Causal inference weeks 2 & 3: Selection on observables

Regression, matching, and sub-classification

Andy Eggers

Oxford DPIR

HT 2018

Motivation

If treatment is randomized, the **difference in group means (DIGM)** (naive estimator) is an unbiased estimator of the **average treatment effect (ATE)**:

Motivation

If treatment is randomized, the **difference in group means (DIGM)** (naive estimator) is an unbiased estimator of the **average treatment effect (ATE)**:

$$\begin{aligned} E\Big[\mathrm{DIGM}\Big] &\equiv E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] && \text{(definition)} \\ &= E[Y_{1i}] - E[Y_{0i}] && \text{(randomization)} \\ &= \mathrm{ATE} && \text{(definition)} \end{aligned}$$

But more generally (e.g. in observational study),

$$E\Big[\mathrm{DIGM}\Big] = \text{ATT} + \text{selection bias}$$

Motivation

If treatment is randomized, the **difference in group means (DIGM)** (naive estimator) is an unbiased estimator of the **average treatment effect (ATE)**:

$$
\begin{aligned}
E\Big[\mathrm{DIGM}\Big] &\equiv E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] && \text{(definition)} \\
&= E[Y_{1i}] - E[Y_{0i}] && \text{(randomization)} \\
&= \mathrm{ATE} && \text{(definition)}
\end{aligned}
$$

But more generally (e.g. in observational study),

$$E\Big[\mathrm{DIGM}\Big] = \text{ATT} + \text{selection bias}$$

This is why, instead of just calculating DIGM, we **control for covariates** using e.g. regression.

Motivation

If treatment is randomized, the **difference in group means (DIGM)** (naive estimator) is an unbiased estimator of the **average treatment effect (ATE)**:

$$\begin{aligned} E\Big[\mathrm{DIGM}\Big] &\equiv E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] && \text{(definition)} \\ &= E[Y_{1i}] - E[Y_{0i}] && \text{(randomization)} \\ &= \mathrm{ATE} && \text{(definition)} \end{aligned}$$

But more generally (e.g. in observational study),

$$E\Big[\mathrm{DIGM}\Big] = \text{ATT} + \text{selection bias}$$

This is why, instead of just calculating DIGM, we **control for covariates** using e.g. regression.

DIGM is also known as the **naive estimator**. Let's not be so naive!

Plan

- Introduction to **covariate adjustment** and the key assumption behind it: the **conditional independence assumption (CIA)**

Plan

- Introduction to **covariate adjustment** and the key assumption behind it: the **conditional independence assumption (CIA)**
- With running example of “MPs for Sale?” (2009), illustration using three methods of **covariate adjustment**:

Plan

- Introduction to **covariate adjustment** and the key assumption behind it: the **conditional independence assumption (CIA)**
- With running example of “MPs for Sale?” (2009), illustration using three methods of **covariate adjustment**:
  - sub-classification
  - matching
  - regression

Plan

- Introduction to **covariate adjustment** and the key assumption behind it: the **conditional independence assumption (CIA)**
- With running example of “MPs for Sale?” (2009), illustration using three methods of **covariate adjustment**:
  - sub-classification
  - matching
  - regression

Ultimately use regression, but understand others.

Introduction to covariate adjustment

Covariate adjustment based on categorical variables

Covariate adjustment using the propensity score

Covariate adjustment in sparse data without the propensity score

Two important facts about regression

Setup

You are given a very large (population-level) dataset with three columns, labeled Y_i, D_i, and X_i, and asked to assess the effect of D_i on Y_i.

Setup

You are given a very large (population-level) dataset with three columns, labeled Y_i, D_i, and X_i, and asked to assess the effect of D_i on Y_i.

You calculate DIGM: $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 1.5 - 0.75 = 0.75$.

Setup

You are given a very large (population-level) dataset with three columns, labeled Y_i, D_i, and X_i, and asked to assess the effect of D_i on Y_i.

You calculate DIGM: $E[Y_{1i}|D_i = 1] – E[Y_{0i}|D_i = 0] = 1.5 – 0.75 = 0.75$.

But using `table()` and `prop.table()` in `R` you observe that D_i is related to X_i:

Joint distribution of X_i and D_i in the dataset

	$D_i = 0$	$D_i = 1$
$X_i = 0$	3/8	1/8
$X_i = 1$	1/8	3/8

Setup

You are given a very large (population-level) dataset with three columns, labeled Y_i, D_i, and X_i, and asked to assess the effect of D_i on Y_i.

You calculate DIGM: $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 1.5 - 0.75 = 0.75$.

But using `table()` and `prop.table()` in `R` you observe that D_i is related to X_i:

Joint distribution of X_i and D_i in the dataset

	$D_i = 0$	$D_i = 1$
$X_i = 0$	3/8	1/8
$X_i = 1$	1/8	3/8

To discuss: Think of an example where D_i and X_i might be related in this way. What does Y represent in your example?

Outcomes by X and D

You calculate the mean outcomes by X_i and D_i:

	$D_i = 0$	$D_i = 1$
$X_i = 0$	0	0
$X_i = 1$	3	2

Outcomes by X and D

You calculate the mean outcomes by X_i and D_i:

	$D_i = 0$	$D_i = 1$
$X_i = 0$	0	0
$X_i = 1$	3	2

So, what is the effect of D_i on Y_i?

Conditional independence assumption (CIA)

Note: Cannot proceed without assumptions.

Conditional independence assumption (CIA)

Note: Cannot proceed without assumptions.

One possible assumption: treatment “as-if random” within levels of X_i.

Conditional independence assumption (CIA)

Note: Cannot proceed without assumptions.

One possible assumption: treatment “as-if random” within levels of X_i.

Equivalent (but more technical) assumptions:

- potential outcomes independent of D_i given X_i: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$
- no selection bias conditional on X_i:
  $E[Y_{0i}|D_i = 1, X_i = x] = E[Y_{0i}|D_i = 0, X_i = x] \ \forall x \in \{0, 1\}$

Conditional independence assumption (CIA)

Note: Cannot proceed without assumptions.

One possible assumption: treatment "as-if random" within levels of X_i.

Equivalent (but more technical) assumptions:

- potential outcomes independent of D_i given X_i: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$
- no selection bias conditional on X_i: $E[Y_{0i}|D_i = 1, X_i = x] = E[Y_{0i}|D_i = 0, X_i = x] \; \forall x \in \{0, 1\}$

This is the **conditional independence assumption (CIA)**, aka "selection on observables".

Conditional independence assumption (CIA)

Note: Cannot proceed without assumptions.

One possible assumption: treatment "as-if random" within levels of X_i.

Equivalent (but more technical) assumptions:

- potential outcomes independent of D_i given X_i: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$
- no selection bias conditional on X_i:
  $E[Y_{0i}|D_i = 1, X_i = x] = E[Y_{0i}|D_i = 0, X_i = x] \; \forall x \in \{0, 1\}$

This is the **conditional independence assumption (CIA)**, aka "selection on observables".

If CIA holds, DIGM gives us ATE *within levels of* X_i.

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE_x).

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE$_x$).

$$\text{CATE}_x \equiv E[Y_{1i} - Y_{0i} | X_i = x]$$

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE_x).

$$\text{CATE}_x \equiv E[Y_{1i} - Y_{0i} | X_i = x]$$

Given CIA we have

	$D_i = 0$	$D_i = 1$	CATE_x
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE_x).

$$\text{CATE}_x \equiv E[Y_{1i} - Y_{0i} | X_i = x]$$

Given CIA we have

	$D_i = 0$	$D_i = 1$	CATE_x
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

(Thinking of examples, why would CATE_x vary with x?)

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE_x).

$$\text{CATE}_x \equiv E[Y_{1i} - Y_{0i} | X_i = x]$$

Given CIA we have

	$D_i = 0$	$D_i = 1$	CATE_x
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

(Thinking of examples, why would CATE_x vary with x?)

We summarize average effect of D_i on Y_i by calculating **weighted averages** of CATE_xs.

CATE, ATE, ATT, ATC

ATE where $X_i = x$ is called **Conditional Average Treatment Effect** (CATE$_x$).

$$\text{CATE}_x \equiv E[Y_{1i} - Y_{0i} | X_i = x]$$

Given CIA we have

	$D_i = 0$	$D_i = 1$	CATE$_x$
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

(Thinking of examples, why would CATE$_x$ vary with x?)

We summarize average effect of D_i on Y_i by calculating **weighted averages** of CATE$_x$s.

When CATE$_x$ weighted by distribution of X_i in the

- population: ATE
- treatment group: ATT (ATE on the treated)
- control group: ATC (ATE on the control)

CATE, ATE, ATT, ATC in this example

Reminder: our CATE$_x$s

	$D_i = 0$	$D_i = 1$	CATE$_x$
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

Reminder: Joint distribution of X_i and D_i:

	$D_i = 0$	$D_i = 1$
$X_i = 0$	3/8	1/8
$X_i = 1$	1/8	3/8

CATE, ATE, ATT, ATC in this example

Reminder: our CATE$_x$s

	$D_i = 0$	$D_i = 1$	CATE$_x$
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

Reminder: Joint distribution of X_i and D_i:

	$D_i = 0$	$D_i = 1$
$X_i = 0$	3/8	1/8
$X_i = 1$	1/8	3/8

Reminder: When CATE$_x$ weighted by distribution of X_i in the

- population: ATE
- treatment group: ATT (ATE on the treated)
- control group: ATC (ATE on the control)

CATE, ATE, ATT, ATC in this example

Reminder: our CATE_xs

	$D_i = 0$	$D_i = 1$	CATE_x
$X_i = 0$	0	0	0
$X_i = 1$	3	2	-1

Reminder: Joint distribution of X_i and D_i:

	$D_i = 0$	$D_i = 1$
$X_i = 0$	3/8	1/8
$X_i = 1$	1/8	3/8

Reminder: When CATE_x weighted by distribution of X_i in the

- population: ATE
- treatment group: ATT (ATE on the treated)
- control group: ATC (ATE on the control)

So what is ATE, ATT, ATC in this example (given CIA)?

Illustration for this example

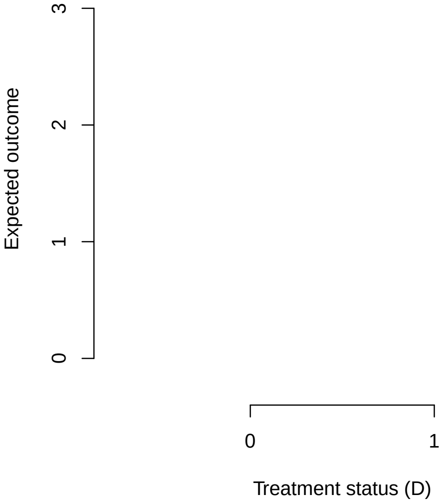


Illustration for this example

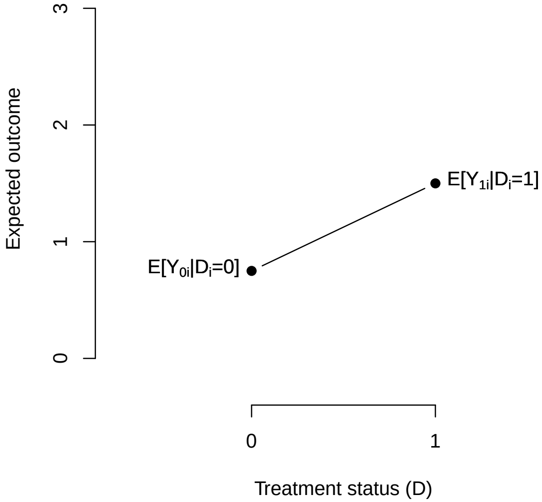


Illustration for this example

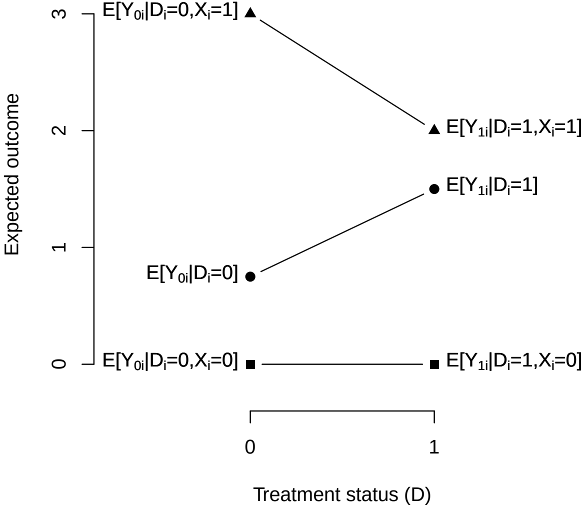


Illustration for this example

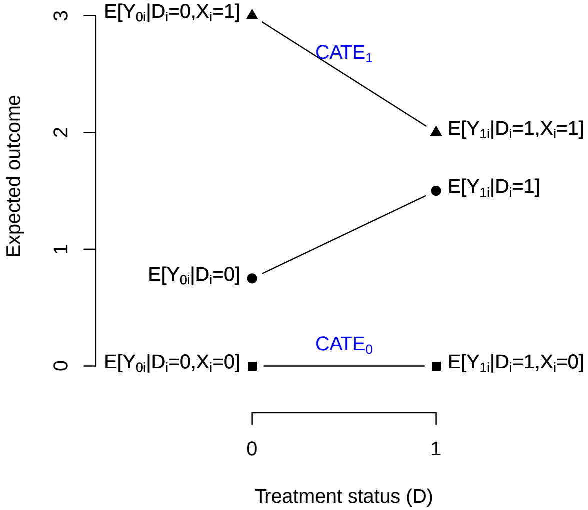


Illustration for this example

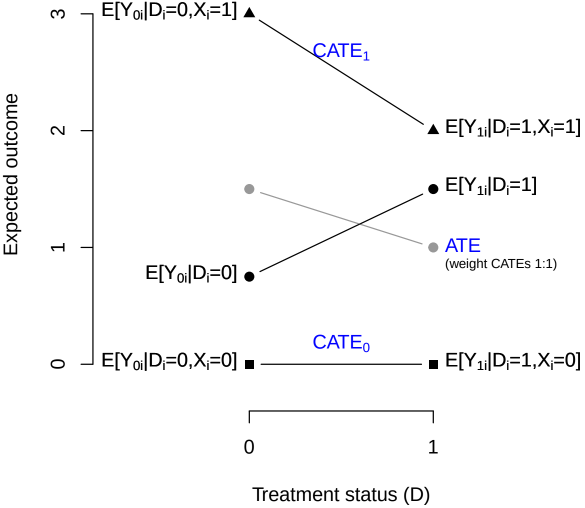


Illustration for this example

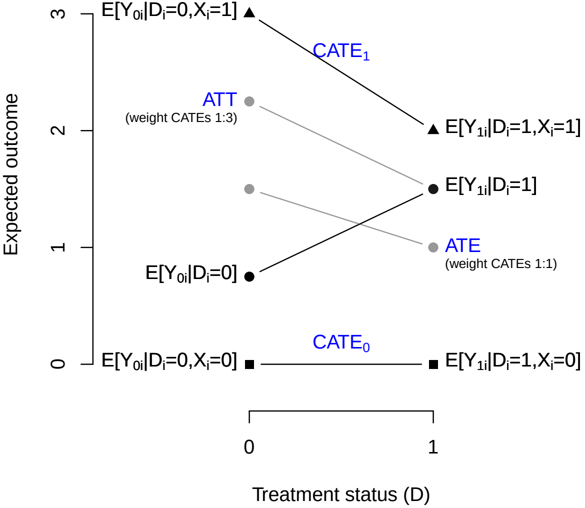


Illustration for this example

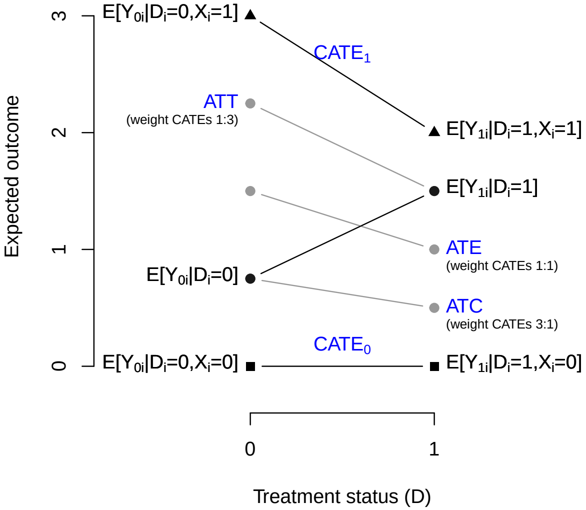
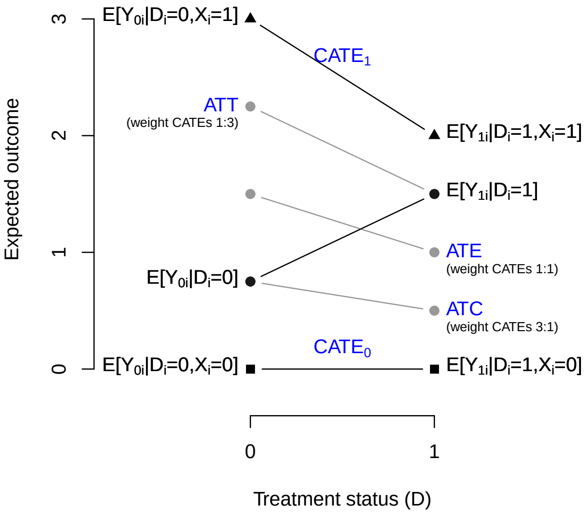


Illustration for this example (2)



Check understanding:

- Why are ATE, ATT, ATC all the same in a randomized experiment?
- How can the DIGM be positive when neither CATE is positive?

Mathier explanation

The ATT is the weighted average of the CATEs, where the weights reflect the distribution of X_i in the treatment group.

$$\begin{aligned} \text{ATT} &= \sum_{x=0,1} E\Big[Y_i(1) - Y_i(0) | X_i = x\Big] \Pr\Big(X_i = x | D_i = 1\Big) && (1) \\ &= \sum_{x=0,1} \text{CATE}_x \Pr\Big(X_i = x | D_i = 1\Big) && (2) \end{aligned}$$

In this case, this is

$$\text{ATT} = 0 \times 1/4 + -1 \times 3/4 = -3/4.$$

The weights of $1/4$ and $3/4$ come from the joint distribution: the probability of $X_i = 1$ given that $D_i = 1$ can be calculated as the ratio $\frac{3/8}{3/8+1/8}$, which is a simple application of Bayes' Theorem.

Does the CIA hold?

The CIA is an untestable assumption: it relies on “theory”, some indirect evidence.

Does the CIA hold?

The CIA is an untestable assumption: it relies on “theory”, some indirect evidence.

Two key ways it might fail in this example:

- There might be another variable Z_i related to D_i that affects Y_i (**selection bias** persists).

Does the CIA hold?

The CIA is an untestable assumption: it relies on “theory”, some indirect evidence.

Two key ways it might fail in this example:

- There might be another variable Z_i related to D_i that affects Y_i (**selection bias** persists).
- X_i might be an outcome, so controlling for it introduces **post-treatment bias**.

Does the CIA hold?

The CIA is an untestable assumption: it relies on “theory”, some indirect evidence.

Two key ways it might fail in this example:

- There might be another variable Z_i related to D_i that affects Y_i (**selection bias** persists).
- X_i might be an outcome, so controlling for it introduces **post-treatment bias**. e.g. suppose D_i had been randomly assigned – how would we interpret this data?

What do we condition on/control for?

For the CIA to hold, we must control for every variable that

- is not an effect of D_i
- affects D_i and, conditional on other control variables, affects Y_i.

What do we condition on/control for?

For the CIA to hold, we must control for every variable that

- is not an effect of D_i
- affects D_i and, conditional on other control variables, affects Y_i.

(For guidance on what to control for, see Morgan & Winship and the "backdoor criterion".)

What do we condition on/control for?

For the CIA to hold, we must control for every variable that

- is not an effect of D_i
- affects D_i and, conditional on other control variables, affects Y_i.

(For guidance on what to control for, see Morgan & Winship and the "backdoor criterion".)

To discuss: In regressions, generally only one coefficient (at most) can be interpreted as a (causal) effect. Why is that?

How do we condition/control?

We'll discuss three approaches:

How do we condition/control?

We'll discuss three approaches:

- Sub-classification
  - put units into cells according to X_i
  - calculate CATE_x within cells and average

How do we condition/control?

We'll discuss three approaches:

- Sub-classification
  - put units into cells according to X_i
  - calculate CATE$_x$ within cells and average
- Matching
  - for each treated unit (or each control unit) (or each unit), find unit(s) same/similar on X_i with opposite D_i
  - calculate DIGM

How do we condition/control?

We’ll discuss three approaches:

- Sub-classification
  - put units into cells according to X_i
  - calculate CATE$_x$ within cells and average
- Matching
  - for each treated unit (or each control unit) (or each unit), find unit(s) same/similar on X_i with opposite D_i
  - calculate DIGM
- Regression
  - estimate $E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i$

How do we condition/control?

We'll discuss three approaches:

- Sub-classification
    - put units into cells according to X_i
    - calculate CATE_x within cells and average
- Matching
    - for each treated unit (or each control unit) (or each unit), find unit(s) same/similar on X_i with opposite D_i
    - calculate DIGM
- Regression
    - estimate $E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i$

For simple cases, basically identical.

How do we condition/control?

We'll discuss three approaches:

- Sub-classification
  - put units into cells according to X_i
  - calculate CATE$_x$ within cells and average
- Matching
  - for each treated unit (or each control unit) (or each unit), find unit(s) same/similar on X_i with opposite D_i
  - calculate DIGM
- Regression
  - estimate $E[Y_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i$

For simple cases, basically identical.

Link to CIA more obvious for sub-classification and matching, but regression more flexible and common.

Running example

Eggers and Hainmueller, “MPs for Sale” (2009)

Running example

Eggers and Hainmueller, “MPs for Sale” (2009)

Basic question: Was election to the UK House of Commons financially rewarding?

Running example

Eggers and Hainmueller, “MPs for Sale” (2009)

Basic question: Was election to the UK House of Commons financially rewarding?

- **Units:** Candidates who ran for parliament at some point between 1950 and 1970
- **Treatment:** Winning at least one election
- **Outcome:** Wealth at death (probate value)

Running example

Eggers and Hainmueller, “MPs for Sale” (2009)

Basic question: Was election to the UK House of Commons financially rewarding?

- **Units:** Candidates who ran for parliament at some point between 1950 and 1970
- **Treatment:** Winning at least one election
- **Outcome:** Wealth at death (probate value)

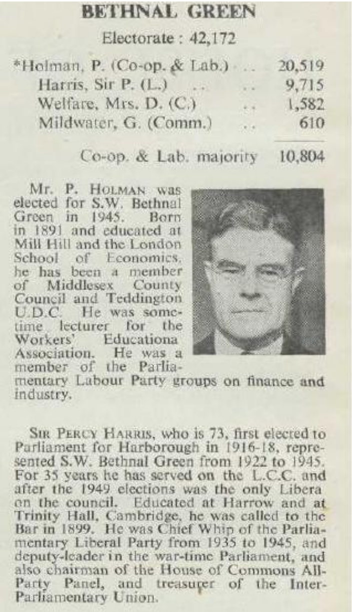
What about the DIGM as an estimator for ATE?

Covariates

For each candidate, we have

- party
- electoral results
- year of birth
- secondary education
- university education
- profession

How can we use these to make better comparisons?



BETHNAL GREEN

Electorate : 42,172

*Holman, P. (Co-op. & Lab.)	20,519
Harris, Sir P. (L.)	9,715
Welfare, Mrs. D. (C.) ..	1,582
Mildwater, G. (Comm.) ..	610
Co-op. & Lab. majority	10,804

Mr. P. HOLMAN was elected for S.W. Bethnal Green in 1945. Born in 1891 and educated at Mill Hill and the London School of Economics, he has been a member of Middlesex County Council and Teddington U.D.C. He was sometime lecturer for the Workers' Educationa Association. He was a member of the Parliamentary Labour Party groups on finance and industry.

SIR PERCY HARRIS, who is 73, first elected to Parliament for Harborough in 1916-18, represented S.W. Bethnal Green from 1922 to 1945. For 35 years he has served on the L.C.C. and after the 1949 elections was the only Libera on the council. Educated at Harrow and at Trinity Hall, Cambridge, he was called to the Bar in 1899. He was Chief Whip of the Parliamentary Liberal Party from 1935 to 1945, and deputy-leader in the war-time Parliament, and also chairman of the House of Commons All-Party Panel, and treasurer of the Inter-Parliamentary Union.

Introduction to covariate adjustment

Covariate adjustment based on categorical variables

Covariate adjustment using the propensity score

Covariate adjustment in sparse data without the propensity score

Two important facts about regression

CIA based on categorical variables

Suppose we believe that CIA holds given candidate's

- party (Labour, Conservative)
- type of secondary education (Eton, other public, other, not mentioned), and
- type of university education (Oxbridge, other, not mentioned)

CIA based on categorical variables

Suppose we believe that CIA holds given candidate's

- party (Labour, Conservative)
- type of secondary education (Eton, other public, other, not mentioned), and
- type of university education (Oxbridge, other, not mentioned)

i.e. For pairs of candidates with the same party, school, and university, MP status as-if randomly assigned – not related to potential outcomes.

CIA based on categorical variables

Suppose we believe that CIA holds given candidate's

- party (Labour, Conservative)
- type of secondary education (Eton, other public, other, not mentioned), and
- type of university education (Oxbridge, other, not mentioned)

i.e. For pairs of candidates with the same party, school, and university, MP status as-if randomly assigned – not related to potential outcomes.

Plausible?

Number of candidates by party & education

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	16	4	5	2	0	0
Other public	40	25	31	21	10	3
Other	10	25	31	19	49	31
Not mentioned	7	12	17	5	27	37

Number of MPs and unsuccessful candidates by cell

Note: (2,1) indicates 2 elected candidates and 1 unelected candidate

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	(14, 2)	(3, 1)	(4, 1)	(0, 2)	(0, 0)	(0, 0)
Other public	(18, 22)	(11, 14)	(11, 20)	(6, 15)	(4, 6)	(1, 2)
Other	(2, 8)	(8, 17)	(10, 21)	(2, 17)	(15, 34)	(13, 18)
Not mentioned	(4, 3)	(7, 5)	(12, 5)	(1, 4)	(8, 19)	(11, 26)

Difference in group means (of log wealth at death) by cell

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Sub-classification

Sub-classification: calculate DIGMs in each cell; average them.

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Sub-classification

Sub-classification: calculate DIGMs in each cell; average them.

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Average DIGM across cells

Sub-classification

Sub-classification: calculate DIGMs in each cell; average them.

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Average DIGM across cells

weighting by #**candidates** in each cell: .40 (ATE)

Sub-classification

Sub-classification: calculate DIGMs in each cell; average them.

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Average DIGM across cells

weighting by #**candidates** in each cell:	.40	(ATE)
weighting by #**MPs** in each cell:	.54	(ATT)

Sub-classification

Sub-classification: calculate DIGMs in each cell; average them.

	Party: Con.			**Party:** Lab.		
University:	Oxbridge	Other	None	Oxbridge	Other	None
School: Eton	2.61	2.66	-0.67	-	-	-
Other public	0.35	0.15	0.48	0.65	-0.27	0.58
Other	1.05	0.33	0.51	-0.02	-0.01	0.45
Not mentioned	0.06	1.48	0.25	-0.6	0.27	0.06

Average DIGM across cells

weighting by #**candidates** in each cell:	.40	(ATE)
weighting by #**MPs** in each cell:	.54	(ATT)
weighting by #**non-MPs** in each cell:	.31	(ATC)

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status.

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Matching

Matching: fill in missing potential outcomes using "nearest neighbor" with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	?	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	
Labour	Other Public	Not Mentioned	0	12.5	12.5	

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	
Labour	Other Public	Not Mentioned	0	12.5	12.5	

Matching

Matching: fill in missing potential outcomes using "nearest neighbor" with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	?
Labour	Other Public	Not Mentioned	0	12.5	12.5	?

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	12.7
Labour	Other Public	Not Mentioned	0	12.5	12.5	12.7

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	12.7
Labour	Other Public	Not Mentioned	0	12.5	12.5	12.7

To get ATE: take difference in mean of Y_{1i} and Y_{0i} (with imputations).

Matching

Matching: fill in missing potential outcomes using “nearest neighbor” with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	12.7
Labour	Other Public	Not Mentioned	0	12.5	12.5	12.7

To get ATE: take difference in mean of Y_{1i} and Y_{0i} (with imputations).
To get ATT: same, but only using treated rows.

Matching

Matching: fill in missing potential outcomes using "nearest neighbor" with opposite treatment status. (e.g. for Tory **MP** born in 1928 who went to Oxford, use Tory **non-MP** born in 1927 who went to Cambridge)

Exact matching: fill in missing potential outcomes using average of exact matches with opposite treatment status.

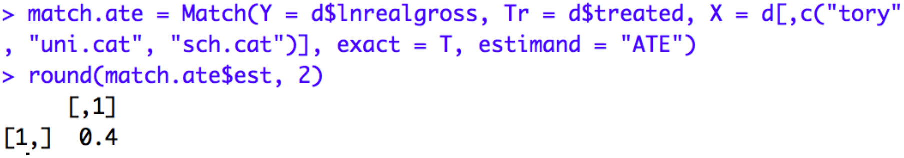
For example:

Party	School	University	Treated	ln(Wealth)	Y_{0i}	Y_{1i}
Labour	Other Public	Not Mentioned	1	12.7	12.15	12.7
Labour	Other Public	Not Mentioned	0	11.8	11.8	12.7
Labour	Other Public	Not Mentioned	0	12.5	12.5	12.7

To get ATE: take difference in mean of Y_{1i} and Y_{0i} (with imputations).
To get ATT: same, but only using treated rows.
To get ATC: same, but only using control rows.

Exact matching: implementation

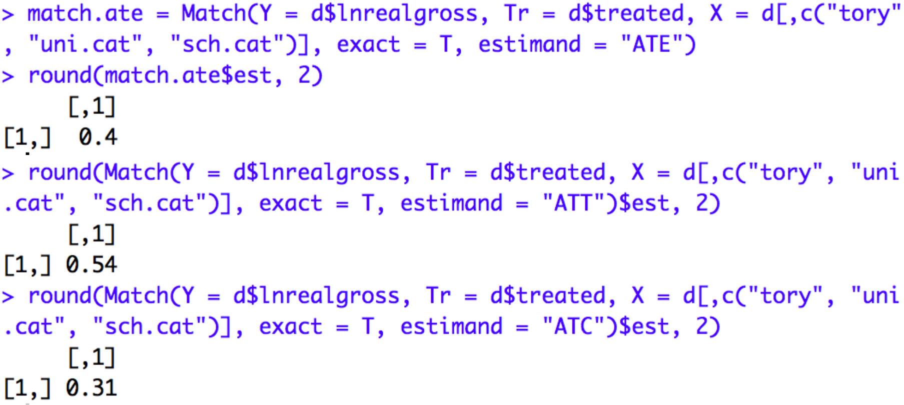
Using `Matching` library:



```
> match.ate = Match(Y = d$lnrealgross, Tr = d$treated, X = d[,c("tory"
, "uni.cat", "sch.cat")], exact = T, estimand = "ATE")
> round(match.ate$est, 2)
     [,1]
[1,]  0.4
```

Exact matching: implementation

Using `Matching` library:



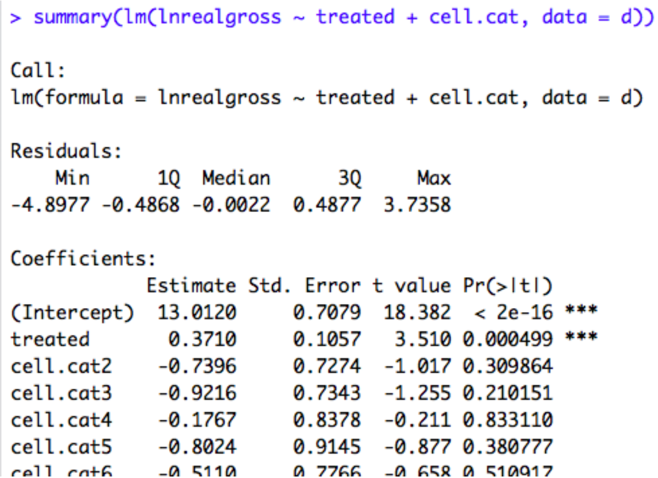
```
> match.ate = Match(Y = d$lnrealgross, Tr = d$treated, X = d[,c("tory"
, "uni.cat", "sch.cat")], exact = T, estimand = "ATE")
> round(match.ate$est, 2)
     [,1]
[1,]  0.4
> round(Match(Y = d$lnrealgross, Tr = d$treated, X = d[,c("tory", "uni
.cat", "sch.cat")], exact = T, estimand = "ATT")$est, 2)
     [,1]
[1,] 0.54
> round(Match(Y = d$lnrealgross, Tr = d$treated, X = d[,c("tory", "uni
.cat", "sch.cat")], exact = T, estimand = "ATC")$est, 2)
     [,1]
[1,] 0.31
```

Regression

Given CIA, **regression** implies regressing Y_i on an indicator for treatment and a dummy for every cell.

Regression

Given CIA, **regression** implies regressing Y_i on an indicator for treatment and a dummy for every cell.



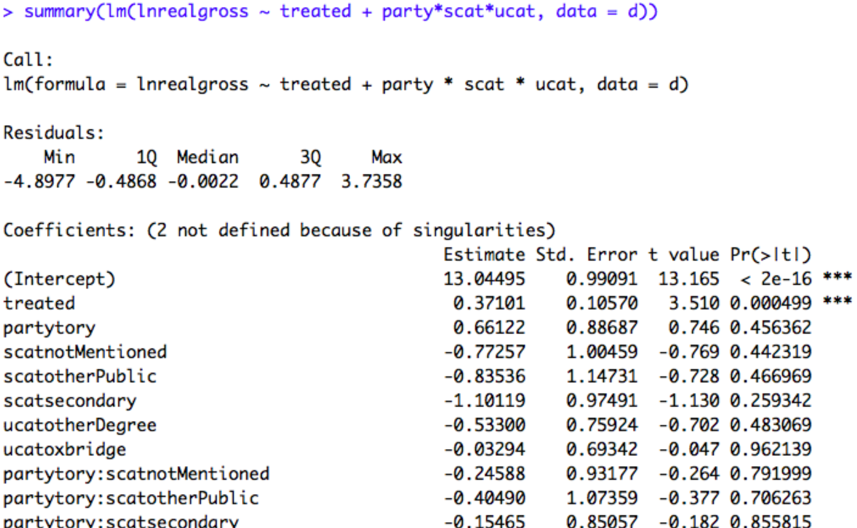
```
> summary(lm(lnrealgross ~ treated + cell.cat, data = d))

Call:
lm(formula = lnrealgross ~ treated + cell.cat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8977 -0.4868 -0.0022  0.4877  3.7358 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.0120     0.7079  18.382  < 2e-16 ***
treated       0.3710     0.1057   3.510 0.000499 ***
cell.cat2    -0.7396     0.7274  -1.017 0.309864    
cell.cat3    -0.9216     0.7343  -1.255 0.210151    
cell.cat4    -0.1767     0.8378  -0.211 0.833110    
cell.cat5    -0.8024     0.9145  -0.877 0.380777    
cell.cat6    -0.5110     0.7766  -0.658 0.510917    
```

Regression (2)

Equivalent: regressing Y_i on each of the categorical variables and their interactions:

Regression (2)

Equivalent: regressing Y_i on each of the categorical variables and their interactions:



```
> summary(lm(lnrealgross ~ treated + party*scat*ucat, data = d))

Call:
lm(formula = lnrealgross ~ treated + party * scat * ucat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8977 -0.4868 -0.0022  0.4877  3.7358 

Coefficients: (2 not defined because of singularities)
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            13.04495    0.99091  13.165  < 2e-16 ***
treated                                 0.37101    0.10570   3.510 0.000499 ***
partytory                               0.66122    0.88687   0.746 0.456362    
scatnotMentioned                       -0.77257    1.00459  -0.769 0.442319    
scatotherPublic                        -0.83536    1.14731  -0.728 0.466969    
scatsecondary                          -1.10119    0.97491  -1.130 0.259342    
ucatotherDegree                        -0.53300    0.75924  -0.702 0.483069    
ucatoxbridge                           -0.03294    0.69342  -0.047 0.962139    
partytory:scatnotMentioned             -0.24588    0.93177  -0.264 0.791999    
partytory:scatotherPublic              -0.40490    1.07359  -0.377 0.706263    
partytory:scatsecondary                -0.15465    0.85057  -0.182 0.855815    
```

Saturated regression models

This is a **saturated model**, i.e. one with a dummy for each possible combination of the explanatory variables

Saturated regression models

This is a **saturated model**, i.e. one with a dummy for each possible combination of the explanatory variables:

- all “**main effects**” (Labour, Eton, . . .), and
- all **interactions** (Labour × Eton, Labour × OtherPublic, . . .)).

Saturated regression models

This is a **saturated model**, i.e. one with a dummy for each possible combination of the explanatory variables:

- all “**main effects**” (Labour, Eton, . . .), and
- all **interactions** (Labour × Eton, Labour × OtherPublic, . . .)).

Given categorical variables, a saturated model is the most flexible possible functional form.

Saturated regression models

This is a **saturated model**, i.e. one with a dummy for each possible combination of the explanatory variables:

- all “**main effects**” (Labour, Eton, . . .), and
- all **interactions** (Labour × Eton, Labour × OtherPublic, . . .)).

Given categorical variables, a saturated model is the most flexible possible functional form.

To discuss: Why would you prefer a saturated model to a model with only main effects, no interactions?

Comparison of techniques

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.54	.31
Matching (exact)	.40	.54	.31
Regression	.37	-	-

Comparison of techniques

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.54	.31
Matching (exact)	.40	.54	.31
Regression	.37	-	-

To note:

- ATE from sub-classification is average of cell DIGMs weighted by {#units in cell} (definition).

Comparison of techniques

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.54	.31
Matching (exact)	.40	.54	.31
Regression	.37	-	-

To note:

- ATE from sub-classification is average of cell DIGMs weighted by {#units in cell} (definition).
- ATE from (exact) matching is exactly the same thing.

Comparison of techniques

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.54	.31
Matching (exact)	.40	.54	.31
Regression	.37	-	-

To note:

- ATE from sub-classification is average of cell DIGMs weighted by {#units in cell} (definition).
- ATE from (exact) matching is exactly the same thing.
- ATE from saturated regression is the average of cell DIGMs weighted by {#units in cell × variance of treatment in cell} (Angrist and Pischke MHE p. 75).

Introduction to covariate adjustment

Covariate adjustment based on categorical variables

Covariate adjustment using the propensity score

Covariate adjustment in sparse data without the propensity score

Two important facts about regression

CIA based on further categorical variables

Perhaps we don’t quite believe the CIA based only on party, secondary education, and university education.

CIA based on further categorical variables

Perhaps we don’t quite believe the CIA based only on party, secondary education, and university education.

The candidate’s profession before entering politics is another likely confounder.

CIA based on further categorical variables

Perhaps we don't quite believe the CIA based only on party, secondary education, and university education.

The candidate's profession before entering politics is another likely confounder.

But here we run into a problem of **sparse data** (curse of dimensionality):

- Sub-classification: many empty cells
- Matching: few exact matches
- (Saturated) regression: many empty groups, `NA` coefficients

CIA based on further categorical variables

Perhaps we don't quite believe the CIA based only on party, secondary education, and university education.

The candidate's profession before entering politics is another likely confounder.

But here we run into a problem of **sparse data** (curse of dimensionality):

- Sub-classification: many empty cells
- Matching: few exact matches
- (Saturated) regression: many empty groups, `NA` coefficients

You may get an estimate, but it will be based on an unrepresentative subset (far from true ATE).

How to proceed when many cells are empty

What can we do?

- Sub-classification: propensity score methods
- Matching: propensity score methods, nearest neighbor, coarsened exact matching
- Regression: propensity score methods, stronger CIA (i.e. less flexible functional form, e.g. drop interactions)

Propensity score methods

The propensity score is the probability of treatment, given covariates:

$$p(X_i) \equiv \Pr(D_i = 1|X_i) = E[D_i|X_i]$$

This can be estimated with OLS (linear probability model) or logistic regression (logit).

Propensity score methods

The propensity score is the probability of treatment, given covariates:

$$p(X_i) \equiv \Pr(D_i = 1|X_i) = E[D_i|X_i]$$

This can be estimated with OLS (linear probability model) or logistic regression (logit).

Ideally, the propensity score summarizes covariates that differ between treated and untreated units.

Propensity score methods

The propensity score is the probability of treatment, given covariates:

$$p(X_i) \equiv \Pr(D_i = 1|X_i) = E[D_i|X_i]$$

This can be estimated with OLS (linear probability model) or logistic regression (logit).

Ideally, the propensity score summarizes covariates that differ between treated and untreated units.

The CIA becomes: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i \,|\, p(X_i)$

Propensity score methods

The propensity score is the probability of treatment, given covariates:

$$p(X_i) \equiv \Pr(D_i = 1|X_i) = E[D_i|X_i]$$

This can be estimated with OLS (linear probability model) or logistic regression (logit).

Ideally, the propensity score summarizes covariates that differ between treated and untreated units.

The CIA becomes: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i \,|\, p(X_i)$

Having estimated the propensity score, we can

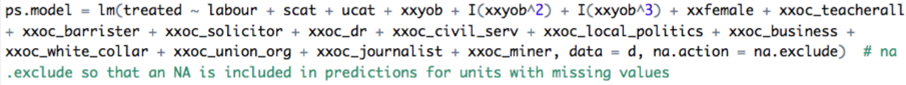
- Sub-classification: calculate DIGM within bands of the propensity score
- Matching: match units based on nearby propensity scores
- Regression: regress outcome on treatment controlling flexibly for the propensity score

Propensity score example

I regress the treatment indicator on party, secondary school category, university category, year of birth (yob), yob^2, yob^3, gender, 11 profession indicators.

Propensity score example

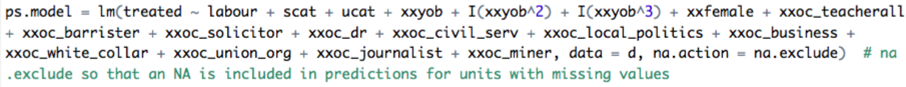
I regress the treatment indicator on party, secondary school category, university category, year of birth (yob), yob^2, yob^3, gender, 11 profession indicators.



```
ps.model = lm(treated ~ labour + scat + ucat + xxyob + I(xxyob^2) + I(xxyob^3) + xxfemale + xxoc_teacherall
+ xxoc_barrister + xxoc_solicitor + xxoc_dr + xxoc_civil_serv + xxoc_local_politics + xxoc_business +
xxoc_white_collar + xxoc_union_org + xxoc_journalist + xxoc_miner, data = d, na.action = na.exclude)  # na
.exclude so that an NA is included in predictions for units with missing values
```

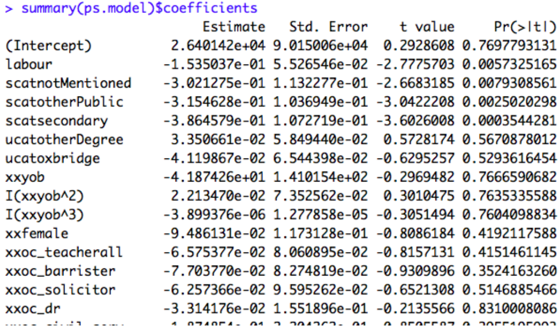
Propensity score example

I regress the treatment indicator on party, secondary school category, university category, year of birth (yob), yob^2, yob^3, gender, 11 profession indicators.



```
ps.model = lm(treated ~ labour + scat + ucat + xxyob + I(xxyob^2) + I(xxyob^3) + xxfemale + xxoc_teacherall
+ xxoc_barrister + xxoc_solicitor + xxoc_dr + xxoc_civil_serv + xxoc_local_politics + xxoc_business +
xxoc_white_collar + xxoc_union_org + xxoc_journalist + xxoc_miner, data = d, na.action = na.exclude)  # na
.exclude so that an NA is included in predictions for units with missing values
```

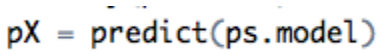
Take a look at the results:



```
> summary(ps.model)$coefficients
                       Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)        2.640142e+04 9.015006e+04  0.2928608 0.7697793131
labour            -1.535037e-01 5.526546e-02 -2.7775703 0.0057325165
scatnotMentioned  -3.021275e-01 1.132277e-01 -2.6683185 0.0079308561
scatotherPublic   -3.154628e-01 1.036949e-01 -3.0422208 0.0025020298
scatsecondary     -3.864579e-01 1.072719e-01 -3.6026008 0.0003544281
ucatotherDegree    3.350661e-02 5.849440e-02  0.5728174 0.5670878012
ucatoxbridge      -4.119867e-02 6.544398e-02 -0.6295257 0.5293616454
xxyob             -4.187426e+01 1.410154e+02 -0.2969482 0.7666590682
I(xxyob^2)         2.213470e-02 7.352562e-02  0.3010475 0.7635335588
I(xxyob^3)        -3.899376e-06 1.277858e-05 -0.3051494 0.7604098834
xxfemale          -9.486131e-02 1.173128e-01 -0.8086184 0.4192117588
xxoc_teacherall   -6.575377e-02 8.060895e-02 -0.8157131 0.4151461145
xxoc_barrister    -7.703770e-02 8.274819e-02 -0.9309896 0.3524163260
xxoc_solicitor    -6.257366e-02 9.595262e-02 -0.6521308 0.5146885466
xxoc_dr           -3.314176e-02 1.551896e-01 -0.2135566 0.8310008086
```

Propensity score example (2)

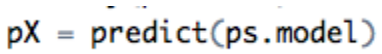
The propensity score is the prediction from this model:



```
pX = predict(ps.model)
```

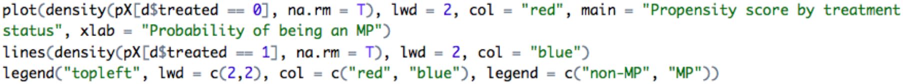
Propensity score example (2)

The propensity score is the prediction from this model:

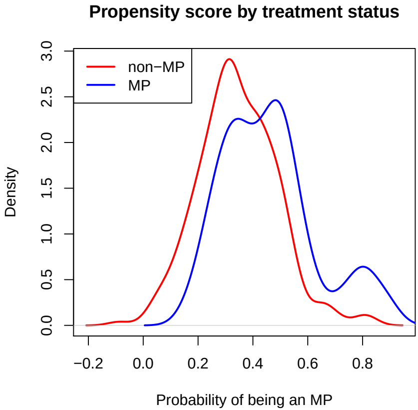


```
pX = predict(ps.model)
```

Compare the distribution by treatment status:

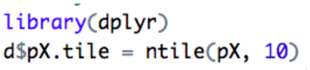


```
plot(density(pX[d$treated == 0], na.rm = T), lwd = 2, col = "red", main = "Propensity score by treatment
status", xlab = "Probability of being an MP")
lines(density(pX[d$treated == 1], na.rm = T), lwd = 2, col = "blue")
legend("topleft", lwd = c(2,2), col = c("red", "blue"), legend = c("non-MP", "MP"))
```



Sub-classification on the propensity score

Let's start with 10 sub-classes of the propensity score:



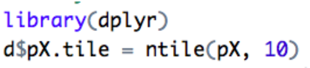
```
library(dplyr)
d$pX.tile = ntile(pX, 10)
```

Counts and DIGMs in each sub-class:

Subclass	#units	#MPs	#non-MPs	DIGM
1	43	4	39	-0.24
2	43	13	30	0.01
3	43	13	30	0.9
4	42	12	30	0.33
5	43	18	25	0.24
6	43	15	28	-0.13
7	42	15	27	0.35
8	43	20	23	0.5
9	43	26	17	0.47
10	40	29	11	1.69

Sub-classification on the propensity score

Let's start with 10 sub-classes of the propensity score:



```
library(dplyr)
d$pX.tile = ntile(pX, 10)
```

Counts and DIGMs in each sub-class:

Subclass	#units	#MPs	#non-MPs	DIGM
1	43	4	39	-0.24
2	43	13	30	0.01
3	43	13	30	0.9
4	42	12	30	0.33
5	43	18	25	0.24
6	43	15	28	-0.13
7	42	15	27	0.35
8	43	20	23	0.5
9	43	26	17	0.47
10	40	29	11	1.69

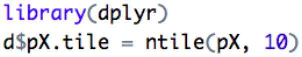
ATE: 0.40

ATT: 0.57

ATC: 0.30

Sub-classification on the propensity score

Let's start with 10 sub-classes of the propensity score:



```
library(dplyr)
d$pX.tile = ntile(pX, 10)
```

Counts and DIGMs in each sub-class:

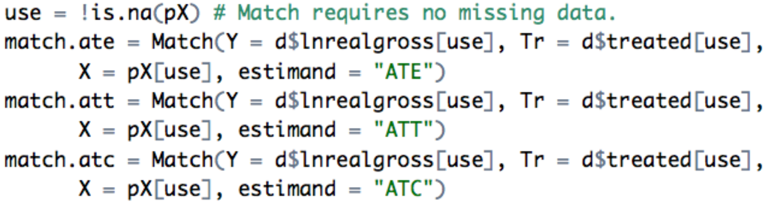
Subclass	#units	#MPs	#non-MPs	DIGM
1	43	4	39	-0.24
2	43	13	30	0.01
3	43	13	30	0.9
4	42	12	30	0.33
5	43	18	25	0.24
6	43	15	28	-0.13
7	42	15	27	0.35
8	43	20	23	0.5
9	43	26	17	0.47
10	40	29	11	1.69

ATE: 0.40
ATT: 0.57
ATC: 0.30

How many sub-classes? **Bias-variance tradeoff.**

Nearest-neighbor matching on the propensity score

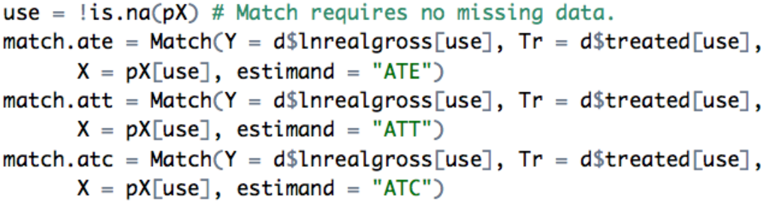
Using defaults in `Matching`:



```
use = !is.na(pX) # Match requires no missing data.
match.ate = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATE")
match.att = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATT")
match.atc = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATC")
```

Nearest-neighbor matching on the propensity score

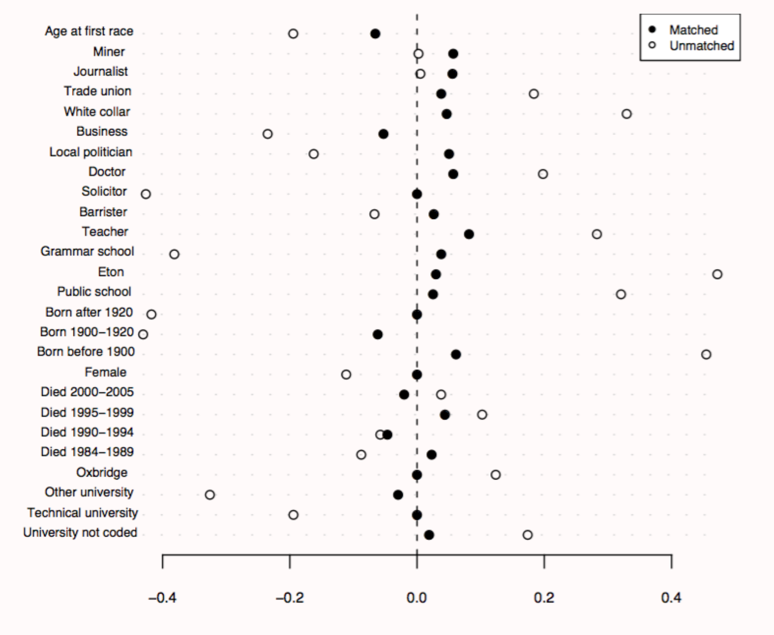
Using defaults in `Matching`:



```
use = !is.na(pX) # Match requires no missing data.
match.ate = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATE")
match.att = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATT")
match.atc = Match(Y = d$lnrealgross[use], Tr = d$treated[use],
      X = pX[use], estimand = "ATC")
```

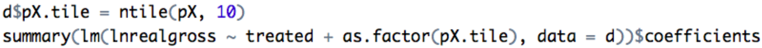
ATE: 0.42
ATT: 0.50
ATC: 0.38

Balance statistics for matching



Regression controlling for the propensity score

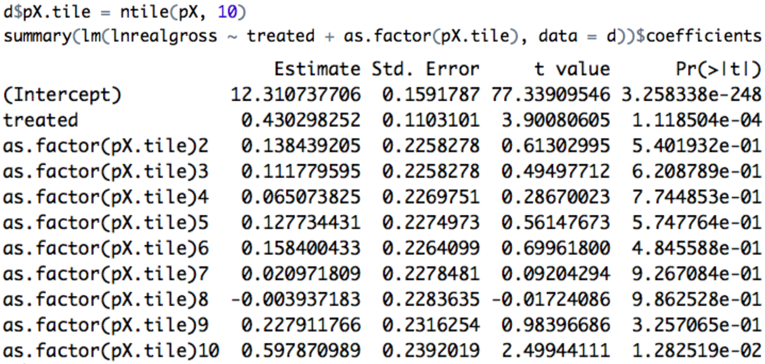
Controlling via dummies for 10 sub-classes:



```
d$pX.tile = ntile(pX, 10)
summary(lm(lnrealgross ~ treated + as.factor(pX.tile), data = d))$coefficients
```

Regression controlling for the propensity score

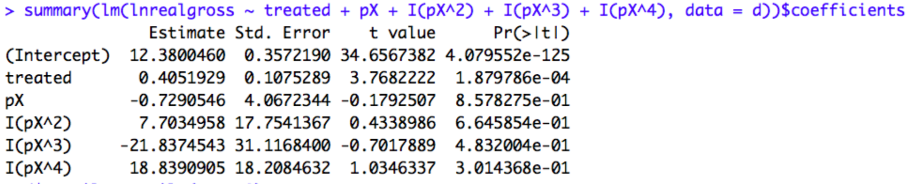
Controlling via dummies for 10 sub-classes:



```
d$pX.tile = ntile(pX, 10)
summary(lm(lnrealgross ~ treated + as.factor(pX.tile), data = d))$coefficients

                       Estimate Std. Error     t value      Pr(>|t|)
(Intercept)         12.310737706  0.1591787 77.33909546 3.258338e-248
treated              0.430298252  0.1103101  3.90080605  1.118504e-04
as.factor(pX.tile)2  0.138439205  0.2258278  0.61302995  5.401932e-01
as.factor(pX.tile)3  0.111779595  0.2258278  0.49497712  6.208789e-01
as.factor(pX.tile)4  0.065073825  0.2269751  0.28670023  7.744853e-01
as.factor(pX.tile)5  0.127734431  0.2274973  0.56147673  5.747764e-01
as.factor(pX.tile)6  0.158400433  0.2264099  0.69961800  4.845588e-01
as.factor(pX.tile)7  0.020971809  0.2278481  0.09204294  9.267084e-01
as.factor(pX.tile)8 -0.003937183  0.2283635 -0.01724086  9.862528e-01
as.factor(pX.tile)9  0.227911766  0.2316254  0.98396686  3.257065e-01
as.factor(pX.tile)10 0.597870989  0.2392019  2.49944111  1.282519e-02
```

Regression controlling for the propensity score (2)

Controlling via polynomials of propensity score:



```
> summary(lm(lnrealgross ~ treated + pX + I(pX^2) + I(pX^3) + I(pX^4), data = d))$coefficients
               Estimate Std. Error    t value      Pr(>|t|)
(Intercept)  12.3800460  0.3572190 34.6567382 4.079552e-125
treated       0.4051929  0.1075289  3.7682222  1.879786e-04
pX           -0.7290546  4.0672344 -0.1792507  8.578275e-01
I(pX^2)       7.7034958 17.7541367  0.4338986  6.645854e-01
I(pX^3)     -21.8374543 31.1168400 -0.7017889  4.832004e-01
I(pX^4)      18.8390905 18.2084632  1.0346337  3.014368e-01
```

Comparison of techniques (propensity score version)

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.57	.30
Matching	.42	.50	.38
Regression: bins	.41	-	-
Regression: polynomials	.43	-	-

Comparison of techniques (propensity score version)

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.57	.30
Matching	.42	.50	.38
Regression: bins	.41	-	-
Regression: polynomials	.43	-	-

To note:

- Propensity score is estimated, which should be considered in variance

Comparison of techniques (propensity score version)

Comparison of estimates

	ATE	ATT	ATC
Sub-classification	.40	.57	.30
Matching	.42	.50	.38
Regression: bins	.41	-	-
Regression: polynomials	.43	-	-

To note:

- Propensity score is estimated, which should be considered in variance
- Our model for the propensity score has no interactions – stronger assumptions than previous exercise

Introduction to covariate adjustment

Covariate adjustment based on categorical variables

Covariate adjustment using the propensity score

Covariate adjustment in sparse data without the propensity score

Two important facts about regression

Sub-classification without the propensity score

To avoid empty cells, you might try playing around with how cells are defined.

Sub-classification without the propensity score

To avoid empty cells, you might try playing around with how cells are defined.

This is basically how I think about **coarsened exact matching (CEM)** by King et al.

Nearest-neighbor matching with the covariates

We can do nearest-neighbor matching with the covariates themselves, rather than propensity score.

Nearest-neighbor matching with the covariates

We can do nearest-neighbor matching with the covariates themselves, rather than propensity score.

But which units are “near” each other?

Nearest-neighbor matching with the covariates

We can do nearest-neighbor matching with the covariates themselves, rather than propensity score.

But which units are “near” each other?

e.g. Should I match a Tory born in 1928 who went to Oxford to a

- Tory born in 1927 with no university listed?
- Labour candidate born in 1928 who went to the LSE?
- Tory born in 1942 who went to Oxford?

Nearest-neighbor matching with the covariates

We can do nearest-neighbor matching with the covariates themselves, rather than propensity score.

But which units are “near” each other?

e.g. Should I match a Tory born in 1928 who went to Oxford to a

- Tory born in 1927 with no university listed?
- Labour candidate born in 1928 who went to the LSE?
- Tory born in 1942 who went to Oxford?

Some of the options:

- scale distance on each variable by inverse of the variable’s sample variance (default in `Matching` when not exact)
- scale distance by the inverse of the covariance matrix (Mahalanobis distance)
- genetic matching: search for a weight matrix that yields overall covariate balance (Diamond and Sekhon)

Regression

Adding control variables to a regression model is very straightforward.

Regression

Adding control variables to a regression model is very straightforward.

With sparser data, can no longer use saturated models; rely more on linearity, additivity (as with the estimation of the propensity score).

Regression

Adding control variables to a regression model is very straightforward.

With sparser data, can no longer use saturated models; rely more on linearity, additivity (as with the estimation of the propensity score).

These two approaches should give you very similar results:

Regression

Adding control variables to a regression model is very straightforward.

With sparser data, can no longer use saturated models; rely more on linearity, additivity (as with the estimation of the propensity score).

These two approaches should give you very similar results:

1. Estimate propensity score $p(X_i)$ based on model:

$$D_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \ldots + \alpha_K X_{iK}$$

   Regress Y_i on D_i and a flexible function of the propensity score.

Regression

Adding control variables to a regression model is very straightforward.

With sparser data, can no longer use saturated models; rely more on linearity, additivity (as with the estimation of the propensity score).

These two approaches should give you very similar results:

1. Estimate propensity score $p(X_i)$ based on model:

   $$D_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \ldots + \alpha_K X_{iK}$$

   Regress Y_i on D_i and a flexible function of the propensity score.
2. Regress Y_i on D_i and covariates X_{i1} to X_{iK}.

Regression: propensity score vs. covariates

In the “MPs for Sale” example:

Approach	**ATE**
Regress Y_i on D_i and 10 bins of propensity score	.43
Regress Y_i on D_i and 4 polynomials of propensity score	.41
Regress Y_i on D_i and covariates from propensity score model	.41

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
  - Link to CIA is transparent: units with same/similar values of X_i are comparable

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
  - Link to CIA is transparent: units with same/similar values of X_i are comparable
  - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
    - Link to CIA is transparent: units with same/similar values of X_i are comparable
    - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging
- But regression should be your default tool

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
  - Link to CIA is transparent: units with same/similar values of X_i are comparable
  - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging
- But regression should be your default tool:
  - Saturated regression replicates sub-classification and matching on categorical variables (up to reweighting)

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
  - Link to CIA is transparent: units with same/similar values of X_i are comparable
  - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging
- But regression should be your default tool:
  - Saturated regression replicates sub-classification and matching on categorical variables (up to reweighting)
  - If estimating propensity score, can replicate sub-classification (up to reweighting)

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
    - Link to CIA is transparent: units with same/similar values of X_i are comparable
    - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging
- But regression should be your default tool:
    - Saturated regression replicates sub-classification and matching on categorical variables (up to reweighting)
    - If estimating propensity score, can replicate sub-classification (up to reweighting)
    - Assuming linear relationship between X_i and Y_i can be useful; with bins and polynomials (and splines and GAMs . . .) can be as flexible as we want

Comparing methods of covariate adjustment: bottom line

My view:

- Sub-classification and matching are useful for developing understanding
    - Link to CIA is transparent: units with same/similar values of X_i are comparable
    - Math is simple: basically just grouping, calculating DIGMs, weighting and averaging
- But regression should be your default tool:
    - Saturated regression replicates sub-classification and matching on categorical variables (up to reweighting)
    - If estimating propensity score, can replicate sub-classification (up to reweighting)
    - Assuming linear relationship between X_i and Y_i can be useful; with bins and polynomials (and splines and GAMs . . .) can be as flexible as we want
    - Statistical inference straightforward

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Rather than spending weeks/months/years on mastering matching, sub-classification, MLE, Bayesian models, hierarchical models, etc.

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Rather than spending weeks/months/years on mastering matching, sub-classification, MLE, Bayesian models, hierarchical models, etc., you should

- collect better covariates relevant to your question

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Rather than spending weeks/months/years on mastering matching, sub-classification, MLE, Bayesian models, hierarchical models, etc., you should

- collect better covariates relevant to your question,
- find questions/settings for which there are very good covariates (RDD as extreme example)

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Rather than spending weeks/months/years on mastering matching, sub-classification, MLE, Bayesian models, hierarchical models, etc., you should

- collect better covariates relevant to your question,
- find questions/settings for which there are very good covariates (RDD as extreme example) , and
- look for questions/settings where you don’t need such good covariates (e.g. randomized experiment, natural experiment, IV, diff-in-diff)

Covariate adjustment: bottom line

To make the “selection on observables” assumption credible, you need **good observables**, i.e. good measures of characteristics that affect the outcome and differ between treatment and control groups.

Rather than spending weeks/months/years on mastering matching, sub-classification, MLE, Bayesian models, hierarchical models, etc., you should

- collect better covariates relevant to your question,
- find questions/settings for which there are very good covariates (RDD as extreme example) , and
- look for questions/settings where you don’t need such good covariates (e.g. randomized experiment, natural experiment, IV, diff-in-diff)

Keep the statistics simple, focus on the data, and be opportunistic.

Introduction to covariate adjustment

Covariate adjustment based on categorical variables

Covariate adjustment using the propensity score

Covariate adjustment in sparse data without the propensity score

Two important facts about regression

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}.

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}. **Definitions** (**MM** pg. 93):

The **long regression** includes both:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}. **Definitions** (**MM** pg. 93):

The **long regression** includes both:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

The **short regression** includes only X_{1i}:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s.$$

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}. **Definitions** (**MM** pg. 93):

The **long regression** includes both:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

The **short regression** includes only X_{1i}:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s.$$

The **auxiliary regression** (my term) describes the relationship between the two covariates:

$$X_{2i} = \alpha^a + \pi_{21} X_{1i} + e_i^a.$$

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}. **Definitions** (**MM** pg. 93):

The **long regression** includes both:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

The **short regression** includes only X_{1i}:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s.$$

The **auxiliary regression** (my term) describes the relationship between the two covariates:

$$X_{2i} = \alpha^a + \pi_{21} X_{1i} + e_i^a.$$

Then the **omitted variables bias formula** tells us

$$\beta^s = \beta^l + \pi_{21}\gamma.$$

Omitted variable bias formula

Suppose two covariates, X_{1i} and X_{2i}. **Definitions** (**MM** pg. 93):

The **long regression** includes both:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

The **short regression** includes only X_{1i}:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s.$$

The **auxiliary regression** (my term) describes the relationship between the two covariates:

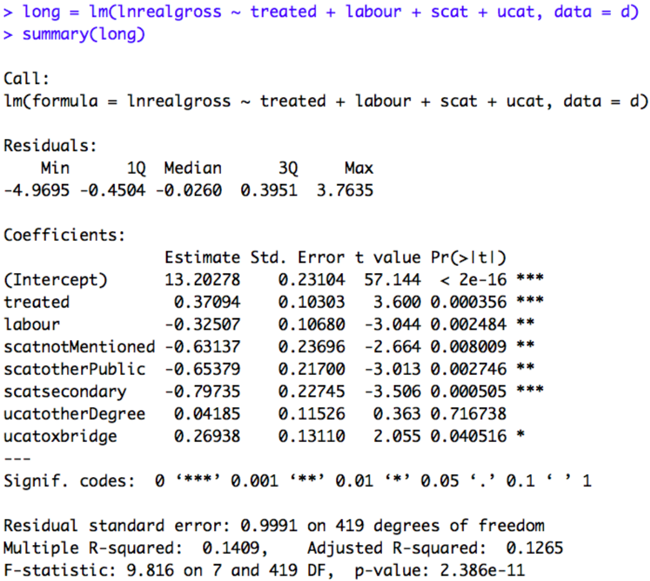
$$X_{2i} = \alpha^a + \pi_{21} X_{1i} + e_i^a.$$

Then the **omitted variables bias formula** tells us

$$\beta^s = \beta^l + \pi_{21}\gamma.$$

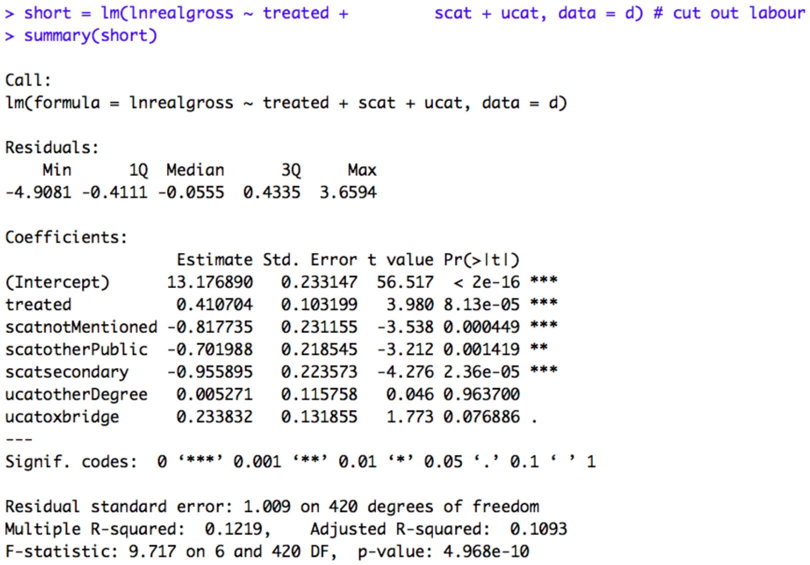
Explain the last line in plain English.

OVB formula example: long regression



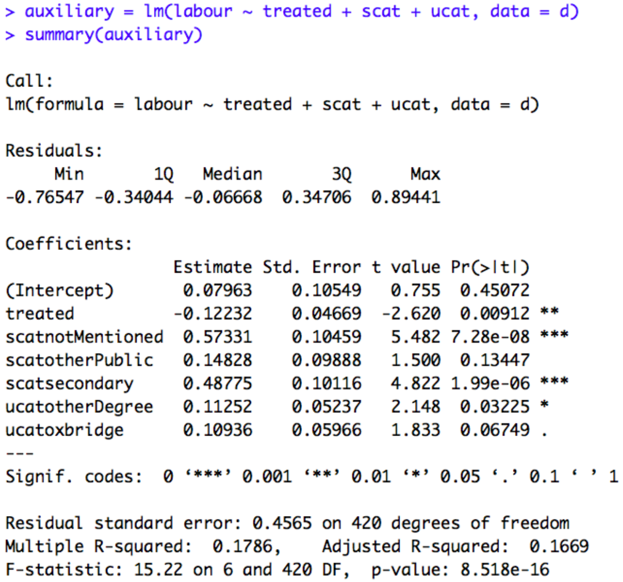
```
> long = lm(lnrealgross ~ treated + labour + scat + ucat, data = d)
> summary(long)

Call:
lm(formula = lnrealgross ~ treated + labour + scat + ucat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9695 -0.4504 -0.0260  0.3951  3.7635 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     13.20278    0.23104  57.144  < 2e-16 ***
treated          0.37094    0.10303   3.600 0.000356 ***
labour          -0.32507    0.10680  -3.044 0.002484 ** 
scatnotMentioned -0.63137    0.23696  -2.664 0.008009 ** 
scatotherPublic -0.65379    0.21700  -3.013 0.002746 ** 
scatsecondary   -0.79735    0.22745  -3.506 0.000505 ***
ucatotherDegree  0.04185    0.11526   0.363 0.716738    
ucatoxbridge     0.26938    0.13110   2.055 0.040516 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9991 on 419 degrees of freedom
Multiple R-squared:  0.1409,	Adjusted R-squared:  0.1265 
F-statistic: 9.816 on 7 and 419 DF,  p-value: 2.386e-11
```

OVB formula example: short regression



```
> short = lm(lnrealgross ~ treated +         scat + ucat, data = d) # cut out labour
> summary(short)

Call:
lm(formula = lnrealgross ~ treated + scat + ucat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9081 -0.4111 -0.0555  0.4335  3.6594 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.176890   0.233147  56.517  < 2e-16 ***
treated           0.410704   0.103199   3.980 8.13e-05 ***
scatnotMentioned -0.817735   0.231155  -3.538 0.000449 ***
scatotherPublic  -0.701988   0.218545  -3.212 0.001419 ** 
scatsecondary    -0.955895   0.223573  -4.276 2.36e-05 ***
ucatotherDegree   0.005271   0.115758   0.046 0.963700    
ucatoxbridge      0.233832   0.131855   1.773 0.076886 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.009 on 420 degrees of freedom
Multiple R-squared:  0.1219,	Adjusted R-squared:  0.1093 
F-statistic: 9.717 on 6 and 420 DF,  p-value: 4.968e-10
```

OVB formula example: auxiliary regression



```
> auxiliary = lm(labour ~ treated + scat + ucat, data = d)
> summary(auxiliary)

Call:
lm(formula = labour ~ treated + scat + ucat, data = d)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76547 -0.34044 -0.06668  0.34706  0.89441 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.07963    0.10549   0.755  0.45072    
treated          -0.12232    0.04669  -2.620  0.00912 ** 
scatnotMentioned  0.57331    0.10459   5.482 7.28e-08 ***
scatotherPublic   0.14828    0.09888   1.500  0.13447    
scatsecondary     0.48775    0.10116   4.822 1.99e-06 ***
ucatotherDegree   0.11252    0.05237   2.148  0.03225 *  
ucatoxbridge      0.10936    0.05966   1.833  0.06749 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4565 on 420 degrees of freedom
Multiple R-squared:  0.1786,	Adjusted R-squared:  0.1669 
F-statistic: 15.22 on 6 and 420 DF,  p-value: 8.518e-16
```

OVB formula example: confirming equality

So is it true that

$$\beta^s = \beta^l + \pi_{21}\gamma.$$

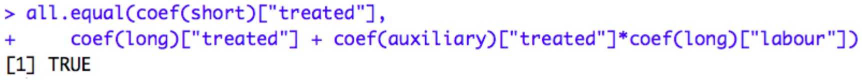
i.e. “short equals long plus the effect of omitted times the regression of omitted on included”?

OVB formula example: confirming equality

So is it true that

$$\beta^s = \beta^l + \pi_{21}\gamma.$$

i.e. “short equals long plus the effect of omitted times the regression of omitted on included”?



```
> all.equal(coef(short)["treated"],
+     coef(long)["treated"] + coef(auxiliary)["treated"]*coef(long)["labour"])
[1] TRUE
```

Yes.

Lessons from the OVB formula

Omitting a variable causes bias in our estimate of ATE if and only if

- it is related to the treatment, conditional on other covariates, and
- it is related to the outcome, conditional on other covariates.

Lessons from the OVB formula

Omitting a variable causes bias in our estimate of ATE if and only if

- it is related to the treatment, conditional on other covariates, and
- it is related to the outcome, conditional on other covariates.

This is why

- you don’t have to control for anything in a randomized experiment
- you don’t have to control for everything you can think of that affects Y_i - only variables related to D_i (and Y_i) conditional on other covariates
- you don’t have to control for anything other than the running variable in an RDD

“Regression anatomy”

The coefficient β^l in the **long** regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

"Regression anatomy"

The coefficient β^l in the **long** regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

can be calculated by performing the **reverse auxiliary regression** (my term)

$$X_{1i} = \alpha^a + \pi_{12} X_{2i} + e_i^a,$$

“Regression anatomy”

The coefficient β^l in the **long** regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

can be calculated by performing the **reverse auxiliary regression** (my term)

$$X_{1i} = \alpha^a + \pi_{12} X_{2i} + e_i^a,$$

getting the residuals,

$$\tilde{X}_{1i} = X_{1i} - \left(\hat{\alpha}^a + \hat{\pi}_{12} X_{2i}\right),$$

“Regression anatomy”

The coefficient β^l in the **long** regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

can be calculated by performing the **reverse auxiliary regression** (my term)

$$X_{1i} = \alpha^a + \pi_{12} X_{2i} + e_i^a,$$

getting the residuals,

$$\tilde{X}_{1i} = X_{1i} - \left(\hat{\alpha}^a + \hat{\pi}_{12} X_{2i}\right),$$

and regressing Y_i on those (“outcome-on-residuals” regression):

$$Y_i = \alpha^* + \beta^* \tilde{X}_{1i} + e_i^*.$$

“Regression anatomy”

The coefficient β^l in the **long** regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

can be calculated by performing the **reverse auxiliary regression** (my term)

$$X_{1i} = \alpha^a + \pi_{12} X_{2i} + e_i^a,$$

getting the residuals,

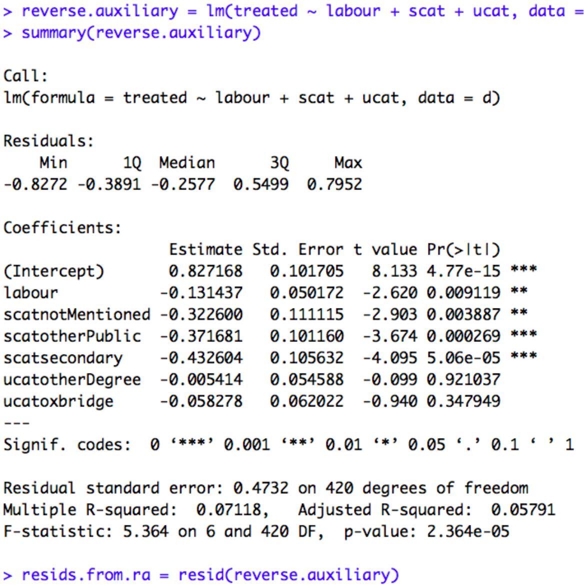
$$\tilde{X}_{1i} = X_{1i} - \left(\hat{\alpha}^a + \hat{\pi}_{12} X_{2i}\right),$$

and regressing Y_i on those (“outcome-on-residuals” regression):

$$Y_i = \alpha^* + \beta^* \tilde{X}_{1i} + e_i^*.$$

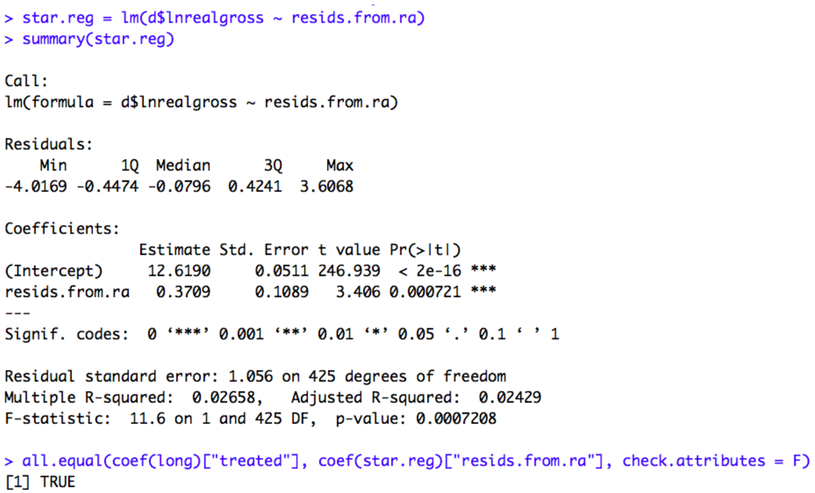
i.e., $\beta^* = \beta^l$.

Regression anatomy: “reverse auxiliary” regression



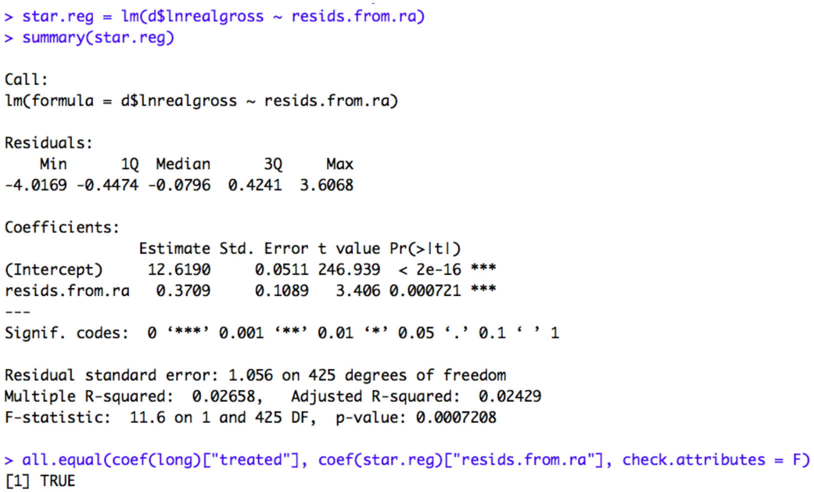
```
> reverse.auxiliary = lm(treated ~ labour + scat + ucat, data = d)
> summary(reverse.auxiliary)

Call:
lm(formula = treated ~ labour + scat + ucat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8272 -0.3891 -0.2577  0.5499  0.7952 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.827168   0.101705   8.133 4.77e-15 ***
labour           -0.131437   0.050172  -2.620 0.009119 ** 
scatnotMentioned -0.322600   0.111115  -2.903 0.003887 ** 
scatotherPublic  -0.371681   0.101160  -3.674 0.000269 ***
scatsecondary    -0.432604   0.105632  -4.095 5.06e-05 ***
ucatotherDegree  -0.005414   0.054588  -0.099 0.921037    
ucatoxbridge     -0.058278   0.062022  -0.940 0.347949    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4732 on 420 degrees of freedom
Multiple R-squared:  0.07118,	Adjusted R-squared:  0.05791 
F-statistic: 5.364 on 6 and 420 DF,  p-value: 2.364e-05

> resids.from.ra = resid(reverse.auxiliary)
```

Regression anatomy: “outcome-on-residuals” regression



```
> star.reg = lm(d$lnrealgross ~ resids.from.ra)
> summary(star.reg)

Call:
lm(formula = d$lnrealgross ~ resids.from.ra)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0169 -0.4474 -0.0796  0.4241  3.6068 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     12.6190     0.0511 246.939  < 2e-16 ***
resids.from.ra   0.3709     0.1089   3.406 0.000721 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 425 degrees of freedom
Multiple R-squared:  0.02658,	Adjusted R-squared:  0.02429 
F-statistic:  11.6 on 1 and 425 DF,  p-value: 0.0007208

> all.equal(coef(long)["treated"], coef(star.reg)["resids.from.ra"], check.attributes = F)
[1] TRUE
```

Regression anatomy: “outcome-on-residuals” regression



```
> star.reg = lm(d$lnrealgross ~ resids.from.ra)
> summary(star.reg)

Call:
lm(formula = d$lnrealgross ~ resids.from.ra)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0169 -0.4474 -0.0796  0.4241  3.6068 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     12.6190     0.0511 246.939  < 2e-16 ***
resids.from.ra   0.3709     0.1089   3.406 0.000721 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 425 degrees of freedom
Multiple R-squared:  0.02658,	Adjusted R-squared:  0.02429 
F-statistic:  11.6 on 1 and 425 DF,  p-value: 0.0007208

> all.equal(coef(long)["treated"], coef(star.reg)["resids.from.ra"], check.attributes = F)
[1] TRUE
```

It works!

Lessons from regression anatomy

The OLS coefficient on D_i measures the relationship between Y_i and the part of D_i not “explained” by X_i.

Lessons from regression anatomy

The OLS coefficient on D_i measures the relationship between Y_i and the part of D_i not “explained” by X_i.

What is the CIA in any regression that claims to measure the effect of D_i on Y_i?

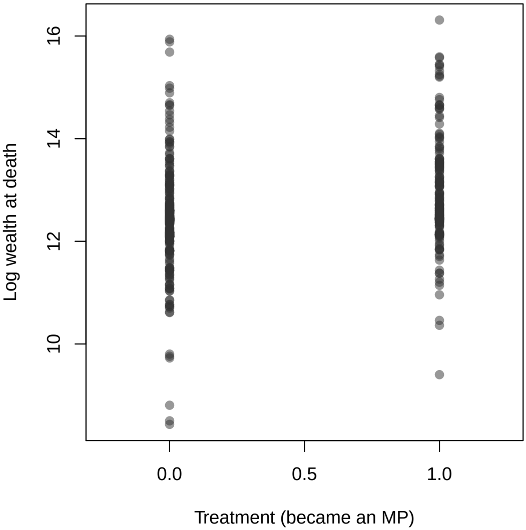
Lessons from regression anatomy

The OLS coefficient on D_i measures the relationship between Y_i and the part of D_i not "explained" by X_i.

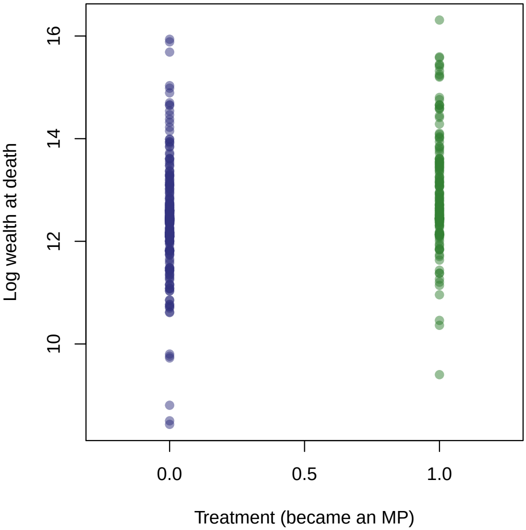
What is the CIA in any regression that claims to measure the effect of D_i on Y_i?

The part of D_i not "explained" by X_i (the residual from the "reverse auxiliary regression") is not related to the potential outcomes, i.e. as-if random.

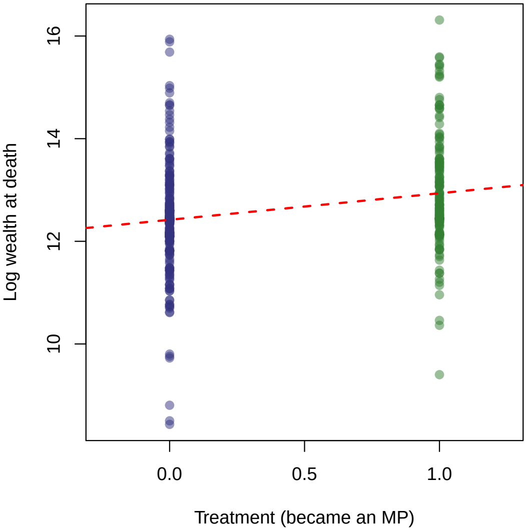
Regression CIA: illustration



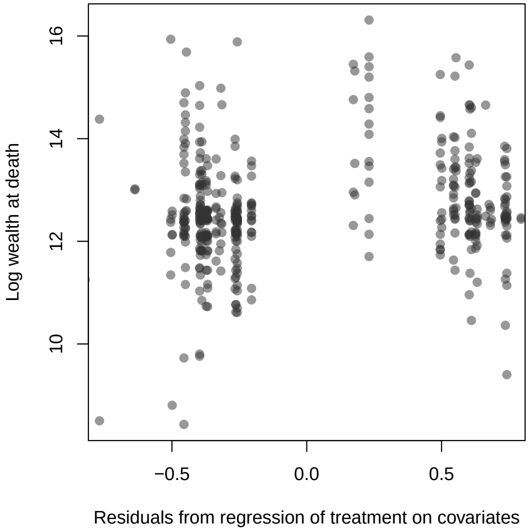
Regression CIA: illustration



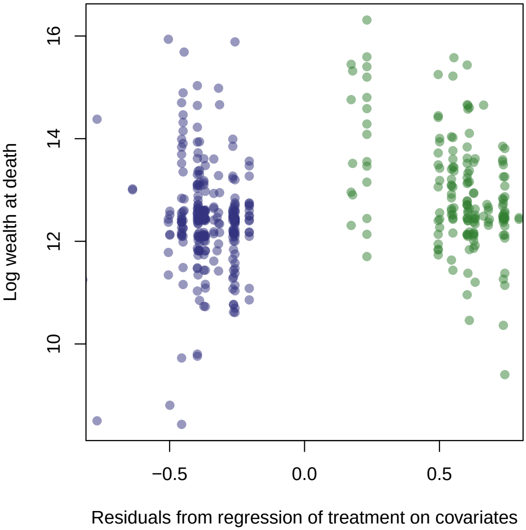
Regression CIA: illustration



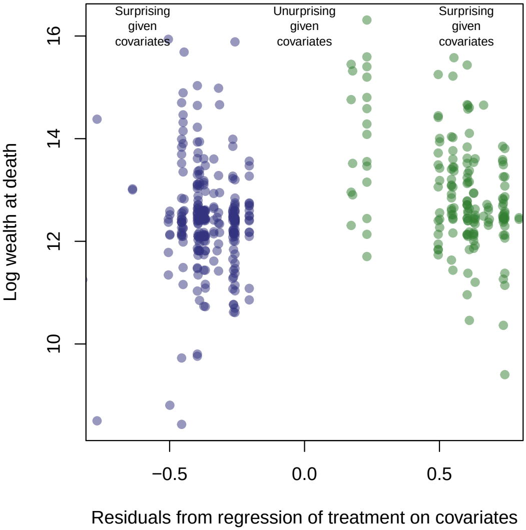
Regression CIA: illustration



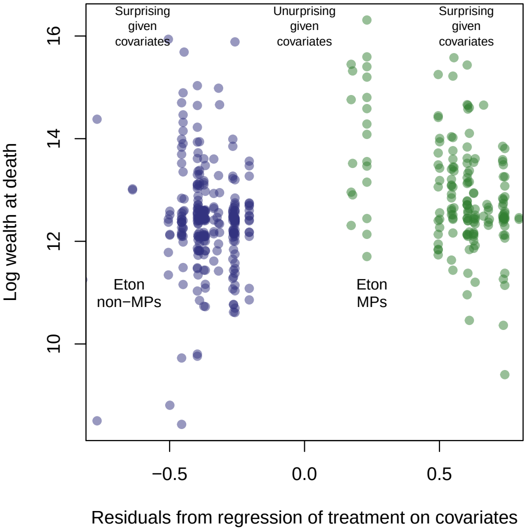
Regression CIA: illustration



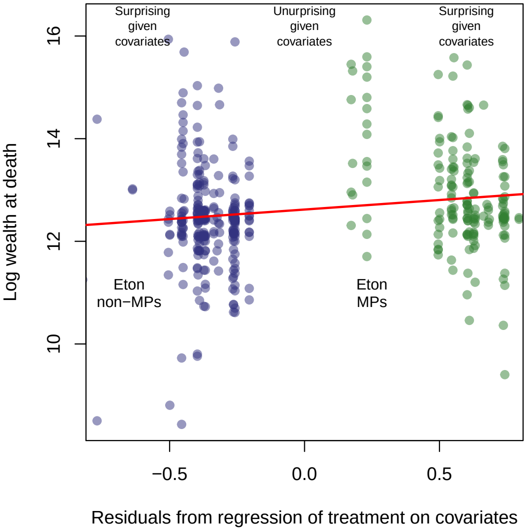
Regression CIA: illustration

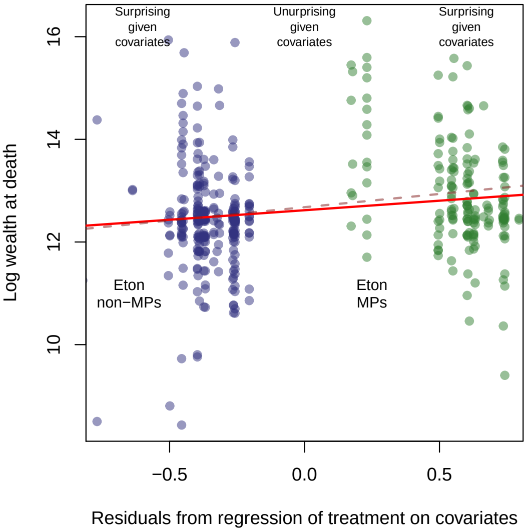


Regression CIA: illustration



Regression CIA: illustration




Surprising
given
covariates
Unurprising
given
covariates
Surprising
given
covariates
Eton
non−MPs
Eton
MPs
Log wealth at death
10
12
14
16
−0.5
0.0
0.5
Residuals from regression of treatment on covariates