estimating individual causal effects

patrick lam

april 10, 2013

roadmap

- the what and why of ICEs
- estimation
- simulations
- application

contributions

1. reorient our thinking away from estimating average treatment effects first
2. coherent framework to think about individual causal effects
3. model that combines existing methods
4. practical applications to discover treatment effect heterogeneity and recover any causal quantity

the typical empirical paper

- "the effect of W on Y is $\hat{\beta}$"
- "the treatment effect is $\hat{\tau}$"

q: what is $\hat{\beta}$ or $\hat{\tau}$ estimating?

a: β or τ is usually the **average treatment effect** (ATE)

sometimes, it's an ATE on a subset of the population:

- average treatment effect on the treated (ATT)
- conditional average treatment effect (CATE)
- local average treatment effect (LATE)

but really, what exactly is an average treatment effect?

potential outcomes framework (again)

suppose a binary treatment variable W.

each individual i has a potential outcome associated with treatment $Y_i(1)$ and control $Y_i(0)$:

$$\tau_i = Y_i(1) - Y_i(0)$$

τ_i is an **individual causal effect** (ICE).

fundamental problem of causal inference: at most one potential outcome is ever observed for each individual

the average treatment effect is...

the average of the individual treatment effects:

$$\tau_{ATE} = E[\tau_i] = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$$

i	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	$Y_1(1)$	$Y_1(0)$	τ_1
2	$Y_2(1)$	$Y_2(0)$	τ_2
3	$Y_3(1)$	$Y_3(0)$	τ_3
4	$Y_4(1)$	$Y_4(0)$	τ_4
5	$Y_5(1)$	$Y_5(0)$	τ_5
6	$Y_6(1)$	$Y_6(0)$	τ_6

the ATE is the difference between the averages of the second and third columns OR equivalently the average of the fourth column

the average treatment effect is NOT...

- the treatment effect of any specific individual
- the treatment effect of the average individual

but we probably care more about these quantities!

q: given this, why do we use the average treatment effect?

a: probably because

- the ATE is *probably* the "best" *general* one number summary
- the ATE is usually the easiest to estimate
- the ATE is equal to the treatment effect for everybody IFF one makes the constant treatment effects assumption:

$$\tau_{ATE} = \tau_1 = \tau_2 = \cdots = \tau_N$$

although often implicit in language, rarely assumed explicitly and almost never reasonable

estimating the average treatment effect

observed data:

i	W	$Y(1)$	$Y(0)$
1	1	$Y_1(1)$	
2	0		$Y_2(0)$
3	0		$Y_3(0)$
4	1	$Y_4(1)$	
5	1	$Y_5(1)$	
6	0		$Y_6(0)$

assume ignorability of treatment assignment and SUTVA.

$$\hat{\tau}_{ATE} = E[Y(1)|W = 1] - E[Y(0)|W = 0]$$

observed outcomes are a random sample from each column so $\hat{\tau}_{ATE}$ is an unbiased estimate of τ_{ATE}.

what does the ATE miss?

suppose we observe the following data.

i	W	Y
1	1	15
2	0	10
3	0	15
4	1	8
5	1	10
6	0	8

$=$

i	W	$Y(1)$	$Y(0)$
1	1	15	?
2	0	?	10
3	0	?	15
4	1	8	?
5	1	10	?
6	0	?	8

$$\begin{aligned}\hat{\tau}_{ATE} &= E[Y(1)|W=1] - E[Y(0)|W=0] \\ &= 11 - 11 \\ &= 0\end{aligned}$$

what does the true underlying world look like?

what does the ATE miss?

with $\tau_{ATE} = 0$, the true underlying world can be

i	W	$Y(1)$	$Y(0)$	τ
1	1	15	15	**0**
2	0	10	10	**0**
3	0	15	15	**0**
4	1	8	8	**0**
5	1	10	10	**0**
6	0	8	8	**0**

(a) world 1

OR

i	W	$Y(1)$	$Y(0)$	τ
1	1	15	10	**5**
2	0	15	10	**5**
3	0	8	15	**-7**
4	1	8	15	**-7**
5	1	10	8	**2**
6	0	10	8	**2**

(b) world 2

- we often naturally think ($\tau_{ATE} = 0$) $\equiv$ world 1 (constant effects assumption); because of our priors? some type of cognitive bias?
- without additional information, $P(\text{world 1}) = P(\text{world 2})$
- world 1 and world 2 have very different implications academically and policy-wise

what can be done?

this is a problem of **treatment effect heterogeneity**.

possible solutions:

1. estimate conditional ATEs, where we estimate the ATE for a subset of individuals defined by some covariate(s)
   - need to define the covariate(s) ourselves based on what we think affects the heterogeneity
   - possibly run lots of models to search for heterogeneity
2. estimate the **individual causal effects** (ICEs)

difference:

- estimating CATEs (top-down): define a subset and estimate a subsetted ATE
- estimating ICEs (bottom-up): estimate effects for every individual and aggregate/explore different subsets

individual causal effect: $\tau_i = Y_i(1) - Y_i(0)$

why estimate ICEs?

- we usually care about the effect on specific individuals, the average individual, or groups of individuals, but not the ATE
- discover and explore treatment effect heterogeneity
- bridges the gap between quantitative and qualitative research
- every other causal quantity is a simple function of ICEs, so we can calculate other estimands directly

why not estimate ICEs?

- strictly speaking, not identified
- hard to estimate
- only an in-sample quantity, hard to generalize to the population of individuals not in data without additional assumptions

if we had the ICEs (τ_i)...

- ATE: $\frac{\sum_{i=1}^{N} \tau_i}{N}$
- ATT: $\frac{\sum_{i \in \{W_i=1\}} \tau_i}{N_t}$
- $\text{CATE}_{\{X=1\}}$: $\frac{\sum_{i \in \{X_i=1\}} \tau_i}{\sum \mathbb{I}(X_i=1)}$
- $\text{P}(\tau_i > 0)$
- relationship between X and τ: scatterplot of X and τ_i

estimating ICEs

problem: ICEs are not identified in the data

- the data does not give any information to distinguish whether τ_i = -1000, 0, or 9999.8 since we do not observe both potential outcomes.

strategy: get a sense of the plausible values for the ICEs by

1. assuming that similar observations (on covariates) have similar potential outcomes (**matching**)
2. using a bayesian model to combine possible prior beliefs with information about potential outcomes from these observations to derive a posterior distribution for the ICEs

not really a new idea but focus has rarely ever been on ICEs before

the idea (more simply)

observed and unobserved (missing) data:

i	W	$Y(1)$	$Y(0)$	τ_i
1	1	Y_1	Y_1^{mis}	?
2	0	Y_2^{mis}	Y_2	?
3	0	Y_3^{mis}	Y_3	?
4	1	Y_4	Y_4^{mis}	?
5	1	Y_5	Y_5^{mis}	?
6	0	Y_6^{mis}	Y_6	?

- fill in missing potential outcomes (Y^{mis}) by imputation
- τ_i and any other causal estimand can be calculated given Y_i and Y_i^{mis}
- builds on something Rubin has done in a number of papers
  - Rubin (2005), Rubin and Waterman (2006), Jin and Rubin (2008), Pattanayak, Rubin, and Zell (2012), Gutman and Rubin (2012)

spoiler

embed in a bayesian model {

1. match
2. impute Y_i^{mis}
3. calculate τ_i
4. repeat for all i

}

none of these are new ideas!

how nature generated the data

1. draw and fix values of $X_i^{(p)}$, W_i, and τ_i for $i = 1, \ldots, N$

$$X_i^{(p)} = \{X_i, X_i^{(u)}\}$$

where X_i and $X_i^{(u)}$ are our observed and unobserved prognostic covariates (that predict the outcomes)

2. generate outcomes by

$$\begin{aligned} Y_i &= h(X_i^{(p)}) \\ Y_i^{mis} &= h(X_i^{(p)}, \tau_i) \quad \text{for } W_i = 0 \end{aligned}$$

$$\begin{aligned} Y_i &= h(X_i^{(p)}, \tau_i) \\ Y_i^{mis} &= h(X_i^{(p)}) \quad \text{for } W_i = 1 \end{aligned}$$

where $h(\cdot)$ is an unknown function

assumptions

- data are a finite sample of size N drawn from the data generating process described, so only look at sample estimands
- ignorability of treatment assignment

$$(Y(1), Y(0)) \perp W|X$$
$$\tau \perp W|X$$
$$X^{(u)} \perp W|X$$

- SUTVA: no interference & same version of treatment across i

estimation framework

model the missing potential outcomes as

$$Y_i^{mis} \sim f(\cdot|\theta_i^{mis}, X_i, W_i)$$

where randomness is derived from not observing $X_i^{(u)}$ and θ_i^{mis} is the mean of the distribution $f(\cdot)$

translation: assume that observations j which have the same values on X and the opposite treatment assignment as i have observed outcomes that follow the same distribution as Y_i^{mis}

$$\begin{aligned} Y_j &\sim f(\cdot|\theta_j, X_j, W_j) \\ \theta_j &= \theta_i^{mis} \end{aligned}$$

if $X_i = X_j$ and $W_i \neq W_j$

- find these "donor" observations via matching
- same process regardless of whether i is treated or control

estimation overview

1. think of estimating each τ_i as a separate "study" where we have data consisting of observation i and all observations j where $W_i \neq W_j$
2. choose a matching procedure $\mathcal{M}$
3. using $\mathcal{M}$, construct a donor pool for i consisting of observations j that are "close" on the covariates X
4. model the mean of the donor pool
5. draw an imputation for $\tilde{Y}_i^{mis}$ from $f(\cdot)$ given the mean
6. calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$
7. repeat for all i

- incorporate in a bayesian sampler and repeat to simulate from the entire posterior distribution of τ_i for uncertainty
- each observation can be used in multiple donor pools but only once within any particular donor pool

bayesian model

$$p(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y, X, W) \propto p(Y|X, W, \theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}) p(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M})$$

- θ^{mis} is the vector of θ_i^{mis}, which is of interest
- $\theta_{\mathcal{M}}$ is a vector of parameters within the matching method
- data does not tell us anything about the choice of $\mathcal{M}$ so it is purely prior driven

steps:

1. simulate from the posterior via MCMC
2. draw values of $\tilde{Y}_i^{mis}$ from the posterior predictive distribution $f(\cdot)$ given the marginal posterior for θ^{mis}
3. simulate the posterior for τ (vector of τ_i) by calculating $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$

deriving the posterior

augment data with N binary variables $D^{(i)}$ for $i = 1, \ldots, N$

$$D_j^{(i)} = \begin{cases} 1 & \text{if } W_j \neq W_i \ \& \ j \text{ is a match to } i \\ 0 & \text{otherwise.} \end{cases}$$

example: suppose we want to match 1-to-1 on a single variable X

i	W	X	Y	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$	$D^{(4)}$	$D^{(5)}$	$D^{(6)}$
1	1	5	Y_1	0	0	0	0	0	1
2	0	3	Y_2	0	0	0	1	0	0
3	0	2	Y_3	0	0	0	0	1	0
4	1	3	Y_4	0	1	0	0	0	0
5	1	2	Y_5	0	0	1	0	0	0
6	0	5	Y_6	1	0	0	0	0	0

$D^{(i)}$ is an indicator for whether an observation is a match to the ith observation when estimating τ_i

how data augmentation helps

the data likelihood (conditional on X and W):

$$\begin{aligned} \mathcal{L}(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y) &= p(Y|\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}) \\ &= \text{intractable} \end{aligned}$$

augment with the variables D to get complete data likelihood:

$$\begin{aligned} \mathcal{L}_{comp}(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y, D) &= p(Y, D|\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}) \\ &= p(Y|D, \theta^{mis})p(D|\theta_{\mathcal{M}}, \mathcal{M}) \end{aligned}$$

actual (observed) likelihood averages over D:

$$\begin{aligned} \mathcal{L}(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y) &= \int p(Y|D, \theta^{mis})p(D|\theta_{\mathcal{M}}, \mathcal{M})dD \\ &= \text{tractable} \end{aligned}$$

the complete data likelihood

- complete likelihood is likelihood if we observed D
- uncertainty in D comes only from matching uncertainty

$$\begin{aligned}\mathcal{L}_{comp}(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y, D) &= p(Y|D, \theta^{mis})p(D|\theta_{\mathcal{M}}, \mathcal{M}) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N}\left[p(Y_j|\theta_j^{mis})^{D_j^{(i)}} p(D_j^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})\right]\end{aligned}$$

- for any specific τ_i, observed Y_i is fixed, Y_j is random for donor j because of unobserved $X^{(u)}$, non-donor observations don't matter
- any particular Y_j can appear in the likelihood multiple times or not at all
- outer product assumes independence of each "study" (each ICE is estimated independently)

priors

$$p(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}) = \left[\prod_{i=1}^{N} p(\theta_i^{mis})\right] p(\theta_{\mathcal{M}}) p(\mathcal{M})$$

- usually use improper uniform priors (bayesian model that approximates non-bayesian results)
- can easily incorporate qualitative priors
- prior on $\mathcal{M}$ reflects uncertainty over matching specification
- current use of matching almost always settles on one single specification $\equiv$ spike prior on $\mathcal{M}$ and $\theta_{\mathcal{M}}$
- possible to incorporate information from data and let $\mathcal{M}$ enter into likelihood via balance measures??

simulating from the posterior via MCMC

use a Gibbs sampler: **algorithm**: repeat the following n_{sim} times

1. draw a matching procedure $\tilde{\mathcal{M}}$ from

$$p(\mathcal{M}|Y, X, W, \theta_{\mathcal{M}}, D, \theta^{mis}) = p(\mathcal{M})$$

2. draw a value $\tilde{\theta}_{\mathcal{M}}$ from

$$p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M}, D, \theta^{mis}) = p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M})$$

captures estimation uncertainty of matching and of the parameters of the matching procedures

simulating from the posterior via MCMC

for (i in 1:N) $\{$

3. draw $\tilde{D}^{(i)}$ from

$$p(D^{(i)}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M}, D^{(-i)}, \theta^{mis}) = m(\theta_{\mathcal{M}}, \mathcal{M})$$

$D^{(i)}$ is a deterministic function of $\theta_{\mathcal{M}}$ and $\mathcal{M}$ (**matching**)

4. draw $\tilde{\theta}_i^{mis}$ from

$$p(\theta_i^{mis}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M}, D, \theta_{-i}^{mis}) = p(\theta_i^{mis}|Y_{\{D^{(i)}=1\}}, D^{(i)})$$

can use conjugacy here to model the mean of the donor pool

$\}$

end of Gibbs sampler steps here gives us one draw from the joint posterior $p(\theta^{mis}, \theta_{\mathcal{M}}, \mathcal{M}|Y, X, W)$

simulating from the posterior via MCMC

given $\tilde{\theta}^{mis}$, impute from the posterior predictive distribution:

for (i in 1:N) {

5. draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\tilde{\theta}_i^{mis})$ (**imputation**); captures uncertainty from not observing $X^{(u)}$
6. calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$

}

end up with n_{sim} draws $\tilde{\tau}_i$ (matrix of size $N \times n_{sim}$) from the posterior distribution of τ_i

after simulation

theoretically should check

- MCMC convergence
  - lots of parameters ($> N$)
  - can check convergence on ATE, ATT, etc.:
    non-convergence on aggregations $\Rightarrow$ overall non-convergence
    convergence on aggregations $\not\Rightarrow$ overall convergence
- balance on matching
  - lots of balance to check ($> N \times n_{sim}$)
  - unlike usual matching, we're not comparing distributions but rather one observation versus multiple observations
- number of times each observation is used as a donor
  - need to make sure results are not too reliant on very few unique observations as donors

need more research into these areas!

summary

- estimating ICEs are a good idea but very hard
- in the absence of identification, want to get at least some idea of the causal effects for each individual (τ_i)
- use semi-parametric approach: matching + bayesian model
- relies on the typical causal inference assumptions plus some parametric model assumptions
- can be used to predict ICEs for unobserved or future individuals but need assumptions about similarities of future to current individuals or some more parametric assumptions

some questions:

- hidden assumptions about the smoothness and/or variance of the distribution of ICEs in the data?
- matching in a bayesian framework logical?

simulation study

want to test:

- how well does the model recover ICEs under normal conditions
- horse race to compare how well different matching procedures (hold other parametric model assumptions constant)

the idea:

1. generate fake data with ICEs known
2. consider different ways of generating outcomes and different ways of generating treatment assignments (unconfounded and various confounded)
3. evaluate performance of model and different matching procedures on various metrics

choosing the matching procedure $\mathcal{M}$: method

- choice of method: distance metric and how to choose matches given distance
  - match on nearest neighbor mahalanobis distance
  - match on nearest neighbor predictive mean
  - match on nearest neighbor (linear) propensity score
  - subclassification on (linear) propensity score
- also compare to
  - (bayesian) linear regression imputation
  - no matching: use all observations with different treatment as donors

mahalanobis matching

mahalanobis distance for two observations i and j on X:

$$\Delta_M(x_i, x_j) = (X_i - X_j)^T S^{-1} (X_i - X_j)$$

where S^{-1} is the sample covariance matrix of X

- calculate $\Delta_M(x_i, x_j)$ for every i and j pair
- $D^{(i)} = 1$ for the M observations of the opposite treatment that have the smallest mahalanobis distance to i
- no variation in donor pool across iterations unless M or X varies
- most posterior variation likely coming from imputation step

predictive mean matching

embed two linear regression steps within the algorithm

1. regression of Y on X for treated observations: $Y_t = X_t \beta_t$
2. regression of Y on X for control observations: $Y_c = X_c \beta_c$

- for treated i, calculate $\tilde{\mu}_i = X_i \tilde{\beta}_c$ and $\tilde{\mu}_j = X_j \tilde{\beta}_c$ for all control observations j
- β_c is the estimated contribution of X on Y
- Y_i^{mis} is the outcome with only contributions from X
- $\tilde{\mu}_i$ is initial best guess of Y_i^{mis}
- match to control observations with similar "guesses"
- $D^{(i)} = 1$ for the M control observations with $\tilde{\mu}_j$ closest to $\tilde{\mu}_i$
- do the same for control i using $\tilde{\beta}_t$ instead

propensity score matching

propensity score: $e_i = P(W_i = 1|X_i)$

embed logistic regression of W on X within the algorithm

- calculate **linear** propensity score for all observations

$$\ln\left(\frac{\tilde{e}_i}{1-\tilde{e}_i}\right) = X_i\tilde{\beta}$$

- $D^{(i)} = 1$ for the M observations of the opposite treatment with the closest linear propensity scores to i
- variation in donor pools due to variation in $\theta_{\mathcal{M}}$

subclassification on propensity score

- calculate linear propensity score for all observations with the same process as before
- sort linear propensity scores and divide into M subclasses
- $D^{(i)} = 1$ for observations of the opposite treatment in the same subclass as i
- restrict each subclass to have at least two treated and two control observations
- if within an iteration, a subclass does not meet the restriction, reduce M by one for that iteration only

choosing the matching procedure $\mathcal{M}$: M and X

in addition to method, a specification of $\mathcal{M}$ also includes

- set of X variables to match on
  - should match on all confounding variables to satisfy ignorability assumption
  - possibly match on other prognostic variables (tradeoff between worse matches but more precise imputations)
- choice of number of matches or subclasses M (can be fixed or random)

performance metrics

- traditional performance metrics (coverage, bias, mse, etc.) do not really exist for bayesian models
- bayesian posteriors characterize probability of parameters
- results are distributions rather than point estimates and standard errors
- leverage bayesian calibration and decision theory: bayesian counterparts to traditional metrics
- no repeated sampling of data since theoretically individuals always have the same ICE (bayesian rather than frequentist)

i use the following performance metrics:

1. posterior mean "bias" ("bias")
2. expected error loss ("root mean squared error")
3. proportion of ICE credible intervals not including 0 ("power")
4. calibration of ICEs ("coverage")

posterior mean "bias" ("bias")

traditional bias: $E[\hat{\theta}] - \theta$

versus

posterior mean "bias": $E[\theta|X] - \theta$

- how far off from the truth is our "best" estimate?
- for ICEs, calculate the average posterior mean "bias"

expected error loss ("root mse")

traditional root mean squared error:

$$\sqrt{E[(\hat{\theta} - \theta)^2]} = \sqrt{\text{variance} + \text{bias}^2}$$

versus

expected error loss:

$$\sqrt{\int ((\theta|X) - \theta)^2 p(\theta|X) d\theta} \approx \sqrt{\frac{\sum([\tilde{\theta}|X] - \theta)^2}{n_{sim}}}$$

- akin to average distance from the truth for each of our posterior draws $\tilde{\theta}$
- for ICEs, calculate the average expected error loss

proportion of ICE credible intervals including 0 ("power")

- ask what proportion of the N 95% credible intervals contain 0
- true τ_i in my simulations vary, but are never exactly equal to 0
- traditional definition of power: given the null hypothesis is false, what is the probability of rejecting the null?
- here: given a non-zero τ_i, what is the probability of 0 being in the 95% credible interval?
- key differences:
  - calculate probability across i rather than across repeated samples
  - τ_i is different across i
- nevertheless, gets at some notion of "power"

calibration of ICEs ("coverage")

bayesian calibration:

- 95% credible interval represents 0.95 posterior probability of parameter being with the interval
- model calibration by testing whether future observations are within 95% credible interval 95% of the time

calibration:

- ask what proportion of the N 95% credible intervals contain the true τ_i
- model is well calibrated if proportion is close to 0.95

fake data generation

- 10 prognostic covariates:
  - $x_1 \sim \mathcal{N}(0, 2^2)$
  - $x_2 \sim \mathcal{N}(0, 1)$
  - $x_3 \sim \mathcal{N}(0, 1)$
  - $x_4 \sim \mathcal{U}(-3, 3)$
  - $x_5 \sim \chi^2_1$
  - $x_6 \sim \mathit{Bernoulli}(.5)$
  - $x_7 \sim \mathcal{N}(0, 1)$
  - $x_8 \sim \mathcal{N}(0, 1)$
  - $x_9 \sim \mathcal{N}(0, 1)$
  - $x_{10} \sim \mathcal{N}(0, 1)$
- linear, moderately non-linear, and very non-linear outcome equations
- unconfounded treatment assignment and confounded treatment assignment with linear and non-linear equations
- sample sizes of 100, 1000, and 5000
- 27 different datasets

9 different fake data generating processes

- outcome equations:
  1. $Y(0) = x_1 + x_2 + x_3 - x_4 + x_5 + x_6 + x_7 - x_8 + x_9 - x_{10}$
  2. $Y(0) = x_1 + x_2 + 0.2x_3x_4 - \sqrt{x_5} + x_7 + x_8 - x_9 + x_{10}$
  3. $Y(0) = (x_1 + x_2 + x_5)^2 + x_7 - x_8 + x_9 - x_{10}$

- treatment assignments:
  1. $p(W = 1) = 0.5$
  2. $\eta = x_1 + 2x_2 - 2x_3 - x_4 - 0.5x_5 + x_6 + x_7$
     $W = 1$ if $\eta > 0$; otherwise $W = 0$
  3. $\eta = 0.5x_1 + 2x_1x_2 + x_3^2 - x_4 - 0.5\sqrt{x5} - x_5x_6 + x_7$
     $W = 1$ if $\eta > 0$; otherwise $W = 0$

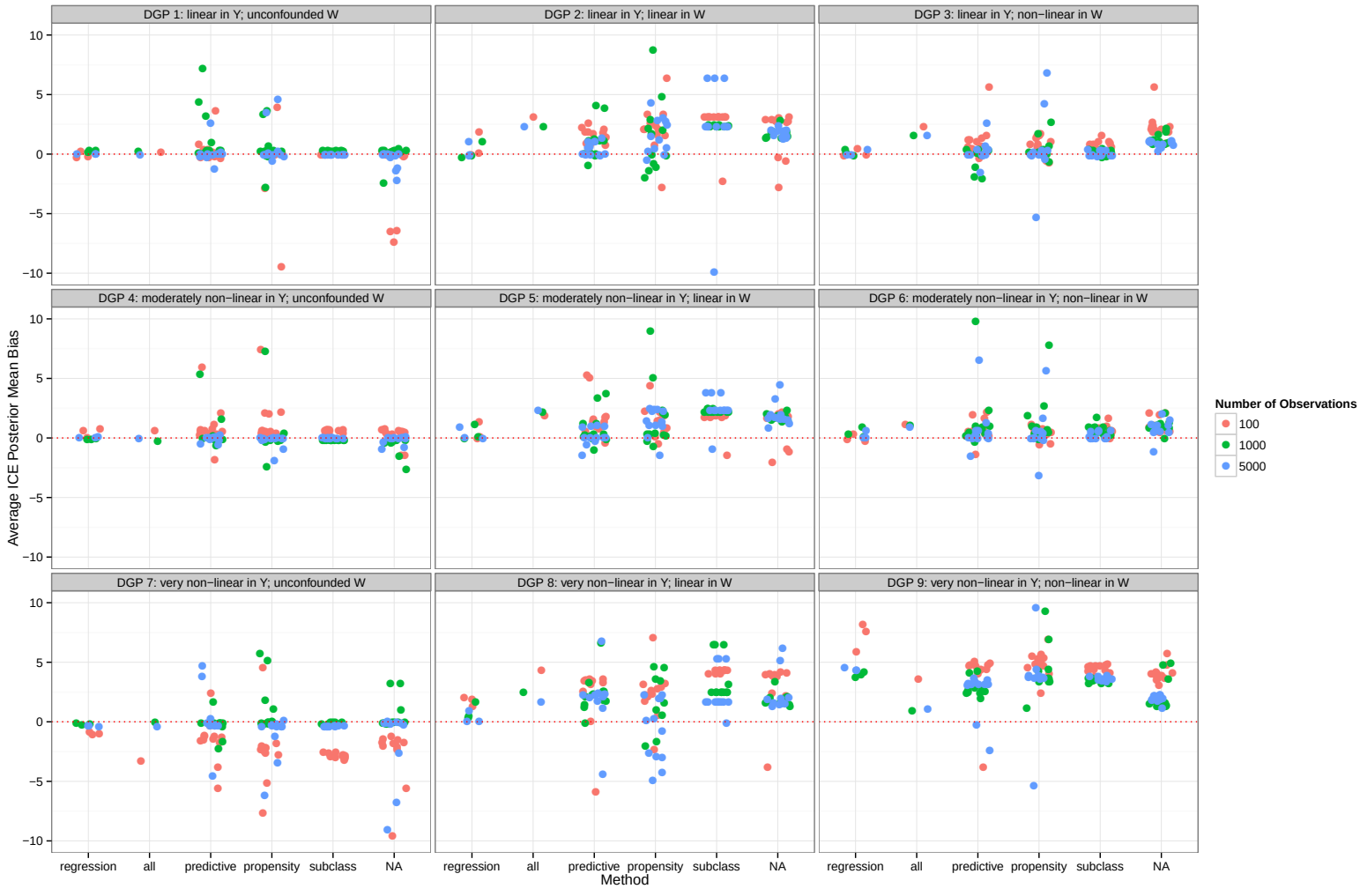
- generate true ICEs: $\tau_i \sim \mathcal{N}(2, (\sqrt{3})^2)$; also consider
  - $\tau_i \sim \mathcal{N}(20, (\sqrt{3})^2)$
  - $\tau_i \sim \mathcal{N}(2, (\sqrt{100})^2)$
  - $\tau_i \sim \mathcal{N}(20, (\sqrt{100})^2)$
  - mixtures

- $Y_i(1) = Y_i(0) + \tau_i$

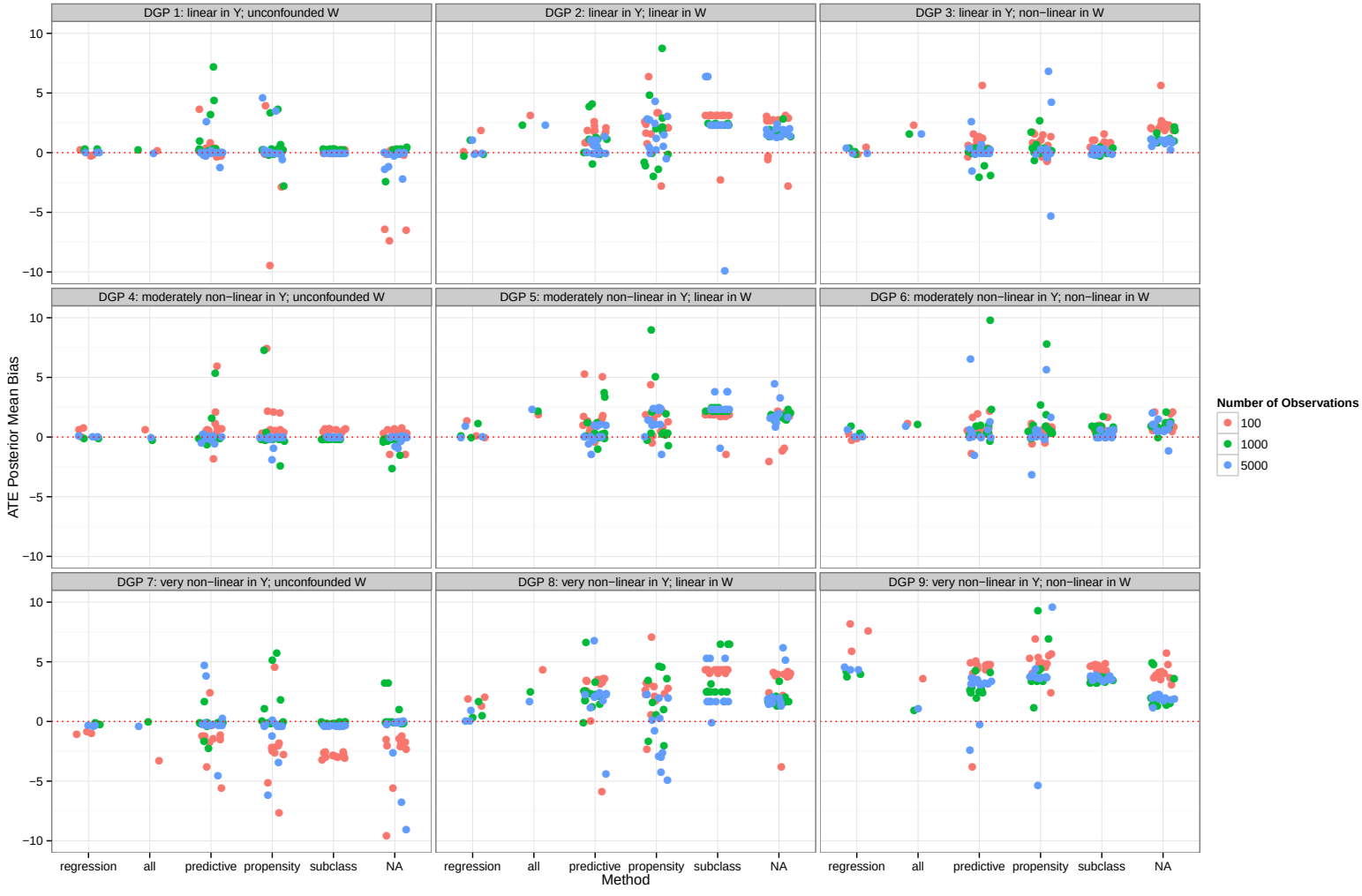
simulation comparison details

- 6 methods: regression imputation, no matching (all), 4 matching methods
- size of X: 5,**7**,10
- M: small, medium, large, random
- for matching, also consider M as a percentage of size smaller treatment group
- compare performance metrics for recovering ICEs and ATE
- 1816 different specifications
- MCMC chain length of $n_{sim} = 2000$

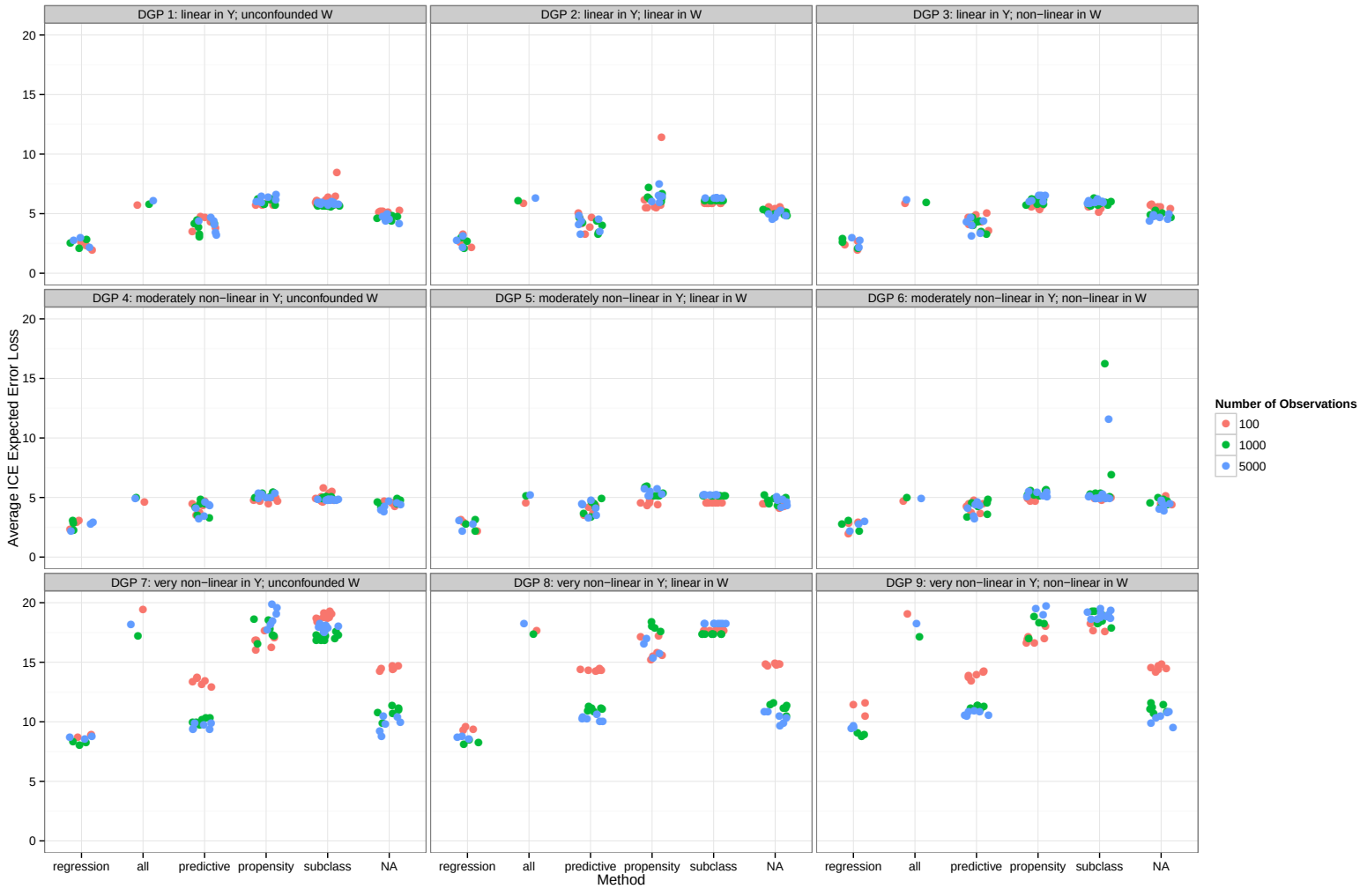
“bias” of ICEs


DGP 1: linear in Y; unconfounded W
DGP 2: linear in Y; linear in W
DGP 3: linear in Y; non−linear in W
DGP 4: moderately non−linear in Y; unconfounded W
DGP 5: moderately non−linear in Y; linear in W
DGP 6: moderately non−linear in Y; non−linear in W
DGP 7: very non−linear in Y; unconfounded W
DGP 8: very non−linear in Y; linear in W
DGP 9: very non−linear in Y; non−linear in W
Average ICE Posterior Mean Bias
10
5
0
−5
−10
regression
all
predictive
propensity
subclass
NA
Method
Number of Observations
100
1000
5000

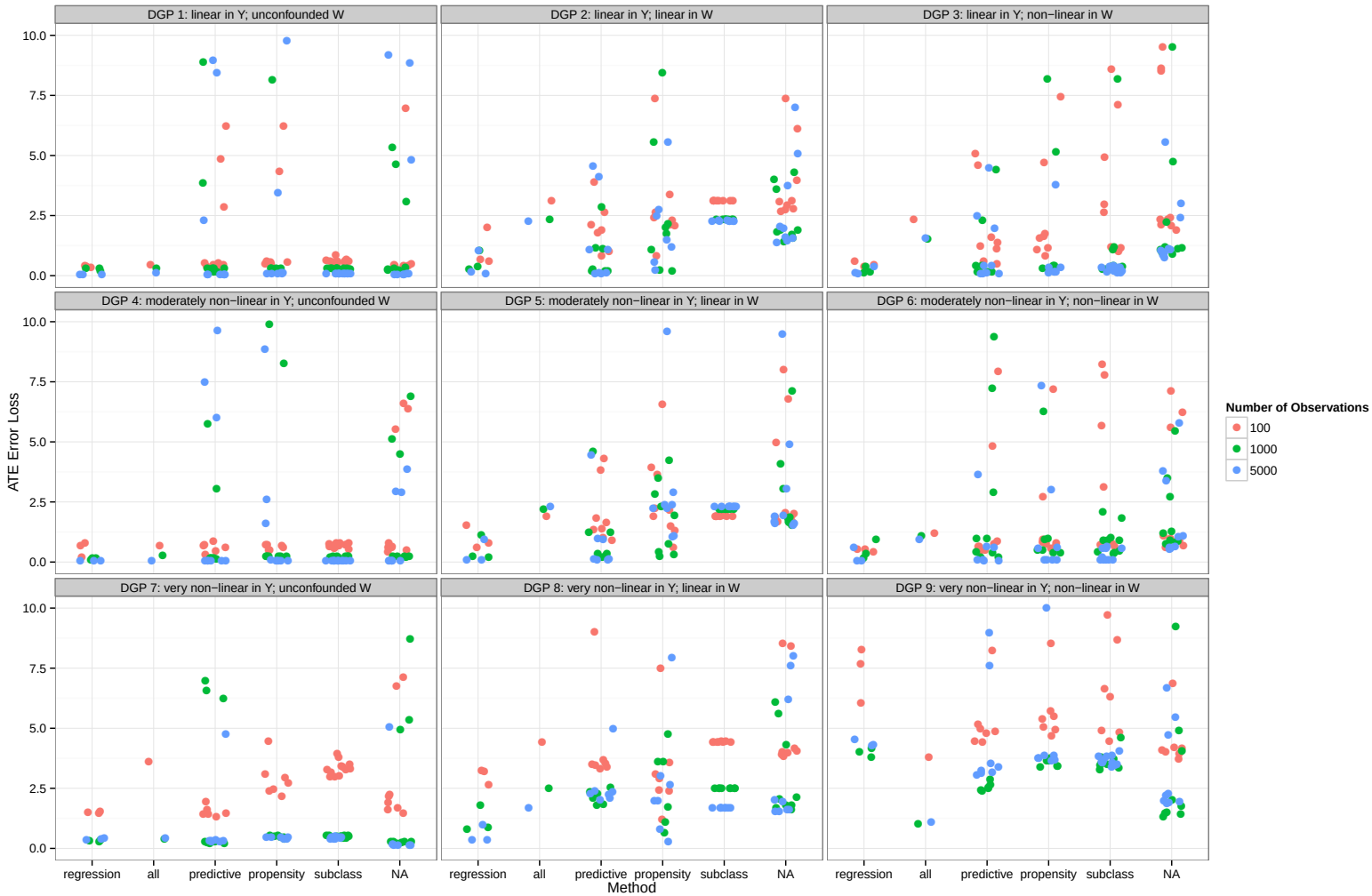
“bias” of ATE


DGP 1: linear in Y; unconfounded W
DGP 2: linear in Y; linear in W
DGP 3: linear in Y; non−linear in W
DGP 4: moderately non−linear in Y; unconfounded W
DGP 5: moderately non−linear in Y; linear in W
DGP 6: moderately non−linear in Y; non−linear in W
DGP 7: very non−linear in Y; unconfounded W
DGP 8: very non−linear in Y; linear in W
DGP 9: very non−linear in Y; non−linear in W
ATE Posterior Mean Bias
10
5
0
−5
−10
regression
all
predictive
propensity
subclass
NA
Method
Number of Observations
100
1000
5000

“root mse” of ICEs



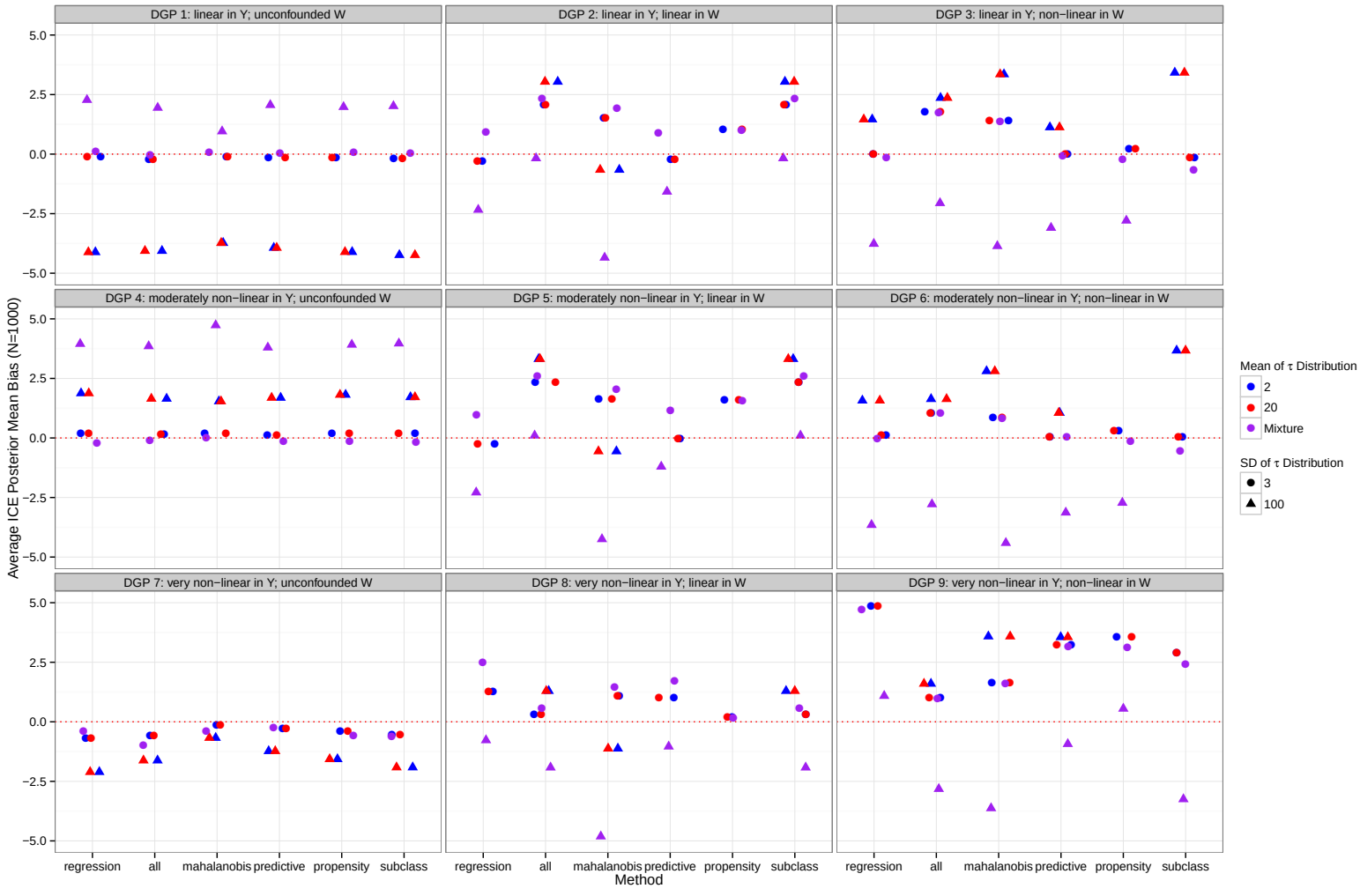
“root mse” of ATE



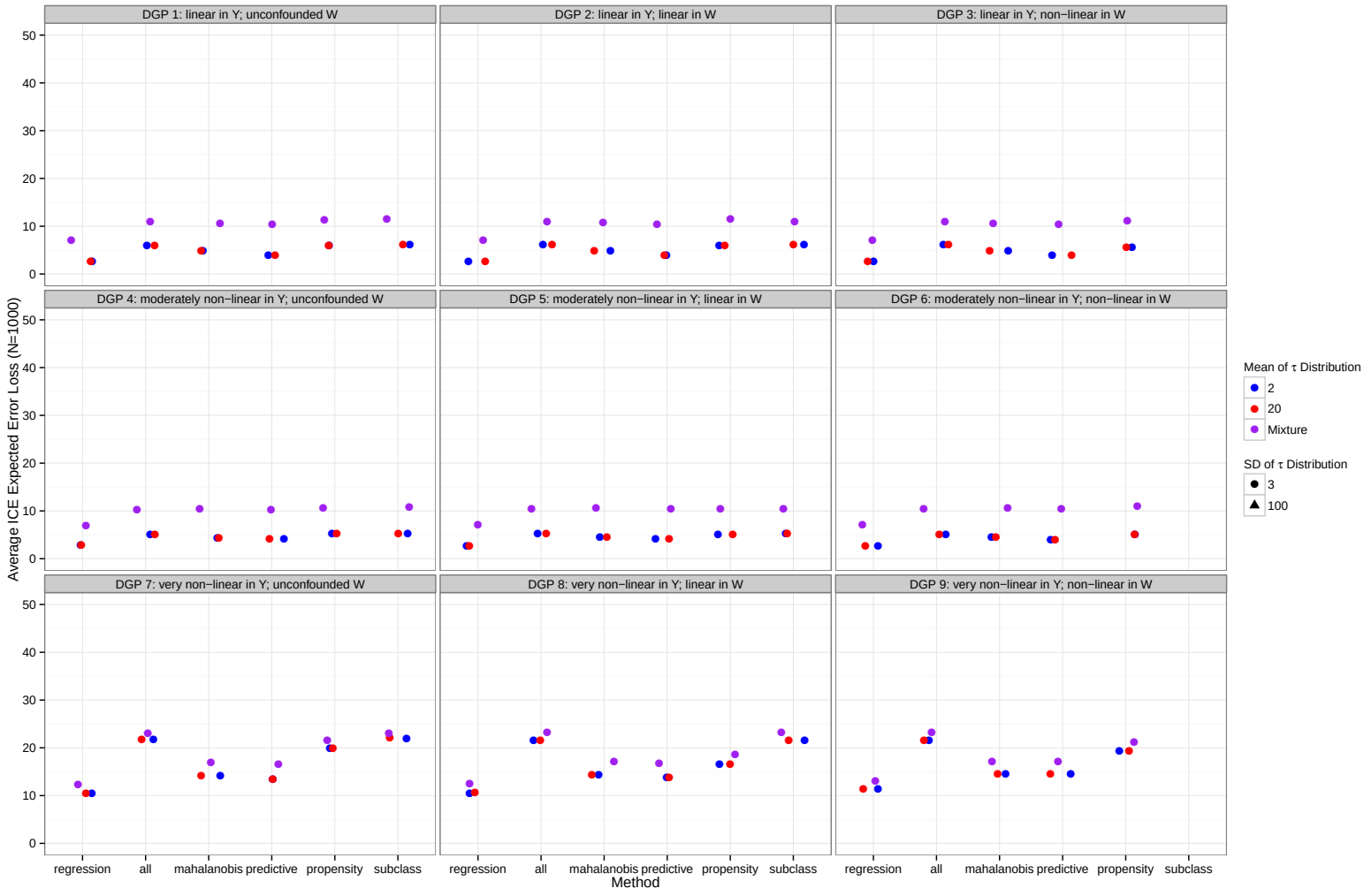
"power" of ICEs

“coverage” of ICEs

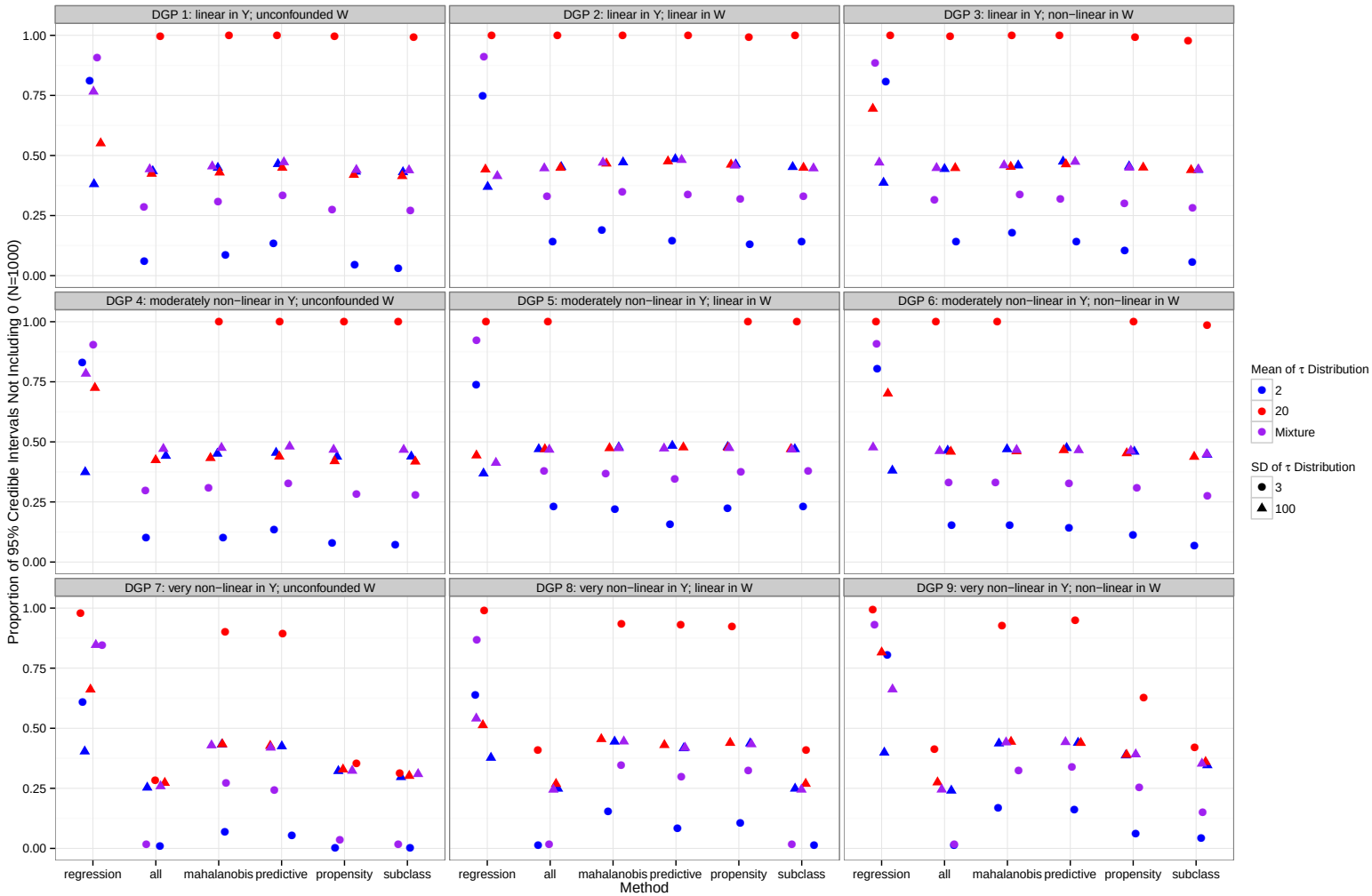
"bias" of ICEs: different τ_i distributions



"root mse" of ICEs: different τ_i distributions



"power" of ICEs: different τ_i distributions



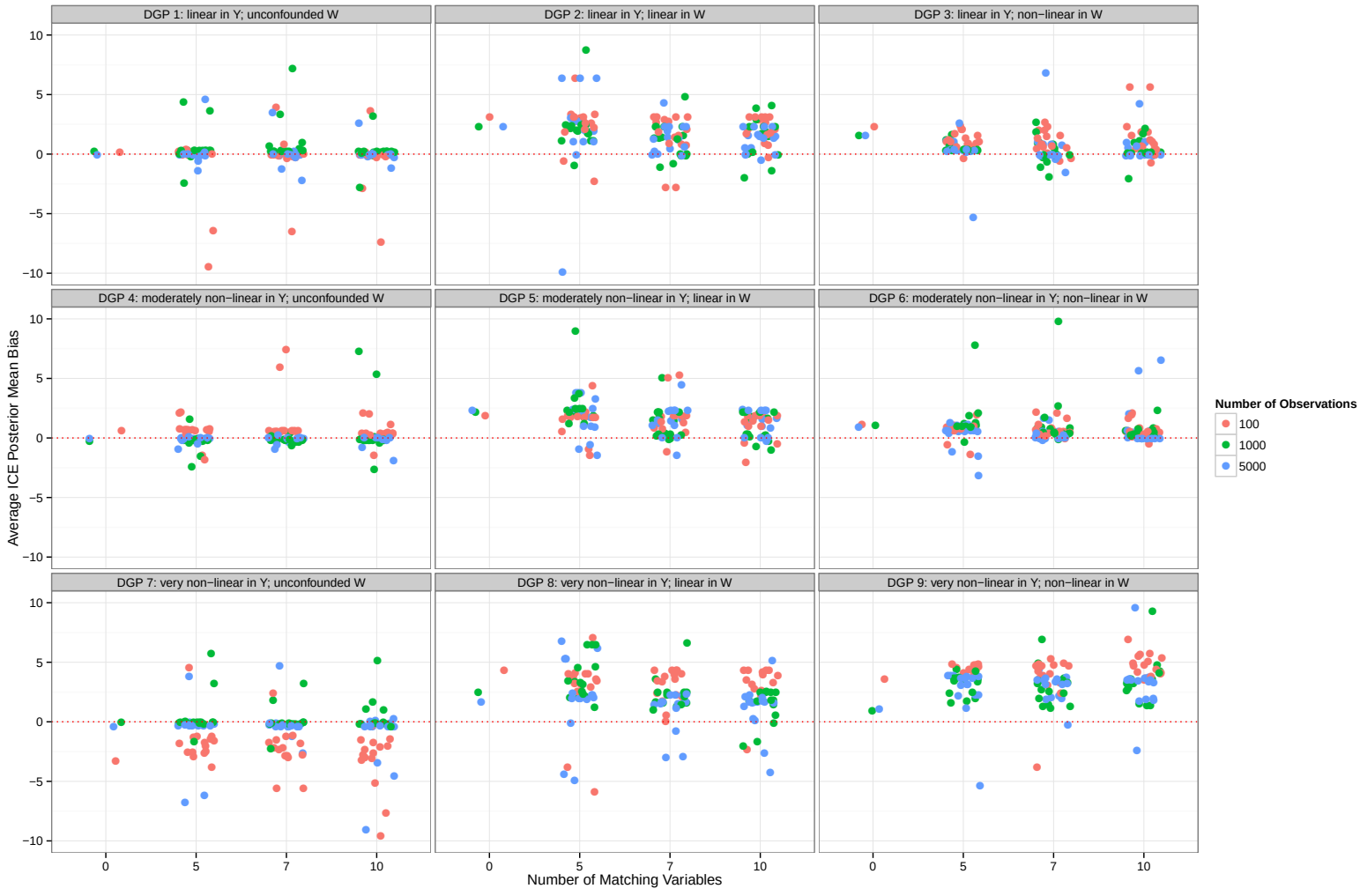
"coverage" of ICEs: different τ_i distributions

comparing methods summary

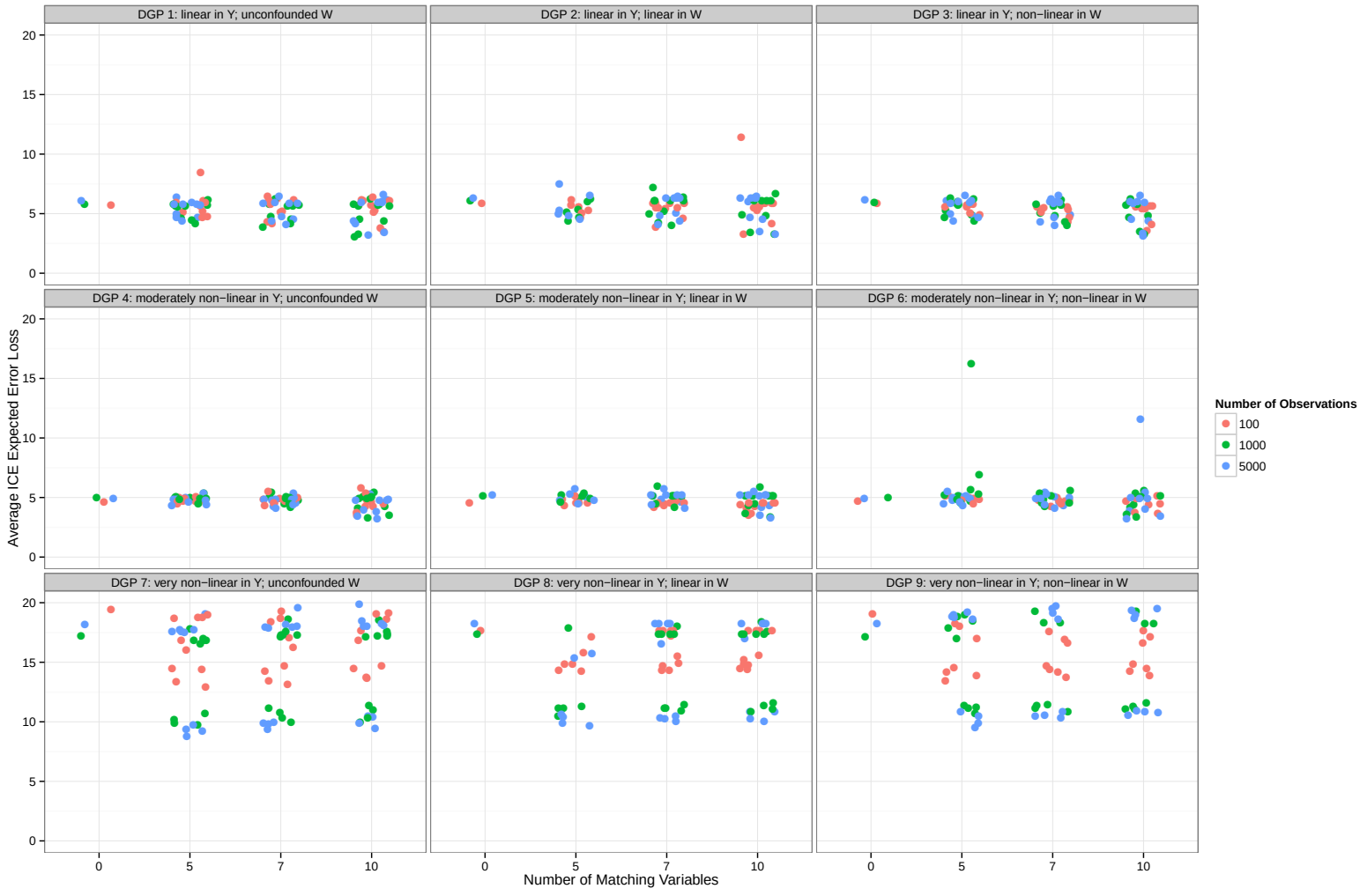
- of the matching methods, predictive mean matching usually performs as well or better than the others
- matching methods for ICEs have fairly low "power"
- regression has high power but very poor calibration "coverage"

conclusion: regression performs well if only interested in "averages"; for better performance at the individual level, use predictive mean matching

comparing X: "bias" of ICEs

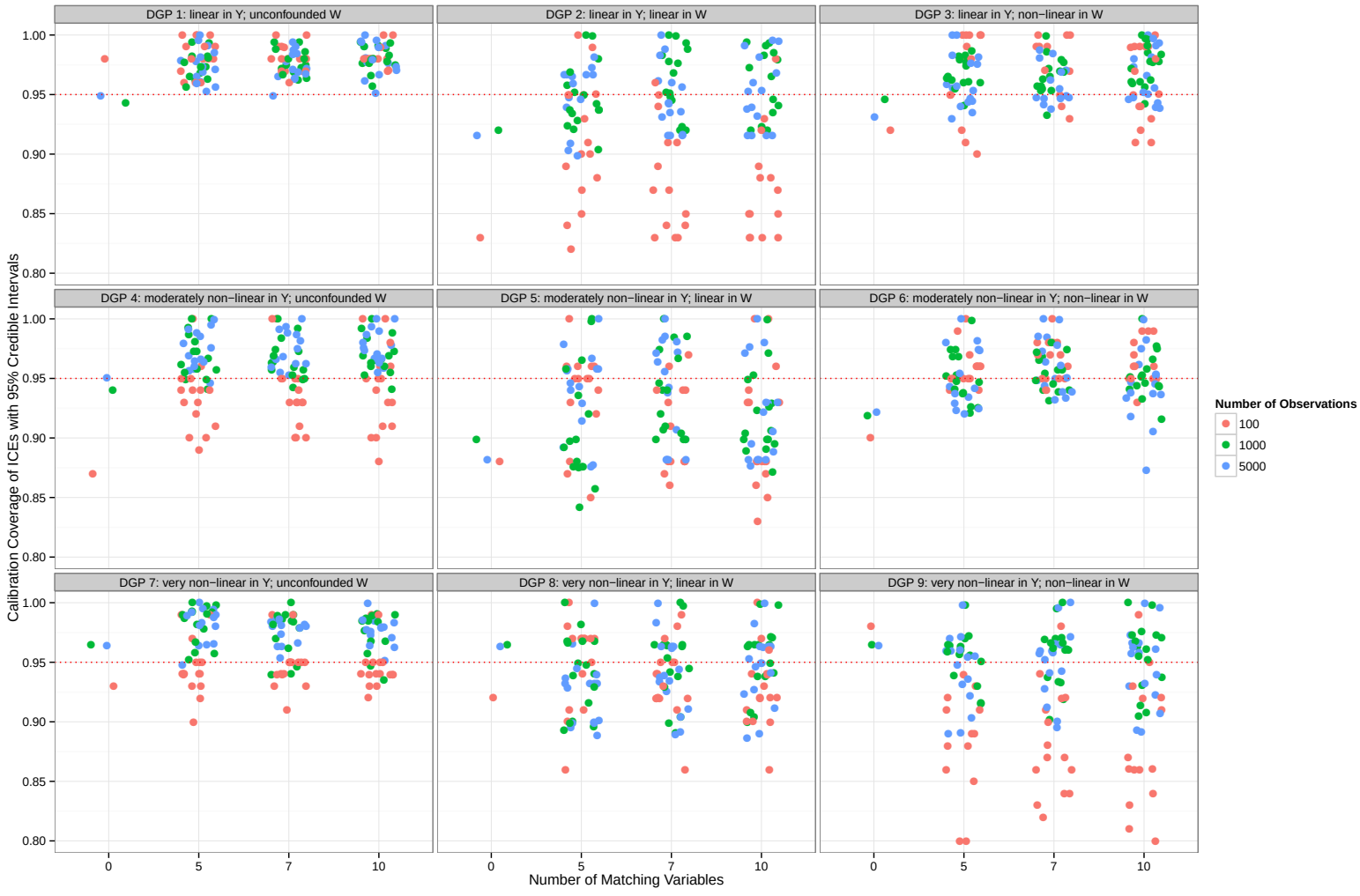


comparing X: "root mse" of ICEs



comparing X: "power" of ICEs

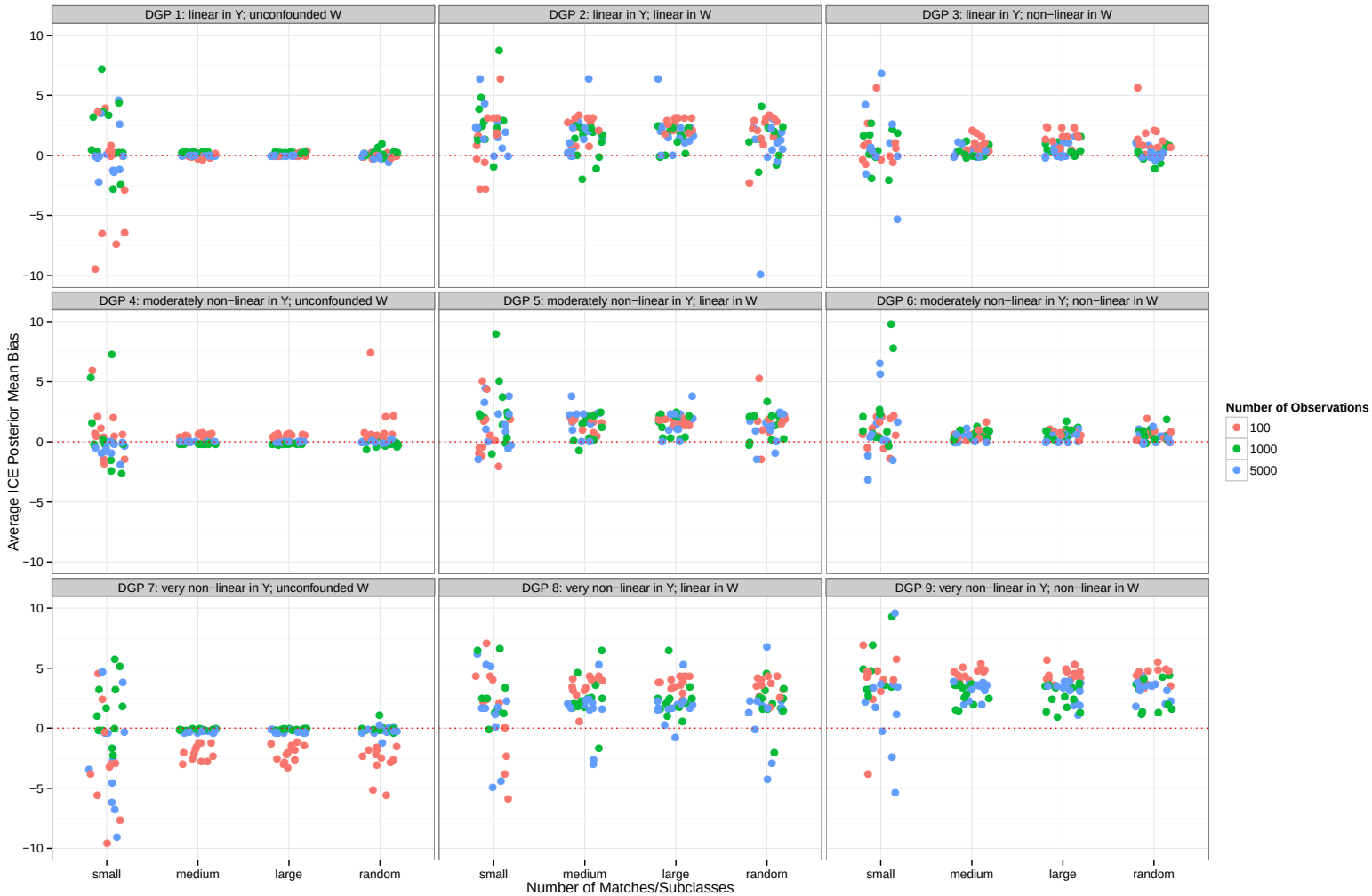
comparing X: "coverage" of ICEs



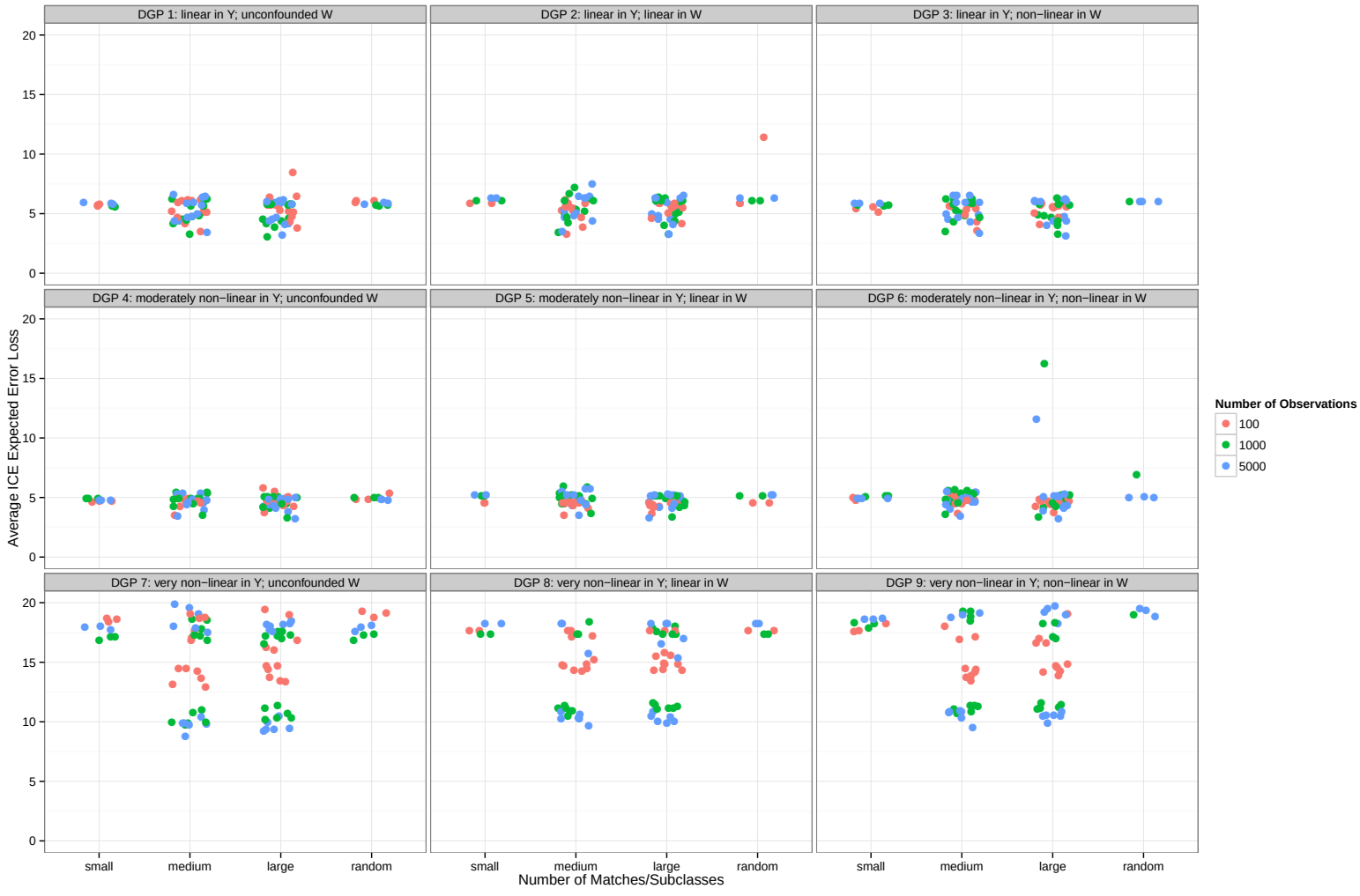
comparing X summary

- include all confounders
- no huge gain to including more

comparing M: "bias" of ICEs



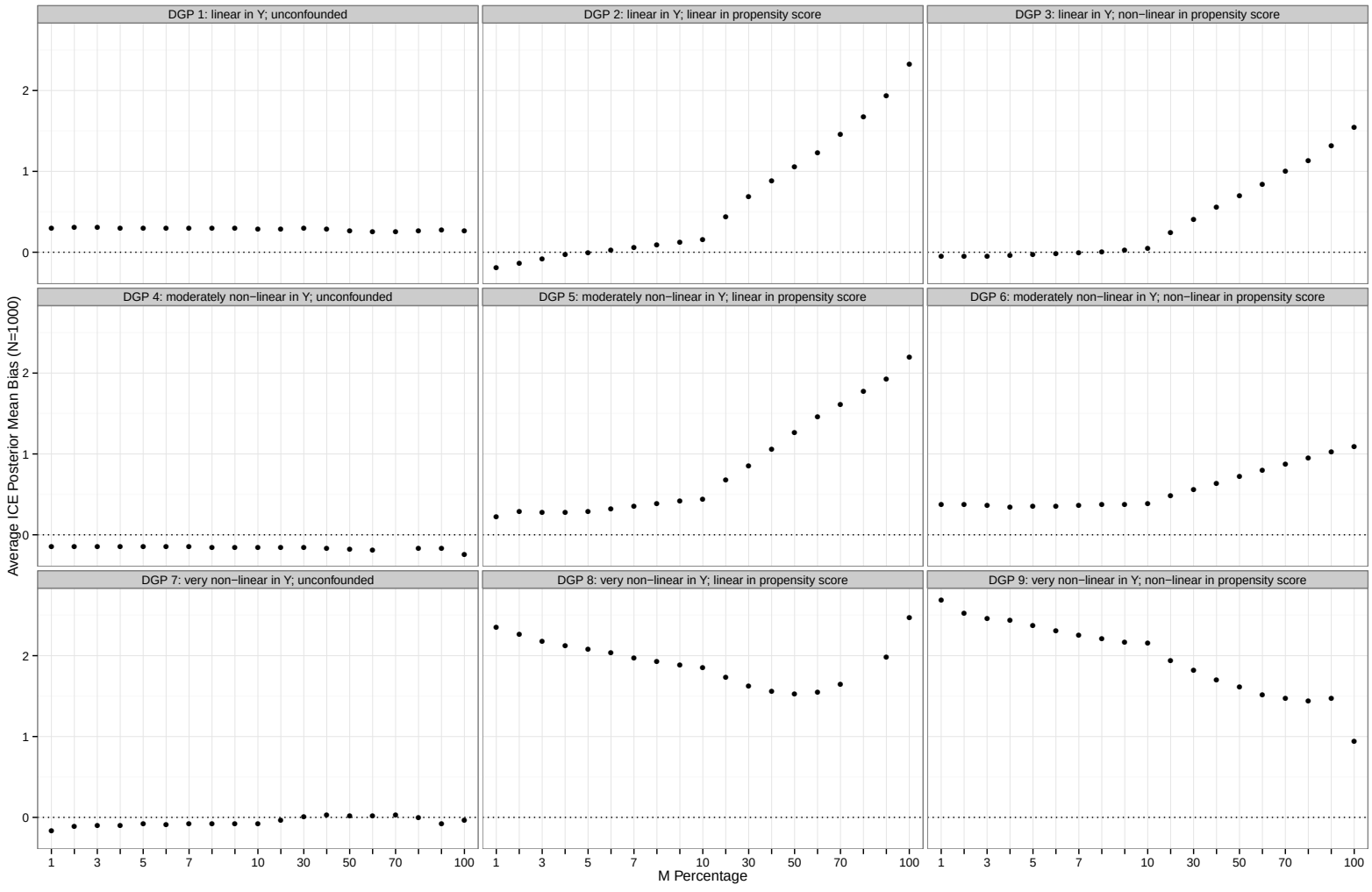
comparing M: "root mse" of ICEs



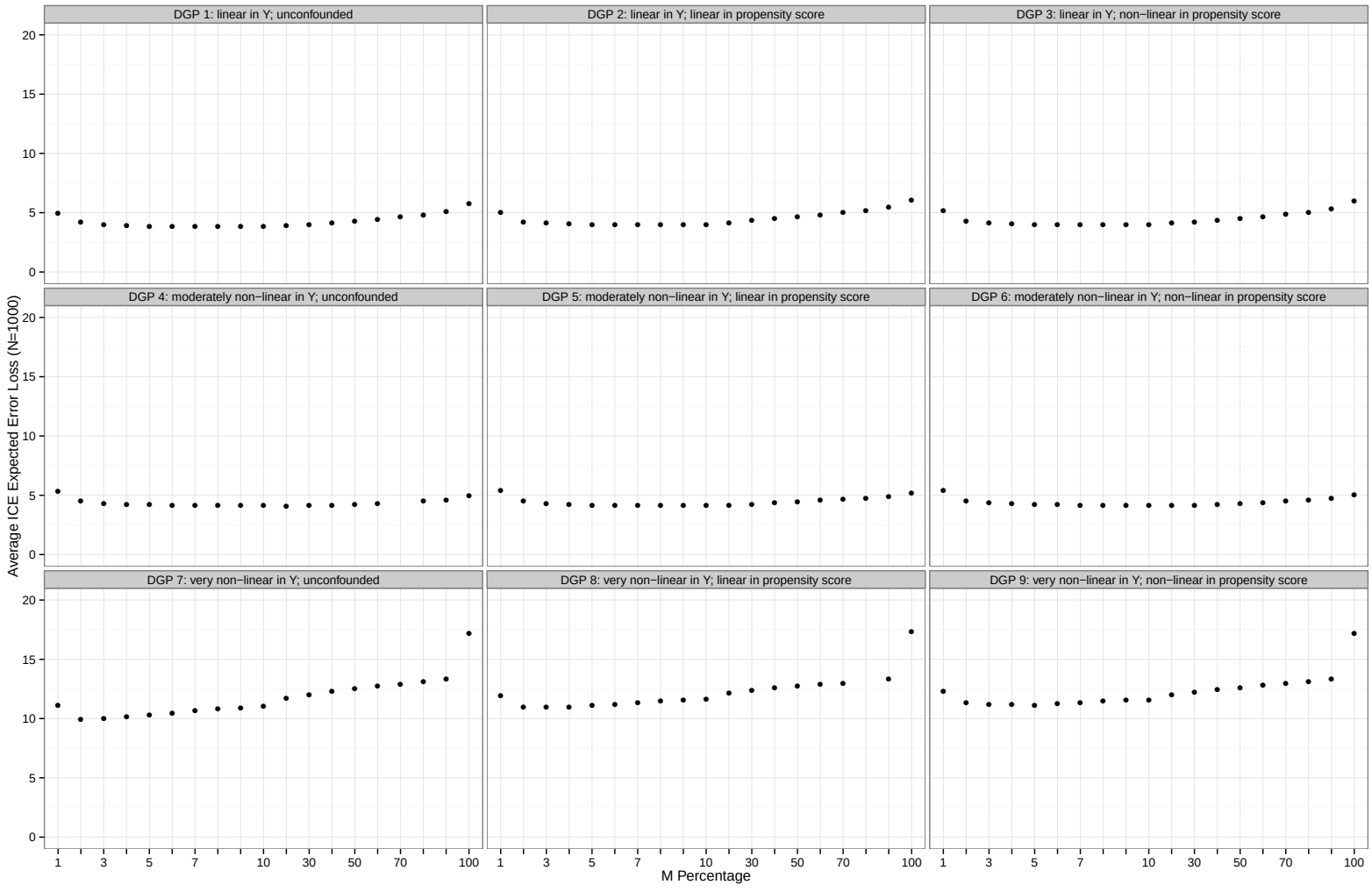
comparing M: "power" of ICEs

comparing M: "coverage" of ICEs

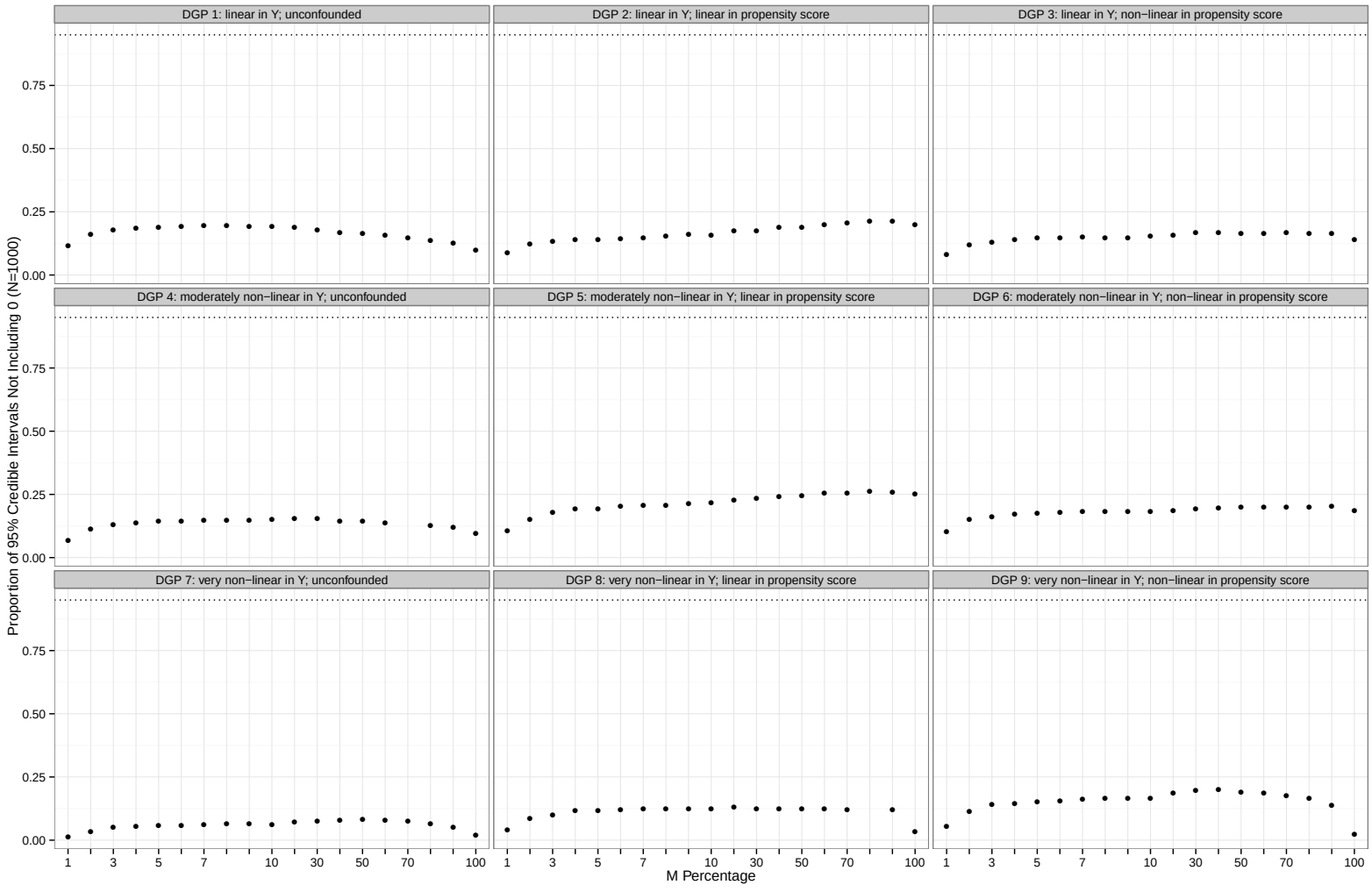
comparing M percentages: "bias" of ICEs



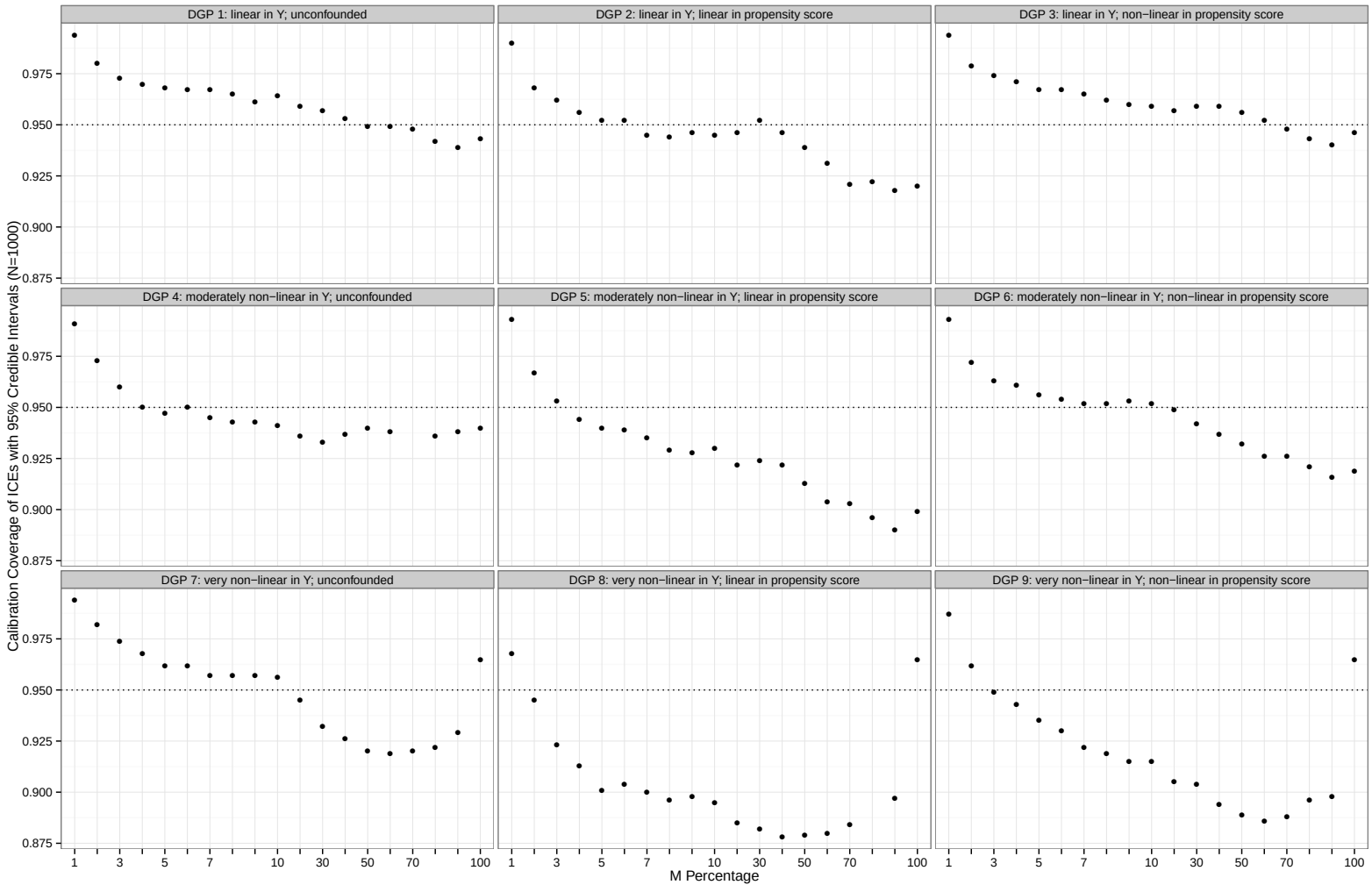
comparing M percentages: "root mse" of ICEs



comparing M percentages: "power" of ICEs



comparing M percentages: "coverage" of ICEs



comparing M summary

- larger donor pools better up to a certain point
- no clear optimal size (depends on data and application)
- random M introduces more uncertainty for little "bias" gain
- possibly use a smaller range of random M

simulation results lessons

- decent calibration coverage which improves with larger samples
- generally poor power
- performs well in recovering "unbiased" estimates of ICEs
- predictive mean matching generally performs as well or better than the other methods
- larger M is better up to a certain point (around 10% of smaller treatment group size), although there is no ideal M
- fairly robust to functional form misspecifications in the outcome or treatment assignment

choice: predictive mean matching with approximate M size of 10 percent of smaller treatment group

application: monitoring corruption

Olken (2007) field experiment in Indonesia

question: can top-down or grassroots bottom-up monitoring reduce corruption?

the setting:

- over 600 Indonesian villages received funds for road projects
- villages were randomly assigned monitoring mechanisms
- all villages hold three public project-accountability meetings
- corruption was measured by taking the difference between reported spending and an independent assessment of costs

Olken's main findings:

- top-down monitoring effective in reducing corruption
- grassroots participation in monitoring had little effect

three randomly assigned treatments

project **audit** from government agency (top down)

- baseline of 4% chance of audit; treated villages increased audit chance to 100%
- results of audit reported in village accountability meetings
- audit treatment cluster randomized at the subdistrict level

invitations to attend accountability meetings (bottom-up)

- invitations distributed through schools or neighborhood heads
- some villages randomly received additional treatment of **anonymous comment forms** in addition to the invitations
- comment forms summarized at accountability meetings
- classify both types into the "participation" treatment

measuring corruption

corruption can occur through overreporting of costs

$$\begin{aligned} Y &= \log(\text{reported cost}) - \log(\text{actual cost}) \\ &\approx \text{percent missing} \end{aligned}$$

$Y1$: major items (sand, rock, gravel, labor) in road project

$Y2$: major items in roads and ancillary projects

$Y3$: materials in road project

$Y4$: unskilled labor in road project

actual costs estimated by

- estimating quantity of materials used by digging up road
- estimating hours worked and prices through worker and supplier surveys

treatment assignment issues

- audit treatment cluster randomized at subdistrict level while participation treatments randomized at village level
- missing data for various reasons (listwise deleted)
- overlapping treatments: 606 total villages of which 264 received audit treatment, 185 received invites treatment, 189 received invites + comments treatment, 106 received no treatment

less than ideal randomization. . .

. . . but interesting scenarios for causal inference and ICEs

1. binary treatment on continuous outcome

$$\text{audit } (W) \rightarrow \text{corruption } (Y)$$

$$\text{participation } (W) \rightarrow \text{“outsider” meeting attendance } (A)$$

2. continuous treatment on continuous outcome

$$\text{attendance } (A) \rightarrow \text{corruption } (Y)$$

3. two stage design of “instrument” on outcome

$$\text{participation } (W) \rightarrow \text{attendance } (A) \rightarrow \text{corruption } (Y)$$

we can look at all three in an ICE framework!

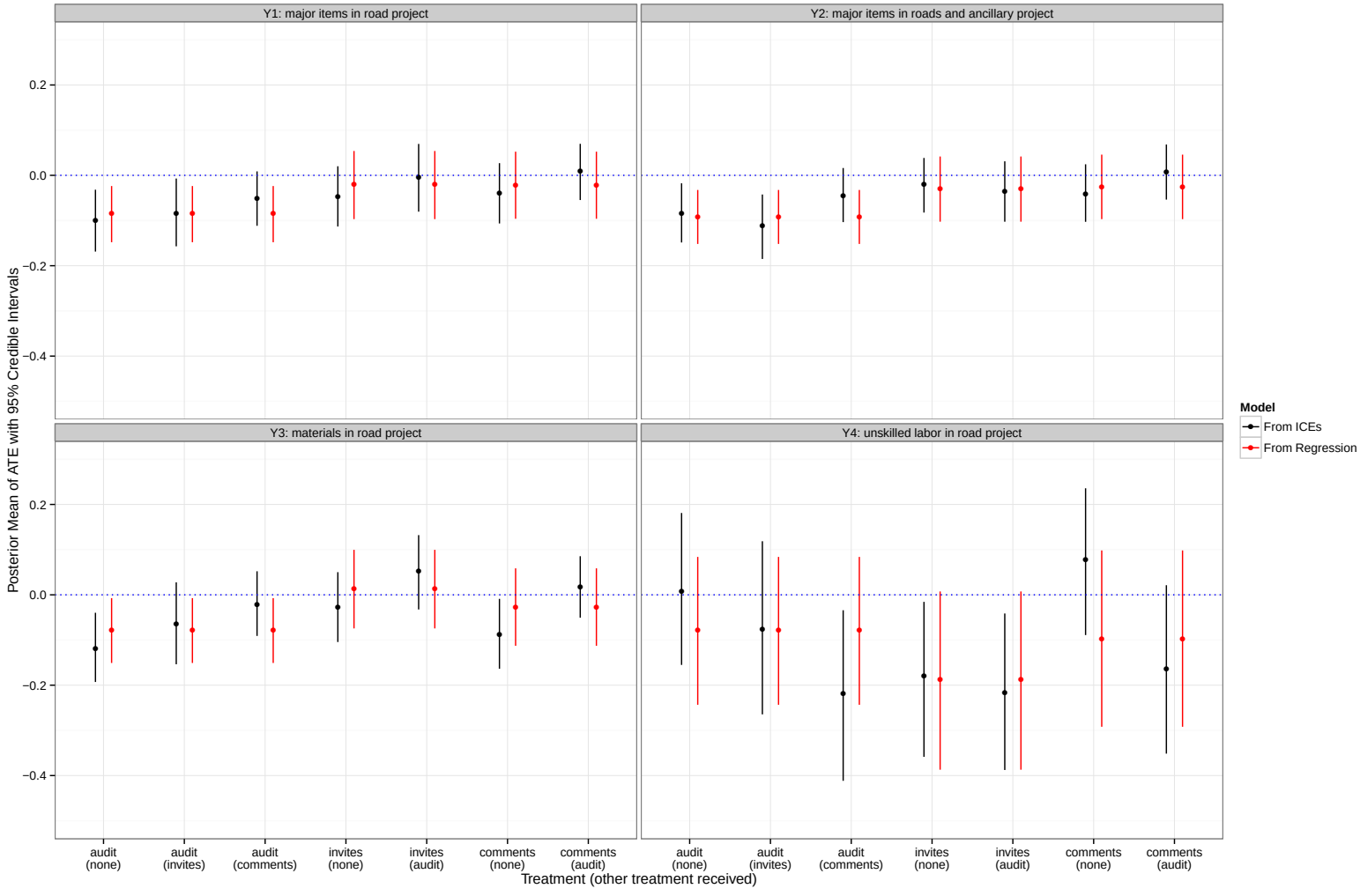
other variables

observations at the village level with covariates:

- distance to subdistrict
- education of village head
- age of village head
- salary of village head
- percent of households poor
- village population
- mosques per 1,000 population
- mountainous village dummy
- total budget
- number of subprojects

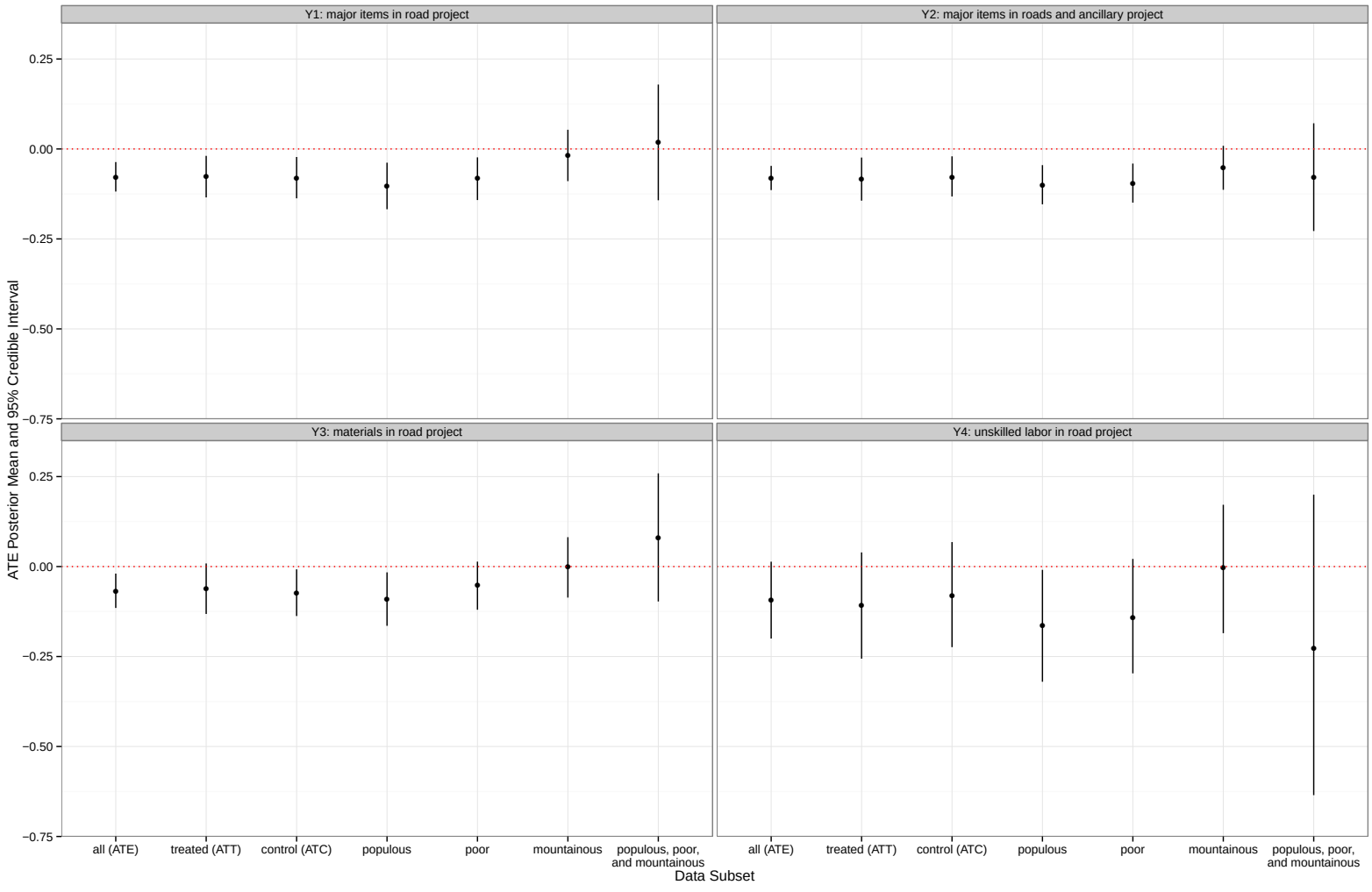
also measure average "outsider" meeting attendance and average "outsider" meeting attendance percent

ATE: W on Y

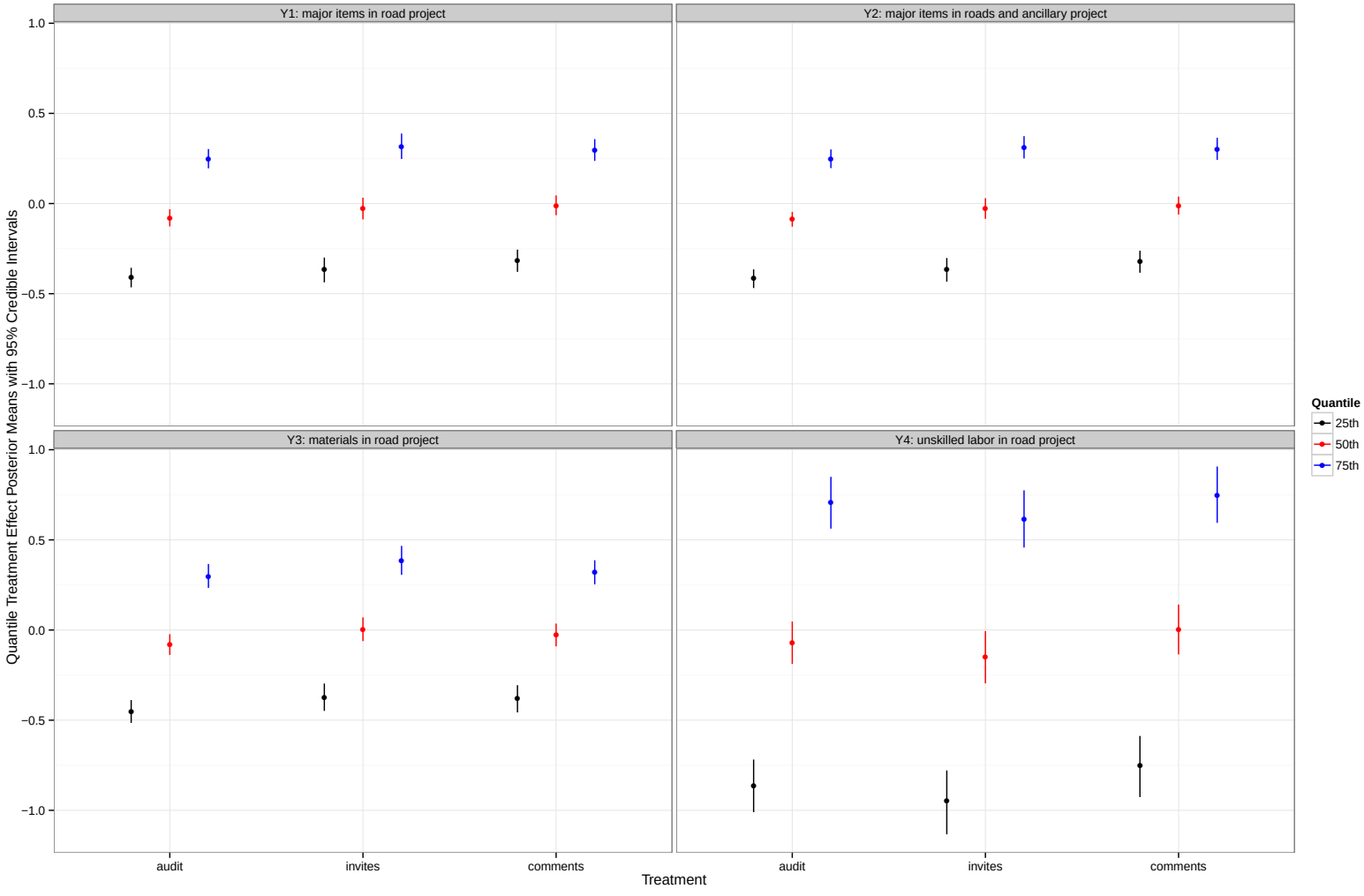


audit treatment works; participation treatments don't really

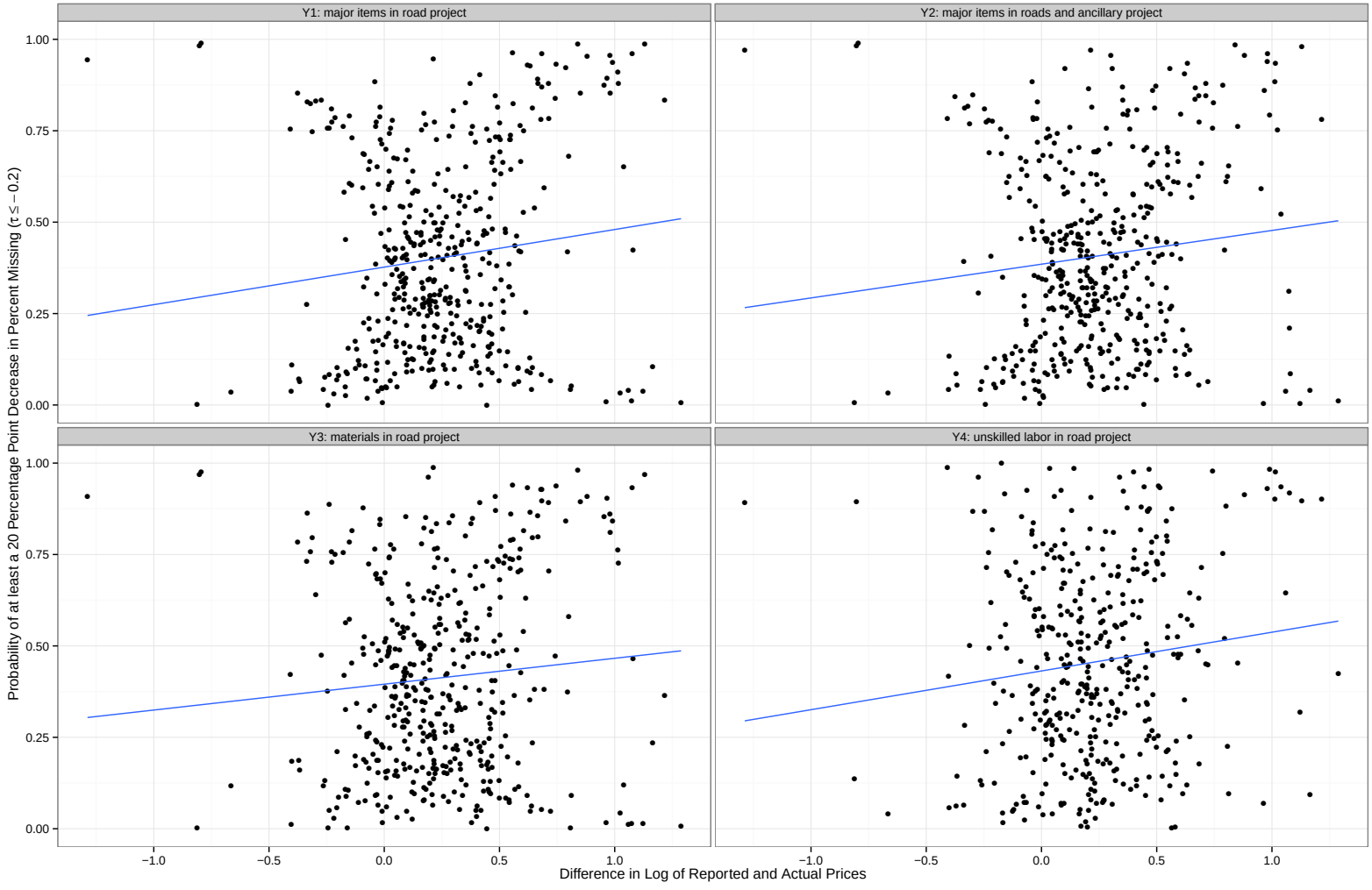
W on Y: different types of average treatment effects



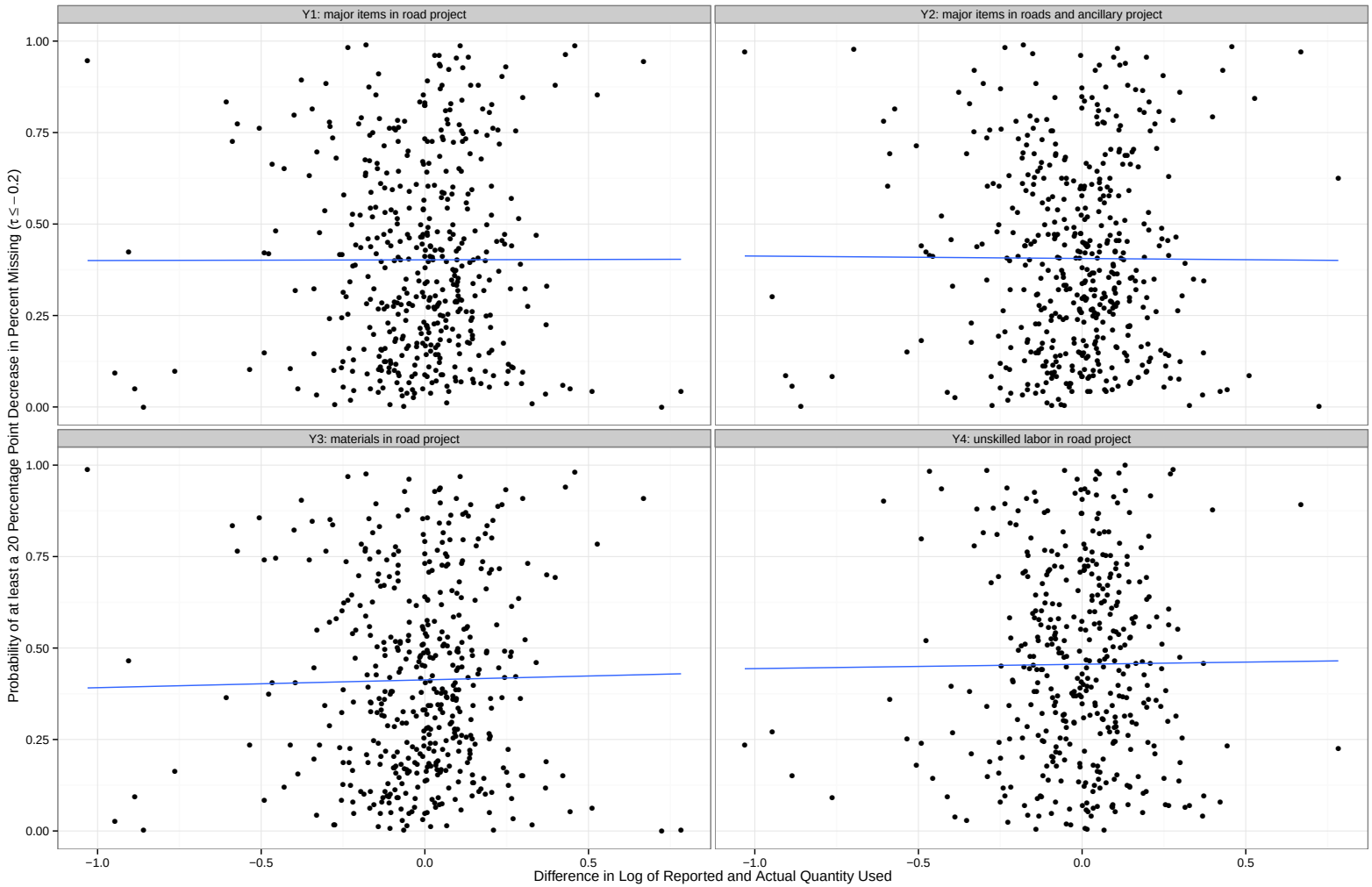
W on Y: quantiles of treatment effects



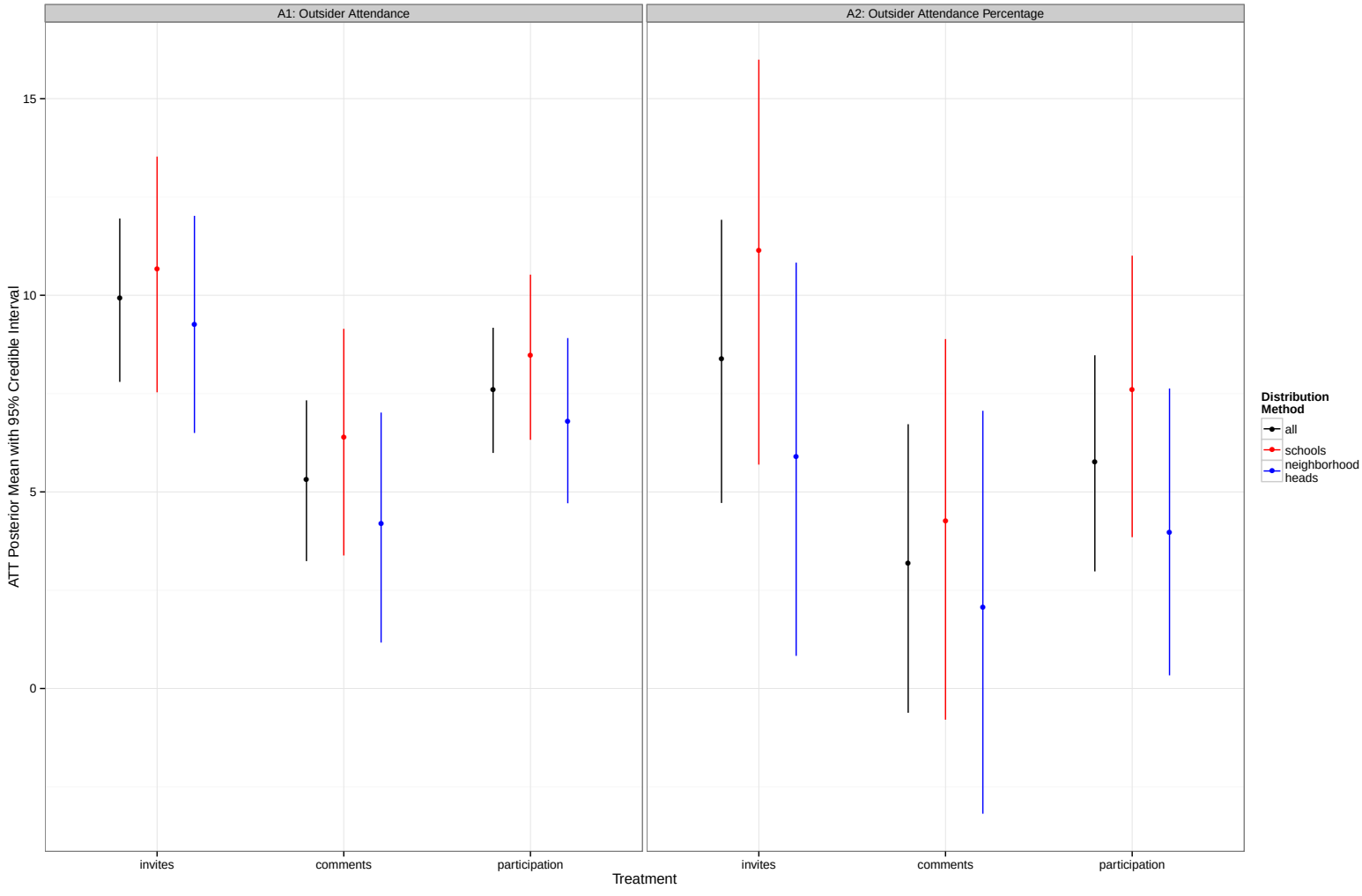
W on Y: audits have bigger effect on price corruption...



W on Y: . . . than on quantities used corruption



W on A: comments substitute for attendance



distribution through schools is slightly better

unique application: testing SUTVA

SUTVA usually violated through

1. interference across individuals OR
2. **multiple versions of treatment** (dosage issue)

Here: multiple versions of treatment (τ^a and τ^b)

- two participation treatments (invites and invites + comments)
- two distribution methods (schools and neighborhood heads)

SUTVA violated if

$$\begin{aligned} Y_i(1^a) &\neq Y_i(1^b) \\ \tau_i^a &\neq \tau_i^b \end{aligned}$$

for every i assuming $Y_i(0^a) = Y_i(0^b)$

testing SUTVA

SUTVA violation if any $\tau_i^a \neq \tau_i^b$ so

$$P(\text{SUTVA violated}) \approx P(\tau_i^a \neq \tau_i^b)$$

define various violation criteria to estimate $P(\text{SUTVA violated})$:

- one-sided violation:

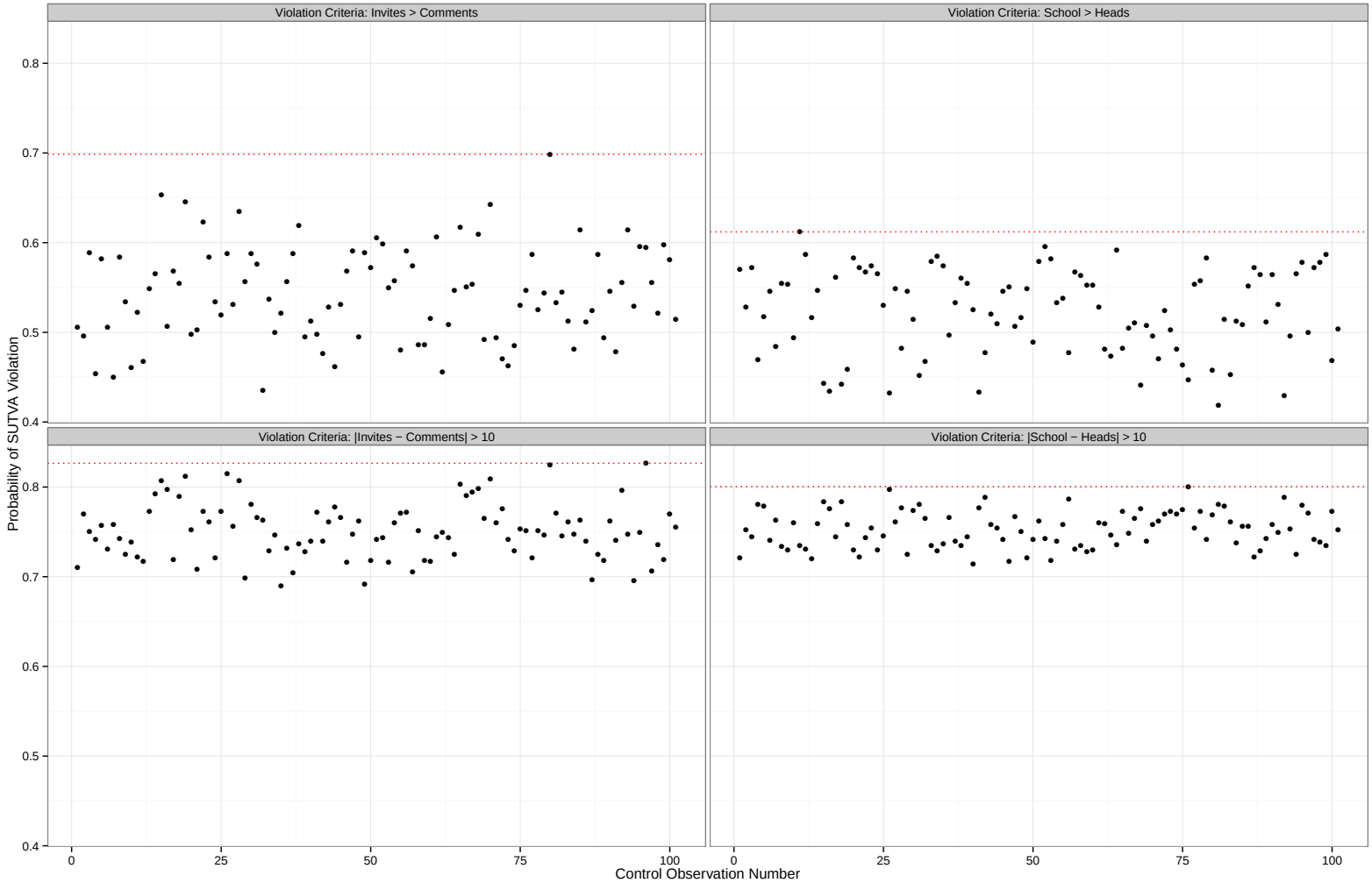
$$P(\text{SUTVA violated}) = \max(P(\tau_i^a > \tau_i^b))$$

- posterior range:

$$P(\text{SUTVA violated}) = \max(P(|\tau_i^a - \tau_i^b| > \epsilon))$$

- others?

W on *A*: testing SUTVA



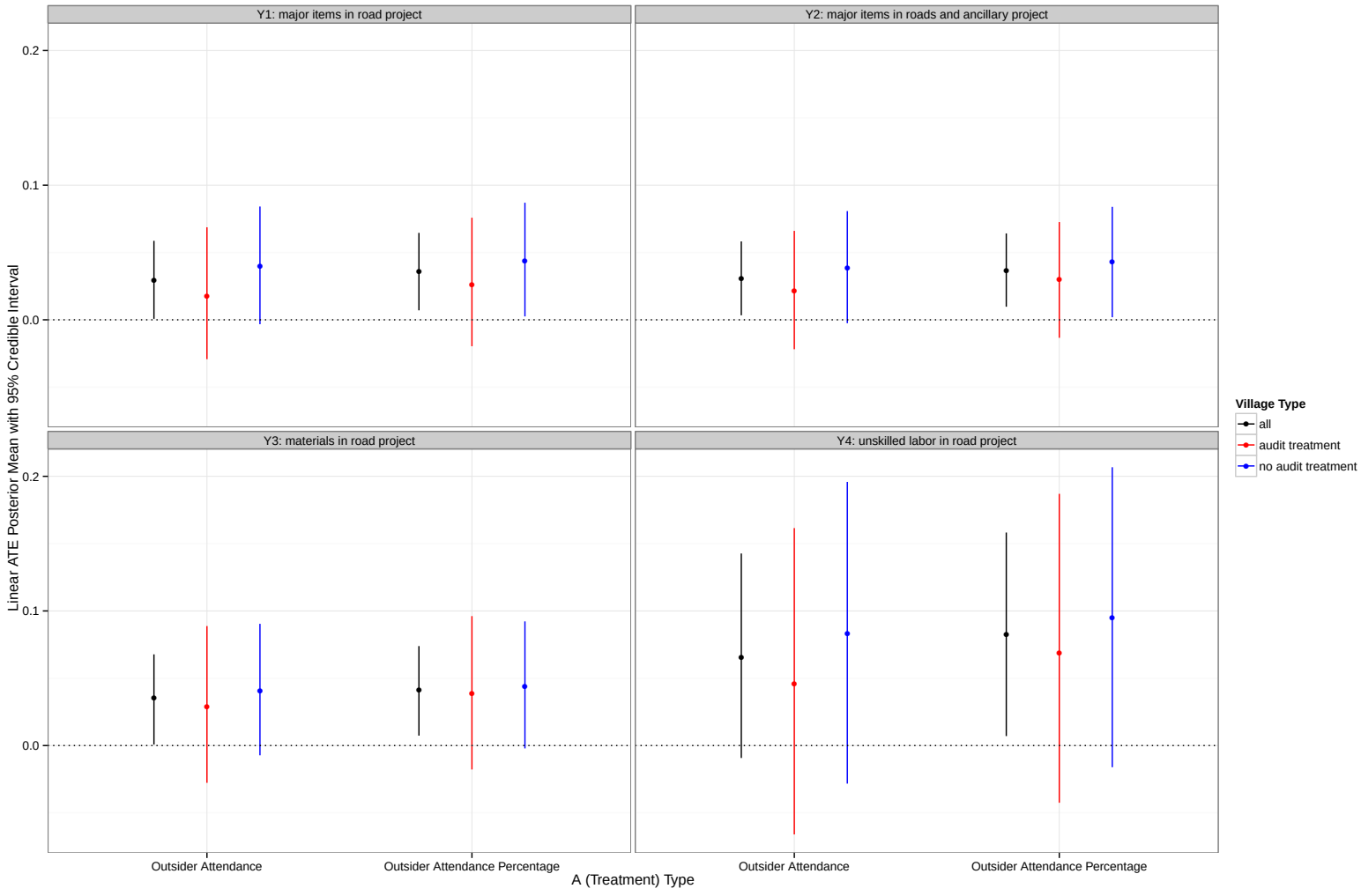
A on Y: continuous treatments

for continuous treatment variable A, assume linear effect:

1. calculate predictive means with one regression of Y on X
2. match with possible donor pool of all observations j where $A_i \neq A_j$
3. run linear regression of Y on A with donor pool and i to get $\tilde{\beta}_{D^{(i)}}$
4. draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\theta_i^{mis})$ where $\tilde{\theta}_i^{mis} = \tilde{\beta}_{D^{(i)}}(A_i - 1)$ (so assume i is always "treated" and calculate its outcome under "control" $(A_i - 1)$)
5. calculate $\tilde{\tau}_i^{mis}$ as $Y_i - \tilde{Y}_i^{mis}$

τ_i is a linear ICE

A on Y: no effect of attendance on corruption



2-stage W on Y

W is an "instrument" for (continuous) A

- monotonicity assumption: $A_i(1) \geq A_i(0)$
- exclusion restriction: if $A_i(1) = A_i(0)$, then $Y_i(1, A_i(1)) = Y_i(0, A_i(0))$

two sets of ICEs:

1. first stage ICE: $\delta_i = A_i(1) - A_i(0)$
2. second stage ICE:
   - if $\delta_i > 0$ (compliers), then $\tau_i^{comp} = Y_i(1, A_i(1)) - Y_i(0, A_i(0))$
   - if $\delta_i = 0$ (non-compliers), then $\tau_i^{ncomp} = Y_i(1) - Y_i(0)$

typical estimand: local (complier) average treatment effect

$$E[\tau_i^{comp} | \delta_i > 0] = \frac{\sum_{i:\delta_i>0} Y_i(1, A_i(1)) - Y_i(0, A_i(0))}{\sum_{i=1}^{N} \mathbb{I}(\delta_i > 0)}$$

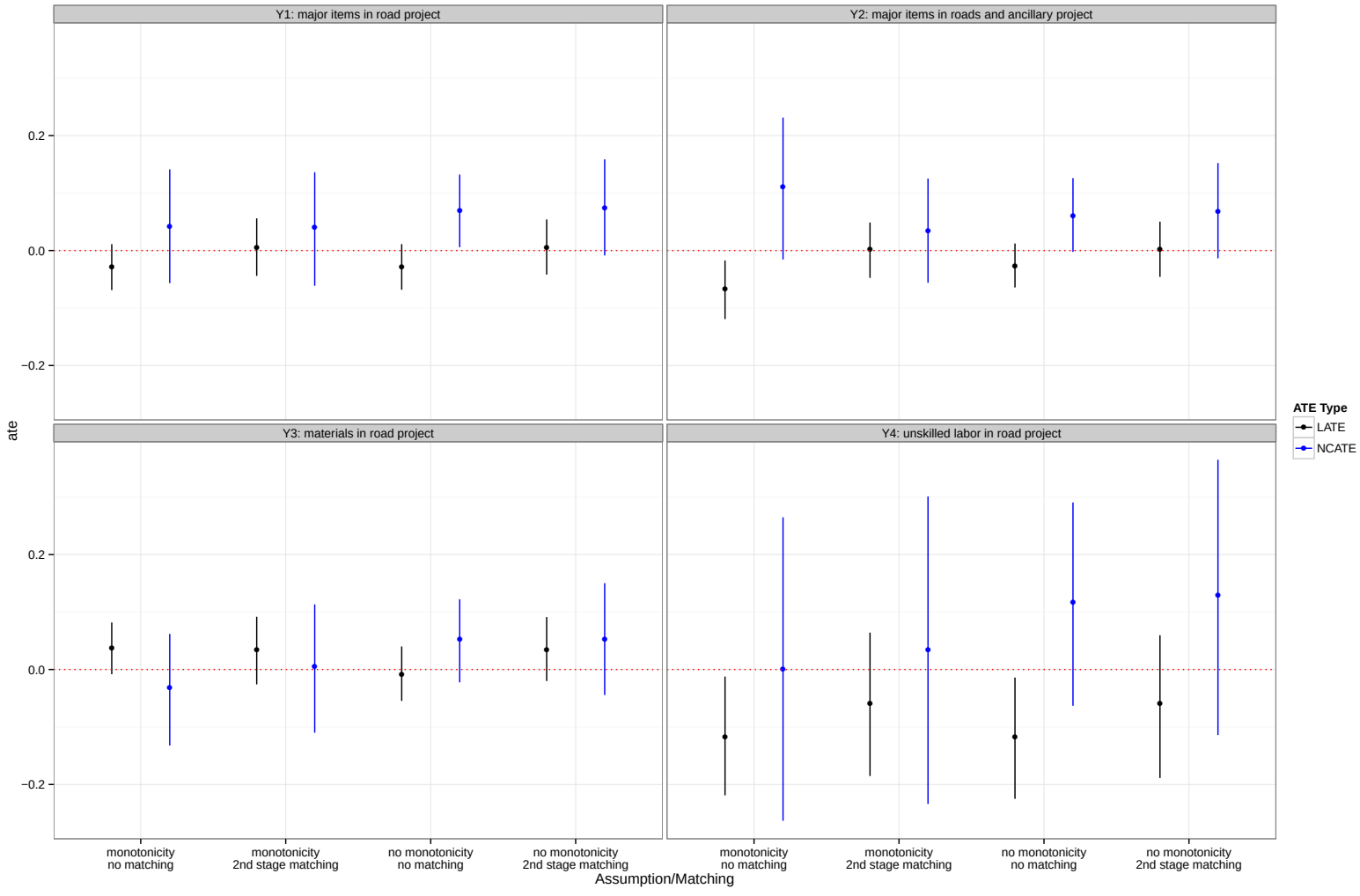
2-stage W on Y: estimation

need to impute Y^{mis} and A^{mis}:

1. draw $\tilde{A}_i^{mis}$ without matching: donor pool = all observations from opposite treatment (can match if desired)
2. calculate $\tilde{\delta}_i = W_i(A_i - \tilde{A}_i^{mis}) + (1 - W_i)(\tilde{A}_i^{mis} - A_i)$
3. determine compliance status: $\tilde{G}_i = 1$ if $\tilde{\delta}_i > 0$
4. draw $\tilde{Y}_i^{mis}$ (with or without covariate matching) as follows:
   - always match on $\tilde{G}_i$ for $D^{(i)}$
   - without monotonicity: same process as ICEs for continuous treatments with $\tilde{\theta}_i^{mis} = \tilde{\beta}_{D^{(i)}}\tilde{\delta}_i$
   - with monotonicity: $\tilde{\theta}_i^{mis} = \tilde{\beta}_{D^{(i)}}\tilde{\delta}_i$ for compliers and draw $\tilde{\theta}_i^{mis}$ from $p(\theta_i^{mis}|Y_{\{D^{(i)}=1\}}, D^{(i)})$ as normal for non-compliers
5. calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$

$$\tilde{\tau}_i = \tilde{\tau}_i^{comp} \text{ if } \tilde{G}_i = 1 \text{ and } \tilde{\tau}_i = \tilde{\tau}_i^{ncomp} \text{ if } \tilde{G}_i = 0$$

2-stage W on Y: similar results; exclusion restriction possibly okay



conclusion

- argument for estimating ICEs
- combining matching with bayesian model
- enormous flexibility in discover treatment heterogeneity and recover any causal quantity
- adaptable to different data structures
- extensions:
  1. relaxing monotonicity assumption in IV estimation
  2. testing causal inference assumptions