How wrong are we? Using middle initials to estimate mismatch rates and reduce bias in regression coefficeints

Joshua R. Goldstein
BPC Mini-Conference

February 10, 2020

Known Unknowns

We can tolerate false matches, if we know how often we are are wrong.

Our Case: exact matching in CenSoc project

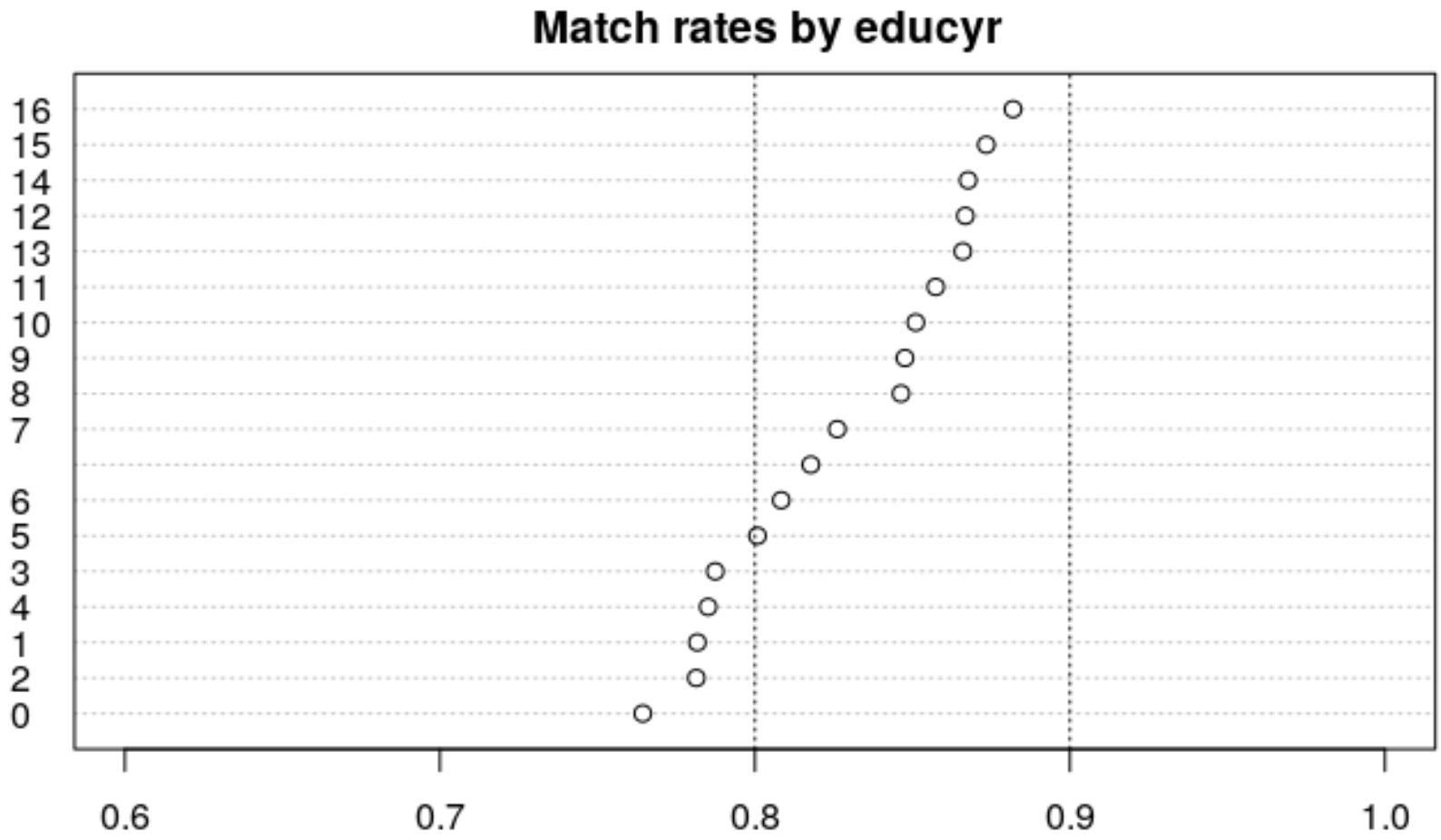
- We match 1940 census to Social Security Death File deaths 1975-2004
- Exact, unique, matches on first name, last name, year of birth, (place of birth)
- Because we don't use middle name, can use to check false match rate

A self-centered example

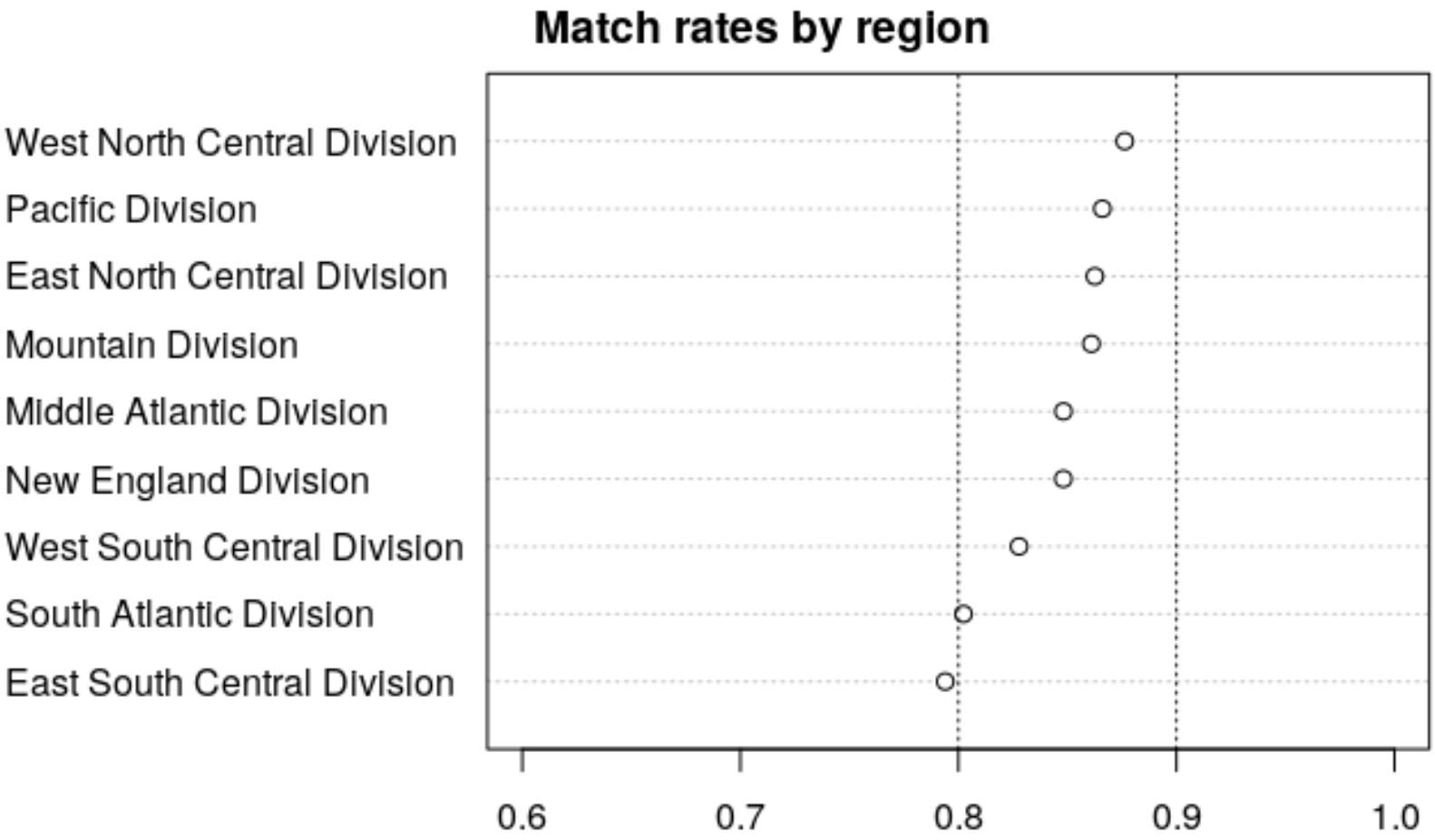
```
Joshua [R.] Goldstein
Josh [A.] Goldstein
```

```
Joshua [A.] Goldstein
Josh [R.] Goldstein
```

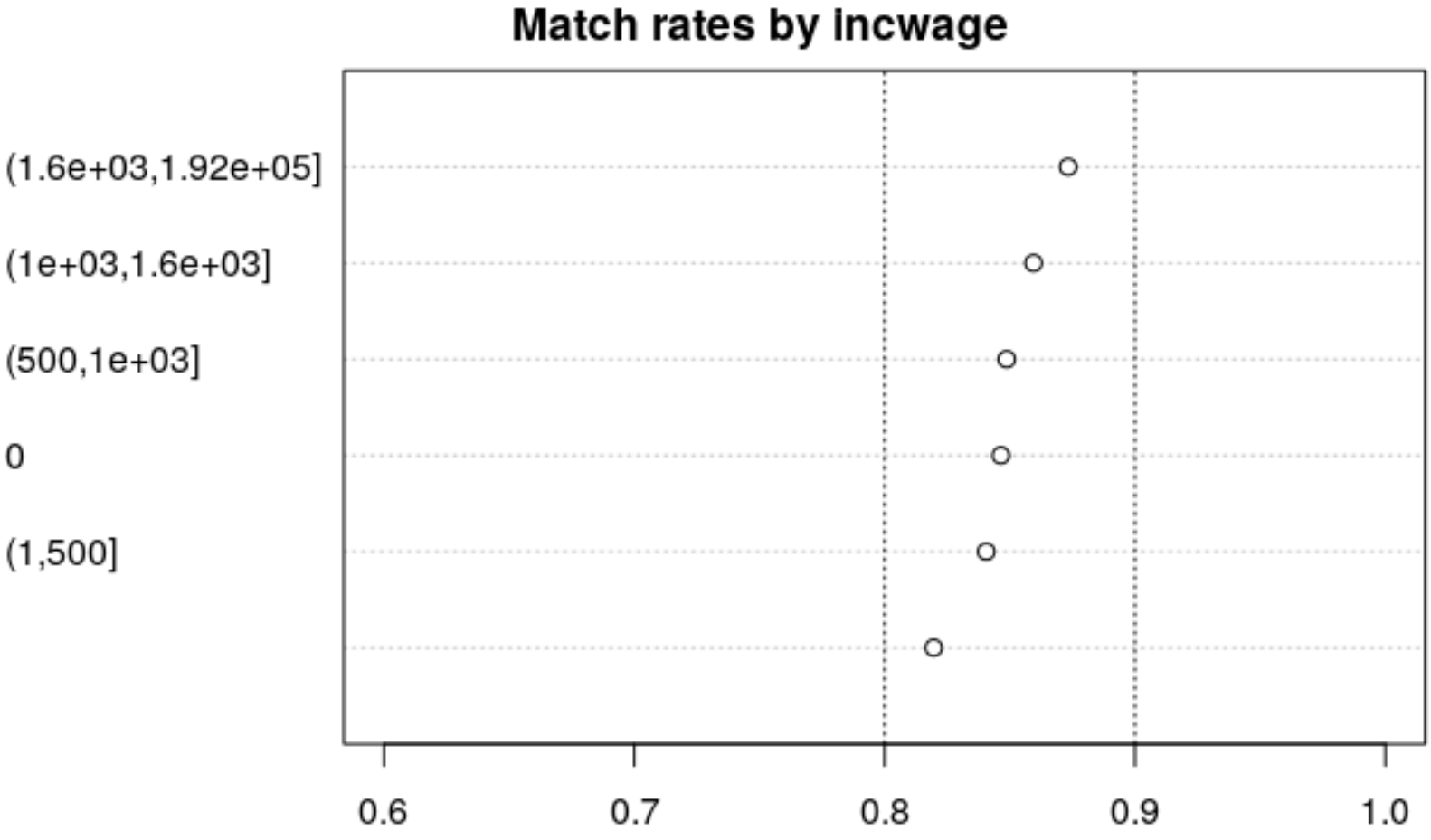
Patterns: Education



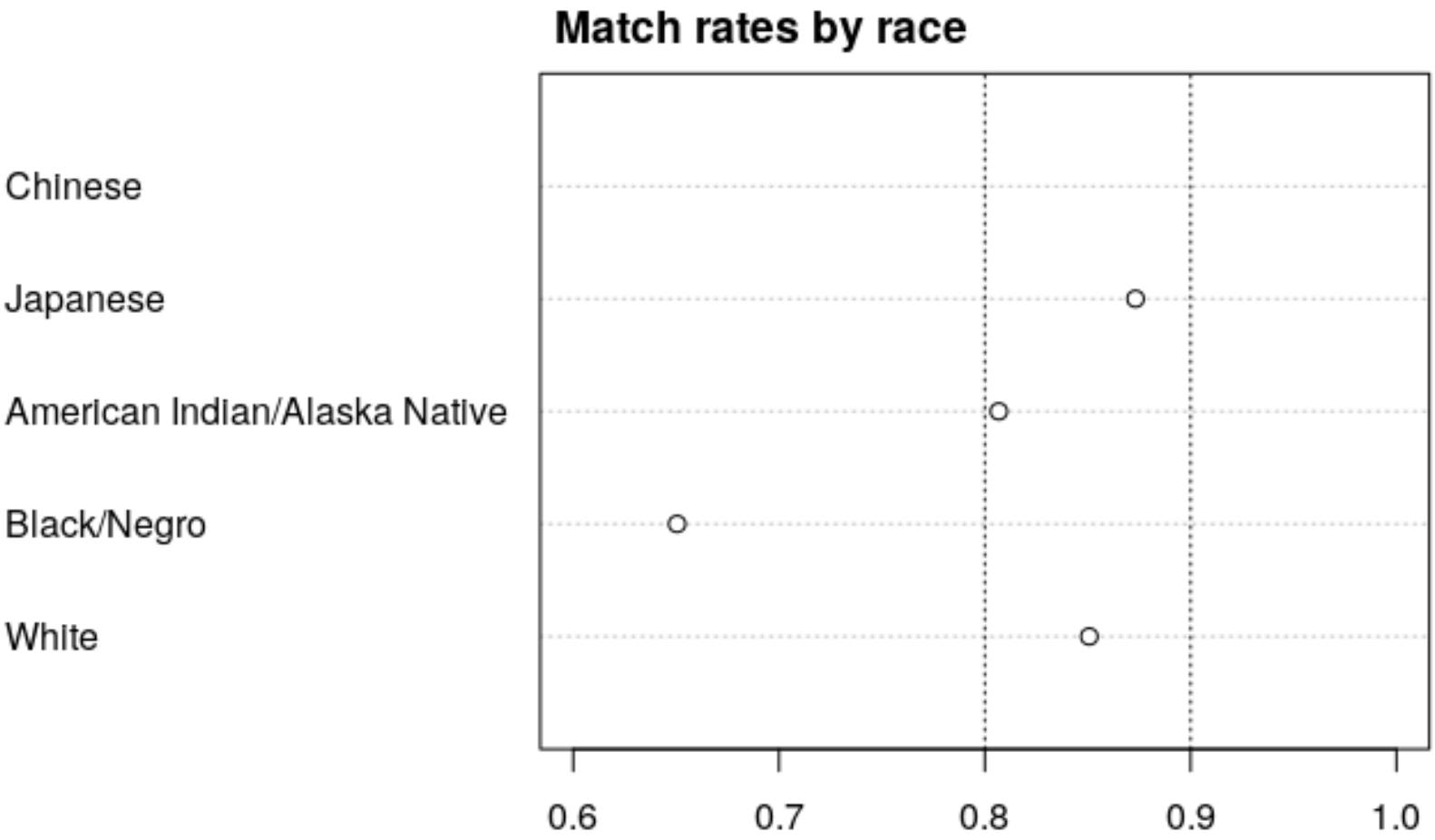
Patterns: Region



Patterns: Income



Patterns: Race



Take-away: Big Black-White Disparity

- Not sure why
- Regression analysis suggests it's not due to name frequencies
- Unstable reporting over time? (Name, birthyear?)
- Enumerator issues?

An Application

A regression of age at death on education

$$Y_i = \beta_0 + \beta_{ED} ED_i + \epsilon_i$$

But what if we have wrong person's education?
Can model as measurement error:

$$ED_j = ED_i + u_i$$

Can "unbias" the coefficients by dividing them by proportion "true matches"
The formula turns out to be

$$\hat{\beta}_{true} = \beta_{bias} \times \frac{1}{1 - \alpha_{\text{mismatch rate}}}$$

Black-White differences in the effect of education

	White	Black
β_{bias}	0.140	0.055

Black-White differences in the effect of education

	White	Black
β_{bias}	0.140	0.055
$\hat{\alpha}_j$	0.150	0.350

Black-White differences in the effect of education

	White	Black
β_{bias}	0.140	0.055
$\hat{\alpha}_j$	0.150	0.350
$\hat{\beta}_{true}$	0.165	0.085

Black-White differences in the effect of education

	White	Black
β_{bias}	0.140	0.055
$\hat{\alpha}_j$	0.150	0.350
$\hat{\beta}_{true}$	0.165	0.085

- So, difference appears not due to measurement error.
- "Real" explanations required to understand why education has smaller pay-off for Blacks than whites (e.g., lower quality schooling)

Conclusions

- Trade-offs: effort vs. sample bias vs. false-match rate, ...
- Perhaps false-matches not such a problem, if we can get good estimates of how often they occur.