

Week 10: Causality with Measured Confounding

Brandon Stewart¹

Princeton

November 12–14, 2024

¹These slides are heavily influenced by Matt Blackwell, Jens Hainmueller, Erin Hartman, Kosuke Imai, Gary King, Ian Lundberg, Richard Nielsen, and Matt Salganik.

Where We've Been and Where We're Going...

- Last Week

- ▶ diagnostics

- This Week

- ▶ review to this point
- ▶ parametric g methods
- ▶ matching as nonparametric preprocessing
- ▶ fixed effects
- ▶ sensitivity analysis

- Next Week

- ▶ approaches with unmeasured confounding

- Long Run

- ▶ causal frameworks → inference → regression → causal inference

Administrative Notes

- Summary of the Rest of Class:

- ▶ Nov 12–14: Measured Confounding (Week 10), Pset 9 due 11/14, Precept 11/15
- ▶ Nov 19–21: Unmeasured Confounding (Week 11), Pset 10 due 11/21, Precept 11/22
- ▶ Nov 26: No class or precept (although technically it follows a Friday schedule)
- ▶ Dec 3–5: Diff-in-Diff + Wrap-Up (Week 12), Pset 11 due 12/5, Precept 12/6
- ▶ For Dec 3: Read Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86.3 (2021): 532-565.
- ▶ Dec 6: Review Session during "Normal" Precept Time
- ▶ Dec 14?–19: Final Exam Period (Discussion)

- 1 Regression with Measured Confounding
 - Regression with Constant Effects
 - Additive Regression with Heterogeneous Effects
 - Imputation Estimators

- 2 Matching
 - Fundamentals of Matching
 - Two Approaches to Matching

- 3 Fixed Effects
 - Non-parametric Identification and Fixed Effects

From Stratification to Regression

- Nonparametric identification yielded this estimand:

$$E[Y_i(1) - Y_i(0)] = E\left[E[Y_i \mid T_i = 1, \mathbf{X}_i] - E[Y_i \mid T_i = 0, \mathbf{X}_i]\right]$$

- What if we estimated the inner conditional expectations with a **regression** model like we have been doing all semester?
- We will sometimes write the empirical quantity of interest as

$$\mu_t(\mathbf{x}) = E[Y_i(t) \mid T = t, \mathbf{X}_i = \mathbf{x}]$$

- Ignorability tell us that the **counterfactual** quantities can be written in terms of these quantities based on **observed** data, but we still need to estimate those conditional expectations!

The parametric g-formula: Simple case

$$X \rightarrow T \rightarrow Y$$


Nonparametric identification $E[Y(t)] = E[E[Y \mid X, T = t]]$

Parametric model for $\mu(\mathbf{x})$ $\alpha + \gamma X + \beta T$

Regression estimator for the ATE:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)$$

Estimator in words:

- 1 Estimate the regression model(s)
- 2 Change all treatment values to 1 and 0
- 3 Predict for everyone
- 4 Take the sample mean of the difference.

Connection to $\hat{\beta}$ for Additive Regression

Estimator for the effect $\tau = E[Y(1)] - E[Y(0)]$:

$$\begin{aligned}\hat{\tau} &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 1 \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 0 \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\beta} \\ &= \hat{\beta}\end{aligned}$$

With additive regression, the parametric g-formula estimator simplifies to $\hat{\beta}$.

The parametric g-formula is more general

Suppose we add the $T \times \mathbf{X}$ interaction to the model.

$$E(Y \mid T, \mathbf{X}) = \alpha + \gamma \mathbf{X} + \beta T + \eta T \mathbf{X}$$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{\tau} &= \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 1 + \hat{\eta} \times 1 \times \mathbf{X}_i \right) \right) \\ &\quad - \left(\frac{1}{n} \sum_{i=1}^n \left(\hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 0 + \hat{\eta} \times 0 \times \mathbf{X}_i \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\hat{\beta} + \hat{\eta} \mathbf{X}_i \right)\end{aligned}$$

This no longer collapses to a coefficient!

The parametric g-formula: Word of warning

The parametric g-formula does not replace any assumptions.

You need:

Identification Assumptions (causal)

Consistency $Y_i = Y_i(T_i)$

Ignorability $T \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid \mathbf{X}$

Positivity $P(T = t \mid \mathbf{X} = \mathbf{x}) > 0$

Estimation Assumptions (statistical)

Estimation $E(Y \mid T, \mathbf{X})$ follows the assumed model

To use a parametric model, we **added** the bottom assumption.

- Estimation assumptions have implications in the data and thus can be somewhat checked
- Pragmatically you need **common support** (i.e. treated data where control units are and vice-versa).

Let's consider the implications of estimation with regression under different assumptions about the model.

Identification under Selection on Observables: Regression

Consider the estimation model: $Y_i = \beta_0 + \tau T_i + X_i' \beta + u_i$.

Given **selection on observables**, there are mainly three identification scenarios:

- ① Constant treatment effects and outcomes are linear in X
 - ▶ τ will provide unbiased and consistent estimates of ATE.
- ② Constant treatment effects and unknown functional form
 - ▶ τ will provide well-defined linear approximation to the average causal response function $E[Y|T=1, X] - E[Y|T=0, X]$. Approximation may be very poor if $E[Y|T, X]$ is misspecified and then τ may be biased for the ATE.
- ③ Heterogeneous treatment effects (τ differs for different values of X)
 - ▶ even If outcomes are linear in X , τ converges to the conditional-variance-weighted average of the underlying causal effects rather than the ATE.

Ideal Case: Constant Effects Model With Linearly Separable Confounding

Assume a data generating process with **constant effects** and **linearly separable confounding**:

$$Y_i(t) = Y_i = \beta_0 + \tau T_i + \eta_i$$

- **Linearly separable confounding:** assume that $E[\eta_i|X_i] = X_i'\beta$, which means that $\eta_i = X_i'\beta + u_i$ where $E[u_i|X_i] = 0$.

$$\begin{aligned} Y_i &= \beta_0 + \tau T_i + \eta_i \\ &= \beta_0 + \tau T_i + X_i'\beta + u_i \end{aligned}$$

- Thus, a regression where T_i and X_i are entered linearly can recover the ATE. (The regression model matches the data generating process)

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned}Y_i &= T_i Y_i(1) + (1 - T_i) Y_i(0) \\&= Y_i(0) + (Y_i(1) - Y_i(0)) T_i \\&= \mu_0 + \tau_i T_i + (Y_i(0) - \mu_0) \\&= \mu_0 + \tau T_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot T_i \\&= \mu_0 + \tau T_i + \varepsilon_i\end{aligned}$$

- Error term now includes two components:
 - ① “Baseline” variation in the outcome: $(Y_i(0) - \mu_0)$
 - ② Variation in the treatment effect, $(\tau_i - \tau)$
- We can verify that under experiment, $E[\varepsilon_i | T_i] = 0$
- Thus, OLS estimates the ATE with no covariates.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.
- ATE is the weighted sum of Conditional Average Treatment Effects (CATEs):

$$\tau = \sum_x \tau(x) P[X_i = x]$$

- ATE/ATT are weighted averages of CATEs.
- What about the regression estimand, τ_R ? How does it relate to the ATE/ATT?

Heterogeneous effects and regression

- Let's investigate this under a saturated regression model:

$$Y_i = \sum_x B_{xi} \alpha_x + \tau_R T_i + e_i.$$

- Use a dummy variable for each unique combination of X_i :
 $B_{xi} = \mathbb{I}(X_i = x)$
- Linear in X_i by construction!

Investigating the regression coefficient

- How can we investigate τ_R ? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, T_i - E[T_i|X_i])}{\text{Var}(T_i - E[T_i|X_i])}$$

- $T_i - E[T_i|X_i]$ is the residual from a regression of T_i on the full set of dummies.
- With a little work we can show:

$$\tau_R = \frac{E[\tau(X_i)(T_i - E[T_i|X_i])^2]}{E[(T_i - E[T_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_t^2(X_i)]}{E[\sigma_t^2(X_i)]}$$

- $\sigma_t^2(x) = \text{Var}[T_i|X_i = x]$ is the conditional variance of treatment assignment.

ATE versus OLS

$$\tau_R = E[\tau(X_i)W_i] = \sum_x \tau(x) \frac{\sigma_t^2(x)}{E[\sigma_t^2(X_i)]} P[X_i = x]$$

- Compare to the ATE:

$$\tau = E[\tau(X_i)] = \sum_x \tau(x) P[X_i = x]$$

- Both weight strata relative to their size ($P[X_i = x]$)
- OLS weights strata higher if the treatment variance in those strata ($\sigma_t^2(x)$) is higher in those strata relative to the average variance across strata ($E[\sigma_t^2(X_i)]$).
- The ATE weights only by their size.

Regression weighting

$$W_i = \frac{\sigma_t^2(X_i)}{E[\sigma_t^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a **minimum-variance estimator** $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If T_i is binary, then we know the conditional variance will be:

$$\sigma_t^2(x) = P[T_i = 1|X_i = x] (1 - P[T_i = 1|X_i = x])$$

- Maximum variance with $P[T_i = 1|X_i = x] = 1/2$.

OLS weighting example

- Binary covariate:

Group 1	Group 2
$P[X_i = 1] = 0.75$	$P[X_i = 0] = 0.25$
$P[T_i = 1 X_i = 1] = 0.9$	$P[T_i = 1 X_i = 0] = 0.5$
$\sigma_t^2(1) = 0.09$	$\sigma_t^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $E[\sigma_t^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$\begin{aligned}\tau_R &= E[\tau(X_i)W_i] \\ &= \tau(1)W(1)P[X_i = 1] + \tau(0)W(0)P[X_i = 0] \\ &= 1 \times 0.692 \times 0.75 + -1 \times 1.92 \times 0.25 \\ &= 0.039\end{aligned}$$

When will OLS estimate the ATE?

- When does $\tau = \tau_R$? (where τ_R is the estimate from a regression)
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = P[T_i = 1|X_i = x] = e$.
 - ▶ Implies that the OLS weights are 1.
- Incorrect linearity assumption in X_i will lead to more bias.

Implications

- Knowing the weights gives us **substantive insight**
- When some observations have no weight, this means that the covariates **completely** explain their treatment condition.
- This is a **feature**, not a **bug**, of regression: we can't learn anything from those cases anyway (i.e. it is automatically handling issues of **common support**).
- The downside is that we have to be aware of what happened!

Example Replication from Aronow and Samii (2015)

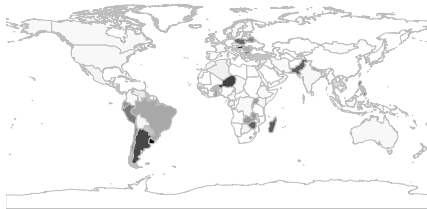
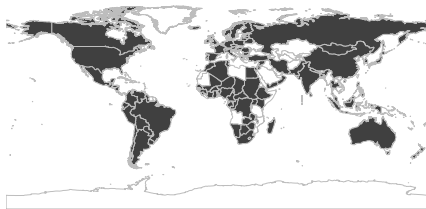
Jensen (2003), “Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment.”

Jensen presents a large- N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

The nominal sample: 114 countries from 1970 to 1997.

Jensen estimates that a 1 unit increase in polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of GDP ($p < 0.001$).

Nominal and Effective Samples



Over 50% of the weight goes to just 12 (out of 114) countries.

Broader Implications

Aronow and Samii argue that analyses which use regression, **even with a representative sample**, have no greater claim to external validity than do [natural] experiments.

- When a treatment is “as-if” randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.
- The **effective sample** (upon which causal effects are estimated) may have radically different properties than the nominal sample.
- When there is an underlying natural experiment in the data, a properly specified regression model may reproduce the internally valid estimate associated with the natural experiment.

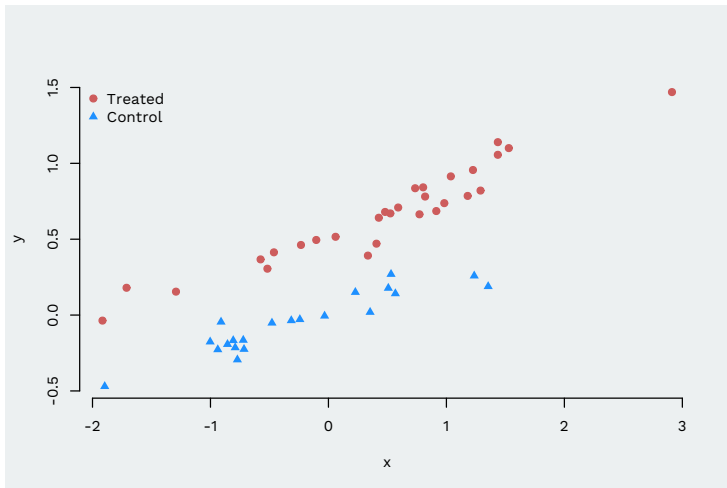
Parametric g-Formula With Flexible Models

- Impute the treated potential outcomes with $\hat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\hat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
 - ▶ Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control) using a flexible model.
 - ▶ Regress Y_i on X_i in the control group and get predicted values for all units (treated or control) using a flexible model.
 - ▶ Take the average difference between these predicted values (uncertainty via bootstrap)
- More mathematically, look like this:

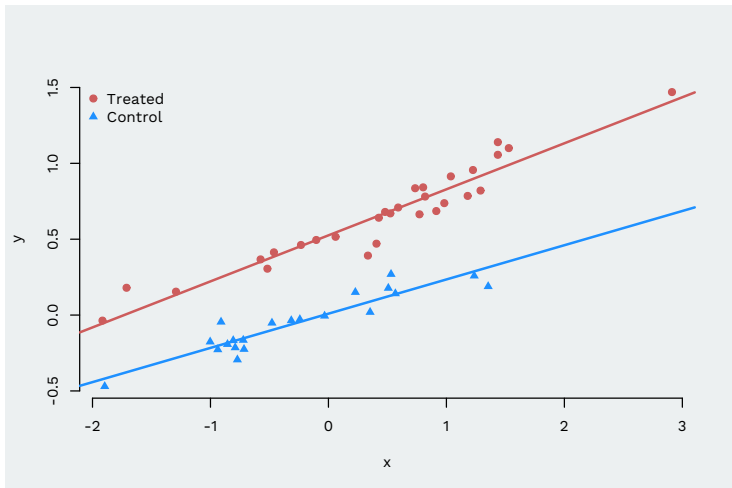
$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- If the models are **consistent** for the conditional expectation (and identification assumptions hold) will be consistent for the treatment effect!

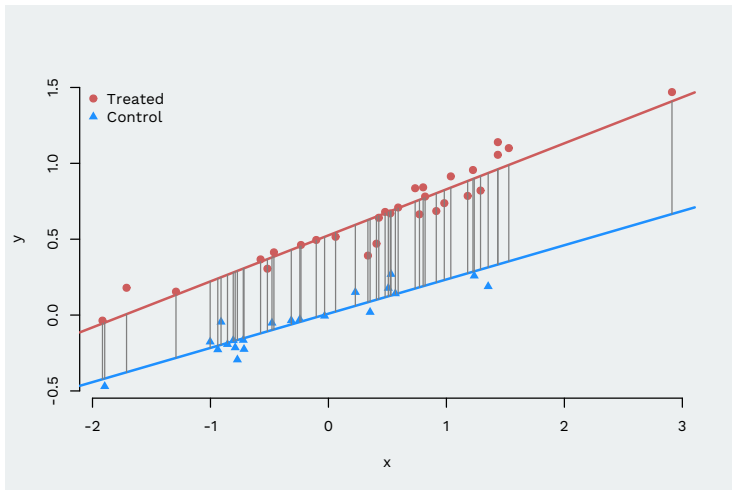
Imputation estimator visualization



Imputation estimator visualization

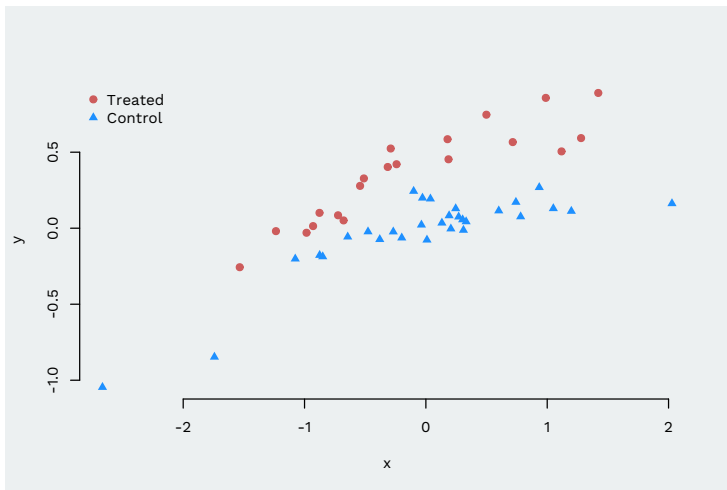


Imputation estimator visualization



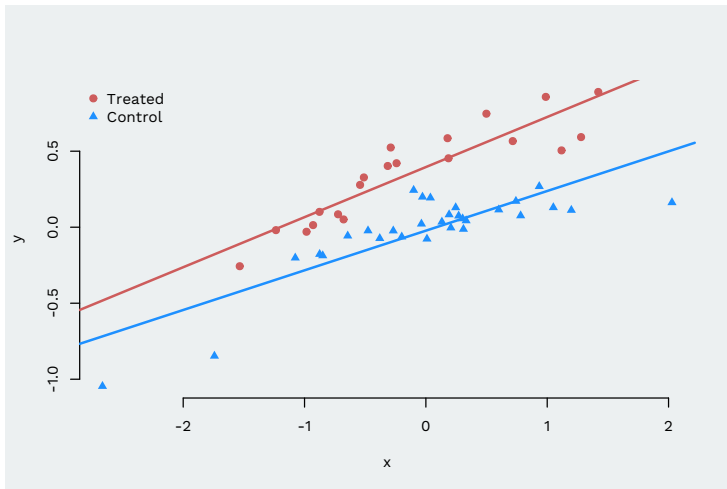
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i :



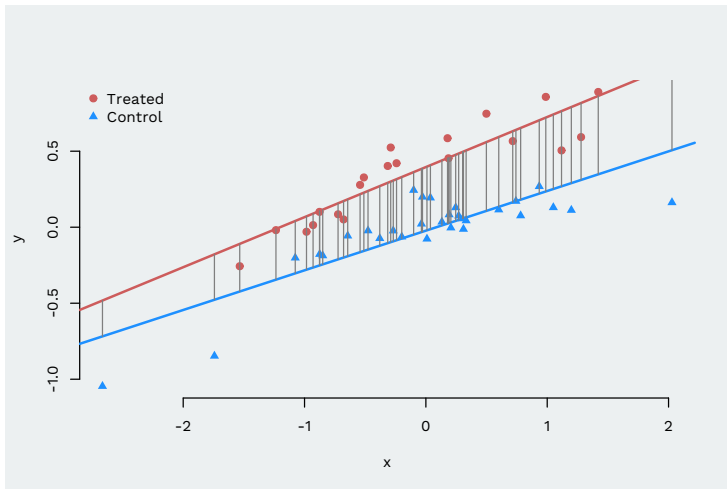
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i :



Nonlinear relationships

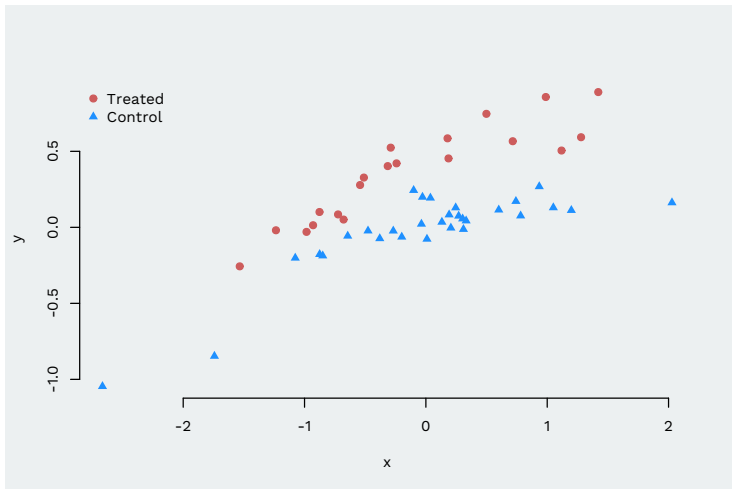
- Same idea but with nonlinear relationship between Y_i and X_i :



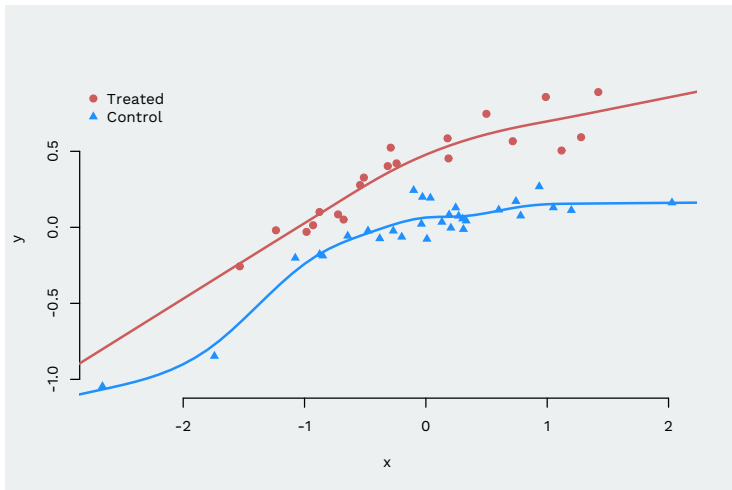
Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use Generalized Additive Models (GAMs) from the `mgcv` package for flexible estimate (a kind of machine learning).
- More generally we can use general machine learning estimators for this kind of adjustment (see e.g. Double Machine Learning from Chernozhukov et. al.)

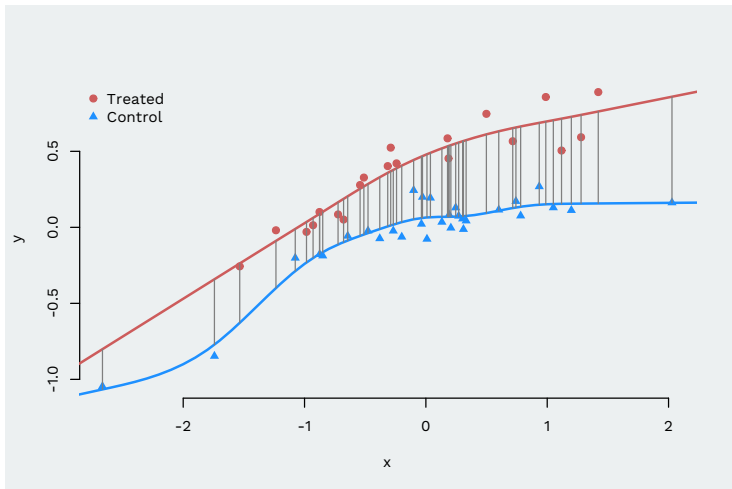
Using GAMs



Using GAMs



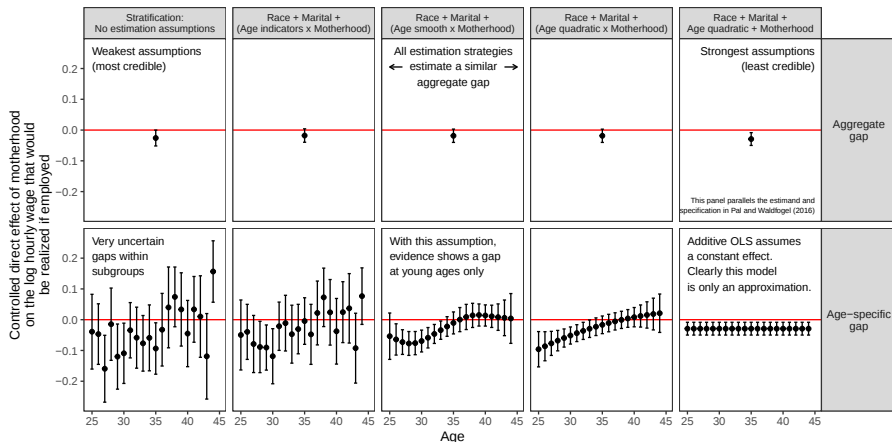
Using GAMs



Parametric g -Formula Estimators

- Imputation estimators (via the parametric g -formula) are a great and **underutilized** technique (to learn more, check out Naimi, Cole, and Kennedy, 2017).
- It is harder to implement than vanilla OLS particularly for uncertainty estimation, but you can always bootstrap!
- If $\hat{\mu}_t(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- To be flexible, people are increasingly using machine learning techniques like: kernel regression, neural networks, regression trees, etc.
- As we just saw, GAMs are a nice trade-off of the ease vs. flexibility side.
- These kinds of things will tend to matter a lot more for conditional treatment effects than the overall aggregate treatment effect, but you also don't know for sure until you try.

Example from Lundberg, Johnson and Stewart

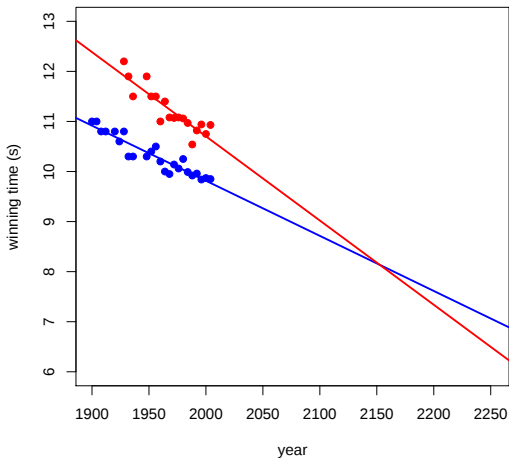


Advantages of the Regression Framework

- Learning how **regression** behaves as an estimator is helpful for many reasons:
 - ▶ people use regression all the time and it is useful to know what the coefficient represents
 - ▶ the variance-weighting is a key ingredient in understanding how many additively separable models behave.
 - ▶ it lends itself nicely to various kinds of sensitivity analysis (e.g. Cinelli and Hazlett 2019) and diagnostics
- Basic linear regression is an efficient estimator for relatively small datasets that in practice is probably at least going to get the sign correct.
- Remember a fancier regression only makes the estimation assumptions more credible, not the identification assumption!

- 1 Regression with Measured Confounding
 - Regression with Constant Effects
 - Additive Regression with Heterogeneous Effects
 - Imputation Estimators
- 2 Matching
 - Fundamentals of Matching
 - Two Approaches to Matching
- 3 Fixed Effects
 - Non-parametric Identification and Fixed Effects

Remember This?



Tatem et al.'s predictions of sprint times with alternate models. Men's times are in blue, women's times are in red.

Model Dependence

Model Free Inference

To estimate $E(Y|X = x)$ at x , average many observed Y with value x

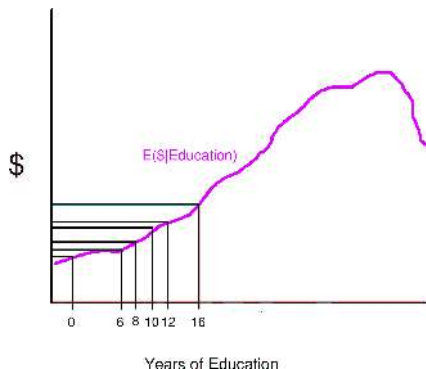
Assumptions (Model-Based Inference)

- ① Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
- ② The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

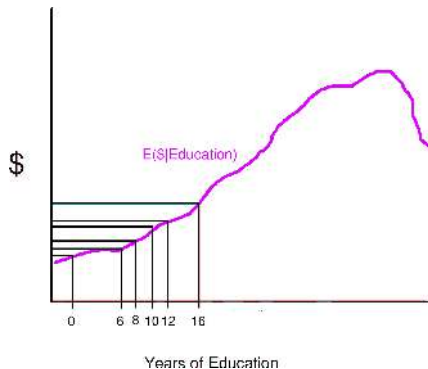
The maximum degree of model dependence: solely a function of the **distance from the counterfactual to the data**

What Inferences Would You Be Willing to Make?



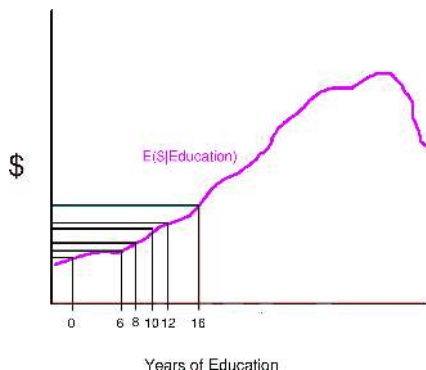
- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?
- The **model-free estimate**: $\text{mean}(Y)$ among those with $X = 12$.
- The **model-based estimate**: $\hat{Y} = X\hat{\beta} = 12 \times \$1,000 = \$12,000$

Counterfactual Inferences with Interpolation



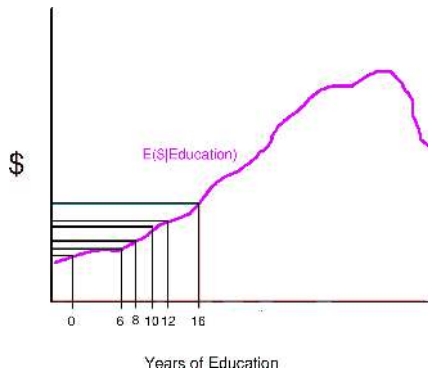
- How much salary would someone receive with **14** years of education (an Associates Degree)?
- **Model free estimate:** impossible
- **Model-based estimate:** $\hat{Y} = X\hat{\beta} = 14 \times \$1,000 = \$14,000$

Counterfactual Inference with Extrapolation



- How much salary would someone receive with 24 years of education (a Ph.D.)?
- $\hat{Y} = X\hat{\beta} = 24 \times \$1,000 = \$24,000$

Another Counterfactual Inference with Extrapolation



- How much salary would someone receive with **53** years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$
- What's changed? How would we recognize it when the example is less extreme or multidimensional?

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \dots, 10$.
- **Model-free** method: average observations on Y for each value of X
- **Model-based** method: regress Y on X , summarizing 10 parameters with 2 (intercept and slope).
- The **difference** between the 10 we need and the 2 we estimate with regression is **pure assumption**.
- (If X were continuous, we would be reducing ∞ to 2, also by assumption)

Model Dependence with **Two** Explanatory Variables

Variables: X (education) and Z , parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the **curse of dimensionality**: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).
- But what about including an interaction? Right, so now we're summarizing 100 parameters with 4.
- The difference: an enormous assumption based on convenience, not evidence or theory.

Model Dependence with **Many** Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
 - ▶ need to estimate 10^{15} (a quadrillion) parameters with how many observations?
 - ▶ Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.
 - ▶ 10^{80} is more than the number of atoms in the universe.
 - ▶ Yet, with a few simple assumptions, we can still run a regression and estimate only 81 parameters.
- The curse of dimensionality introduces huge assumptions, often unrecognized.

Overview of Matching

- Goal: **reduce model dependence** in our matching approach
- Makes parametric models work better rather than substitute for them (i.e., matching is not an estimator; it's a preprocessing method).
- It also prunes away data where we don't have **common support** (both treatment and control at same level of x),
- Apply model to preprocessed (pruned) rather than raw data
- Overall idea:
 - ▶ If each treated unit **exactly matches** a control unit w.r.t. X , then: (1) treated and control groups are identical, (2) X is no longer a confounder, (3) no need to worry about the functional form ($\bar{Y}_T - \bar{Y}_C$ is good enough).
 - ▶ If treated and control groups are **better balanced** than when you started, due to pruning, model dependence is reduced

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i :

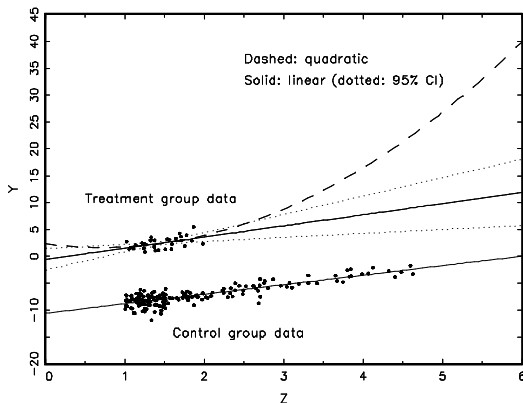
$$\begin{aligned} TE_i &= Y_i(1) - Y_i(0) \\ &= \text{observed} - \text{unobserved} \end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence
- Follow preprocessing with whatever statistical method you'd have used without matching

(Warning: Pruning nonmatches can change your feasible estimand.)

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 [Political Analysis](#))

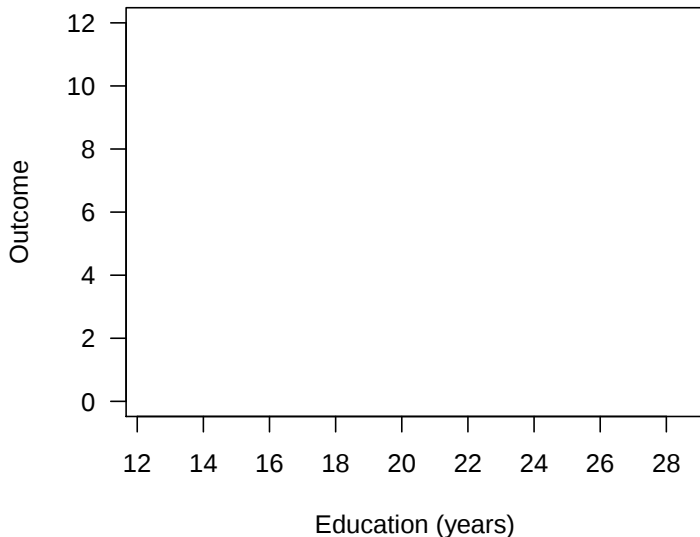


What to do?

- **Preprocess I:** Eliminate extrapolation region
- **Preprocess II:** Match (prune) within interpolation region
- **Model** remaining imbalance (as you would w/o matching)

Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, [Political Analysis](#))



Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).
- lognormal survival model.
- seven oversight variables (median adjusted ADA scores for House and Senate Committees as well as for House and Senate floors, Democratic Majority in House and Senate, and Democratic Presidency).
- 18 control variables (clinical factors, firm characteristics, media variables, etc.)

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).
- run 262,143 possible specifications and calculates ATE for each.
- Look at variability in ATE estimate across specifications.
- (Normal applications would only use one or a few specifications.)

Reducing Model Dependence

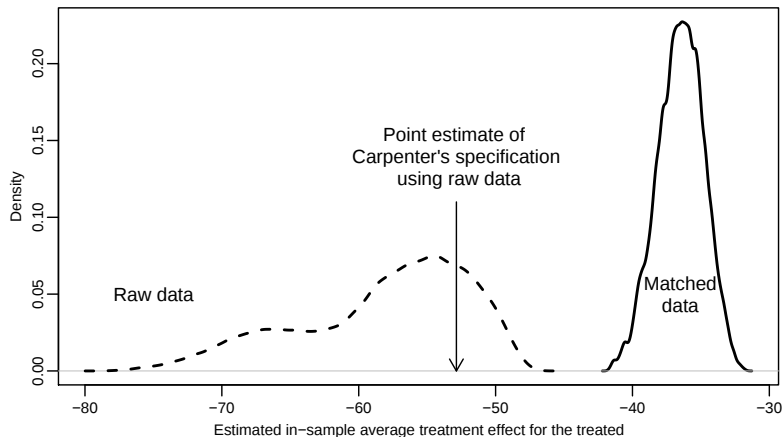


Figure: SATT Histogram: Effect of Democratic Senate majority on FDA drug approval time, across 262,143 specifications.

- 1 Regression with Measured Confounding
 - Regression with Constant Effects
 - Additive Regression with Heterogeneous Effects
 - Imputation Estimators
- 2 Matching
 - Fundamentals of Matching
 - Two Approaches to Matching
- 3 Fixed Effects
 - Non-parametric Identification and Fixed Effects

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \dots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
 - ▶ Find the set of unmatched control units j such that $X_i = X_j$
 - ▶ Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \dots, j(N_t)\}$ be the set of matched controls.
- Last, discard all unmatched control units.
- The distribution of X_i will be **exactly** the same for treated and matched control:

$$P(X_i = x | T_i = 1) = P(X_i = x | T_i = 0, \mathbb{I}_c)$$

Weakening the identification assumptions

- No unmeasured confounders, SUTVA, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to **conditional mean independence** (CMI):

$$E[Y_i(0)|X_i, T_i = 1] = E[Y_i(0)|X_i, T_i = 0]$$

- Two nice features of CMI:
 - 1 Only have to make assumptions about $Y_i(0)$ not $Y_i(1)$
 - 2 Only places restrictions on the means, not other parts of the distribution (variance, skew, kurtosis, etc)

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\hat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_i - \frac{1}{N_c} \sum_{j \in \mathbb{I}_c} Y_j$$

- Notice that we matched 1 treated to 1 control exactly, so we have:

$$\hat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - Y_{j(i)})$$

- $\rightsquigarrow$ average of the within matched-pair differences.

Beyond exact matching

- With high-dimensional X_i , not feasible to exact match.
- Let S be a **matching solution**: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$T_i \perp\!\!\!\perp X_i | S$$

- With no unmeasured confounders we have:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i | S$$

- Balance is checkable $\rightsquigarrow$ are T_i and X_i related in the matched data?

The matching procedure

- 1 Choose a number of matches
- 2 Choose a distance metric
- 3 Find matches (drop non-matches)
- 4 Check balance
- 5 Repeat (1)-(4) until balance is acceptable
- 6 Calculate the effect of the treatment on the outcome in the matched dataset.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
 - ▶ $P(X_i = x | T_i = 1) = P(X_i = x | T_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to “impute” missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\hat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

- New estimator for the effect:

$$\hat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \hat{Y}_i(0))$$

- Under no unmeasured confounding, $\hat{Y}_i(0)$ is a good predictor of the true potential outcome under control, Y_i .

Number of matches

- How many control matches should we include?
 - ▶ Small $M \rightsquigarrow$ small sample sizes
 - ▶ Large $M \rightsquigarrow$ worse matches (each additional match is further away).
- If M varies by treated unit, need to weight observations to ensure balance.

With or without replacement

- **Matching with replacement:** a single control unit can be matched to multiple treated units
- Benefits:
 - ▶ Better matches!
 - ▶ Order of matching does not matter.
- Drawbacks:
 - ▶ Inference is more complicated.
 - ▶ $\rightsquigarrow$ need to account for multiple appearances with weights.
 - ▶ Potentially higher uncertainty (using the same data multiple times = relying on less data).

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i .
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a **distance metric** which maps two covariates vectors into a single number.
 - ▶ Lower values $\rightsquigarrow$ more similar values of X_i .
 - ▶ Choice of distance metric will lead to different matches.

Exact distance metric

- **Exact**: only match units to other units that have the same exact values of X_i .

$$D_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

Euclidean distance

- The normalized **Euclidean distance metric** just uses the sum of the normalized distances for each covariate.
 - ▶ “Closeness” is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \dots, X_{iK})'$, so that there are K covariates.
- Then the Euclidean distance metric is:

$$D_{ij} = \sqrt{\sum_{k=1}^K \frac{(X_{ik} - X_{jk})^2}{\hat{\sigma}_k^2}}$$

- Here, $\hat{\sigma}_k^2$ is the variance of the k th variable:

$$\hat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^N (X_{ik} - \bar{X}_k)^2$$

Mahalanobis distance

- **Mahalanobis distance:** Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
 - ▶ Easy to get close on correlated covariates $\rightsquigarrow$ downweight.
 - ▶ Harder to get close on uncorrelated covariates $\rightsquigarrow$ upweight.
- Metric:

$$D_{ij} = \sqrt{(X_i - X_j)' \hat{\Sigma}^{-1} (X_i - X_j)}$$

- $\hat{\Sigma}$ is the estimated variance-covariance matrix of the observations:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})^T$$

Complications

- Combining distance metrics:
 - ▶ Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
 - ▶ Dropping those matches (treated and control) improves balance.
 - ▶ Dropping treated units changes the quantity of interest.
- Implementation: a **caliper**, which is the maximum distance we would accept

Estimands

- Matching easiest to justify for the Average Treatment Effect on the Treated.
 - ▶ Dropping control units doesn't affect this identification.
- Can also identify the Average Treatment Effect on Control by finding matched treated units for the controls.
- Combine the two to get the ATE:

$$\tau = \tau_{ATT}P(T_i = 1) + \tau_{ATC}P(T_i = 0)$$

- Estimated:

$$\hat{\tau} = \hat{\tau}_{ATT} \left(\frac{N_t}{N} \right) + \hat{\tau}_{ATC} \left(\frac{N_c}{N} \right)$$

Moving the goalposts

- **Common support:** finding the subspace of X_i where there is overlap between the treated and control groups.
 - ▶ Hard to extrapolate outside region.
 - ▶ Theoretical: effect of voting for those under 18 ($P(D_i = 1|X_i < 18) = 0$).
 - ▶ Empirical: no/extremely few treated units in a sea of controls.
 - ▶ Solution: restrict analysis to common support (dropping treated and controls).
- **Moving the goalposts:** dropping treated units.
 - ▶ We move away from being able to identify the Average Treatment Effect on the Treated (ATT).
 - ▶ Now it's the ATT in the matched subsample (sometimes called the feasible ATT).
 - ▶ Good to be clear about this.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg \min_{j \in \mathbb{J}_c} D_{ij}$$

- ▶ $\mathbb{J}_c$ are the available controls for matching.
- This is **nearest neighbor**: “Find the control unit with the smallest distance metric.”
- Do the same for all treated units.
- What about ties?
 - ▶ Randomly choose between them.
- Note: in nearest neighbor without replacement the **order matters**!

Assessing balance

- All matching methods seek to maximize balance:

$$P(X_i = x | T_i = 1, S) = P(X_i = x | T_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
 - ▶ If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
 - ▶ Differences-in-means/medians, standardized.
 - ▶ Quantile-quantile plots/KS statistics for comparing the entire distribution of X_i .
 - ▶ L_1 : multivariate histogram.

- 1 Regression with Measured Confounding
 - Regression with Constant Effects
 - Additive Regression with Heterogeneous Effects
 - Imputation Estimators
- 2 Matching
 - Fundamentals of Matching
 - Two Approaches to Matching
- 3 Fixed Effects
 - Non-parametric Identification and Fixed Effects

Two Approaches to Matching

- There are **many** approaches to matching. We will cover just two for the sake of time.
- This isn't a statement that these are the best two, just a set which are straightforward to learn.
- Which is the best method? The one that produces the best balance!

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1 Preprocess (Matching)

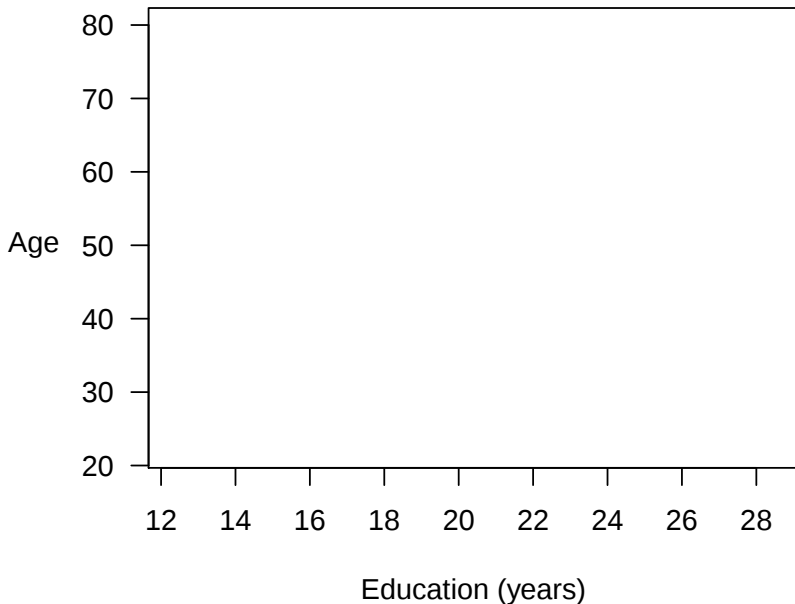
- ▶ $\text{Distance}(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
- ▶ Match each treated unit to the nearest control unit
- ▶ Control units: not reused; pruned if unused
- ▶ Prune matches if $\text{Distance} > \text{caliper}$

2 Checking Measure imbalance, tweak, repeat, ...

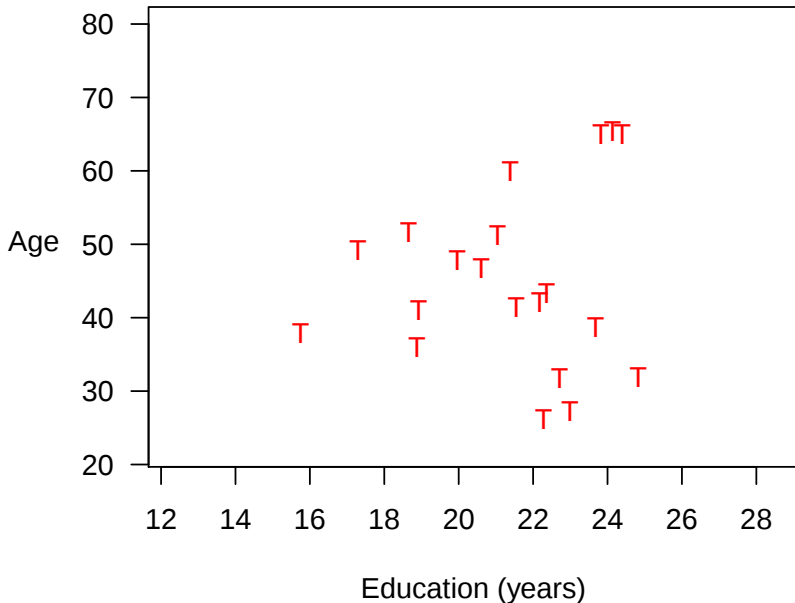
3 Estimation Difference in means or a model



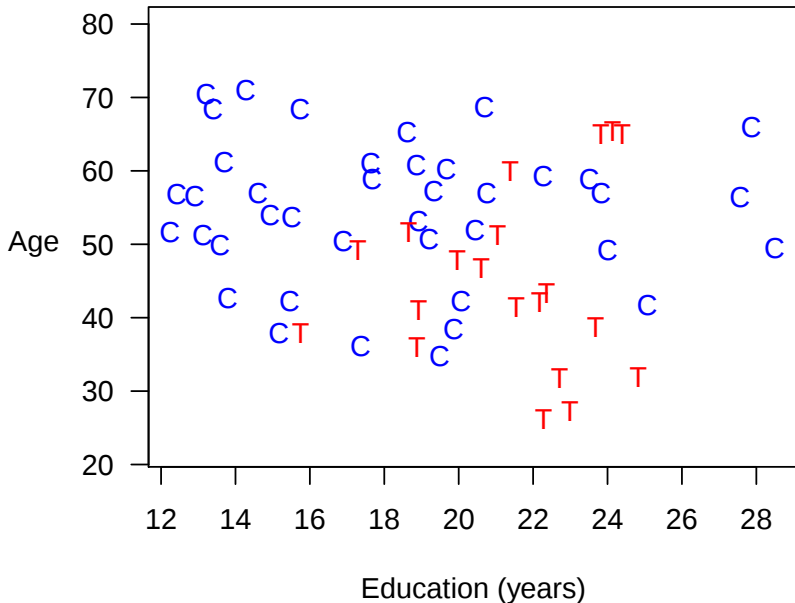
Mahalanobis Distance Matching



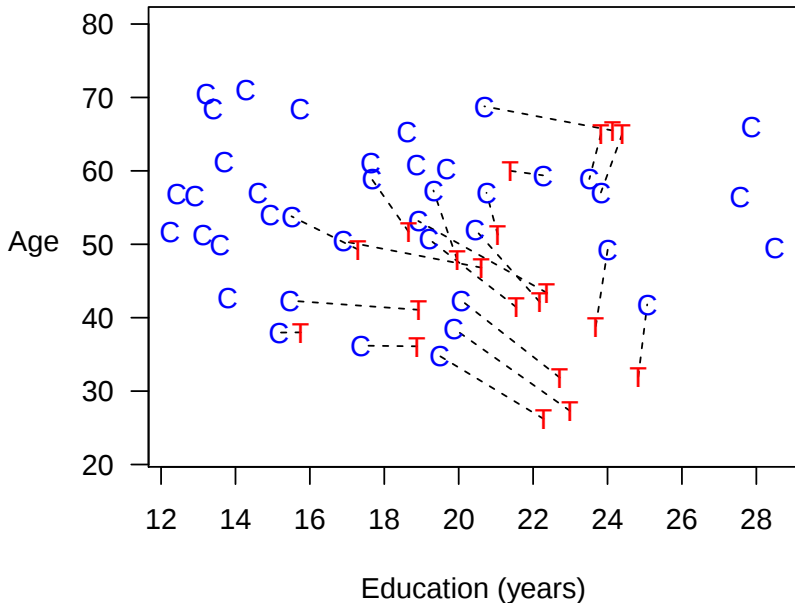
Mahalanobis Distance Matching



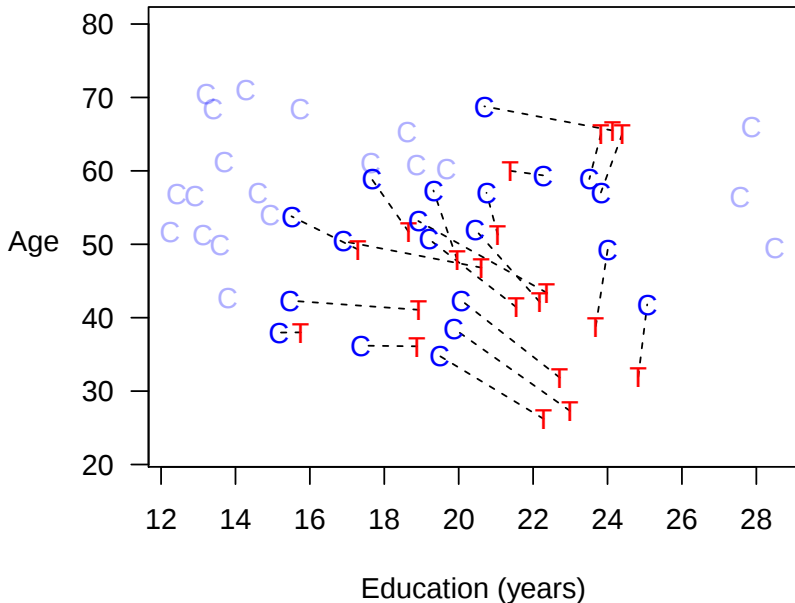
Mahalanobis Distance Matching



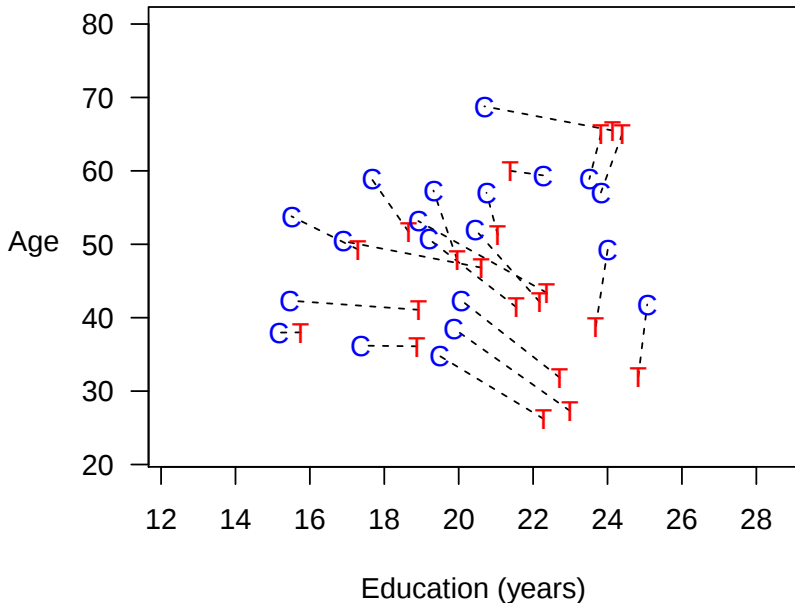
Mahalanobis Distance Matching



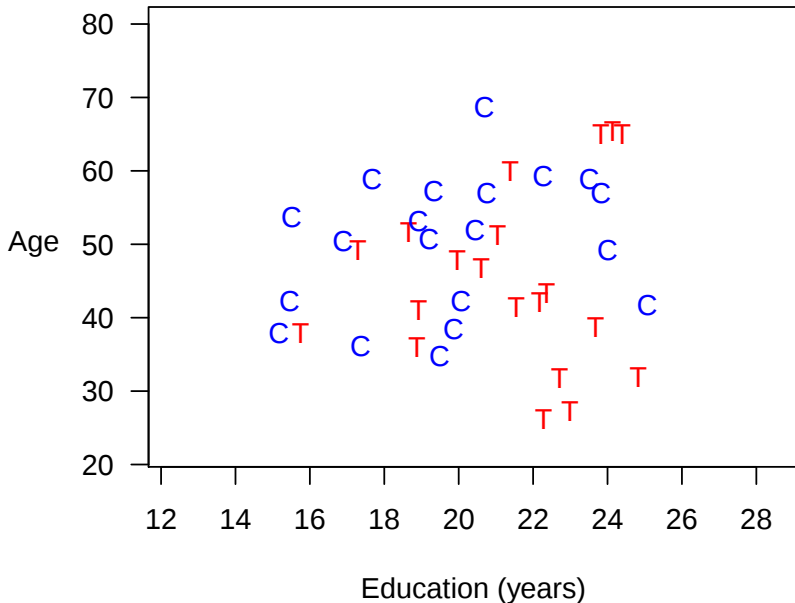
Mahalanobis Distance Matching



Mahalanobis Distance Matching



Mahalanobis Distance Matching



Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1 Preprocess (Matching)

- ▶ Temporarily coarsen X as much as you're willing
 - ★ e.g., Education (grade school, high school, college, graduate)
- ▶ Apply exact matching to the coarsened X , $C(X)$
 - ★ Sort observations into strata, each with unique values of $C(X)$
 - ★ Prune any stratum with 0 treated or 0 control units
- ▶ Pass on original (uncoarsened) units except those pruned

2 Checking Determine matched sample size, tweak, repeat, ...

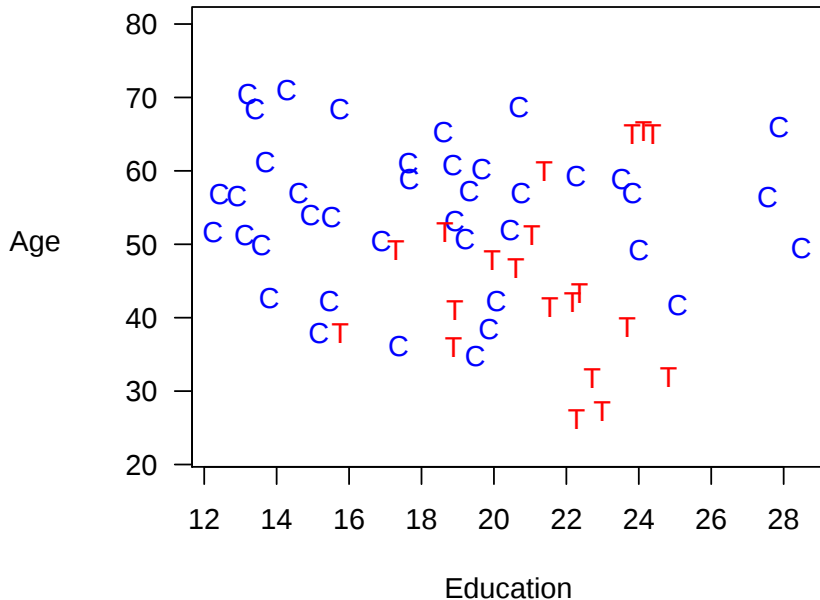
- ▶ Easier, but still iterative

3 Estimation Difference in means or a model

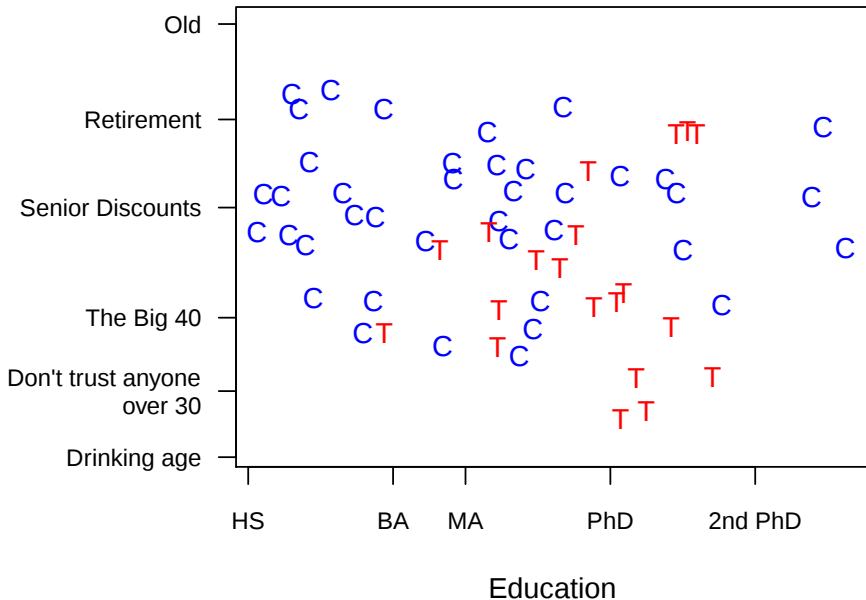
- ▶ Need to weight controls in each stratum to equal treateds

Coarsened Exact Matching

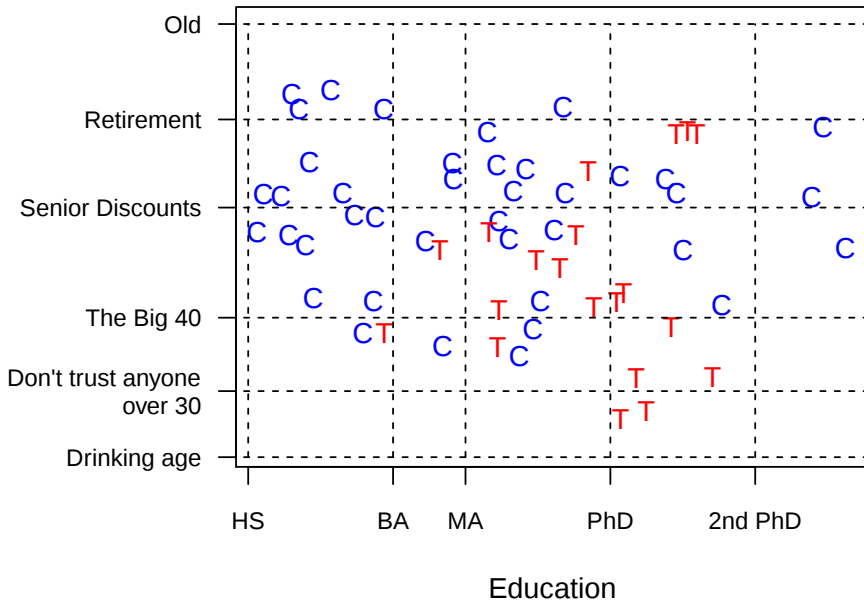
Coarsened Exact Matching



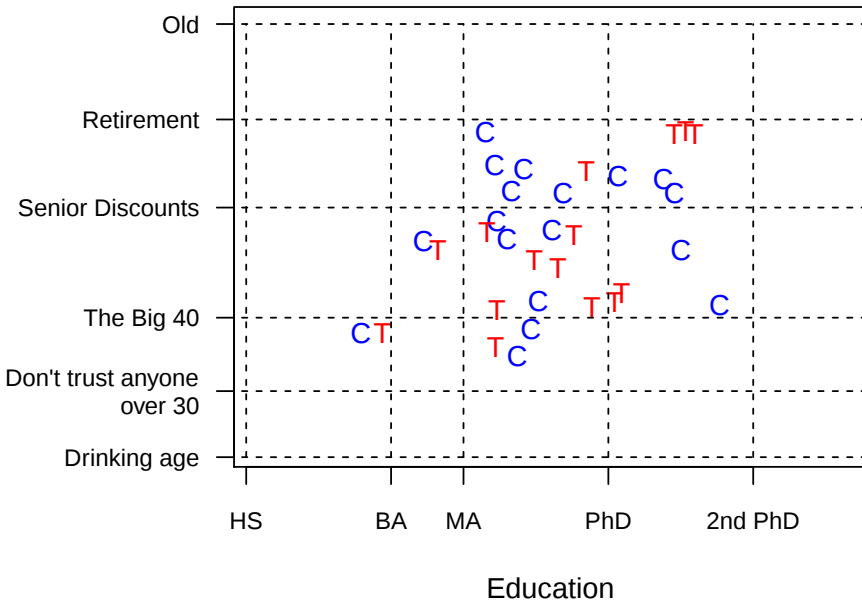
Coarsened Exact Matching



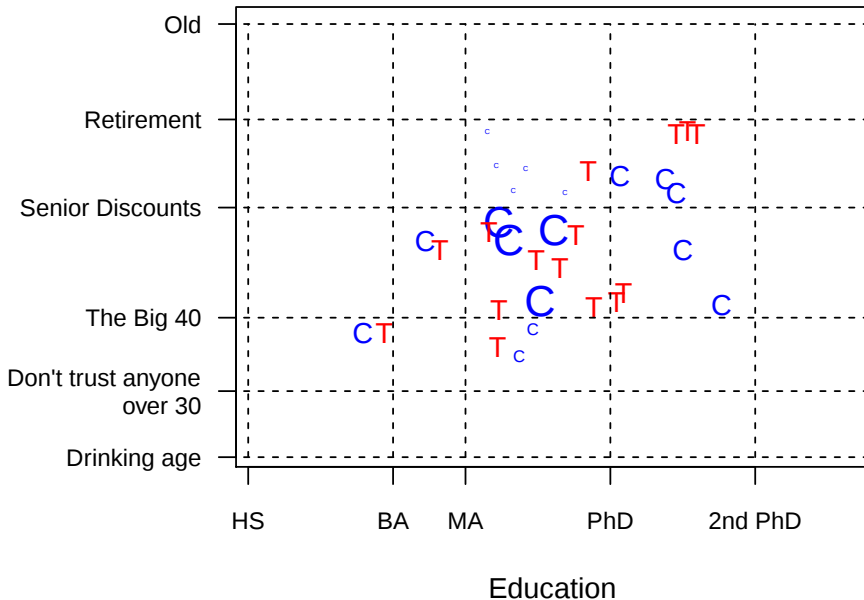
Coarsened Exact Matching



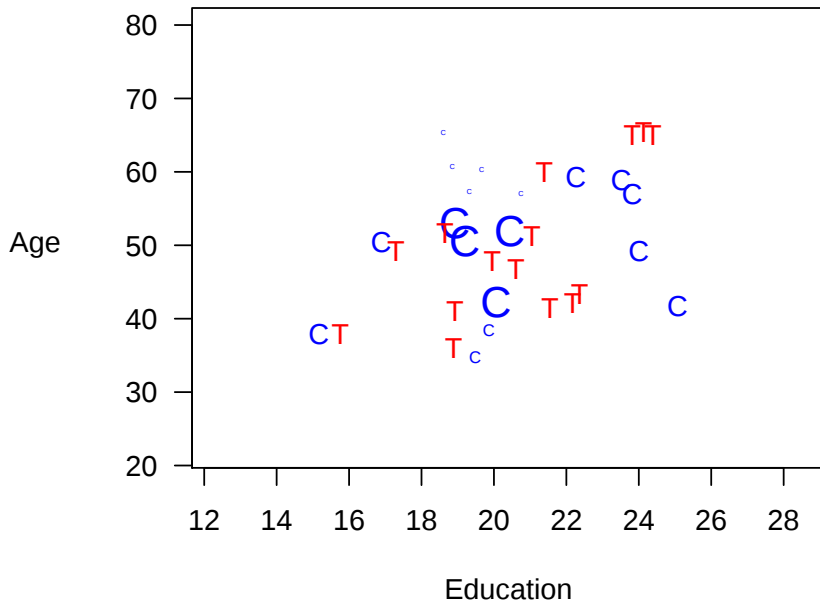
Coarsened Exact Matching



Coarsened Exact Matching



Coarsened Exact Matching



Final Thoughts on Matching

- Matching is an entire literature unto itself (you could probably teach a whole class just on that). There is so much we didn't cover here including bias corrections and how to get variance estimators.
- The central advantage of matching is that it is **transparent** about where the counterfactual estimates are coming from (you can look at the matched unit!).
- Technically the thing it is buying you relative to regression methods we will talk about next is that it limits **extrapolation**.
- But there is no need for these techniques to compete—we can match and then use regression!
- Importantly, there is **nothing magic** about matching, it is just another way of conditioning.

- 1 Regression with Measured Confounding
 - Regression with Constant Effects
 - Additive Regression with Heterogeneous Effects
 - Imputation Estimators
- 2 Matching
 - Fundamentals of Matching
 - Two Approaches to Matching
- 3 Fixed Effects
 - Non-parametric Identification and Fixed Effects

Basic Model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

- Consider the above basic model where a_i is an intercept associated with unit i and u_{it} satisfy strict exogeneity.
- The **fixed effects model** is a way to remove time-invariant unmeasured confounding from a_i
- We will start by assuming the model and discussing properties and in the next section, we will consider non-parametric identification.

Fixed Effects Models

- Core idea is to focus on **within-unit comparisons**: changes in y_{it} and x_{it} relative to their within-group means
- First note that taking the average of the y 's over time for a given unit leaves us with a very similar model:

$$\begin{aligned}\bar{y}_i &= \frac{1}{T} \sum_{t=1}^T [\mathbf{x}'_{it}\boldsymbol{\beta} + a_i + u_{it}] \\ &= \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}'_{it} \right) \boldsymbol{\beta} + \frac{1}{T} \sum_{t=1}^T a_i + \frac{1}{T} \sum_{t=1}^T u_{it} \\ &= \bar{\mathbf{x}}'_i \boldsymbol{\beta} + a_i + \bar{u}_i\end{aligned}$$

- Key fact: because it is **time-constant** the mean of a_i is just a_i
- This regression is sometimes called the “between regression”

Within Transformation

- The “fixed effects,” “within,” or “time-demeaning” transformation is when we subtract off the over-time means from the original data:

$$(y_{it} - \bar{y}_i) = (\mathbf{x}'_{it} - \bar{\mathbf{x}}'_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i)$$

- If we write $\ddot{y}_{it} = y_{it} - \bar{y}_i$, then we can write this more compactly as:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}'_{it}\boldsymbol{\beta} + \ddot{u}_{it}$$

- Degrees of freedom: $nT - n - k - 1$, which accounts for within transformation (i.e. either use a package like `plm` or adjust the degrees of freedom manually).
- We are now modeling observations as deviation from their group mean.
- NB: you **must** demean the X variables not just the Y variables.

Fixed Effects Example with data from Ross

```
fe.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross, index = c("id", "year"),
  model = "within")
summary(fe.mod)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = "within", index = c("id", "year"))
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##      Min.   1st Qu.   Median   3rd Qu.    Max.
## -0.70500 -0.11700  0.00628  0.12200  0.75700
##
## Coefficients :
##              Estimate Std. Error t-value Pr(>|t|)
## democracy    -0.143233   0.033500  -4.2756 2.299e-05 ***
## log(GDPcur)  -0.375203   0.011328 -33.1226 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    81.711
## Residual Sum of Squares: 23.012
## R-Squared      : 0.71838
##      Adj. R-Squared : 0.53242
## F-statistic: 613.481 on 2 and 481 DF, p-value: < 2.22e-16
```

Strict Exogeneity

- FE models are valid if $E[\mathbf{u}|\mathbf{X}] = 0$: all errors are uncorrelated with covariates in every period:

$$E[\ddot{u}_{it}|\ddot{\mathbf{X}}] = E[u_{it}|\ddot{\mathbf{X}}] - E[\bar{u}_i|\ddot{\mathbf{X}}] = 0 - 0 = 0$$

- This is because the composite errors, $\ddot{u}_{it}$ are function of the errors in every time period through the average, $\bar{u}_i$
- This rules out, for instance, lagged dependent variables, since $y_{i,t-1}$ has to be correlated with $u_{i,t-1}$. Thus it can't be a covariate for y_{it} .

Fixed Effects and Time-Invariant Covariates

- What if there is a covariate that doesn't vary over time?
- Then $x_{it} = \bar{x}_i$ and $\ddot{x}_{it} = 0$ for all periods t .
- If the time-demeaned covariate is always 0, then it will be perfectly collinear with the intercept and will violate full rank. R/Stata will **drop** it from the regression.
- Basic message: any time-constant variable gets “absorbed” by the fixed effect. It has nothing to contribute because the comparison is **within the units**.
- Can include interactions between time-constant and time-varying variables, but lower order term of the time-constant variables get absorbed by fixed effects too

Time-constant variables

- Pooled model with a time-constant variable, proportion Islamic:

```
library(lmtest)
p.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
              data = ross, index = c("id", "year"), model = "pooling")
coeftest(p.mod)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30607817 0.35951939 28.6663 < 2.2e-16 ***
## democracy   -0.80233845 0.07766814 -10.3303 < 2.2e-16 ***
## log(GDPcur) -0.25497406 0.01607061 -15.8659 < 2.2e-16 ***
## islam        0.00343325 0.00091045  3.7709 0.0001794 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time-constant variables

- FE model, where the islam variable drops out, along with the intercept:

```
fe.mod2 <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,  
               data = ross, index = c("id", "year"), model = "within")  
coeftest(fe.mod2)
```

```
##  
## t test of coefficients:  
##  
##              Estimate Std. Error  t value  Pr(>|t|)  
## democracy    -0.129693   0.035865  -3.6162 0.0003332 ***  
## log(GDPcur)  -0.379997   0.011849 -32.0707 < 2.2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternate Computation: Least Squares Dummy Variable

- As an alternative to the within transformation, we can also include a series of $n - 1$ dummy variables for each unit:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + d_i^{(1)}\alpha_1 + d_i^{(2)}\alpha_2 + \cdots + d_i^{(n)}\alpha_n + u_{it}$$

- Here, $d_i^{(1)}$ is a binary variable which is 1 if $i = 1$ and 0 otherwise—just a unit dummy.
- Gives the **exact** same estimates/standard errors as with time-demeaning
 - Advantage: easy to implement in base R (so is the de-meaning but you have to recompute standard errors by changing the degrees of freedom manually).
 - Disadvantage: computationally difficult with large data sets, since we have to run a regression with $n + k$ variables.
- Why are these equivalent? (remember partialing out strategy and Frisch-Waugh-Lovell theorem)

Example with Ross data

```
library(lmtest)
lsdv.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur) +
               as.factor(id), data = ross)
coeftest(lsdv.mod)[1:6,]
coeftest(fe.mod)[1:2,]
```


##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	13.7644887	0.26597312	51.751427	1.008329e-198
##	democracy	-0.1432331	0.03349977	-4.275644	2.299393e-05
##	log(GDPcur)	-0.3752030	0.01132772	-33.122568	3.494887e-126
##	as.factor(id)AGO	0.2997206	0.16767730	1.787485	7.448861e-02
##	as.factor(id)ALB	-1.9309618	0.19013955	-10.155498	4.392512e-22
##	as.factor(id)ARE	-1.8762909	0.17020738	-11.023558	2.386557e-25

##		Estimate	Std. Error	t value	Pr(> t)
##	democracy	-0.1432331	0.03349977	-4.275644	2.299393e-05
##	log(GDPcur)	-0.3752030	0.01132772	-33.122568	3.494887e-126

Applying Fixed Effects

- We can use fixed effects for other data structures to restrict comparisons to within unit variation
 - ▶ Matched pairs
 - ★ Twin fixed effects to control for unobserved effects of family background
 - ▶ Cluster fixed effects in hierarchical data
 - ★ School fixed effects to control for unobserved effects of school

Moving Beyond Linear Separable Confounding

- One reason we like DAGs is that the identification results don't have to start with a statement like, assume the following linear model:

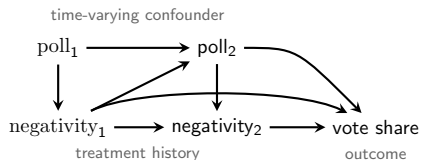
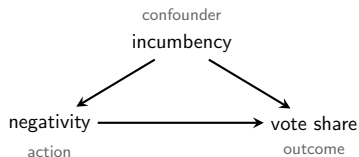
$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

- What assumptions have we made so far?
 - ▶ constant effects
 - ▶ linearity
 - ▶ strict exogeneity
- We've seen the trouble with constant effects before, it goes back to Lecture 10 and results on regression with heterogeneous treatment effects more generally.

Contemporaneous, Cumulative and Dynamic Effects

- Another assumption we have been making is that our interest is in a single contemporaneous effect: $\mathbf{x}'_{it}\beta$
- What if we want to consider the history of a treatment or the effect of a treatment regime (i.e. a treatment that varies over time)?
- Opens up new estimands, but have to think about how time-varying confounders affect treatment assignment.

Examples of static and dynamic causal inference problems:



Core Conundrum

There is a (possibly irresolvable) tension: modeling **causal dynamics** between treatment and outcomes **OR** addressing **unobserved time-invariant confounders**. Three great recent papers:

A Framework for Dynamic Causal Inference in Political Science

Nathaniel Blackwell | [I analyzed](#) | [Twitter](#)

Abstract: "This paper presents a framework for dynamic causal inference in political science. It begins by reviewing the limitations of existing methods for causal inference in political science, including the use of randomized controlled trials, regression discontinuity designs, and difference-in-differences designs. It then introduces a new framework for dynamic causal inference, which is based on the use of structural equation models and Bayesian inference. This framework allows researchers to estimate the causal effects of political events over time, while accounting for the possibility of unobserved time-invariant confounders. The framework is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States." [Full paper](#)

When we think about the core conundrum of causal inference in political science, we often think about the tension between modeling causal dynamics and addressing unobserved time-invariant confounders. This tension is at the heart of many of the most important debates in the field. In this paper, I propose a new framework for dynamic causal inference that aims to address this tension. The framework is based on the use of structural equation models and Bayesian inference. It allows researchers to estimate the causal effects of political events over time, while accounting for the possibility of unobserved time-invariant confounders. The framework is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

Journal of Economic Surveys (2024) 100, 105912
doi:10.1016/j.econsur.2024.105912
© 2024 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables
MATTHEW BLACKWELL ^{a,*}, HARSHAD UNIVERSITY
ADAM N. GLYNN ^b, ZEWITY UNIVERSITY

Repeated measurement of the same countries, people, or groups over time are often a major source of political science data. These measurements, however, often reflect time-varying unobserved confounders. This paper develops a new method for estimating causal effects from time-series cross-sectional data under selection on observables. The method is based on the use of structural equation models and Bayesian inference. It allows researchers to estimate the causal effects of political events over time, while accounting for the possibility of unobserved time-invariant confounders. The method is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

INTRODUCTION

Many inquiries in political science involve the study of repeated measurement of the same countries, people, or groups at several points in time. This type of data, sometimes called time-series cross-sectional (TSCS) data, allows researchers to study a large range of phenomena over time, while accounting for the possibility of unobserved time-invariant confounders. This paper develops a new method for estimating causal effects from TSCS data under selection on observables. The method is based on the use of structural equation models and Bayesian inference. It allows researchers to estimate the causal effects of political events over time, while accounting for the possibility of unobserved time-invariant confounders. The method is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

Repeated measurement of the same countries, people, or groups over time are often a major source of political science data. These measurements, however, often reflect time-varying unobserved confounders. This paper develops a new method for estimating causal effects from time-series cross-sectional data under selection on observables. The method is based on the use of structural equation models and Bayesian inference. It allows researchers to estimate the causal effects of political events over time, while accounting for the possibility of unobserved time-invariant confounders. The method is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

^a Matthew Blackwell is an Assistant Professor of Government and Economics at Harvard University. He is also a senior research advisor at the Center for Global Development. ^b Adam N. Glynn is an Assistant Professor of Economics at Zewity University. He is also a senior research advisor at the Center for Global Development.

Journal of the American Statistical Association (2024) 119(587), 1-15
doi:10.1080/01621459.2024.2311111
© 2024 The Author(s). Published by the American Statistical Association.
When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?
Kosuke Imai ^a, Nathan K. Elfrink ^b

Unit fixed effects regression models are a popular method for causal inference with longitudinal data. However, these models are often misapplied, leading to biased estimates of causal effects. This paper develops a new method for determining when to use unit fixed effects regression models for causal inference. The method is based on the use of structural equation models and Bayesian inference. It allows researchers to determine when to use unit fixed effects regression models for causal inference, while accounting for the possibility of unobserved time-invariant confounders. The method is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

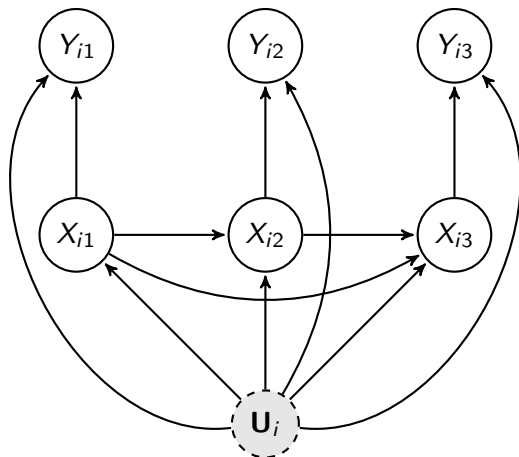
Unit fixed effects regression models are a popular method for causal inference with longitudinal data. However, these models are often misapplied, leading to biased estimates of causal effects. This paper develops a new method for determining when to use unit fixed effects regression models for causal inference. The method is based on the use of structural equation models and Bayesian inference. It allows researchers to determine when to use unit fixed effects regression models for causal inference, while accounting for the possibility of unobserved time-invariant confounders. The method is applied to a study of the effect of the 2008 financial crisis on the political economy of the United States.

We are going to focus on addressing unobserved time-invariant confounders using the last paper. Next several slides are based on slides graciously provided by In Song Kim and Kosuke Imai.

Directed Acyclic Graph (DAG)

Non-parametric identification assumptions for fixed effects:

$$Y_{it} = g(X_{it}, \mathbf{U}_i, \epsilon_{it}) \quad \text{and} \quad \epsilon_{it} \perp\!\!\!\perp \{\mathbf{X}_i, \mathbf{U}_i\}$$



Assumptions:

- 1 No unobserved time-varying confounders
- 2 Past outcomes do not directly affect current outcome
- 3 Past outcomes do not directly affect current treatment
- 4 Past treatments do not directly affect current outcome

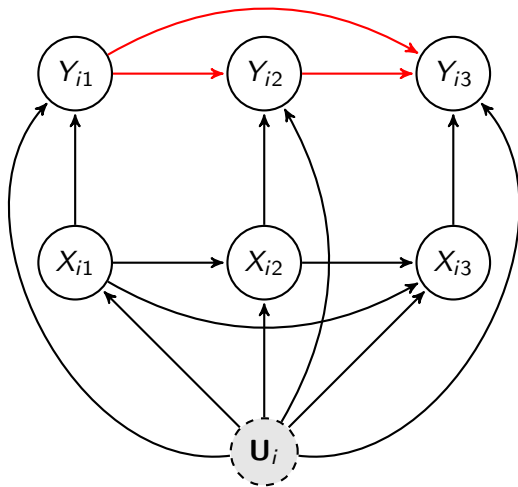
*the result implies that the **counterfactual** outcome for a treated observation in a given time period is estimated using the **observed outcomes of different time periods of the same unit**. Since such a comparison is **valid only when no causal dynamics exist**, this finding underscores the important limitation of linear regression models with unit fixed effects.*

- Imai and Kim (2019)

What Ideal Experiment Corresponds to the Fixed Effects Model?

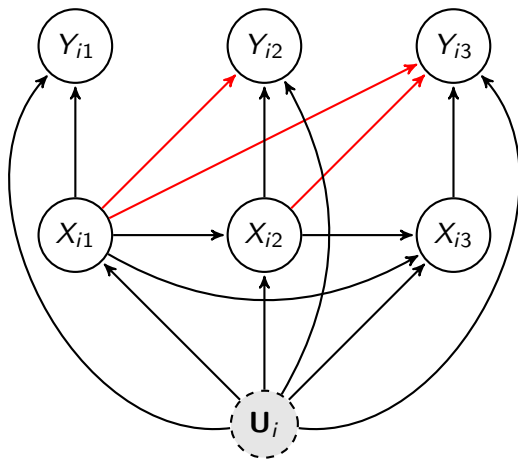
- Experiment that satisfies the model assumptions:
 - ① randomize X_{i1} given $\mathbf{U}_i$
 - ② randomize X_{i2} given X_{i1} and $\mathbf{U}_i$
 - ③ randomize X_{i3} given X_{i2} , X_{i1} , and $\mathbf{U}_i$
 - ④ and so on
- Experiment that does not satisfy the model assumptions:
 - ① randomize X_{i1}
 - ② randomize X_{i2} given X_{i1} and Y_{i1}
 - ③ randomize X_{i3} given X_{i2} , X_{i1} , Y_{i1} , and Y_{i2}
 - ④ and so on
- Now let's consider each assumption in turn.

Past Outcomes Don't Directly Affect Current Outcome (A2)



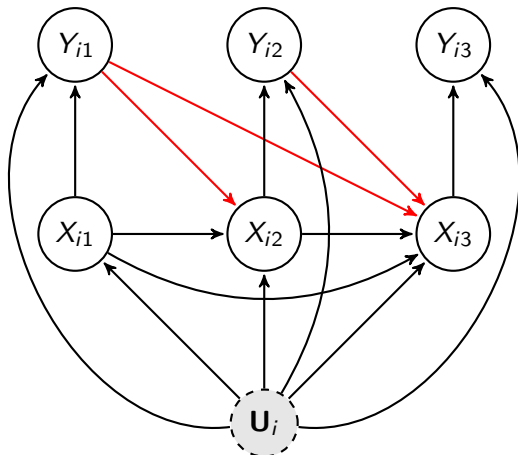
- Strict exogeneity still holds.
- Past outcomes do not confound $X_{it} \rightarrow Y_{it}$ given U_i .
- No need to adjust for past outcomes.
- Should use cluster robust standard errors for inference.
- Conclusion: **The assumption can be relaxed**

Past Treatments Don't Directly Affect Current Outcome (A4)



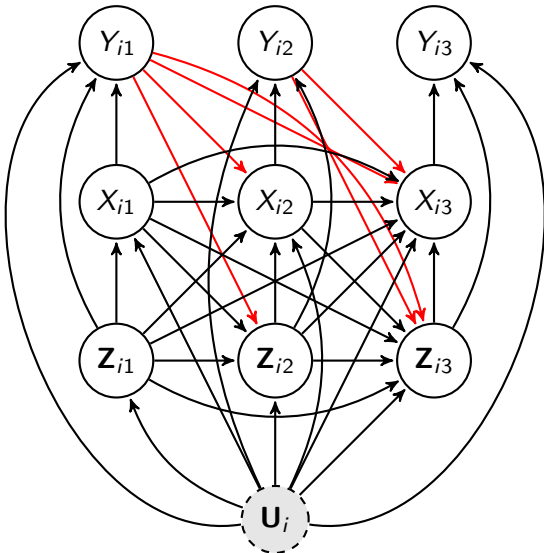
- Need to adjust for past treatments
- Strict exogeneity holds given past treatments and U_i
- Impossible to adjust for an entire treatment history and U_i at the same time
- Adjust for a small number of past treatments $\rightsquigarrow$ often arbitrary
- Conclusion: The assumption can be partially relaxed

Past Outcomes Don't Directly Affect Current Treatment (A3)



- Correlation between error term and future treatments
- Violation of strict exogeneity
- No adjustment is sufficient
- Implication: No dynamic causal relationships between treatment and outcome variables
- Conclusion: **The assumption cannot be relaxed**

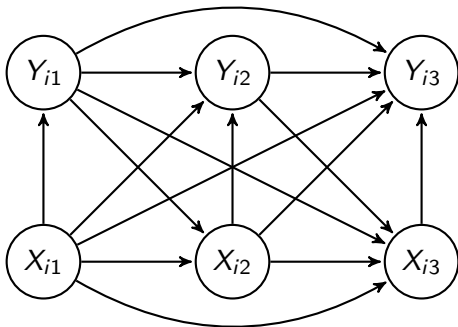
Can't We Just Adjust for Time-Varying Confounders?



- $Y_{it} = \alpha_i + \beta X_{it} + \gamma^\top \mathbf{Z}_{it} + \epsilon_{it}$
- past outcomes cannot directly affect current treatment
- past outcomes cannot *indirectly* affect current treatment through $\mathbf{Z}_{it}$

But What If I Have Causal Dynamics?

Alternative: **Marginal Structural Models** (Robins, Hernán and Brumback, 2000) — see Blackwell 2013 and Blackwell and Glynn 2018 for accessible introductions.



- Absence of unobserved time-invariant confounders $\mathbf{U}_i$
- past treatments can directly affect current outcome
- past outcomes can directly affect current treatment

- Comparison across units within the same time rather than across different time periods within the same unit
- Can identify the average effect of an entire treatment sequence
- Trade-off $\rightsquigarrow$ no free lunch

Conclusions and Nonparametric Estimation

- Imai and Kim (2019) offer a matching framework for fixed effects models which exploits an equivalence to weighted unit fixed effects estimators (see `wfe` package in R as well).
- The paper clarifies assumptions for fixed effects and first difference estimators.
- Follow-up papers Imai and Kim (2021) and Imai, Kim and Wang (2022+) extends to two-way fixed effects estimator.
- Tradeoff:
 - 1) unobserved time-invariant confounders $\rightsquigarrow$ fixed effects
 - 2) causal dynamics between treatment and outcome $\rightsquigarrow$ selection-on-observables

Summary Table (Imai and Kim 2019)

TABLE 1 Identification Assumptions of Various Estimators

	Linearity	Time-Invariant Unobservables	Past Outcomes Affect Current Treatment	Past Treatments Affect Current Outcome
$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$	Yes	Allowed	Not allowed	Not allowed
$Y_{it} = \alpha_i + \beta X_{it} + \rho Y_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Allowed	Not allowed
$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Not allowed	Allowed
$Y_{it} = \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \rho Y_{i,t-1} + \epsilon_{it}$	Yes	Allowed	Partially allowed	Partially allowed
Marginal structural models	No	Not allowed	Allowed	Allowed

What to read next?

- Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects
- Imai and Kim (2019) “When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?” *American Journal of Political Science*,
<http://dx.doi.org/10.1111/ajps.12417>
- Liu, Wang and Xu “A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data” *American Journal of Political Science*.

We Covered

- Regression based estimation of causal effects under measured confounding.
- Imputation estimators which generalize to broader class of machine learning estimators for the conditional expectation function.
- Matching as nonparametric preprocessing
- Fixed Effects
- Sensitivity Analysis

Next Time: What to Do About Unmeasured Confounding (Week 11)