Causal Inference with General Treatment Regimes:

Generalizing the Propensity Score

David van Dyk
Department of Statistics, University of California, Irvine
vandyk@stat.harvard.edu

Joint work with
Kosuke Imai
Department of Politics, Princeton University

Outline

1. Review the propensity score of Rosenbaum and Rubin (1983) for causal inferences in observational studies.
2. Introduce the **Propensity Function**, a generalization of the propensity score: applicable beyond binary treatment regimes.
3. Establish the key theorem: Strong Ignorability of Treatment Assignment Given Propensity Function.
4. Monte Carlo experiments: comparison with standard regression approaches.
5. Examples:
   - ⋆ Effect of smoking on medical expenditure (bivariate continuous treatment).
   - ⋆ Effect of education on wages (ordinal instrumental variable).

Salk's Polio Vaccine Trials

- Background
  - ⋆ What is polio?
  - ⋆ When was the epidemic?
  - ⋆ How did it end?
  - ⋆ How would people respond to the vaccine before it was proven effective?
- Goal: *Compare* polio rates between
  1. Children given the vaccine (*treatment group*)
  2. Children not given the vaccine (*control group*)
- National Foundation of Infantile Paralysis 1954 Polio Vaccine Trials

  Treatment Group: Children whose parents give consent

  Control Group: Children whose parents **do not** give consent

What problems are there with this design?

The Results

The randomized controlled double-blinded experiment			*The NFIP study*		
	Size	Rate[a]		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

- The Treatment/consent and No-Treatment/No-Consent groups are comparable between the studies.
- Using the No-Consent group as a Control *biases* the causal effect.

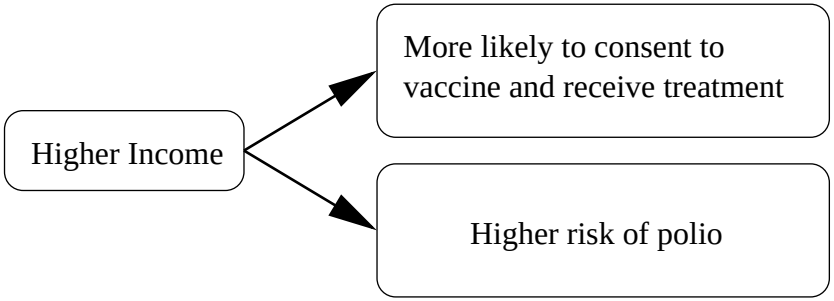
Consent (treatment indicator) is correlated with polio rates in the absence of treatment (potential outcome).

Causal Inference in Observational Studies

1. The general setup with a binary treatment:

		Treatment	Potential Outcomes	
	Covariates	Group	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	$\mathbf{Y_3(0)}$	$Y_3(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

2. E.g.: National Foundation of Infantile Paralysis 1954 Polio Vaccine Trials



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

Formalizing Causal Inference

- Causal effect: e.g., $Y(t_1^P) - Y(t_2^P)$.
  - $\star$ $Y(t^P)$ is potential outcome with potential treatment, $t^P \in \mathcal{T}$.
- Problem: Treatment assignment, T^A, is not random and is correlated with potential outcomes.
  - $\star$ $Y_i(T_i^A = t_1^P) - Y_j(T_j^A = t_2^P)$ does not correspond to causal effect.
- Key assumption (Strong Ignorability of Treatment Assign.; Rubin, 1978):

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

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

Adjusting for Covariates in Causal Inference

- Standard regression approach: $Y(t^P) \sim N(\alpha + X\beta + t^P\gamma, \sigma^2)$.
  - ⋆ common assumptions, e.g. linearity, are often violated.
  - ⋆ leads to biased causal inferences.
- Matching and Subclassification reduce bias more effectively:
  - ⋆ non/semi-parametric methods.
  - ⋆ diagnostics directly related to causal inferences.
  - ⋆ but, difficult when the dimensionality of X is large.

Propensity Score of Rosenbaum and Rubin (1983)

- If the treatment is binary, the propensity score,

$$e(X) = \Pr(T^A = 1 \,|\, X),$$

  fully characterizes $p(T^A \,|\, X)$.

- Key Theorems:

  1. The propensity score is a balancing score:

$$\Pr\{T^A = 1 \,|\, X, e(X)\} = \Pr\{T^A = 1 \,|\, e(X)\}.$$

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

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

- Unbiased estimate of causal effect is possible conditional on $e(X)$.

- Match or subclassify on $e(X)$, a scalar variable.

The Power of Conditioning on Propensity Scores

The problem with (non-randomized) observational studies:

Treatment assignment is 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 assignment is UNcorrelated with potential outcomes.

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

Use of the Propensity Score in Observational Studies

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

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

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

General Treatment Regimes

- 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

Goal: Generalization of propensity score to non-binary treatment.

Existing literature:

1. ordinal treatment (Joffe & Rosenbaum, 1999),
2. categorical treatment (Imbens, 2000).

Our method encompasses both and other types of treatments.

Framework of Causal Inference via Potential Outcomes

Set of potential outcomes: $\mathcal{Y} = \{Y(t^P), t^P \in \mathcal{T}\}$.

Assumptions:

1. Stable Unit Treatment Value (Rubin, 1990):

$$p\{Y_i(t_i^P) \,|\, t_j^P, T_j^A, X_i\} \;=\; p\{Y_i(t_i^P) \,|\, X_i\} \quad \forall i \neq j \text{ and } t_i^P, t_j^P \in \mathcal{T}.$$

2. Strongly Ignorable Treatment Assignment (Rubin 1978):

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

Quantity of primary interest:

$$p\{Y(t^P)\} \;=\; \int p\{Y(t^P) \,|\, X\}\, p(X)\, dX.$$

Propensity Function: Generalization of Propensity Score

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

Uniquely Parameterized Propensity Function Assumption:

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

θ fully characterizes $p(T^A \mid X)$ and is typically of low-dimension.

Examples:

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

Theoretical Properties of Propensity Function

Theorem 1: Propensity Function is a Balancing Score.

$$p(T^A \,|\, X) \;=\; p\{T^A \,|\, e(\cdot \,|\, X), X\} \;=\; p\{T^A \,|\, e(\cdot \,|\, X)\}.$$

Theorem 2: Strongly Ignorable Treatment Assignment Given Propensity Function.

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

Estimation of causal effects via subclassification:

$$\begin{aligned} p\{Y(t^P)\} &= \int p\{Y(t^P) \,|\, T^A = t^P, \theta\}\, p(\theta)\, d\,\theta \\ &\approx \sum_{j=1}^{J} p\{Y(t^P) \,|\, T^A = t^P, \theta_j\}\, W_j. \end{aligned}$$

Monte Carlo Experiment 1: Continuous Treatment

- Number of simulations: 5,000.
- Sample size per simulation: n =1,000.
- Observed covariates:

$$X_1 \overset{\text{indep.}}{\sim} N(1,1), \quad X_2 \overset{\text{indep.}}{\sim} N(2,1).$$

- Treatment variable:

$$T^A \,|\, X_1, X_2 \overset{\text{indep.}}{\sim} N\left(1 + X_1 X_2 + X_1^2 + X_2^2,\ 1\right).$$

- Outcome variable:

$$Y \,|\, T^A, X_1, X_2 \overset{\text{indep.}}{\sim} N\left(1 + T^A + X_1 X_2 + X_1^2 + X_2^2,\ 1\right).$$

Robustness to Misspecification

Two (misspecified) models for estimating the causal effects:

1. Direct regression: $Y \mid T^A, X \overset{\text{indep.}}{\sim} N(\alpha T^A + X^\top \beta, \sigma^2)$ where $X = (1, X_1, X_2)$.
2. Propensity Function:
   - Model: $T^A \mid X \overset{\text{indep.}}{\sim} N(X^\top \beta, \sigma^2)$.
   - Subclassify data into 10 blocks based on $\hat{\theta} = X^\top \hat{\beta}$.
   - Within each block, regress Y on T^A and $\hat{\theta}$.
   - Calculate weighted average of 10 within-block estimates.

Results:

	Average Causal Effect of T^A	
	Bias	MSE
Direct Regression	0.832	0.692
Propensity Function	0.390	0.153

Reduce MSE by 75%.

Monte Carlo Experiment 2: Interaction of Two Treatments

- Number of simulations: 5,000.
- Sample size per simulation: n =2,000.
- Observed covariates: $X_1 \overset{\text{indep.}}{\sim} N(1,1), \quad X_2 \overset{\text{indep.}}{\sim} N(2,1)$.
- Treatment variables:

$$\begin{pmatrix} T^{A_1} \\ T^{A_2} \end{pmatrix} \Big| X_1, X_2 \overset{\text{indep.}}{\sim} N_2 \left\{ \begin{pmatrix} 1 + X_1^2 + X_2^2 \\ 1 + X_1 X_2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\}.$$

- Outcome variable:

$$Y \overset{\text{indep.}}{\sim} N\left(1 + T^{A_1} + T^{A_2} + T^{A_1} T^{A_2} + X_1 X_2 + X_1^2 + X_2^2,\ 1\right).$$

Robustness to Misspecification

Two (misspecified) models for estimating the causal effects:

1. Direct Linear Regression: Regress Y on X_1, X_2, T^{A_1}, T^{A_2}, and, $T^{A_1}T^{A_2}$.
2. Propensity Functions:
   - Model Specification of Propensity Function: Independently regress both treatment variables on X_1 and X_2
   - Stratify on $\hat{\theta} = (X^\top \hat{\beta}_1, X^\top \hat{\beta}_2)$.
   - Within Strata Model: Regress Y on T^{A_1}, T^{A_2}, $T^{A_1}T^{A_2}$, and $\hat{\theta}$.

	Direct Regression		Propensity Function	
	Bias	MSE	Bias	MSE
Effect of T^{A_1}	0.517	0.268	0.159	0.026
Effect of T^{A_2}	-0.340	0.118	-0.102	0.014
Effect of $T^{A_1}T^{A_2}$	0.0452	0.0021	-0.0074	0.0001

Reduce MSE by 90%.

Subclassification with Two Propensity Functions

Propensity Function For T_1^A	**Propensity Function For T_1^A** lower 1/3	middle 1/3	upper 1/3
lower 1/3	Block I	Block II	Block III
middle 1/3	Block IV	Block V	Block VI
upper 1/3	Block VII	Block VIII	Block IX

Example 1: 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:

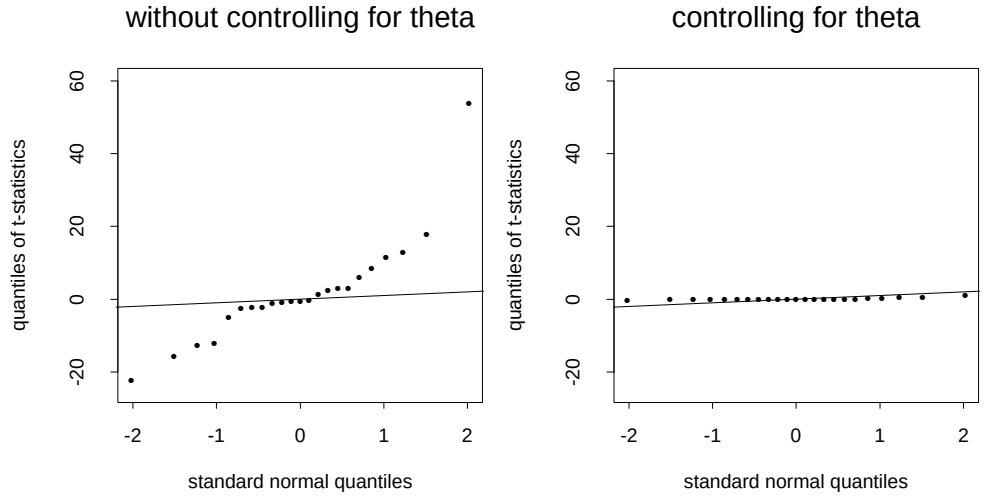
$$\textit{packyear} \;=\; \frac{\text{number of cigarettes per day}}{20} \;\times\; \text{number of years smoked.}$$

  - ⋆ previous studies apply propensity score to compare smokers with non-smokers (Larsen, 1999; Rubin, 2001).
  - ⋆ alternative strategy: frequency and duration of smoking as bivariate treatment variable.
- Outcome variable: self-reported annual medical expenditure.
- Covariates: age at the times of the survey, age when the individual started smoking, gender, race, marriage status, education level, census region, poverty status, and seat belt usage.

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.

Balance



- First panel: T-statistics for predicting the covariates from the log(packyears).
- Second panel: Same, but controlling for the linear predictor.

The Balance is Improved

The balance serves as a model diagnostic for the propensity function.

Estimated Effects of Smoking

- Within each block, use a Two-Part model for the semi-continuous response:
  1. Logistic regression for $\Pr(Y > 0 \,|\, T^A, \hat{\theta}_j)$.
  2. Gaussian linear regression for $p\{\log(Y) \,|\, Y > 0, T^A, \hat{\theta}_j)$.

		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

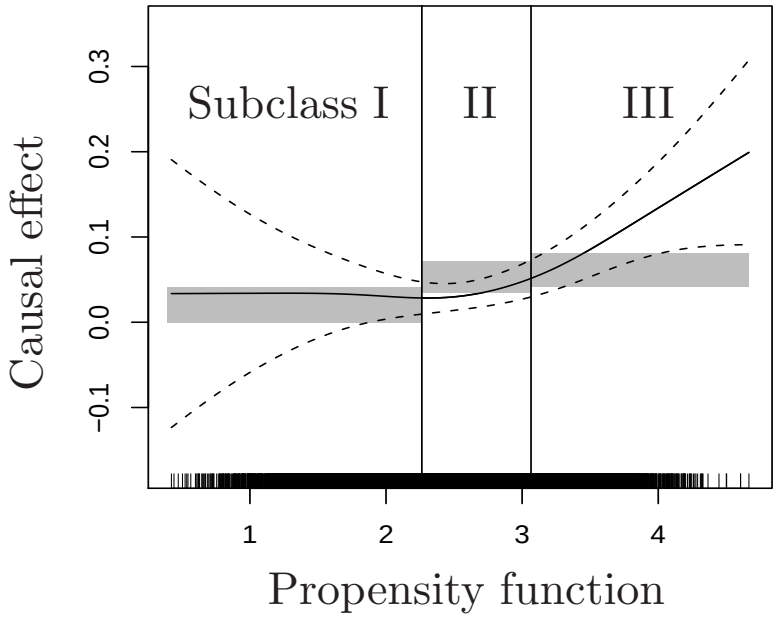
A Smooth Coefficient Model

- In our analysis we fit a separate regression in each subclass, allowing for different treatment effects in each subclass.
- An alternate strategy allows the treatment effect (and perhaps other regression coefficients 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.

The Smooth Coefficient Model Fit



- The propensity score is the linear predictor for log(packyears).
- Important covariates include age and age when began smoking.
- For older people the effect of smoking on medical expenses is greater.

Analysis with Bivariate Treatment

Now we consider the bivariate treatment

$$\begin{aligned} T_1^A &= \log(\text{average number of cigarettes per day}) \\ T_2^A &= \log(\text{number of years smoked}) \end{aligned}$$

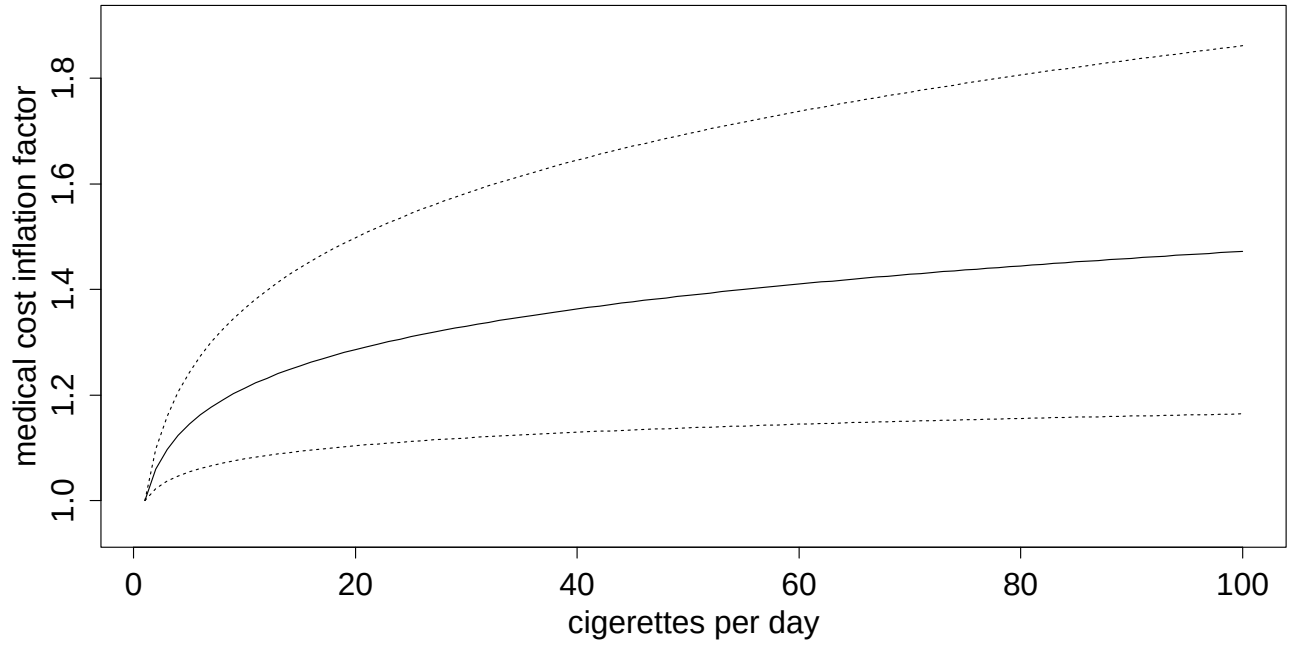
Analysis

- Model Specification of Propensity Function: two independent regressions with same covariates as before (including squared terms).
- Stratify on $\hat{\theta} = (X^\top \hat{\beta}_1, X^\top \hat{\beta}_2)$.
- Within Strata Model: two-part model.

Results:

	T_1^A	T_2^A
Logistic Linear Regression Model	-0.358 (se=7.110)	0.075 (se=4.527)
Gaussian Linear Regression Model	0.084 (se=0.026)	0.011 (se=0.036)

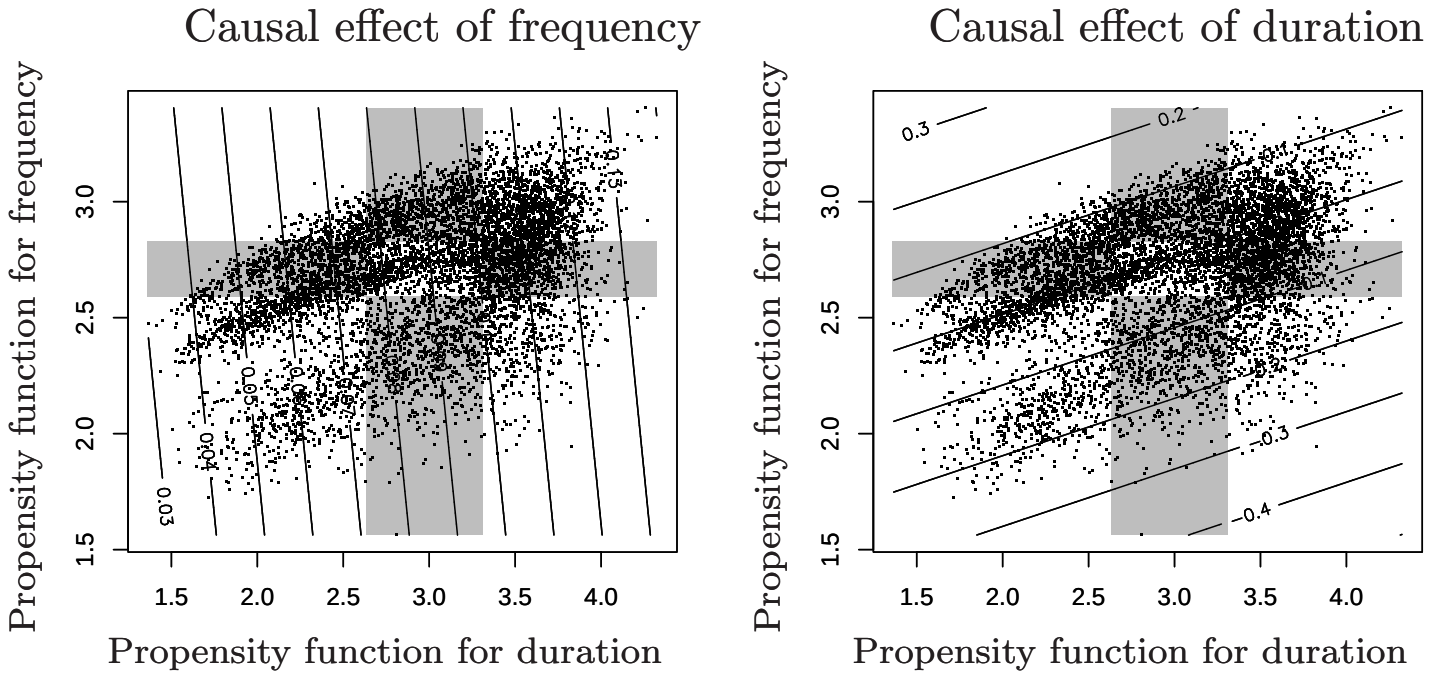
Fitted effect of Smoking Frequency with 95% error bars



Fitted effect by subclass

Propensity Function for Frequency	Propensity Function for Duration: lower third	Propensity Function for Duration: middle third	Propensity Function for Duration: upper third
upper third	**Subclass I** dur: 0.317 (0.221) freq: -0.223 (0.143) $n = 324$	**Subclass II** dur: 0.075 (0.092) freq: 0.125 (0.075) $n = 1160$	**Subclass III** dur: 0.016 (0.078) freq: 0.093 (0.067) $n = 1542$
middle third	**Subclass IV** dur: 0.020 (0.105) freq: 0.009 (0.075) $n = 1162$	**Subclass V** dur: -0.011 (0.092) freq: 0.123 (0.076) $n = 910$	**Subclass VI** dur: -0.182 (0.100) freq: 0.208 (0.080) $n = 952$
lower third	**Subclass VII** dur: -0.079 (0.099) freq: 0.105 (0.058) $n = 1538$	**Subclass VIII** dur: -0.178 (0.096) freq: 0.016 (0.072) $n = 954$	**Subclass XI** dur: 0.018 (0.138) freq: 0.026 (0.106) $n = 532$

The Smooth Coefficient Model Fit



Example 2: Effects of Education on Wages

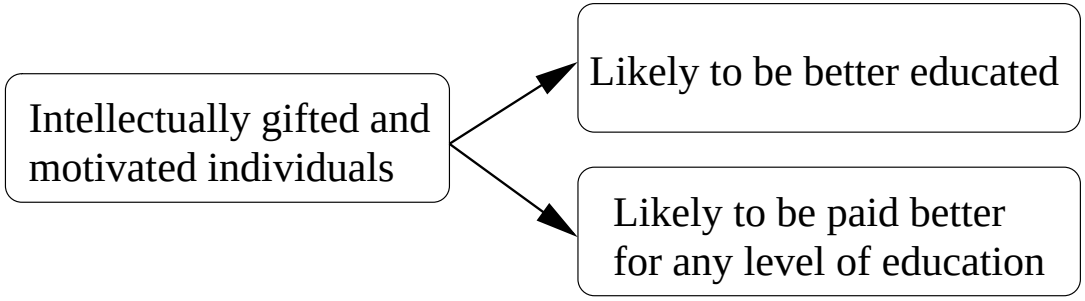
- Data: 16,193 men born between 1949 and 1953 from U.S. Current Population Surveys used in Angrist and Krueger (1995).
- Treatment variable: highest grade completed (0–18).
- Outcome variable: annual wage in 1978 dollars.
- Covariates: race, year of birth, marital status, veteran status, Vietnam draft lottery code, region of residence, and indicator variables for residence in a central city and employment in a Standard Metropolitan Statistical Area.
- Following Angrist and Krueger (1995), we exclude men who did not work and/or recorded zero earning as well as those who have missing values for at least one variable. This yields the sample size of 13,900 for our analysis.

Propensity Function Method for Education Data

1. Model specification for propensity function: Ordinal logistic model.
   - $\hat{\theta} = X^\top \hat{\beta}$ completely identifies propensity function.
2. Stratify data based on the scalar $\hat{\theta}$.
3. Within strata model: Gaussian linear regression, $p(Y \mid T^A, \hat{\theta})$.

But what about the ignorability assumption?

Are T^A and $Y(t^P)$ independent given X?



Covariates do not measure native intelligence or work ethic.

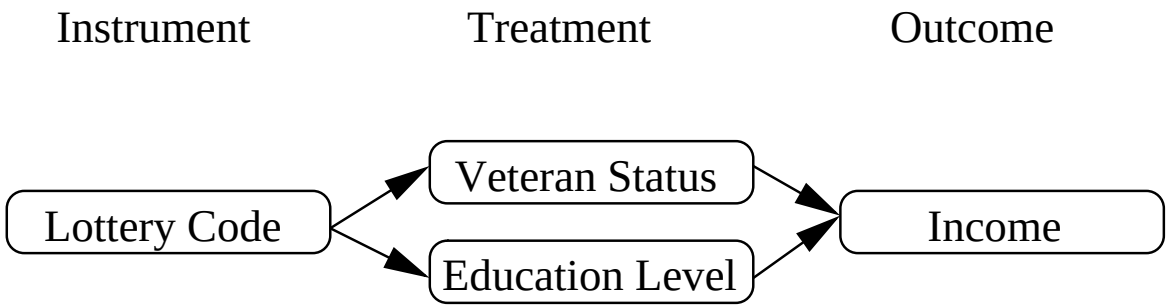
An Instrumental Variables Analysis

An instrumental variable:

- is independent of the potential treatments and the potential outcomes given the covariates (i.e., ignorability),
- does not affect the potential outcome given the treatment and the covariates,
- is monotonically predictive of the treatment.

(These assumptions are required for causal interpretation, see Angrist, Imbens and Rubin, JASA, 1996.)

The Vietnam Draft Lottery:

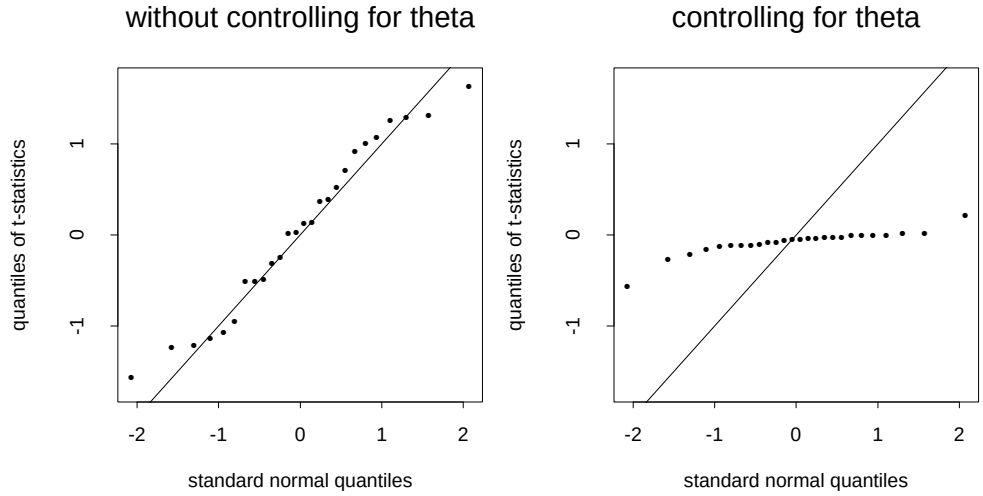


Balancing the Covariates Across the Instrument

The Lottery Code is recorded as an 14-category ordinal variable.

- We use a ordinal logistic model for the propensity function
- The model is characterized by the scalar linear predictor, $\hat{\theta} = X\hat{\beta}$.

Balance



- First panel: T-statistics for predicting the covariates from the instrument
- Second panel: Same, but controlling for the linear predictor.

The Balance is Improved

Estimated Effects of Education

The average effect of a one year increase in the highest grade on log weekly wage:

	Direct Models		Propensity Function	
	TSLS	SSIV	5 blocks	10 blocks
average causal effect	0.109	0.040	0.062	0.063
standard error	0.034	0.037	0.015	0.010

(SSIV=Split-Sample Instrumental Variables, Angrist and Krueger (1995))

Non-Constant Treatment Effects

The treatment effect appears to vary with the covariates, as measured by $\hat{\theta}$:

	Block 1	Block 2	Block 3	Block 4	Block 5	Overall
estimate	0.0839	0.0628	0.0200	0.0541	0.0896	0.0621
std. error	0.0278	0.0348	0.0284	0.0357	0.0358	0.0146

Unfortunately, the heterogeneity in treatment effect is difficult to make sense of because $\hat{\theta}$ is not easily interpretable.

Current Research aims at using propensity functions to identify the linear combination of the covariates that best predicts "compliance" with the randomized treatment. We expect this linear combination to be predictive of the magnitude of the *intention to treat* causal effect.

This strategy

- uses propensity functions to (partially) identify heterogeneity of treatment effects in IV analyses rather than to balance the instrument, and,
- provides a more interpretable causal effect than TSLS when the treatment is not binary. (See Angrist and Imbens, JASA, 1995).

Angrist and Imbens' Result

Under the above IV assumptions, suppose

- Z is a binary instrumental variable,
- T_1 and T_2 are the two potential treatments taking values in $\{0, 1, \ldots, J\}$,
- $Y_1, \ldots, Y_J$ are the potential outcomes, and
- there are no covariates.

Angrist and Imbens showed that a population version of the TSLS estimate,

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{E(T|Z=1) - E(T|Z=0)} = \sum_{j=1}^{J} \omega_j E(Y_j - Y_{j-1}|T_1 \geq j > T_0), \qquad (1)$$

where

$$\omega_j = \frac{\Pr(T_1 \geq j > T_0)}{\sum_{i=1}^{J} \Pr(T_1 \geq i > T_0).}$$

Eqn. (1) is a weighted average of per unit change in treatment effects. Unfortunately, the average is over overlapping subpopulations. (E.g., individuals with $T_1 = J$ and $T_0 = 0$ *are members of all the subpopulations.*)

Concluding Remarks

- Our generalization widens the range of potential applications of propensity score methods.
- Subclassification on propensity function allows flexible modeling strategies.
- Propensity function methods retain attractive features of propensity score methods:
  1. Low dimensional summary of high dimensional covariates.
  2. Useful diagnostics: balance of covariates.
  3. Relatively robust to misspecification.
  4. Reduces bias and mean squared error relative to linear models.