Causal Inference in Observational Studies with Non-Binary Treatments

David A. van Dyk

Statistics Section, Imperial College London

Joint work with Shandong Zhao and Kosuke Imai

Cass Business School, October 2013

Imperial College London

Outline

Imperial College London

Outline

Imperial College London

Salk's Polio Vaccine Trials

- NFIP 1954 Polio Vaccine Trials
  Treatment Group: Children whose parents give consent
  Control Group: Children whose parents **do not** consent
- Results:

The randomized controlled double-blinded experiment			*The NFIP study*		
	Size	Rate[1]		Size	Rate
Treatment	200,000	28	Vaccine (consent)	225,000	25
Control	200,000	71			
No Consent	350,000	46	No Vaccine/consent	125,000	44

[a] per 100,000

- Rows are comparable between the studies.
- Using No-Consent group as Control *biases* results.

Consent (treatment indicator) is correlated with polio rates in the absence of treatment (potential outcome).

Imperial College London

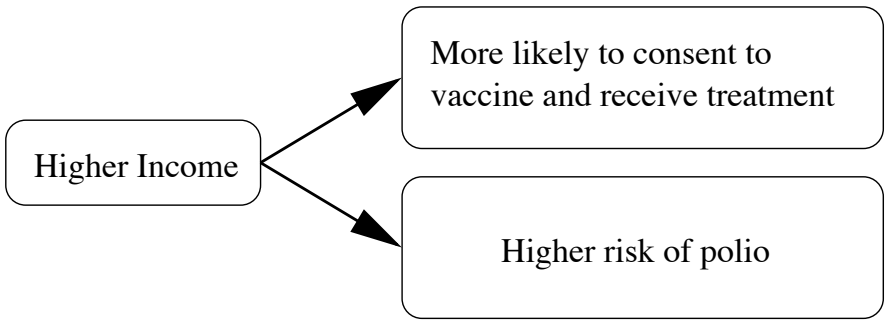
Causal Inference in Observational Studies

1. The Rubin Causal Model:

	Covariates	Treatment Group	Potential Outcomes	
			Control	Treatment
1	X_1	1	$Y_1(0)$	$\mathbf{Y_1(1)}$
2	X_2	1	$Y_2(0)$	$\mathbf{Y_2(1)}$
3	X_3	0	$\mathbf{Y_3(0)}$	$Y_3(1)$
⋮	⋮	⋮	⋮	⋮

Children with treatment more likely to contract polio in the absence of treatment.

2. E.g.: NFIP 1954 Polio Vaccine Trials



- ⋆ The treatment is correlated with the potential outcomes.
- ⋆ Biases results.
- ⋆ Key: Control for the (correct) covariates.

Imperial College London

Formalizing Causal Inference

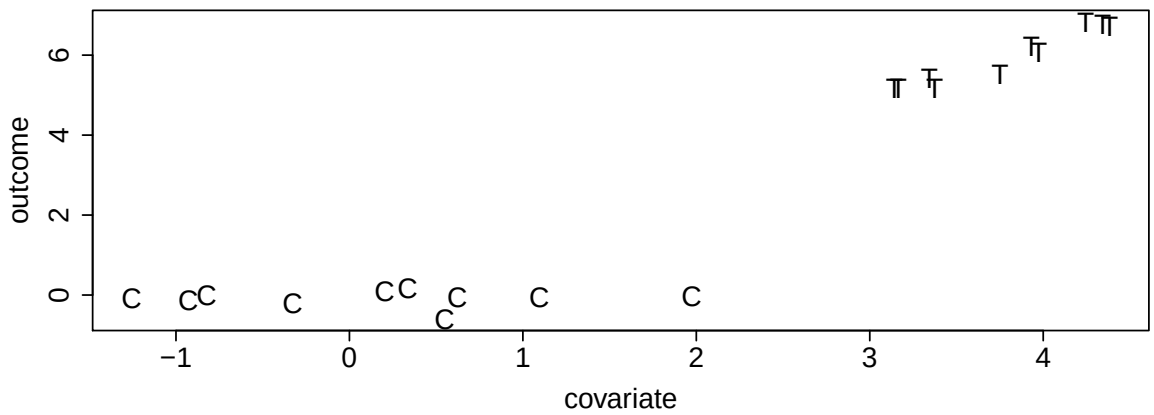
- Causal effect: e.g., $Y(t_1^P) - Y(t_2^P)$.
  - $Y(t^P)$ is potential outcome with potential treatment, $t^P \in \mathcal{T}$.
- Problem: Treatment, T, correlated w/ potential outcomes.
  - $Y_i(T_i = t_1^P) - Y_j(T_j = t_2^P)$ is not the causal effect.
- Key assumption (Strong Ignorability of Treatment Assign.):

$$p\{T \mid X\} \;=\; p\{T \mid Y(t^P), X\} \quad \forall\, t^P \in \mathcal{T}.$$

Under this assumption, we must adjust for the covariates in our analysis

Imperial College London

Adjusting for Covariates in Causal Inference



- Standard regression: $Y(t^P) \sim N(\alpha + X\beta + t^P\gamma, \sigma^2)$.
  - ⋆ common assumptions, e.g. linearity, are often violated.
  - ⋆ leads to biased causal inferences.

- Matching and Subclassification reduce bias:
  - ⋆ non/semi-parametric methods.
  - ⋆ but, difficult when the dimensionality of X is large.

Imperial College London

Propensity Score of Rosenbaum and Rubin (1983)

- If the treatment is binary, the propensity score,

$$e(X) = \Pr(T = 1 \mid X),$$

  fully characterizes $p(T \mid X)$.
- Key Theorems:
  1. The propensity score is a balancing score:

$$\Pr\{T = 1 \mid X, e(X)\} = \Pr\{T = 1 \mid e(X)\}.$$

  2. Strong Ignorability of Treatment Assignment Given $e(X)$:

$$E\{Y(t^P) \mid e(X)\} = E\{Y(t^P) \mid T = t^P, e(X)\} \quad \text{for} \quad t^P = 0, 1.$$

- Unbiased estimate of causal effect is possible given $e(X)$.
- Match or subclassify on $e(X)$, a scalar variable.

Imperial College London

The Power of Conditioning on Propensity Scores

The problem with (non-randomized) observational studies:

Treatment assignment correlated with potential outcomes.

E.g., subjects who are more likely to respond well *without* treatment are more likely to be in the control group.

The power of propensity scores:
In a subclass with the *same* value of the the propensity score,

Treatment is UNcorrelated with potential outcomes.

We can classify subjects based on their propensity score, and analyze the data separately in each class.

Imperial College London

Use of the Propensity Score in Observational Studies

- The propensity score is unknown in observational studies.
- Estimate $e(X)$, e.g., via logistic regression.
- Advantages:
  1. Diagnostics: check balance of X between treatment and control groups after matching or subclassifying on $e(X)$.

$$p\{X \mid T = 0, e(x)\} = p\{X \mid T = 1, e(x)\}$$

  2. Robust to misspecification of functional forms for estimating propensity score (Drake, 1993; Dehejia and Wahba, 1999).
- Disadvantage: vulnerable to unobserved confounders.

Imperial College London

Covariance Adjustment

- In addition to matching and subclassification, Rosenbaum and Rubin (1983) suggested *covariance adjustment:*
  - ⋆ Regress Y on $e(X)$ separately for treatment and control.

    Contol: $Y \sim \alpha_C + \beta_C\, e(X)$

    Treatment: $Y \sim \alpha_T + \beta_T\, e(X)$
  - ⋆ Average difference of fitted potential outcomes for each unit

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\alpha_T + \beta_T\, e(X_i) - \{\alpha_C + \beta_C\, e(X_i)\}\Big]$$

$$= \alpha_T - \alpha_C + (\beta_T - \beta_C)\frac{1}{n}\sum_{i=1}^{n} e(X_i)$$

- Covariance Adjustment seems less robust than matching of subclassifiation.

Imperial College London

Outline

Imperial College London

General Treatment Regimes

Goal: *Generalize propensity score to non-binary treatment.*

- Propensity score is confined to binary treatment scenarios.
- Researchers have no control over treatment in observational studies.
  - ⋆ Continuous treatment: dose response function.
  - ⋆ Ordinal treatment: effects of years in school on income.
  - ⋆ Event-count, duration, semi-continuous, etc. . . .
  - ⋆ Multivariate treatments

Two highly-cited methods

1. Generalized Propensity Score (Hirano and Imbens, 2004)
2. The Propensity Function (Imai and van Dyk, 2004)

Imperial College London

The Propensity Function

Definition: Conditional density of the actual treatment given observed covariates, $e(\cdot|X) \equiv p_\psi(T \mid X)$.

Uniquely Parameterized Propensity Function Assumption:

- $e(\cdot \mid X)$ depends on X only through $\theta_\psi(X)$.
- $\theta = \theta_\psi(X)$ uniquely represents $e\{\cdot \mid \theta_\psi(X)\}$.

Examples:

1. Continuous treatment: $T \mid X \sim N(X^\top \beta, \sigma^2)$.
   $\psi = (\beta, \sigma^2)$ and $\theta_\psi(X) = X^\top \beta$.
2. Categorical treatment: Multinomial probit for $\Pr(T \mid X)$.
   $\psi = (\beta, \Sigma)$ and $\theta_\psi(X) = \mathbf{X}^\top \beta$.
3. Ordinal treatment: Ordinal logistic model for $\Pr(T \mid X)$.
   $\psi = \beta$ and $\theta_\psi(X) = X^\top \beta$.

Imperial College London

The Propensity Function in Practice

Properties:

1. Balance: $p\{T \mid e(\cdot \mid X), X\} = p\{T \mid e(\cdot \mid X)\}$.
2. Ignorablity: $p\{Y(t) \mid T, e(\cdot \mid X)\} = p\{Y(t) \mid e(\cdot \mid X)\}\ \forall t$

Estimation of Causal Effects:

- Subclassification:

$$p\{Y(t)\} = \int p\{Y(t) \mid T = t, \theta\} p(\theta) d\theta \approx \sum_{j=1}^{J} p\{Y(t) \mid T = t, \hat{\theta}_j\}\, W_j.$$

  - ⋆ Subclassify observations into J subclasses based on $\hat{\theta}$.
  - ⋆ Regress $Y(T)$ on T (and $\hat{\theta}$) within each subclass.
  - ⋆ Average the within subclass average treatment effects.

Imperial College London

The Propensity Function in Practice

Estimation of Causal Effects:

- Subclassification:

$$p\{Y(t)\} = \int p\{Y(t) \mid T = t, \theta\} p(\theta) d\theta \approx \sum_{j=1}^{J} p\{Y(t) \mid T = t, \hat{\theta}_j\} W_j.$$

- Smooth Coefficient Model:

$$E(Y(t) \mid T = t, \hat{\theta}) \;=\; f(\hat{\theta}) + g(\hat{\theta}) \cdot T.$$

  - ⋆ $f(\cdot)$ and $g(\cdot)$ are unknown smooth continuous functions.
  - ⋆ Average treatment effect: $\frac{1}{n} \sum_{i=1}^{n} \hat{g}(\hat{\theta}_i)$

Imperial College London

The Generalized Propensity Score

Definition: Conditional density of treatment given observed covariates evaluated at observed treatment, $R = r(T|X)$.

Properties:

1. Balance: $I\{T = t\}$ is conditionally indep. of X given $r(t|X)$.
2. Ignorability: $p_T\{t|r(t,X), Y(t)\} = p_T\{t|r(t,X)\}$ for every t.

Estimation of Dose Response Function:

$$E(Y(t)|T = t, \hat{R}) = \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \alpha_3 \cdot \hat{R} + \alpha_4 \cdot \hat{R}^2 + \alpha_5 \cdot T \cdot \hat{R}.$$

$$\hat{E}\{Y(t)\} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot t + \hat{\alpha}_2 \cdot t^2 + \hat{\alpha}_3 \cdot \hat{r}(t, X_i) + \hat{\alpha}_4 \cdot \hat{r}(t, X_i)^2 + \hat{\alpha}_5 \cdot t \cdot \hat{r}(t, X_i) \right).$$

Note similarity with covariance adjustment.

Imperial College London

Comparing GPS with Covariance Adjustment

Covariance Adjustment with Binary Treatment:

- Regress Y on $e(X)$ separately for treatment and control.

  Contol: $Y \sim \alpha_C + \beta_C\, e(X)$

  Treatment: $Y \sim \alpha_T + \beta_T\, e(X)$

- Average difference of fitted potential outcomes

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\alpha_T + \beta_T\, e(X_i) - \big\{\alpha_C + \beta_C\, e(X_i)\big\}\Big]$$

Estimation of Dose Response Function with GPS:

$$E(Y(t)|T=t,\hat{R}) \;=\; \alpha_0+\alpha_1\cdot T+\alpha_2\cdot T^2+\alpha_3\cdot \hat{R}+\alpha_4\cdot \hat{R}^2+\alpha_5\cdot T\cdot \hat{R}.$$

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\alpha}_0+\hat{\alpha}_1\cdot t+\hat{\alpha}_2\cdot t^2+\hat{\alpha}_3\cdot \hat{r}(t,X_i)+\hat{\alpha}_4\cdot \hat{r}(t,X_i)^2+\hat{\alpha}_5\cdot t\cdot \hat{r}(t,X_i)\Big).$$

Imperial College London

Dose Response Functions

- Imai & van Dyk (2003) compute average treatment effect.
  Subclassification: Weighted average of within subclass coefficients of T.
  SCM: Average of individual coefficients of T, $\hat{g}(\hat{\theta}_i)$.

- Hirano & Imbens (2003) compute full Dose Response.

- The subclassification formula of Imai and van Dyk, however, gives a *parametric* Dose Response Function:

$$p\{Y(t)\} = \int p\{Y(t) \mid T = t, \theta\} p(\theta) d\theta \approx \sum_{j=1}^{J} p\{Y(t) \mid T = t, \theta_j\} \, W_j.$$

Imperial College London

Simulation 1

A very simple simulation:

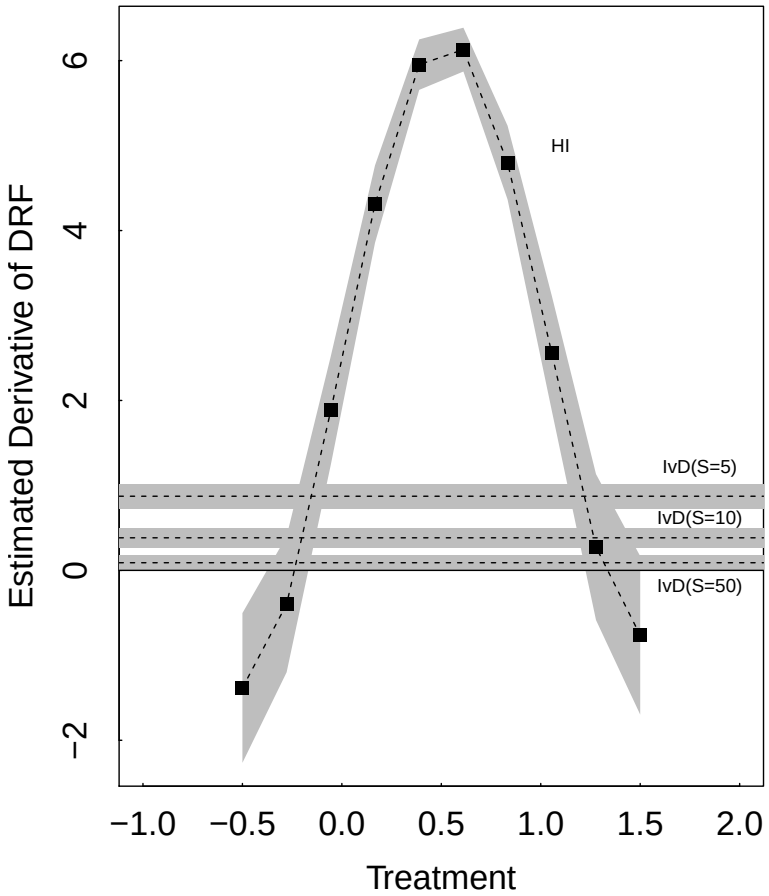
- $X \overset{\text{indep.}}{\sim} \mathcal{N}(0.5, 0.25)$, $T|X \overset{\text{indep.}}{\sim} \mathcal{N}(X, 0.25)$, and $n = 2000$
- $Y(t) \mid t, X \overset{\text{indep.}}{\sim} \mathcal{N}(10X, 1)$ for all $t \in \mathcal{T}$
- Linear Regression $Y \sim T$ gives a treatment effect of 5.

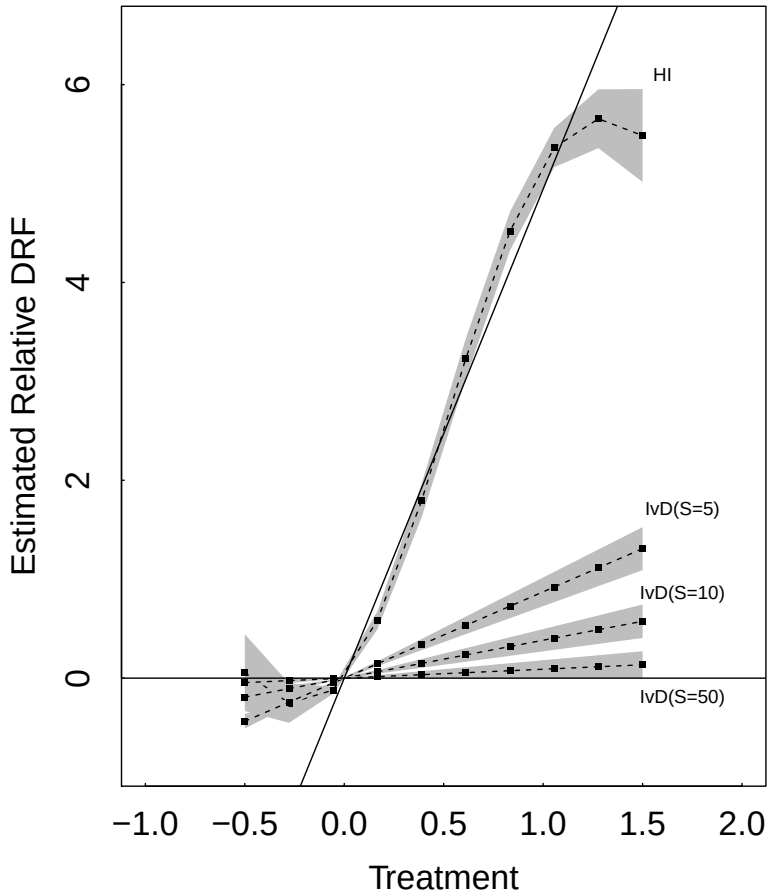
Fits

- Treatment model is correctly specified.
- IvD: Linear regression ($Y \sim T$) within S subclasses.
- Do not adjust for $\hat{\theta}$ in within subclass models.
  Doing so would dramatically improve preformence.

Imperial College London

Simulation 1 Results





Expect any method to work well in this simple setting

Imperial College London

Simulation 2

Frequency Evaluation:

- $X \overset{\text{indep.}}{\sim} \mathcal{N}(0,1)$, $T|X \overset{\text{indep.}}{\sim} \mathcal{N}(X + X^2, 1)$, and $n = 2000$.
- Generative model for $Y(t)$:

  Linear DRF: $Y(t) \mid t, X \overset{\text{indep.}}{\sim} \mathcal{N}(X + t, 9)$

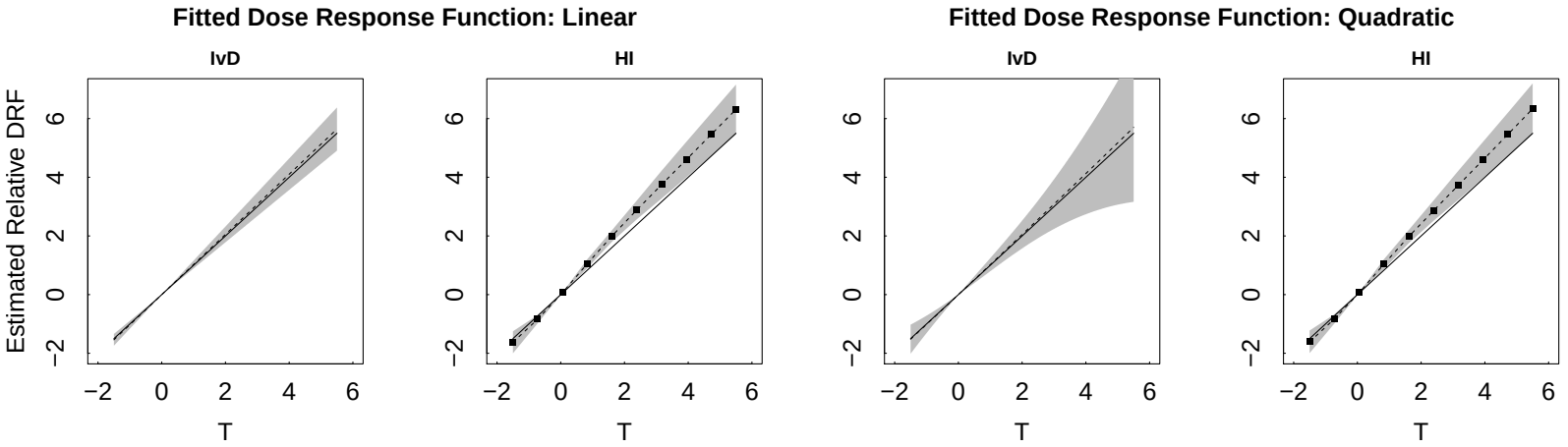
  Quadratic DRF: $Y(t) \mid t, X \overset{\text{indep.}}{\sim} \mathcal{N}((X + t)^2, 9)$
- Correctly specified treatment models used for all fits.
- Linear and quadratic (in T) response models used with both methods.
- Used 10 subclasses with IvD.

Entire procedure was replicated for 1000 simulated data sets.

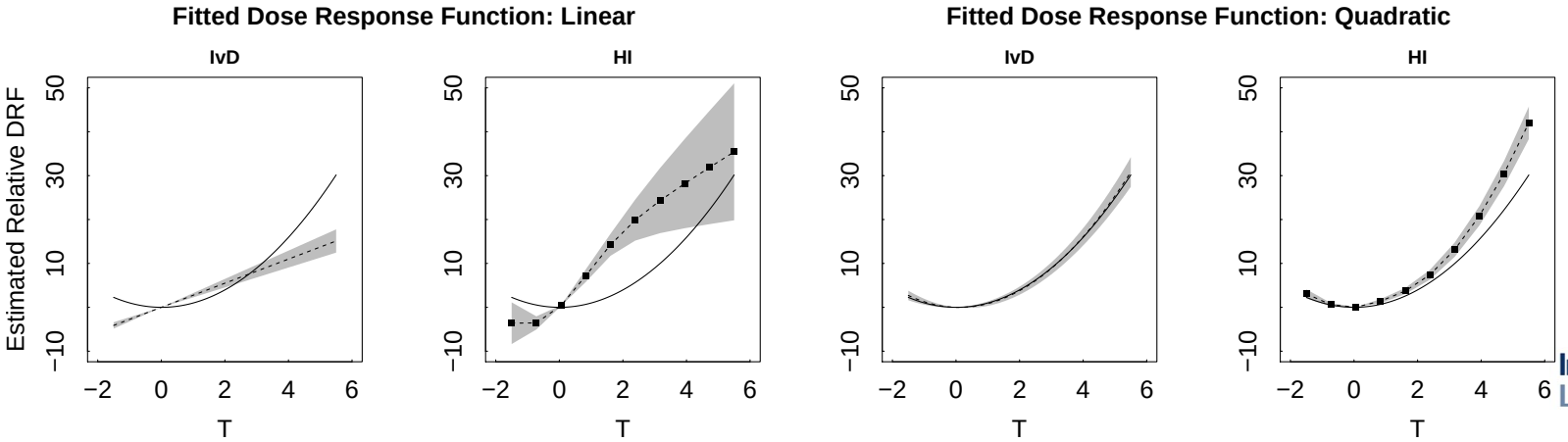
Imperial College London

Simulation 2 Results

Generative Dose Response Function: Linear



Generative Dose Response Function: Quadratic



Imperial College London

Outline

Imperial College London

Covariance Adjustment for a Categorical Treatment

Estimating Dose Response using Covariance Adjustment:

- Regress Y on $\hat{R} = \hat{r}(T = t, X)$ separately for units in each treatment group:

  Group t: $Y \sim \alpha_t + \beta_t \hat{R}$ $(\star\star)$

- Average fitted potential outcomes of all units

$$\widehat{\mathrm{E}(Y(t))} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\alpha}_t + \hat{\beta}_t \hat{r}(t, X_i) \right]$$

- Pairwise differences are exactly Rosenbaum and Rubin's estimate of the average treatment effect.
- Hirano and Imbens replace $(\star\star)$ with quadratic regression.

$$Y(t) \sim \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \alpha_3 \cdot \hat{R} + \alpha_4 \cdot \hat{R}^2 + \alpha_5 \cdot T \cdot \hat{R}.$$

Imperial College London

Covariance Adjustment for a Continuous Treatment

- Hirano and Imbens' quadratic regression
  - ⋆ extends easily to continuous treatments,
  - ⋆ but is less robust than treatment-level specific regressions.

- **Solution:** Discretize continuous treatment variable.
  (Zhang, van Dyk, & Imai, 2013)
  - ⋆ Use quantiles of the treatment for discretization.
  - ⋆ Extreme categories tend to have wider range of T and thus we expect more bias.
  - ⋆ Within subclass t, we fit e.g., $Y \sim \alpha_t + \beta_t \hat{R}$
  - ⋆ Estimate Dose Response: $\widehat{\mathrm{E}(Y(t))} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\alpha}_t + \hat{\beta}_t\,\hat{r}(t, X_i)\right]$

- **Note:** In contrast to standard methods, we subclassify on the *treatment* rather than on the propensity score.

Imperial College London

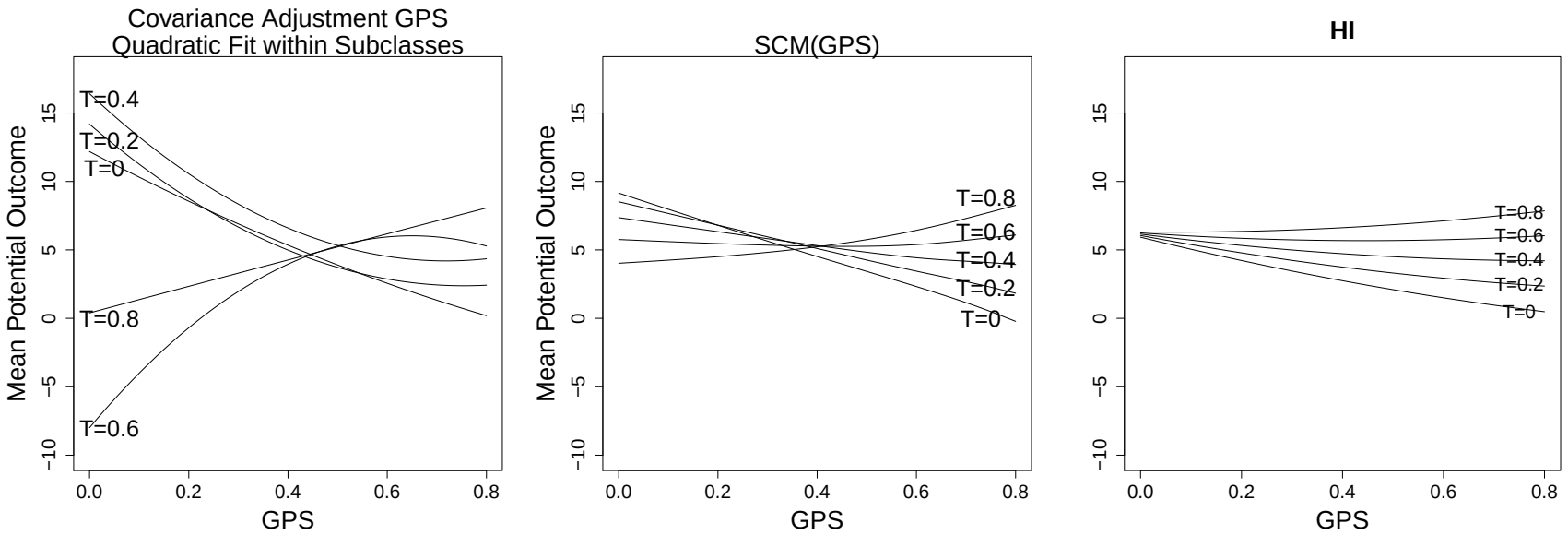
Using a SCM with the GPS

- Fitting $Y \sim \alpha_t + \beta_t \hat{R}$ within each of several subclasses can be viewed as a flexible regression of Y on R and T.
- Flores et al. (2012) proposed another solution:
  - ⋆ Non-parametrically model $Y \sim f(R, T)$
  - ⋆ Estimate Dose Response: $\widehat{\mathrm{E}(Y(t))} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}\{r(t, X_i), t\}\right]$
- Flores et al. use a non-parametric kernel estimator.
- We use a SCM to facilitate comparisons.
- The three GPS-based methods vary only in the choice of response model (quadratic regression, subclassification, or SCM)

Imperial College London

Importance of the Choice of Response Model

Fitted $E(Y(t)|T = t, R)$ for dataset in Simulation 1



Covariance Adjustment GPS is much more flexible than Quadratic Regression

Imperial College London

Using the Propensity Function to Estimate the DRF

Extend Imai and van Dyk (2003) to estimate Dose Response

- Begin by writing the Dose Response Function as

$$E[Y(t)] = \int E[Y(t) \mid \theta]\, p(\theta) d\theta = \int E[Y(T) \mid \theta, T = t]\, p(\theta) d\theta$$

- Use Smooth Coefficient Model for rightmost integrand

$$E[Y(T) \mid \theta, T = t] \;=\; f(\theta, T),$$

- Average over units to obtain fitted Dose Response

$$\hat{E}[Y(t)] \;=\; \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\hat{\theta}_i, t),$$

Imperial College London

The Problem with Extrapolation

Estimated Dose Response: $\hat{E}[Y(t)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\theta}_i, t).$

- Evaluating at t_0 involves evaluating $\hat{f}(\hat{\theta}_i, t_0)$ at every observed value of $\hat{\theta}_i$.
- Invariably, the range of $\hat{\theta}$ among units with T near t_0 is smaller than the total range of $\hat{\theta}$, at least for some t_0.
- *This estimate involves some degree of extrapolation!!*
- We can diagnose using a scatterplot of T vs $\hat{\theta}$.

Imperial College London

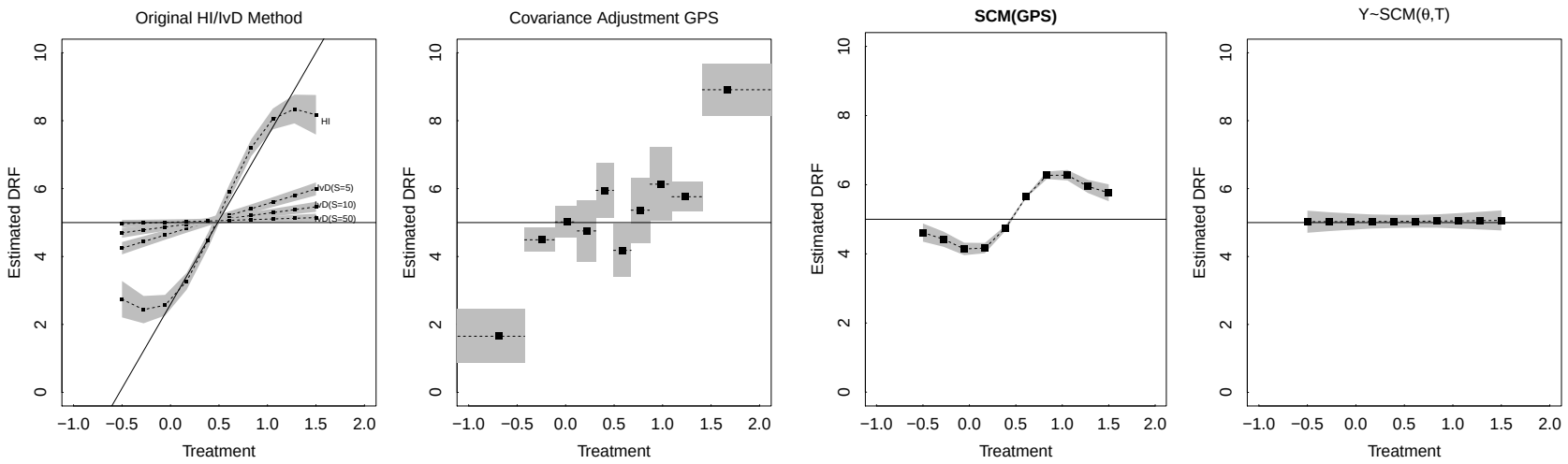
The Problem with Extrapolation

Similarly with the GPS: $\hat{E}[Y(t)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\{r(t, X_i), t\}$.

- This problem is more acute: The range of R among units with T near t_0 may not even overlap with range of $r(t_0, X)$.
- This is true for all three response models.
- *The estimate of the DRF at t_0 may depend entirely on extrapolation!!*
- We can diagnose by comparing scatterplots of T vs R and T vs $r(T, X)$

Imperial College London

Simulation 1: Revisit

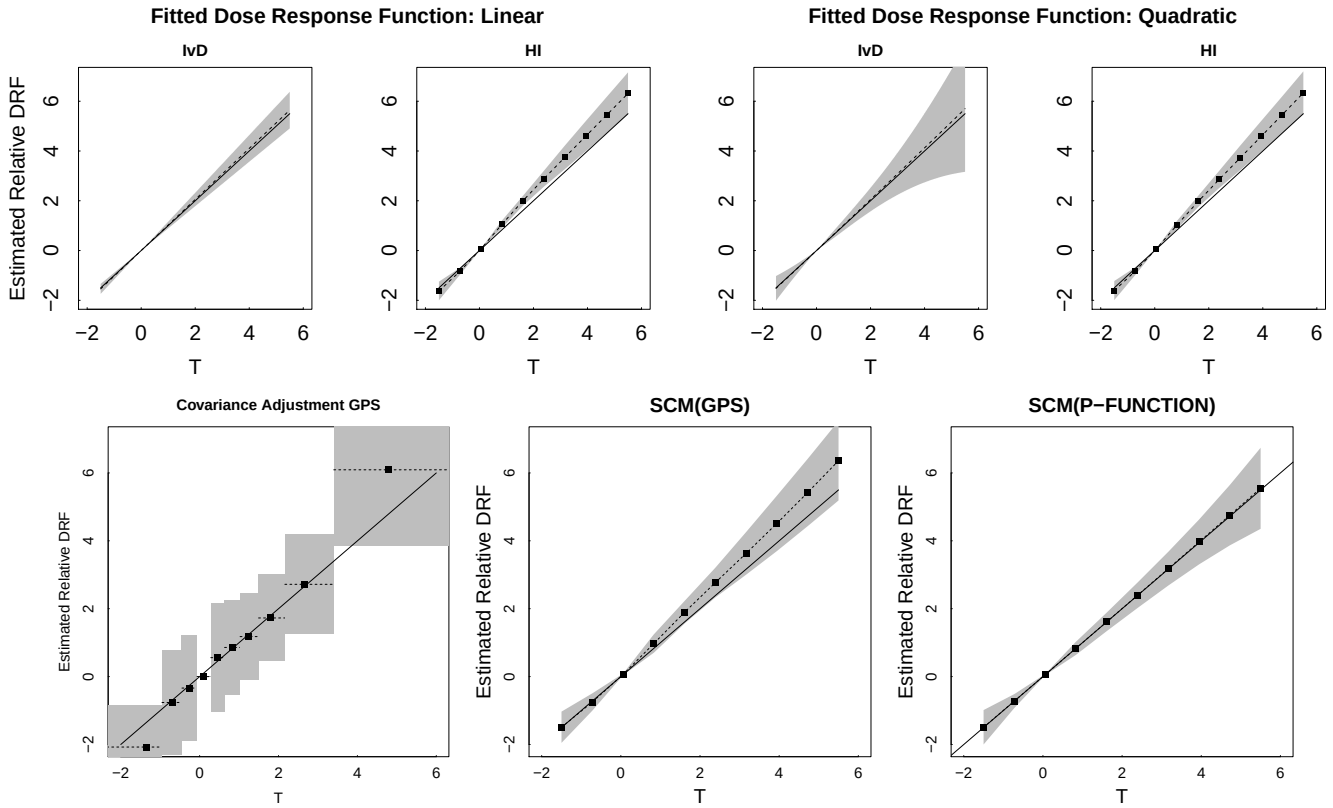


- SCM(GPS) exhibits an odd cyclic pattern
- SCM(P-Function) results in a much smoother fit.
- As expected, covariance adjustment GPS is biased in extreme subclasses.

Imperial College London

Simulation 2: Revisited

Linear Generative DRF

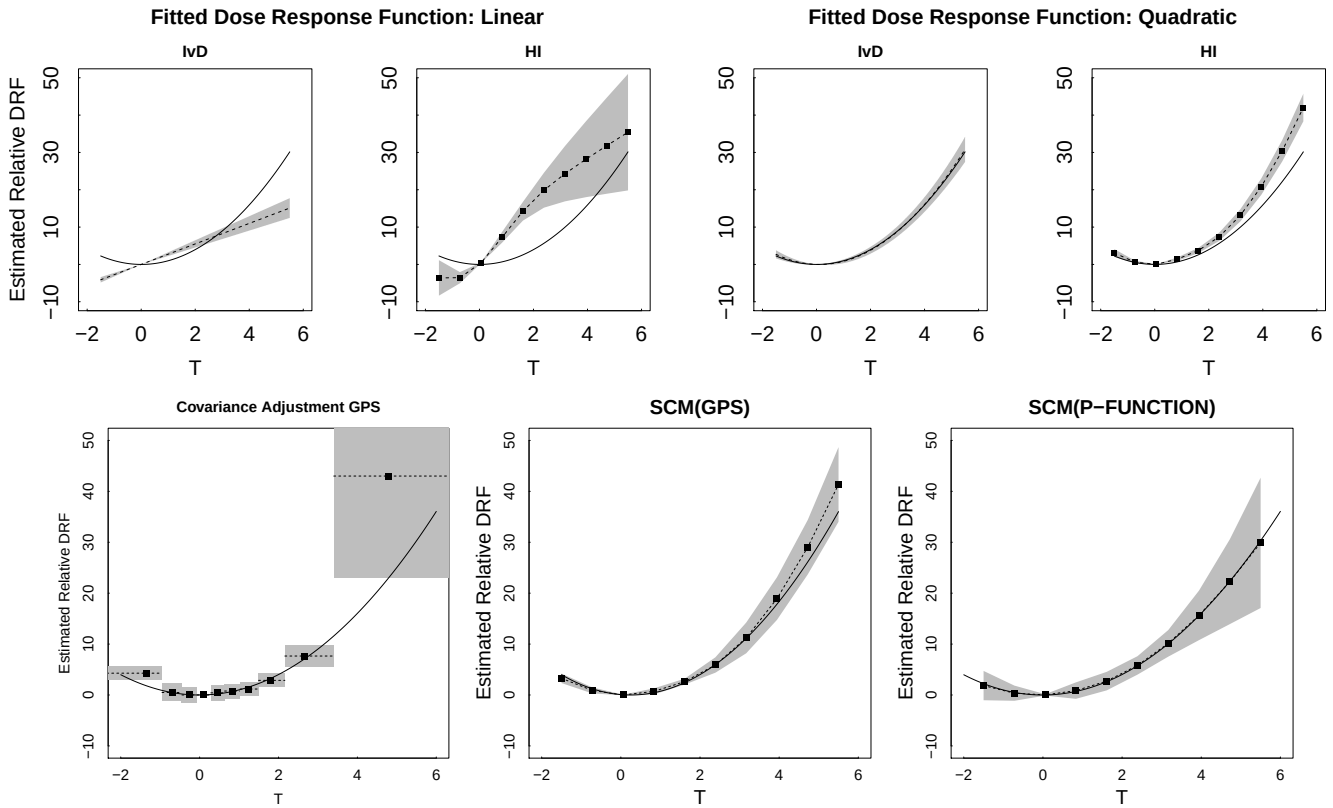


Reduced bias without parametric specification

Imperial College London

Simulation 2: Revisited

Quadratic Generative DRF



Reduced bias without parametric specification

Imperial College London

Example: Effects of Smoking on Medical Expenditure

- Data: 9,708 smokers from 1987 National Medical Expenditure Survey (Johnson *et al.*, 2002).
- Treatment variable, $T^A = \log(\text{packyear})$: continuous measure of cumulative exposure to smoking:

$$\textit{packyear} = \frac{\#\text{ of cigarettes per day}}{20} \times \#\text{ of years smoked.}$$

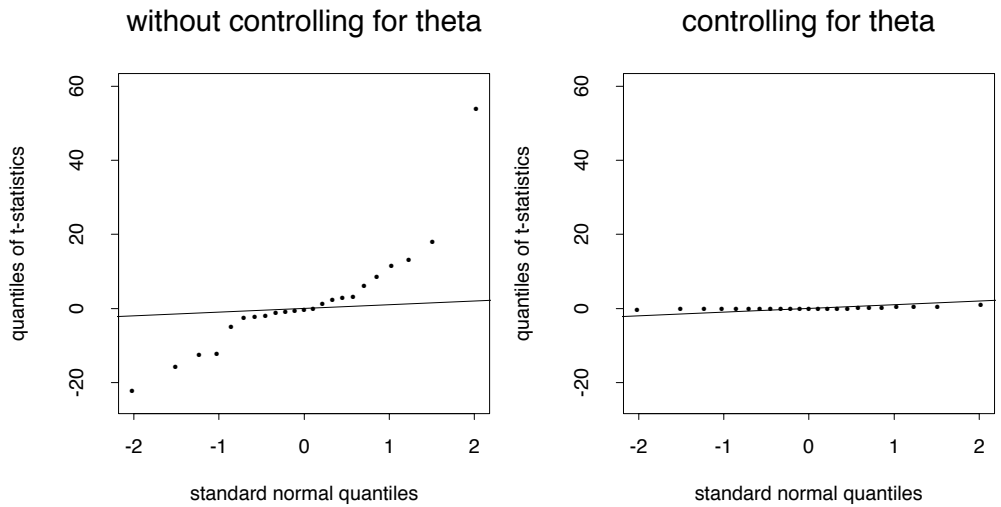
  - ⋆ alternative strategy: frequency and duration of smoking as bivariate treatment variable.

- Outcome: self-reported annual medical expenditure.
- Covariates: age at time of survey, age when the individual started smoking, gender, race, marriage status, education level, census region, poverty status, and seat belt usage.

Imperial College London

Model Specification of the Propensity Function

- Model: $T^A \mid X \overset{\text{indep.}}{\sim} N(X^\top \beta, \sigma^2)$ and $\theta = X^\top \beta$ where X includes some square terms in addition to linear terms.



- First panel: T-stats for predicting covariates from log(packyear).
- Second panel: Same, but controlling for propensity funt'n.

The Balance is Improved

Balance serves as a model diagnostic for propensity function.

Imperial College London

Using the Method of Imai and van Dyk (2004)

Within each block, use a Two-Part model for the semi-continuous response:

1. Logistic regression for $\Pr(Y > 0 \mid T^A, \hat{\theta}_j)$.
2. Gaussian linear regression for $p\{\log(Y) \mid Y > 0, T^A, \hat{\theta}_j\}$.

		Propensity Function	
	Direct Models	3 blocks	10 blocks
Logistic Linear Regression Model			
coefficient for T^A	-0.097	-0.060	-0.065
standard error	3.074	3.031	3.074
Gaussian Linear Regression Model			
average causal effect	0.026	0.051	0.053
standard error	0.016	0.017	0.018

Imperial College London

A Smooth Coefficient Model

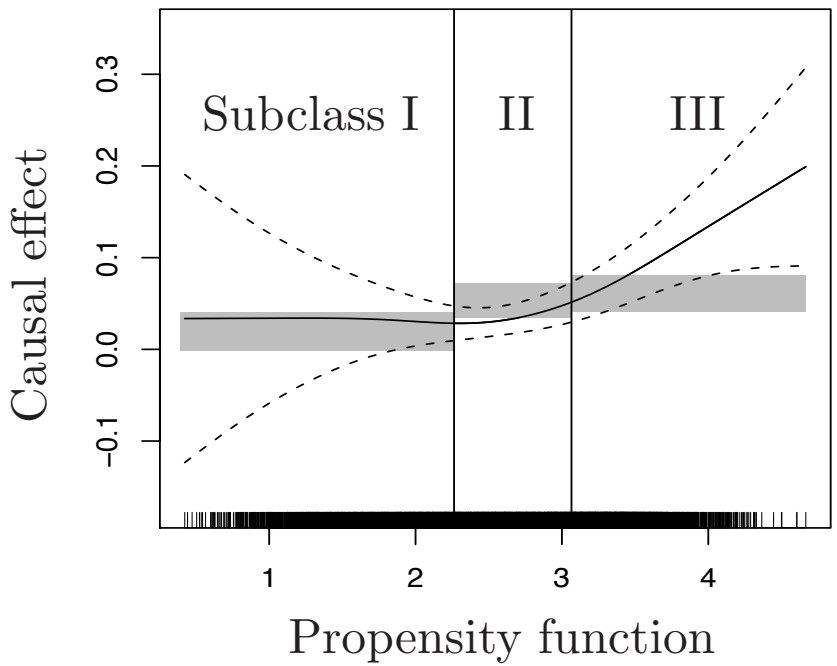
- In the subclassification-based analysis, we fit a separate regression in each subclass, allowing for different treatment effects in each subclass.
- An alternate strategy allows the treatment effect to vary smoothly with θ.
- For example, in stage two

$$log(Y) \sim \text{Normal}(\alpha(\theta) + \beta(\theta) T^A + \gamma X, \sigma^2),$$

where α and β are smooth functions of θ.

Imperial College London

The Smooth Coefficient Model Fit



- Propensity funct'n is linear predictor for log(packyears).
- Propensity function is hard to interpret.
- *Dose response function* would be much more useful.

Imperial College London

Estimating the Dose Response Function

We begin with a simulation study:

- Use observed X and T and simulate Y using each of
  1. Quadratic DRF:

$$E[\log(Y_i(t))|X] = \frac{4}{25} \cdot t^2 + [\log(\text{age}_i)]^2$$

  2. Piecewise Linear DRF:

$$E[\log(Y_i(t))|X] = \begin{cases} -4 - 0.5 \cdot t + [\log(\text{age}_i)]^2, & t \leq 2 \\ -5 + 2.3 \cdot (t-2) + [\log(\text{age}_i)]^2, & t > 2 \end{cases}$$

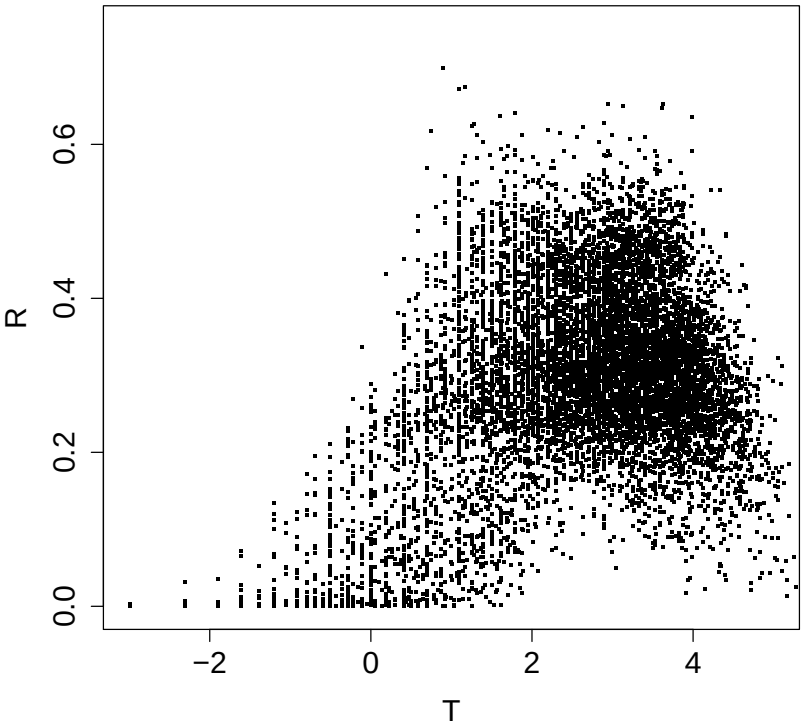
  3. Hockey-Stick DRF:

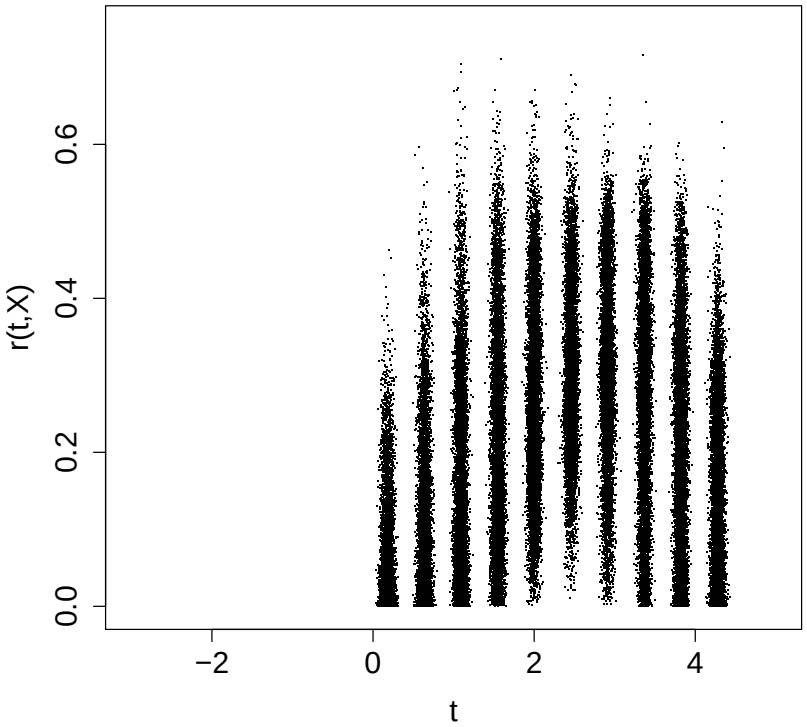
$$E[\log(Y_i(t))|X] = \begin{cases} -8.1 + [\log(\text{age}_i)]^2, & t \leq 3 \\ -8.1 + 1.5 \cdot (t-3)^2 + [\log(\text{age}_i)]^2, & t > 3 \end{cases}$$

Imperial College London

Extrapolation in Fitting the GPS-Based DRF

Recall: $Y \sim f(R, T)$ and $\widehat{\mathrm{E}(Y(t))} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}\{r(t, X_i), t\}\right]$





Extrapoloation for:

Small t ($r > 0.3$), Midrange t ($r < 0.2$) and Large t ($r < 0.1$)

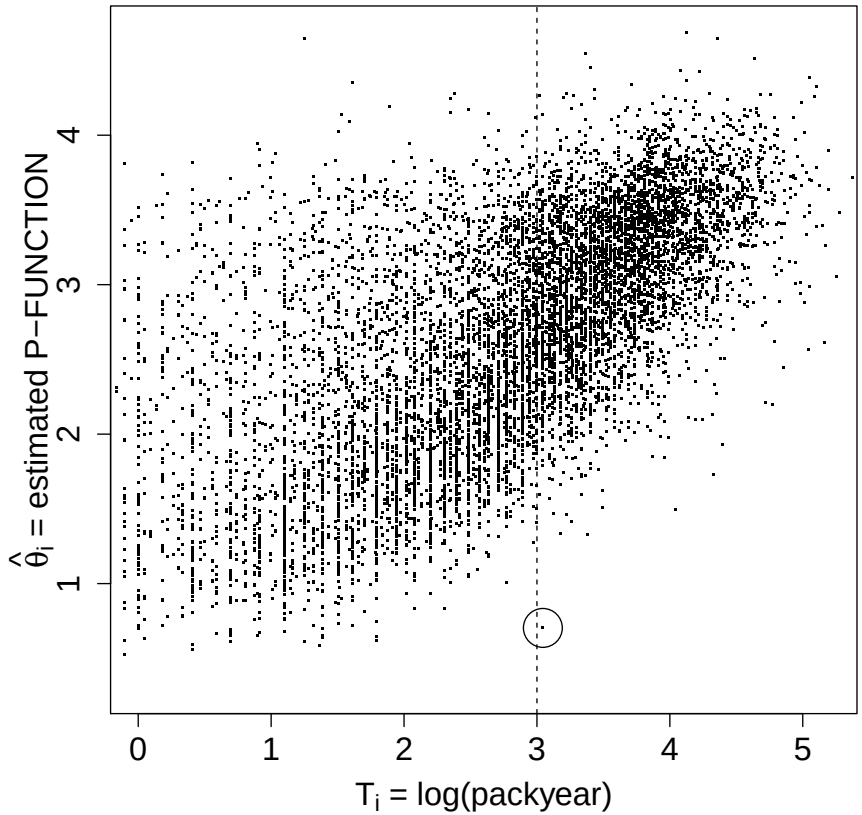
Imperial College London

Extrapolation in Fitting the P-Function based DRF

Recall:

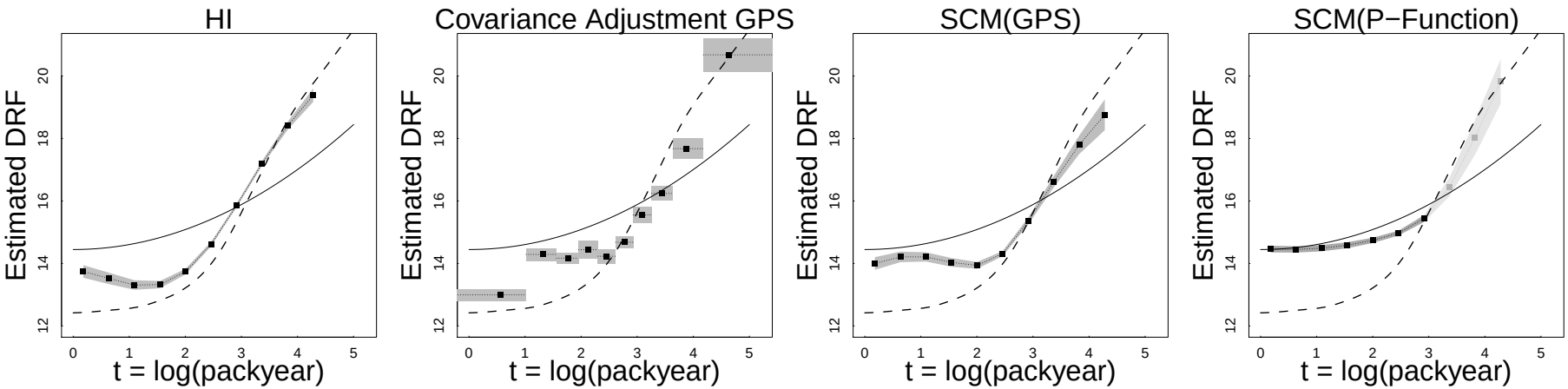
- $Y \sim f(\hat{\theta}, T)$
- $\hat{E}[Y(t)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\theta}_i, t).$

We expect bias in $\hat{E}[Y(t)]$ *for* $t > 3$.



Imperial College London

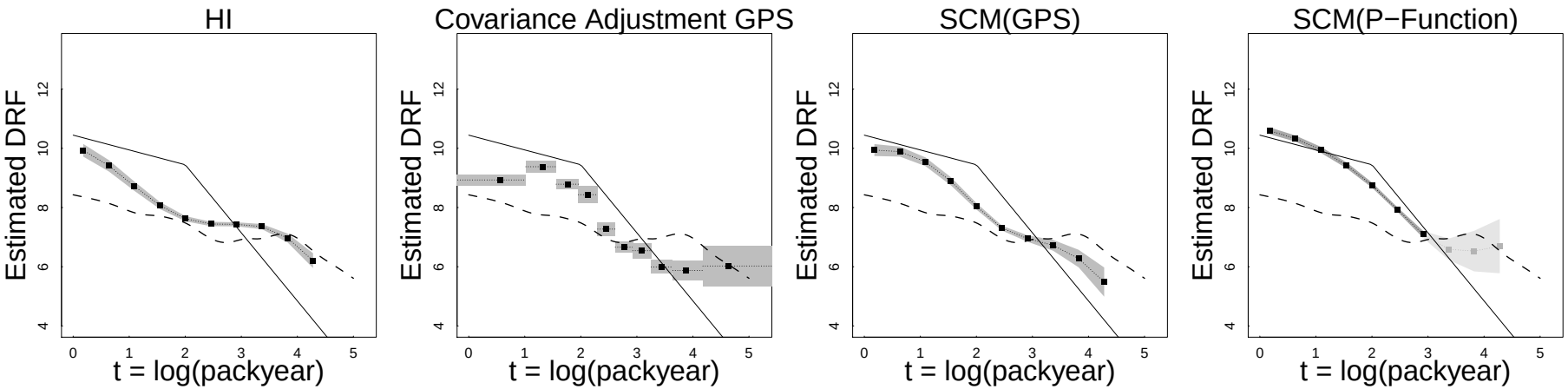
Results: Quadratic DRF



1. Hirano and Imben's estimate tends to follow unadjusted fit.
2. The two other GPS-based methods do better.
3. SCM(GPS) again exhibits a small cyclic pattern
4. SCM(P-Functions) does the best, at least for $t < 3$.

Imperial College London

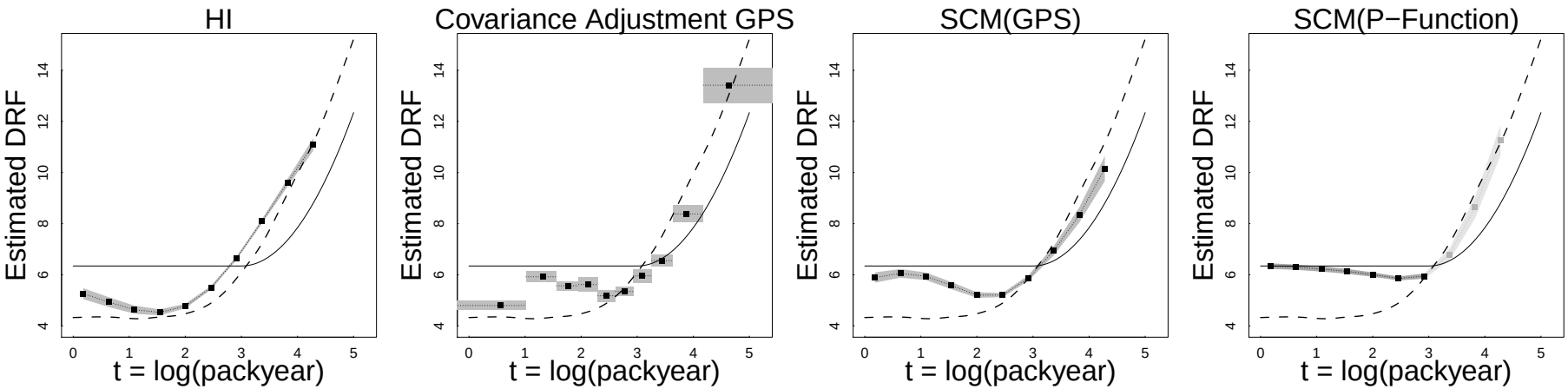
Results: Piecewise Linear DRF



1. Hirano and Imben's estimate tends to follow unadjusted fit.
2. The two other GPS-based methods do better.
3. SCM(GPS) again exhibits a small cyclic pattern
4. SCM(P-Functions) does the best, at least for $t < 3$.

Imperial College London

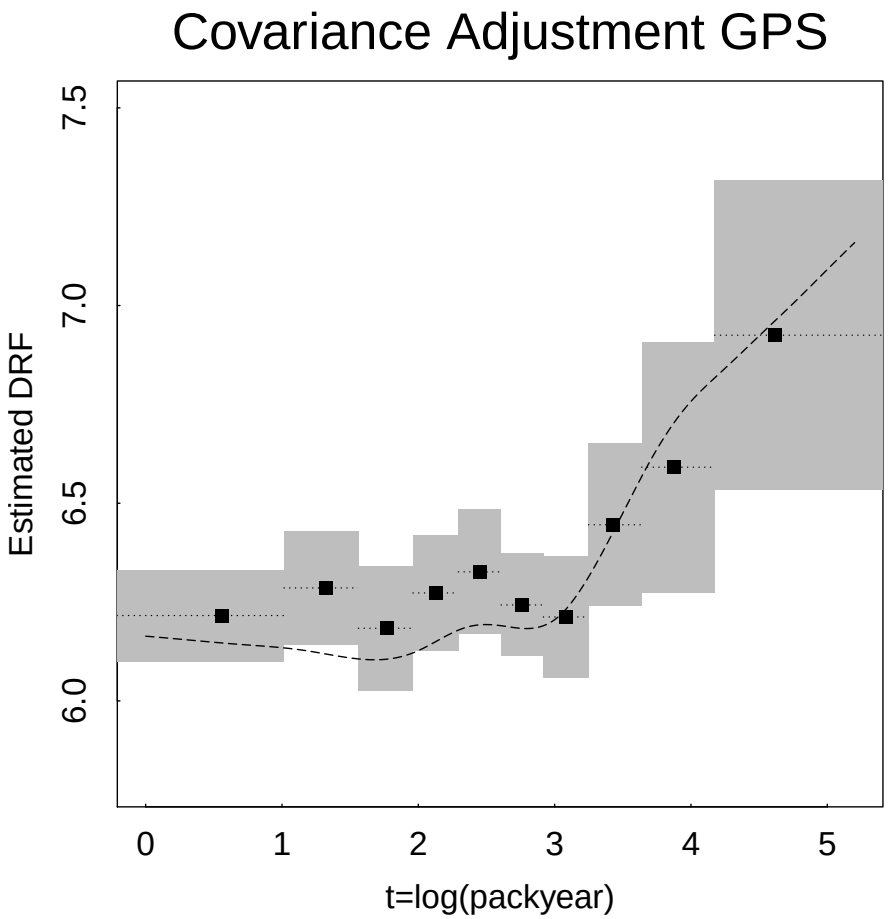
Resuls: Hockey-Stick DRF

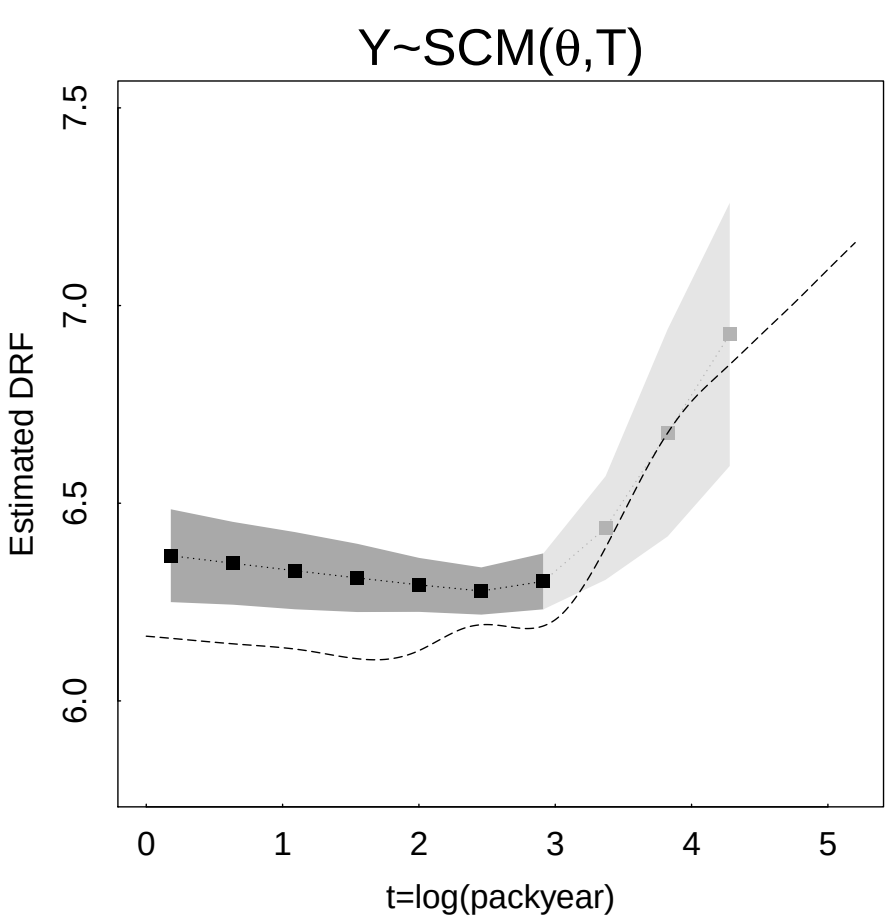


1. Hirano and Imben's estimate tends to follow unadjusted fit.
2. The two other GPS-based methods do better.
3. SCM(GPS) again exhibits a small cyclic pattern
4. SCM(P-Functions) does the best, at least for $t < 3$.

Imperial College London

The Estimated DRF: Using the Data





Imperial College London

Summary and Conclusions

- Fitted Dose Response Function of Hirano and Imbens seems unreliable.
- Imai & van Dyk is more limited in scope but more reliable.
- We aim to derive reliable estimates of dose response in observational studies.
- P-function based estimate appear to offer significant improvement.
- Nonetheless we advise caution.
  - ⋆ Large samples are required.
  - ⋆ In smaller studies, *average treatment effect* is preferable.

Imperial College London