

Reliable Causal Inference

Alexis J. Diamond

Key Points

- New matching method for causal inference
- Gives reliable results (in this data set)
- Use it to estimate all types of causal effects

I want to change the way you think about and do causal inference

Outline

- Intellectual history
 - Data: The NSW Experiment
 - Rubin Model and Matching
-
- My attempts to replicate prior work
 - Quest for balance: a new matching method
 - Results for the NSW data
 - Future research

History Part I: Lalonde '86 Sparks a Crisis

- NSW job training experiment (1976-1977)
 - Random selection for treatment
 - Given the experiment, easy to make inferences about causal effects
- Experiments are rare in social science
- Do we get it right? Usually we don't know.
- NSW data could test econometrics as it's typically used: in *non-experimental* settings
 1. Get the true, experimentally-determined causal effect
 2. Substitute survey data for the control group
 3. Try to recover the correct answer with OLS, IV, etc.

History Part II: The Great Debate

- Lalonde (1986): OLS and IV regression fail
- Heckman and others make excuses
- Dehejia and Wahba 1999 (DW)
 - Propensity score matching gets it right
- Smith and Todd (2001)
 - DW cherrypicked their models to get it right
 - Propensity score matching is unreliable
- The debate rages, and the game continues

NSW Data Set

- Outcomes: income in 1978, two years after training (the treatment)
- Covariates
 - Age, years of education, marital status, pre-treatment income, black, hispanic, high-school degree
- Two different samples under consideration
 - Lalonde's original sample
 - 297 treated, 425 control
 - Dehejia & Wahba's sample (2 yrs of pretreat income)
 - 185 treated, 260 control
- Six different nonexperimental controls
 - CPS-1 (n = 16,000), CPS-2 (n = 2,300), & 3 (n = 500), and PSID-1, 2, and 3
 - Not very similar to the experimental sample: outcomes are measured differently

Randomized Experiments

- Gold standard for causal inference
- Balances all background characteristics: observed & unobserved
 - Treatment assignment is “strongly ignorable”
 - Treatment assignment and outcomes are conditionally independent of all **confounders**: characteristics used to assign treatment and related to outcomes
- Rubin Model adopts this experimental framework
 - What are the units?
 - What is the treatment & how is it assigned?
 - What are the outcomes?

Causal Inference as Missing Data Problem

True Potential Outcomes

<i>Unit</i>	<i>Tmt Status</i>	<i>Yobs</i>	<i>Y(0)</i>	<i>Y(1)</i>
Al	1	\$ 10	\$ 4	\$ 10
Beth	0	\$ 2	\$ 2	\$ 5
Carl	0	\$ 4	\$ 4	\$ 10

Observed Outcomes

<i>Unit #</i>	<i>Tmt Status</i>	<i>Yobs</i>	<i>Y(0)</i>	<i>Y(1)</i>
Al	1	\$ 10	?	\$ 10
Beth	0	\$ 2	\$ 2	?
Carl	0	\$ 4	\$ 4	?

Causal Inference

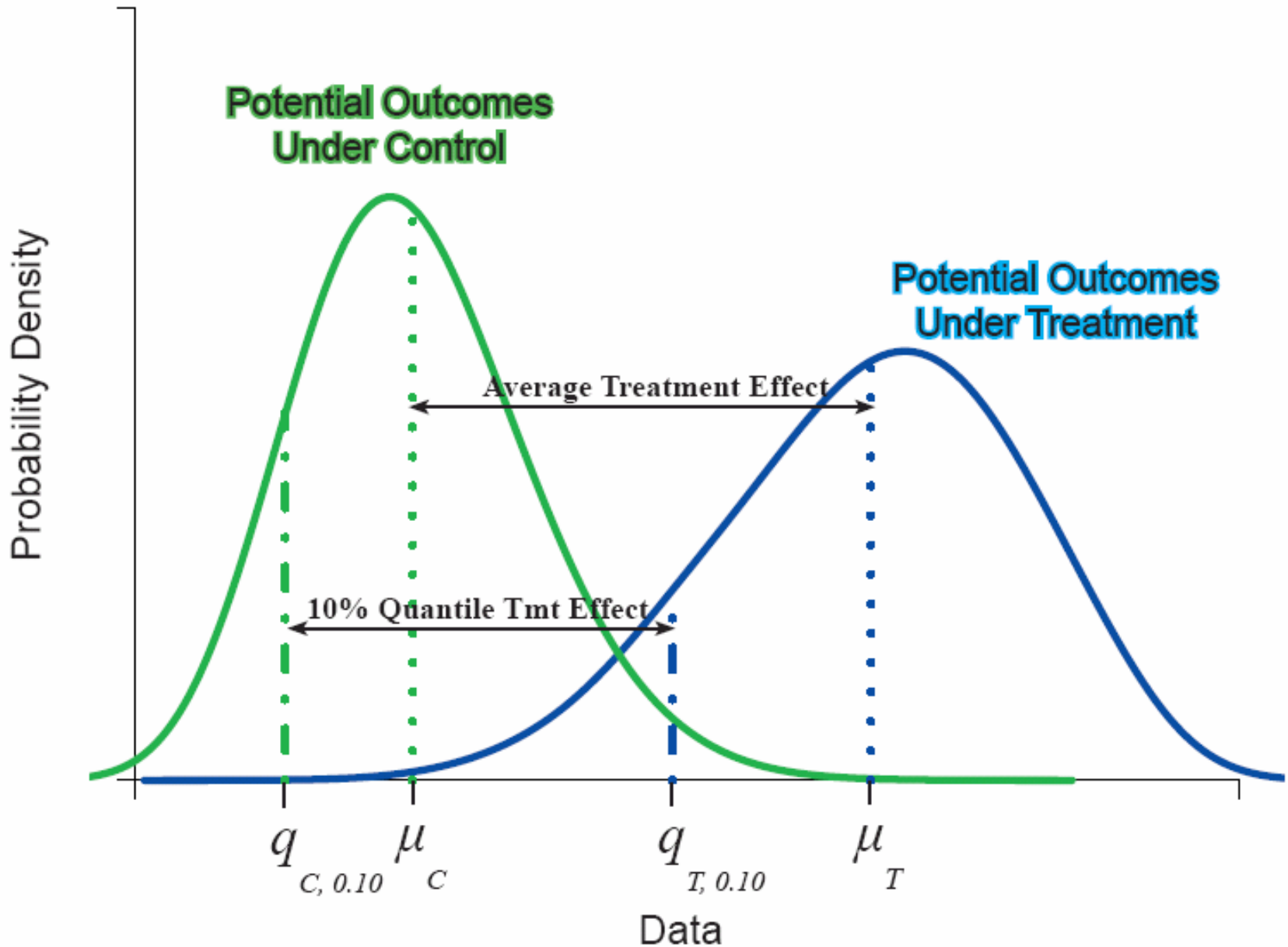
- Two sets of **potential outcomes**
 - Potential outcomes for the control
 - Potential outcomes for the treated

<i>Unit #</i>	<i>Tmt Status</i>	<i>Yobs</i>	<i>Y(0)</i>	<i>Y(1)</i>
Al	1	\$10	\$ 4	\$ 10
Beth	0	\$ 2	\$ 2	\$ 5
Carl	0	\$ 4	\$ 4	\$ 10

- Average treatment effect: $E [Y(1)] - E [Y(0)]$
- To get quantile treatment effects, same thing:
 - $\text{Quantile}_q [Y (1)] - \text{Quantile}_q [Y(0)]$

ATE vs. 10% Quantile Treatment Effect

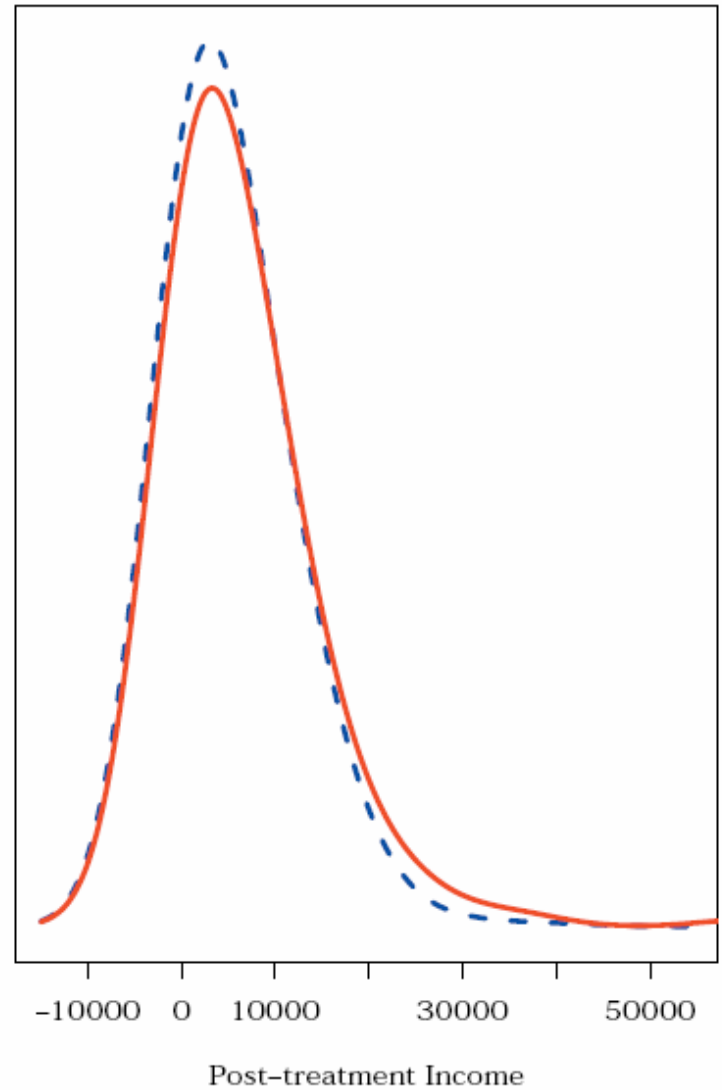
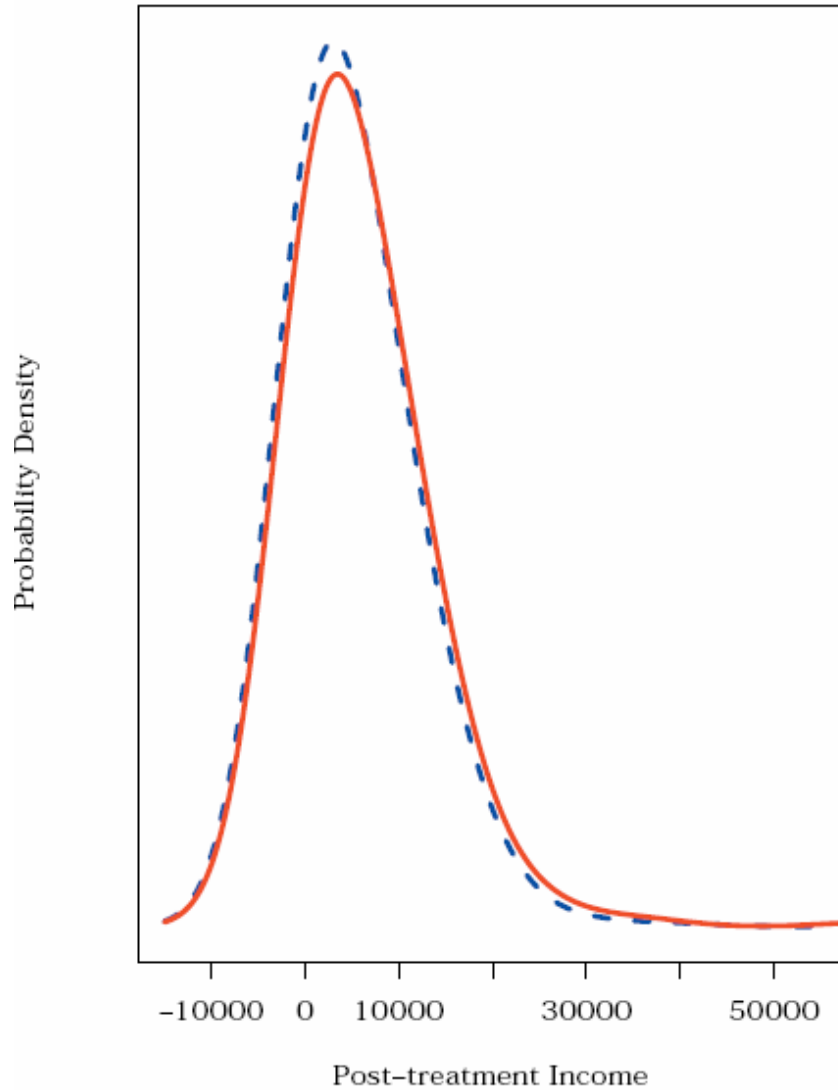
(10% chosen as an arbitrary example)



Distributions of Potential Outcomes in the NSW Experiment (Red = Treatment, Blue = Control)

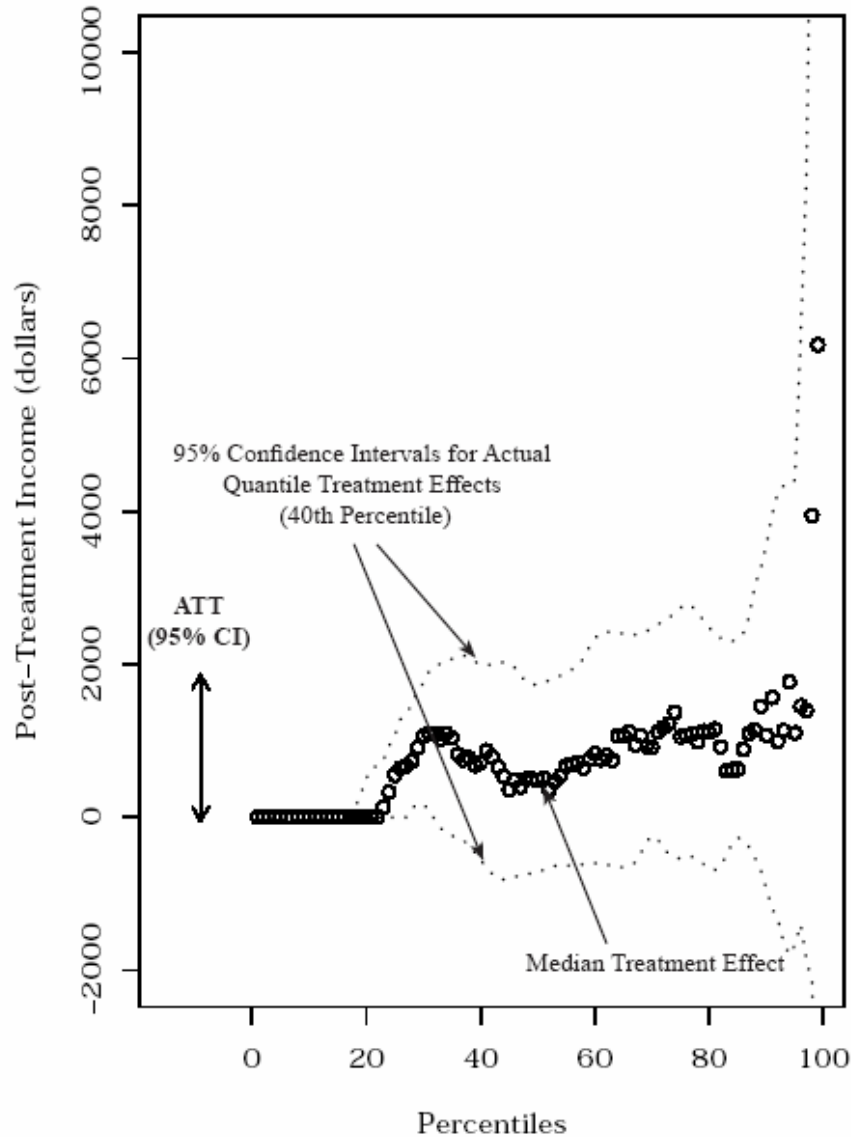
Lalonde Sample

Dehejia-Wahba Sample

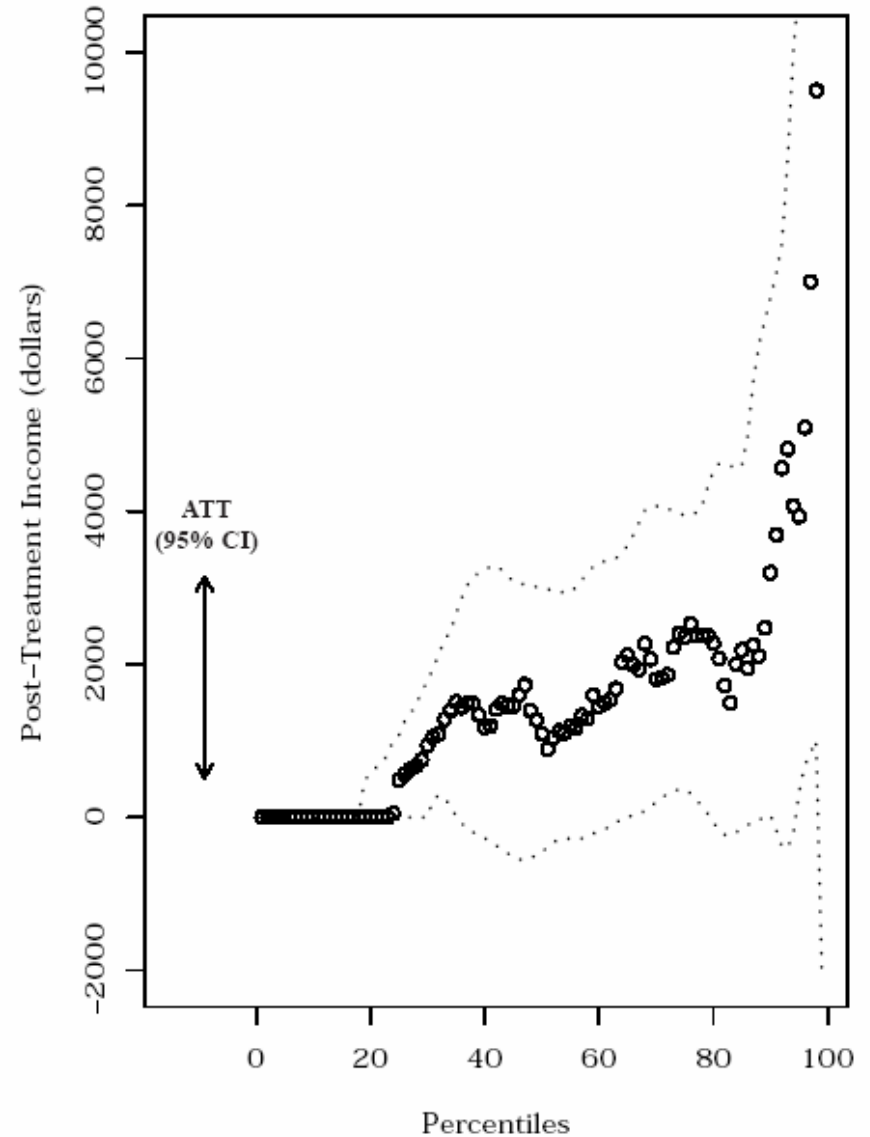


Average and Quantile Treatment Effects for the Treated Units Results of the Randomized NSW Experiment

Lalonde Sample



Dehejia and Wahba Sample



Inference in Nonexperimental Settings

- Selection bias and confounding
 - Assignment is non-random
 - No balance on all background characteristics
 - But inferential missing-data problem persists
 - Assume assignment involves only *observables* (X)

Nearest-neighbor matching on X : *Estimating the Effect for the Treated*

<i>Unit #</i>	<i>Tmt?</i>	<i>Age (X)</i>	<i>Yobs</i>	<i>Y(0)</i>	<i>Y(1)</i>
Al	1	28 years	\$ 10	\$ 4	\$10

Bill	0	56 years	\$ 2	\$ 2	?
Cathy	0	27 years	\$ 4	\$ 4	?

Propensity score matching

- Easier to match on one characteristic than many
 - Confounders include age, education, and maybe age*education, black*marriage, ... who knows?
- Rosenbaum & Rubin (1983) propensity score
 - A propensity score is a scalar balancing score
 - Balance on this, and you balance on all observed confounders, given enough data
 - Often estimated by fitting a logit for treatment assignment

DW: Balance Achieved?

***** V1 re74

before matching:
mean treatment. 2095
mean control... 5619
T-test pval.... 1.7478e-12
var ratio Tr/Co 0.51813

after matching:

mean treatment. 2095
mean control... 2664
T-test pval.... 0.20043
var ratio Tr/Co 1.0797

***** V3 married *****

before matching:
mean treatment. 0.18919
mean control... 0.51282
T-test pval.... < 2.22e-16
var ratio Tr/Co. 0.61589

after matching:

mean treatment. 0.18919
mean control... 0.19586
T-test pval.... 0.8264
var ratio Tr/Co 0.97397

** V5 educ **

before matching:
mean treatment.. 10.3
mean control.... 10.2
T-test pval. 0.58
var ratio Tr/Co.. 0.49

after matching:

mean treatment... 10.3
mean control 10.7
T-test pval 0.06
var ratio Tr/Co.. 0.57

**** V2 re75 ****

before matching:
mean treatment. 1532.1
mean control... 2466.5
T-test pval.... 0.0011501
var ratio Tr/Co.0.9563

after matching:

mean treatment. 1532.1
mean control... 1732.8
T-test pval.... 0.53607
var ratio Tr/Co 1.0647

***** V4 black *****

before matching:
mean treatment. 0.84324
mean control... 0.20280
T-test pval.... < 2.22e-16
var ratio Tr/Co 0.82014

after matching:

mean treatment. 0.84324
mean control... 0.84324
T-test pval.... 1
var ratio Tr/Co 1

** V6 black*educ

before matching:
mean treatment. 8.6
mean control... 2.04
T-test pval.. 0
var ratio Tr/Co 0.98

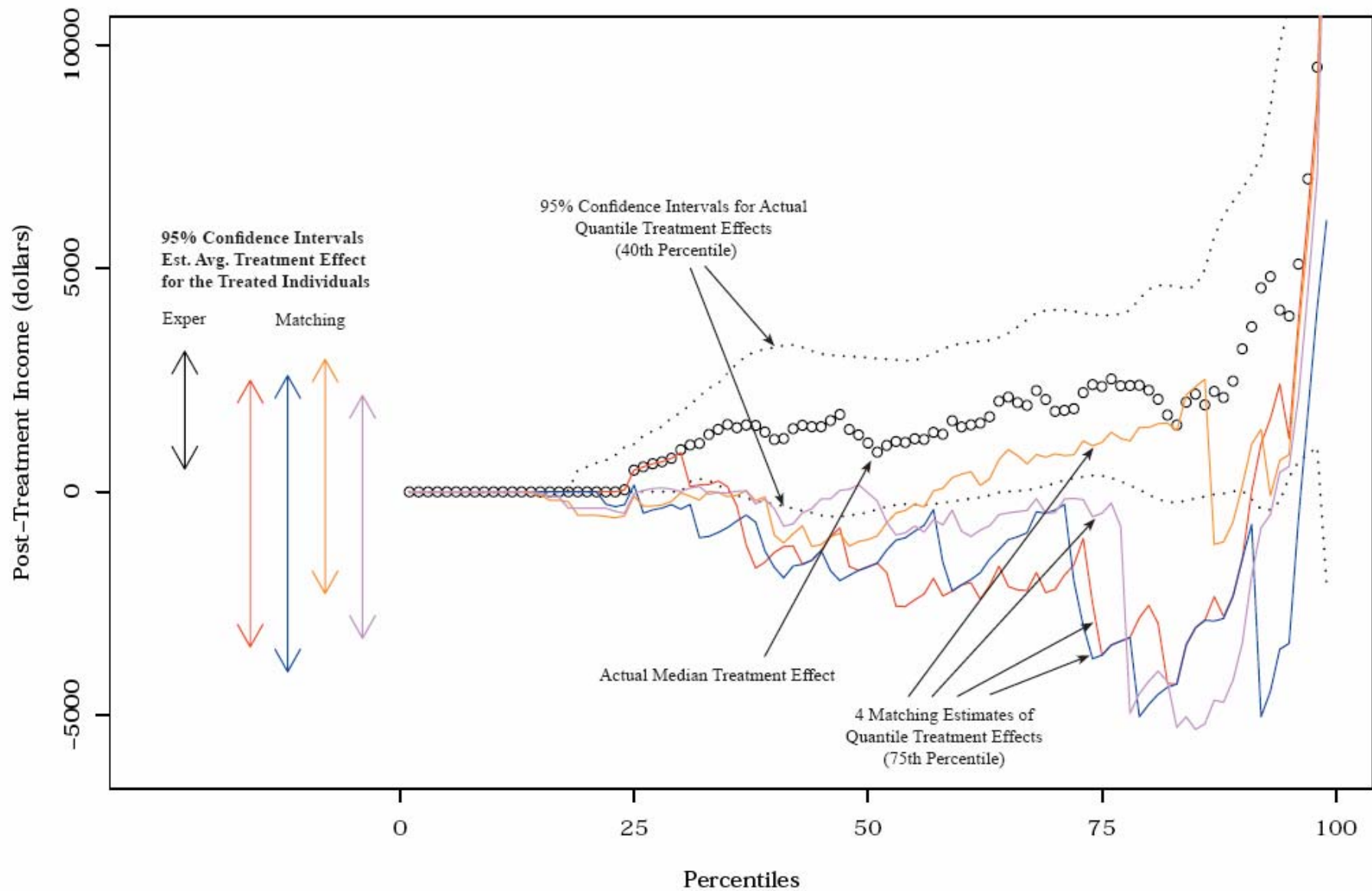
after matching:

mean treatment. 8.7
mean control.... 9.2
T-test pval 0.04
var ratio Tr/Co. 0.82

No Balance, No Reliability

- DW and Smith/Todd literature suggests 144 treated/control matches
 - 2 treated samples (Lalonde vs. DW)
 - 6 control data sets
 - 6 different models (confounders)
 - 2 different ways to match (full vs. p-score)
- Only 8 matched groups satisfy a very weak balance threshold
 - P-values on each confounder covariate must exceed 0.05
 - Variance ratio on each confounder covariate must be btw 0.5 and 1.5
- If this is the threshold, matching is unreliable

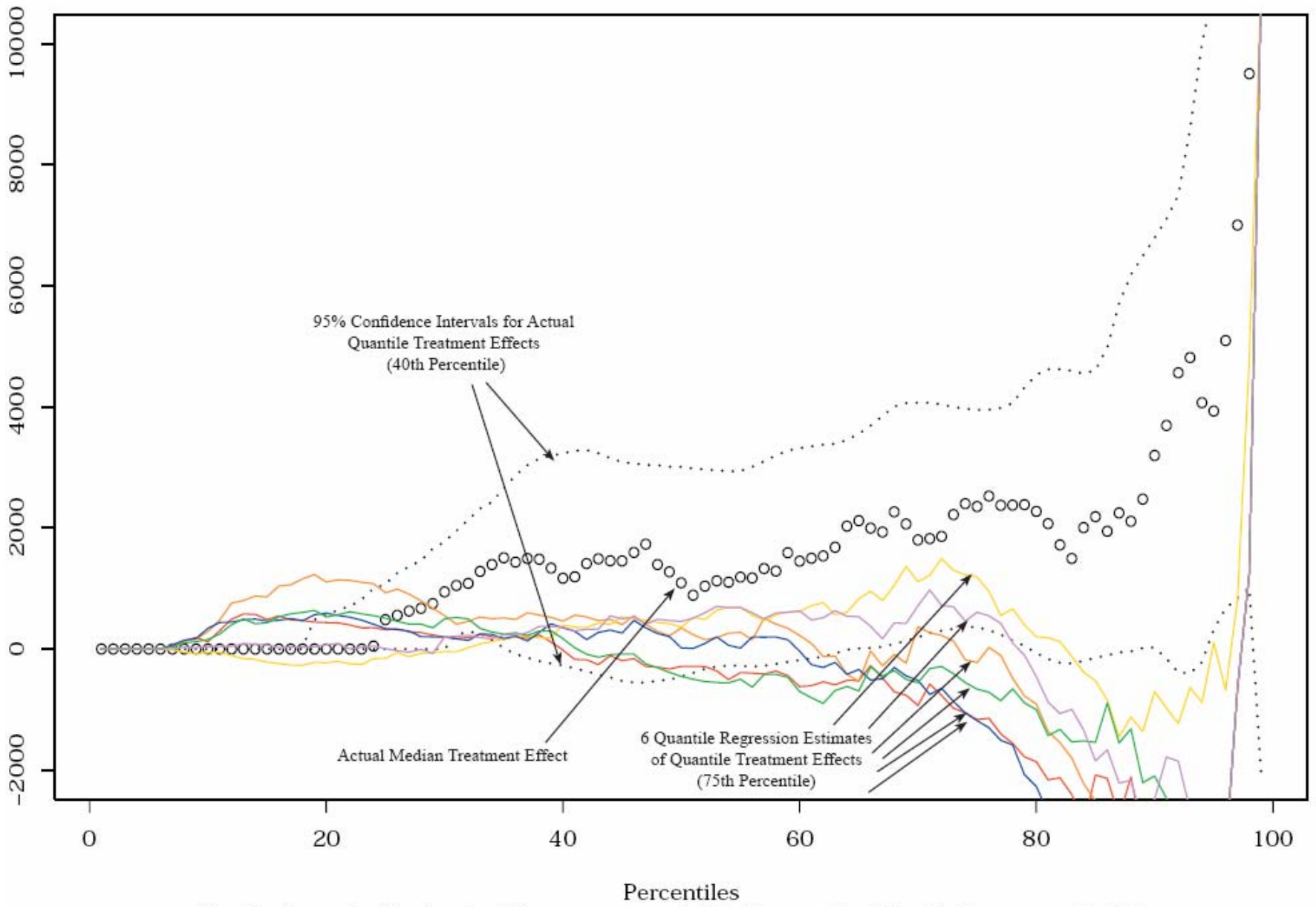
Results after Achieving a False Sense of Balance: Dehejia and Wahba Sample



Quantile Regression

- Alternative way to estimate quantile treatment effects
- Very much like regular OLS regression
 - Slightly different loss function
 - Median effects result from minimizing sum of absolute deviations
- Like OLS, quantile regression is vulnerable to misspecification error
 - Change the model, change the result

Quantile Regression Results: Dehejia and Wahba Sample



Results from the Randomized Experiment are in Black --- Matching Estimates are in Color

Is Reliable Inference Possible?

- Hypothesis: Yes, if balance truly obtains (& ignorability)
- It's a hard problem:
 - Of 16,000 control units, pick the correct 200 units that will create satisfactory balance
 - Must create balance across the distributions of all covariates, all their interactions, and all higher-order terms (very conservative)
 - Too many combinations to do exhaustive search
 - Need a smart agent to pick the correct control units for you

What Balance is Possible?

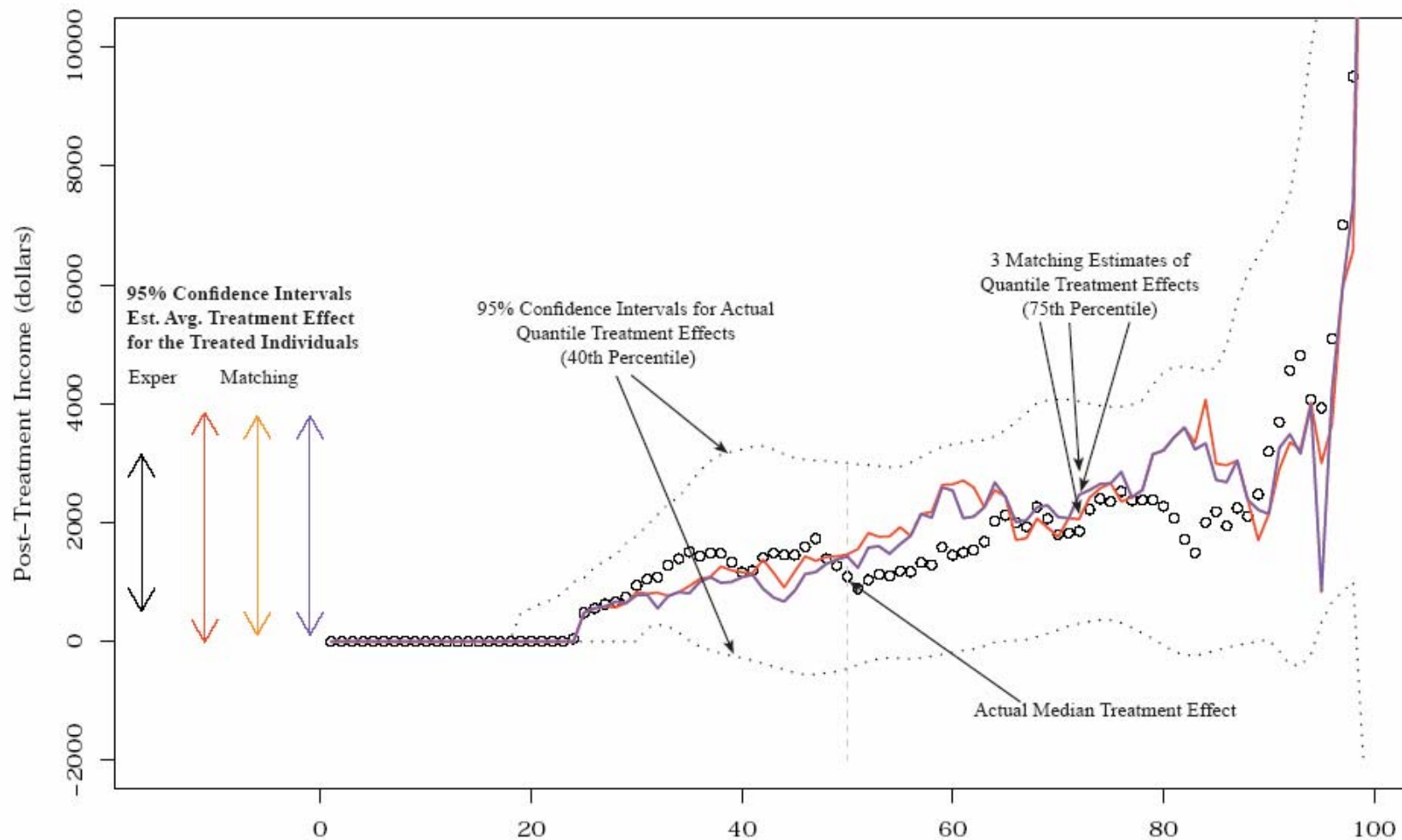
- Across all covariates
- All squared terms
- All interaction terms
- **Total of 36 variables**

Lowest p-value for t-test/KS test = 0.31

Better balance than with random assignment*

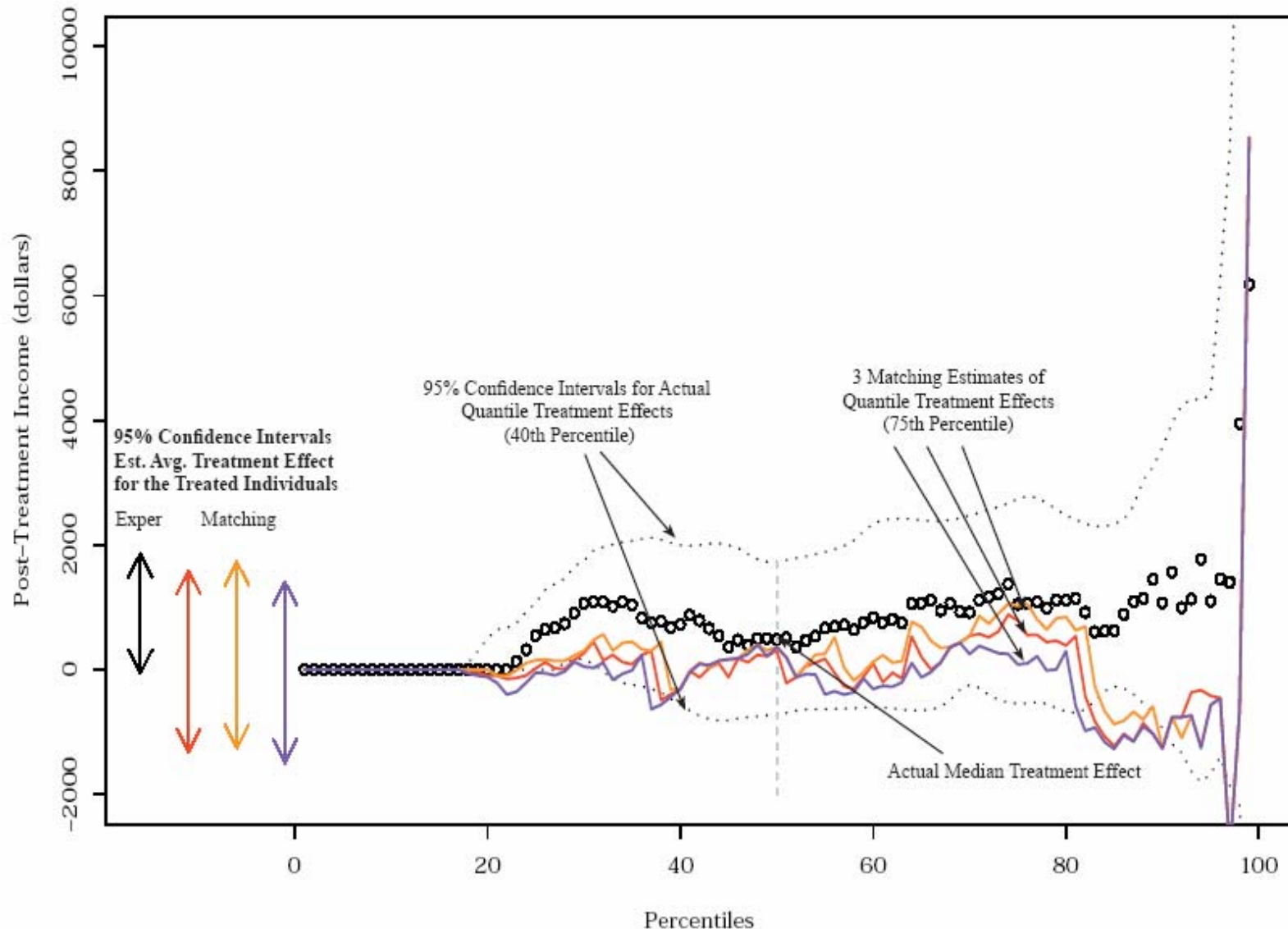
*Assuming assignment was ignorable– selection on observables

Average and Quantile Treatment Effects For the Treated: DW Sample (Controlling for 1974 Income)



Results from the Randomized Experiment are in Black --- Matching Estimates are in Color

Average and Quantile Treatment Effects for the Treated: Lalonde Sample (Not Controlling for 1974 Income)



Results from the Randomized Experiment are in Black --- Different Matching Estimates That Achieve Balance are in Color

New Method vs. Propensity Score

- New method does not estimate propensity scores or balancing scores
- Places observations in a complex space, and then matches
- The ends justify the means
- 2 thoughts
 - Propensity score matching and this new method are parallel paths: both can work
 - Which is easier to optimize for balance?
 - Symmetry: distorts the space to create a mirror image in the space of control units for cloud of treated units

Next Steps

- Complete paper (10 days, hopefully)
- Optimize the algorithm
- Incorporate into Jas's matching software
- Rebalance and re-estimate everything
- Figure out how to estimate standard errors