Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 11: Causal inference

Feb 21, 2017

Linear/logistic regression

Logistic regression

$$P(y = 1 \mid x, \beta) = \frac{\exp\left(\sum_{i=1}^{F} x_i \beta_i\right)}{1 + \exp\left(\sum_{i=1}^{F} x_i \beta_i\right)}$$

Linear regression

$$y = \sum_{i=1}^{F} x_i \beta_i + \varepsilon$$

x = feature vector

Feature	Value
follow clinton	0
follow trump	0
“benghazi”	0
negative sentiment + “benghazi”	0
“illegal immigrants”	0
“republican” in profile	0
“democrat” in profile	0
self-reported location = Berkeley	1

β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

$$\frac{P(y \mid x, \beta)}{1 - P(y \mid x, \beta)} = \exp\left(x_0\beta_0\right)\exp\left(x_1\beta_1\right)$$

Let's increase the value of x by 1 (e.g., from 0 → 1)

$$\exp(x_0\beta_0)\exp((x_1 + 1)\beta_1)$$

$$\exp(x_0\beta_0)\exp(x_1\beta_1 + \beta_1)$$

$$\exp(x_0\beta_0)\exp\left(x_1\beta_1\right)\exp\left(\beta_1\right)$$

exp(β) represents the factor by which the **odds** change with a 1-unit increase in x

$$\frac{P(y \mid x, \beta)}{1 - P(y \mid x, \beta)}\exp\left(\beta_1\right)$$

Prediction vs. Understanding

- Two main uses of statistical models:
- Prediction: inferring the most likely values (+ prediction intervals) for data where you don't know the answer
- Understanding: estimating the relationship between a predictor variable and some outcome (+ quantifying uncertainty about that relationship)

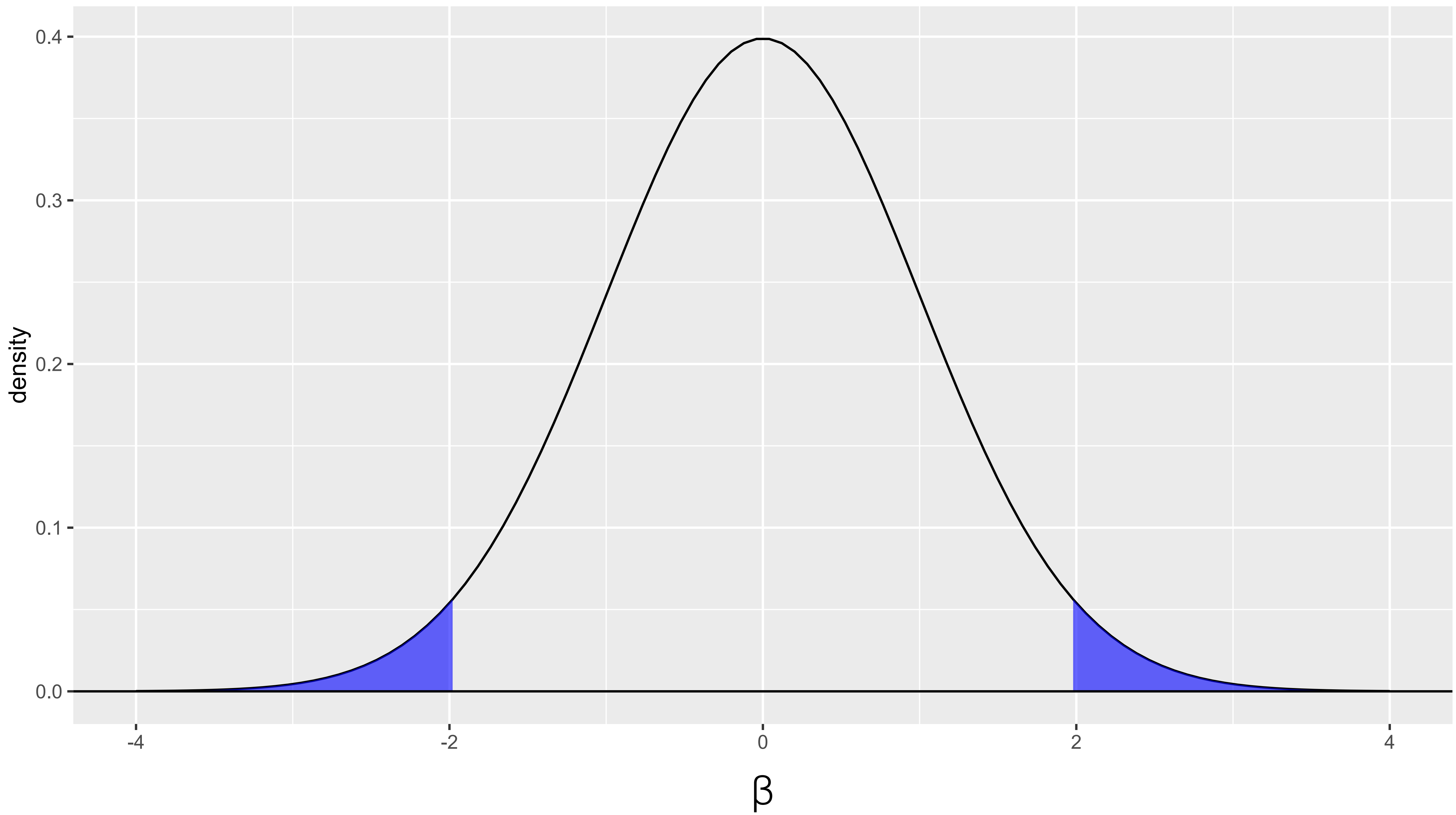
Significance

- When we estimate coefficients in linear/logistic regression, we do so from a sample. Different samples can lead to variability in our estimates.

- We can assess how significant is the relationship between a predictor and its response with a hypothesis test.

- Null hypothesis: All $\beta = 0$.

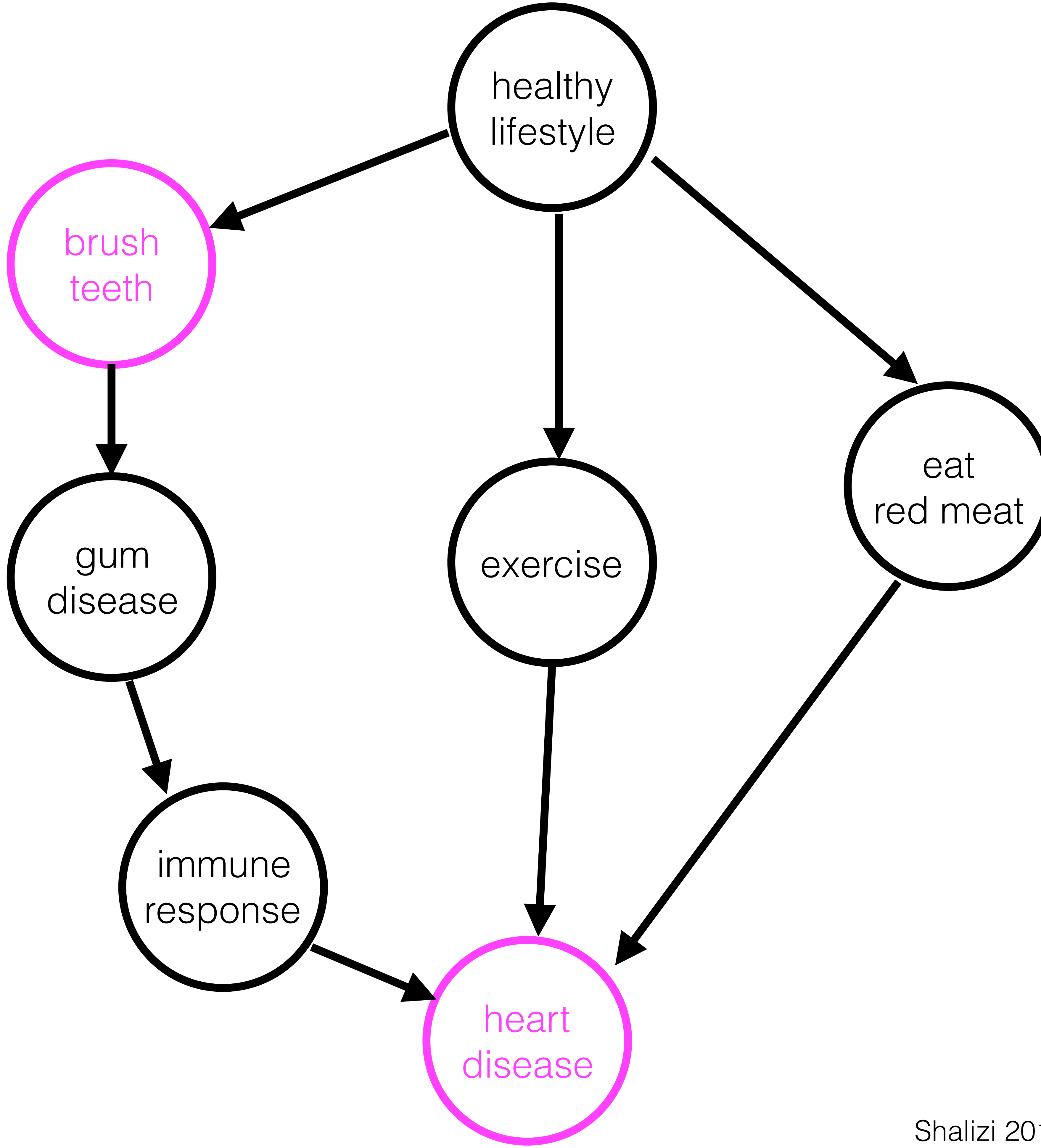
Significance



Correlation vs. Causation

- We want to understand the causal relationship of a treatment T on some outcome Y

Treatment	Outcome
take a drug	cured of disease
graduate high school	earnings
cast John Goodman	box office
living in Berkeley	political preference



Shalizi 2014

Terminology

- Treatment. $T(0)$, $T(1)$. The predictor variable whose causal relationship we're interested in.

- Potential outcomes. $Y{=}0$, $Y{=}1$. The dependent variable.

- We're interested in the causal relationship between the treatment T and the outcome Y.

Counterfactual

- John doesn't brush his teeth (T=0) and developed heart disease (Y=1). What would have happened if he did brush his teeth (T=1)?

Fundamental problem

- For any data point, we only ever get to observe one outcome. We never observe the counterfactual.

Treatment	Outcome
take a drug	cured of disease
graduate high school	earnings
cast John Goodman	box office
living in Berkeley	political preference

β = coefficients

With linear/logistic regression, we can assess the statistical significance of the effect of the features (i.e., with hypothesis test that $\beta \neq 0$)

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
***“illegal immigrants”	8.7
***“republican” in profile	7.9
***“democrat” in profile	-3.0
*self-reported location = Berkeley	-1.7

Observational data

- A survey of the political affiliation of Berkeley residents is observational data
    - the independent variable (living in Berkeley) is not under our control
- Tweets, books, surveys, the web, the census etc. — is all observational.

Observational data

- Hypothesis tests for observational data assess the relationship between variables but don't establish causality.

- Example: if we intervened and relocated someone to Berkeley, would they become liberal?

Experimental data

- Data that allows you to perform an intervention and determine the value of some variable
    - Clinical data: treatment vs. placebo
    - Web design: one of two homepage designs
    - Political email campaigns: one of two (differently worded) solicitations

Experimental data

- A potential confound exists if any other variable is correlated with your intervention decision:

- e.g., users volunteering to receive a drug (and not the placebo)

Randomization experiments

- Users are randomly assigned an outcome (which web page), which allows us to better establish causality

- A/B testing = significance test in randomized experiment with two outcomes

Randomization experiments

- We can run a standard regression, but now if the β_{design_A} is significant, we can interpret it causally.

	user 1	user 2
age	?	?
prior visit	1	0
gender	?	?
design A	1	0
y	$37	$16

Randomization experiments

- By randomly assigning the treatment, we are ensuring that its value is uncorrelated with any other variable.

	user 1	user 2
age	?	?
prior visit	1	0
gender	?	?
design A	1	0
y	$37	$16

Causal inference

If we can ensure that no other variables are correlated with the treatment, we can interpret its effect on an outcome causally.

Observational data

- With randomized experiments, we can perform an intervention, and set the value of a treatment for a given data point.
- With observational data, we can't intervene.
- Instead, we believe there is a randomization experiment lurking in the data; we just need to find it.

- Estimating the effect of graduating high school on future earnings [Angrist and Krueger 1991; Esarey 2015]
- Use census data (= observational)

years of school	≥ 12 years?	weekly earnings
12	1	$158
15	1	$151
7	0	$197
16	1	$217
18	1	$177

Linear regression

$$y = \sum_{i=1}^{F} x_i \beta_i + \epsilon$$

x

graduate high school
1
1
0
1
1

β(graduating high school) = .401

= 1.5 times increase in salary

y

log(weekly earnings)
5.062
5.014
5.283
5.378
5.179

More features

graduate	race	y.o.b.	married	metro area	$$$
1	0	1927	1	1	980
1	1	1921	1	0	312
1	0	1923	0	0	77
1	0	1927	0	1	95
1	1	1928	1	1	123
0	0	1924	1	1	150

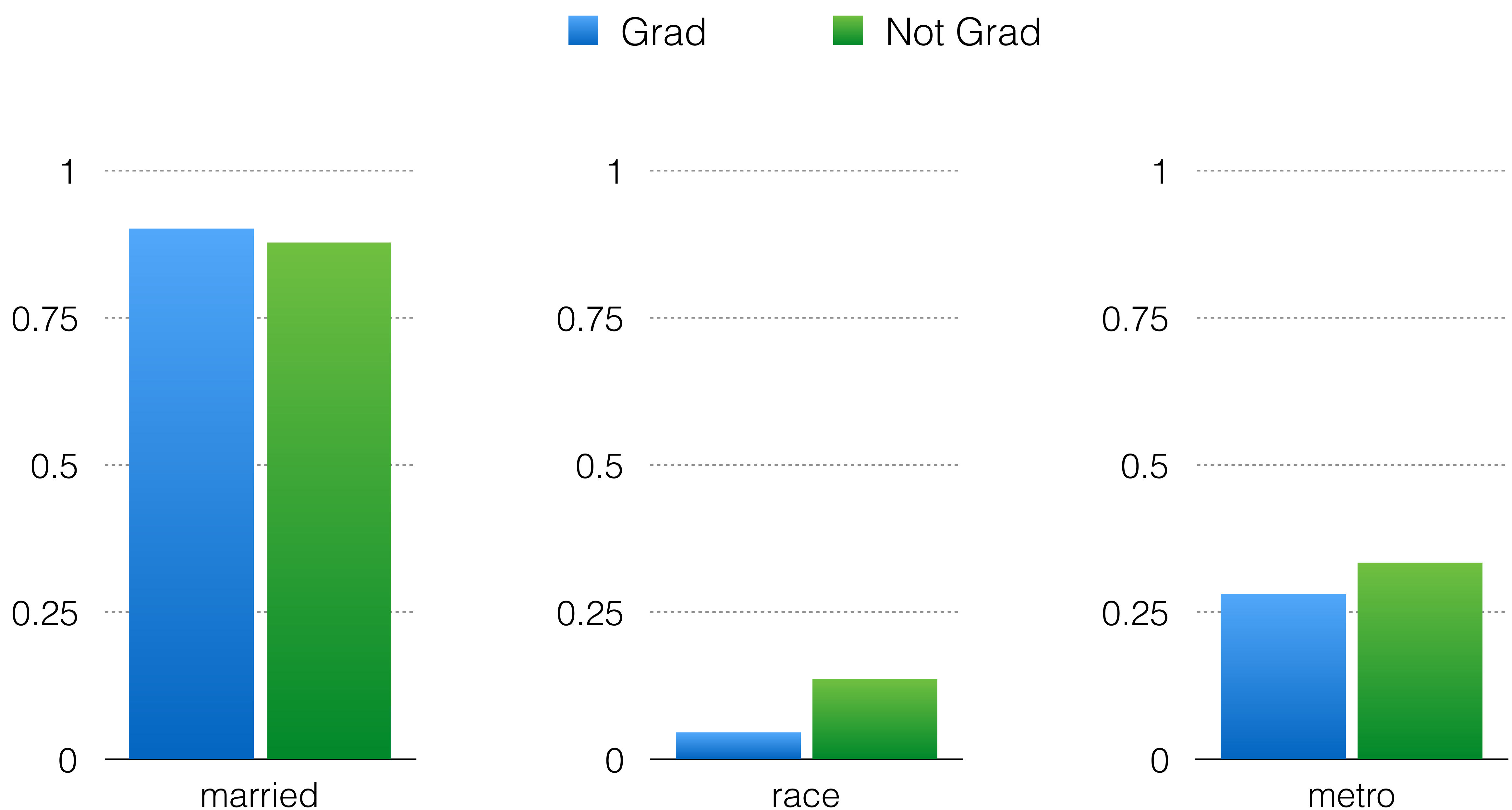
$$y = \sum_{i=1}^{F} x_i \beta_i + \epsilon$$

	β	exp(β)	\$200
graduate	0.35	1.42	\$284
race	-0.38	0.68	\$137
y.o.b.	~	~	~
married	0.31	1.36	\$272
metro area	-0.16	0.85	\$170

Causal inference

If we can ensure that no other variables are correlated with the treatment, we can interpret its effect on an outcome causally.

Balance



Balance

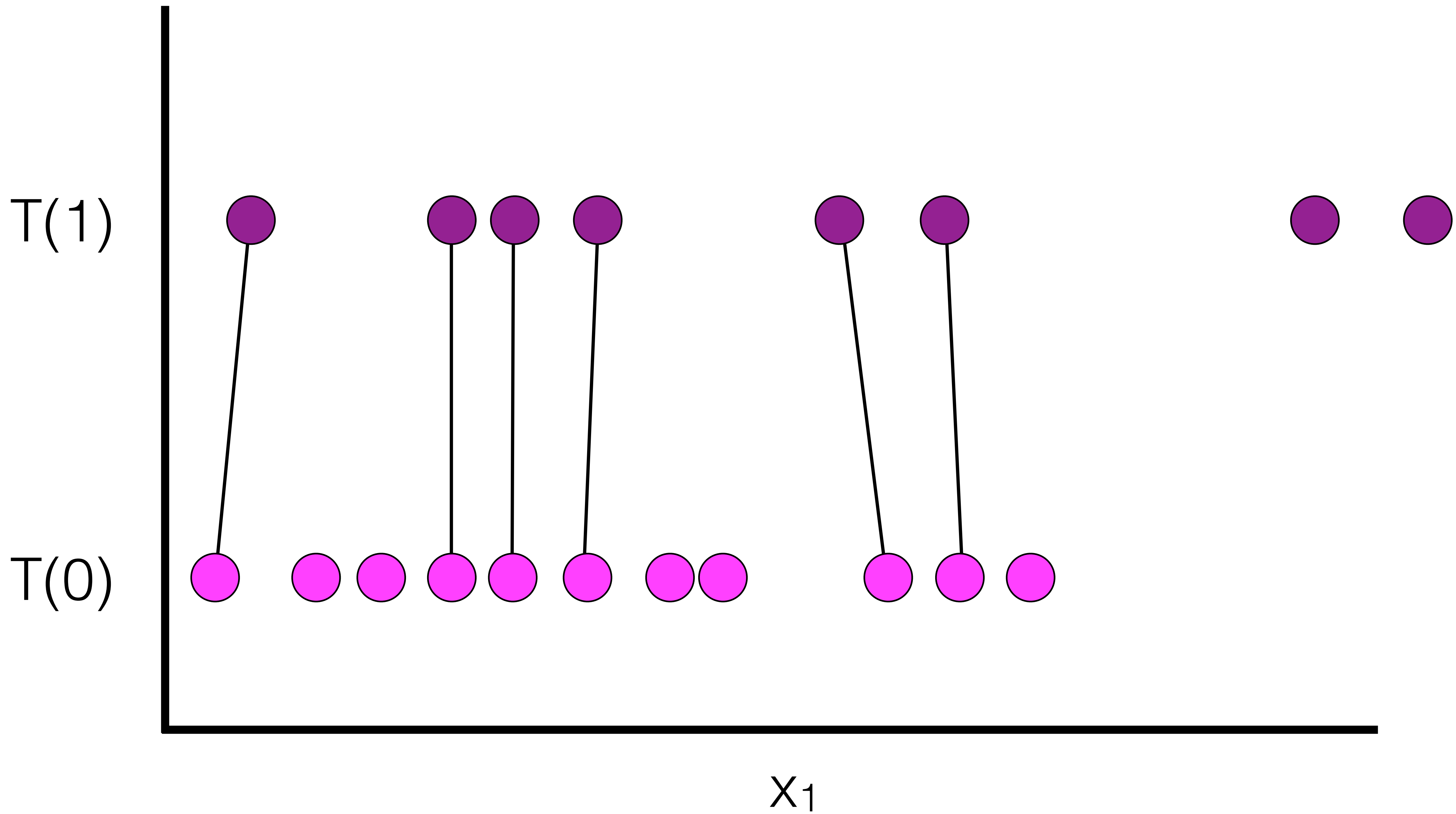
	balance
married	-0.081
race	0.450
metro	0.116

$$\frac{\bar{x}_t - \bar{x}_c}{\sigma_t}$$

Matching

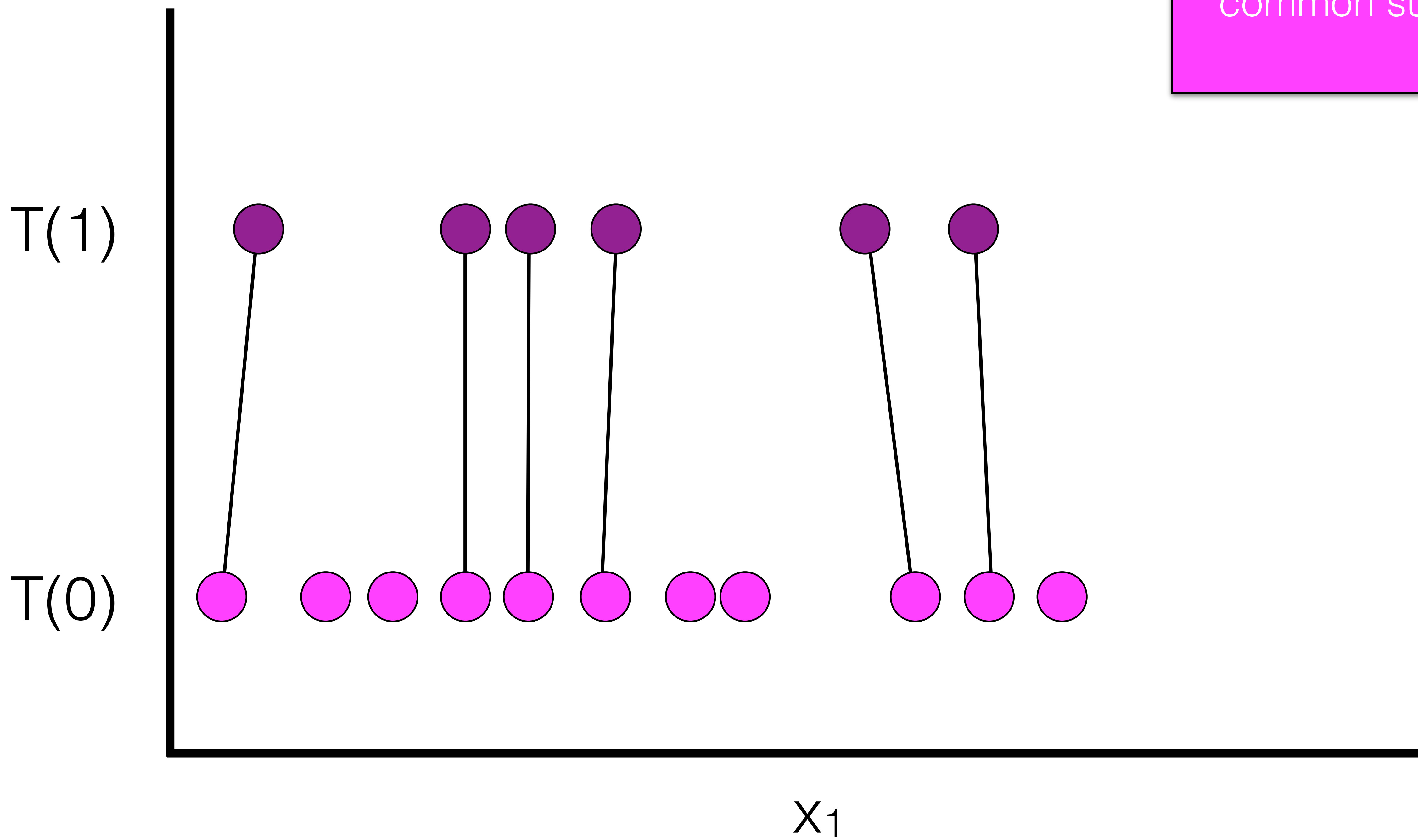
- We'll ensure balance of the covariates by pairing each data point in the treatment with another similar data point in the control
- Ideally: every other feature is the same except the treatment value

Matching

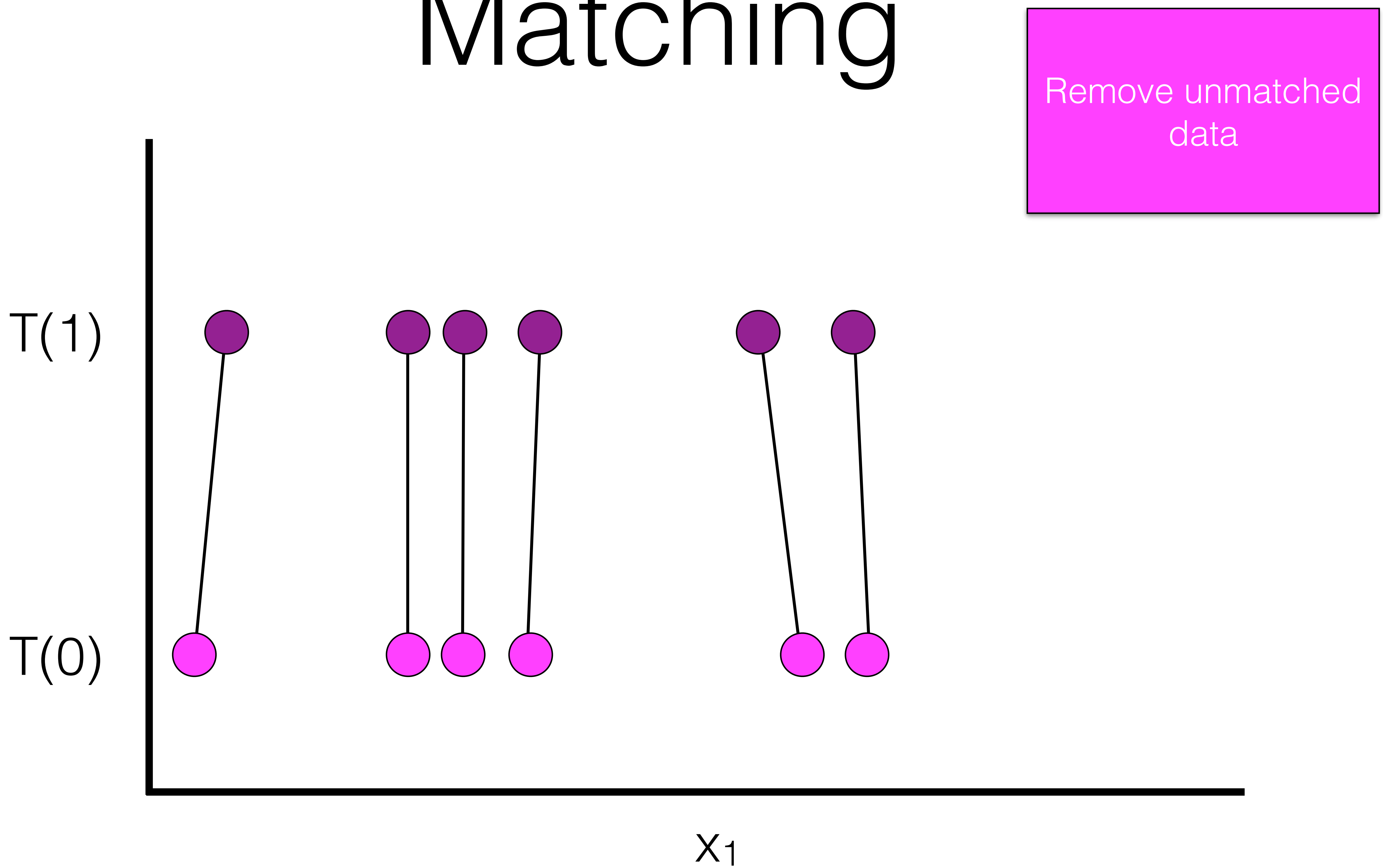


Matching

Remove data without common support



Matching



Matching

- After matching, we need to assess balance again (since the entire point is to improve covariate balance).

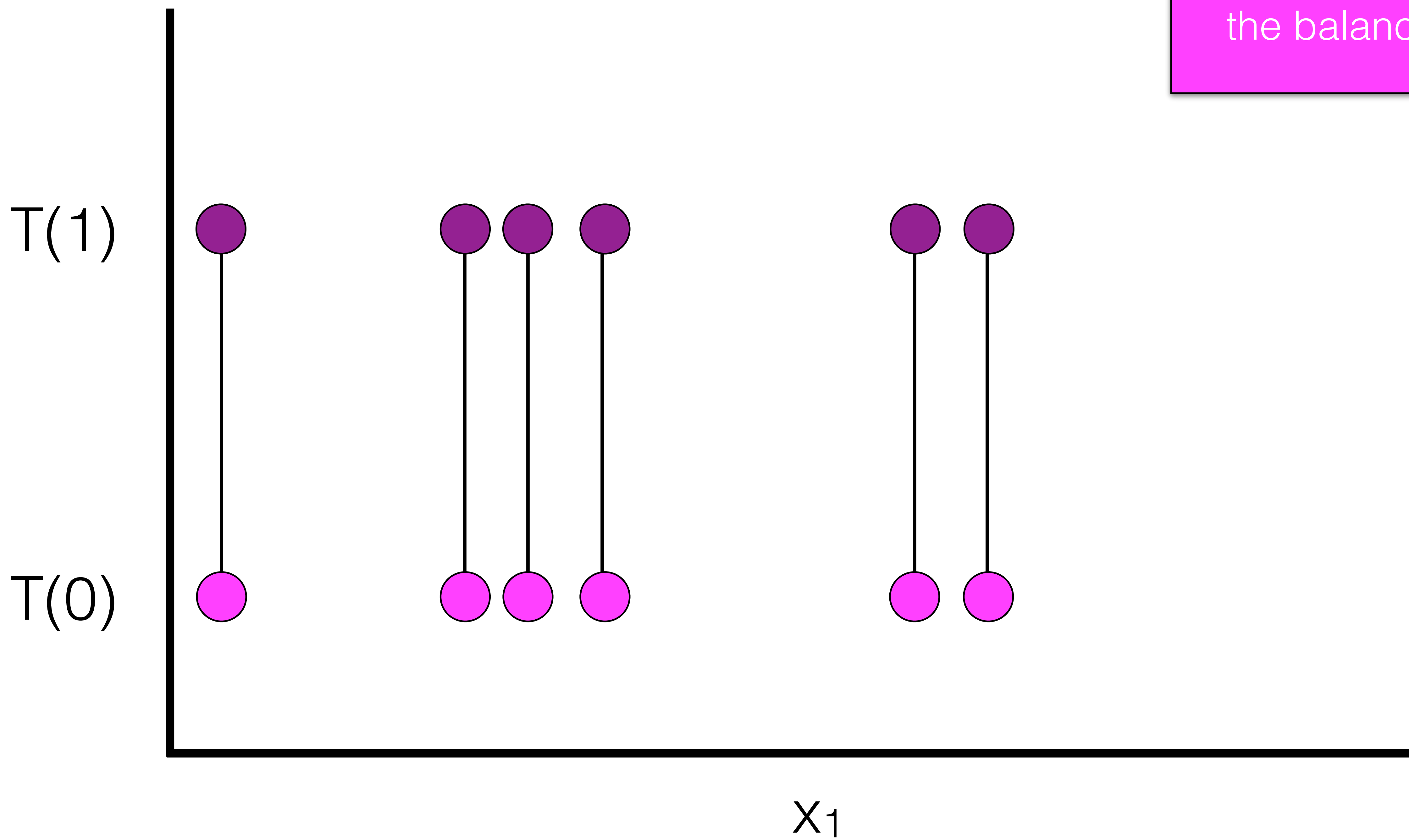
Distance measurements

- Exact matching: match a treatment data point to a data point with exactly the same values for all of its features

graduate	race	y.o.b.	married	metro area
1	0	1927	1	1
1	1	1921	1	0
1	0	1923	0	0
0	0	1927	1	1
1	1	1928	1	1

Matching

If matching was exact, what would the balance be?



Coarsened Exact Matching

- Preprocessing: “coarsen” each variable (e.g., into buckets) and define strata of variables that have exact coarsened values
- Throw out all strata that don’t have at least 1 treatment and control data point
- Rebalance treatment and control within each strata so each strata has the same distribution of treatment and control units as the entire dataset.

Coarsened Exact Matching

- How do we coarsen?

graduate	city	y.o.b.	siblings	metro area	politics
1	Berkeley	1990	3	1	very liberal
1	Boise	1987	1	0	liberal

Mahalanobis Distance

- Distance metric between two points x_i and x_j that accounts for different features having different degrees of variability

- Σ = covariance matrix

$$MDM(x_i, x_j) = (x_i - x_j)\Sigma^{-1}(x_i - x_j)$$

Propensity scores

- Propensity scores generate a single summary number for all covariates: the probability of the treatment

y   x

graduate	race	y.o.b.	married	metro area
1	0	1927	1	1
1	1	1921	1	0
1	0	1923	0	0
0	0	1927	1	1
1	1	1928	1	1

Propensity scores

- Propensity scores generate a single summary number for all covariates: the probability of the treatment

$$T \perp Y \mid X$$

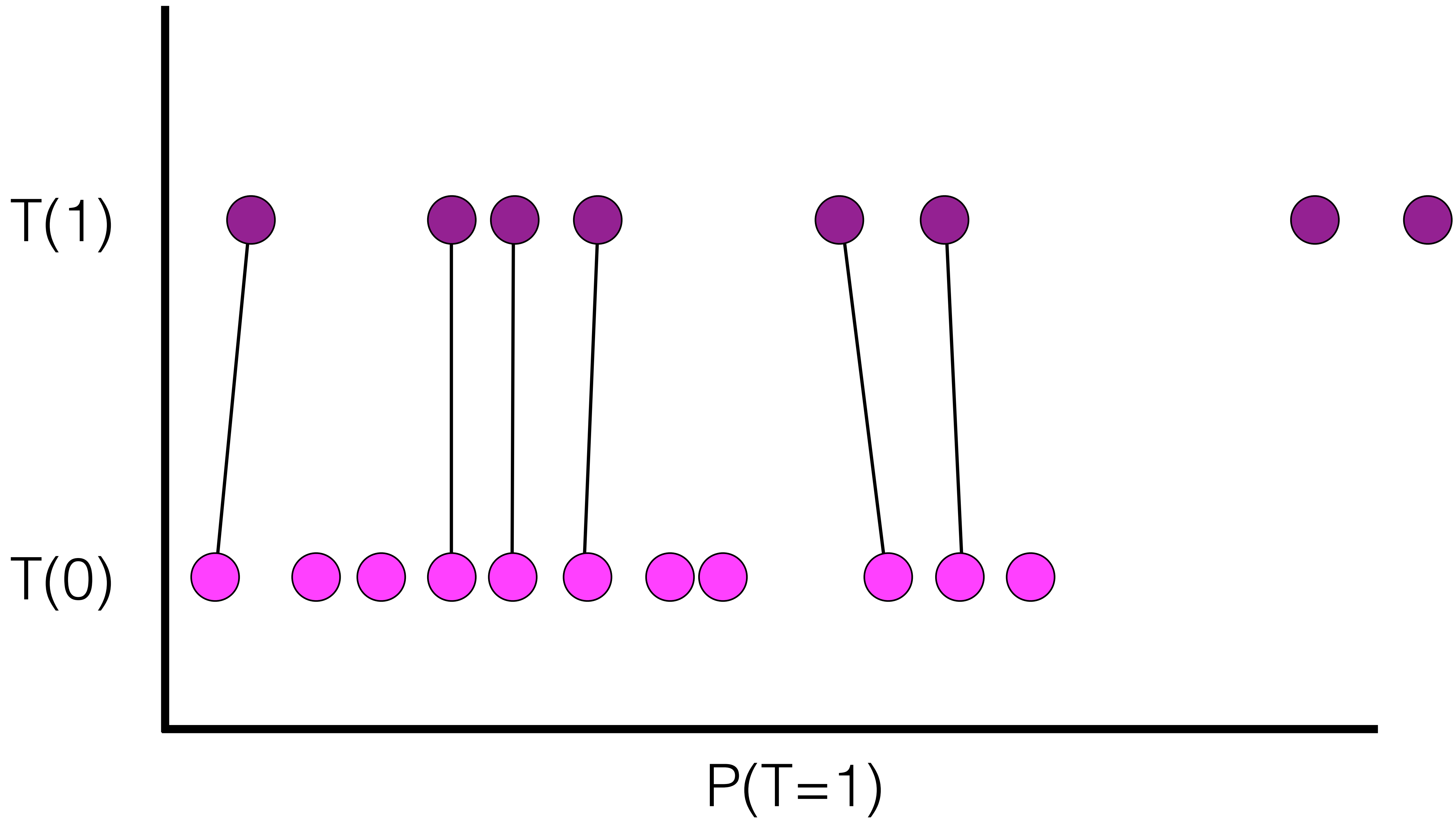
$$\Rightarrow T \perp Y \mid P(T = 1 \mid X)$$

Propensity scores

- We can use any model that generates a probability as part of its decision
- The accuracy of the model does not matter as much as the covariate balance after matching

$$P(y = 1 \mid x, \beta) = \frac{\exp\left(\sum_{i=1}^{F} x_i \beta_i\right)}{1 + \exp\left(\sum_{i=1}^{F} x_i \beta_i\right)}$$

Matching



Balance

- With matching, we are identifying a subset of our original data to use for analysis

- The entire point of matching is to reduce imbalance among the covariates. We need to check that it worked.

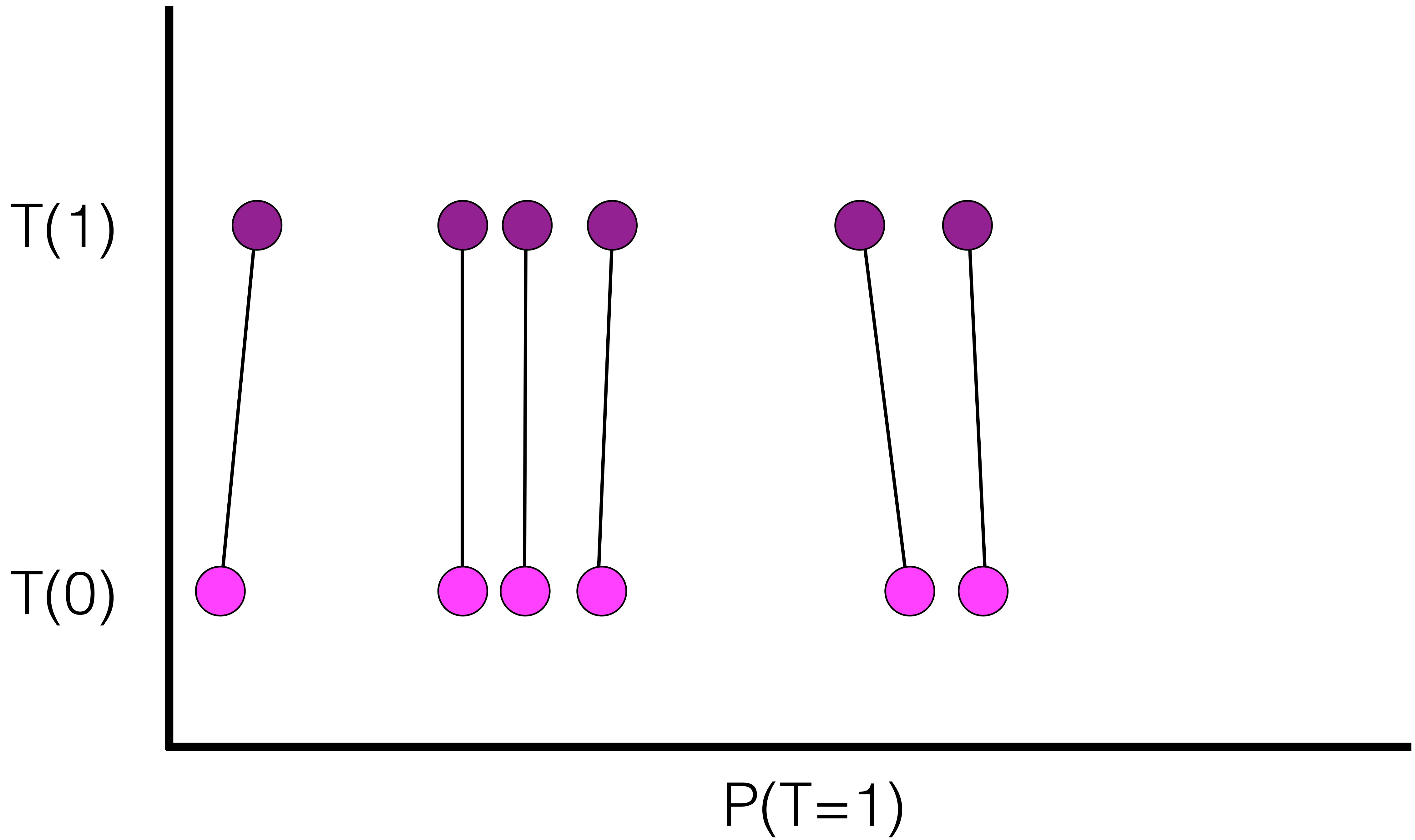
Balance

Want post-matching balance < 0.25

$$\frac{\bar{x}_t - \bar{x}_c}{\sigma_t}$$

	balance before matching	balance after matching
married	-0.081	-0.007
race	0.450	0.01
metro	0.116	0.005

Matching



Analysis

- Matching methods constitute a design phase for causal analysis: identifying the subset of observational data that can be thought of as a latent randomization experiment.
- Once we identify the subset, we simply apply the original analysis to it — e.g., linear/logistic regression and analyzing the coefficients for significance.

Analysis

$$y = \sum_{i=1}^{F} x_i \beta_i + \epsilon$$

	β	$\beta_{matching}$	\$200
graduate	0.35	0.34	\$281
race	-0.38	-0.36	\$140
y.o.b.	~	~	
married	0.31	0.31	\$284
metro area	-0.16	-0.14	\$174

Assumptions

- Ignorability
- Positive probability of treatment
- SUTVA

Ignorability

- The treatment T is independent of the potential outcomes Y given the observed covariates X.

$$T \perp Y \mid X$$

Positivity

- The probability of receiving a treatment is positive (i.e., non-zero) for all values of X

$$P(T = 1 \mid X) > 0$$

SUTVA

- Stable unit treatment value assumption
- The outcome for one data point is not affected the treatment for another

$$T_i \perp Y_j$$

Issues

- What about high-dimensional problems?