Behavioral Data Mining

Lecture 9

Intro. to Causal Analysis and Observational Studies

Outline

- Confounds
- Matching
- Mahalanobis distance matching
- Propensity Scores
- Checking Balance

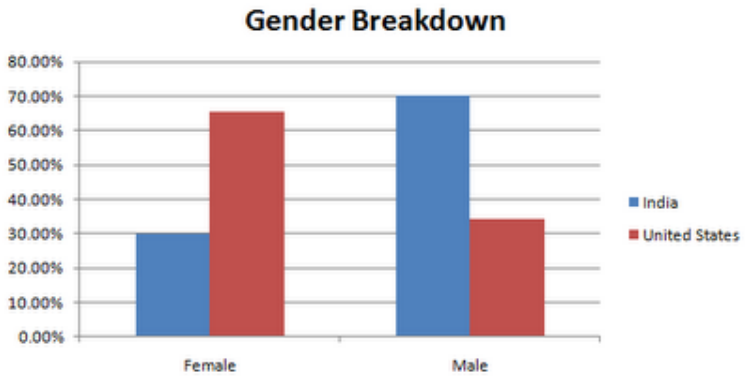
A Mechanical Turk study revisited

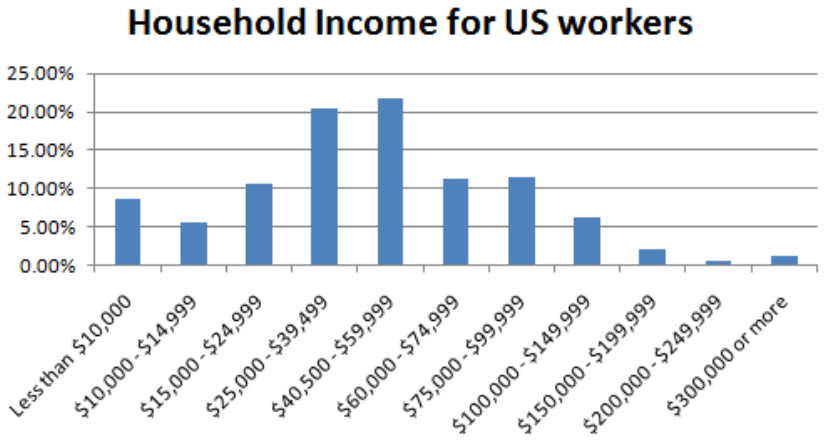
- Please state your income:
- Please state your gender:

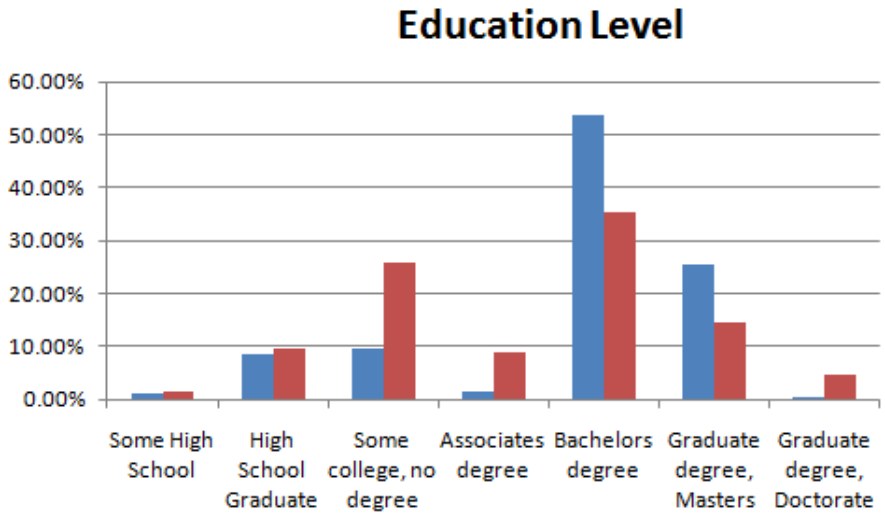
A Mechanical Turk study revisited

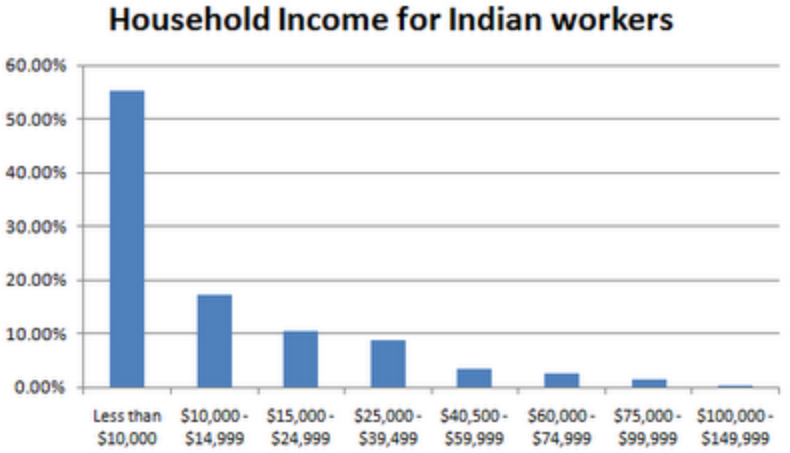
Result: On average women earn 10x what men do ??

Mechanical Turker Demographics

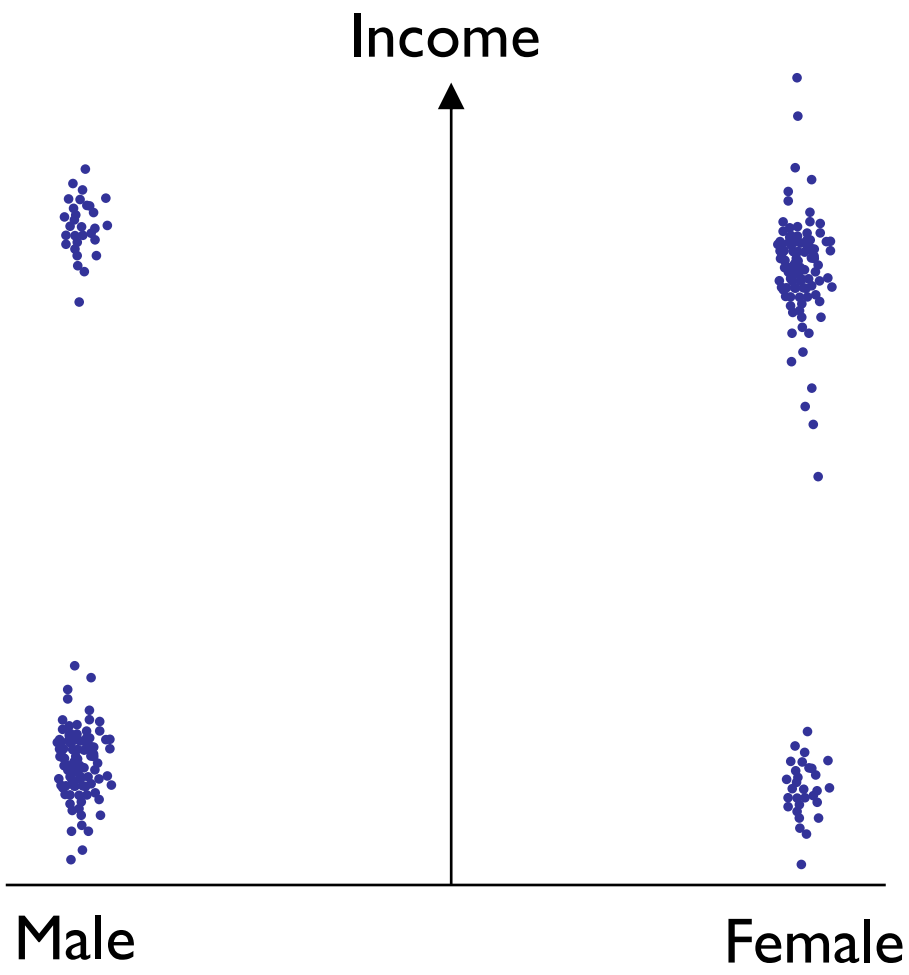




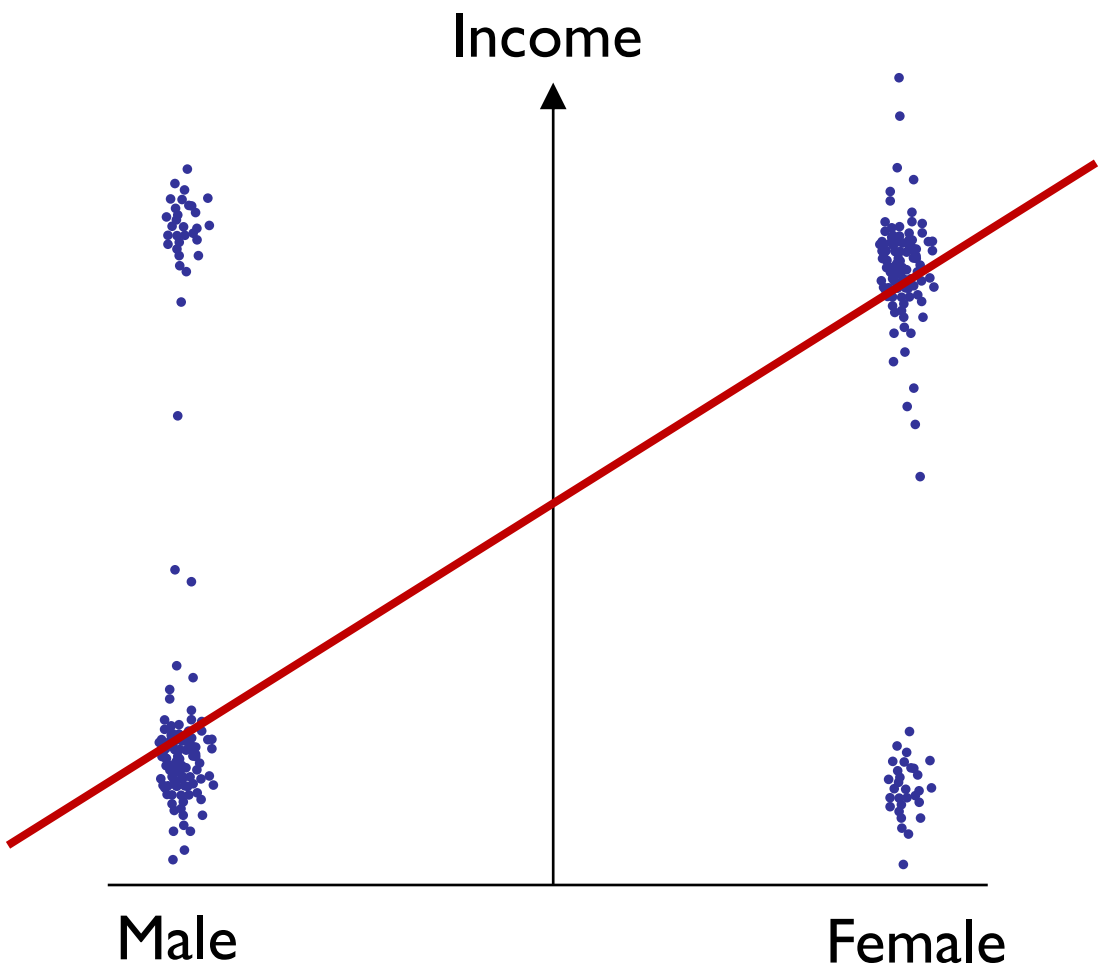




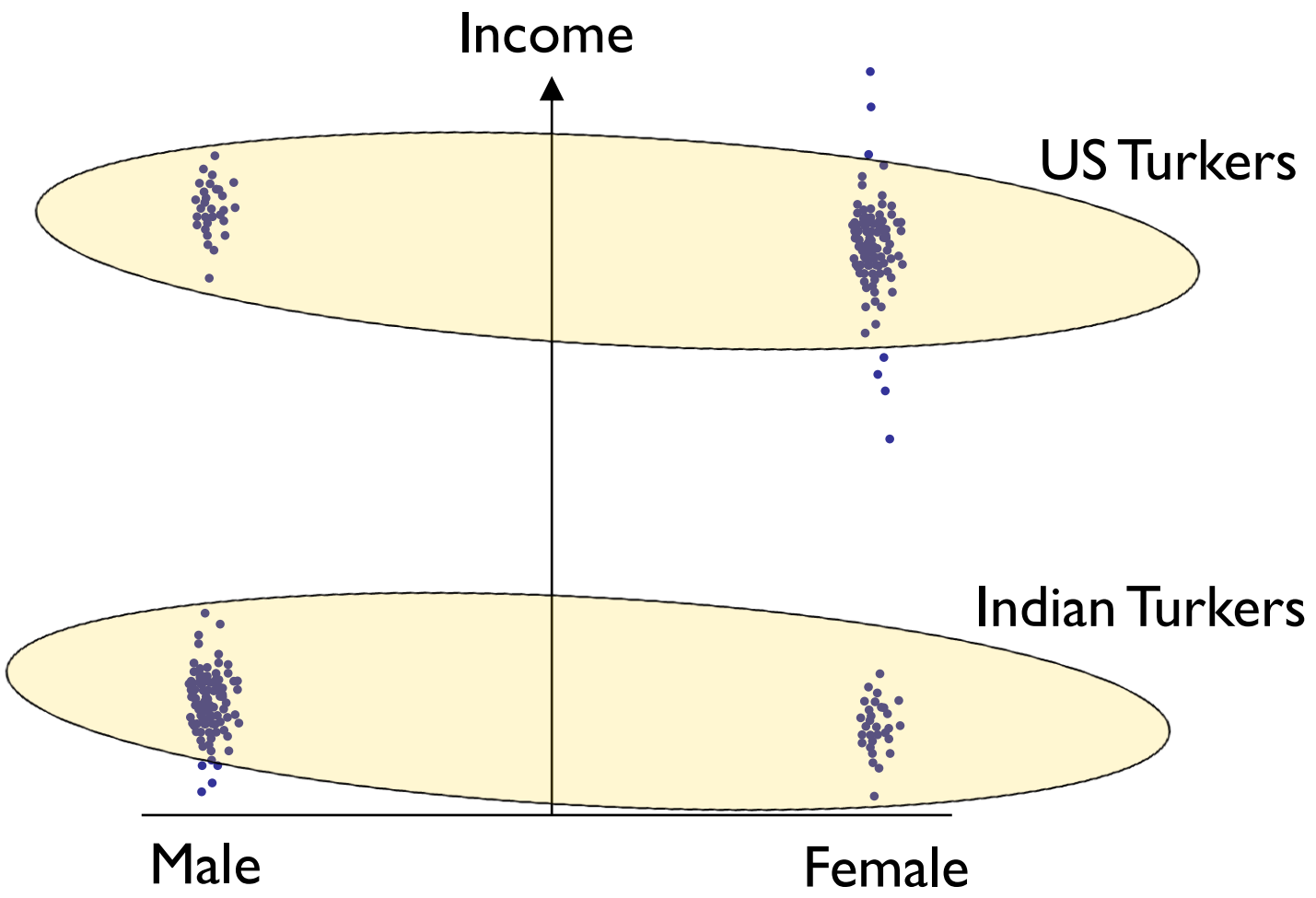
Scatter Plot



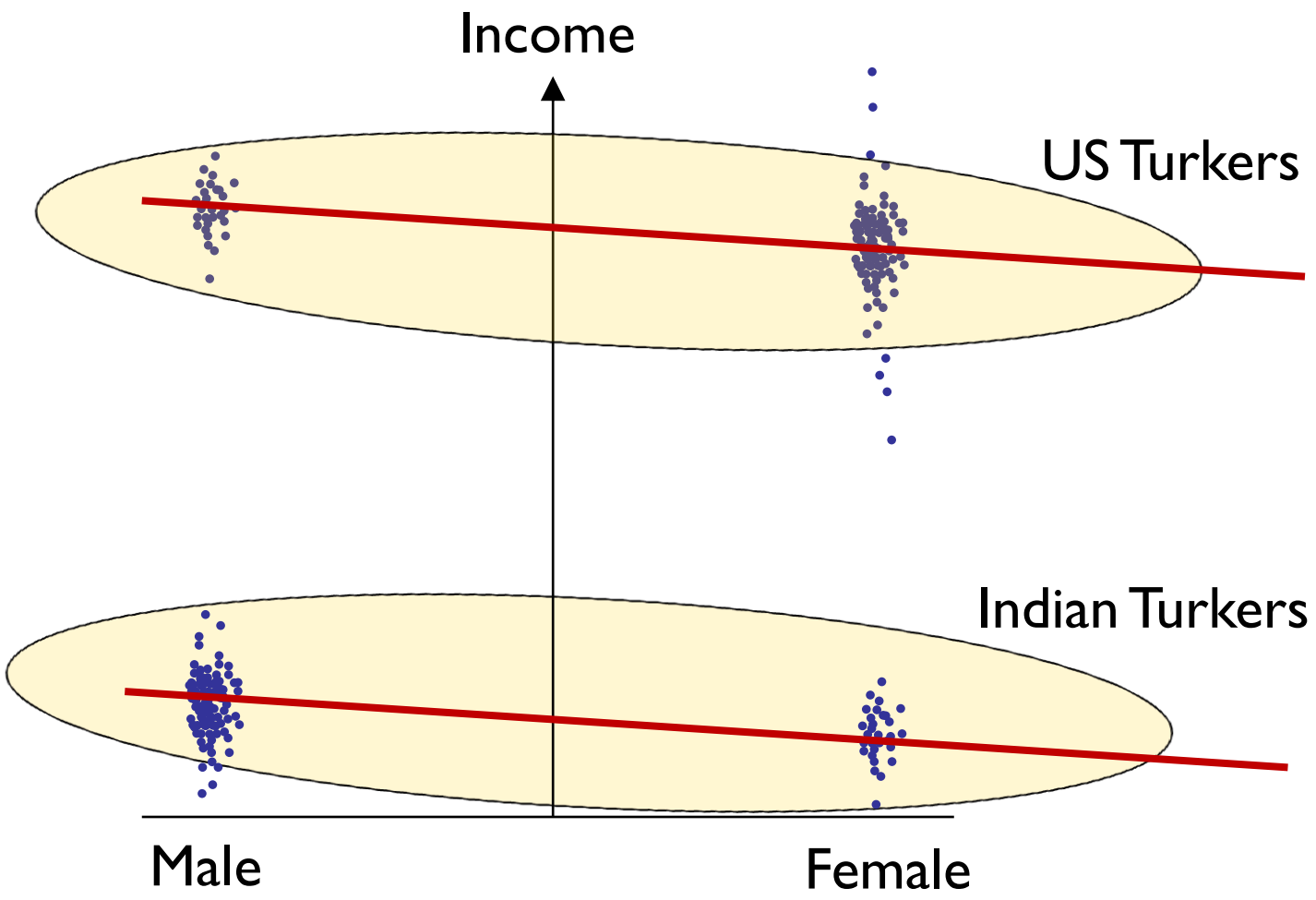
Regression



Scatter Plot



Matching



Observational Studies

Observational Study: Umbrella term for studies conducted without randomization.

We assume there is **treatment** and a **control group**, but that the assignment cannot be done randomly. Examples of generalized treatments:

- Drugs already prescribed
- Smoking, Chronic Illness
- Genetic effects
- Marriage/divorce
- Nationality, Gender, Education
- Life Events
- Values

Natural Experiments

Even if the treatment assignment isn't randomized by the experimenter, it can still be random by "nature", and independent of any confound:

- Gender of human children is a good example.

An experiment designed with such a treatment is called a "**Natural Experiment**".

But the Turker example shows that natural treatment of a population is not enough. The ***dataset sampling*** must also be random across the treatment conditions, but here it clearly is not.

Causal Analysis

Much of behavioral data analysis is concerned with **causal** or **counterfactual** questions:

- Did observable A cause B ?
- What would happen if A=b (not observed) instead of A=a (observed)?

- Will showing this ad promote a purchase?
- Does spending time on Facebook enhance job prospects?
- Does playing video games postpone cognitive decline?

Correlation ≠ Causation

Retargeting: serve ads that correlate with purchases.

Do they *enhance* purchases?

Causal Analysis

Non-treatment variables which correlate with either the treatment or the outcomes are potential ***confounds***.

Causal analysis deals with eliminating or reducing the effects of those variables.

In the Turk study these are:

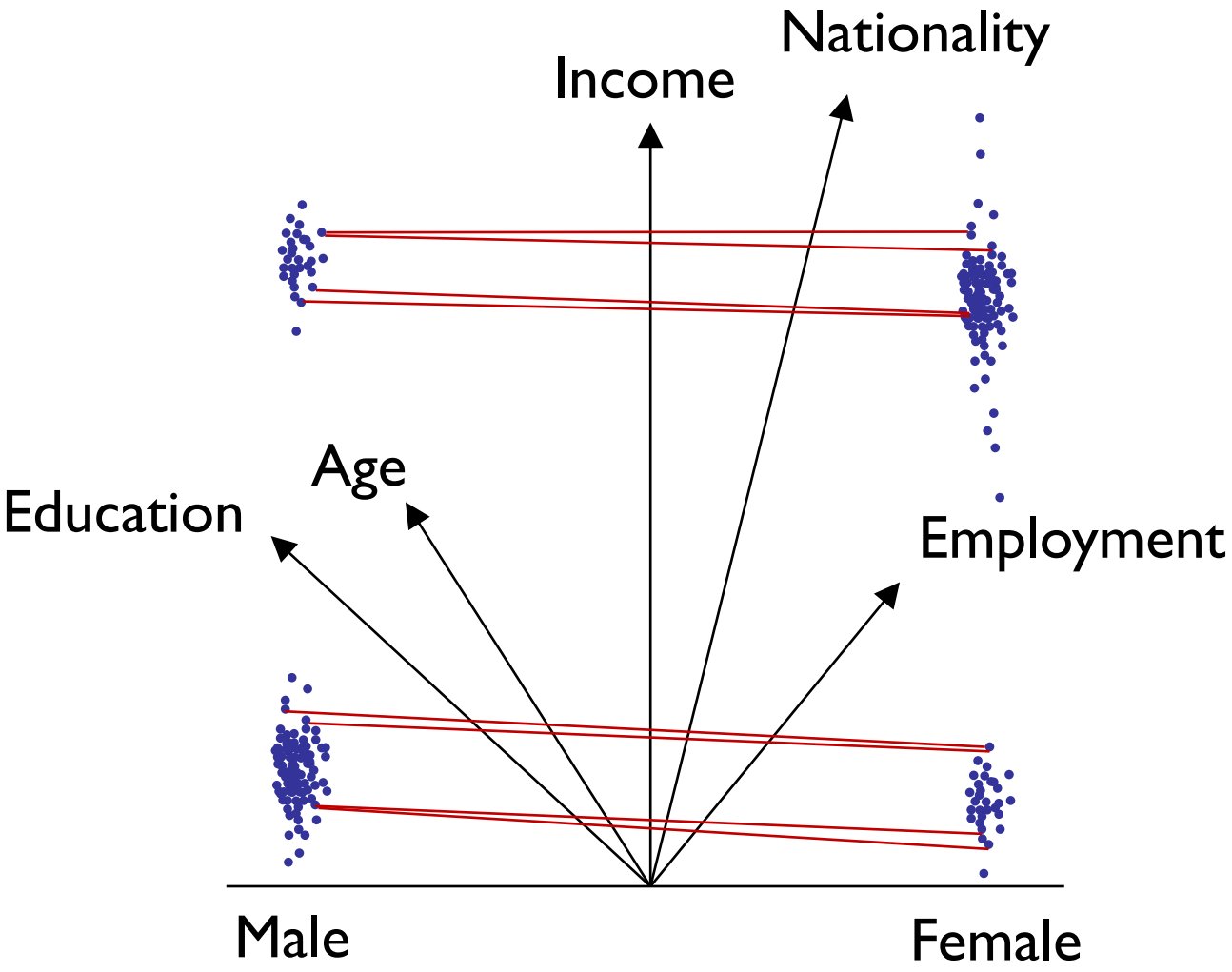
- Nationality
- Age
- Education
- Employment area

Note these only affect admission of individuals into the study, not their gender.

Outline

- Confounds
- Matching
- Mahalanobis distance matching
- Propensity Scores
- Checking Balance

Multivariate Matching



Multivariate Matching

Let the two populations be P_1 and P_2 and take samples G_1 and G_2 from them.

- Find "matching" subsamples G_{1*} and G_{2*} that agree on all the non-outcome (matching) variables *X*.
- Compare values of outcome (dependent) variables between G_{1*} and G_{2*}.

For discrete matching variables, the matching is normally simple agreement on those variables (e.g. Nationality).

For continuous variables, the goal is to reduce the bias in the dependent variable estimates, **EPBR** or **Equal Percent Bias Reducing**.

Multivariate Matching

What measure of similarity to use for matching?:

- Mahalanobis distance (covariance-corrected euclidean distance).
- Propensity score.
- Genetic matching.

Mahalanobis Distance Matching

For a multivariate data vector $(X_1, \ldots, X_N)^T$ with means $(\mu_1, \ldots, \mu_N)^T$, the **covariance matrix** is

$$S_{ij} = \mathrm{E}\left[(X_i - \mu_i)^T (X_j - \mu_j)\right]$$

The **Mahalanobis distance** between points x and y is

$$d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

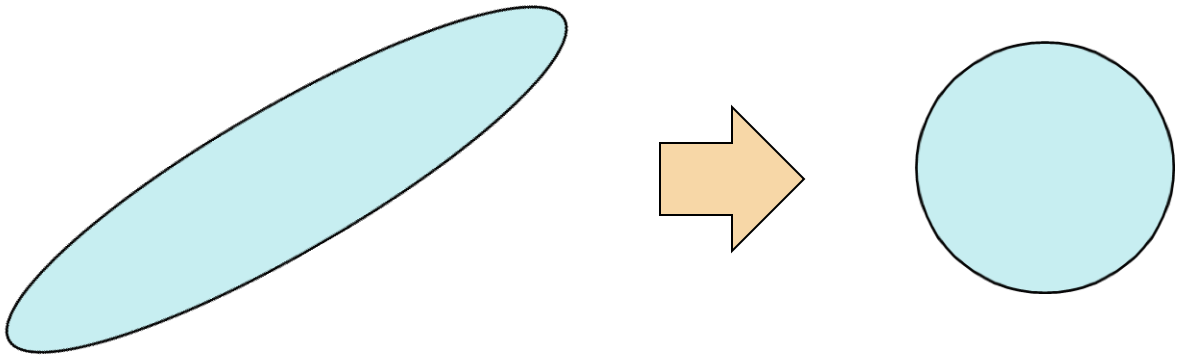
or equivalently the euclidean distance in a projection space:

$$p(x) = S^{-1/2}(x - \mu)$$

Mahalanobis Distance Matching

If the original data are ellipsoidally-distributed, the Mahalanobis projection produces independent, equal-variance coordinates:

$$p(x) = S^{-1/2}(x - \mu)$$



Under these conditions, Mahalanobis distance matching is EPBR

Multivariate Matching Difficulties

Multivariate matching has many difficulties.

In particular there is a "curse of dimensionality".

As d (number of covariates) goes up, the number of cells in a d-dimensional partition of space grows exponentially with d.

Intuitively, we care most about variables which are true confounds, i.e. **whose values differ between treatment and control groups**.

e.g. If gender were balanced between US and Indian Turker groups, we would not see a confounding between nationality and gender.

Outline

- Confounds
- Matching
- Mahalanobis distance matching
- Propensity Scores
- Checking Balance

Propensity Score

It may be possible to use a simpler function of the covariates for matching. Define a **balancing score** $b(x)$ such that

$$x \perp\!\!\!\perp z \,|\, b(x)$$

i.e. the conditional distribution of x given b(x) is the same for different values of z (treatments).

The balancing score removes coordinates that don't help (that don't covary with the treatment).

It turns out there is a single coarsest balancing score called the **propensity score**.

Propensity Score

Define

$$e(x) = \Pr(z = 1 | x)$$

then $e(x)$ is the **propensity score**.

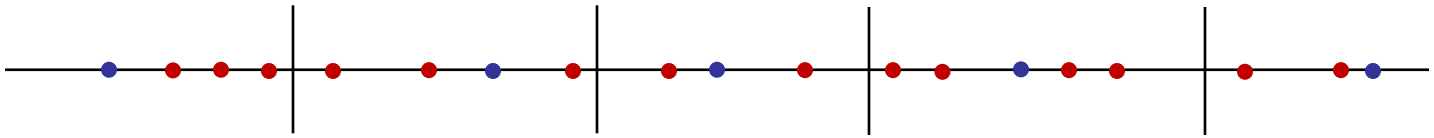
$e(x)$ is a univariate function which succinctly predicts the treatment condition from the covariates.

While there is no closed form in general for $e(x)$, it is often approximated with a logistic model.

Propensity Score Matching

After estimating $e(x)$, the univariate values of $e(x)$ are used for matching. Several matching strategies are possible:

- Nearest neighbor matching: from one side (e.g. treatment) find the nearest neighbor in the other side.
- This results in a **stratification** of the treatment values.
  - = control
  - = treatment

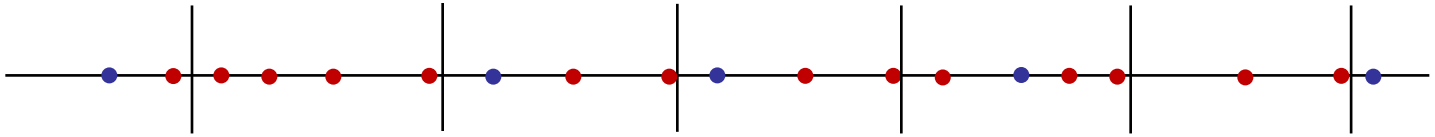


Causal effects (differences) in outcomes between control and treatment cases treatment values are computed in each stratum.

Stratification

Break the score range into strata of some size (oblivious to the sample values)

- = control
- = treatment



Propensity Score

Table 1. Comparison of Mortality Rates for Three Smoking Groups in Three Databases*

Variable	Canadian Study			United Kingdom Study			United States Study		
	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers
Mortality rates per 1000 person-years, %	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
Average age, *y*	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
Adjusted mortality rates using subclasses, %									
2 subclasses	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3 subclasses	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9–11 subclasses	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

The propensity score method is probably the most widely used technique in causal analysis. It has contributed much of the knowledge base for public health and economics.

Outline

- Confounds
- Matching
- Mahalanobis distance matching
- Propensity Scores
- Checking Balance

Strong Ignorability

If (r_1, r_0) are some observations, v some covariates, then if

$$(r_1, r_0) \perp\!\!\!\perp z \mid v \quad 0 < \Pr(z = 1|v) < 1$$

we say the treatment assignment is **strongly ignorable.**

This requires that every combination of covariates and treatments actually occur.

It also assumes that all covariates are observable.

These are strong conditions, as is the requirement to compute an accurate $e(x)$. So matching is still a preferable alternative in many cases.

Checking Balance

If T is the treatment variable, X the observed covariates and U the unobserved covariates, **selection on observables** gives:

$$T \perp\!\!\!\perp U \mid X$$

And if propensity scoring using $\pi(X)$ is used, it should be that:

$$X \perp\!\!\!\perp T \mid \pi(X)$$

If the propensity score balances, the distribution of covariates in strata of $\pi(X)$ should be identical between treatment and control.

Kolmogorov-Smirnoff Test

Is a non-parametric test to determine if two distributions could be the same.

It tests a variety of forms of difference, and is a good choice for checking balance in a univariate matching method.

This is the basis of the "Genetic Match" algorithm from the reading.