xtvfreg

A Stata Command for Modeling Mean and Variance in Panel Data

Version 0.4.8

Tim F. Liao
University of Illinois Urbana-Champaign

Stata Sociology Virtual Symposium
20 October 2026



Outline (≈42 Minutes)

5 min	**Motivation & the Research Behind xtvfreg** Why model variance; the Mooi-Reci & Liao (2025) study
8 min	**The Method** Mean-variance framework, the iterative algorithm, estimation details
3 min	**What's New in v0.4.8** Syntax, options, and recent changes to the command
18 min	**Hands-On Walkthrough** Data prep, running the command, reading output, five worked examples
4 min	**Comparisons, Tips & Reporting** vs. standard FE/RE; troubleshooting; how to report results
4 min	**Resources & Additional ado** Reference, help files, i_ineq,

Why Model Variance?

The Problem

Traditional panel models assume constant variance (homoscedasticity).

But inequality is often heterogeneous:
- Wage dispersion varies by education
- Income volatility differs by labor-market history over time
- Risk heterogeneity across groups

Standard mean models miss these variance patterns entirely.

The xtvfreg Solution

Models BOTH mean AND variance, jointly, within a panel-data (fixed-effects) framework:

Mean equation:
- Who has higher/lower outcomes?

Variance equation:
- Who has more/less dispersion?

Captures the complete inequality picture.

Methodological Foundation

In Sociology

We have a long tradition in inequality research on:

- between-group inequality

&

- within-group inequality

In Methodology

For panel data:

- The xt series of regression in Stata & Paul Allison's hybrid/between-within model

For cross-sectional data:

- the variance function regression (Western & Bloome, Sociological Methodology, 2009)

xtvfreg is the marriage of the two.

The Research Behind xtvfreg

Mooi-Reci, I., and T. F. Liao. 2025. “Unemployment: A Hidden Source of Wage Inequality?” *European Sociological Review* 41(3): 382–394. https://doi.org/10.1093/esr/jcae029

Research Question

Does unemployment leave a lasting mark not just on the level of wages, but on how unequal those wages become — both between people and within a person's own career over time?

Data

Dutch OSA Labour Supply Panel
1985–2008

Longitudinal wage histories with unemployment spells

Contribution

Introduces the variance function panel regression — the method that xtvfreg implements in Stata

Key Findings from the Article

Lower AND More Dispersed

Wages of previously unemployed workers are both lower on average and more dispersed than wages of continuously employed workers.

A Gender Difference

These wage disparities — in level and in spread — are more pronounced among previously unemployed men than among women.

A Hidden Source of Inequality

Unemployment operates as a hidden driver of wage inequality: its effect on dispersion is easy to miss with mean-only models.

Method Validated

The variance function panel regression proved useful for identifying sources of wage inequality within persons and over time.

From Findings to Method

A mean-only model would have shown only half the story:

Mean-Only Model Would Show

✓ Previously unemployed earn less on average

✗ Nothing about how much MORE variable those wages are

Variance Function Panel Model Shows

✓ Previously unemployed earn less on average

✓ AND their wages are substantially more dispersed — a second, independent inequality channel

This is exactly the gap xtvfreg is built to close: it lets any researcher — not just this study's authors — jointly estimate mean and variance equations on their own panel data, group by group, using a single Stata command.

The Mean-Variance Framework

Mean Equation

$$E(Y_{it} \mid X_{it}) = X'_{it}\beta$$

The expected outcome for individual i at time t is a linear function of covariates X. Estimated by GLM with a Gaussian family and identity link.

Variance Equation

$$\mathrm{Var}(e_{it} \mid Z_{it}) = \exp\left(Z'_{it}\gamma\right)$$

The variance of the within-person residual is an exponentiated (always-positive) function of covariates Z. Estimated by GLM with a Gamma family and log link.

Z can equal X, differ from X, or partially overlap — theory should guide the choice for each equation.

Between vs. Within Covariates

Every time-varying covariate is split into two parts before it enters meanvars() or varvars():

$$x_{it} = \bar{x}_i + (x_{it} - \bar{x}_i)$$

Between Component ($\bar{x}_i$)

The individual's own mean across all t — constant over time.
Captures stable, cross-sectional (between-person) differences.
Example: mage, mtenure

Within Component ($x_{it} - \bar{x}_i$)

The deviation from the individual's own mean at time t.
Captures change within a person over time — the fixed-effects part.
Example: dage, dtenure

This is the between-within ("hybrid") specification for panel data (Allison 2009).
xtvfreg extends it to variance: including both parts in meanvars() and varvars() separates between- and within-person contributions to BOTH the level and the dispersion of the outcome.

The Iterative Algorithm

1. **Initial Mean Model**
   GLM (Gaussian, identity link) of Y on meanvars()

2. **Within Transformation**
   xtreg, fe on residuals → within-group (fixed-effects) residuals

3. **Variance Model**
   GLM (Gamma, log link) of squared FE residuals on varvars()

4. **Weighted Re-estimation**
   Re-estimate mean model with analytic weight = $1/S^2$

5. **Iterate to Convergence**
   Repeat 2–4 until $|\Delta$ log-likelihood$|$ < converge()

Estimation Details: Why Gamma GLM?

Squared residuals are the dependent variable in the variance equation

- Squared residuals are strictly non-negative and typically right-skewed — the Gaussian family is a poor fit.
- The Gamma family naturally handles non-negative, skewed outcomes.
- The log link guarantees the predicted variance S^2 is always positive, whatever the covariate values.

Convergence criterion

Convergence is judged on the change in the summed log-likelihood of the variance equation between iterations: $|LL(t) - LL(t-1)| <$ converge(). Defaults: converge(1e-6), maxiter(100). Most real applications converge within 5–10 iterations.

What's New in Version 0.4.8

No more beta_[group]

mean_[group] now carries the coefficients AND all diagnostics/decomposition scalars directly — one set of estimates instead of two.

Sanitized group names

Stored estimates use mean_<gname> / var_<gname>, where <gname> strips spaces and converts '.' and '-' to '_' (e.g., group 3.5 → mean_3_5).

table vs. combined, clarified

table = one wide diagnostics table (mean eq. only). combined = two separate SE tables (mean, then variance). They can be used together.

New tvar() option

Accepts a time-variable name for future extensions; not yet used in estimation.

Syntax

```
xtvfreg depvar [if] [in] [pweight], \
  groupvar(varname) panelid(varname) \
  meanvars(varlist) varvars(varlist) \
  [tvar(varname) converge(#) maxiter(#) nolog table combined]
```

Required

groupvar() – grouping variable
panelid() – panel identifier (used with xtreg, fe)
meanvars() – covariates, mean eq.
varvars() – covariates, variance eq.

Optional

converge() – tolerance, default 1e-6
maxiter() – default 100
nolog – suppress iteration log
table / combined – see next slide

table vs. combined: Which to Use?

table

One wide table of the MEAN equation only, across all groups.

Includes diagnostic scalars:
- n_iter, vf_converged
- ll_final
- var_total
- prop_mean, prop_var

Best for a quick diagnostic overview.

combined

TWO separate tables, each with standard errors:

1) Mean equation, all groups
2) Variance equation, all groups

No diagnostic scalars — just coefficients and SEs, ready for a paper.

Best for publication-style comparison.

Both options can be combined in one call — xtvfreg will display all three tables in sequence.

1 Prepare Your Data

Load Your Dataset

```
webuse nlswork, clear
```

Decompose Variables: Between / Within

```
* Between (individual means)
egen mage = mean(age), by(idcode)
egen mtenure = mean(tenure), by(idcode)

* Within (deviations from means)
gen dage = age - mage
gen dtenure = tenure - mtenure
```

Declare Panel Structure (Required!)

```
xtset idcode year
```

2 Run the Basic Command

```
xtvfreg ln_wage, \
  groupvar(south) panelid(idcode) \
  meanvars(collgrad mage mtenure dage dtenure) \
  varvars(collgrad mage mtenure dage dtenure)
```

What Each Part Does

ln_wage → dependent variable (log wage)

groupvar(south) → separate models for south=0 and south=1

panelid(idcode) → panel identifier (individual ID)

meanvars(...) → predictors of the mean (expected wage)

varvars(...) → predictors of the variance (wage dispersion)

3 Understanding the Output — Iteration Log

What You See First: Convergence Progress

```
Iteration History:
──────────────────────────────────────────────────
Iter     Log-likelihood      Change         Criterion
──────────────────────────────────────────────────
1          -1234.5678        0.123456        0.000001
2          -1234.4444        0.012340        0.000001
3          -1234.4321        0.001230        0.000001
──────────────────────────────────────────────────
Converged in 3 iterations
```

What This Tells You

✓ **Converged quickly**
→ The algorithm found a stable joint solution for mean and variance
✓ **Change < Criterion at the final iteration**
→ Results are numerically stable and reliable

3 Understanding the Output — Mean Equation

Mean Equation Results (Non-South, illustrative)

	Coef.	Std.Err.	z	P>|z|
collgrad	0.2989	0.0234	12.77	0.000
mage	0.0041	0.0009	4.56	0.000
dage	0.0038	0.0007	5.43	0.000
_cons	1.2345	0.0567	21.78	0.000

How to Read It

collgrad = 0.30 → college grads earn about 30% more (additive scale, standard OLS-style interpretation)
mage vs. dage → both between- and within-person age effects are positive and similar in size here

3 Understanding the Output — Variance Equation

Variance Equation Results (Log Scale!)

	Coef.	Std.Err.	z	P>|z|
collgrad	0.2989	0.0456	6.56	0.000
mage	0.0203	0.0089	2.28	0.023
_cons	-1.5678	0.1234	-12.71	0.000

CRITICAL: Exponentiate for Interpretation

collgrad = 0.2989 → exp(0.2989) = 1.35
→ College graduates have about 35% MORE wage variance — more heterogeneity in returns to education
mage = 0.0203 → exp(0.0203) = 1.02 → about 2% more variance per year of age

3 Understanding the Output — Variance Decomposition

```
Variance Decomposition:
──────────────────────────────

Total variance:  0.234567
  Mean model:        0.057008  (24.3%)
  Variance model:    0.120139  (51.2%)
  Unexplained:       0.057420  (24.5%)
```

24.3%

Mean Model

Systematic between-group differences

51.2%

Variance Model

Modeled heteroscedasticity

24.5%

Unexplained

Individual FE + residual noise

Here, heteroscedasticity (51%) explains MORE of total wage variance than the mean model (24%) — a pattern mean-only models would completely miss.

4 Interpretation Framework

Mean Equation → Between-Group Inequality

Effect is ADDITIVE. Shifts the average level of the outcome. Interpreted like standard OLS.

Variance Equation → Within-Group Inequality

Effect is MULTIPLICATIVE (log scale). Changes the spread of the outcome. Always exponentiate before interpreting.

Between (mage) vs. within (dage): one is stable across people, the other tracks change within a person over time.

5 Working with Stored Results (Updated)

For each group, xtvfreg (v0.4.8) stores two estimation sets:

mean_<gname> – mean equation coefficients + ALL diagnostics/decomposition (main result)
var_<gname> – variance equation coefficients

<gname> is the sanitized group value: spaces removed, '.' and '-' → '_' (e.g., south=1 → mean_1; a group coded 3.5 → mean_3_5).

Access and Export

```
estimates replay mean_0        // full results, group 0
estimates replay var_1         // variance eq., group 1

esttab mean_*, se star(* 0.10 ** 0.05 *** 0.01)

etable, estimates(mean_0 mean_1) ///
    title(Table 1) export(Results.docx, replace)
```

Hands-On Example 2: Grouping by Race

Any categorical variable can be groupvar() — here, three race categories:

```
xtvfreg ln_wage, \
  groupvar(race) panelid(idcode) \
  meanvars(collgrad mage mtenure dage dtenure) \
  varvars(collgrad mage mtenure dage dtenure) \
  combined nolog
```

Result: three estimate pairs are stored automatically:
mean_1 / var_1 (White) mean_2 / var_2 (Black) mean_3 / var_3 (Other)

combined then prints the mean-equation table for all three groups side by side, followed by the variance-equation table — ideal for spotting where returns to education differ most in level vs. dispersion.

Hands-On Example 3: the table Option

Use table for a fast diagnostic scan across all groups:

```
xtvfreg ln_wage, \
  groupvar(south) panelid(idcode) \
  meanvars(collgrad mage mtenure) varvars(collgrad mage mtenure) \
  table nolog
```

```
Mean Equation Results with Diagnostics (All Groups):
_____________________________________

                    mean_0          mean_1
collgrad          0.284***        0.312***
n_iter                   6               8
vf_converged             1               1
prop_mean            0.243           0.229
prop_var             0.512           0.548
```

One glance tells you: did every group converge, and where does variance dominate the mean model?

Hands-On Example 4: Weights & Subsamples

Probability Weights

```
gen sampwgt = 1
xtvfreg ln_wage [pweight=sampwgt], groupvar(south) panelid(idcode) \
  meanvars(collgrad mage mtenure) varvars(collgrad mage mtenure) nolog
```

Internally, the combined weight used in the mean equation is pweight / S^2 — sampling design and heteroscedasticity are both accounted for.

Subsample with if

```
xtvfreg ln_wage if race==1, groupvar(south) panelid(idcode) \
  meanvars(collgrad mage mtenure) varvars(collgrad mage mtenure) combined
```

Restricts estimation to White respondents (race==1) before splitting by region — useful for sensitivity checks.

Hands-On Example 5: Theory-Driven Specification

meanvars() and varvars() need not match — let theory decide:

```
xtvfreg ln_wage, \
  groupvar(south) panelid(idcode) \
  meanvars(collgrad mage mtenure) \
  varvars(mage dage dhours) \
  nolog
```

This model says:
- Mean depends on education and stable characteristics (between)
- Variance depends on age variability and hours variability (within)

Mirrors the article's logic: dispersion can be driven by a different set of mechanisms than the average level.

xtvfreg vs. Standard FE / RE

Feature	Standard FE / RE	xtvfreg
Variance assumption	Constant (homoscedastic)	Function of covariates
Estimation	OLS / GLS within transform	Iterative weighted GLS
Inequality insight	Between-group only	Between- AND within-group
Efficiency	Standard	Weighted by $1/S^2$ — more efficient when heteroscedastic
Output	Single equation	Mean eq. + variance eq. + variance decomposition

Linking Back to the Article

In the original study…

groupvar() split the sample by prior unemployment history; varvars() modeled how spell history and other covariates shaped within-person wage volatility over the 1985–2008 panel.

In today's demonstration…

south and race stand in as illustrative grouping variables on the public nlswork data — the mechanics are identical to what the published analysis used on the Dutch OSA panel.

Caution: the numeric examples in this talk are illustrative for teaching syntax and interpretation — not a restatement of the published findings.

Practical Tips & Reporting Guidelines

Practical Tips

- Always decompose vars into between/within first
- Check convergence via vf_converged
- If not converging: check collinearity, sparse groups
- Always exponentiate variance-eq. coefficients
- Use table for a quick scan, combined for a paper
- Let theory guide meanvars() vs. varvars()

Reporting Checklist

☐ Report BOTH mean and variance tables
☐ State the link function (log) for variance eq.
☐ Report variance decomposition percentages
☐ Note convergence (iterations, tolerance)
☐ Cite Mooi-Reci & Liao (2025) for the method
☐ Note version of xtvfreg used (e.g., 0.4.8)

Reference & Resources

Based on Published Research

Mooi-Reci, I., and T. F. Liao. 2025. “Unemployment: A Hidden Source of Wage Inequality?” *European Sociological Review* 41(3): 382–394.

https://doi.org/10.1093/esr/jcae029

Stata Help

```
help xtvfreg  (v0.4.8)
```

Author

Tim F. Liao — tfliao@illinois.edu

xtvfreg reveals the complete inequality picture: who has more (mean) and who has more dispersion (variance) — the very insight that motivated its creation.

Another Stata ado, runmixedregls, that implements Hedecker's MIXEDREGLS, by George Leckie, University of Bristol, can fits mixed effects locale scale models, though slower, obtain similar results

Another Stata ado: i_ineq

Implements Published Methodology

T. F. Liao. 2022. "Individual Components of Three Inequality Measures for Analyzing Shapes of Inequality." Sociological Methods & Research 51(3): 1325–1356.

https://doi.org/10.1177/0049124119875961

What it does

It takes an outcome variable (like income), a grouping variable, and optional sampling weights to return individual contributions to an inequality index along with their between-group and within-group components.

The i_ineq ado calculates such individual between- and within-group components of the Gini, the Theil-L, and the Theil-T inequality measures.

Published Application Articles

Assortative Mating: iGini

Y. Shen. 2021. "The Nonlinear Linkage Between Earnings Homogamy and Earnings Inequality Among Married Couples." Demography 58(2): 527–550.

https://doi.org/10.1215/00703370-8996374

City-Level Spatiotemporal Inequality: iGini

C. Xu et al. 2025. *"Inequality and Future Trends in Agricultural Net Primary Productivity Across Chinese Cities Based on Satellite Data." Applied Spatial Analysis & Policy 18: 100.*

https://doi.org/10.1007/s12061-025-09700-3

Civil Engineering: Electricity Distribution Inequality: iTheilT

A. Beck et al. 2026. *"A metric to support higher-resolution equity-based infrastructure prioritization." International Journal of Disaster Risk Reduction 137: 10683.*

https://doi.org/10.1016/s12061-025-09700-3.j.ijdrr.2026.10683

Thank You

Questions and Comments

xtvfreg: Uncovering hidden sources of inequality

`help xtvfreg` • `Version 0.4.8`

Contact information: Tim Liao, tfliao@illinois.edu