Institute for
Fiscal Studies

Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies

Arun Advani and Tymon Słoczyński

30 June 2014

Motivation

- Monte Carlo Simulations are a useful way of assessing finite sample performance of estimators.
- However, performance will vary with the DGP used.
    - low external validity.
- Hence might want to use DGP based on the data to be used (Empirical MC Study).
    - *e.g.* Busso *et al.* (2013) suggest 'conduct a small-scale simulation study designed to mimic their empirical context'.

Motivation

- Stronger suggestion by Huber *et al.* (2013) that 'the advantage [of an empirical Monte Carlo study] is that it is valid in at least one relevant environment'.
  - *i.e.* its internal validity is high by construction: should tell you about performance of estimators in your setting.
- Clearly if we can accurately replicate joint distribution of data then this is true.
- However, if we could do this then would be able to directly calculate effects of interest.

Research Questions

- Do the recently proposed EMCS procedures (Busso *et al.* 2013, Huber *et al.* 2013) provide a sufficiently good approximation to the true joint distribution of the data?
    - Answer: Not generally.
- Can we identify what are the key features of the DGP that EMCS must match?
    - Answer: We think so, and they are not testable in most data.

Outline

- EMCS
  - What is it?
  - Different designs
  - When might it work (in theory)
- Application
  - Test Design
  - Data
  - Some Results
- Conclusions

What is EMCS?

- Empirical Monte Carlo Studies (EMCS) are studies where:
  - We have an initial dataset of interest.
  - We want to somehow generate samples from the same DGP that created the initial data.
  - We can then test the performance of estimators of a particular statistic relative to the true effect in that sample.
  - We use the results on performance to inform us about which estimators are most useful in the original data.
- **Key issue:** how to generate these samples from the same DGP.

EMCS designs

- Suppose we have an original dataset with outcome Y, covariates $\boldsymbol{X}$, and treatment status T.
    - N observations: N_T treated, N_C control.
- Want to draw data from the DGP the created this, and estimate, *e.g.* the ATT.
- Two approaches suggested in the literature:
    - 'Structured design' (Abadie and Imbens, 2011; Busso *et al.* 2013).
    - 'Placebo design' (Huber *et al.* 2013).

Structured design

- Generate N observations, and assign treatment status s.t N_T are treated.
- Draw covariates $\boldsymbol{X}$ from a distribution which mimics the empirical distribution, conditional on T. ‣ Correlation
    - For a binary variable, match $\Pr(X^{(1)} = 1 | T = t)$ in generated sample to $\frac{\sum_i X_i^{(1)}.1(T=t)}{\sum_i 1(T=t)}$.
    - For a continuous variable, draw from normal/log-normal with appropriate mean and variance.
- Estimate conditional mean of $Y|\boldsymbol{X}$ on the original data.
    - Use this to construct fitted values for the new observations.
    - Generate new outcome as the fitted value plus an error with variance that matches that of the residuals.

Placebo design

- In original data, estimate a treatment status equation.
  - Run logit of T on relevant part of $\boldsymbol{X}$.
  - Store fitted value.
- Draw N observations, with replacement, from the control sample of the original data to create new samples.
- Assign 'placebo' treatment status to observations in this sample:
  - $T_i = 1(T_i^* > 0)$, where $T_i^* = \alpha + \lambda \boldsymbol{X}_i \boldsymbol{\beta} + \varepsilon_i$ and $\varepsilon_i \sim$ *iid logit*.
  - Choose α s.t. $\Pr(T = 1)$ in sample is same as in original data.
  - Choose $\lambda = 1$, as HLW. ▸ Calibrating Lambda
- By construction all treatment effects will be zero.

When might we expect EMCS to work?

- Suppose we ...
  - observed all the variables determining treatment and the outcome, and
  - knew the functional forms for their relationships with the covariates.
- Then clearly could generate data from the distribution...
  - ... but would also already know what the treatment effect is, so no need.

When might we expect EMCS to work?

- Treatment effect estimators we consider assume we observe all the relevant covariates, so we can assume this for our DGP as well.
    - Already a big assumption.

When might we expect EMCS to work?

- Treatment effect estimators we consider assume we observe all the relevant covariates, so we can assume this for our DGP as well.
- ‘Structured’ makes strong functional form assumptions.
  - Reasonable likelihood of misspecification.
  - Proposition in literature is implicitly that EMCS is more informative about the performance of estimators than a stylised DGP would be, *even if estimated structured DGP were misspecified*.

When might we expect EMCS to work?

- Treatment effect estimators we consider assume we observe all the relevant covariates, so we can assume this for our DGP as well.
- ‘Structured’ makes strong functional form assumptions.
- ‘Placebo’ avoids functional form assumptions for outcome.
    - Only uses subsample of data and has treatment effect of zero by construction.
    - Not clear when this might work.

Test Design

- Use a variety of different estimators (regression, matching, weighting...) on a dataset.
- Use EMCS to create replications of the data.
- Use same estimators on EMCS replications, and see if performance of estimators here matches performance in original data.
    - can choose any performance metric: *e.g.* bias, absolute bias, RMSE of ATT.
- **Key issue:** need to be able to calculate performance of estimators in the original dataset.

Data

Two solutions:

- Use the National Supported Work data (LaLonde 1986, ...), where experimental estimate is known.
  - but, only one specific sample – may for some reason be unrepresentative of underlying population.
  - RMSE not defined on a single sample, can only compute bias as performance measure.
- Use a “Monte Carlo of Monte Carlo” procedure.
  - Define a DGP for the population, and draw a population.
  - Draw samples from this population – we know how each sample compares to the population.
  - Perform EMCS on each sample.

NSW Results – Structured

- ‘Structured’ EMCS can replicate bias.
    - *i.e.* estimates from original data and EMCS samples are positively correlated.

NSW Results – Structured PSID

Table: **Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-PSID Data Set**

	"True biases"			
	Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)
Correlations				
Bias–Mean bias	0.371**	0.256	0.643***	0.549***
	(0.031)	(0.189)	(0.000)	(0.002)
Abs. bias–Abs. mean bias				
Rank–Rank				
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y
Number of estimators	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first columns).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

NSW Results – Structured

- ‘Structured’ EMCS can replicate bias.
    - *i.e.* estimates from original data and EMCS samples are positively correlated.
- Can’t generally replicate *absolute* bias.

NSW Results – Structured PSID

Table: **Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-PSID Data Set**

	"True biases"			
	Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)
Correlations				
Bias–Mean bias	0.371**	0.256	0.643***	0.549***
	(0.031)	(0.189)	(0.000)	(0.002)
Abs. bias–Abs. mean bias	−0.363**	−0.217	−0.435***	−0.216
	(0.035)	(0.267)	(0.009)	(0.260)
Rank–Rank	−0.357**	−0.169	−0.380**	−0.142
	(0.038)	(0.391)	(0.025)	(0.461)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y
Number of estimators	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first columns).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

NSW Results – Structured

- ‘Structured’ EMCS can replicate bias.
- Can’t generally replicate *absolute* bias.
  - True when in-sample bias is comparing to bias in original data.
  - Bias in original data for an estimator is difference between the estimate and the true effect.
- If instead we compare in-sample bias to a hypothetical bias, calculated as difference between estimate and *predicted value of the model* in the original data, performance is much better.

NSW Results – Structured PSID

Table: **Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-PSID Data Set**

	"Hypothetical biases"			
	Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)
Correlations				
Bias–Mean bias	0.371**	0.256	0.643***	0.549***
	(0.031)	(0.189)	(0.000)	(0.002)
Abs. bias–Abs. mean bias	0.408**	0.297	0.698***	0.616***
	(0.017)	(0.125)	(0.000)	(0.000)
Rank–Rank	0.408**	0.222	0.693***	0.599***
	(0.017)	(0.256)	(0.000)	(0.001)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y
Number of estimators	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first column).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

NSW Results – Structured

- ‘Structured’ EMCS can replicate bias.
- Can’t generally replicate *absolute* bias.
- If instead we compare in-sample bias to a hypothetical bias, calculated as difference between estimate and *predicted value of the model* in the original data, performance is much better.
  - In PSID data, true effect on unemployment is 11.06pp, but ‘predicted value of model (“structured”)’ estimates effect of 25.68pp.
  - This is because DGP was based on Oaxaca-Blinder LPM, which doesn’t perform well here.
- In CPS we know that OB LPM does perform well (estimated effect is 11.74pp), so ‘true’ absolute bias results should be good.

NSW Results – Structured CPS

Table: **Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-CPS Data Set**

	"True biases"			
	Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)
Correlations				
Bias–Mean bias	0.390**	0.259	0.530***	0.379**
	(0.023)	(0.184)	(0.001)	(0.042)
Abs. bias–Abs. mean bias	0.458***	0.420**	0.396**	0.333*
	(0.007)	(0.026)	(0.019)	(0.078)
Rank–Rank	0.484***	0.428**	0.426**	0.334*
	(0.004)	(0.023)	(0.011)	(0.077)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y
Number of estimators	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first column).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

NSW Results – Placebo

- In 'placebo' design we always know the true effect.
- But, it isn't clear that only using the control data to test for a placebo treatment effect is a relevant comparison to the original data.
  - Only using a subset of the data.
  - Treatment effect used is generally different to truth.
- In general we find it is unable to even replicate biases let alone absolute biases

NSW Results – Placebo

Table: **Correlations Between the Biases in the Uncalibrated and Calibrated Placebo Designs and in the Original NSW-CPS and NSW-PSID Data Sets**

	Uncalibrated		Calibrated	
	NSW-PSID	*NSW-CPS*	*NSW-PSID*	*NSW-CPS*
Correlations				
Bias–Mean bias	–0.337**	–0.353**	–0.403**	0.470***
	(0.048)	(0.041)	(0.018)	(0.004)
Abs. bias–Abs. mean bias	–0.022	0.045	0.273	–0.015
	(0.900)	(0.801)	(0.119)	(0.930)
Rank–Rank	0.061	–0.187	0.351**	–0.178
	(0.730)	(0.289)	(0.042)	(0.307)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Number of estimators	35	34	34	35

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: matching on the propensity score, $N = 40$ (second column) and bias-adjusted matching on covariates, $N = 40$ (third column).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Conclusions

- Recent proposals that some form of EMCS might overcome the design dependence issues common to MCS, at least for dataset of interest.
- We considered two forms of EMCS:
  - ‘Structured’.
  - ‘Placebo’.
- Structured design only informative if treatment effect in data is same as that implied by DGP.
  - Clearly untestable, and if we knew the true treatment effect then we would stop there.
- Placebo design appears to be even more problematic.
- Unfortunately only very negative results:
  - Don’t find any silver bullet for choosing estimator in particular circumstance.
  - For now best to continue using multiple approaches.

Additional Appendix Slides

Structured design (correlated)

- As before, but now we want to allow covariates to be correlated in a way that matches original data.
- In particular, want to draw each binary $X^{(n)}$, from the distribution suggested by the data conditional on T *and* $\{X^{(1)}, ..., X^{(n-1)}\}$, and draw continuous outcomes jointly conditional on the discrete covariates, so that we just need mean, variance and covariance.

▸ Back

Placebo design (calibrated)

- In 'uncalibrated' placebo design, $\lambda = 1$.
  - Huber *et al.* (2013) suggest this should guarantee 'selection [into treatment] that corresponds roughly to the one in our "population".'
- Only true if degree of covariate overlap between treated and controls in original data were same as overlap between placebo treated and placebo control in sample.
  - No reason we should expect this to be true.

Placebo design (calibrated)

- Can grid search $\lambda \in \{0.01, 0.02, ..., 0.99\}$ and find value of λ that minimises RMSD between simulated overlap and overlap in data.
  - ‘Overlap’ defined here as proportion of placebo treated individuals whose estimated propensity score is between the minimum and maximum pscore among placebo controls.

▸ Back