z)Y_c. \tag{1}$$

- Want to estimate $\mathbb{E}(Y_t - Y_c)$.
- Fundamental Problem of Causal Inference

Peters-Belson method

- Use control group to predict Y_c.

$$Y_c = X\beta + \epsilon, Z = 0. \tag{2}$$

- Estimate treatment effect in treatment group,

$$Y_t - \hat{Y}_c, Z = 1. \tag{3}$$

Formalization

Model is two-stage regression. X are covariates, Y response ($Y = ZY_t + (1-Z)Y_c$), Z treatment indicator.
Stage 1: In the control group ($Z = 0$),

$$Y = X\beta_c + \delta. \tag{4}$$

Stage 2: In the treatment group ($Z = 1$),

$$Y - X\hat{\beta}_c = \tau + \eta X\hat{\beta}_c + \epsilon. \tag{5}$$

(Note that $Y = Y_c$ in Stage 1 and $Y = Y_t$ in Stage 2.)

Formalization

Model is two-stage regression. X are covariates, Y response ($Y = ZY_t + (1-Z)Y_c$), Z treatment indicator.
Stage 1: In the control group ($Z = 0$),

$$Y = X\beta_c + \delta. \tag{4}$$

Stage 2: ~~In the treatment group ($Z = 1$),~~

$$Y - X\hat{\beta}_c = \tau Z + \eta Z X\hat{\beta}_c + \epsilon. \tag{5}$$

(Note that $Y = Y_c$ in Stage 1 and $Y = Y_t$ in Stage 2.)

Formalization

Model is two-stage regression. X are covariates, Y response ($Y = ZY_t + (1 - Z)Y_c$), Z treatment indicator.
Stage 1: In the control group ($Z = 0$),

$$Y = X\beta_c + \delta. \tag{4}$$

Stage 2: In the treatment group ($Z = 1$),

$$Y - X\hat{\beta}_c = \tau + \eta X\hat{\beta}_c + \epsilon. \tag{5}$$

(Note that $Y = Y_c$ in Stage 1 and $Y = Y_t$ in Stage 2.)

Feasible Values of η

$$Y - X\hat{\beta}_c = \tau + \eta X\hat{\beta}_c + \epsilon, Z = 1 \tag{6}$$

- $\eta = 0$

$$Y - X\hat{\beta}_c = \tau + \epsilon \tag{7}$$

- $\eta = -1$

$$Y = \tau + \epsilon \tag{8}$$

- $\eta < -1$: effect of X is reversed in treatment group.
- η large & positive: effect of X substantially stronger in treatment group.

Fixes in Literature

- Cross-validation
  - Abadie, Chingos, and West (2013) show this produces proper coverage
  - Computationally "intensive."
- Out-of-sample first stage
  - Hayward et al. (2006)
  - Not feasible in most settings.

Correcting Standard Error

- Standard OLS does not take into account variance in $\hat{\beta}_c$ from Stage 1.
- Instead use a robust Sandwich estimator based upon estimating equations from M-estimators.

$$\begin{pmatrix} \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sum_{\{i:Z_i=0\}} (Y_i - X_i'\beta_c)X_i \\ \sum_{\{i:z_i=1\}} (Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c) \begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix} \end{pmatrix} \tag{9}$$

Correcting Standard Error

- Long story short, end with a Sandwich form of the covariance

$$B(\theta)^{-1}M(\theta)B(\theta)^{-1}, \tag{10}$$

where $\theta = (\beta_c, \tau, \eta)$, B is derivative of estimating equation and M is the variance of the estimating equation.

Test Inversion

- Wald-style confidence interval fails.
- Use old idea of inverting a hypothesis test.
- Hypothesize that $H_0 : \eta = \eta_0$. Confidence region is all η_0 such that we fail to reject H_0.
- Reject if

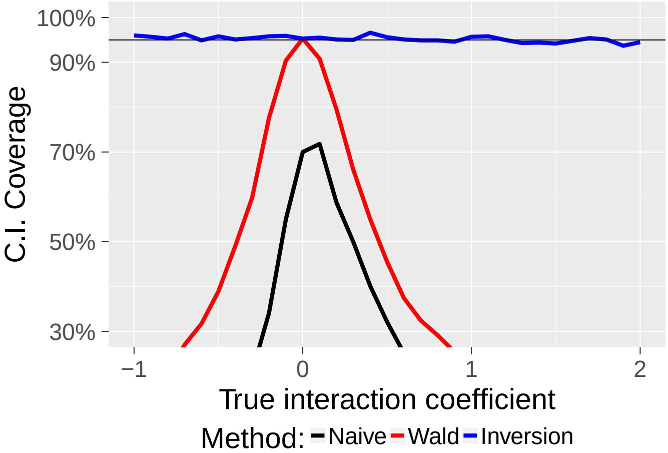
$$w_\alpha(\eta_0) := (\hat{\eta} - \eta_0)^2 - \left(\chi^2_{(1-\alpha)}(1)\right)^* \mathrm{Var}(\eta_0) \geq 0. \qquad (11)$$

- Squaring makes $w_\alpha(\eta_0)$ quadratic in η_0, so confidence region $(\eta_0 : w_\alpha(\eta_0) \leq 0)$ will be continuous.

Solving for Confidence Region

- Since $w_\alpha(\eta_0)$ is quadratic in η_0, in theory we can solve for and interpret its roots.
- In practice, quite infeasible.
  - Coefficient on η_0^2 is half a page long.
- Can solve $w_\alpha(\eta_0)$ for 3 values of η_0, fit linear regression with quadratic η_0 term.

Simulation



Coverage with $n = 100$ and various η, with 1,000 iterations per η.

Method advice

- Focus on first stage model fit
- Poor fitting first stage yields wide to infinite condfidence intervals.

Correcting Giné et al.

	Estimate	Standard Error	Confidence Interval
Published	-.896	.043	$(-0.980, -0.812)$
Corrected	-.896	.054	$(-0.998, -0.781)$

Further Extensions to PBPH

GLM First Stage

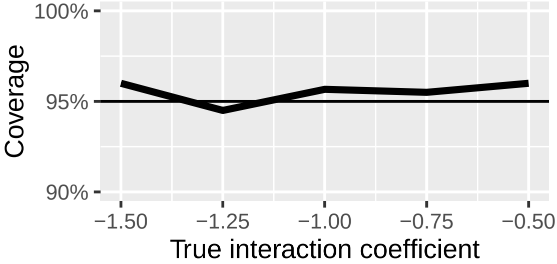
- First stage (relationship between Y_c and X) allowed to be GLM.
- Second stage remains least squares.
  - E.g. $Y \in \{0, 1\}$ but $Y - \hat{Y}_c \ \in [-1, 1]$.
  - $Y - \hat{Y}_c$ unlikely to have same error as Y.
- Similar SE calculations, use test inversion.
- Works for any GLM with a canonical link.

Clustered Random Trials

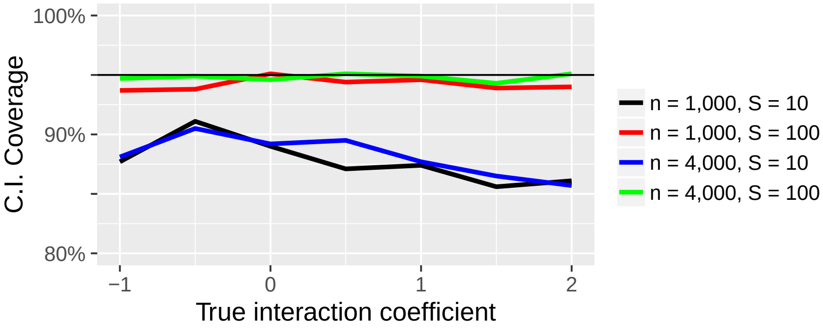
- Randomization at the cluster level can be convenient.
    - Loss of power due to intracluster correlation.
- Typically addressed with Sandwich estimators, so an easy adjustment.
- Again similar SE calculations and test inversion.

Logistic First Stage First Stage

- $n = 1{,}000$ over 1,000 replicates.
- Restrictions on η, τ
  - Dependent on link function, based on bounds of $(Y - \hat{Y}_C)$



Clustered Random Trials



Giné et al. Results

- Giné et al. had randomized clusters (clubs).

	Estimate	Standard Error	Confidence Interval
Published	-.896	.043	$(-0.980, -0.812)$
Corrected	-.896	.054	$(-0.998, -0.781)$
with Clusters	-.896	.109	$(-1.110, -0.635)$

We can no longer reject $H_0 : \eta = -1$.

Giné et al. Results with Binary Response

Now consider response as full repayment.

	Estimate	Standard Error	Confidence Interval
Published	-.994	.051	$(-1.094, -0.894)$
Corrected	-.994	.052	$(-1.100, -0.888)$
with Clusters	-.994	.111	$(-1.237, -0.748)$

Linear Treatment Effect with Binary Response

Motivation

- With a binary response, treatment effect may be linear on the probability scale.
    - Logistic regression forces treatment effect to be linear on logit scale.
    - Linear regression forces covariates X to be on linear scale as well.
- Two-stage regression can separate the relationships.

Linear vs Logistic

- First stage, only in control group,

$$\text{logit}(Y_c) = X\beta_c. \tag{12}$$

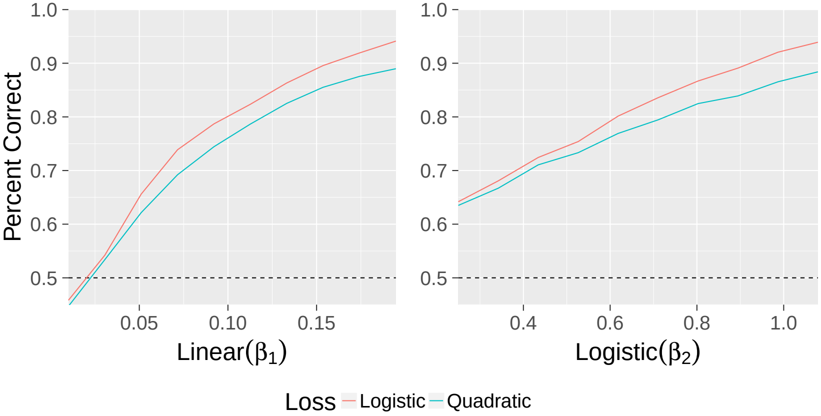
- Two forms of second stage,

$$\begin{aligned} Y &= Z\beta_1 + \hat{Y}_c, \\ \text{logit}(Y) &= Z\beta_2 + \text{logit}\left(\hat{Y}_c\right). \end{aligned} \tag{13}$$

Comparing Linear vs Logistic

- Model comparison using estimated risk:
  - Linear regression: Quadratic loss
  - Logistic regression: Logistic loss
- Need to choose a loss to compare across models

Linear vs Logistic



Stratified Data with Binary Response

- Common method of analysis is Conditional Logistic Regression
  - Likelihood is maximized conditional on nuisance parameters.
- Conditioning on matched sets lacks meaning.
- Suggest two-stage model that accounts for strata.

Heuristic Test of Linearity in Probability

- Let $\hat{Y}_s$ be the proportion of 1 responses in strata s, let

$$\lambda_{is} = \frac{1}{\hat{Y}_s(1 - \hat{Y}_s)}. \tag{14}$$

  - λ_{is} upweights observations in strata with $\hat{Y}_s$ closer to 0 or 1.
- $Z\lambda$ roughly emulates a treatment effect linear in probability.
- Both stages logistic, two variations of second stage with Z or $Z\lambda$.

Linear Treatment Effect with Stratification

- First stage remains logistic on control group only.
- Possible forms of second stage:

1. Ignoring stratification:

$$Y = Z\tau + \hat{Y}_c \tag{15}$$

2. Stratification with fixed effects:

$$Y = Z\tau_f + S\kappa_f + \hat{Y}_c \tag{16}$$

3. (15) with weights

Weights

- Let

$$\delta_i = \begin{cases} \frac{\sum_{j:S_j=S_i} Z_j}{\sum_{j:S_j=S_i} 1}, & Z_i = 1, \\ \frac{\sum_{j:S_j=S_i} (1-Z_j)}{\sum_{j:S_j=S_i} 1}, & Z_i = 0. \end{cases} \quad (17)$$

- δ_i is proportion of observations in strata which observation i belongs to that share the same treatment status as observation i.
- Weight (15) by $w_i = \delta_i^{-1} / \sum_j \delta_j^{-1}$.

Weighted model

- Second stage model is weighted least squares,

$$Y = Z\tau_w + \hat{Y}_c, \tag{18}$$

where

$$\hat{\tau}_w = \frac{\sum_i w_i Z_i (Y_i - \hat{Y}_{ci})}{\sum_i w_i Z_i}. \tag{19}$$

- Standard error associated with any of these second stages requires Sandwich estimation.

$\hat{\tau}_f$ vs $\hat{\tau}_w$

- If treatment effect is constant across matched sets,
  - both are estimates of that constant treatment effect.
- If treatment effect varies across matched sets,
  - $\hat{\tau}_f$ is some weighted average of set-specific estimated treatment effects.
  - $\hat{\tau}_w$ is still estimating average treatment effect.

Ignoring Decision Criteria

- We've advocated a data-driven choice of second stage (linear vs logistic).
- There are additional benefits to each approach:
  - 2nd stage linear:
    - Estimate of treatment effect is consistent regardless of whether treatment effects are actually linear on the probability scale.
  - 2nd stage logistic:
    - Odds ratios are reversible (e.g. OR of rate given exposure = OR of exposure given rates).

Gurm et al. (2013)

- Authors estimate effect of using vascular closure devices on the risk of vascular complications following arterial access.
    - Conditional Logistic Regression on matched sets.
- Found effect size of odds ratio 0.78 and standard error of 0.06.

Gurm et al. (2013)

- Fitting two stage logistic models with Z and $Z\lambda$,

	Z	$Z\lambda$
Estimated Risk	0.0917	0.0808
AIC	12080	10660

- Fit second stage linear model with weights w_i, linear treatment effect estimated as -0.002 with standard error of 0.0019.

Thank you!

S.E. Derivation, 1

$$
\begin{pmatrix} \mathbf{0} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{\{i:Z_i=0\}} (Y_i - X_i'\beta_c)X_i \\ \displaystyle\sum_{\{i:z_i=1\}} (Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c)\begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix} \end{pmatrix}
$$

$$
= \begin{pmatrix} \Phi(Y;\beta_c) \\ \Psi(Y,\beta_c;\tau,\eta) \end{pmatrix}
$$

$$
B^{(c)}(\beta_c,\tau,\eta) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}
$$

$$
= \begin{bmatrix} \mathbb{E}\,\frac{\partial}{\partial\beta_c}\Phi(Y;\beta_c) & \mathbb{E}\,\frac{\partial}{\partial(\tau,\eta)}\Phi(Y;\beta_c) \\ \mathbb{E}\,\frac{\partial}{\partial\beta_c}\Psi(Y,\beta_c;\tau,\eta) & \mathbb{E}\,\frac{\partial}{\partial(\tau,\eta)}\Psi(Y,\beta_c;\tau,\eta) \end{bmatrix}
$$

Same form for M, the meat.

S.E. Derivation, 2

$$
\begin{aligned}
B_{11} &= \sum_{\{i:Z_i=0\}} X_i X_i', \\
B_{12} &= 0, \\
B_{21} &= \sum_{\{i:Z_i=1\}} \mathbb{E}\left(\begin{array}{c} -(1+\eta)X_i \\ (Y_i - \tau - 2(1+\eta)X_i\beta_c)X_i \end{array} \right), \\
B_{22} &= \sum_{\{i:Z_i=1\}} \mathbb{E}\left[\begin{array}{cc} 1 & X_i\beta_c \\ X_i\beta_c & (X_i\beta_c)^2 \end{array} \right].
\end{aligned}
$$

S.E. Derivation, 3

$$M_{11} = \sum_{\{i:Z_i=0\}} \text{Var}\left(Y_i - X_i\beta_c\right) X_i X_i',$$

$$M_{12} = 0,$$

$$M_{21} = 0,$$

$$M_{22} = \sum_{\{i:Z_i=1\}} \text{Var}\left((Y_i - X_i'\beta_c - \tau - \eta X_i'\beta_c) \begin{pmatrix} 1 \\ X_i'\beta_c \end{pmatrix} \right).$$

S.E. Derivation, 4

Covariance in second stage is lower right 2 × 2 sub-matrix of $B_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})^{-1} M_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta}) B_{n_t}^{(c)}(\hat{\beta}_c, \hat{\tau}, \hat{\eta})^{-T}$.

$$\widehat{\mathrm{Var}}(\tau, \eta) = \hat{B}_{22}^{-1} \left(\hat{M}_{22} + \hat{B}_{21} \hat{B}_{11}^{-1} \hat{M}_{11} \hat{B}_{11}^{-T} \hat{B}_{21}^{T} \right) \hat{B}_{22}^{-T}.$$

Showing $w_\alpha(\eta_0)$ is quadratic

$$w_\alpha(\eta_0) := (\hat{\eta} - \eta_0)^2 - \left(\chi^2_{(1-\alpha)}(1)\right)^* \mathrm{Var}(\eta_0).$$

$$\mathrm{Var}(\eta_0) = B(\theta)^{-1} M(\theta) B(\theta)^{-1},$$

- Four variations (Lindsay and Qu (2003))
    1. $B(\theta_0)^{-1} M(\theta_0) B(\theta_0)^{-1}$
    2. $B(\hat{\theta})^{-1} M(\hat{\theta}) B(\hat{\theta})^{-1}$
    3. $B(\hat{\theta})^{-1} M(\theta_0) B(\hat{\theta})^{-1}$
    4. $B(\theta_0)^{-1} M(\hat{\theta}) B(\theta_0)^{-1}$
- 4 performed best in simulations.

n vs p

- Develop a finite sample rule of thumb for size of p
- He and Shao (2000) proved asymptotic rule

$$p^2 \log(p) = o(n). \tag{20}$$

- We consider slower growth rate

$$p^2 \log(p)^2 < Cn \tag{21}$$

  for some fixed C.
- Simulations chose $C = 2.5$.
- For $n = 100$, use $p = 7$.

Abadie, Alberto, Matthew M Chingos, and Martin R West. 2013. "Endogenous Stratification in Randomized Experiments." National Bureau of Economic Research.

Dynarski, Susan, Joshua M. Hyman, and Diane Schanzenbach. 2011. "Experimental Evidence on the Effect of Childhood Investments on Post-secondary Attainment and Degree Completion."

Giné, Xavier, Jessica Goldberg, and Dean Yang. 2012. "Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi." *American Economic Review* 102 (6): 2923–54.

Goldrick-Rab, Sara, Douglas N. Harris, James Benson, and Robert Kelchen. 2011. "Conditional cash transfers and college persistence: Evidence from a randomized need-based grant program."

Gurm, Hitinder S, Carrie Hosman, David Share, Mauro Moscucci, and Ben B Hansen. 2013. "Comparative Safety of Vascular Closure Devices and Manual Closure Among Patients Having Percutaneous