A Brief Introduction to Matching Estimators

Brenton Kenkel

University of Rochester
The Star Lab

21 October 2011

Roadmap

Purposes of this talk:

- Introduce you to the philosophy and lingo of causal inference
- Familiarize you with the most popular matching estimators
- Point you to software and further reading if you want to use matching in your research
- Convince you that this stuff isn't strange or suspicious

Potential outcomes

Notation:

- Units of observation $i = 1, \ldots, N$
- Outcome of interest, $Y_i \in \Re$
- Treatment indicator, $T_i \in \{0, 1\}$

For each unit, there are two potential outcomes:

$$Y_i = \begin{cases} Y_i(0) & T_i = 0, \\ Y_i(1) & T_i = 1. \end{cases}$$

Only one of these is observed; the other is counterfactual.

Estimand

We care about the average treatment effect,

$$\tau := \mathbb{E}\left(Y_i(1) - Y_i(0)\right) = \mathbb{E}\left(Y_i(1)\right) - \mathbb{E}\left(Y_i(0)\right)$$

Why can't we just use a difference of means test? This would give us

$$\hat{\tau} \xrightarrow{p.} \mathbb{E}(Y_i(1) \mid T_i = 1) - \mathbb{E}(Y_i(0) \mid T_i = 0),$$

But if there are confounding variables related to both T_i and Y_i,

$$\mathbb{E}(Y_i(1) \mid T_i = 1) \neq \mathbb{E}(Y_i(1))$$
$$\mathbb{E}(Y_i(0) \mid T_i = 0) \neq \mathbb{E}(Y_i(0))$$

Potential outcomes example

Do seatbelts save lives?

Units: Individuals in traffic accidents

Treatment (T_i): Seatbelt use

Outcome (Y_i): Mortality

- $Y_i(0)$: whether i will die if she isn't wearing a seatbelt
- $Y_i(1)$: whether i will die if she is wearing a seatbelt

Potential confounders: Speed at time of the accident

Potential outcomes example

If seatbelt users are slower drivers, a naive difference of means test will overstate the benefits of seatbelt use.

To estimate the average effect of seatbelt use, we should only seatbelt users to non-users who were traveling at similar speeds.

Regression and model dependence

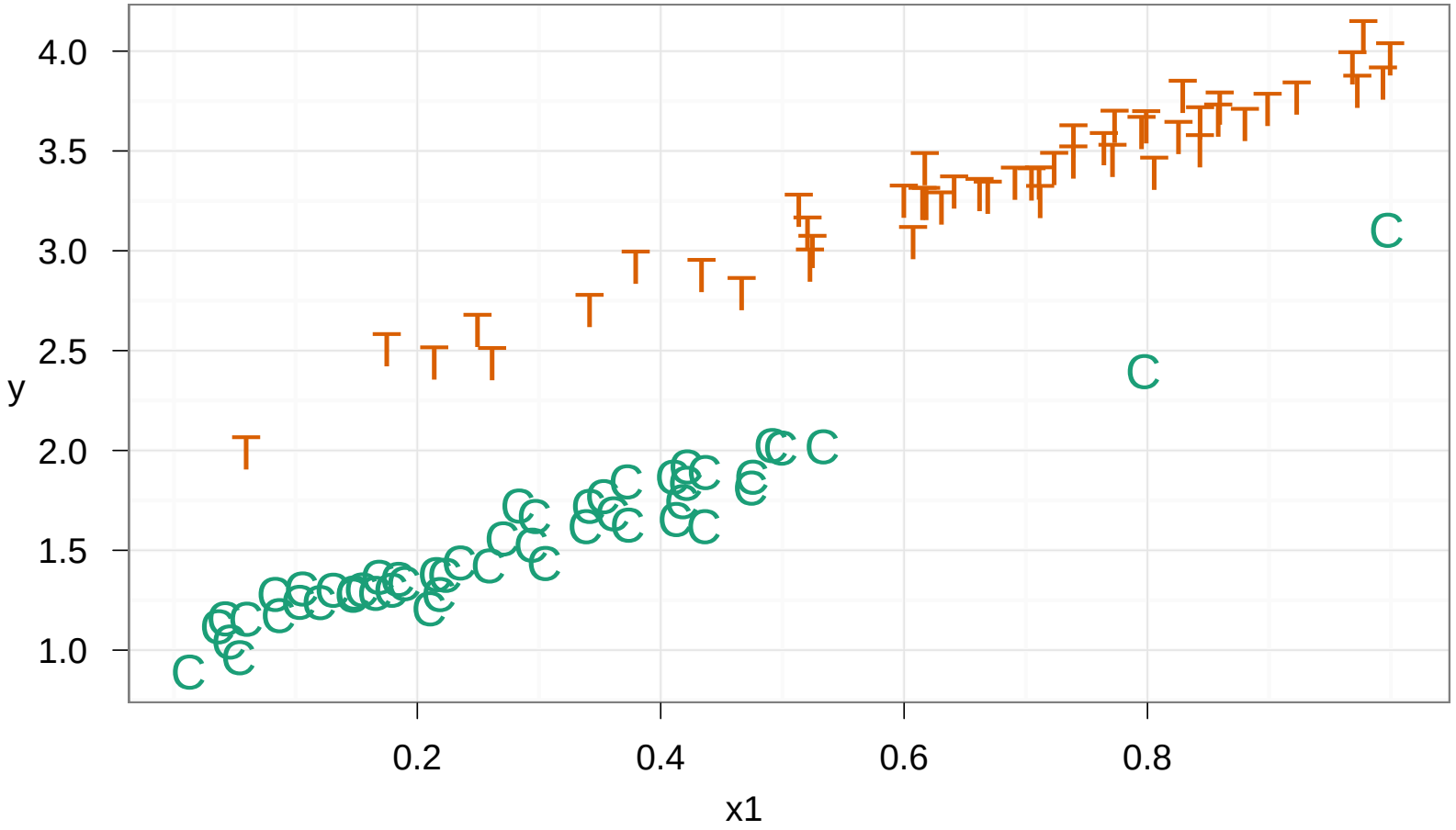
What about linear regression? This will yield an unbiased estimate of τ if

$$Y_i = \alpha + \mathbf{x}_i'\beta + \tau T_i + \epsilon_i$$

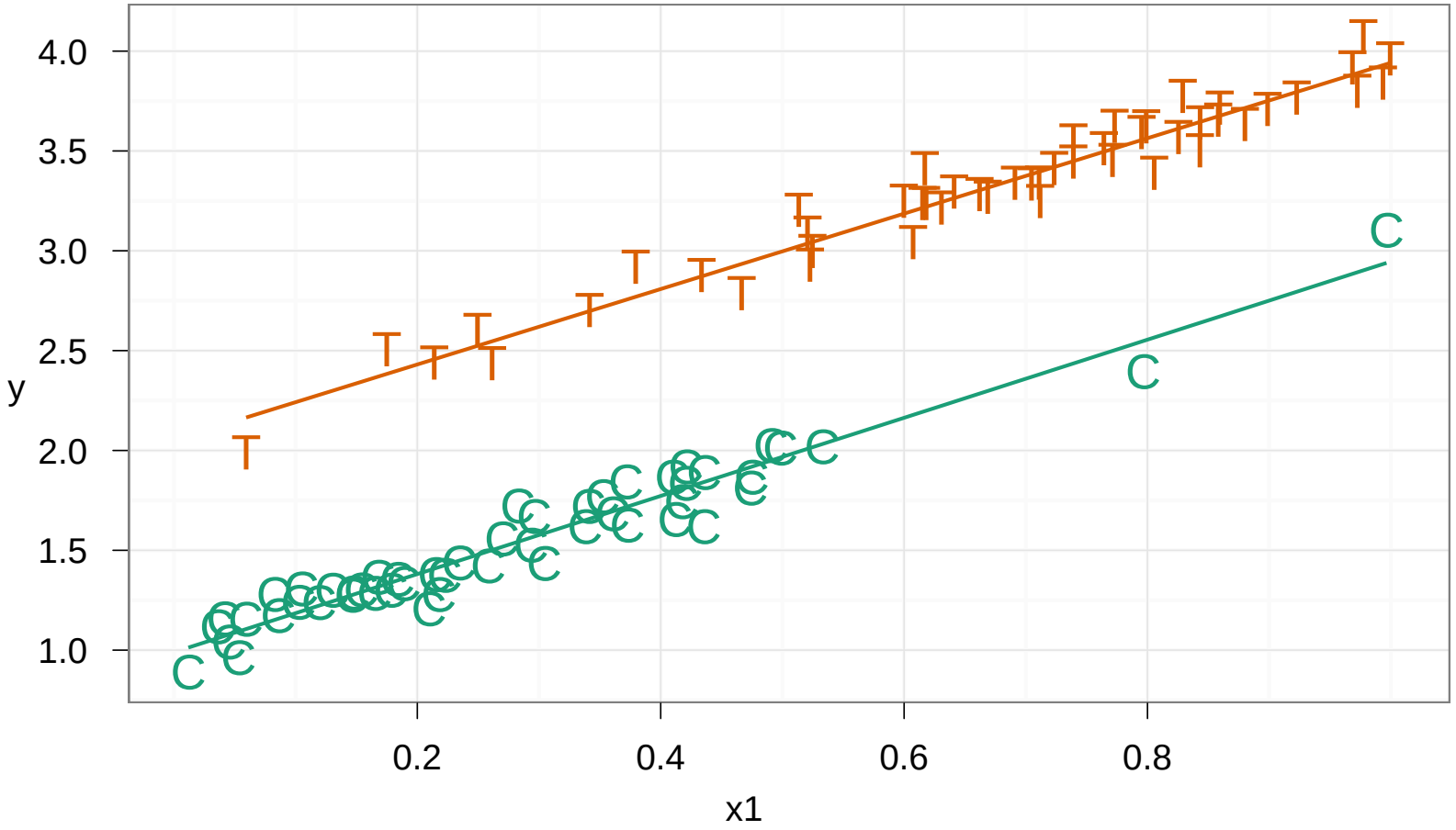
is the true model, but not otherwise.

We want an unbiased estimator of τ that doesn't depend on knowing the true form of the relationship among $\mathbf{X}$, Y and T.

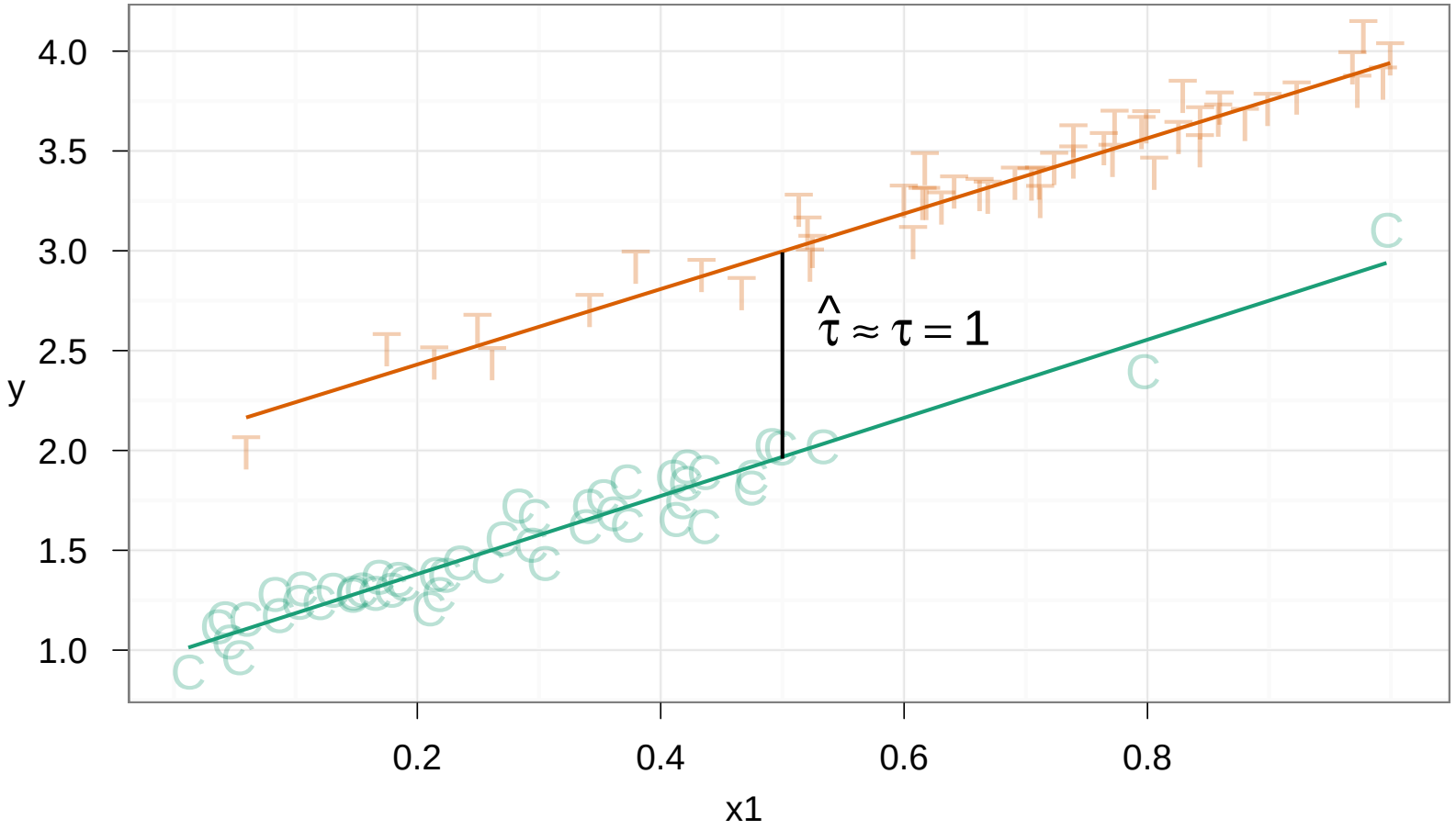
Example where OLS works



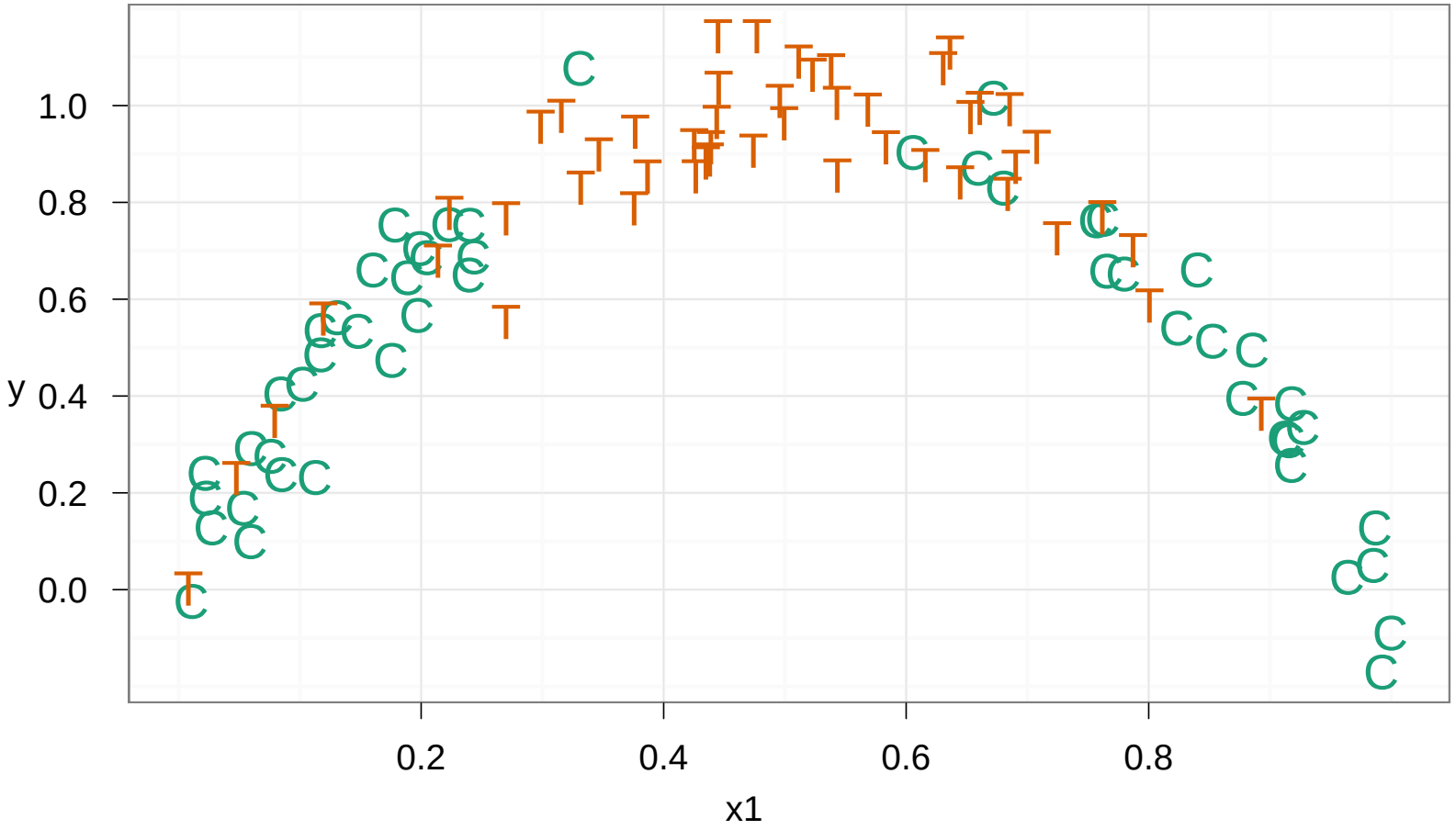
Example where OLS works



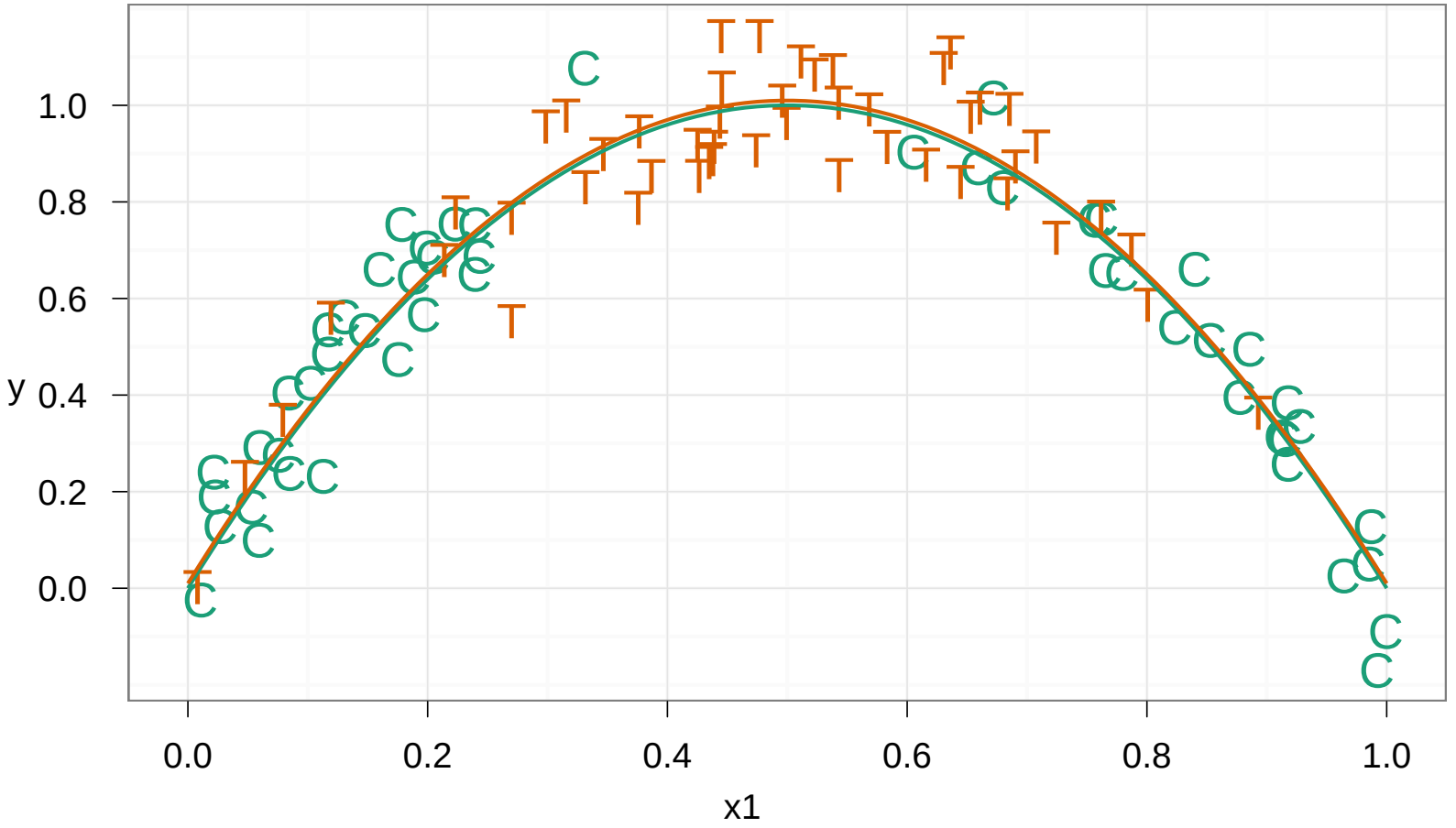
Example where OLS works



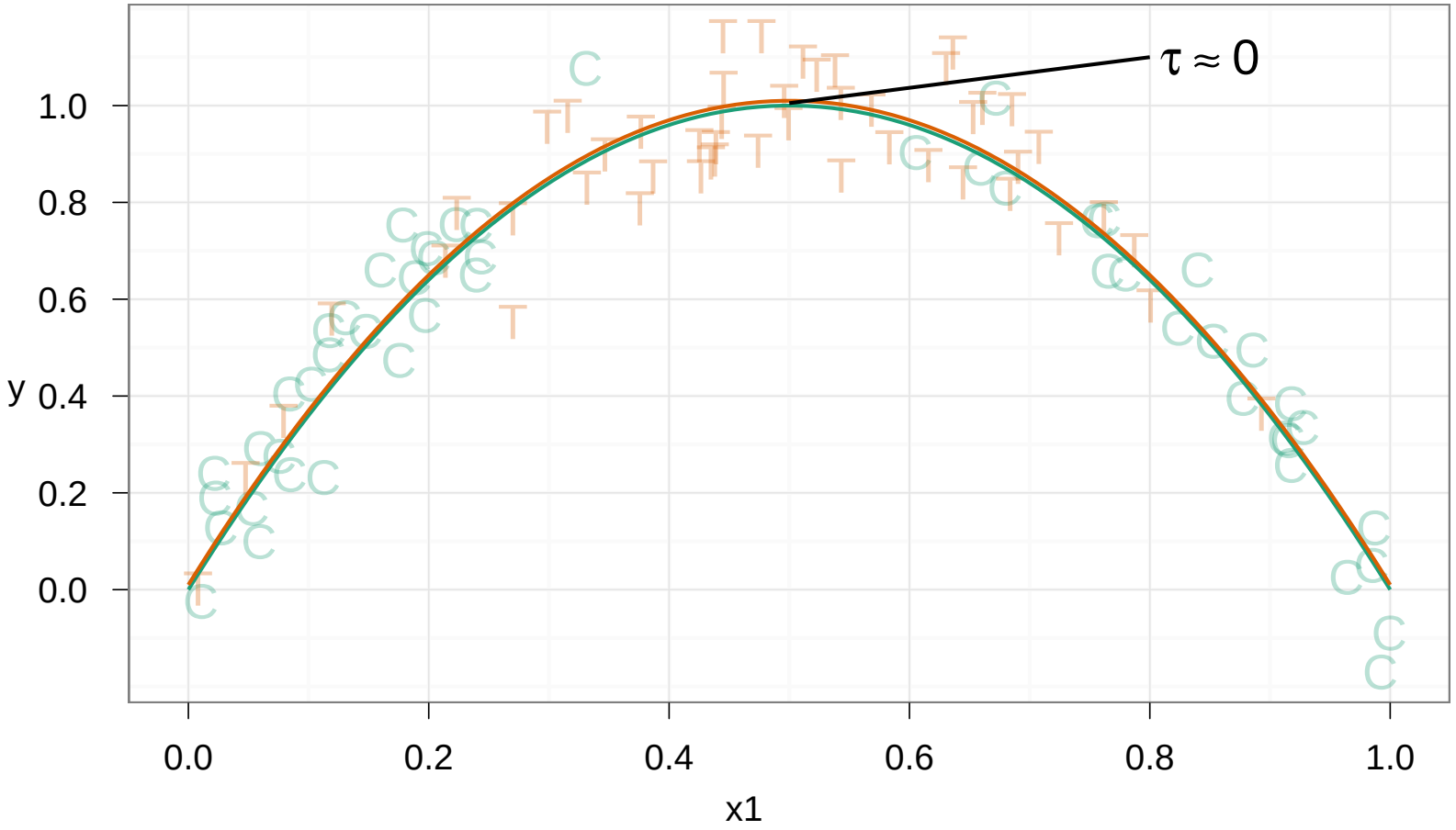
Example where OLS fails



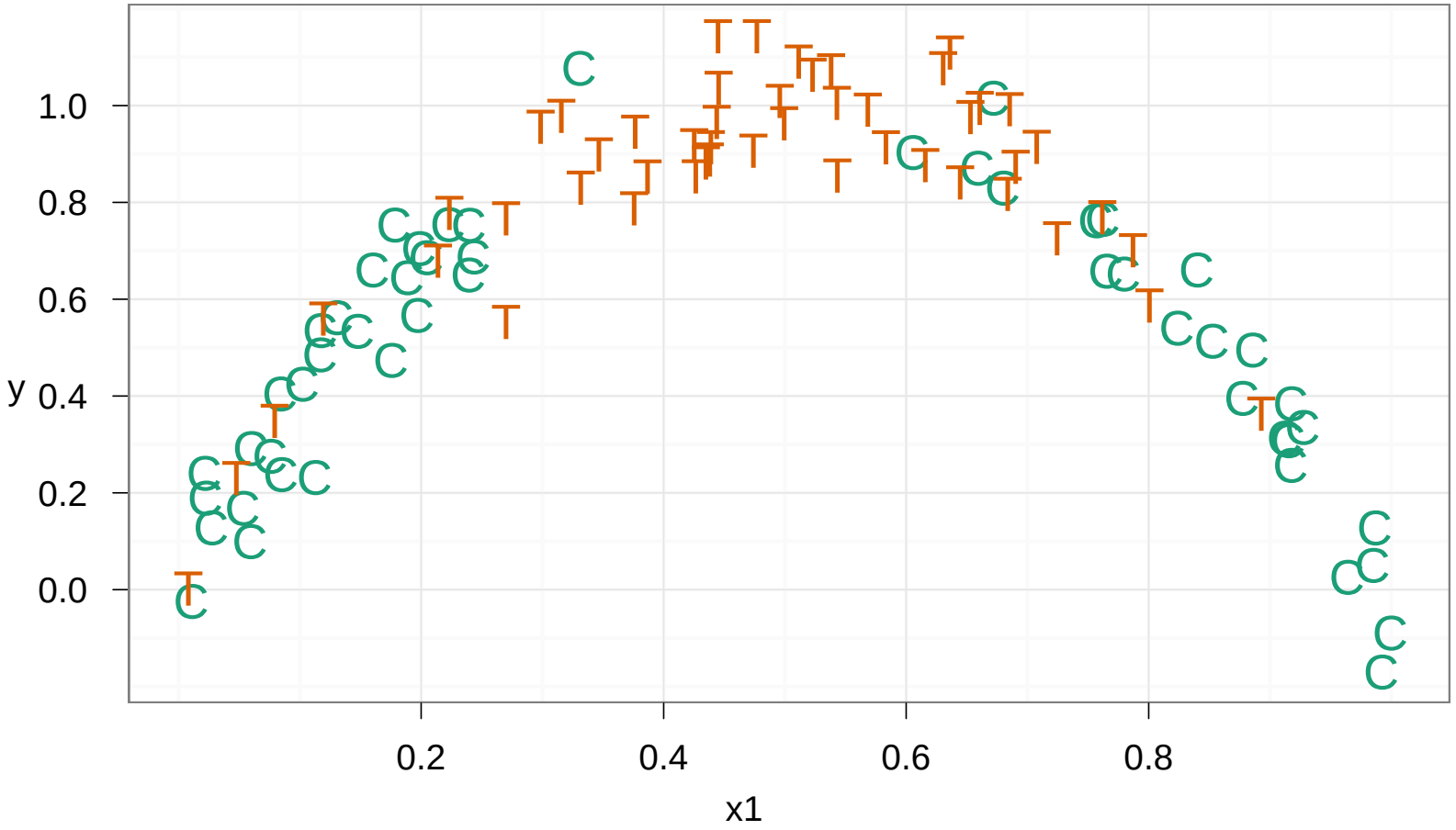
Example where OLS fails



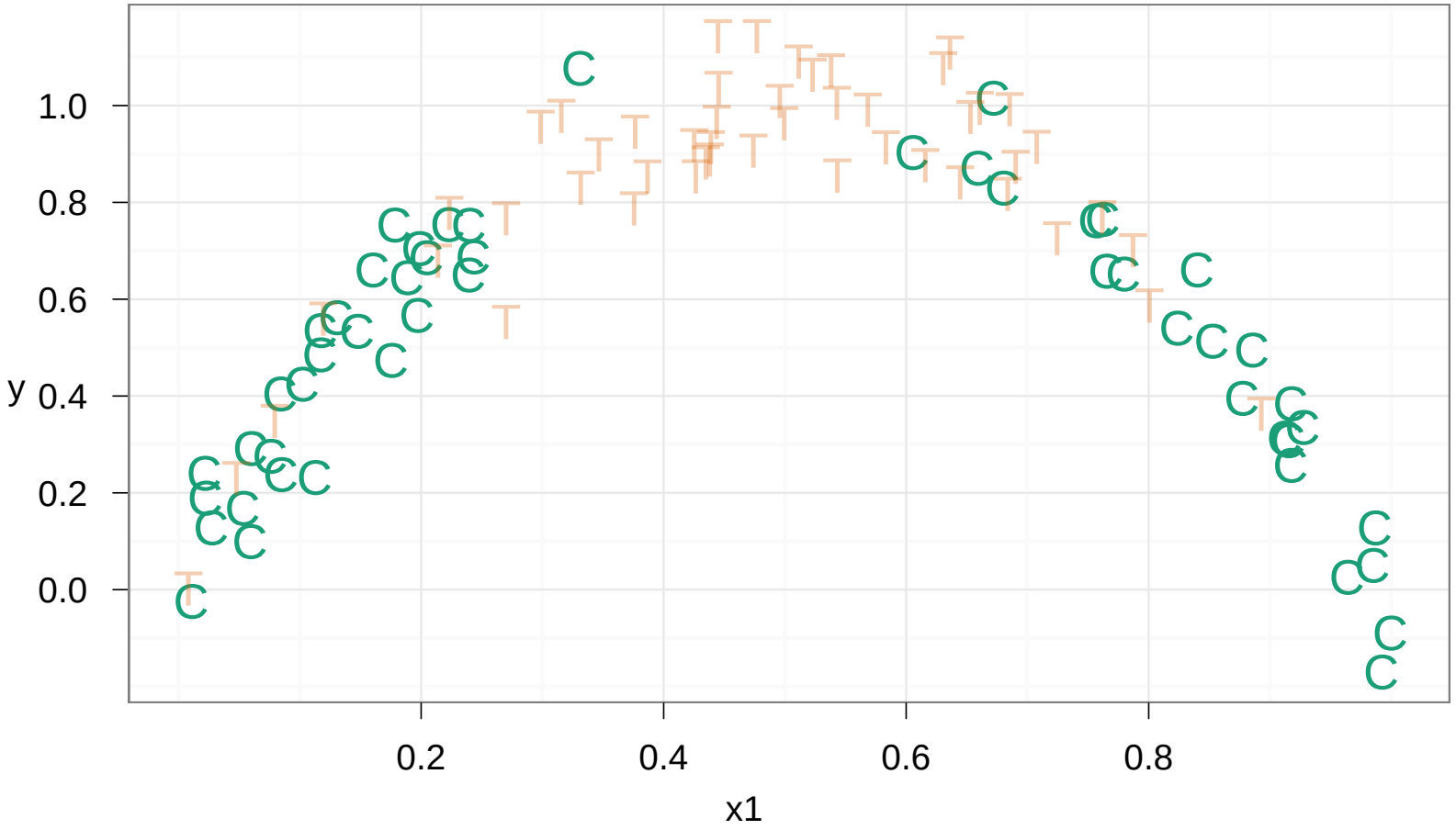
Example where OLS fails



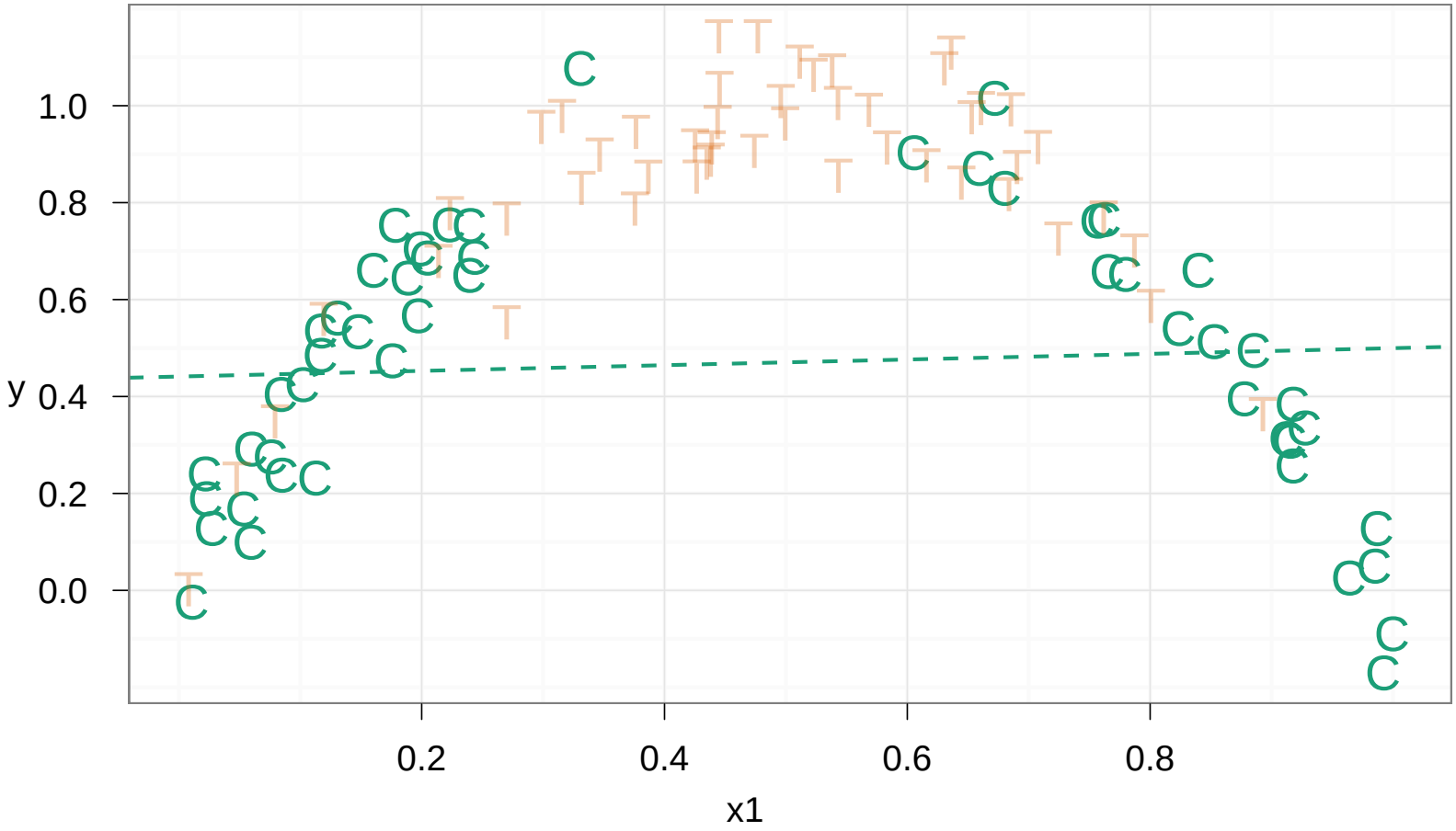
Example where OLS fails



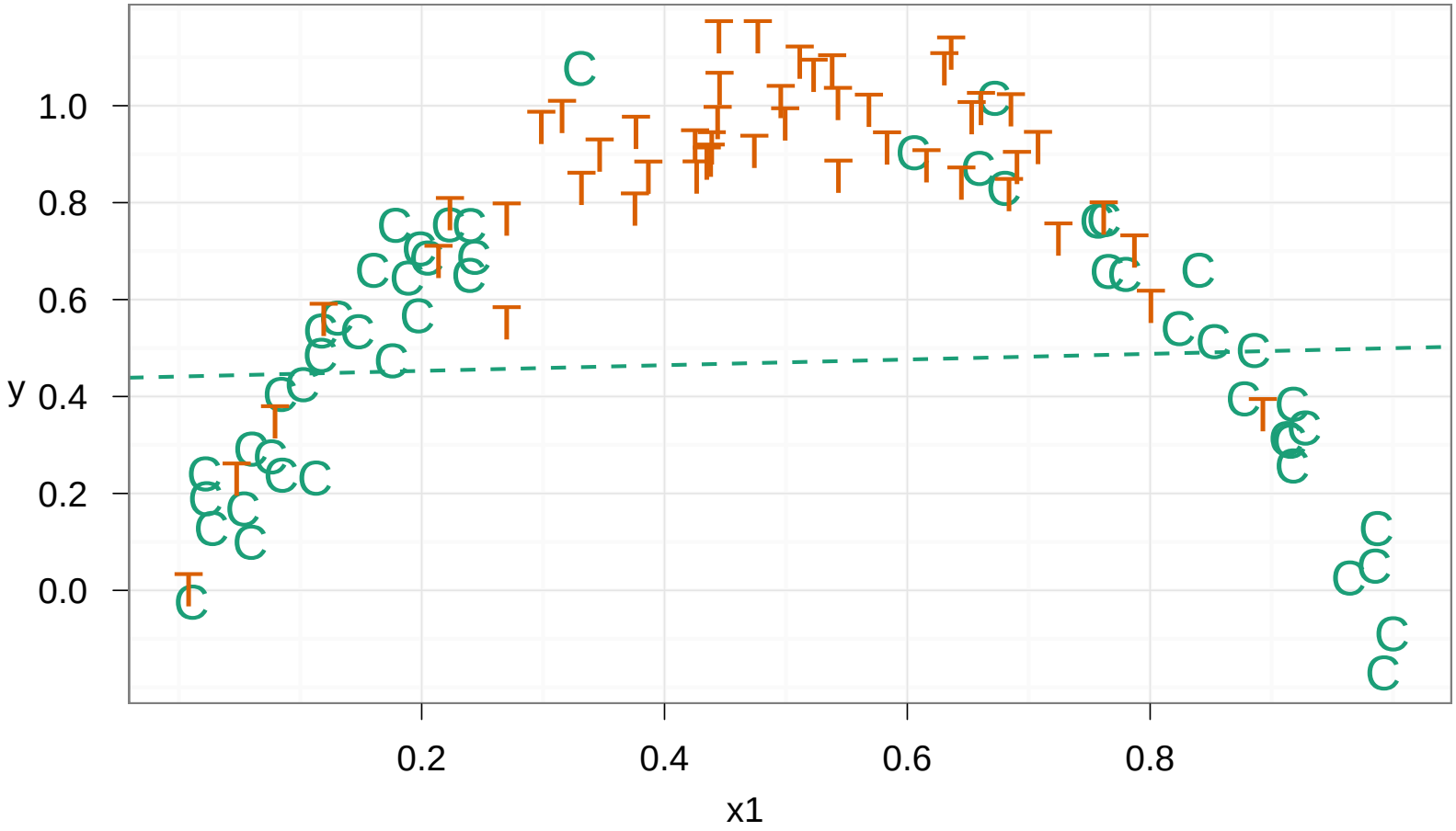
Example where OLS fails



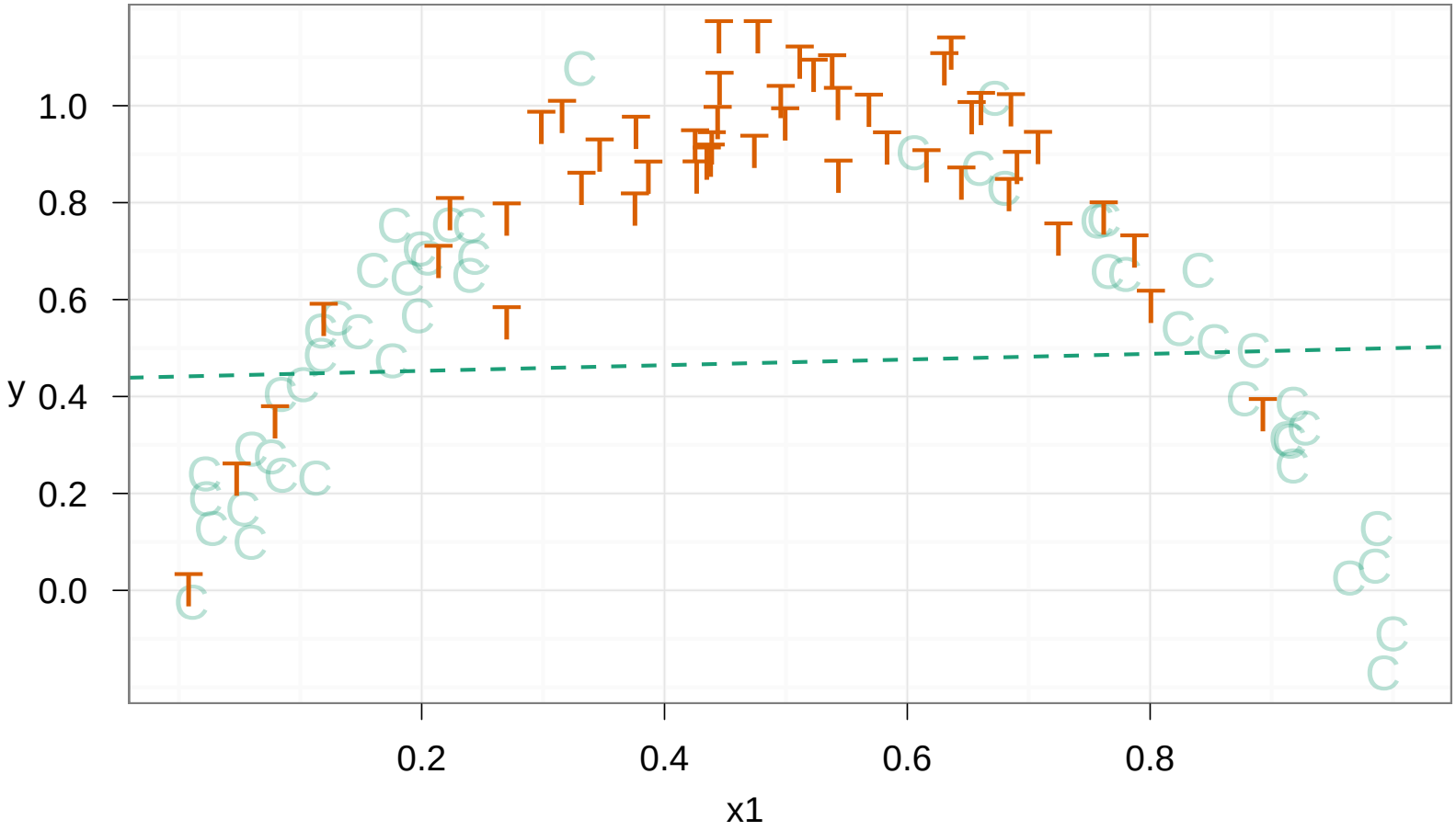
Example where OLS fails



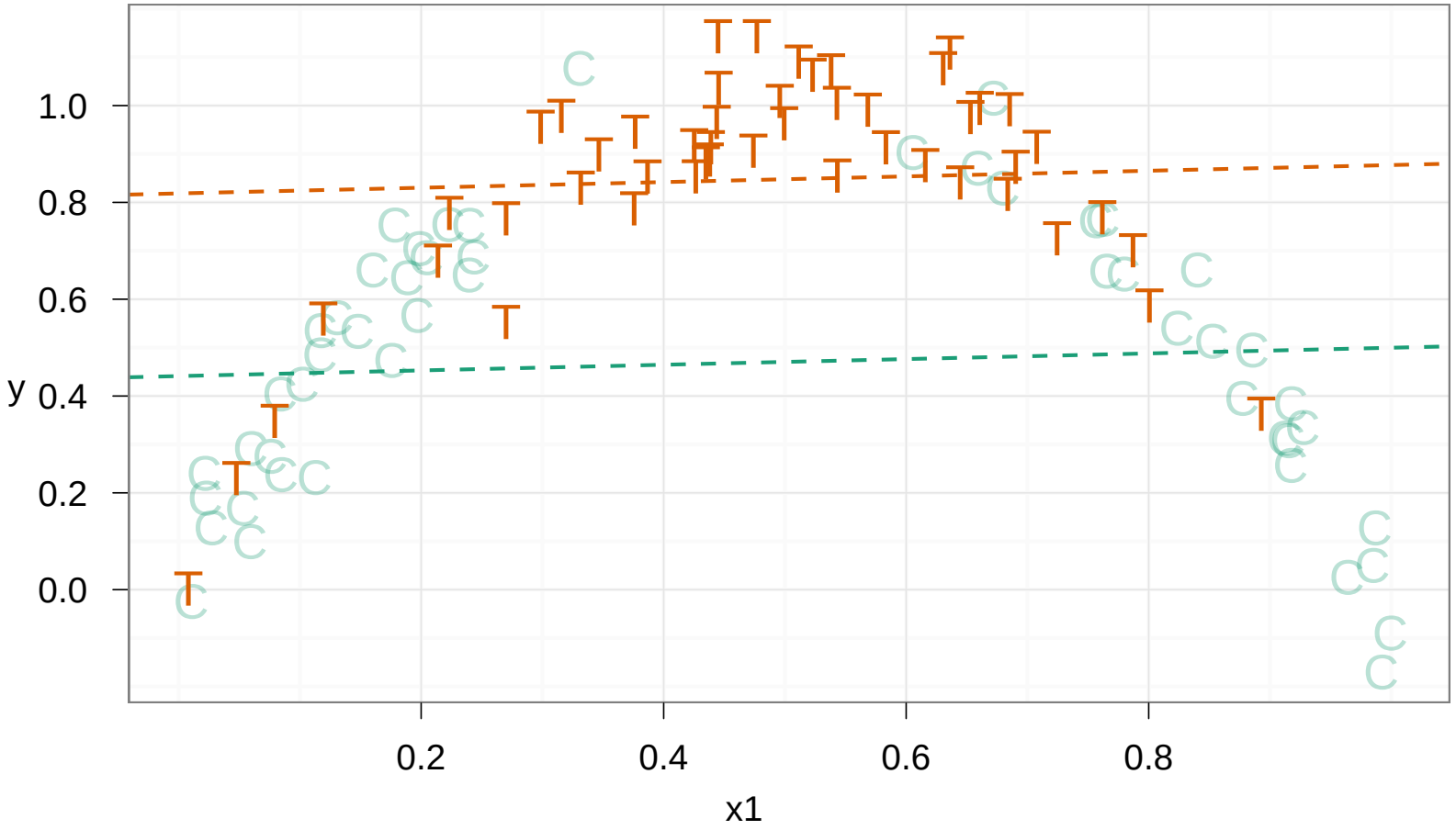
Example where OLS fails



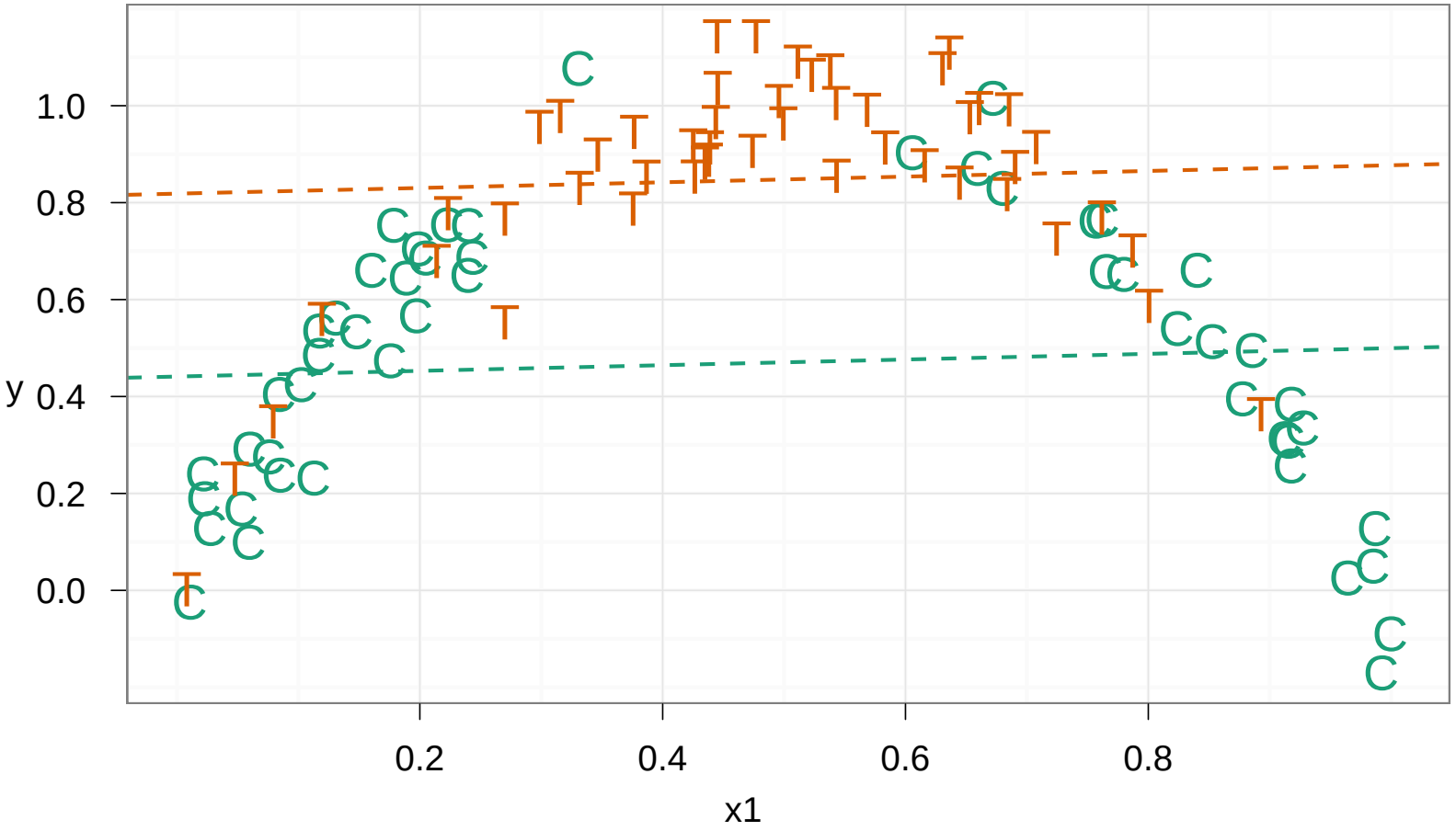
Example where OLS fails



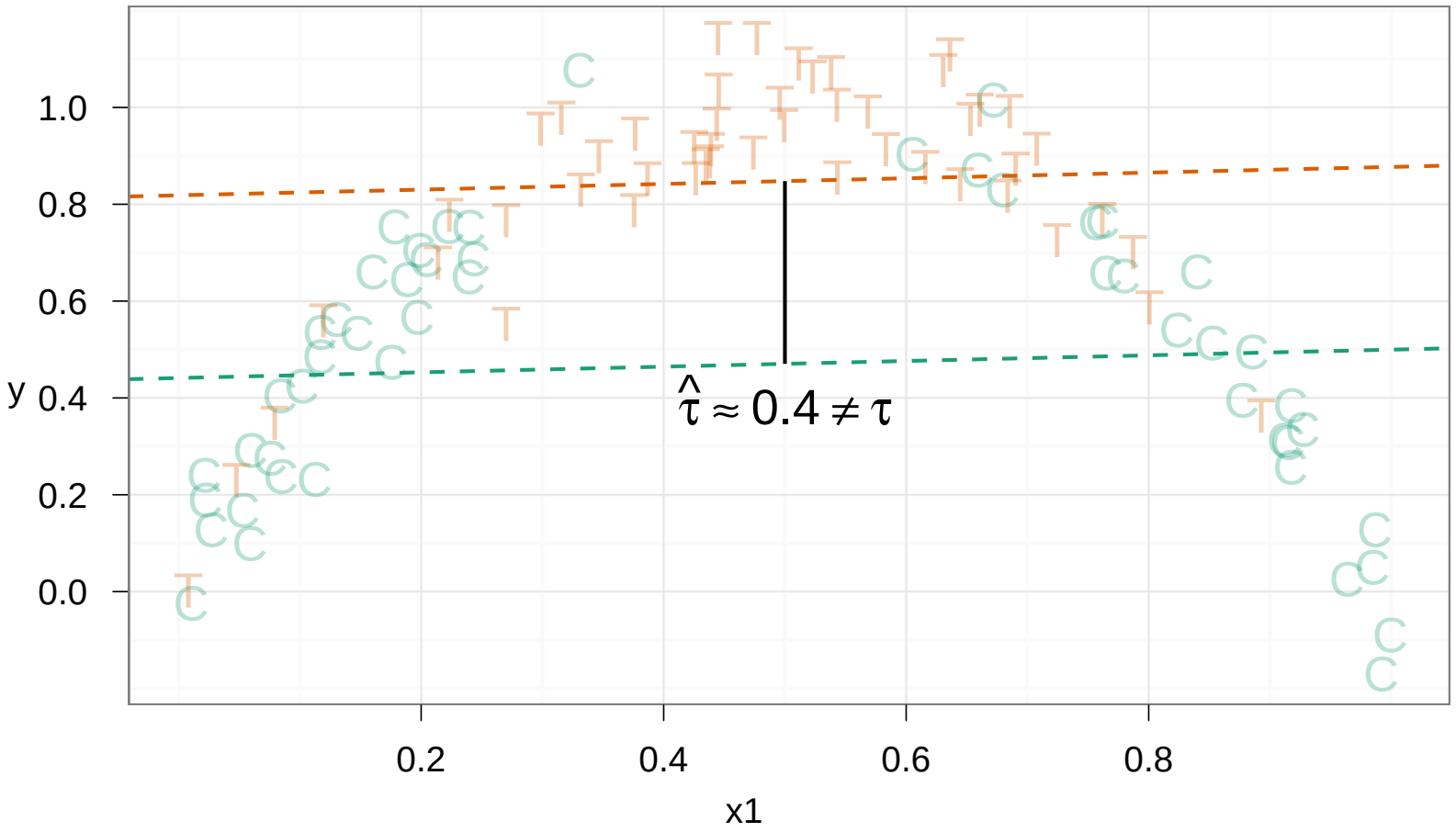
Example where OLS fails



Example where OLS fails



Example where OLS fails



Background assumptions

Strong ignorability.

1. No omitted confounders. This is equivalent to the conditional independence condition $(Y(0), Y(1)) \perp T \,|\, \mathbf{x}$.
2. Positive overlap. For all $\mathbf{x}$, $0 < \Pr(T = 1 \,|\, \mathbf{x}) < 1$.

Stable unit treatment values (SUTVA). Each unit's potential outcomes are independent of whether others are treated and the mechanism of treatment assignment.

Background assumptions

Strong ignorability.

1. No omitted confounders. This is equivalent to the conditional independence condition $(Y(0), Y(1)) \perp T \,|\, \mathbf{x}$.
2. Positive overlap. For all $\mathbf{x}$, $0 < \Pr(T = 1 \,|\, \mathbf{x}) < 1$.

Stable unit treatment values (SUTVA). Each unit's potential outcomes are independent of whether others are treated and the mechanism of treatment assignment.

Background assumptions

Strong ignorability.

1. No omitted confounders. This is equivalent to the conditional independence condition $(Y(0), Y(1)) \perp T \,|\, \mathbf{x}$.
2. Positive overlap. For all $\mathbf{x}$, $0 < \Pr(T = 1 \,|\, \mathbf{x}) < 1$.

Stable unit treatment values (SUTVA). Each unit's potential outcomes are independent of whether others are treated and the mechanism of treatment assignment.

Background assumptions

Strong ignorability.

1. No omitted confounders. This is equivalent to the conditional independence condition $(Y(0), Y(1)) \perp T \,|\, \mathbf{x}$.
2. Positive overlap. For all $\mathbf{x}$, $0 < \Pr(T = 1 \,|\, \mathbf{x}) < 1$.

Stable unit treatment values (SUTVA). Each unit's potential outcomes are independent of whether others are treated and the mechanism of treatment assignment.

Are these assumptions unrealistic?

No more so than the regression assumptions.

OLS assumption		**Matching assumption**
no omitted variables	$\Rightarrow$	no omitted confounders
i.i.d. errors no measurement error	$\Rightarrow$	SUTVA

The two least plausible assumptions are implied by OLS assumptions. Positive overlap isn't, but only thanks to linear extrapolation.

Are these assumptions unrealistic?

No more so than the regression assumptions.

OLS assumption		**Matching assumption**
no omitted variables	$\Rightarrow$	no omitted confounders
i.i.d. errors no measurement error	$\Rightarrow$	SUTVA
???	$\not\Rightarrow$	positive overlap

The two least plausible assumptions are implied by OLS assumptions. Positive overlap isn't, but only thanks to linear extrapolation.

Philosophy of matching

Two related goals:

1. Only compare similar cases
2. Don't extrapolate outside the data

Statistical version of Mill's method: To determine whether T causes Y, examine cases that are identical on all but T, and see if Y differs.

What to match on

You only need to match on true confounders, which affect *both* T and Y.

Do match on pre-treatment variables, exogenous variables whose values are determined before treatment assignment, T_i.

Don't match on intervening variables, endogenous variables whose values are determined after treatment assignment.

Matching and selection issues

Self-selectivity bias: You want to estimate the effect of T on Y, but individuals self-select into T.

- Example: job training program
- Matching *can* help you

Sample selection bias: You want to estimate the effect of X on Y, but you only observe X and Y for individuals where $T_i = 1$.

- Example: effect of education on wages (only have those who entered workforce)
- Matching *can't* help you

Exact matching

Procedure:

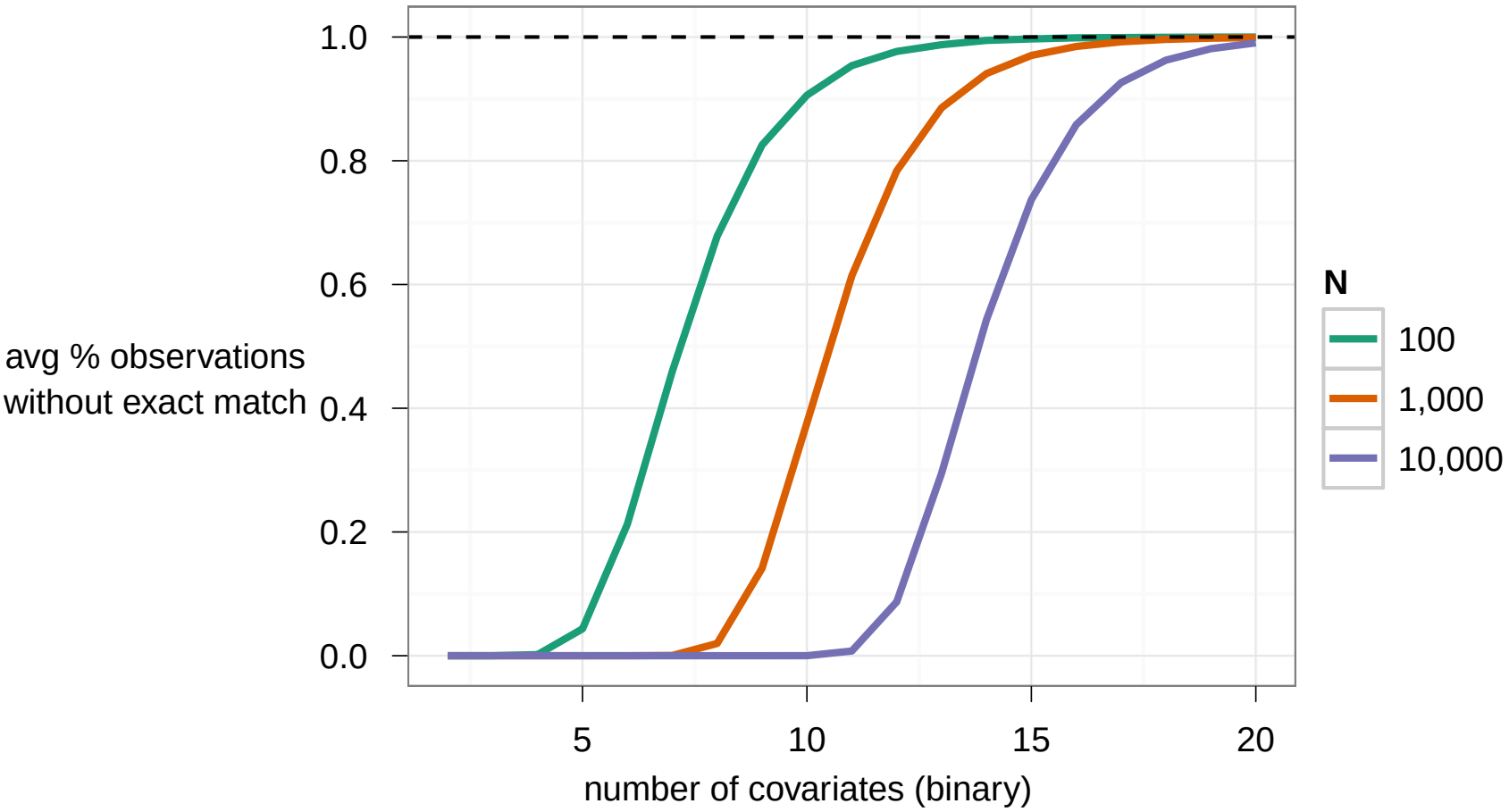
1. For each i such that $T_i = 1$:
   1. Set $\pi(i) = j$, for some j such that $T_j = 0$ and $\mathbf{x}_j = \mathbf{x}_i$.
   2. If no such j exists, remove i from the dataset.
2. Estimate the ATE as the difference of means,

$$\hat{\tau} = \underbrace{\overline{\{Y_i : T_i = 1\}}}_{\substack{\text{avg outcome in} \\ \text{matched treated obs}}} - \underbrace{\overline{\{Y_{\pi(i)} : T_i = 1\}}}_{\substack{\text{avg outcome in} \\ \text{matched control obs}}}$$

Problems:

- Next to impossible with continuous variables...
- ...or more than a few categorical variables (curse of dimensionality)

Curse of dimensionality



Nearest-neighbor matching

Suppose there is only one confounding variable, X.

Procedure:

1. For each i such that $T_i = 1$, set

$$\pi(i) = \underset{j:T_j=0}{\operatorname{argmin}} \|X_i - X_j\|.$$

2. Estimate the ATE the same as in exact matching.

Problem: When do you only have one confounding variable?

Mahalanobis distance matching

Let S be the sample covariance matrix of $\mathbf{X}$. The Mahalanobis distance between rows $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)S^{-1}(\mathbf{x}_i - \mathbf{x}_j)'}.$$

Procedure: Nearest-neighbor matching on $d_M(\cdot, \cdot)$.

Problems:

- No reason to prefer Mahalanobis distance over other metrics, especially for non-Gaussian data
- In high dimensions, closest match may still be distant

Propensity score matching

For each unit, let the propensity score be its ex ante probability of treatment, $e_i = \Pr(T_i = 1) = \Pr(T = 1 \mid \mathbf{x}_i)$.

Procedure:

1. Estimate $\hat{e}_i$ via logistic regression or a similar procedure
2. Perform nearest-neighbor matching on the distance between propensity scores

Propensity score matching

This is the most popular procedure in political science and economics because it's easy and appears to avoid the curse of dimensionality.

Problems:

- Properties are proven for exact matching on e_i, not nearest-neighbor matching on $\hat{e}_i$
- How to select a model to estimate propensity scores?

Many, many more. . .

- Exact matching on propensity score deciles
- Mahalanobis distance matching within propensity score calipers
- Matching with replacement, without replacement, with multiple matches per unit, etc.
- Genetic matching (Sekhon)
- Coarsened exact matching (King), combines elements of exact and Mahalanobis distance matching

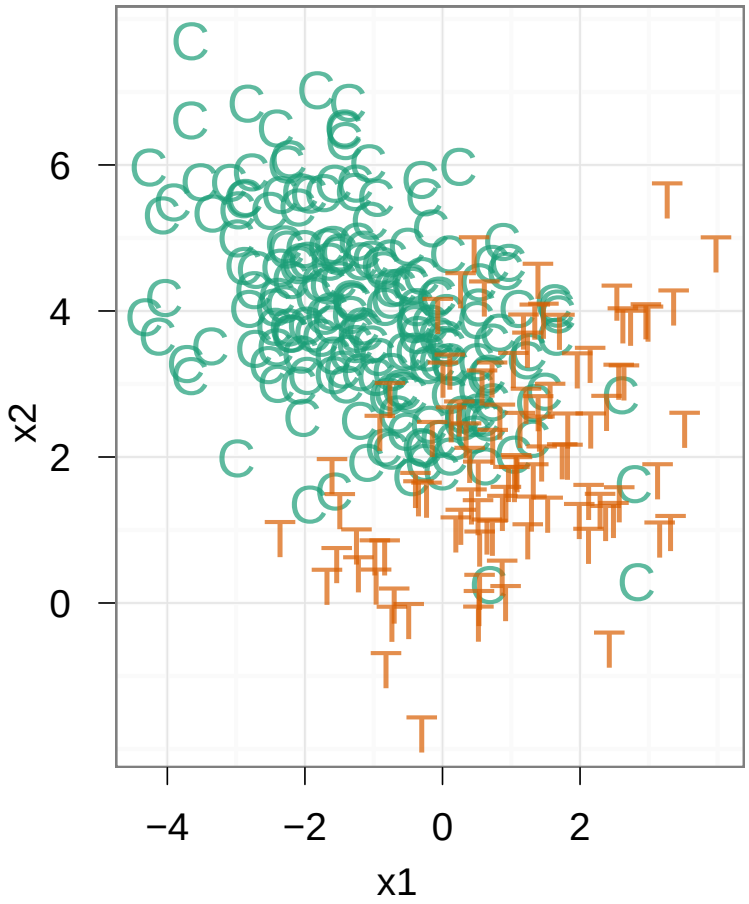
Comparing matched samples

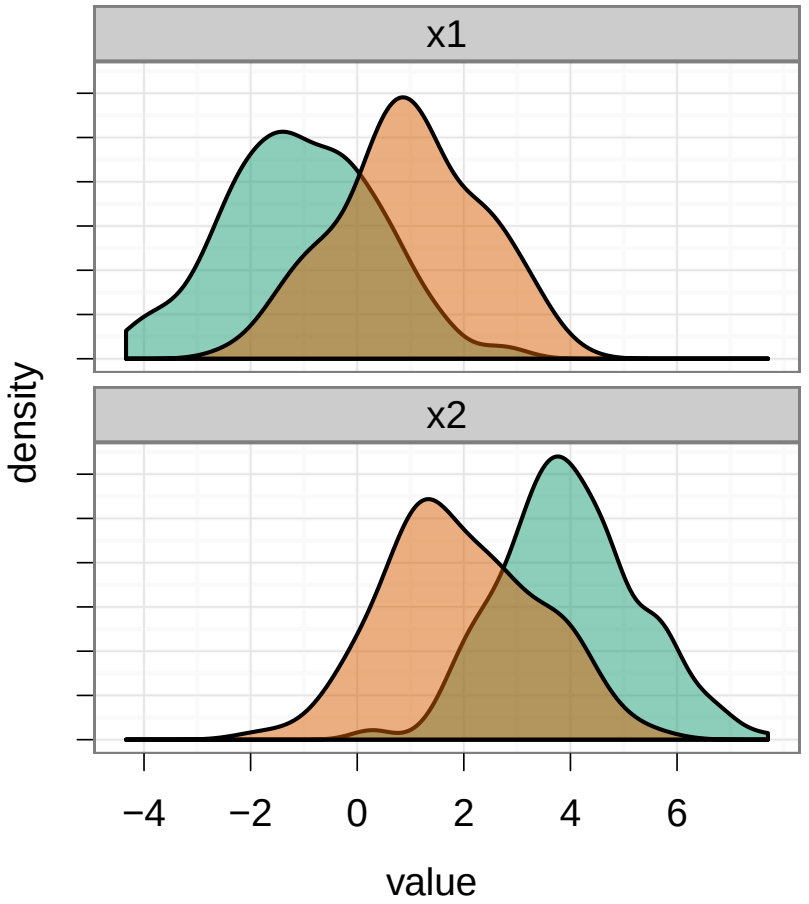
How do we know which matching method is best for a particular sample?

The standard recommendation is to achieve maximal balance: the distribution of $\mathbf{x}$ in the treated group should be approximately the same as in the matched control group.

Balance: An illustration

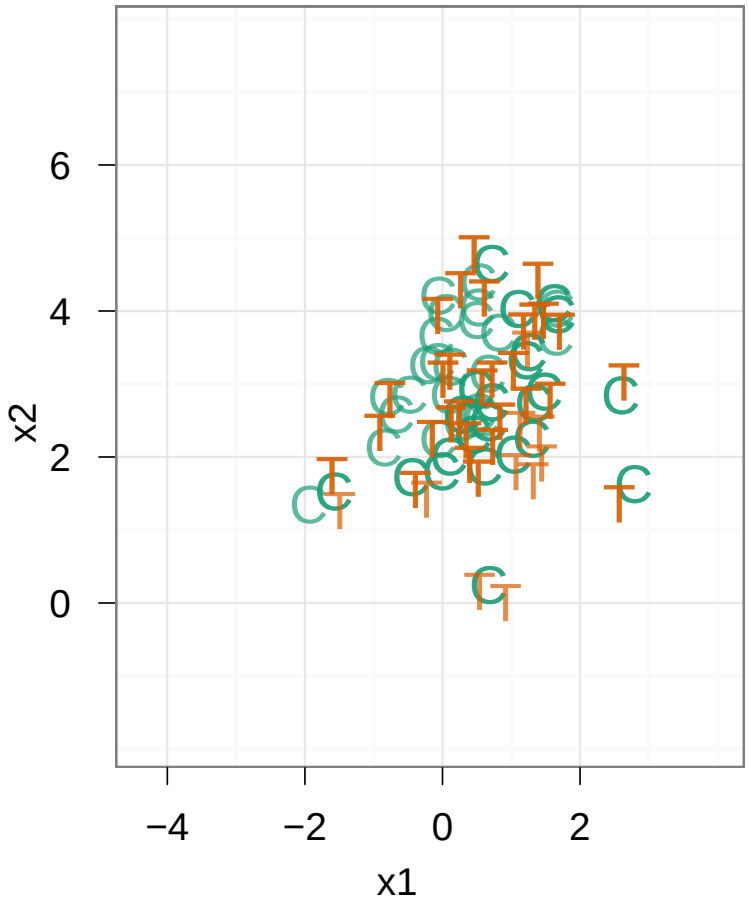
Before matching

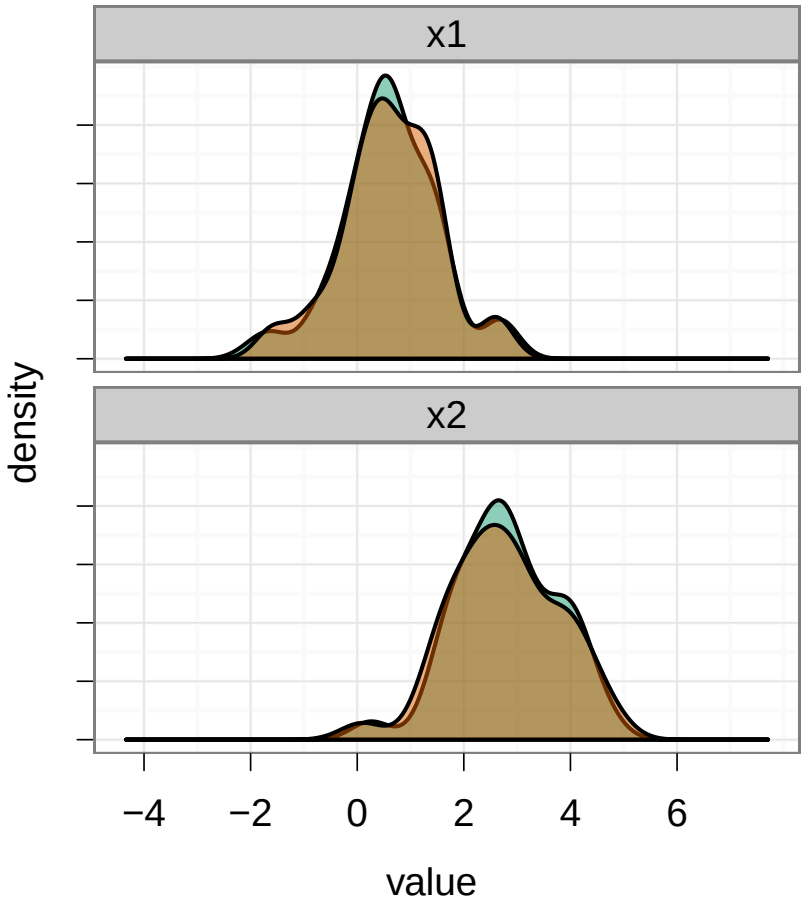




Balance: An illustration

After matching





Balance checks

The general idea:

- Match using many different methods or propensity score specifications
- Check balance on covariates in each matched sample using t tests, Kolmogorov-Smirnov tests, or other metrics
- Use the matched sample that does best

CEM and genetic matching both automate this procedure, in different ways.

R software for matching

The main package is Jas Sekhon's `Matching`:

- Matching methods implemented: exact, Mahalanobis distance, propensity score, genetic
- Numerous options
  - 1:M matching
  - With or without replacement
  - Weighted matching
  - Matching within calipers
- Balance checking
- Standard error estimation

Summing up

Matching is just a nonparametric estimator of a population effect — generically less biased, but less efficient, than the regression coefficient.

So if your goal is to estimate a population effect, the choice between matching and regression comes down to

- how much data you have
- your belief in linearity of the relationship
- your loss function for bias vs. variance

Summing up

- Matching isn't scary
- Matching isn't evil
- Matching doesn't require strange assumptions
- Matching isn't that different from what "we" do (in practice)

Causal inference

- Paul Rosenbaum (1984), "From Association to Causation in Observational Studies," *JASA*.
- Paul Holland (1986), "Statistical and Causal Inference," *JASA* (and responses).
- Judea Pearl (1995), "Causal Diagrams for Empirical Research," *Biometrika* (and responses).
- Donald Rubin (2007), "The Design *versus* the Analysis of Observational Studies for Causal Effects," *Statistics in Medicine*.
- Angrist and Pischke (2008), *Mostly Harmless Econometrics*.
- Judea Pearl (2009), *Causality*.

Matching

- Donald Rubin (1973), "Matching to Remove Bias in Observational Studies," *Biometrics*.
- Heckman, Ichimura and Todd (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Rev. Econ. Stud*.
- Guido Imbens (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Rev. Econ. & Stats*.
- Alberto Abadie and Guido Imbens (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*.
- Jas Sekhon (2009), "Opiates for the Matches: Matching Methods for Causal Inference," *Ann. Rev. Poli. Sci*.

Propensity scores

- Paul Rosenbaum and Donald Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*.
- Rajeev Dehejia and Sadek Wahba (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *JASA*.
- By political scientists:
  - Kosuke Imai (2004), "Do Get-Out-The-Vote Calls Reduce Turnout?" *APSR*.
  - Kosuke Imai and David van Dyk (2004), "Causal Inference with General Treatment Regimes," *JASA* (methods for non-binary treatments).
  - Ho, Imai, King and Stuart (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *PA*.