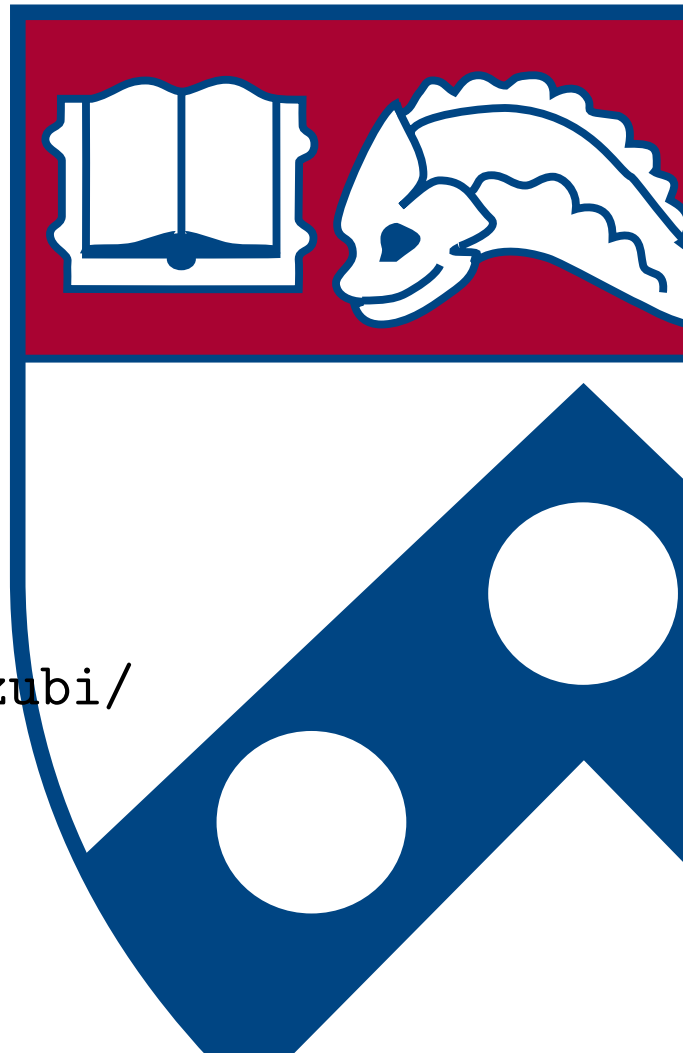
Using Mixed Integer Programming for Matching in Observational Studies

José R. Zubizarreta
Department Statistics
The Wharton School
University of Pennsylvania



http://www-stat.wharton.upenn.edu/~josezubi/

Key takeaway points

- Optimal matching method
  - Get the balance you want
  - Know it is infeasible
    - Eliminate guesswork
- Directly balance several statistics beyond means
- Keep the adjustments simple enough
  - People can talk about them
  - Sensitivity analysis to unobserved biases

Outline

Outline

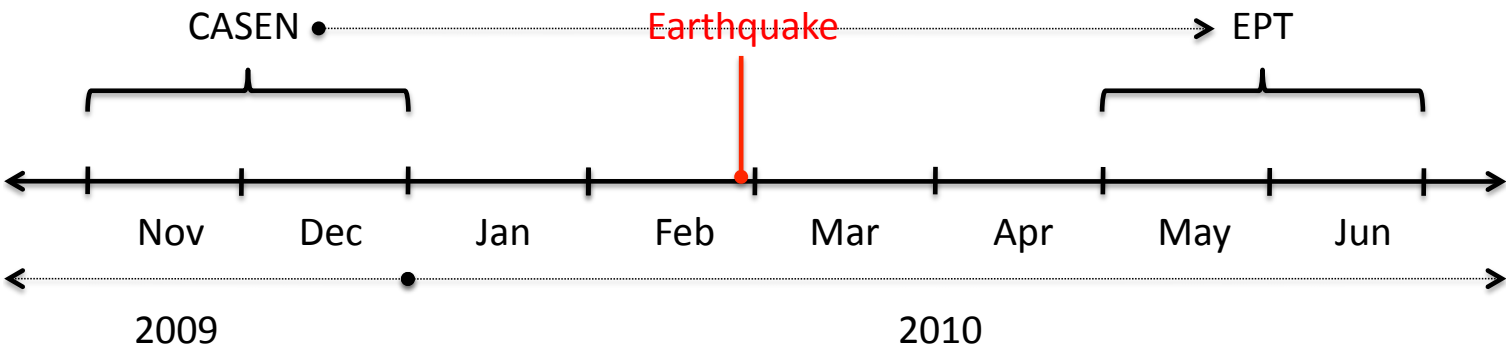
The 2010 Chilean earthquake

4th strongest earthquake in the world in the last 50 years

Sebastián Martínez/AP Photo

Effect of the earthquake

- Effect of the earthquake on posttraumatic stress?
- The post earthquake survey (EPT)

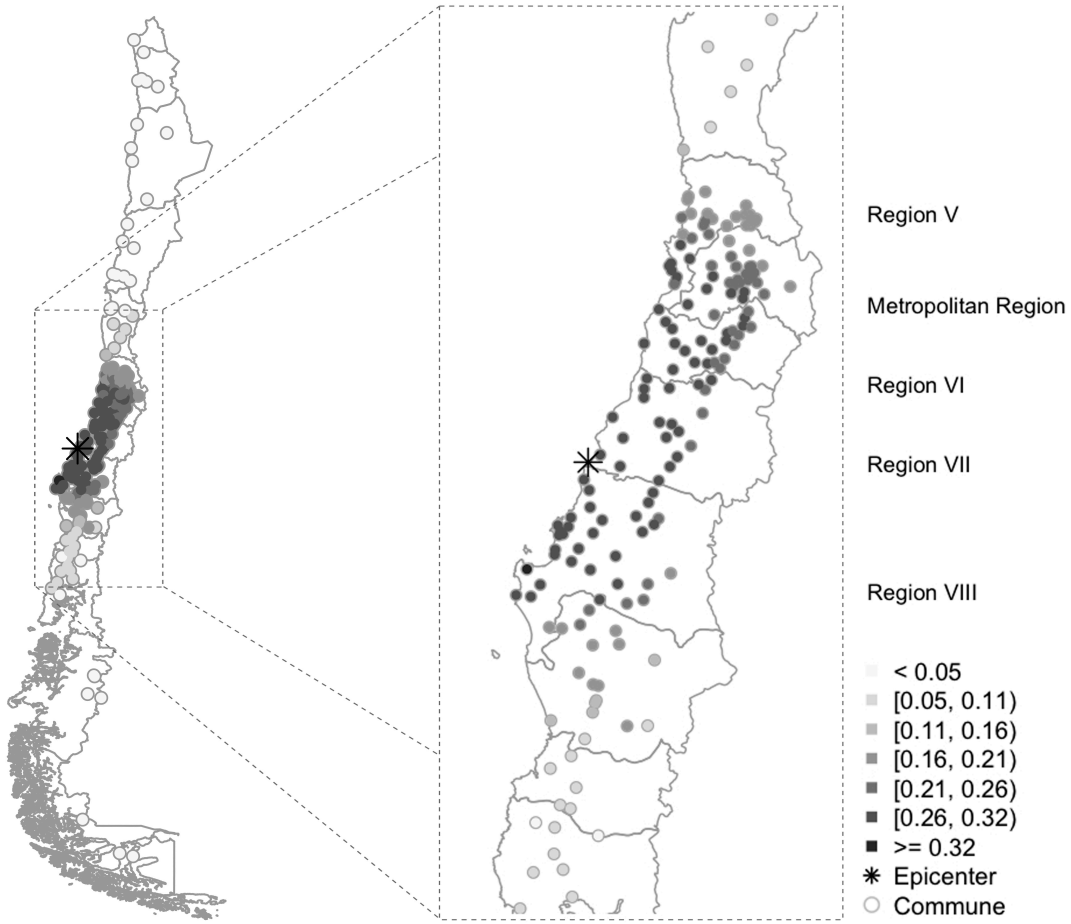


Re-interviewed 22,456 households from CASEN 2009

Detailed measurements of the **same** individuals before and after

Intensity of the earthquake

- Peak ground acceleration (PGA) in the communes of the EPT



Matched design

- Matched respondents with PGA < 0.014 to those with PGA ≥ 0.275
- We matched
  - exactly for sex, age and ethnic groups
  - with fine balance for self-rated health, quality of the housing
  - balancing the entire empirical distributions of income
  - mean balancing the 46 covariates in the study

Outline

Notation

- Let $\mathcal{T} = \{t_1, ..., t_T\}$ be the set of treated units, and $\mathcal{C} = \{c_1, ..., c_C\}$, the set of potential controls, with $T \leq C$
- Define $\mathcal{P} = \{p_1, ..., p_P\}$ as the set of observed covariates
- Each treated unit $t \in \mathcal{T}$ has a vector of observed covariates $\mathbf{x}_{t,\cdot} = \{\mathbf{x}_{t,p_1}, ..., \mathbf{x}_{t,p_P}\}$, and each control $c \in \mathcal{C}$ has a similar vector $\mathbf{x}_{c,\cdot} = \{\mathbf{x}_{c,p_1}, ..., \mathbf{x}_{c,p_P}\}$
- Based on these covariates there is a distance $\mathbf{0} \leq \delta_{t,c} < \infty$ between treated and control units
- Decision variable

$$a_{t,c} = \begin{cases} 1 & \text{if treated } t \text{ is assigned to control } c \\ 0 & \text{otherwise} \end{cases}$$

The assignment algorithm

$$\begin{aligned}
\underset{\mathbf{a}}{\text{minimize}} \quad & \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \delta_{t,c} a_{t,c} \\
\text{subject to} \quad & \sum_{c \in \mathcal{C}} a_{t,c} = m, \ \ t \in \mathcal{T} \\
& \sum_{t \in \mathcal{T}} a_{t,c} \leq 1, \ \ c \in \mathcal{C} \\
& a_{t,c} \in \{0, 1\}, \ \ t \in \mathcal{T}, c \in \mathcal{C}
\end{aligned}$$

A MIP with direct balance via the objective function

$$\underset{\mathbf{a}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \delta_{t,c}a_{t,c} + \sum_{j\in\mathcal{J}} \omega_i\mu_i(\mathbf{a})$$

$$\text{subject to} \quad \sum_{c\in\mathcal{C}} a_{t,c} = m, \;\; t \in \mathcal{T}$$

$$\sum_{t\in\mathcal{T}} a_{t,c} \leq 1, \;\; c \in \mathcal{C}$$

$$a_{t,c} \in \{0,1\}, \;\; t \in \mathcal{T}, c \in \mathcal{C}$$

A MIP with direct balance via the constraints

$$\underset{\mathbf{a}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \delta_{t,c} a_{t,c}$$

$$\text{subject to} \quad \sum_{c\in\mathcal{C}} a_{t,c} = m, \;\; t \in \mathcal{T}$$

$$\sum_{t\in\mathcal{T}} a_{t,c} \leq 1, \;\; c \in \mathcal{C}$$

$$a_{t,c} \in \{0,1\}, \;\; t \in \mathcal{T}, c \in \mathcal{C}$$

$$\nu_j(\mathbf{a}) \leq \varepsilon_j, \; j \in \mathcal{J}$$

Outline

Balancing the means of the covariates (1)

$$\underset{\mathbf{a}}{\text{minimize}} \quad \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \delta_{t,c} a_{t,c} + \sum_{j \in \mathcal{J}} \omega_j \left| \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \frac{x_{c,j} a_{t,c}}{mT} - \overline{x}_{\mathcal{T},j} \right|$$

$$\text{subject to} \quad \sum_{c \in \mathcal{C}} a_{t,c} = m, \;\; t \in \mathcal{T}$$

$$\sum_{t \in \mathcal{T}} a_{t,c} \leq 1, \;\; c \in \mathcal{C}$$

$$a_{t,c} \in \{0, 1\}, \;\; t \in \mathcal{T}, c \in \mathcal{C}$$

Balancing the means of the covariates (1)

$$\underset{\mathbf{a},\mathbf{z}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \delta_{t,c} a_{t,c} + \sum_{j\in\mathcal{J}} \omega_j z_j$$

$$\text{subject to} \quad \sum_{c\in\mathcal{C}} a_{t,c} = m, \;\; t \in \mathcal{T}$$

$$\sum_{t\in\mathcal{T}} a_{t,c} \leq 1, \;\; c \in \mathcal{C}$$

$$a_{t,c} \in \{0,1\}, \;\; t \in \mathcal{T}, c \in \mathcal{C}$$

$$z_j \geq \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \frac{x_{c,j} a_{t,c}}{mT} - \overline{x}_{\mathcal{T},j}, \; j \in \mathcal{J}$$

$$z_j \geq -\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \frac{x_{c,j} a_{t,c}}{mT} + \overline{x}_{\mathcal{T},j}, \; j \in \mathcal{J}$$

Balancing the means of the covariates (1)

$$\underset{\mathbf{a},\mathbf{z}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \delta_{t,c} a_{t,c}$$

$$\text{subject to} \quad \sum_{c\in\mathcal{C}} a_{t,c} = m, \;\; t \in \mathcal{T}$$

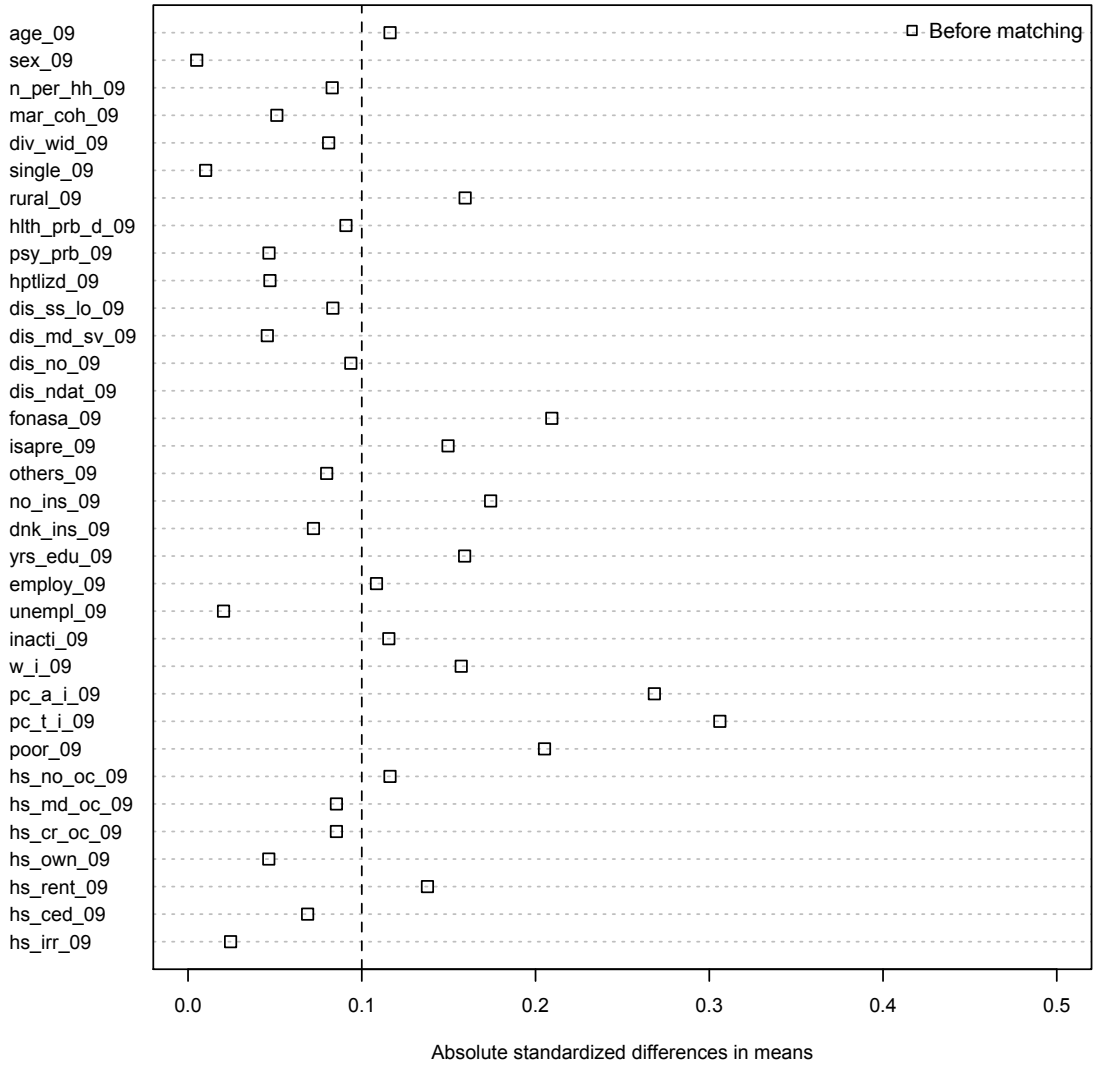
$$\sum_{t\in\mathcal{T}} a_{t,c} \leq 1, \;\; c \in \mathcal{C}$$

$$a_{t,c} \in \{0,1\}, \;\; t \in \mathcal{T}, c \in \mathcal{C}$$

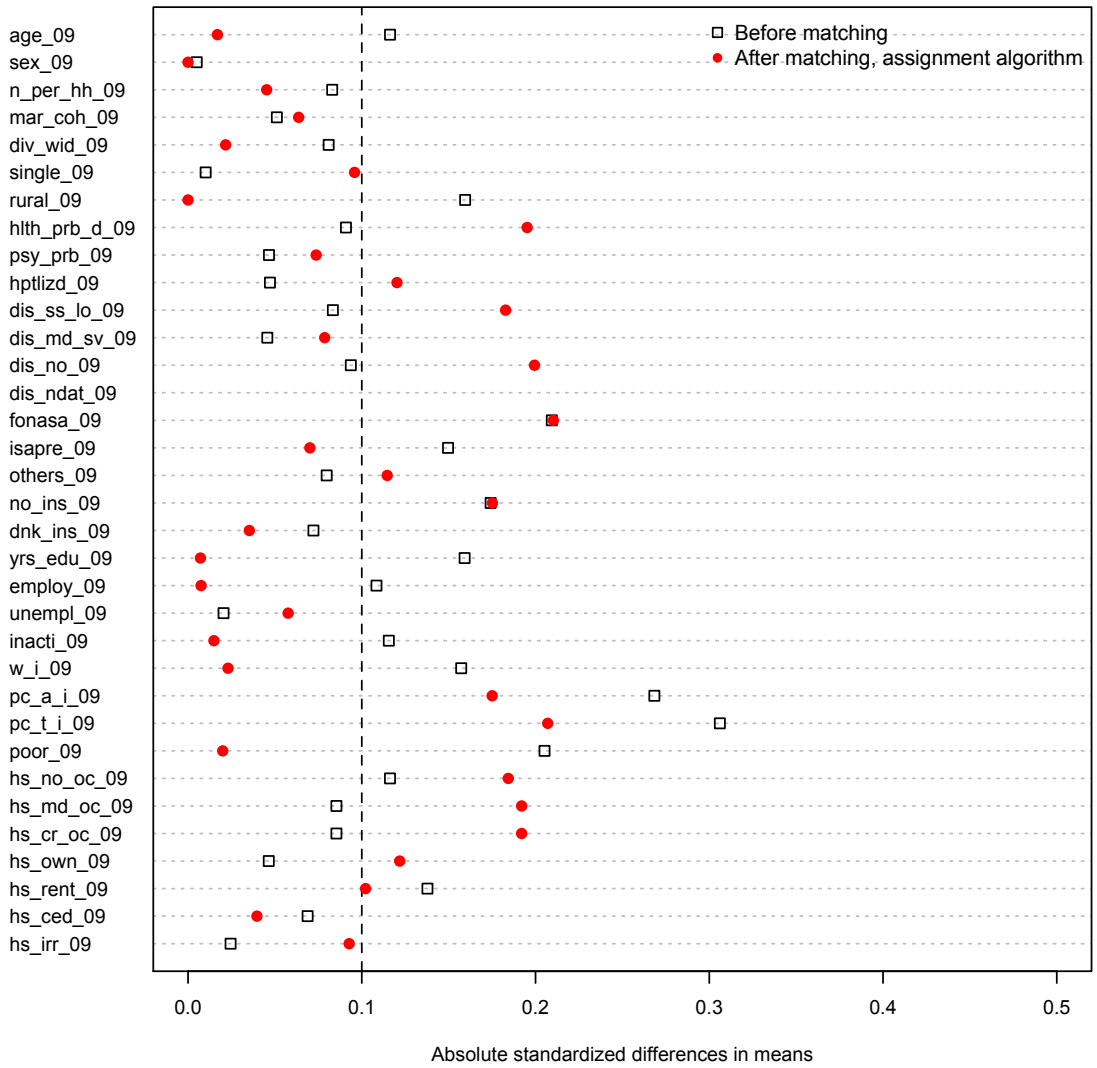
$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \frac{x_{c,j} a_{t,c}}{mT} - \overline{x}_{\mathcal{T},j} \leq \varepsilon_j, \; j \in \mathcal{J}$$

$$-\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \frac{x_{c,j} a_{t,c}}{mT} + \overline{x}_{\mathcal{T},j} \leq \varepsilon_j, \; j \in \mathcal{J}$$

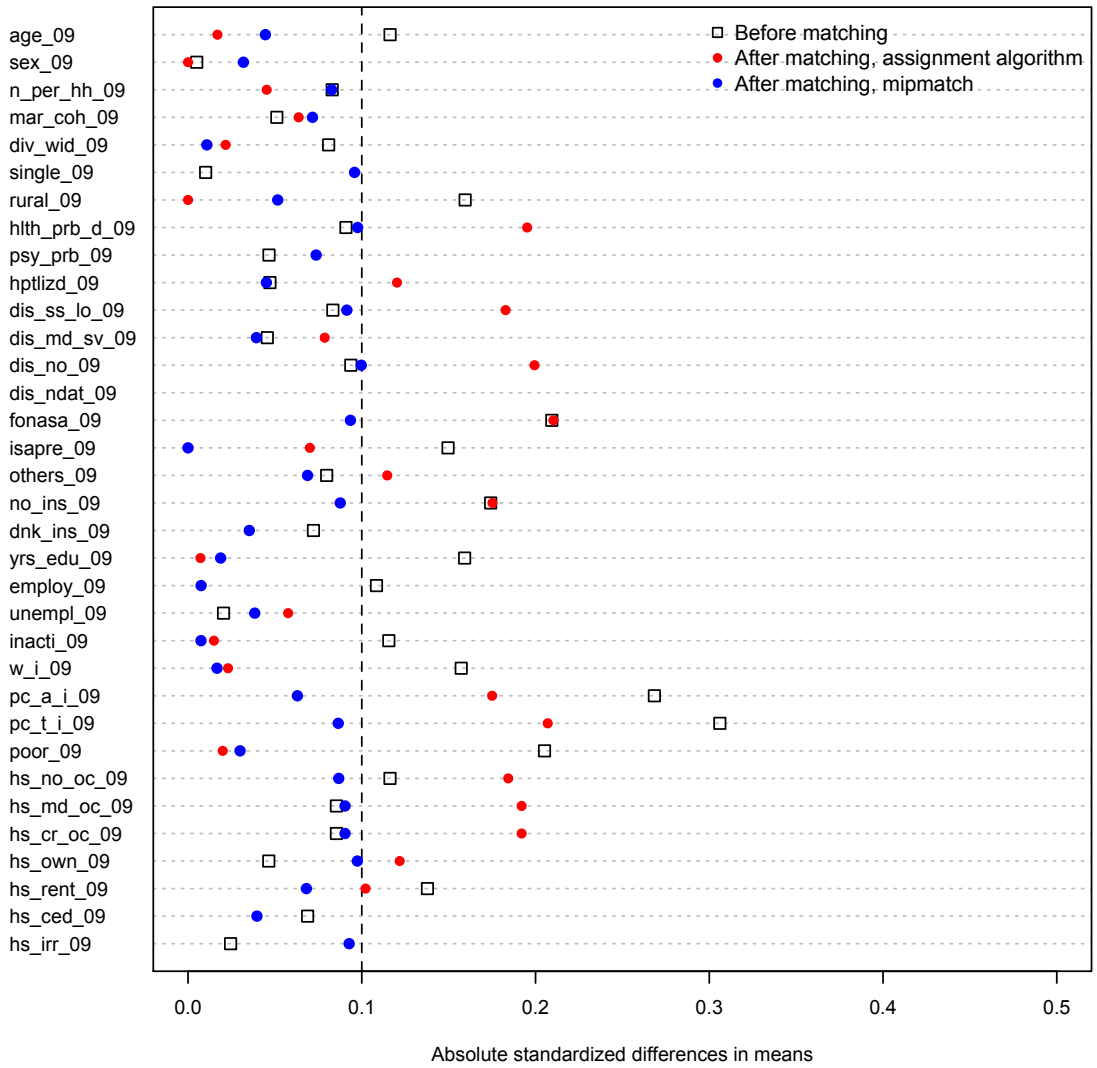
Balancing the means of the covariates (2)



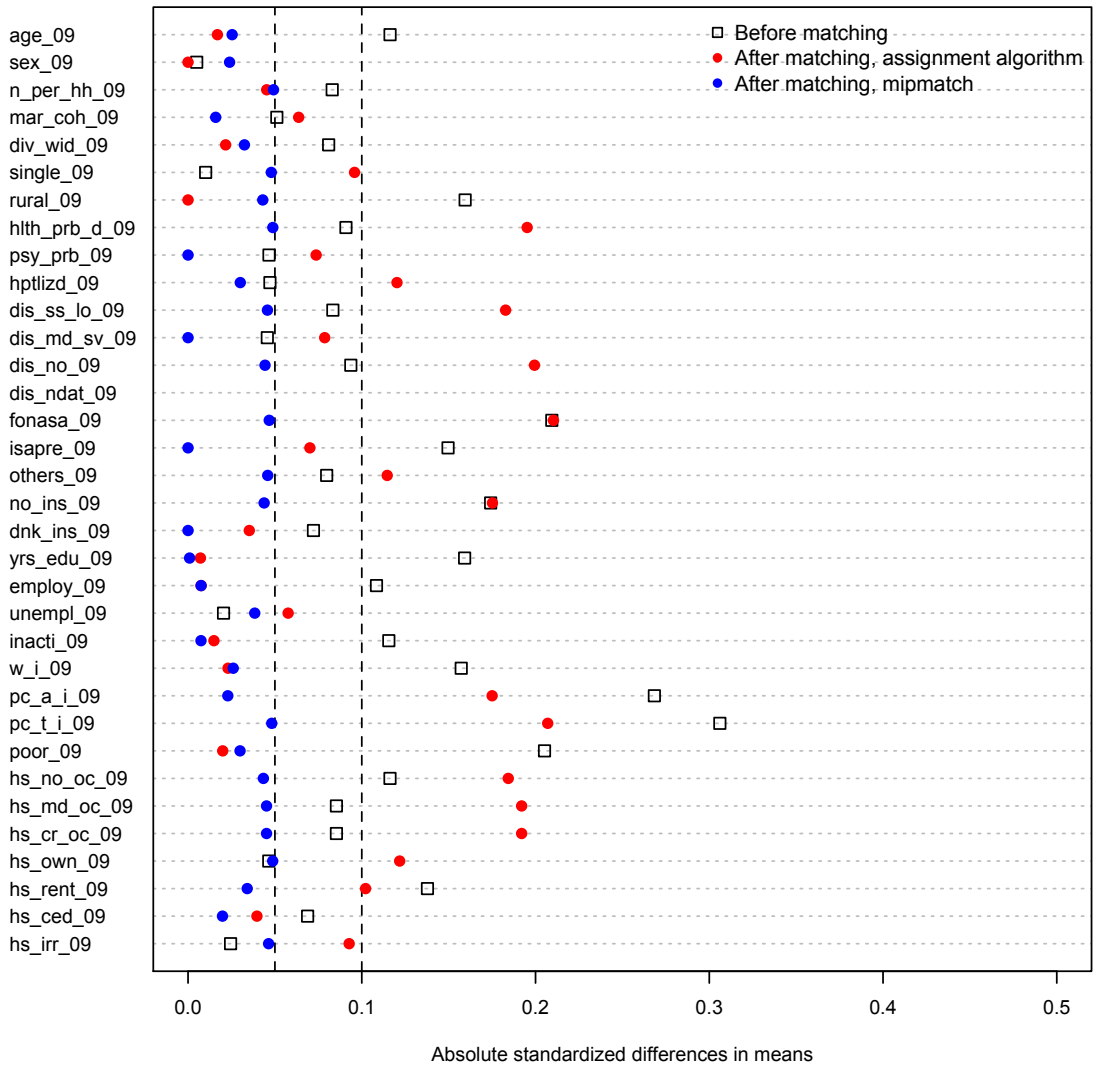
Balancing the means of the covariates (2)



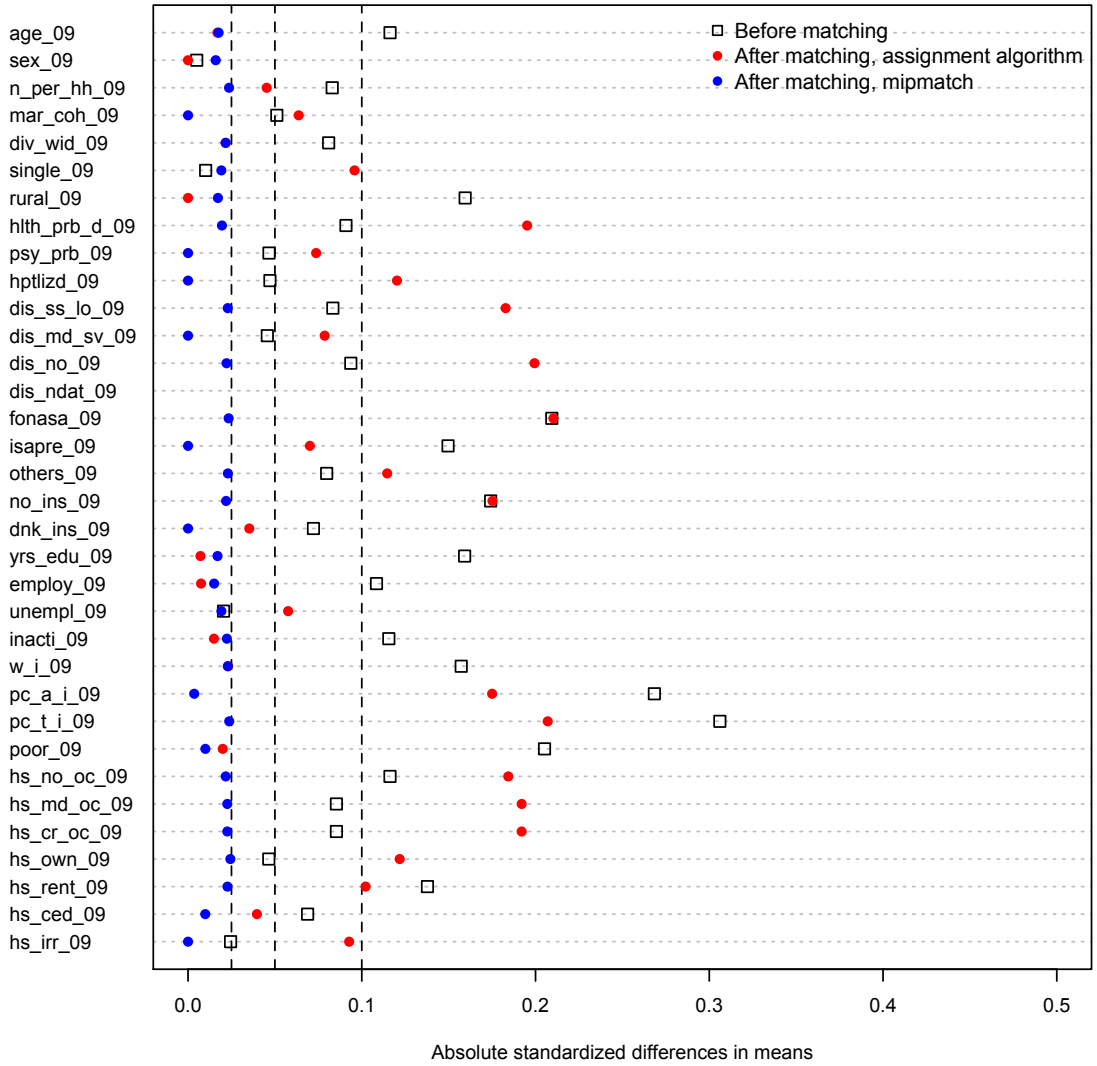
Balancing the means of the covariates (2)



Balancing the means of the covariates (2)



Balancing the means of the covariates (2)



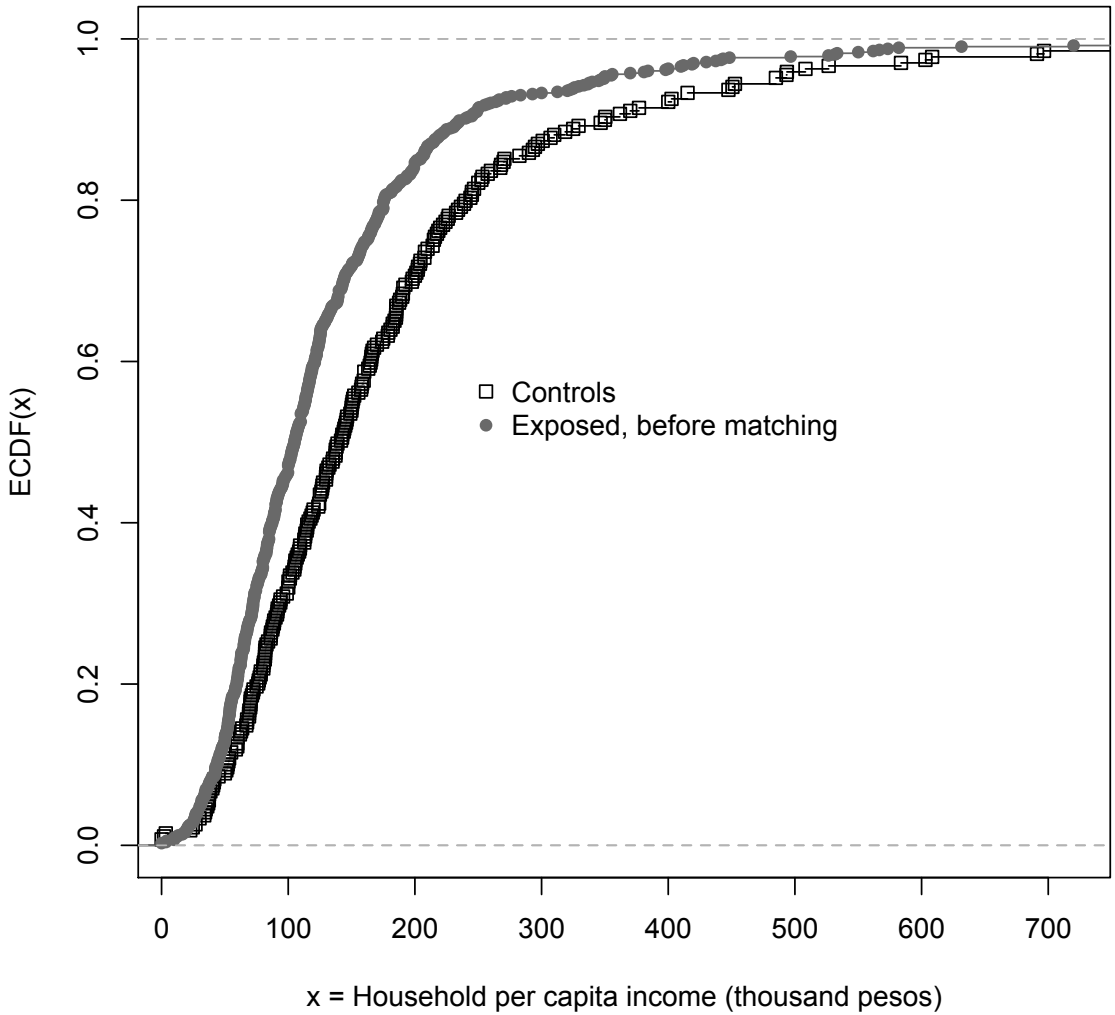
Balancing Kolmogorov-Smirnov statistics (1)

$$\sum_{j \in \mathcal{J}} \omega_j \mu_j(\mathbf{a}) = \omega_j \sup_{x_{c,p} \in \mathcal{G}(\mathbf{x}_{\mathcal{T},p})} \left| F_{\mathcal{T}}(x_{c,p}) - F_{\mathcal{C}}(x_{c,p}, \mathbf{a}) \right|$$

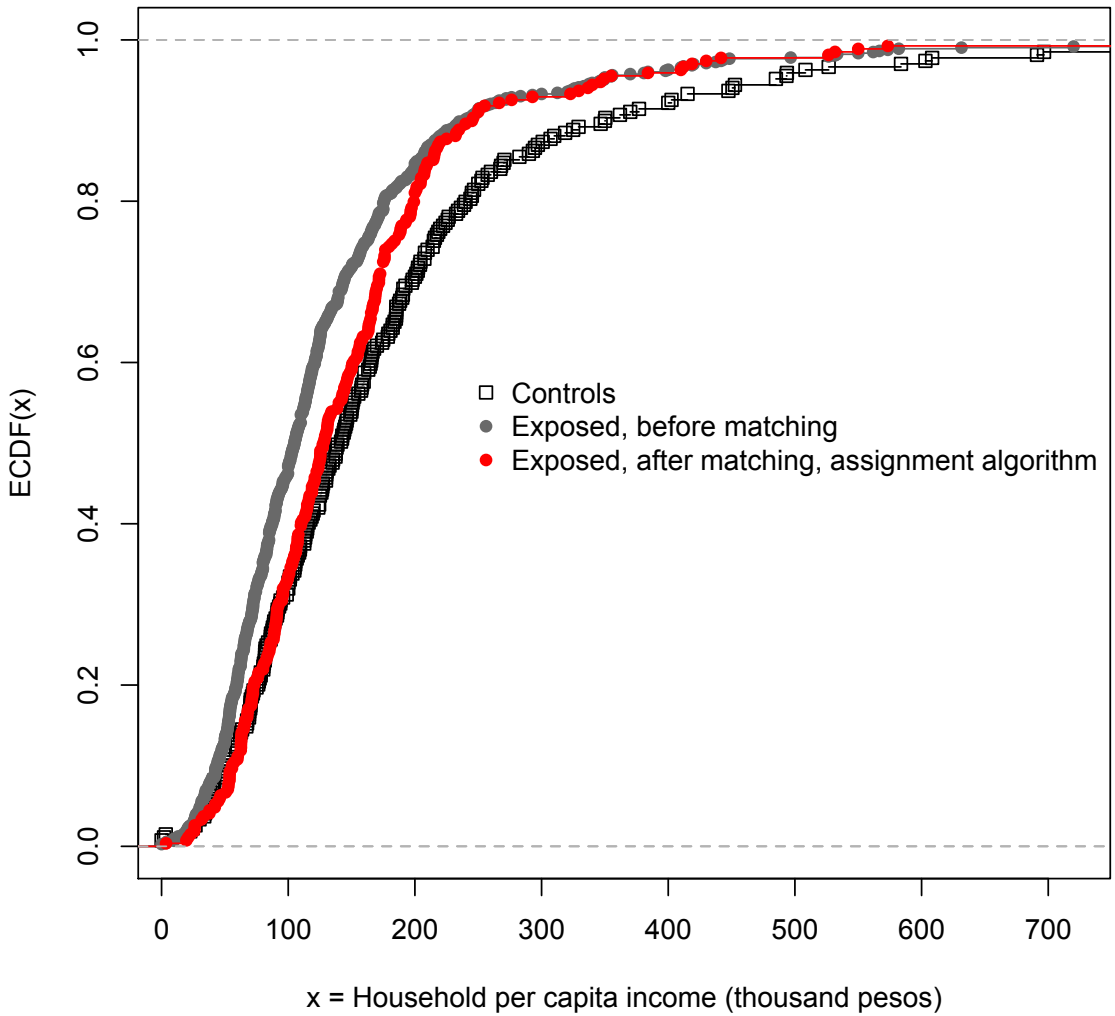
$$= \omega_j z_j$$

$$\geq \left| \frac{1}{|\mathcal{G}(\mathbf{x}_{\mathcal{T},p})|} - \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \frac{\mathbb{1}_{\{x_{g-1;p} \leq x_{c,p} < x_{g;p}\}} a_{t,c}}{mT} \right| \quad \forall x_{g;p} \in \mathcal{G}(\mathbf{x}_{\mathcal{T},p})$$

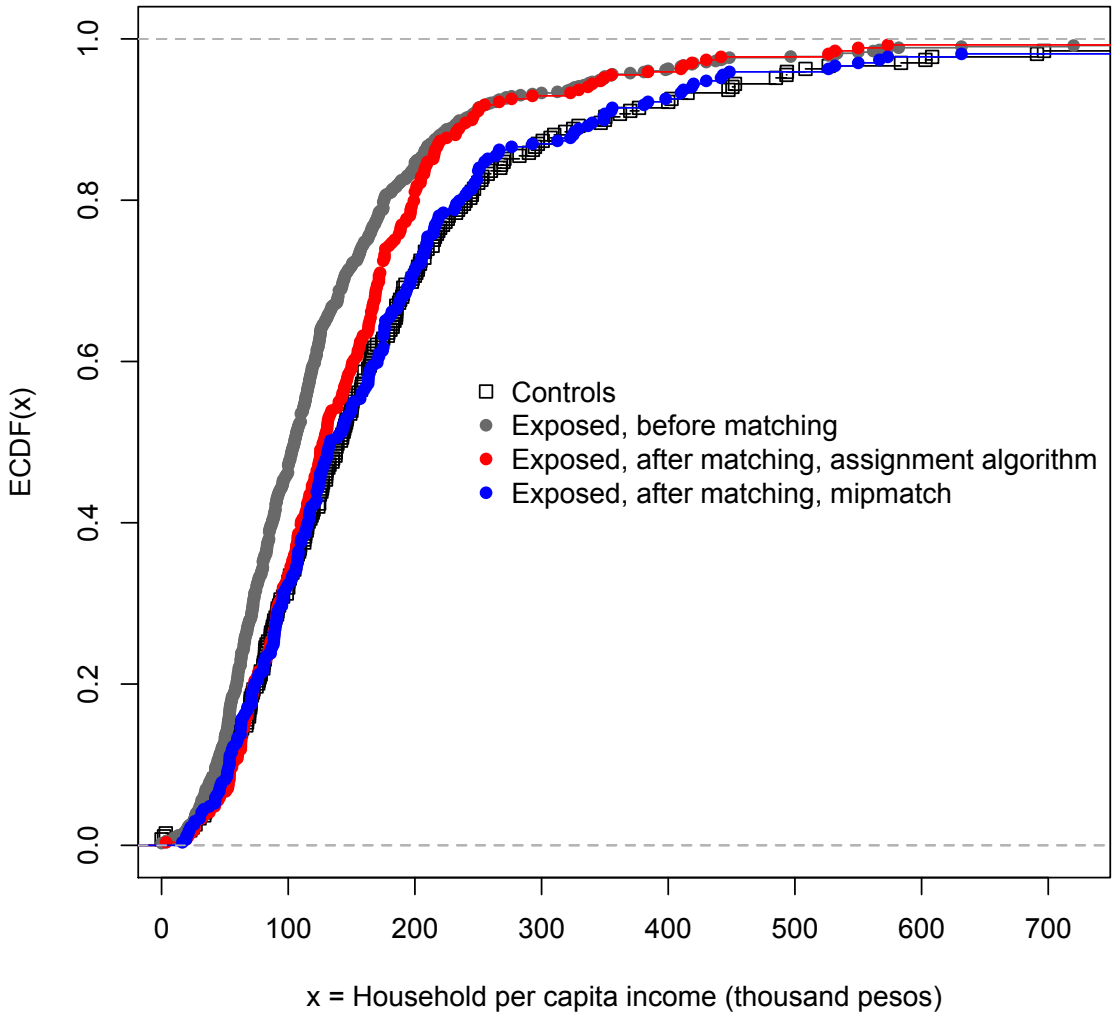
Balancing Kolmogorov-Smirnov statistics (2)



Balancing Kolmogorov-Smirnov statistics (2)



Balancing Kolmogorov-Smirnov statistics (2)



Fine and near-fine balance for several covariates (1)

Fine balance:

$$\sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} a_{t,c} \mathbb{1}_{\{x_{c,p}=b\}} = m \sum_{t \in \mathcal{T}} \mathbb{1}_{\{x_{t,p}=b\}} \qquad \forall b \in \mathcal{B}$$

Near-fine balance:

$$\left| \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} a_{t,c} \mathbb{1}_{\{x_{c,p}=b\}} - m \sum_{t \in \mathcal{T}} \mathbb{1}_{\{x_{t,p}=b\}} \right| \leq \xi \qquad \forall b \in \mathcal{B}$$

Fine and near-fine balance for several covariates (2)

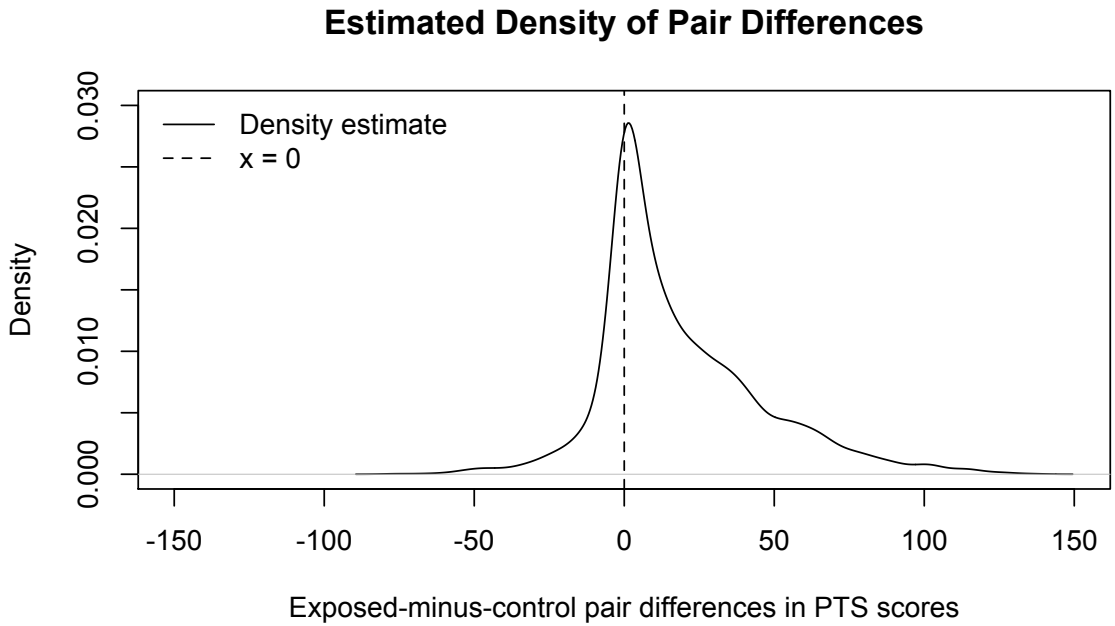
Table: Fine balance for self-rated health

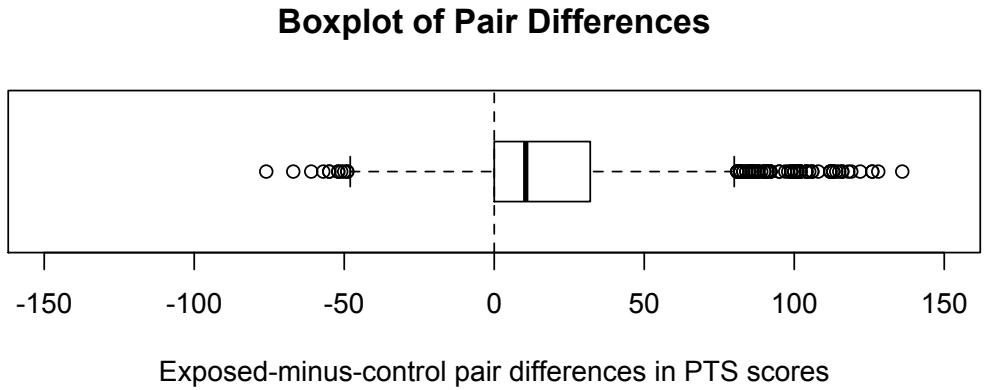
	Exposed	Controls
Poor	122	122
Good	1487	1487
Fair	911	911

Table: Fine balance for material quality of the housing

	Exposed	Controls
Acceptable	1738	1738
Unacceptable	739	739
Beyond repair	43	43

Density plot of PTS scores





Outline

Summary and remarks

- Explicitly optimize or constrain the criteria used to assess the quality of the match
  - Meet the criteria
  - Know that the criteria is infeasible
- Directly balance
  - Means
  - Variances and skewness
  - Correlations
  - Quantiles
  - Kolmogorov-Smirnov statistic
- While matching with exact, near-exact, fine and near-fine balance for more than one covariate
- A systematic method for improving covariate balance

Extensions

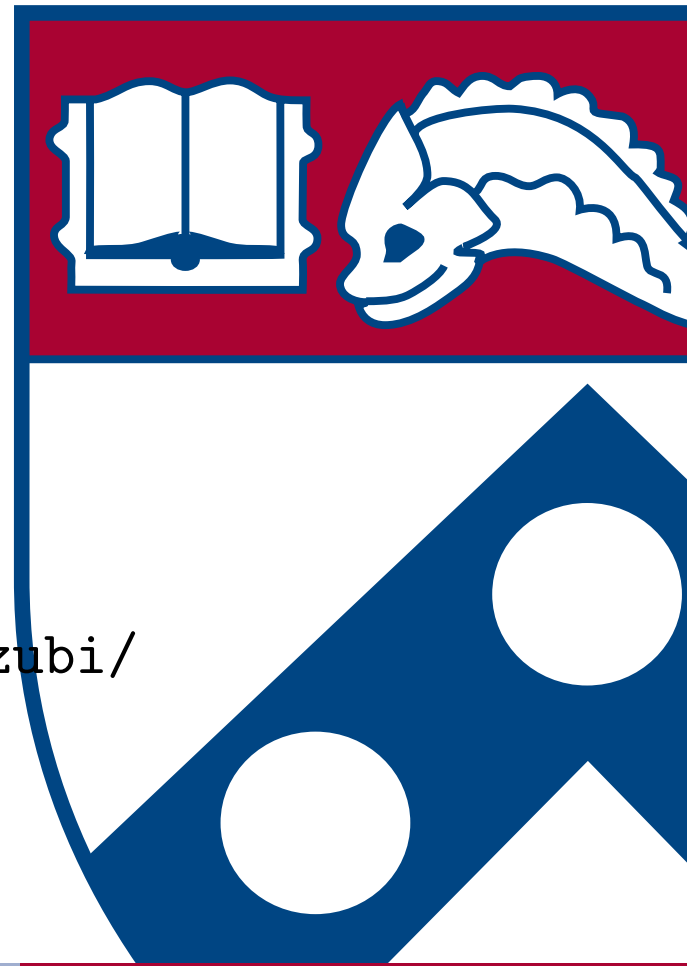
- Optimal subset matching
- Building a stronger instrumental variable
- Enhancing regression discontinuity designs
- `R` package `mipmatch`

References

- Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Acute Kidney Injury after Surgery," under revision.
- Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2012), "Effect of the 2010 Chilean Earthquake on Posttraumatic Stress: Designing an Observational Study to be Less Sensitive to Unmeasured Biases," under revision.
- Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2011), "Matching for Several Sparse Nominal Variables in a Case-Control Study of Readmission Following Surgery," *The American Statistician*, 65, 229-238.

Using Mixed Integer Programming for Matching in Observational Studies

José R. Zubizarreta
Department Statistics
The Wharton School
University of Pennsylvania



`http://www-stat.wharton.upenn.edu/~josezubi/`