`datamirror`: coefficient-preserving synthetic data for restricted microdata

Jeffrey Clark

Department of Economics, Stockholm University

registream.org

Northern European Stata Conference, Oslo, 24 September 2026

Replication with restricted data

- Journals expect a replication package: data and code.
- Restricted microdata cannot be shared. It stays on an authorized server; only code and aggregates leave.
- Without access, a reader cannot re-run the paper.
- Inside, development is slow: submit, wait, read the log, resubmit.

A precedent

- Schlafmann, Setty and Vestman (*Review of Financial Studies*, 2024) ship a synthetic version of their Swedish register data in the replication package.
- Built ad hoc for that paper: a mean vector and a correlation matrix from inside the secure environment, drawn from a multivariate normal.
- So a synthetic stand-in for register data can clear a top journal.

What if there were a standard method that

1. reproduces the paper's published coefficients,
2. exports only aggregates an output checker can review,
3. and lets anyone verify the result?

The coefficient-recovery gap

- Synthetic-data and disclosure-control tools preserve distributions (`synthpop`, `sdcMicro`, deep generators). Regressions on their output hold *in expectation*; on the one draw you ship, $\hat{\beta}$ moves by sampling noise.
- Snoke et al. (2018) recommend recording coefficient fidelity separately from general utility, and supply no generator that delivers it.
- A scripted scan of all of SSC ($\approx$3,800 Stata packages) and the Stata Journal ($\approx$1,360 articles): one synthetic-data generator, built from preset profiles. None synthesizes from the researcher's own microdata; none preserves coefficients.

datamirror: the workflow

Inside the secure environment

```
datamirror init, checkpoint_dir("out")
regress y x1 x2, vce(cluster id)
datamirror checkpoint
...
datamirror extract
```

A bundle of aggregates. No records.

bundle
(8 CSVs)



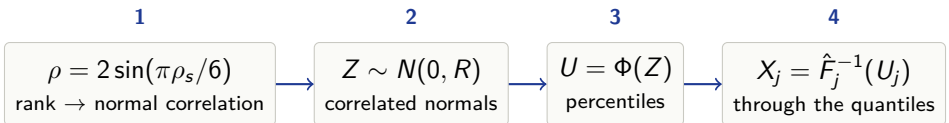
Anywhere

```
datamirror rebuild using "out"
datamirror check using "out"
```

A synthetic dataset on which every checkpointed regression returns its published coefficient.

Two uses: a replication package for data that cannot be deposited, and a working copy to develop on outside.

The rebuild: a Gaussian copula on the bundle



- Dependence is set on the normal scale; each variable gets its own distribution back from its quantiles.
- Categoricals: cut at their cumulative frequencies. With strata: the same, within each group.

The pin

Rebuild pins each checkpointed regression to its published coefficients β^*.

$$\delta y = X(\beta^* - \hat{\beta})$$

- Refit on $y + \delta y$: $\hat{\beta} = \beta^*$ exactly. The residuals do not change ($M_X \, \delta y = 0$).
- Exact: `regress`, `reghdfe`, `ivregress 2sls` (the instrumented coefficient).
- Within sampling noise: `logit`, `probit`, `poisson`, `nbreg`; y is redrawn from the model at β^*.

The bundle and its disclosure controls

- Eight small CSV files: quantile tables, category frequencies, a rank-correlation matrix, and the checkpointed coefficients with their standard errors.
- About $101V + C + V(V-1)/2 + K$ numbers in total, set by the number of variables and checkpoints. An output checker can read every one.

Three disclosure controls, at extract:

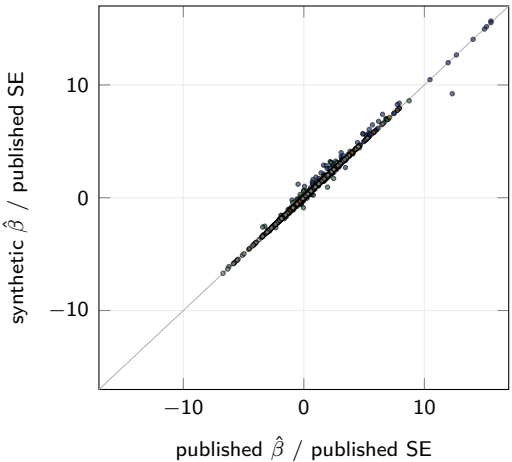
- `min_cell_size(50)`: rare categories, small strata and small-category coefficients are suppressed.
- `quantile_trim(1)`: minimum and maximum never leave.
- `min_resid_df(10)`: near-saturated regressions are not checkpointed.

Whether a bundle may leave is the data holder's decision.

Live demonstration

`checkpoint` $\rightarrow$ `extract` $\rightarrow$ `rebuild` $\rightarrow$ `check`

Evidence: four AEA replication packages



1,892 scored coefficients, all four packages.

Package	Pass	Evaluable
Duflo et al. (2012)	61	62
Dupas, Robinson (2013)	82	82
Banerjee et al. (2015)	162	162
Autor et al. (2019)	72	72
All	**377**	**378**

Pass: every scored coefficient within 3 published SEs of its published value. The linear ones land at 0.000.

Scored: every coefficient except constants and fixed effects; in IV, the instrumented variable.

Limitations

- **Standard errors are not targeted.** Synthetic over published standard error, scored coefficients, four packages:

	OLS iid	OLS robust	OLS clustered	IV clustered
median ratio	0.94	0.58	0.53	1.79

- **Only checkpointed regressions are pinned.** Other relationships are as good as the copula.
- **Samples can differ.** An `if` on a variable the copula does not reproduce jointly can select a different N. `check` reports N and clusters for every regression.
- **It certifies the computation.** An error in the original analysis passes through.

Conclusion

None of the steps is new on its own. What is new is the standard:

- built from aggregates that can leave the secure environment,
- a whole table of regressions, consistent on one dataset,
- disclosure controls, and a check anyone can run,
- one format an output checker reviews once.

Restricted data has been the standing exception to reproducible publication. For the regression coefficients a paper reports, it no longer has to be.

```
net install registream, from("https://registream.org/install/stata/registream/latest")
net install datamirror, from("https://registream.org/install/stata/datamirror/latest")
```

registream.org · Stata Journal, under review · jeffrey.clark@su.se

Demo 1: inside the secure environment

```
. regress annual_income age female i.educ_level, vce(cluster municipality)
             |      Coef.   Std. Err.      t    P>|t|
         age |   870.5335   11.64887    74.73   0.000
      female |  -7063.925   337.7418   -20.92   0.000
  ...
. datamirror checkpoint, tag("income")
Checkpoint 1: income
  N = 50000   Clusters = 50 (municipality)
  Disclosure: omitting 2 small-category coefficient(s) from the recipe:
    6.educ_level 7.educ_level
✓ Checkpoint 1 saved
. reghdfe wellbeing employed married age, absorb(year region)
. datamirror checkpoint, tag("wellbeing")
✓ Checkpoint 2 saved
. ivregress 2sls earnings_index (trained = voucher) age educ_level
. datamirror checkpoint, tag("earnings_iv")
✓ Checkpoint 3 saved
```

Demo 2: what leaves

```
. datamirror extract
✓ schema.csv
✓ marginals_cont.csv
✓ marginals_cat.csv
  Privacy: Suppressed 23/230 categories (< 50 obs)
✓ correlations.csv (full matrix: continuous + categorical)
✓ checkpoints.csv
✓ checkpoints_coef.csv
✓ metadata.csv

. type "demo_bundle/checkpoints_coef.csv"
cp_num,varname,coef,se
1,age,870.5334871760692,11.64887247338222
1,female,-7063.924866586461,337.7418463074135
2,employed,8.827783623492376,.0892252082425922
3,trained,4.925282158171824,.1853858099711584
...
```

Demo 3: rebuild outside, and check

```
. clear
. datamirror rebuild using "demo_bundle", clear seed(2409)
  N = 50000   Variables = 18   Checkpoints = 3
REBUILD COMPLETE: 50000 observations, 18 variables

. datamirror check using "demo_bundle"
Checkpoint 1: income
  N:        original    50,000   synthetic    49,975
  Clusters: original        50   synthetic        50 (municipality)
  Variable       Original     Synthetic   |Δβ|     Δβ/SE
  age            870.5335     870.5335    0.0000  0.0000
  female         -7.1e+03     -7.1e+03    0.0000  0.0000
  Headline Max Δβ/SE = 0.0000
Checkpoint 2: wellbeing        Headline Max Δβ/SE = 0.0000
Checkpoint 3: earnings_iv
  trained          4.9253       4.9253    0.0000  0.0000
  Headline Max Δβ/SE = 0.0000
```