# `xtgamfe`: Fixed-effects gamma regression with a log link for strictly positive panel outcomes in Stata

M. Christopher Roebuck*

Working paper. This version: November 2026 (command version 1.1.1).

**Abstract**

Positive health care costs are heavily skewed to the right, gamma regression with a log link is a leading choice for modeling them in cross-section, and panels that follow the same persons over time (multi-year claims extracts, linked survey panels) let cost analyses control for unobserved person-level characteristics. This paper describes `xtgamfe`, a Stata command for fixed-effects gamma regression with a log link, which also reports two estimates of the gamma shape and average partial effects in the units of the outcome. Its coefficients are consistent with the number of periods fixed when each period's error has the same expected share of its panel's total, given the fixed effect and the regressors and offsets of all periods. Errors that are exchangeable across periods satisfy this condition, whereas serial dependence, dispersion that changes over time, and a variance not proportional to the squared mean can violate it; in balanced panels whose error vectors are independent of the regressors, period indicators can absorb a common serial or dispersion pattern. In simulations in which the variance was proportional to the squared mean and panel levels differed widely, fixed-effects Poisson had 1.7 to 2.7 times the root-mean-squared error of the gamma estimator, whereas with other powers of the mean the gamma estimator was biased and Poisson, which is consistent there, much less so. An illustration with Medical Expenditure Panel Survey data relates positive nondrug spending to the days supplied of medication in the prior round. Under the main coding, the `xtgamfe` estimate is within one standard error of zero in both a diabetes and an antipsychotic cohort, whereas the diabetes estimate of fixed-effects Poisson differs from it (nominal bootstrap $p$-values of 0.017 and 0.009 in the two models), and same-round medication use also predicts spending in that cohort, which casts doubt on the lag-only specification. For applied work on health care costs, the results argue for reporting not just the gamma estimates but also fixed-effects Poisson on the same observations, and for reading a material difference between them as evidence against the share condition, the mean specification, or strict exogeneity.

**Keywords:** xtgamfe; gamma regression; fixed effects; panel data; health care costs

*RxEconomics LLC, Miami, FL, USA. Email: mcr@rxeconomics.com. Web: www.rxeconomics.com.

## 1. Introduction

Positive health care expenditures are continuous and heavily skewed to the right. In cross-sectional work, generalized linear models have become a standard tool for such data because they model the mean of costs directly in dollars, with no transformation of the outcome (Blough et al. 1999), and so avoid the retransformation step that makes log-linear models fragile under heteroskedasticity (Duan 1983; Manning 1998). Among them, the gamma regression with a log link, whose variance is proportional to the square of the mean, is "a leading choice in published models of healthcare expenditures" (Deb et al. 2017, 76). Manning and Mullahy (2001) included it among the estimators from which their algorithm for skewed outcomes chooses, and Deb et al. (2017, 4) advise choosing the link and the distribution for the data at hand rather than by default.

Repeated observations of the same individuals (multi-year claims extracts, linked survey panels, enrollment files) allow cost analyses to control for unobserved person-level characteristics that are roughly constant over time: baseline severity, health behaviors, the propensity to seek care. The panel counterpart of the cross-sectional model is a gamma regression with a fixed effect for each person, and the alternatives are pooled or random-effects gamma models, correlated random effects (Mundlak 1978), log-linear fixed effects, and fixed-effects Poisson regression (Hausman et al. 1984; Wooldridge 1999). Fixed effects in nonlinear models generally raise an incidental-parameter problem (Neyman and Scott 1948; Lancaster 2000), to which Poisson is a familiar exception. Official Stata does not offer fixed-effects gamma regression as an ordinary panel estimator, in the way that `xtpoisson, fe` offers the Poisson.

The fixed effects of the gamma model with a log link have a closed-form solution for any value of the coefficients (Hinz et al. 2019; Weidner and Zylkin 2021). Under the gamma model the coefficients are estimated consistently as the number of panels grows with the number of periods fixed, and, among the common parameters, only the shape suffers from the incidental-parameter problem (Dhaene and Jochmans 2019, 7–8). The reason, which goes back to Chamberlain (1985), is that the within-panel shares of the outcome have a distribution that does not involve the fixed effect. Greene (2004, 102) lists the exponential special case among the models whose conditional likelihood gives a consistent estimator

of the coefficients. Wooldridge (1999, sec. 6) treats the same multiplicative model under assumptions on its first two conditional moments alone: a variance proportional to the squared mean and no correlation over time, given the regressors and the effect. He estimates it from the orthogonality conditions that those moments imply, which yields a different estimator.

Weidner and Zylkin (2021, 2) advise against gamma pseudo-likelihood in gravity models with three sets of fixed effects and few periods. Their reason, as they state it, is that it is consistent only if the conditional variance is proportional to the squared mean, whereas Poisson requires the conditional mean alone (2021, 8, after their Proposition 2). The one-way argument behind that statement uses independent lognormal errors (2021, supplementary appendix, 25–29). For the one-way model of this paper, Section 2.3 shows that the estimator is centered by a condition written directly in terms of the within-panel shares of the errors. Called *share centering* below, condition (1) requires that, given the regressors, each period's error have the same expected share of its panel's total. A variance proportional to the squared mean does not guarantee it: independent errors with that variance in every period can violate it. Exchangeable errors (errors whose joint distribution, given the regressors, does not change when the periods are reordered), by contrast, satisfy it even without a finite variance. The variance function still matters: a variance proportional to another power of the mean makes the relative dispersion of the errors depend on the regressors, which can violate the condition.

Two fixed-effects Poisson regressions appear below. With a constant as the dependent variable and the log of the outcome as offset, the auxiliary regression of Section 2.5 reproduces the gamma coefficients with their signs reversed. With the outcome itself as the dependent variable, fixed-effects Poisson (FE Poisson in the tables) is a different estimator of the same conditional mean and the robust alternative with which the gamma estimator is compared. Within Stata, existing commands give the same coefficients by three other routes (Section 2.5); the second is the auxiliary regression just described:

1. `glm` with an indicator for every panel gives the same coefficients, clustered standard errors, fitted means, and average partial effects, but takes about half a minute with 1,000 panels, and the shape implied by the scale it reports is not consistent when the number of periods is fixed (Sections 2.5, 4.1, and 4.9).
2. The auxiliary regression fit with `ppmlhdfe` (Correia et al. 2020) also gives the clustered

standard errors of `xtgamfe`, at the panel or a higher level and with sampling weights, but the fitted means and the shape only with further lines of code (Section 2.5).

3. `xtspj` (Sun and Dhaene 2019) fits gamma and exponential models with fixed effects and information-based standard errors, but whether its optimizer reaches the estimate depends on the units of the outcome, and its estimate of the shape is not consistent when the number of periods is fixed (Sections 2.5 and 4.1).

Outside Stata, gamma models with fixed effects can be fit in R with the `fixest` package (Bergé et al. 2026), the `glmhdfe` package (Hinz et al. 2019, 9), and the `feglm()` function of the `alpaca` package (Stammann and Czarnowske 2025), which supports several of R's model families. They can also be fit in Julia with `GLFixedEffectModels.jl` (Boehm and contributors 2025), which accepts the distributions of `GLM.jl`, the gamma among them. For their simulations of gravity models with three sets of fixed effects, Weidner and Zylkin (2021, supplementary appendix, 30n12 and 31) computed fixed-effects gamma pseudo-likelihood estimates with a modified version of the HDFE-IRLS algorithm used in Correia et al. (2020). None of these is compared here.

This paper contributes to applied work with panel cost data in two ways. First, it provides `xtgamfe`, which fits the estimator as an ordinary Stata panel command. The command

- reports the coefficients on their natural scale in one step, with a cost per iteration that is linear in the number of observations for a given number of regressors;
- clusters the standard errors on panels or on a level in which they are nested;
- accepts frequency, sampling, and importance weights, factor variables, and time-series operators;
- estimates the gamma shape in two ways that, under the gamma model, remain consistent when the number of periods is fixed; and
- returns fitted means, fixed effects, and average partial effects, computed from the fitted means or from the observed outcome, with standard errors that allow for the estimation of the fixed effects.

One of the two shape estimators is the marginal-likelihood estimator of Chamberlain (1985), to which the adjusted profile score of Dhaene and Jochmans (2019) also leads.

Second, the paper sets out when the estimator can be relied on with the number of periods fixed, and when fixed-effects Poisson regression should be preferred.

Share centering is sufficient for consistency. It is the step on which Dhaene and Jochmans (2019) rest their proof for the gamma model, and Section 2.3 reads that step as a condition on errors of any positive distribution. Log-linear fixed effects is consistent instead when the log error has the same mean in every period, a condition that a stationary error process independent of the regressors satisfies even when share centering fails. In the simulations of Section 4.3, a variance proportional to a power of the mean other than two (1, 1.5, 2.5, or 3) biased the coefficients by 4 to 36 percent of their value with two to eight periods, upward below two and downward above it, and period indicators did not help. Serial dependence biased the coefficient on a treatment that switches on in the last period by 5 to 17 percent, about as much with eight periods as with four when the autocorrelation was 0.4 and nearly twice as much when it was 0.8. That bias, and the bias from dispersion that changes over time, disappear with period indicators in balanced panels when the error vectors are independent of the regressors (Proposition 2), and in unbalanced panels under further conditions.

The advantage of the gamma estimator over fixed-effects Poisson is precision, but a smaller standard deviation is no guide by itself. When the variance was proportional to a power of the mean other than two, or when the dispersion of the errors depended on a regressor, the gamma estimator was biased and its intervals often missed the true value, although in some of these cases its standard deviation was the smaller (Section 4.3). How large the gain is depends on how much levels differ across panels (Section 4.4), and a Park test does not settle which case applies (Section 6).

Three quantities should also be kept apart: the coefficients; the fitted mean of an individual panel, which at the true coefficients is unbiased but noisy, because its fixed effect is estimated from a few observations however many panels there are; and averages over observations, such as average partial effects. These averages are consistent under share centering, but computed from fitted means they generally are not when the coefficients owe their consistency to period indicators; `estat ape, ybar` then computes them from the observed outcome (Section 2.7).

Several of the results below restate existing ones, credited where they appear: the closed form of the fixed effects, the share argument of Proposition 1, and the share likelihood of

Proposition 3 with its estimator of the shape. For fixed-effects Poisson, Santos Silva and Kemp (2016, slide 17) and Martin's (2017b) working paper, a version of which appeared as Martin (2017a), give the identity behind effects computed from the observed outcome. Martin (2017b) also shows that averages over the estimated fixed effects are consistent with the number of periods fixed and derives their asymptotic variance; the theorems cited below follow the working paper's numbering. The paper's own results are:

1. the reading of the share argument as a condition on errors of any positive distribution, with the ways in which the condition fails (Section 2.3);
2. the period-indicator result of Proposition 2 and its extensions to unbalanced panels;
3. the second-order approximation (3);
4. effects computed from the observed outcome when period indicators may absorb a pattern in the errors, with their standard errors (Section 2.7); and
5. the standard error of the shape (Appendix A.4).

The auxiliary Poisson regression of Section 2.5 is a computational device assembled from standard results; as far as the author knows, it has not been stated before for fixed-effects gamma or exponential regression.

Section 2 sets out the model, the condition, and the role of period indicators, with the proofs in Appendix A. Section 3 describes the command; its worked example (Section 3.4) shows how a coefficient of 0.499 becomes an average partial effect of 6.50 outcome units per unit of the regressor (6.35 when computed from the fitted means). Section 4 reports the comparisons with the other routes and the Monte Carlo evidence, Section 5 applies the command to data from the Medical Expenditure Panel Survey, and Section 6 draws the implications for applied work.

## 2. Model and estimator

The argument of this section has five steps. First, the fixed effects can be solved out, so that the estimator depends on the outcome only through its within-panel shares (Section 2.2). Second, the estimator is consistent under share centering: given the regressors, the error of each period has the same expected share of its panel's total, one over the number of periods

(Section 2.3). Third, that holds whenever the errors of a panel are exchangeable across periods given the regressors, whatever their distribution, so that the gamma assumption is not needed for the coefficients. Fourth, it can fail when the distribution of the errors depends on a regressor (as it does when the variance is proportional to a power of the mean other than two), when the errors are serially correlated, or when their dispersion changes over time. Fifth, in balanced panels, period indicators restore the consistency of the other coefficients in the last two cases when the errors are independent of the regressors (Proposition 2), although average partial effects computed from fitted means generally remain inconsistent.

### 2.1 Setup

Let $i = 1, \dots, G$ index panels and $t = 1, \dots, T_i$ the observations of panel $i$, with $n = \sum_i T_i$. The outcome $y_{it}$ is strictly positive, and its conditional mean is

$$\mathrm{E}[\, y_{it} \mid \mathbf{X}_i, \alpha_i \,] = \mu_{it} = \exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}_0 + o_{it}),$$

where $\alpha_i$ is a fixed effect, $\mathbf{X}_i$ collects the regressors (and offsets) of every period of panel $i$, and $o_{it}$ is an optional known offset. Conditioning on the whole of $\mathbf{X}_i$ imposes strict exogeneity. In multiplicative form, $y_{it} = \mu_{it}\,\varepsilon_{it}$ with $\varepsilon_{it} > 0$ and $\mathrm{E}[\varepsilon_{it} \mid \mathbf{X}_i, \alpha_i] = 1$: the error of each period has mean one given the regressors of every period. Panels are assumed to be independent of one another (or, when the standard errors are clustered on a level that nests them, clusters of panels are), whereas observations within a panel may be dependent. Panels with a single observation carry no information about $\boldsymbol{\beta}_0$ and are dropped.

The *gamma model* adds a distributional assumption to this mean: given $(\mathbf{X}_i, \alpha_i)$, the $y_{it}$ are independent gamma variates with means $\mu_{it}$ and a common shape $\kappa$, so that $\mathrm{Var}(y_{it} \mid \mathbf{X}_i, \alpha_i) = \mu_{it}^2/\kappa$ (McCullagh and Nelder 1989). The estimator below is maximum likelihood for $\boldsymbol{\beta}_0$ under the gamma model and, outside it, a pseudo-likelihood estimator whose consistency depends on the condition given in Section 2.3. Robust standard errors do not relax that condition, because they correct the variance of an estimator and cannot re-center one.

### 2.2 The profiled criterion

Apart from terms that do not involve the mean, the gamma log likelihood is $-\kappa$ times

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_i \sum_t \Big[ \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + o_{it} + y_{it} \exp(-\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta} - o_{it}) \Big],$$

so that the coefficient estimates minimize $Q$ whatever the value of $\kappa$. For given $\boldsymbol{\beta}$, the minimizing fixed effect is

$$\hat{\alpha}_i(\boldsymbol{\beta}) = \ln\Big( \tfrac{1}{T_i} \sum_t y_{it} \exp(-\mathbf{x}_{it}'\boldsymbol{\beta} - o_{it}) \Big),$$

a closed form that Dhaene and Jochmans (2019, 5–7) display, in their scale parameterization, with the profile log likelihoods of the exponential and gamma models, that Hinz et al. (2019) use for estimation, and that Weidner and Zylkin (2021) give for a class of pseudo-likelihoods that includes the gamma. Substituting it into $Q$ leaves, up to a constant, a criterion in the coefficients alone:

$$Q^p(\boldsymbol{\beta}) = \sum_i \Big[ T_i \, \hat{\alpha}_i(\boldsymbol{\beta}) + \sum_t \mathbf{x}_{it}'\boldsymbol{\beta} \Big].$$

Let $w_{it}(\boldsymbol{\beta}) = y_{it} \exp(-\mathbf{x}_{it}'\boldsymbol{\beta} - o_{it}) \big/ \sum_s y_{is} \exp(-\mathbf{x}_{is}'\boldsymbol{\beta} - o_{is})$, the share of period $t$ in the panel's total of mean-scaled outcomes. The score (the gradient of the panel's contribution to $Q^p$, which is minimized) and the Hessian of panel $i$ are

$$\mathbf{g}_i(\boldsymbol{\beta}) = \sum_t \mathbf{x}_{it} \Big( 1 - T_i \, w_{it} \Big), \qquad \mathbf{H}_i(\boldsymbol{\beta}) = T_i \sum_t w_{it} \, (\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)',$$

with $\bar{\mathbf{x}}_i^w = \sum_t w_{it} \, \mathbf{x}_{it}$. Since $T_i w_{it} = y_{it}/\hat{\mu}_{it}$, the score is the sum of the regressors times $1 - y_{it}/\hat{\mu}_{it}$, the negative of the relative residual of Section 2.6. Those scalar terms sum to zero within every panel at any $\boldsymbol{\beta}$ as a consequence of profiling, not of optimization, whereas the vector terms $\mathbf{x}_{it}(1 - y_{it}/\hat{\mu}_{it})$ need not.

Three properties of the criterion follow directly. First, each $\mathbf{H}_i$ is $T_i$ times a weighted within-panel covariance matrix of the regressors, so that $Q^p$ is convex, and strictly convex whenever the regressors, centered within panels, have full column rank. Starting values do not change the unique minimizer, although they can affect numerical convergence. Second, adding a panel-specific constant to a regressor, or multiplying every outcome of a panel by a positive constant, leaves the estimate unchanged, so that only within-panel variation in

the regressors and only within-panel shares of the outcome matter. Third, a finite minimizer exists whenever that rank condition holds: because the centered regressors take both signs within some panel in any direction, the criterion increases without bound in every direction, and no analogue of separation in logit or Poisson models can arise when the outcomes are strictly positive.

### 2.3 When the estimator is consistent

Under standard regularity conditions for estimators that minimize a sample criterion (Newey and McFadden 1994), $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal as $G \to \infty$ with the $T_i$ fixed if the regressors are identified by within-panel variation and the panel score has mean zero at $\boldsymbol{\beta}_0$ (Appendix A.1 states the conditions). No further moment of the errors is needed to center the score. At $\boldsymbol{\beta}_0$ the shares $w_{it}$ equal the shares of the errors,

$$W_{it} = \frac{\varepsilon_{it}}{\sum_s \varepsilon_{is}},$$

because the fixed effect cancels from the ratio. The mean-zero requirement is therefore $\mathrm{E}\Big[\sum_t \mathbf{x}_{it}(1 - T_i W_{it})\Big] = 0$, and a sufficient condition, which restricts the errors alone, given the regressors, is share centering:

$$\mathrm{E}[\, W_{it} \mid \mathbf{X}_i, \alpha_i \,] = \frac{1}{T_i}, \qquad t = 1, \ldots, T_i. \tag{1}$$

In an unbalanced panel the expectation in (1), and the exchangeability in Proposition 1, are understood conditional on the set of observed periods as well.

**Proposition 1.** If the errors $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i})$ are exchangeable given $(\mathbf{X}_i, \alpha_i)$, then share centering holds.

*Proof.* Exchangeability gives every $W_{it}$ the same conditional expectation, and the $W_{it}$ sum to one. □

Proposition 1 covers errors that are independent and identically distributed with any positive distribution of mean one, including distributions without a finite variance, so that the gamma distribution is not needed for the coefficients. The proposition restates an existing argument. Dhaene and Jochmans (2019, 6) prove that the profile score of the exponential

regression model has mean zero by exactly this step: the shares of identically distributed errors each have expectation $1/T_i$. They reuse the step for the gamma model (2019, 14–15). Read as a condition on errors of any positive distribution, that step uses only share centering, and so requires neither the gamma family nor exchangeability. Exchangeability delivers it; identical marginal distributions alone do not, although with the independent errors that Dhaene and Jochmans consider they imply exchangeability.

Share centering is not the same as a constant coefficient of variation. Weidner and Zylkin (2021, supplementary appendix, 30n10) state the requirement for the gamma pseudo-likelihood as a conditional variance proportional to the squared mean, and the one-way lemma behind it is proved within a family of independent lognormal errors whose variance is a power of the mean (2021, supplementary appendix, 25–29). Share centering, by contrast, is a first-moment condition on shares and concerns how the errors of a panel relate to one another. It can hold when no variance exists, or when the dispersion of the errors differs from panel to panel. It can also fail when the variance is proportional to the squared mean in every period, even with errors that are independent across periods but lie outside the lognormal family of that lemma. With two periods, for example, an exponential error followed by a lognormal error of the same mean and variance has an expected share of 0.47 in the first, so that the score of a regressor that switches on in the second period does not have mean zero. Three failures are examined below.

1. *An error distribution that depends on the regressors.* If, for example, the dispersion of $\varepsilon_{it}$ rises with a regressor, the expected shares can depend on the regressors and fail to center the score. A variance proportional to a power of the mean other than two is a case in point, because the coefficient of variation of the error then changes with the mean, and so with the regressors (Section 4.3). Fixed-effects Poisson remains consistent in this case, because a correctly specified conditional mean suffices for its consistency (Wooldridge 1999), given the moment conditions of Section 6.
2. *Serial dependence that weakens with distance.* Serial dependence can make the expected share of a period at the edge of a panel differ from that of a period in the middle, despite identical marginal distributions; the positive transformations of a stationary Gaussian autoregression in Section 4.3 provide examples. A regressor with a trend

within panels, such as a treatment that switches on and stays on, can then be correlated with the expected shares. (Applying the same positive transformation to Gaussian variables with a common mean, a common variance, and a common correlation preserves exchangeability and satisfies the condition.)

3. *Dispersion that differs across periods.* Independent errors whose shape changes over time can have expected shares that change over time as well.

The direction and approximate size of these failures follow from a second-order expansion of the expected shares (Appendix A.7). Let $\omega_{it} = \sum_s \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is} \mid \mathbf{X}_i, \alpha_i)$, the sum of row $t$ of the conditional covariance matrix of the panel's errors, and let $\bar{\omega}_i = T_i^{-1} \sum_t \omega_{it}$. Then

$$\mathrm{E}[\, T_i W_{it} \mid \mathbf{X}_i, \alpha_i \,] = 1 - \frac{\omega_{it} - \bar{\omega}_i}{T_i} + r_{it}, \tag{2}$$

where $r_{it}$, given in Appendix A.7, is of third order in the size of the deviations $\varepsilon_{is} - 1$ under the small-dispersion and moment conditions stated there: to second order, a period whose error varies more, or moves more with the other errors of its panel, has the smaller expected share. In a balanced panel, linearizing the expected score at $\boldsymbol{\beta}_0$ then gives the probability limit $\boldsymbol{\beta}^*$ of $\hat{\boldsymbol{\beta}}$ as $G \to \infty$,

$$\boldsymbol{\beta}^* - \boldsymbol{\beta}_0 \approx -\frac{\boldsymbol{\pi}}{T}, \qquad \boldsymbol{\pi} = \left\{ \mathrm{E} \sum_t (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right\}^{-1} \mathrm{E} \sum_t (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\, \omega_{it}, \tag{3}$$

with $\bar{\mathbf{x}}_i = T^{-1} \sum_t \mathbf{x}_{it}$, so that $\boldsymbol{\pi}$ holds the coefficients of the population within-panel regression of $\omega_{it}$ on the regressors. To second order, therefore, the estimated coefficient of a regressor is biased downward when its slope in that regression is positive (that is, when, holding the other regressors fixed, the regressor is high in periods with large $\omega_{it}$), and upward when that slope is negative. The sign of this partial association can differ from that of the simple one. When the errors are independent across periods, $\omega_{it}$ is the conditional variance of $\varepsilon_{it}$. If it depends only on the fixed effect and the regressors of the same period, and the regressors are independent and identically distributed across periods given the fixed effect, as in the designs of Tables 4 and 5, then $\boldsymbol{\pi}$ does not depend on $T$, and to second order the inconsistency falls in proportion to $1/T$.

With stationary errors, independent of the regressors and the fixed effect, whose autoco-

variances are positive and decline with the lag, the first and last periods have neighbors on one side only and so the smallest $\omega_{it}$. An edge indicator, such as a treatment that switches on in the last period, can then be biased when its within-panel variation aligns with the pattern of expected shares; for that switch, because the shortfall at the edge builds up over the range of the dependence, the bias can grow with $T$ before it declines.

The second-order term of (2), and so (3), involves only second moments. It therefore misses failures that come from higher moments alone: for the exponential and lognormal errors above it gives a share of exactly 1/2, against 0.47. In the balanced designs of Section 4.3, (3) has the sign of every bias and is within about 15 percent of it with gamma errors of moderate dispersion. It is up to about 2.2 times the bias when the squared coefficient of variation is large across observations or the errors are strongly skewed, as with the lognormal margins.

In the second and third failures, when the error vectors are independent of the regressors and the fixed effect, $\omega_{it}$ depends on $t$ alone, so that in a balanced panel $\omega_{it} - \bar{\omega}_i$ is a linear combination of the within-panel deviations of period indicators, and (3) assigns it to their coefficients rather than to those of the other regressors. In the first failure, $\omega_{it}$ varies with the regressors themselves. Period indicators can therefore repair the second and third cases, and Proposition 2 shows that the repair is exact under its conditions, which do not require independence from the fixed effect.

**Proposition 2 (period indicators).** Suppose that every panel is observed in the same $T$ periods, that the model includes an indicator for each period but one, that the regressors, together with the period indicators, are identified by within-panel variation in the population, and that, for a panel drawn at random, the error vector $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ is independent of $\mathbf{X}_i$. Then, under conditions (i), (ii), and (iv) of Appendix A.1, the estimator of the other coefficients is consistent, and the estimated period coefficients converge to the true ones plus a vector $\boldsymbol{\delta}^*$ that need not be zero; the proof is in Appendix A.1.

The period coefficients absorb the period pattern of the expected shares, and the price is that they no longer estimate period effects on the mean, which matters for fitted means and average partial effects (Section 2.7). Here $\boldsymbol{\delta}^*$ is the vector at which the errors, rescaled by $e^{-\delta_t^*}$, have expected shares of $1/T$. The distribution of the errors may differ across panels through

unobserved characteristics, provided that these are independent of the regressors; what the proposition excludes is an error distribution related to the regressors. Independence between the errors and the fixed effect is not required as such, because the fixed effect cancels from every share. Since the regressors are generally correlated with the fixed effect, however, an error distribution that varies with it will usually violate the independence from the regressors that the proposition requires. The argument parallels regression M-estimation, in which a free intercept absorbs asymmetric errors that are independent of the regressors and the slopes stay consistent (Carroll and Welsh 1988; Maronna et al. 2019, sec. 4.9.2); here the period effects play the intercept's part.

When, within each pattern of observation (the set of periods in which a panel is observed), the errors are independent of the regressors, the same argument gives two extensions to unbalanced panels (Appendix A.2). First, calendar-period indicators suffice if the period means of the regressors do not depend on the pattern of observation apart from a constant. Second, with an indicator for every period within every pattern of observation (one omitted per pattern), no further condition on the period means of the regressors is needed, provided the slopes remain identified. Indicators for the position of an observation within its panel, by the number of periods observed, adjust for the errors under two further assumptions: that panels are observed for consecutive periods, and that the observed errors have the same distribution in all windows of the same length, as when the error process is stationary and independent of where the window falls. They are added to the model's own specification of the mean, which must still include any real calendar effects. When the first extension does not apply and only calendar-period indicators are included (as with a treatment timed relative to each panel's own window of observation), bias can remain (Section 4.3). Neither proposition helps with the first kind of failure.

Under the usual higher-order regularity conditions, finite-$G$ bias is generally of order $1/G$; no correction is applied.

Log-linear fixed effects, the within regression of $\ln y_{it} - o_{it}$ on the regressors, estimates the same coefficients under a different condition. Because $\ln y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}_0 + o_{it} + \ln \varepsilon_{it}$, log-linear fixed effects is consistent when $\mathrm{E}[\ln \varepsilon_{it} \mid \mathbf{X}_i]$ is finite and constant across periods within each panel, so that the within transformation removes it with the fixed effect. With

period indicators, it is also consistent when that conditional mean differs across periods only by amounts common to all panels. The period coefficients then absorb those amounts while the other slopes remain consistent. When the logarithm of the errors has a finite mean, exchangeable errors satisfy this condition as well as share centering, errors that meet the conditions of Proposition 2 satisfy it with period indicators, and a stationary error process independent of the regressors satisfies it with no period indicators, whereas such a process can violate share centering, as the second failure above shows. When the distribution of the errors depends on the regressors, as in the first failure, neither condition generally holds, and both estimators can be inconsistent (Section 4.3). Section 6 compares the three estimators of the mean.

### 2.4 The likelihood of the within-panel shares

Under the gamma model the criterion has a second interpretation, which supplies both a likelihood that is free of the fixed effects and an estimator of the shape. Let $s_{it} = y_{it}/\sum_r y_{ir}$ be the observed shares and $q_{it} = \exp(-\mathbf{x}_{it}'\boldsymbol{\beta} - o_{it})$.

**Proposition 3 (share likelihood).** Under the gamma model, the density of the shares $(s_{i1}, \ldots, s_{i,T_i-1})$ of panel $i$ is

$$f(\mathbf{s}_i) = \frac{\Gamma(T_i\kappa)}{\Gamma(\kappa)^{T_i}} \prod_t s_{it}^{\kappa-1} \frac{\prod_t q_{it}^{\kappa}}{\left(\sum_t s_{it}\, q_{it}\right)^{T_i\kappa}},$$

which does not involve $\alpha_i$. Its logarithm is $\ln\Gamma(T_i\kappa) - T_i \ln\Gamma(\kappa) + \kappa\Lambda_i(\boldsymbol{\beta}) - \sum_t \ln s_{it}$, where $\Lambda_i(\boldsymbol{\beta}) = \sum_t \ln w_{it}(\boldsymbol{\beta})$, and the part that depends on $\boldsymbol{\beta}$ equals $-\kappa\, Q_i^p(\boldsymbol{\beta})$ up to a constant.

Chamberlain (1985, 25–27) established that the distribution of within-panel shares or ratios is free of the individual effect in Weibull and gamma models. He constructs a marginal likelihood from the ratios of later spells to the first for any number of spells, and illustrates the gamma case with the share of the first of two spells. Dhaene and Jochmans (2019, 8) state that their adjusted profile score, which leaves the equations for the coefficients unchanged and corrects the one for the shape, is the score of that marginal log likelihood of the ratios $y_{it}/y_{i1}$. It follows that the part of the profile log likelihood that depends on the coefficients equals that of the marginal log likelihood, and the maximum likelihood estimate of the shape in Section 2.6 is Chamberlain's marginal-likelihood estimator, to which their adjusted score

also leads. The command uses this density for that estimate, for the model-based covariance matrix of `vce(oim)`, and for the log likelihood of the information criteria. The proof of Proposition 3 is in Appendix A.3.

### 2.5 Other routes to the same coefficients

Three computational routes, which use four existing commands, minimize the criterion of Section 2.2.

*Indicators for every panel.* `glm y x i.id, family(gamma) link(log) vce(cluster id)` estimates the fixed effects as parameters. Fit to the same observations (`xtgamfe` drops panels with a single observation), the coefficients on the regressors and their clustered standard errors equal those of `xtgamfe` (with the default maximum likelihood optimization of `glm`). Its `irls` option builds the sandwich on the expected information of the model with the panel indicators rather than on the observed curvature of the likelihood, which `xtgamfe` and the default use. When each fixed effect is estimated from a few observations the two differ. Under share centering the expected information exceeds the observed curvature on average. In balanced panels under the gamma model with independent panels, the expected information used by `irls` is $(T\kappa + 1)/(T\kappa)$ times the population mean of the observed curvature (Appendix A.6). Because the score covariance is the same, the clustered standard errors of `irls` are too small in large samples (0.029 against 0.034 for $x_1$ in the example of Section 3.4). Outside that case the two sandwiches can differ in either direction, and even within it the observed curvature can be the larger in a particular sample. The default is the one to compare. After that fit, `predict` with the `mu` option gives the fitted means of `xtgamfe`, and `margins, dydx(varlist) vce(unconditional)` its average partial effects, with standard errors that agree with those of `estat ape` to the precision of the numerical derivatives of `margins` (Section 4.1). The scale that `glm` reports, however, is the reciprocal of the Pearson estimate of Section 2.6, which is inconsistent when the number of periods is fixed. The cost is one parameter per panel, so that these lines serve as a check when there are few panels.

*A conditional Poisson regression of a constant.* Consider a Poisson model for a variable that equals one in every observation, with regressors $\mathbf{x}_{it}$, a fixed effect, and $\ln y_{it} - o_{it}$ as

offset. Conditioning on the panel total of that variable, which is $T_i$, removes the fixed effect (Hausman et al. 1984), and the conditional log likelihood of panel $i$ at coefficient vector $\boldsymbol{\gamma}$ is, up to a constant,

$$\ell_i(\boldsymbol{\gamma}) = \sum_t \mathbf{x}_{it}'\boldsymbol{\gamma} - T_i \ln \sum_t y_{it}\, e^{-o_{it}} \exp(\mathbf{x}_{it}'\boldsymbol{\gamma}).$$

Comparison with $Q_i^p$ shows that $\ell_i(-\boldsymbol{\beta}) = -Q_i^p(\boldsymbol{\beta})$ plus a constant, so that the two estimators are the same with the signs reversed. The auxiliary regression combines three standard results. First is the conditioning argument just used, of which Chamberlain (1985, sec. 3.3) gives a version for a gamma duration model, conditioning on the number of complete spells and their total duration. Second is the equivalence of the estimating equations of exponential and Poisson data, which allows an exponential model to be fitted as a Poisson model with the log of the outcome as offset, a device familiar from survival analysis (Holford 1980; Aitkin and Clayton 1980; Laird and Olivier 1981). Third is the fact that the coefficients of the gamma model do not depend on the shape (Section 2.2), so that those of the exponential model serve for every shape. (Here every observation counts as one event after an exposure time equal to the outcome, with no censoring, and the coefficients of the event rate are those of the mean with the opposite sign.) The author has not found these combined for fixed-effects gamma or exponential regression, although the step from them is direct. The auxiliary regression is offered here as a computational device: it requires only official Stata, and it gives an independent check on any implementation of the estimator.

In Stata, for a mean model without an offset (with one, the Poisson offset is the log of the outcome minus that offset),

```
generate byte   one = 1
generate double lny = ln(y)
xtpoisson one x1 x2, fe offset(lny) vce(robust)
```

returns $-\hat{\boldsymbol{\beta}}$, and its robust covariance matrix multiplied by $G/(G-1)$ equals that of `xtgamfe`. The factor is exact, because `xtpoisson, fe` applies no finite-sample correction for the number of clusters (Correia et al. 2020, 107n14). This regression must be distinguished from fixed-effects Poisson regression of the outcome itself (`xtpoisson y x1 x2, fe`), which is a different estimator and the robust alternative discussed throughout. The fitted means

and the shape estimate can be recovered from the auxiliary regression with a few further lines; `xtpoisson, fe` does not accept a cluster variable, sampling weights, or weights that vary within panel. `ppmlhdfe` (Correia et al. 2020) fits the same regression (`ppmlhdfe one x1 x2, absorb(id) offset(lny)`) and accepts both clustering above the panel and sampling weights that vary within panel. With `vce(cluster)`, its sign-reversed coefficients and clustered standard errors equal those of `xtgamfe` with no correction factor (Section 4.1). It can also add further sets of fixed effects, although the conditions of Section 2.3 are not extended to that case here.

*Split-panel jackknife software.* `xtspj y x, model(exponential) method(none)` (Sun and Dhaene 2019) maximizes the same criterion (the exponential model is the gamma model with $\kappa = 1$, and the coefficients do not depend on $\kappa$) and reports standard errors based on the information matrix, with no robust or clustered alternative for its nonlinear models. Those standard errors are $\sqrt{\kappa}$ times the correct ones when the errors are gamma with shape $\kappa$ (Appendix A.6), too small when $\kappa < 1$ and too large when $\kappa > 1$. Its optimizer, moreover, does not always reach the minimum, and whether it does depends on the units of the outcome, although rescaling the outcome only shifts the criterion by a constant (Section 4.1). With `model(gamma)` it estimates the shape jointly with the fixed effects, so that the shape is inconsistent when the number of periods is fixed (Dhaene and Jochmans 2019, 7–8); the information-based standard errors that use it are then inconsistent as well. It is designed for panels whose length grows with their number and allows predetermined regressors (Sun and Dhaene 2019, 335–37), which `xtgamfe` does not. It also offers the split-panel jackknife bias correction of Dhaene and Jochmans (2015), which `xtgamfe` does not. That correction is aimed at models whose coefficients are inconsistent when the number of periods is fixed, and under the conditions of Section 2.3 the coefficients of this model need none.

### 2.6 The shape parameter

The coefficients do not depend on $\kappa$, but the model-based covariance matrix and any statement about dispersion do. The usual shape estimator is inconsistent when the number of periods is fixed. Let $\rho_{it} = y_{it}/\hat{\mu}_{it} - 1$ be the relative residuals, which are also the Pearson residuals for the variance function $\mu_{it}^2$. The estimator that divides the residual degrees of freedom by

their sum of squares, $(n-G-K)/\sum_{it}\rho_{it}^2$ with $K$ the number of coefficients, converges to $\kappa+1/T$ in a balanced panel, because the fixed effects are estimated from $T$ observations each. The command reports that quantity only for reference (`e(kappa_pearson)`) and provides two estimators that are consistent under the gamma model as $G\to\infty$.

The first is a moment estimator, `e(kappa)`, which equates the sum of squared relative residuals to its expectation under the gamma model at the true coefficients, $\sum_i T_i(T_i-1)/(\kappa T_i+1)$ (Appendix A.4).

The second, `e(kappa_ml)`, maximizes the log likelihood of the shares of Proposition 3, evaluated at $\hat{\boldsymbol{\beta}}$, and that log likelihood is strictly concave in $\kappa$. When the fitted shares are exactly equal within every panel, as in an exact fit, neither estimating equation has a finite solution, and the command reports no shape (Section 3.2). Because the slope score is zero at $\hat{\boldsymbol{\beta}}$ for every $\kappa$, the standard error of the shape follows from the derivatives of that log likelihood in $\kappa$ alone: by default a sandwich, clustered like the coefficients, and with `vce(oim)` the inverse of minus the second derivative (Appendix A.4). The two estimators agree in large samples when the gamma model holds (Section 4.7), so a marked difference between them is a sign that it does not.

The fixed effects themselves are not estimated consistently with $T$ fixed. Under the gamma model, at the true coefficients, $\hat{\alpha}_i-\alpha_i$ is the logarithm of the mean of the $T_i$ errors of the panel, and its expectation is $\psi(T_i\kappa)-\ln(T_i\kappa)$, where $\psi$ is the digamma function. The expectation is approximately $-1/(2T_i\kappa)$ unless $T_i\kappa$ is small. In contrast, at the true coefficients $\exp(\hat{\alpha}_i)$ is unbiased for $\exp(\alpha_i)$ whenever the errors have mean one, and that property is what fitted means and average partial effects rely on. The fitted mean of a single panel is unbiased at the true coefficients but not consistent: its ratio to the true mean is the average of the panel's $T_i$ errors, whose variance under the gamma model is $1/(T_i\kappa)$ however many panels are observed. Averages over many panels do not have this problem (Section 2.7).

### 2.7 Average partial effects

For a regressor entering the index linearly, $\beta_j$ is the derivative of the log conditional mean with respect to $x_j$, and $\exp(\beta_j)$ is the ratio of means for a one-unit change. A change in the mean itself (in dollars, say) depends on the level of the mean and therefore on the fixed effect.

The average partial effect of a continuous regressor that enters the index linearly is

$$\theta_j = \frac{\mathrm{E}\left[\sum_t \beta_j \, \mu_{it}\right]}{\mathrm{E}[T_i]}, \qquad \hat{\theta}_j = \hat{\beta}_j \, \frac{1}{n} \sum_{i,t} \hat{\mu}_{it},$$

an average over observations (so that a panel with more observations counts for more) in the population from which the panels are drawn, which, because panels with a single observation are dropped, is that of panels with at least two observations. With weights, the averages are weighted and the sum of the weights replaces $n$. With sampling weights equal to the inverse probabilities of selection, the weighted averages estimate those of the population from which the panels are drawn; with other weights they estimate weighted averages of the partial effects, provided that the errors also have mean one given the weights. For a regressor that enters through polynomial or interaction terms, $\beta_j$ is replaced by the derivative of the index, and for a factor variable the effect of a level is the average difference between the fitted means with every observation set to that level and to the base level, holding the fixed effects at their estimates.

Although the individual $\hat{\alpha}_i$ are noisy, $\hat{\theta}_j$ is consistent as $G \to \infty$ under share centering, provided the averaged quantity has a finite mean. For a regressor that enters linearly, a finite mean of the outcome suffices; when regressors are unbounded, the usual dominance condition in a neighborhood of the true coefficients is also needed. In an unbalanced panel, the errors must further have mean one given the pattern of observation. At the true coefficients $\exp(\hat{\alpha}_i)$ is unbiased for $\exp(\alpha_i)$, so that the fitted means are unbiased for the true ones and their average obeys a law of large numbers across panels; Martin (2017b, Theorem 1) gives the same argument for averages over the estimated fixed effects of fixed-effects Poisson. Consistency of every coefficient entering the fitted mean is sufficient for this argument.

When share centering fails and the coefficients of interest are consistent only by Proposition 2, the period coefficients generally are not. At the limiting common coefficients, the conditional expectation of each profiled fitted mean is the true mean multiplied by $e^{\delta_t^*} \, T^{-1} \sum_s e^{-\delta_s^*}$, and $\hat{\theta}_j$ is generally inconsistent. Averaged over the periods, that factor is at least one (by the Cauchy–Schwarz inequality), so that the effect is too large in magnitude when the true means have no period pattern. In a design of Section 4.8 with period-specific dispersion, the

coefficient is centered and the average partial effect 3 percent too large with four periods.

The remedy is to average the observed outcome in place of the fitted mean. Period indicators can center the other slopes while absorbing a pattern in the errors, and if the coefficients that enter a change are consistent, the average of the outcome times the implied proportional change estimates the effect in outcome units. For a regressor that enters linearly, $\theta_j = \beta_j \,\mathrm{E}[\sum_t y_{it}]/\mathrm{E}[T_i]$, because the conditional mean of $y_{it}$ is $\mu_{it}$; Santos Silva and Kemp (2016, slide 17) and Martin (2017b) give this identity for the exponential model with a linear index, estimated by fixed-effects Poisson, in balanced panels.

The identity needs that conditional mean in every period averaged. In an unbalanced panel the errors must therefore have mean one given the pattern of observation as well, as for fitted means; in a weighted fit they must also have mean one given the weights, and the averages are weighted. Share centering and consistent slopes do not imply this: if which periods are observed depends on a shock common to the errors of a panel, the shock cancels from the shares but not from the mean of the outcomes retained. The product of $\hat{\beta}_j$ and the sample mean of the outcome is therefore consistent whenever $\hat{\beta}_j$ is, the errors have mean one on the periods averaged, and the outcome mean is finite and obeys a law of large numbers across panels. The argument uses nothing about the estimator beyond the consistency of $\hat{\beta}_j$, so that the same product is consistent after fixed-effects Poisson or log-linear fixed effects as well; Martin makes the related point that, because the effect does not depend on the fixed effect, the coefficients can be estimated with regressors that are only weakly exogenous, by sequential moment restrictions.

The same substitution gives every other effect. For a polynomial or interaction, the effect is the average of $y_{it}$ times the derivative of the index; for a factor level, it is the average of $y_{it}\,[\exp(\eta^{\ell}_{it} - \eta_{it}) - \exp(\eta^{0}_{it} - \eta_{it})]$, where $\eta_{it}$ is the observed index and $\eta^{\ell}_{it}$ and $\eta^{0}_{it}$ are the indices with the factor set to the level and to the base level. Each effect is consistent when the coefficients entering the change are consistent, the errors have mean one on the periods averaged, and the absolute value of the outcome times the proportional change is bounded in a neighborhood of the limiting coefficients by an integrable function (the counterpart of the dominance condition for fitted means). For a regressor that enters linearly, a finite mean of the outcome is enough. The period coefficients generally are not consistent here, so that

the argument does not cover effects of the period indicators themselves. `estat ape, ybar` reports the effects of the other regressors in this way, and those of the period indicators with a note to that effect. When share centering holds, the two kinds of effect estimate the same quantity.

The standard error of $\hat{\theta}_j$ must allow for three sources of variation: the coefficients, the fixed effects, and the sample over which the average is taken (under moment conditions on the influence function, given in Appendix A.1, which coefficient consistency does not require). `estat ape` computes this standard error, with or without `ybar`, from the influence functions in Appendix A.5 and applies the one-cluster check of Section 3.2. The variance is clustered on the cluster variable of the fitted model, or on panels after `vce(oim)`, `vce(bootstrap)`, or `vce(jackknife)`. Its intervals use normal critical values unless the fit has posted residual degrees of freedom (after the `small` option or a jackknife), in which case they use $t$ critical values on the number of clusters minus one.

### 2.8 Weights

With weights $a_{it}$ the criterion becomes, up to a constant, $\sum_i [\, A_i \, \hat{\alpha}_i(\boldsymbol{\beta}) + \sum_t a_{it} \mathbf{x}_{it}' \boldsymbol{\beta} \,]$, where $A_i = \sum_t a_{it}$ and $\hat{\alpha}_i(\boldsymbol{\beta})$ is now the logarithm of the weighted mean of $y_{it} \exp(-\mathbf{x}_{it}'\boldsymbol{\beta} - o_{it})$. Sampling weights may vary within panel and enter the sandwich through the weighted panel scores; importance weights are treated in the same way. Frequency weights must be constant within panel and represent that many identical panels. The resulting estimates and analytic standard errors match those from expanded data in which each copy receives its own panel identifier and, when clusters nest panels, retains the original panel's cluster. There is one exception: the check for terms identified by one cluster (Section 3.2) counts all copies of a panel once, whereas expanded data count each copy as a separate cluster. After `vce(jackknife)` the covariance matrix and its degrees of freedom also equal those of the expanded data, although `e(N_clust)` counts the panels as held in the data. The command computes the share likelihood and shape estimates only when the weights are constant within panel.

When the weights are constant within panel, the conditions of Section 2.3 carry over if share centering holds conditional also on the weights, with the moment conditions of

Appendix A.1 stated for the regressors multiplied by the panel's total weight. When they vary within panel, however, the shares in the score become $a_{it}\varepsilon_{it}/\sum_s a_{is}\varepsilon_{is}$, and exchangeability alone no longer guarantees centering at $a_{it}/A_i$, so a weight pattern that aligns with the regressors can bias the coefficients even under the gamma model. With gamma errors of shape 2 and weights that rise from 1 to 4 over four periods, for example, the coefficient on a treatment that switches on in the last period (true value 0.5) is biased by $-0.085$, and by $-0.17$ with two periods and weights of 1 and 2. Period indicators remove that bias for the other identified slopes when the weights follow the same deterministic pattern in every panel and the errors are independent of the regressors, by the argument of Proposition 2. They need not do so when the pattern differs across panels in a way related to the regressors. Fixed-effects Poisson avoids this problem because its score is linear in the outcome, provided the conditional mean remains correctly specified given the weights. The command prints a note when the weights vary within panel.

## 3. The `xtgamfe` command

### 3.1 Syntax

The data must be `xtset`. The syntax is

```
xtgamfe depvar indepvars [if] [in] [weight] [, options]
```

where `indepvars` may contain factor variables and time-series operators, and `fweights`, `pweights`, and `iweights` are allowed. There is no constant, because it is absorbed by the fixed effects. Table 1 lists the options.

A term is omitted in two cases: when its stored values are constant within every panel, and when, after the regressors are centered within panels, it is collinear with earlier terms. An omitted term is noted, displayed as omitted, and listed in `e(omitted)`. A regressor whose variation within panels is small next to its magnitude is kept, with a note on numerical precision. With a time-invariant factor `g`, for example, `c.x##i.g` estimates the coefficient on `x` and the interactions and omits the main effects of `g`, as `xtpoisson, fe` does.

A single selected level of a factor, such as `2.time`, or a selected list of levels, codes every other value of the factor as zero, and `predict` and `estat ape` treat that zero coding as the

base.

The observations selected by `if`, `in`, and the weights are those of the data as the user holds them, and the fit uses one realized set of weights, the values of a weight expression on the rows in the order in which the data are held. An expression whose values visibly depend on that order (through `_n`, a subscript, a running sum, or random numbers) is refused. The realized weights are protected thereafter: `vce(jackknife)` uses them as realized (`vce(bootstrap)` does not allow weights), and `predict`, `estat ape`, and `estat park` check them against a hash, stored at estimation, that binds each weight to its row. A weight expression may use functions such as `F()` (the F distribution function), but not time-series operators such as `F.`.

### 3.2 Estimation and inference

The command minimizes $Q^p$ by Newton's method with analytic derivatives and step halving. The regressors are centered within panels and scaled to unit within-panel root mean square before optimization, and the coefficients and their covariance matrix are mapped back afterward, so that the coefficients are equivariant to the units of the terms of the model and invariant to their origin. This concerns each term as it enters the model, not a variable from which a square or an interaction is built: shifting such a variable changes the model unless the lower-order terms are included. The offset enters through its deviations from a panel reference, so that adding a constant to it, or a different constant in each panel, leaves every result unchanged except the fixed effects, which move by the opposite amount, and the linear prediction, which includes the offset.

When the terms of the index are large (the sum of their absolute values and that of the offset exceeds 1,024 for some observation), fitted means, residuals, scores, and average partial effects are computed from the index centered within each panel, even when large terms cancel, as with two regressors that have large origins and coefficients of opposite sign. These statistics therefore do not depend on the origin of a term either; only the fixed effect and the linear prediction with its standard error, which are defined on the uncentered scale, change with it.

Convergence is declared when the Newton decrement, $(\mathbf{g}'\mathbf{H}^{-1}\mathbf{g})^{1/2}$, falls below the tolerance, or when a stalled line search has a predicted improvement below the numerical

**Table 1.** Options of `xtgamfe`.

| Option | Description |
|---|---|
| `offset(varname)` | offset $o_{it}$ in the log mean |
| `vce(robust)` | sandwich standard errors clustered on panels (the default) |
| `vce(cluster clustvar)` | sandwich standard errors clustered on `clustvar`, in which the panels must be nested |
| `vce(oim)` | observed information of the share likelihood, at the maximum likelihood shape |
| `vce(bootstrap)`, `vce(jackknife)` | resample or delete whole panels |
| `small` | $t$ and $F$ statistics on $M-1$ degrees of freedom, $M$ the number of clusters (with the default or clustered standard errors; not allowed with `vce(oim)` or `vce(bootstrap)`) |
| `nofe` | do not store `e(fe)` and `e(fe_id)` (each panel's fixed effect and identifier, stored by default with a note for up to 5,000 panels and within Stata's limit on matrix size), and store the command line without its `if` or `in` qualifier and its weight expression, for estimates saved or exported from restricted data. A prefix typed by the user stores its own command line, which is not stripped. `e(wexp)` and the factor levels in the coefficient names are still stored and should be checked before the estimates are saved or exported |
| `level(#)`, `eform`, display options | reporting; `eform` reports $\exp(\hat{\beta})$, which for a regressor that enters linearly and without interactions is the ratio of means for a one-unit change (for an interaction term, a ratio of ratios) |
| `tolerance(#)`, `iterate(#)`, `nolog`, `from(matname)` | optimization (by default, a tolerance of $10^{-6}$ for the Newton decrement and at most 1,000 iterations) |

resolution of the objective. The command reports which rule was met. The decrement, unlike the gradient norm, does not depend on the units of the regressors. A fit that does not converge stops with an error (return code 430) and posts no estimates, so that `e()` keeps the results that were there before. The criterion and its derivatives are computed for all panels at once, with no indicator variables, so that, for a given number of regressors, the cost of an iteration is linear in the number of observations.

Shares that are equal within every panel to numerical precision are treated as an exact fit (Section 2.6). The command then reports both shapes as missing, with `e(kappa_ml_status)` set to `infinite_precision`, and withholds the standard error of the shape, the log likelihood and the information criteria, the Wald test of all coefficients, and `vce(oim)`. The coefficient table shows the coefficients without standard errors, tests, or intervals. Weights that vary within panel leave the shapes missing for their own reason (Section 2.8), and `e(kappa_ml_status)` is then `weights_vary_within_panel`.

The default covariance matrix is the sandwich

$$\widehat{\mathbf{V}} = \frac{M}{M-1}\,\mathbf{H}^{-1}\Big(\sum_{c=1}^{M}\mathbf{g}_c\mathbf{g}_c'\Big)\mathbf{H}^{-1},$$

where $\mathbf{g}_c$ is the sum of the panel scores in cluster $c$ and $M$ is the number of clusters (panels by default, or the levels of a variable in which the panels are nested, such as households or sampling units). With frequency weights and standard errors clustered on panels, each copy of a panel counts as a cluster, as in the expanded data. The command subtracts the mean of the $\mathbf{g}_c$ before forming their outer products. That mean is zero at an exact minimum, so the centering changes the result only by the rounding error of the optimization.

The rank of this covariance matrix cannot exceed $M-1$. When the rank is below the number of coefficients, the command does not report the Wald test of all coefficients, because a quadratic form in a generalized inverse of a singular matrix tests only some of them, and which ones depends on the units of the regressors (`xtreg` with clustered standard errors does the same).

When a term varies within panels in only one cluster, the clusters cannot measure its sampling variability, whatever the rank of the covariance matrix (with clusters that nest the

panels the computed rank can stay full). The command then treats the term as follows:

- The coefficient table leaves the term's standard error, test, and interval blank, with a note, and the Wald test of all coefficients is not reported, whereas the standard errors of the other coefficients are not affected.
- `estat ape` shows the effects of the variables in that term without standard errors and says that those of the other effects leave out its sampling variation, which reaches them only through the fitted means of the panels in which the term varies (and not at all with `ybar`).
- `estat fepoisson` compares the term without resampling its difference and tests the other differences jointly. It does not leave out the replications that do not draw that cluster, in which the term drops out of both fits, so that the bootstrap of the other differences is not conditional on drawing it; a replication is left out only if it fails, as when another compared coefficient cannot be estimated in it. The term's own standard error is shown only after `vce(oim)`, which does not use the clusters, and with the panels as clusters its Poisson standard error is not shown either. When every term compared rests on one cluster, it shows the comparison without a bootstrap.
- `predict, stdp` is missing where the term is not zero.

A coefficient can also rest on one cluster even when every term varies in many clusters. Suppose that $r$ varies within panels in one cluster only. Then $x_1$ and $x_1 + r$ are collinear within panels once that cluster is left out, so that the other clusters identify neither coefficient, only their sum. The command therefore also examines every cluster whose leverage, the sum over its observations of the diagonal elements of the hat matrix of the within-panel design, is at least one half; the leverages of all clusters sum to the number of estimated coefficients, and a cluster on which a direction rests has a leverage of at least one. For each cluster examined it finds the directions of the coefficients that the other clusters do not identify (by the collinearity rule of the fit), and it treats every coefficient with a part along such a direction as it treats a term that varies in one cluster. `e(onecl)` names these terms, and a note under the table says that tables made from `e(V)` would show a standard error for them. A combination with no part along those directions, such as the sum of the coefficients of

$x_1 + r$ and $x_1 - r$, keeps a meaningful standard error.

`estat ape` and `estat fepoisson` make this check for the clusters they use, whatever the variance of the fit. After `vce(oim)` (whose fit makes no such check and leaves `e(onecl)` empty), `vce(bootstrap)`, or `vce(jackknife)`, both commands quietly refit the model from the stored estimates with standard errors clustered on those clusters to identify the affected terms, leaving the data and results in memory unchanged. `estat ape` does the same after a bootstrap or jackknife prefix, after which `estat fepoisson` is not available.

The command warns when there are fewer than 30 clusters. In the simulations of Section 4.6, intervals with $t$ critical values on $M - 1$ degrees of freedom, or from `vce(jackknife)`, then came closer to the nominal coverage. The `small` option posts those degrees of freedom, so that the coefficient table, `test`, and `lincom` use $t$ and $F$ statistics, and `estat ape` and `estat fepoisson` use $t$ critical values.

`vce(jackknife)` deletes one panel at a time; when the clusters nest panels, the `jackknife` prefix with its `cluster()` option deletes one cluster at a time instead. Only the coefficients are resampled, and the header says so for the standard error of the shape; `jackknife _b kappa=e(kappa_ml), cluster(clustvar):` resamples the shape as well, but `predict` and the `estat` commands are then not available, because the stored results are those of the prefix, and the model is refitted for them. After a resampling prefix whose clusters cross panels, `estat ape`, `estat park`, and `estat fepoisson` stop with an error, because they would cluster their standard errors on the prefix's clusters, which do not nest the panels.

`vce(oim)` reports $(\hat{\kappa}\mathbf{H})^{-1}$ with the maximum likelihood shape, the inverse information of the share likelihood. It is valid when panels are independent and the share likelihood of Proposition 3 correctly specifies the distribution of the within-panel shares, as it does under the gamma model; only under the gamma model, however, is $\kappa$ the shape of each outcome's own distribution. It is unavailable after an exact fit (Section 2.6) and with sampling or importance weights, and `vce(bootstrap)` is unavailable with weights of any kind.

### 3.3 Postestimation

`predict` returns six statistics: the linear prediction `xb` (the default, $\mathbf{x}_{it}'\hat{\boldsymbol{\beta}} + o_{it}$, without the fixed effect), its standard error `stdp`, the fixed effect `fe`, the fitted mean `mu`, the residual

`residuals` ($y_{it} - \hat{\mu}_{it}$), and the score `score` ($y_{it}/\hat{\mu}_{it} - 1$). Here `stdp` reflects coefficient uncertainty only and is missing where the prediction depends on a coefficient resting on one cluster, while `score` is the negative of the scalar term in $\mathbf{g}_i$ (Section 2.2) and is returned for the estimation sample only. The fixed effect belongs to the panel: it is computed from the estimation sample and returned for every observation of a panel that was in it. The fitted mean and the residual are returned for those observations as well, except where the estimation sample does not identify the panel's fitted mean at the observation's covariate values (a factor level absent from the sample, an empty cell of an interaction, or values that break a relation by which a term was omitted). There the prediction is missing, with a note. With time-series operators, the `xtset` delta must be that of the fit. As Section 2.6 shows, the fitted mean of a single panel is better suited to averages over many panels than to a prediction for one person.

Because the fixed effects are recomputed from the data, these statistics require that the data be unchanged since estimation, which the command verifies. After `estimates use`, which does not restore the estimation sample, `fe`, `mu`, `residuals`, and `score` are therefore not available, nor is `estat ape`; `xb`, `stdp`, and `mu` or `residuals` with `alpha()` are. The exception is a term omitted by a relation that differs between panels, such as a regressor constant within panels: `alpha()` then also needs the estimation sample, and an observation of a panel that has no observations in the estimation sample is missing, with a note, even with `alpha()`, because its fitted mean would depend on which term of the relation was omitted. Without such a relation, `alpha()` gives fitted means in new panels. The option `alpha(varname)` takes the fixed effects from a variable instead, without verifying that the data are unchanged, so that the user is responsible for them. It gives fitted means at general changes in the regressors (save the fixed effects with `predict, fe`, change the regressors, and type `predict` with `mu` and `alpha()`), but because it adds the supplied fixed effect to the uncentered index and offset, it can lose precision when these are very large. For a one-unit increase in a regressor that enters linearly, such as `x1` in the example of Section 3.4, the fitted means are simply multiplied by $\exp(\hat{\beta}_j)$, which needs no change to the data:

```
predict double mu0, mu
generate double mu_plus = mu0*exp(_b[x1]) if e(sample)
```

`estat ape` reports the average partial effects of Section 2.7 with their standard errors, as averages over the observations of the estimation sample; effects averaged over a subpopulation (with `if` or `in`, or by group) are not available. By default it computes each effect from the fitted means, and three options change what is averaged:

- with `ybar` it computes each effect from the observed outcome in place of the fitted mean, which is the consistent choice when period indicators may be absorbing a period pattern in the errors (Section 2.7);
- with `finite` it reports instead, for each continuous variable, the change in the average fitted mean when the variable rises by one unit (or by `delta(#)`), with its standard error (Appendix A.5 gives the case of a regressor that enters linearly); and
- with `finite` and `ybar` together, the observed outcome replaces the fitted mean: the effect averages the outcome times the proportional change in the mean, which for a regressor that enters linearly is $\bar{y}\,(e^{\hat{\beta}_j h} - 1)$ for a rise of $h$ units.

It checks, for each variable and each factor level, that the change is identified by the within-panel design, and reports none, saying so, when it is not, as for a variable that does not vary within panels or is a copy of another variable under a different name. The same term entered twice, as in `x1 x2 x1` or `x1` beside `i.k##c.x1`, is one variable, whose second term the fit omits, and its effect is reported. When every cluster contributes the same to an effect's influence function, as with identical panels, the variance of the effect is zero, and the effect is reported without a standard error, with a note. It does not report effects for variables that enter the model with a time-series operator, including such a variable's term without the operator, although their coefficients can be well identified. Without a list of variables, it skips, with a note, a variable that enters the model both as a factor and as a continuous variable.

After a model with a term that has a level of the time variable of `xtset`, alone or in an interaction with other factor variables, the default table ends with a note that points to `ybar`. Indicators of another period variable, such as calendar year, raise the same question but get no note. When levels of the time variable enter only in interactions with continuous variables, which are not period indicators, the note says instead that an effect computed with `ybar` is

consistent when the coefficients that enter the change are, and one computed from the fitted means when every coefficient is.

`estat ic` reports information criteria based on the share likelihood, which compare only `xtgamfe` models of the same within-panel shares (the same outcome, panel variable, estimation sample, and weights) and not models of the outcome itself. The Bayesian criterion uses the number of observations unless `n()` gives another number or an expression, such as `n(e(N_g))` for the number of panels (the independent contributions to the unweighted likelihood), and with sampling or importance weights the likelihood is a weighted pseudo-likelihood.

`estat fepoisson` fits fixed-effects Poisson regression to the same observations, with the terms, panel variable, and time settings of the fit, and tests the difference between the two coefficient vectors by a panel (or cluster) bootstrap. It does not allow weights. Its refits use the fit's tolerance and iteration cap and start from the default values, so a fit capped by a small `iterate()`, as one started with `from()` can be, should be fitted again without the cap first. Each difference is tested by a $z$ test ($t$ after `small` or `vce(jackknife)`). When the bootstrap covariance matrix has full rank, the differences are also tested jointly: with $D$ the Wald statistic of the $k$ differences and $B$ the number of bootstrap replications, $(B-k)D/\{k(B-1)\}$ is referred to $F$ with $k$ and $B-k$ degrees of freedom, a Hotelling-type $F$ approximation. That law is exact for a normal difference with an independent Wishart estimate of its covariance; here it allows for estimating the covariance from a finite number of resamples. A term whose coefficient rests on one cluster is compared without a bootstrap standard error and left out of the joint test. Replications that do not draw its cluster are not left out for that reason (the term is left out of both refits), so that the bootstrap of the other differences is unconditional. A replication in which another compared coefficient cannot be estimated is left out, as any failed replication is. When every term compared rests on one cluster, nothing is resampled: the differences are shown without standard errors, and neither a bootstrap nor a joint test is run. With fewer than 199 replications used, a note says that the standard errors and tests rest on few draws. The difference is bootstrapped because the command offers no analytic variance from the stacked scores of the two estimators; with 999 replications this took about a minute on one processor on the example of Section 3.4 (five differences), much longer than the fit, and a weighted analysis cannot use it. Section 4.4

reports the size and power of these tests in simulations.

`estat park` computes the Park regression of Section 6: an ordinary least-squares regression of the log squared residuals on the log fitted means, with standard errors clustered as in the fit and a nominal $t$ test of slope 2 shown in the table row. Because the regression treats the fitted means and residuals as data, the test's calibration depends on the design (Section 6). When the weights vary within panel or panels have different numbers of observations in the estimation sample, the slope can differ from 2 even under the gamma model; in those cases `estat park` reports the slope, standard error, and confidence interval, but withholds the test. With frequency weights and standard errors clustered on panels it, too, counts each copy of a panel as a cluster. It is not available after an exact fit, and it stops when the log fitted mean does not vary over the observations it uses.

`test`, `lincom`, `contrast`, and `pwcompare` work as after official estimators; one that involves a coefficient that rests on one cluster is not meaningful, because `e(V)` keeps the computed values for it, unless it has no part along the directions that rest on the cluster. `estat ape` refuses an interaction (`c.x1#c.x2`), a lag or lead (`L.x1`), or a prefix that is not an operator, and lists the variables that can be requested; after an exact fit, a note says that its standard errors reflect only the sampling of the panels over which the effects are averaged. `margins` is not supported because the fitted mean depends on fixed effects that the command recomputes from the data, whereas the default prediction, which `margins` would use, is the linear prediction without them. Santos Silva and Kemp (2016, slide 17) ask whether `margins` should be disabled after `xtlogit` and `xtpoisson` when the `fe` option is used. Instead, `estat ape` reports the effects with standard errors that allow for the estimation of the fixed effects.

The coefficients and their covariance matrix are stored in `e(b)` and `e(V)`, the shape estimates in `e(kappa)` and `e(kappa_ml)` (with `e(kappa_ml_se)` and `e(kappa_ml_status)`), and the fixed effects and the panel identifiers in `e(fe)` and `e(fe_id)`; Table 1 gives the limits on the last two and what `nofe` leaves out of a saved fit. `estat ape` stores its effects in `r(b)` and `r(V)`.

### 3.4 Example

The package includes a simulated dataset of 200 panels observed for four periods, with true coefficients of 0.5 and $-0.3$ and a gamma shape of 2 (it is an ancillary file, which `ssc install xtgamfe, all` copies to the working directory, where `sysuse` finds it). The output shown is from StataNow 19.5; earlier releases of Stata label some columns differently.

```
. sysuse xtgamfe_example, clear
. xtset id time
. xtgamfe cost x1 x2 i.time
Iteration 1: Q =      1552.23, |grad| =      .039419, decrement =    .00133169
Iteration 2: Q =      1552.23, |grad| =  8.14954e-07, decrement =  2.80512e-08
note: e(fe), e(fe_id): one row per panel with its identifier; nofe omits them

Fixed-effects gamma regression (log link)       Number of obs     =        800
Group variable: id                              Number of groups  =        200

                                                Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4

Gamma shape kappa (fixed-T moment):    2.2690   dispersion 1/kappa:    0.4407
Gamma shape kappa (marginal ML):       2.1998   std. err.:             0.1164
Wald chi2(5) =    364.94      Prob > chi2 = 0.0000
Convergence: yes (newton_decrement)       Iterations: 2
Gradient norm:   8.15e-07       Hessian rank: 5       VCE rank: 5


                                  (Std. err. adjusted for 200 clusters in id)
------------------------------------------------------------------------------
             |               Robust
        cost | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          x1 |   .4991795   .0344154    14.50   0.000     .4317267    .5666324
          x2 |  -.2679557    .029299    -9.15   0.000    -.3253806   -.2105307
             |
        time |
          2  |   .0313224   .0722057     0.43   0.664    -.1101982     .172843
          3  |  -.0658478   .0673452    -0.98   0.328    -.1978419    .0661464
          4  |  -.0651344   .0759175    -0.86   0.391    -.2139299    .0836611
------------------------------------------------------------------------------


. estat ape x1 x2, ybar

Average partial effects on the conditional mean       Number of obs =       800
Model: exp(alpha_i + x'b), observed outcomes in place of fitted means
                                  (Std. err. adjusted for 200 clusters in id)
------------------------------------------------------------------------------
             |      dy/dx   Std. err.      z    P>|z|     [95% conf. interval]
```

```
-------------+----------------------------------------------------------------
          x1 |   6.499434   .8385758     7.75   0.000     4.855856    8.143012
          x2 |  -3.488846   .5888958    -5.92   0.000     -4.64306   -2.334631
-------------------------------------------------------------------------------
Note: each effect averages the observed outcome times the change in
      exp(x'b): for a regressor that enters linearly, b times the mean
      outcome, 13.020234.


. estat ape x1 x2

Average partial effects on the conditional mean       Number of obs =       800
Model: exp(alpha_i + x'b), fixed effects as estimated
                                  (Std. err. adjusted for 200 clusters in id)
-------------------------------------------------------------------------------
             |      dy/dx   Std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
          x1 |   6.352671    .801827     7.92   0.000     4.781119    7.924223
          x2 |  -3.410064   .5590915    -6.10   0.000    -4.505864   -2.314265
-------------------------------------------------------------------------------
Note: the model includes indicators of time; if they absorb a period pattern
      in the errors, these effects are not consistent.  Option ybar averages
      the outcome times the implied change, which is consistent when the
      coefficients that enter the change are consistent (see help xtgamfe).
```

The note after the iteration log says that `e(fe)` and `e(fe_id)` hold one row per panel with its identifier (Table 1). The header reports the two shape estimates (2.27 and 2.20, against a true value of 2), the Wald test, and the convergence criterion that was met. Period indicators are included for the reason given in Section 2.3. Because they may absorb a period pattern in the errors, the average partial effects to report are those from `estat ape, ybar`, which uses the observed outcome (Section 6). For `x1`, that effect is 6.50 outcome units per unit of `x1` (standard error 0.84), equal to the coefficient 0.499 times the mean outcome of 13.0. The default `estat ape` computes the effects from the fitted means, 6.35 (0.80) for `x1`, the coefficient times the average fitted mean of 12.7, and its note points to the `ybar` option, because effects computed from the fitted means are not consistent if the indicators absorb a period pattern in the errors (Section 2.7). The errors of this example are exchangeable, so that the two are consistent for the same effect. They differ in the sample because the fitted means of a panel match its outcomes in ratio ($\sum_t y_{it}/\hat{\mu}_{it} = T_i$), not in sum.

The same fit supports the two checks of Section 6. `estat fepoisson, reps(999) seed(20261117)` compares the coefficients with those of fixed-effects Poisson on the same observations: of the five differences, that of `x2` is 0.109 (bootstrap standard error 0.052,

$p = 0.036$, or 0.033 with seed 12345), the other four have $p$-values of 0.24 to 0.92, and the joint test gives $F(5, 994) = 1.77$ ($p = 0.12$). The errors of this example are independent gamma variables with shape 2, so that both estimators are consistent here; the one difference below the 5 percent level is consistent with the size of these tests reported in Section 4.4, and illustrates why a single $p$-value between 0.01 and 0.05 is weak evidence. `estat park` gives a slope of 2.04 (standard error 0.07; test of slope 2, $p = 0.63$).

## 4. Validation and Monte Carlo evidence

The simulations in this section were run on StataNow 19.5 for Linux with pre-release builds of version 1.1.1 of the command, and the timings with its released files; one check of Section 4.1 was also repeated in Stata 15.1 for Windows. Rerun with the released files, the first two replications of every cell of every simulation batch (1,174 replications in 276 batches) reproduce every stored statistic of `xtgamfe` and its postestimation commands bit for bit, except in the designs with sampling weights or with clusters of panels, where the stored coefficients are bit-identical and the average partial effects and their standard errors differ by at most $9 \times 10^{-15}$ relative. The tables are generated from the simulation output by script.

### 4.1 Agreement with the other routes

The author compared `xtgamfe` with each of the routes of Section 2.5. On the example data and on a set of designs that includes unbalanced panels, factor variables, and an offset, the coefficients and clustered standard errors agree with those of `glm` with panel indicators fit to the same observations, and with those of the auxiliary Poisson regression once its coefficients are sign-reversed and its covariance matrix is multiplied by $G/(G-1)$, to $10^{-6}$ or better. In the 10,000 replications of the two designs of Section 4.2 in which both `xtgamfe` and `glm` with indicators were fit, the largest difference in a coefficient was $1.1 \times 10^{-7}$ and no confidence interval differed in whether it covered the true value. The command's other features were checked in the same way (frequency weights against the expanded data, sampling weights and clusters that nest panels against `glm`, lagged regressors against the auxiliary regression, and `vce(oim)` against the information that `glm` reports with a unit scale, divided by the maximum likelihood shape). On the example data, the auxiliary regression fit with `ppmlhdfe`

and `vce(cluster)` matches the sign-reversed coefficients and clustered standard errors of `xtgamfe` with no correction factor, whether clustering on panels or above them and with sampling weights that vary within panel. The relative difference is below $10^{-6}$ at default tolerances and below $10^{-8}$ when both tolerances are tightened. On the same data, `predict, mu` after `glm` with indicators reproduces the fitted means of `xtgamfe` (largest relative difference $1.2 \times 10^{-6}$), and `margins, dydx(x1 x2) vce(unconditional)` its average partial effects of 6.352671 and −3.410064, with standard errors of 0.801832 and 0.559097 against 0.801827 and 0.559092 from `estat ape` (relative differences below $10^{-5}$).

A comparison with `xtspj` used 600 datasets from the gamma design of Section 4.2, with 500 panels of four periods and shapes of 0.5, 1, and 2 (200 datasets each), in which the outcome is of order one (its mean is about 2.5). In every dataset the coefficient on $x_1$ from `xtgamfe` agreed with the sign-reversed one from the auxiliary Poisson regression to within $10^{-7}$. Table 2 counts the fits of `xtspj` (the release on SSC dated 18 May 2019, run in StataNow 19.5 for Linux) with `model(exponential)` and `method(none)` that reach the `xtgamfe` estimate, meaning a coefficient on $x_1$ within $10^{-4}$ of it.

**Table 2.** Fits of `xtspj` in simulated datasets with outcomes of order one: number of datasets in which the fit reaches the `xtgamfe` estimate of the coefficient on $x_1$ (to within $10^{-4}$), reports convergence at another coefficient, or stops with a convergence error. Exponential model with `method(none)` unless stated.

| `xtspj` setting | Datasets | Reaches the `xtgamfe` estimate of $x_1$ | Reports convergence elsewhere | Stops with an error |
|---|---|---|---|---|
| Default settings | 600 | 0 | 590 | 10 |
| Tolerances $10^{-9}$, 1,000 iterations | 600 | 384 | 206 | 10 |
| Default settings, outcome × 1,000 | 600 | 599 | 1 | 0 |
| Tolerances $10^{-9}$, outcome × 1,000 | 600 | 600 | 0 | 0 |
| Tolerances $10^{-9}$, outcome / 1,000 | 120 | 0 | 0 | 120 |
| Tolerances $10^{-9}$, `diagnosis` | 120 | 57 | 0 | 63 |
| `model(gamma)`, tolerances $10^{-9}$ | 600 | 190 | 0 | 410 |

*Notes:* Default settings are tolerances of $10^{-4}$ and at most 100 iterations; the rows with tolerances of $10^{-9}$ also allow 1,000 iterations. The rows with the outcome divided by 1,000 and with `diagnosis` (step halving when an iteration does not improve the objective) use the first 40 datasets of each shape. `model(gamma)` estimates the shape jointly with the fixed effects; all 190 of its fits that reach the estimate are at the shape of 0.5, and at shapes of 1 and 2 every fit stops with an error.

With its default settings `xtspj` never reached the estimate, reporting convergence at a coefficient on $x_1$ 0.18 to 0.40 below it (true value 0.5) in 590 datasets. The 206 fits with the

tightened tolerances that reported convergence at a wrong coefficient (97, 64, and 45 at the three shapes) all had a lower profiled log likelihood at their estimate than at that of `xtgamfe`, by 20 to 99 (it is $-Q^p$ up to a constant, and concave). In this design, a fit that stopped at a wrong coefficient usually showed two symptoms that a user can check: a reported standard error below $2 \times 10^{-5}$, against 0.03 to 0.04 when it reached the estimate, and a reported log likelihood below $-10^{10}$. All 206 of those fits with the tightened tolerances showed both, and 582 of the 590 with the default settings. The symptoms are not a general test, however. On the example data of Section 3.4, in a model without the period indicators, `xtspj` with its default settings reaches the estimate of `xtgamfe` (0.497) with the outcome in its own units. With the outcome multiplied by 1,000 it returns a coefficient on $x_1$ of 0.00005, with a standard error of 0.0008 and a log likelihood of $-3.8 \times 10^6$, which show neither symptom; with the outcome multiplied by 0.01 it returns 4.89.

Rescaling the outcome changes the criterion only by a constant and leaves the estimates unchanged, so that the differences between the rows of Table 2 are properties of the optimizer. Its stopping rules are a tolerance on the change in the parameters and a relative tolerance on the log likelihood: iterations end when the largest change in a parameter falls below `ptol()`, or when the change in the log likelihood, divided by its current absolute value, is nonnegative and below `ltol()` (Sun and Dhaene 2019, 339). At a log likelihood below $-10^{10}$, therefore, a nonnegative improvement of less than about $10^6$ can pass `ltol()` at its default of $10^{-4}$, which is one way in which a wrong solution can be reported as converged; the runs record which fits reported convergence, not which rule stopped them. The rule does not explain why the iterations break down at one scale of the outcome and not at another: in this design it is strictest at the original scale, where the absolute log likelihood is smallest and where the fits with the default settings fail. The `diagnosis` option took up to about four minutes per fit.

The check was repeated in Stata 15.1 MP for Windows for the six rows other than the one with `diagnosis`. In five of them each dataset fell in the same column as in Table 2, and `xtgamfe` gave the same estimates (to within $10^{-16}$). In the sixth, with the tightened tolerances, 212 fits reported convergence at a wrong coefficient rather than 206, and 190 of the 600 datasets changed column (98 from the estimate to a wrong coefficient and 92 the other way), so that which fits fail with those tolerances changes with the Stata release or

platform. These counts describe this design and these settings; they are not an estimate of how often `xtspj` fails in applied work. Where the exponential model, with the tightened tolerances, reached the estimate, the ratio of its standard error for $x_1$ to the clustered one of `xtgamfe` averaged 0.709, 1.001, and 1.426 at the three shapes, against values of $\sqrt{\kappa}$ of 0.707, 1, and 1.414 (Appendix A.6).

### 4.2 Performance when the conditions hold

The first set of experiments uses designs in which share centering holds. The conditional mean is $\exp(\alpha_i + 0.5\,x_{1,it} - 0.3\,x_{2,it})$ with $\alpha_i \sim N(0,1)$. The regressor $x_1$ is correlated with the fixed effect ($x_{1,it} = 0.5\,\alpha_i + \sqrt{0.75}\,u_{it}$, with $u_{it}$ standard normal), so that estimators that ignore the fixed effect are biased, and $x_2$ is standard normal and independent of everything. The errors are independent over time and either gamma with shape 2 or lognormal with a log-scale standard deviation of one, normalized to mean one. Table 3 reports the coefficient on $x_1$ for 100 and 250 panels of 2, 4, and 8 periods, with 500 replications per cell.

The largest absolute bias is 0.005, the average standard error is within 8 percent of the standard deviation of the estimates in every cell, and coverage lies between 0.92 and 0.97, under the lognormal errors as well as under the gamma. No replication failed to converge. (A pooled gamma regression that ignores the fixed effects has a bias of about 0.50.)

The ranges in this paragraph and the last cover all 20 cells of the experiment. Log-linear fixed effects is the closest competitor: its RMSE is 1 to 18 percent larger than that of `xtgamfe` under gamma errors and 1 to 16 percent smaller under lognormal errors, for which it is the efficient estimator. Because the errors here are independent of the regressors, the two estimate the same coefficients. Fixed-effects Poisson, in contrast, has an RMSE 1.7 to 2.7 times that of `xtgamfe`, and its confidence intervals, formed with its cluster-robust standard errors, cover the true value too rarely (0.85 to 0.94). Section 4.4 shows that both features come from the heterogeneity in levels across panels that these designs contain. Their errors are independent and have the same distribution in every observation, so that the variance is proportional to the squared mean. A heavy-tailed variant, in which one error in ten is drawn from a lognormal distribution with a much larger variance, gives the same ordering with 1,000 and 5,000 panels (200 replications per cell): coverage for `xtgamfe` of 0.93 to 0.96,

an RMSE for log-linear fixed effects 1.1 to 1.2 times as large, and one for Poisson 2.6 to 3.9 times as large.

A correlated-random-effects gamma regression (Mundlak 1978) (`meglm` with a normal random intercept on the log scale, the panel means of $x_1$ and $x_2$ as further regressors, and standard errors clustered on panels) did as well as `xtgamfe` in all 20 cells, with bias between $-0.005$ and $+0.005$, coverage between 0.93 and 0.97, and an RMSE 0.94 to 1.04 times that of `xtgamfe`. It did as well when $x_1$ depended on the fixed effect nonlinearly (RMSE 0.93 to 1.00 times that of `xtgamfe`). In both designs the fixed effect depends on the regressors only through the panel mean of $x_1$, which the regression includes, although in the second its distribution given that mean is not the normal one that the regression assumes. The fixed-effects estimator needs no assumption about how the fixed effect is related to the regressors, but no design was simulated in which features of the regressor history other than the panel mean predict the fixed effect.

**Table 3.** Coefficient on $x_1$ (true value 0.5) when the errors are independent and identically distributed: `xtgamfe`, with the root-mean-squared error (RMSE) of two other estimators relative to it. 500 replications per cell.

| Errors | $G$ | $T$ | `xtgamfe`: bias | `xtgamfe`: SD | `xtgamfe`: mean SE | `xtgamfe`: coverage | FE Poisson: RMSE ratio | FE Poisson: coverage | Log-linear FE: RMSE ratio | Log-linear FE: coverage |
|---|---|---|---|---|---|---|---|---|---|---|
| Gamma | 100 | 2 | +0.002 | 0.096 | 0.093 | 0.93 | 1.73 | 0.85 | 1.01 | 0.93 |
| Gamma | 100 | 4 | −0.003 | 0.051 | 0.050 | 0.95 | 1.95 | 0.89 | 1.06 | 0.95 |
| Gamma | 100 | 8 | −0.001 | 0.032 | 0.032 | 0.94 | 2.14 | 0.90 | 1.11 | 0.94 |
| Gamma | 250 | 2 | −0.004 | 0.059 | 0.058 | 0.95 | 1.76 | 0.91 | 1.03 | 0.94 |
| Gamma | 250 | 4 | −0.001 | 0.032 | 0.032 | 0.96 | 2.06 | 0.92 | 1.07 | 0.95 |
| Gamma | 250 | 8 | −0.001 | 0.019 | 0.020 | 0.95 | 2.36 | 0.93 | 1.10 | 0.94 |
| Lognormal | 100 | 2 | −0.005 | 0.124 | 0.119 | 0.92 | 1.98 | 0.87 | 0.99 | 0.93 |
| Lognormal | 100 | 4 | +0.005 | 0.067 | 0.071 | 0.96 | 2.31 | 0.90 | 0.95 | 0.96 |
| Lognormal | 100 | 8 | 0.000 | 0.049 | 0.049 | 0.95 | 2.38 | 0.90 | 0.90 | 0.95 |
| Lognormal | 250 | 2 | −0.002 | 0.078 | 0.074 | 0.94 | 2.26 | 0.92 | 0.98 | 0.94 |
| Lognormal | 250 | 4 | +0.002 | 0.048 | 0.045 | 0.93 | 2.31 | 0.91 | 0.94 | 0.94 |
| Lognormal | 250 | 8 | 0.000 | 0.033 | 0.031 | 0.95 | 2.40 | 0.89 | 0.88 | 0.94 |

*Notes:* SD is the standard deviation of the estimates across replications and mean SE the average panel-clustered standard error. Coverage is of the nominal 95 percent interval; its Monte Carlo standard error near 0.95 is about 0.010. FE Poisson is `xtpoisson, fe vce(robust)` on the outcome and log-linear FE is `xtreg, fe vce(cluster)` on its logarithm. The FE Poisson standard errors carry no finite-sample factor, as `xtpoisson` reports them, whereas the covariance matrix of `xtgamfe` includes $M/(M-1)$. Every coverage rate uses normal critical values, including that of log-linear FE (`xtreg` reports $t$ critical values, which differ little with 100 or more panels). The ranges given in the text also cover panels of 3 and 16 periods, which the table omits.

### 4.3 Performance when the conditions fail

*Dispersion that depends on a regressor.* In the first failure design the errors are lognormal with a log-scale standard deviation of 0.9 when $x_1$ is positive and 0.3 otherwise, normalized to mean one, so that the conditional mean is unchanged and share centering fails. Table 4 reports the results of 1,000 replications.

**Table 4.** Coefficient on $x_1$ (true value 0.5) when the dispersion of the errors depends on $x_1$: bias, with coverage of the nominal 95 percent interval in parentheses. 1,000 replications per cell.

| $G$ | $T$ | `xtgamfe` | Log-linear FE | FE Poisson |
|---|---|---|---|---|
| 250 | 2 | −0.141 (0.17) | −0.142 (0.18) | −0.008 (0.89) |
| 250 | 4 | −0.084 (0.17) | −0.144 (0.00) | −0.006 (0.90) |
| 250 | 8 | −0.048 (0.33) | −0.144 (0.00) | −0.005 (0.91) |
| 1,000 | 2 | −0.142 (0.00) | −0.143 (0.00) | +0.001 (0.93) |
| 1,000 | 4 | −0.083 (0.00) | −0.143 (0.00) | −0.001 (0.94) |
| 1,000 | 8 | −0.046 (0.00) | −0.143 (0.00) | 0.000 (0.94) |

*Notes:* Bias is the mean of the estimates minus the true value, and coverage is the proportion of replications in which the nominal 95 percent interval contains the true value; the same holds in Tables 5 to 8.

`xtgamfe` is biased by −0.14 with two periods, and the bias falls as periods are added, to −0.08 with four periods and −0.05 with eight. Log-linear fixed effects has nearly the same bias with two periods (the two estimators minimize different symmetric loss functions of the same log differences when $T = 2$) and keeps it as periods are added, because it estimates effects on the mean of the logarithm, which here differ from effects on the mean. Fixed-effects Poisson is centered throughout. The estimates of `xtgamfe` nevertheless vary less than those of Poisson (standard deviations of 0.010 to 0.048, against 0.038 to 0.146), so that with 250 panels and four or eight periods its root-mean-squared error is the smaller of the two (0.089 and 0.052, against 0.096 and 0.068), although its intervals cover the true value in at most a third of replications. The cross-sectional advantage of the gamma model over the log-linear model under heteroskedasticity does not carry over to fixed effects with few periods: in this design, only Poisson is consistent.

*A variance that is not proportional to the squared mean.* The usual departure from the variance condition of Weidner and Zylkin is a variance proportional to another power of the mean, $\phi\,\mu_{it}^{p}$, under which the coefficient of variation of the error, proportional to $\mu_{it}^{(p-2)/2}$, changes with the regressors, so that share centering can fail, as in the first case of Section

2.3. In the design below it does. The gamma design of Section 4.2 was rerun with gamma outcomes of variance $0.5\,\mu_{it}^{p}$ for $p$ of 3, 2.5, 2, 1.5, and 1. The shape, $2\mu_{it}^{2-p}$, is 2, as in Table 3, at the median of $\mu_{it}$, and ranges between its 10th and 90th percentiles from 0.84 to 4.8 with $p$ of 1.5 or 2.5 and from 0.35 to 11 with $p$ of 1 or 3. Table 5 reports 1,000 panels.

**Table 5.** Coefficient on $x_1$ (true value 0.5) when the variance of the outcome is $0.5\,\mu^p$: bias, with coverage of the nominal 95 percent interval in parentheses, and the RMSE of FE Poisson relative to that of `xtgamfe`. 1,000 panels; 500 replications per cell.

| Variance | $T$ | `xtgamfe` | FE Poisson | Log-linear FE | CRE gamma | FE Poisson: RMSE ratio |
|---|---|---|---|---|---|---|
| $0.5\mu^3$ | 2 | −0.178 (0.00) | −0.021 (0.90) | −0.490 (0.02) | −0.079 (0.60) | 1.15 |
| $0.5\mu^3$ | 4 | −0.101 (0.00) | −0.008 (0.89) | −0.489 (0.00) | −0.055 (0.43) | 1.57 |
| $0.5\mu^3$ | 8 | −0.057 (0.04) | −0.006 (0.87) | −0.487 (0.00) | −0.035 (0.41) | 2.53 |
| $0.5\mu^{2.5}$ | 2 | −0.068 (0.43) | 0.000 (0.90) | −0.098 (0.29) | −0.044 (0.73) | 1.58 |
| $0.5\mu^{2.5}$ | 4 | −0.037 (0.42) | −0.006 (0.93) | −0.100 (0.01) | −0.027 (0.67) | 2.06 |
| $0.5\mu^{2.5}$ | 8 | −0.020 (0.60) | −0.002 (0.92) | −0.100 (0.00) | −0.016 (0.72) | 2.59 |
| $0.5\mu^2$ | 2 | −0.001 (0.94) | +0.002 (0.94) | −0.001 (0.94) | −0.001 (0.94) | 1.87 |
| $0.5\mu^2$ | 4 | −0.001 (0.96) | −0.005 (0.93) | −0.001 (0.96) | −0.001 (0.95) | 2.23 |
| $0.5\mu^2$ | 8 | −0.001 (0.97) | −0.002 (0.95) | −0.001 (0.96) | −0.001 (0.96) | 2.34 |
| $0.5\mu^{1.5}$ | 2 | +0.070 (0.42) | 0.000 (0.94) | +0.099 (0.27) | +0.037 (0.71) | 0.36 |
| $0.5\mu^{1.5}$ | 4 | +0.039 (0.40) | +0.001 (0.95) | +0.101 (0.00) | +0.026 (0.65) | 0.39 |
| $0.5\mu^{1.5}$ | 8 | +0.020 (0.60) | +0.001 (0.94) | +0.099 (0.00) | +0.014 (0.75) | 0.50 |
| $0.5\mu$ | 2 | +0.179 (0.00) | 0.000 (0.96) | +0.495 (0.02) | +0.017 (0.89) | 0.10 |
| $0.5\mu$ | 4 | +0.103 (0.01) | 0.000 (0.94) | +0.497 (0.00) | +0.034 (0.59) | 0.09 |
| $0.5\mu$ | 8 | +0.056 (0.03) | 0.000 (0.94) | +0.487 (0.00) | +0.027 (0.50) | 0.11 |

*Notes:* The outcome is gamma with mean $\mu_{it}$ and shape $2\mu_{it}^{2-p}$, independent over panels and periods, with $\mu_{it}$ as in Table 3. CRE gamma is the correlated-random-effects regression of Section 4.2. With period indicators, the biases of `xtgamfe` differ from those shown by less than 0.001. The Monte Carlo standard error of a bias is at most 0.002 for `xtgamfe` and CRE gamma, 0.009 for FE Poisson, and 0.007 for log-linear FE; that of a coverage rate is at most 0.022. With $p = 1$, one, three, and four of the outcomes generated in the cells with two, four, and eight periods (each in a different replication), and with $p = 3$ one in each of those cells, were below $10^{-300}$ and were set to that value. In the runs cited in the text, one outcome with 250 panels ($p = 1$) and 13 and 12 with 100,000 panels ($p = 1$ and 3) were set to that value in the same way, and two of the 500 fits of the correlated-random-effects regression with $p = 1$, 250 panels, and two periods failed to converge.

With $p = 2$ the results are those of Section 4.2. With $p = 1.5$ the coefficient from `xtgamfe` is biased upward by 0.070, 0.039, and 0.020 with two, four, and eight periods, or 14, 8, and 4 percent of its value. Its intervals cover the true value in 40 to 60 percent of replications. With $p = 1$ the bias is 0.18, 0.10, and 0.06, and coverage is at most 3 percent. With $p$ above 2 the bias has the other sign and nearly the same size: −0.068, −0.037, and −0.020 with $p = 2.5$ and −0.18, −0.10, and −0.06 with $p = 3$, with coverage of 42 to 60 percent and of

at most 4 percent. The biases are about the same with 250 panels and with 100,000 (five replications per cell). With 16 periods, run with 100,000 panels, they are 0.009 and 0.031 with $p = 1.5$ and 1 and $-0.009$ and $-0.032$ with $p = 2.5$ and 3. This is an inconsistency with the number of periods fixed.

In this design (3) gives $\boldsymbol{\beta}^* - \boldsymbol{\beta}_0 \approx -(p-2)\,\phi\,\mathrm{E}[\mu_{it}^{p-2}]\,\boldsymbol{\beta}_0/T$, where $\phi\,\mathrm{E}[\mu_{it}^{p-2}]$ is the mean squared coefficient of variation of the error (Appendix A.7): to second order, a bias that falls in proportion to $1/T$ and is proportional to the coefficient itself. With $p$ of 1.5 and 2.5 that mean is 0.63, and (3) gives $\pm 0.079$, $\pm 0.039$, $\pm 0.020$, and $\pm 0.010$ with 2, 4, 8, and 16 periods, against simulated biases of 0.068 to 0.070, 0.037 to 0.039, 0.020, and 0.009 in absolute value (1,000 panels; 100,000 with 16 periods). With $p$ of 1 and 3 it is 1.25, and (3) has the right sign but is 1.8 times the simulated bias with two periods and about 1.25 times with sixteen: the simulated bias falls more slowly than $1/T$ ($T$ times it rises from about 0.36 to 0.50) and approaches (3) as $T$ grows. Much of that mean comes from a few observations (15 percent of them carry 63 percent of it), and the expected shares respond less than in proportion to their large variances.

In this design as specified (before the floor that the computation applies to the few outcomes noted under Table 5), the proportionality and the direction of the bias hold exactly, not only to second order. Given the fixed effect, the regressors are normal, independent across periods, and identically distributed, and the distribution of the outcome depends on them only through the index $\mathbf{x}_{it}'\boldsymbol{\beta}_0$. Once their within-period variation is standardized, the part of the regressors orthogonal to the index has mean zero and is independent of the index and the errors, so that at every multiple of $\boldsymbol{\beta}_0$ that part contributes nothing to the expected score. The expected score is therefore zero at the multiple that minimizes the population criterion along $\boldsymbol{\beta}_0$, and by strict convexity the limit $\boldsymbol{\beta}^*$ is that multiple: in the population, the bias of the coefficient on $x_2$ is $-0.60$ times that on $x_1$, as for the true coefficients. The direction follows from the shares: of independent gamma errors with mean one, the one with the larger shape has the larger expected share, for any number of periods (Appendix A.7). When $p < 2$, an observation with a larger mean has the larger shape (with $p = 1.5$ and two periods, the period with the higher $x_1$ has an expected share of 0.513), so that the derivative of the population criterion along $\boldsymbol{\beta}_0$ is negative at the true coefficients, the multiple exceeds

one, and both coefficients are overstated in absolute value. When $p > 2$ the sign reverses.

Period indicators do not help. The bias of fixed-effects Poisson is at most 0.006 in absolute value, and its coverage 0.91 to 0.96, in every cell with $p$ of 2 or less and 250 or 1,000 panels. It is also the more precise: the standard deviation of its estimates is 0.85 to 0.99 times that of `xtgamfe` with $p = 1.5$ and 0.43 to 0.47 times with $p = 1$, against 1.9 to 2.3 times with $p = 2$. With $p$ of 2.5 and 3 its bias is at most 0.021 in absolute value (Monte Carlo standard error 0.009) and its coverage 0.85 to 0.93. The standard deviation of its estimates, however, is 3.6 to 9.5 times that of `xtgamfe`, so that `xtgamfe`, although biased, has the smaller root-mean-squared error: that of Poisson is 1.1 to 3.6 times as large. Log-linear fixed effects and the correlated-random-effects regression are biased as well (Table 5).

*Serial correlation.* In the second failure design the conditional mean is $\exp(\alpha_i + 0.5\, d_{it} - 0.3\, x_{2,it})$, where $d_{it}$ is a treatment indicator, and the errors are a mean-one transformation of a stationary Gaussian first-order autoregressive process with correlation $\varrho$, independent of the regressors, with either a gamma margin of shape 2 or a lognormal margin of unit log variance. (At $\varrho$ of 0.4 and 0.8 the lag-one correlation of the errors themselves is 0.38 and 0.78 with the gamma margin and 0.29 and 0.71 with the lognormal.) Treatment status depends on $\alpha_i$, and treated panels switch on in the last period only (the "late" profile) or from the second period onward (the "early" profile). Each estimator was fit with and without period indicators, with 500 replications for 250 and 1,000 panels, 2, 4, and 8 periods, and $\varrho$ of 0, 0.4, and 0.8. Table 6 reports the late profile with 1,000 panels.

Without period indicators, `xtgamfe` is centered when $\varrho = 0$ (rows that Table 6 omits) and when $T = 2$, because the errors of two periods are exchangeable under any stationary process that is reversible in time, as the Gaussian autoregression used here is. In every cell of the late profile with $\varrho > 0$ and four or eight periods, however, the coefficient is biased upward, by 0.02 to 0.05 with the gamma margin and by 0.04 to 0.08 with the lognormal margin, or 5 to 17 percent of its value. The nominal 95 percent interval then covers the true value as rarely as 44 percent of the time. The early profile gives the mirror image (biases of about the same size and the opposite sign).

Moreover, going from four periods to eight does not help: at $\varrho = 0.8$ the bias nearly doubles. A direct calculation from the population estimating equation, by quasi-Monte Carlo

integration at every panel length from 2 to 48 (estimated integration standard error below 0.0001), is within about two Monte Carlo standard errors of each of the 12 biases of `xtgamfe` without period indicators in Table 6. At $\varrho = 0.8$ the bias is largest at about 13 periods with the gamma margin and about 15 with the lognormal (0.057 and 0.099), changes by less than 0.003 between 12 and 18 periods, and is still 0.032 and 0.067 with 48. At $\varrho = 0.4$ it is largest at five and six periods and declines slowly thereafter.

Approximation (3) accounts for this pattern. The error of the last period has neighbors on one side only, so that its $\omega_{iT}$ falls short of the panel average by $\sum_{k=1}^{T-1}(T-2k)\tau_k/T$, where $\tau_k$ is the autocovariance of the errors at lag $k$. For the late profile, the second-order approximation (3) gives a bias equal to that shortfall divided by $T-1$: zero with two periods, rising with $T$ while the panel is short relative to the range of the dependence, and, if the autocovariances are absolutely summable, close to $\sum_{k\geq 1}\tau_k/T$ in long panels. With the gamma margin it gives 0.026 at $\varrho = 0.4$ with both four and eight periods, 0.025 and 0.048 at $\varrho = 0.8$ with four and eight periods, and, at $\varrho = 0.8$, a maximum of 0.056 at 14 periods and 0.032 at 48, close to the values above. With the lognormal margin, whose squared coefficient of variation is $e - 1 \approx 1.7$, it rises and falls with $T$ in the same way but is 1.4 to 2.2 times the bias of the direct calculation with four and eight periods, and 1.1 and 1.3 times with 48 (with $\varrho$ of 0.4 and 0.8). The gamma margin satisfies the variance condition of Weidner and Zylkin, so that the failure here is one of dependence and not of the variance function.

Fixed-effects Poisson and log-linear fixed effects are consistent in these designs: across the 72 cells their mean signed biases are below 0.001 in absolute value and their mean absolute biases 0.005 and 0.002. The largest, 0.034 for Poisson in one cell with 250 panels, is 4.8 Monte Carlo standard errors from zero, a finite-sample bias, and none for log-linear fixed effects exceeds 0.01.

With period indicators the bias disappears, as Proposition 2 predicts: none of the 72 cell biases exceeds 2.1 Monte Carlo standard errors, and coverage is 0.92 to 0.98. The price is a standard deviation of the treatment coefficient 1.3 to 1.5 times as large in these designs, in which the treatment is itself closely related to the period.

*Dispersion that differs across periods.* In the third failure design the errors are independent gamma variates whose shape rises geometrically from 0.5 in the first period to 4 in the last,

**Table 6.** Treatment coefficient (true value 0.5) with serially correlated errors, late profile, 1,000 panels: bias, with coverage of the nominal 95 percent interval in parentheses. 500 replications per cell.

| Error margin | AR coefficient $\varrho$ | $T$ | `xtgamfe` | `xtgamfe`, period indicators | FE Poisson | Log-linear FE |
|---|---|---|---|---|---|---|
| Gamma | 0.4 | 2 | +0.003 (0.96) | +0.001 (0.96) | −0.002 (0.94) | +0.003 (0.96) |
| Gamma | 0.4 | 4 | +0.025 (0.90) | +0.001 (0.96) | +0.001 (0.96) | −0.002 (0.95) |
| Gamma | 0.4 | 8 | +0.025 (0.92) | −0.001 (0.95) | −0.004 (0.96) | −0.001 (0.95) |
| Gamma | 0.8 | 2 | +0.001 (0.94) | +0.003 (0.94) | 0.000 (0.94) | +0.001 (0.94) |
| Gamma | 0.8 | 4 | +0.027 (0.83) | 0.000 (0.94) | −0.002 (0.93) | 0.000 (0.96) |
| Gamma | 0.8 | 8 | +0.050 (0.60) | −0.001 (0.94) | −0.002 (0.94) | −0.001 (0.97) |
| Lognormal | 0.4 | 2 | 0.000 (0.95) | −0.003 (0.93) | −0.002 (0.92) | 0.000 (0.95) |
| Lognormal | 0.4 | 4 | +0.042 (0.86) | 0.000 (0.95) | +0.002 (0.95) | 0.000 (0.94) |
| Lognormal | 0.4 | 8 | +0.042 (0.88) | −0.002 (0.95) | −0.005 (0.93) | −0.002 (0.94) |
| Lognormal | 0.8 | 2 | −0.001 (0.94) | 0.000 (0.94) | −0.004 (0.94) | −0.001 (0.93) |
| Lognormal | 0.8 | 4 | +0.043 (0.76) | +0.001 (0.95) | +0.002 (0.92) | 0.000 (0.94) |
| Lognormal | 0.8 | 8 | +0.083 (0.44) | 0.000 (0.94) | −0.005 (0.94) | 0.000 (0.93) |

*Notes:* AR coefficient $\varrho$ is that of the stationary Gaussian first-order autoregression from which the errors are formed (Section 4.3). The FE Poisson and log-linear FE columns are fits without period indicators. The Monte Carlo standard error of a bias is 0.001 to 0.003 for `xtgamfe` and log-linear FE and 0.001 to 0.005 for FE Poisson.

with the same two treatment profiles and 1,000 panels (Table 7).

**Table 7.** Treatment coefficient (true value 0.5) when the dispersion of the errors differs across periods: bias, with coverage of the nominal 95 percent interval in parentheses. 1,000 panels; 500 replications per cell.

| Profile | $T$ | `xtgamfe` | `xtgamfe`, period indicators | FE Poisson | Log-linear FE | Log-linear FE, period indicators |
|---|---|---|---|---|---|---|
| Late | 4 | +0.209 (0.00) | +0.001 (0.95) | +0.005 (0.94) | +0.576 (0.00) | 0.000 (0.95) |
| Late | 8 | +0.084 (0.15) | −0.002 (0.95) | +0.001 (0.93) | +0.448 (0.00) | −0.001 (0.96) |
| Early | 4 | +0.348 (0.00) | +0.003 (0.94) | +0.004 (0.93) | +0.943 (0.00) | −0.003 (0.94) |
| Early | 8 | +0.157 (0.37) | 0.000 (0.95) | +0.002 (0.94) | +0.852 (0.00) | −0.007 (0.94) |

Without period indicators the treatment coefficient from `xtgamfe` is biased by +0.21 to +0.35 with four periods and by +0.08 to +0.16 with eight, and its intervals cover the true value in at most 2 of 500 replications with four periods and in 15 and 37 percent of them with eight. Approximation (3), with $\omega_{it} = 1/\kappa_t$, gives +0.23 and +0.09 for the late profile and +0.35 and +0.16 for the early profile with four and eight periods: the late profile switches on in the period whose error is the least dispersed, and the early profile is off in the period whose error is the most dispersed, so that both biases are positive. Log-linear fixed effects without

period indicators fares worse, with biases of +0.45 to +0.94, because the mean of the log error also differs across periods. With period indicators both are centered, as Proposition 2 predicts for the gamma estimator, and fixed-effects Poisson is centered with or without them.

*Unbalanced panels.* Two further designs examine the extensions of Proposition 2 to panels observed for different sets of periods: each of 2,000 panels is observed for a window of two to four consecutive periods placed at random within six calendar periods, with autoregressive errors as above. When treated panels switch on in a common calendar period (the fourth), no estimator shows a bias, with or without indicators (the largest bias of `xtgamfe` in any cell is 0.003). The switch then falls at a uniformly distributed position within the windows that span it, and with a stationary error process that is reversible in time the biases of switches late and early in a window cancel, so that this design does not test the extensions. In the second design the treatment is timed relative to the panel's own window, as when it follows enrollment by a fixed interval: it is on in the panel's last observed period only (late) or from its second observed period (early). Table 8 reports that design.

With $\varrho = 0.8$ the coefficient from `xtgamfe` is biased by about 0.02 in the direction of the profile, and coverage falls to between 0.81 and 0.84. Calendar-period indicators do not help (the bias is, if anything, larger, and coverage is 0.78 to 0.82), because the period means of the treatment now depend on the window and what matters is an observation's position within its own panel, not the calendar period. Indicators for that position, by panel length, remove the bias (it is at most 0.002, with coverage of 0.94 to 0.96), as the pooled form of the second extension of Proposition 2 implies under the two further assumptions of Section 2.3, which this design satisfies, at the price of a standard deviation about 1.4 times as large. Fixed-effects Poisson and log-linear fixed effects are centered without any indicators.

### 4.4 Level heterogeneity and the comparison with fixed-effects Poisson

The comparison with fixed-effects Poisson in Table 3 depends on heterogeneity in panel levels. The fixed effect is standard normal and $x_1$ loads on it, so that the panel-level mean is lognormal with a log standard deviation of about 1.25. Poisson weights each panel by its level, whereas the gamma criterion uses only within-panel shares, so that a few very large panels dominate the Poisson score and its cluster-robust variance. To separate this mechanism from

**Table 8.** Treatment coefficient (true value 0.5) with serially correlated errors in unbalanced panels, treatment timed relative to each panel's window: bias, with coverage of the nominal 95 percent interval in parentheses. 2,000 panels; 500 replications per cell.

| Error margin | Profile | AR coefficient $\varrho$ | `xtgamfe` | `xtgamfe`, calendar-period indicators | `xtgamfe`, position indicators | `xtgamfe`, both | FE Poisson | Log-linear FE |
|---|---|---|---|---|---|---|---|---|
| Gamma | Late | 0.0 | −0.001 (0.94) | 0.000 (0.96) | −0.002 (0.94) | −0.002 (0.94) | −0.001 (0.95) | −0.002 (0.94) |
| Gamma | Late | 0.8 | +0.016 (0.84) | +0.021 (0.82) | 0.000 (0.94) | 0.000 (0.94) | +0.001 (0.94) | 0.000 (0.94) |
| Gamma | Early | 0.0 | 0.000 (0.96) | +0.002 (0.95) | +0.002 (0.96) | +0.002 (0.95) | 0.000 (0.95) | 0.000 (0.96) |
| Gamma | Early | 0.8 | −0.017 (0.83) | −0.021 (0.82) | −0.001 (0.94) | −0.001 (0.94) | −0.002 (0.97) | −0.001 (0.95) |
| Lognormal | Late | 0.0 | −0.001 (0.96) | +0.001 (0.95) | +0.002 (0.94) | +0.002 (0.95) | −0.002 (0.93) | −0.002 (0.95) |
| Lognormal | Late | 0.8 | +0.024 (0.81) | +0.030 (0.78) | −0.002 (0.96) | −0.002 (0.96) | −0.002 (0.95) | −0.001 (0.95) |
| Lognormal | Early | 0.0 | 0.000 (0.95) | −0.001 (0.96) | −0.001 (0.96) | −0.001 (0.95) | −0.002 (0.94) | 0.000 (0.96) |
| Lognormal | Early | 0.8 | −0.024 (0.82) | −0.031 (0.78) | +0.001 (0.96) | +0.001 (0.96) | +0.002 (0.94) | +0.001 (0.95) |

*Notes:* AR coefficient $\varrho$ as in Table 6. Position indicators are indicators for the observation's position within its panel, separately by the number of periods for which the panel is observed (`i.len#i.pos`). FE Poisson and log-linear FE are fit without indicators. The Monte Carlo standard error of a bias is 0.001 to 0.003 for `xtgamfe` and log-linear FE and 0.001 to 0.004 for FE Poisson.

the error distribution, the gamma design was rerun with the direct loading of the fixed effect on the log mean set to 1, 0.5, and 0 (log standard deviations of the panel-level mean of 1.25, 0.75, and 0.25). Table 9 reports the results.

**Table 9.** Level heterogeneity and the comparison with fixed-effects Poisson, gamma errors with shape 2: RMSE of FE Poisson relative to that of `xtgamfe`, and the coverage of the nominal 95 percent interval of each, for the coefficient on $x_1$. 500 replications per cell.

| Panel-level s.d. of log mean | $G$ | RMSE ratio: $T=2$ | RMSE ratio: $T=4$ | RMSE ratio: $T=8$ | FE Poisson coverage: $T=2$ | FE Poisson coverage: $T=4$ | FE Poisson coverage: $T=8$ | `xtgamfe` coverage: $T=2$ | `xtgamfe` coverage: $T=4$ | `xtgamfe` coverage: $T=8$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.25 | 100 | 1.67 | 2.01 | 2.15 | 0.85 | 0.92 | 0.92 | 0.93 | 0.95 | 0.95 |
| 1.25 | 250 | 1.87 | 1.88 | 2.22 | 0.91 | 0.92 | 0.92 | 0.96 | 0.94 | 0.94 |
| 1.25 | 1,000 | 1.93 | 2.23 | 2.42 | 0.93 | 0.95 | 0.93 | 0.93 | 0.96 | 0.96 |
| 0.75 | 100 | 1.30 | 1.35 | 1.44 | 0.89 | 0.91 | 0.92 | 0.94 | 0.92 | 0.93 |
| 0.75 | 250 | 1.26 | 1.32 | 1.42 | 0.93 | 0.94 | 0.94 | 0.93 | 0.94 | 0.94 |
| 0.75 | 1,000 | 1.32 | 1.47 | 1.49 | 0.94 | 0.94 | 0.95 | 0.97 | 0.94 | 0.95 |
| 0.25 | 100 | 0.96 | 1.11 | 1.19 | 0.94 | 0.93 | 0.94 | 0.93 | 0.95 | 0.94 |
| 0.25 | 250 | 0.99 | 1.06 | 1.19 | 0.94 | 0.96 | 0.94 | 0.95 | 0.95 | 0.94 |
| 0.25 | 1,000 | 1.00 | 1.09 | 1.16 | 0.95 | 0.94 | 0.96 | 0.97 | 0.94 | 0.96 |

*Notes:* Coverage is of the nominal 95 percent interval. The first block (1.25) is the level heterogeneity of the designs behind Table 3, in independent replications.

With the heterogeneity of the main designs, the RMSE of fixed-effects Poisson is 1.7 to 2.4 times that of `xtgamfe`, and the coverage of its intervals, with its cluster-robust standard errors, is 0.85 with 100 panels and two periods (where its average standard error is 0.75 of the standard deviation of its estimates) but 0.93 to 0.95 with 1,000. With the smallest heterogeneity, the ratio is 1.0 with two periods and rises only to about 1.2 with eight, and Poisson coverage is 0.93 to 0.96. These results indicate that both the precision gap and the undercoverage in Table 3 are largely consequences of heavy-tailed heterogeneity in levels across panels, and not of gamma errors as such. The first is a matter of weighting by level as against weighting by share, the second a small-sample property of the cluster-robust variance when a few panels dominate. The size of the gap also depends on the variance function and on how the errors of a panel are related. Table 9 holds the error process and the regressors fixed while the spread of levels changes, and every observation has the same relative dispersion. When the variance is proportional to a lower power of the mean, observations with high levels have less dispersed errors and weighting by level is the better choice, as Table 5 shows. With a higher power the reverse holds for precision, although the gamma estimator is then biased.

Nor does equal relative dispersion settle the ordering by itself, because the information in a panel also depends on how its errors are related over time. How much precision the gamma estimator gains over Poisson thus depends on the data and should be judged by fitting both to the same sample.

The same mechanism bears on the test of `estat fepoisson` (Section 3.3), whose panel bootstrap estimates the variance of the difference between the two estimators. Table 10 reports how often its tests rejected, with four periods and 199 bootstrap replications per test, at the three levels of heterogeneity of Table 9, under which both estimators are consistent, and with variances proportional to $\mu^{1.5}$ and $\mu^{2.5}$ as in Table 5, under which only Poisson is.

**Table 10.** Rejection rates of the tests of `estat fepoisson` at the nominal 5 percent level, four periods: the differences between the `xtgamfe` and FE Poisson coefficients on $x_1$ and $x_2$, one at a time and jointly, and the mean bootstrap standard error of each difference relative to the standard deviation of the difference across replications (SE ratio). 199 bootstrap replications per test; 500 replications per cell (1,000 in the first two rows).

| Case | $G$ | Rejection rate: $x_1$ | Rejection rate: $x_2$ | Rejection rate: joint | SE ratio: $x_1$ | SE ratio: $x_2$ |
|---|---|---|---|---|---|---|
| Size: panel-level s.d. 1.25 | 250 | 0.085 | 0.070 | 0.083 | 0.85 | 0.87 |
| Size: panel-level s.d. 1.25 | 1,000 | 0.059 | 0.075 | 0.073 | 0.90 | 0.84 |
| Size: panel-level s.d. 0.75 | 250 | 0.042 | 0.068 | 0.058 | 0.99 | 0.97 |
| Size: panel-level s.d. 0.75 | 1,000 | 0.058 | 0.050 | 0.048 | 1.00 | 0.98 |
| Size: panel-level s.d. 0.25 | 250 | 0.072 | 0.054 | 0.054 | 0.92 | 0.98 |
| Size: panel-level s.d. 0.25 | 1,000 | 0.058 | 0.058 | 0.050 | 0.95 | 0.99 |
| Power: variance $0.5\mu^{1.5}$ | 250 | 0.218 | 0.132 | 0.216 | 0.95 | 0.97 |
| Power: variance $0.5\mu^{1.5}$ | 1,000 | 0.496 | 0.320 | 0.590 | 0.94 | 0.98 |
| Power: variance $0.5\mu^{2.5}$ | 250 | 0.096 | 0.116 | 0.108 | 0.73 | 0.78 |
| Power: variance $0.5\mu^{2.5}$ | 1,000 | 0.102 | 0.086 | 0.110 | 0.77 | 0.79 |

*Notes:* Size: both estimators consistent; gamma errors with shape 2 and the panel-level standard deviation of the log mean of Table 9; the rows with 1.25 pool the design of Table 3 and the first level of Table 9. Power: the outcomes of Table 5 with variance $0.5\mu^p$, under which FE Poisson is consistent and `xtgamfe` is not. The single differences are referred to the normal distribution and the joint test to the Hotelling-type $F$ approximation, as `estat fepoisson` reports them. The Monte Carlo standard error of a rejection rate near 0.05 is 0.010 (0.007 with 1,000 replications), and at most 0.022.

With the heterogeneity of the main designs, the tests at the nominal 5 percent level rejected a true null hypothesis in 7.0 to 8.5 percent of replications with 250 panels and in 5.9 to 7.5 percent with 1,000, because the bootstrap standard error of a difference averaged only 0.84 to 0.90 of the standard deviation of that difference across replications. With less heterogeneity they rejected in 4.2 to 7.2 percent, close to the nominal rate given a Monte

Carlo standard error of about 0.010, and the ratio was 0.92 to 1.00. At the nominal 1 percent level the tests rejected in 1.5 to 2.8 percent of replications with 250 panels and in 0.9 to 1.7 percent with 1,000 in the main designs, and in 0.6 to 1.6 percent with less heterogeneity. (The Monte Carlo standard error of a rate near 0.01 is about 0.004 with 500 replications and 0.003 with 1,000.) With 250 panels and the heterogeneity of the main designs, the excess was therefore larger in proportion to the nominal rate than at the 5 percent level. Using the standard deviation of each difference across replications in place of its bootstrap standard error brought the rejection rates of the single tests to 4.0 to 5.8 percent in every size design. Switching among the reference distributions examined here changed rejection much less: referring the joint statistic to the Hotelling-type $F$ approximation rather than to $\chi^2(2)$ lowered its rejection rate by at most 0.012, and $t$ critical values on $B - 1$ degrees of freedom changed those of the single tests by at most 0.004.

The power of the test was modest with four periods. With a variance proportional to $\mu^{1.5}$, under which `xtgamfe` overstates the coefficient on $x_1$ by about 0.04 (Table 5), the joint test rejected in 22 percent of replications with 250 panels and in 59 percent with 1,000. With a variance proportional to $\mu^{2.5}$, under which `xtgamfe` understates that coefficient by about as much, the joint test rejected in 11 percent of replications with either number of panels, not far above the 7 to 8 percent at which it rejected a true null hypothesis in the main designs. The reason is that Poisson is then imprecise: the standard deviation of the difference on $x_1$ was 0.08 to 0.14, against a mean difference of $-0.031$ to $-0.035$. Even that rate owes something to bootstrap standard errors that were again too small (0.73 to 0.79 of the standard deviations): divided by their standard deviations across replications, the single differences rejected in only 5.4 to 6.6 percent of replications.

### 4.5 Selection into positive outcomes

Because the command requires positive outcomes, it is usually applied to the positive part of an outcome that also has zeros. In a version of the gamma design in which observations are retained according to a probit index, all three fixed-effects estimators are centered on the retained observations when the index depends on $x_1$ and the fixed effect alone, because the retained errors remain exchangeable. About 69 percent of the observations are then

retained (`xtgamfe`, which drops panels left with a single observation, uses 66 percent of all observations with four periods and 69 percent with eight), and with 1,000 panels and four or eight periods the bias of `xtgamfe` is 0.000 and its coverage 0.94 to 0.95. When the index also depends on the logarithm of the error, all three are biased, by −0.07 to −0.08 for `xtgamfe`, −0.05 for Poisson, and −0.08 for log-linear fixed effects, and coverage for `xtgamfe` falls below 0.03. More generally, the score must remain centered among the retained observations, for which it is sufficient that share centering hold given the pattern of retained periods; selection still changes the population described by the estimates.

### 4.6 Few clusters

Cluster-robust standard errors rely on the number of clusters being large (Cameron and Miller 2015). Table 11 reports the gamma design of Section 4.2 with 20, 30, and 50 panels, for the default standard errors with normal critical values and with $t$ critical values on $G-1$ degrees of freedom, and for `vce(jackknife)`, which uses the same $t$ critical values.

**Table 11.** Few clusters, gamma errors with shape 2: ratio of the mean standard error to the standard deviation of the estimates (SE/SD) and coverage of the nominal 95 percent interval, for the coefficient on $x_1$. 1,000 replications per cell.

| $G$ | $T$ | `xtgamfe`, robust: SE/SD | `xtgamfe`, robust: coverage, normal | `xtgamfe`, robust: coverage, $t$ | `xtgamfe`, jackknife: SE/SD | `xtgamfe`, jackknife: coverage, $t$ | FE Poisson: SE/SD | FE Poisson: coverage | Log-linear FE: SE/SD | Log-linear FE: coverage |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 4 | 0.96 | 0.926 | 0.941 | 1.01 | 0.949 | 0.72 | 0.802 | 0.98 | 0.922 |
| 20 | 8 | 0.97 | 0.935 | 0.949 | 0.99 | 0.951 | 0.69 | 0.819 | 0.97 | 0.930 |
| 30 | 4 | 0.97 | 0.931 | 0.939 | 1.00 | 0.947 | 0.76 | 0.862 | 1.00 | 0.929 |
| 30 | 8 | 0.98 | 0.943 | 0.951 | 0.99 | 0.953 | 0.74 | 0.830 | 0.97 | 0.943 |
| 50 | 4 | 1.00 | 0.935 | 0.946 | 1.01 | 0.949 | 0.78 | 0.867 | 0.99 | 0.939 |
| 50 | 8 | 1.00 | 0.940 | 0.945 | 1.01 | 0.948 | 0.80 | 0.869 | 1.03 | 0.958 |

*Notes:* The two coverage rates of the default standard errors come from the same replications. Normal critical values are used for FE Poisson and log-linear FE (the text gives their coverage with $t$ critical values). The FE Poisson standard errors carry no finite-sample factor, whereas the covariance matrix of `xtgamfe` includes $M/(M-1)$, which raises its standard errors by a factor of 1.026 with 20 panels. The Monte Carlo standard error of a coverage rate near 0.95 is about 0.007.

With normal critical values the default intervals of `xtgamfe` cover the true value in 0.926 to 0.943 of replications, slightly less than the nominal rate. With $t$ critical values they cover it in 0.939 to 0.951, and the jackknife intervals in 0.947 to 0.953 (0.936 to 0.945 with normal critical values), so that most of the improvement comes from the critical value, which costs no

refitting, and a little more from deleting panels, which costs one fit per panel. Fixed-effects Poisson, with its cluster-robust standard errors, does poorly in this design (coverage of 0.80 to 0.87, and 0.834 to 0.875 with $t$ critical values), for the reason given in Section 4.4, whereas log-linear fixed effects, with the $t$ critical values that `xtreg` itself uses, covers in 0.940 to 0.960, close to `xtgamfe`. These designs have balanced panels, gamma errors, and independent panels, which are also the units that `vce(jackknife)` deletes; with clusters of panels it is the clusters that should be deleted (Section 3.2), a case not simulated.

### 4.7 The shape estimators

Table 12 reports the two estimators of Section 2.6 under the gamma model.

Both estimators are nearly unbiased (the largest bias is 0.010), and maximum likelihood is the more precise, with a standard deviation 63 to 96 percent of that of the moment estimator, the gain being largest for small shapes and longer panels. Its standard error matches the standard deviation of the estimates, and its interval covers the true value in 0.92 to 0.97 of replications. With 250 panels the biases are somewhat larger (at most 0.017, for the maximum likelihood estimator with a shape of 2 and two periods, whose standard deviation in that cell is 0.165). Under the gamma model the coefficient intervals based on `vce(oim)` have coverage similar to that of the robust ones (0.92 to 0.96 and 0.93 to 0.96; the two disagree on coverage in 128 of the 16,000 paired replications).

### 4.8 Average partial effects

Table 13 examines the effects that `estat ape` reports and their standard errors. The design has a continuous regressor $x_1$ that is correlated with the fixed effect, a second continuous regressor, and a time-varying indicator $d$ whose probability rises with the fixed effect, and the target in each case is the population average effect of Section 2.7, computed from two million simulated panels. Table 13 lists the base gamma-error design and its six variations. The variations include lognormal errors whose log standard deviation rises with the fixed effect from 0.4 to 1.0, informative sampling of panels with inverse-probability weights, and gamma errors whose shape rises linearly from 0.5 in the first period to 4 in the last. Another variation groups panels into clusters of five that share half the fixed-effect variance and a period-specific lognormal shock, with lognormal errors and cluster-level standard errors. The

**Table 12.** Estimators of the gamma shape under the gamma model, 1,000 panels: bias and standard deviation (SD) of the moment and maximum likelihood (ML) estimators, with the mean standard error of the latter and the coverage of its nominal 95 percent interval. 500 replications per cell.

| $\kappa$ | Panel | Moment: bias | Moment: SD | ML: bias | ML: SD | ML: mean SE | ML: coverage |
|---|---|---|---|---|---|---|---|
| 0.5 | $T = 2$ | 0.000 | 0.023 | +0.001 | 0.017 | 0.017 | 0.95 |
| 0.5 | $T = 4$ | +0.001 | 0.015 | +0.001 | 0.011 | 0.010 | 0.94 |
| 0.5 | $T = 8$ | 0.000 | 0.011 | 0.000 | 0.007 | 0.007 | 0.94 |
| 0.5 | $T_i = 2$ to 4 | +0.002 | 0.019 | +0.001 | 0.014 | 0.013 | 0.92 |
| 0.7 | $T = 2$ | +0.005 | 0.029 | +0.004 | 0.025 | 0.025 | 0.94 |
| 0.7 | $T = 4$ | +0.001 | 0.022 | +0.001 | 0.016 | 0.015 | 0.94 |
| 0.7 | $T = 8$ | +0.001 | 0.016 | +0.001 | 0.010 | 0.010 | 0.94 |
| 0.7 | $T_i = 2$ to 4 | +0.001 | 0.025 | +0.001 | 0.019 | 0.018 | 0.95 |
| 1 | $T = 2$ | +0.004 | 0.043 | +0.006 | 0.039 | 0.038 | 0.96 |
| 1 | $T = 4$ | +0.001 | 0.028 | +0.002 | 0.021 | 0.022 | 0.97 |
| 1 | $T = 8$ | +0.001 | 0.021 | 0.000 | 0.015 | 0.015 | 0.96 |
| 1 | $T_i = 2$ to 4 | +0.001 | 0.034 | +0.002 | 0.027 | 0.027 | 0.95 |
| 2 | $T = 2$ | +0.008 | 0.086 | +0.005 | 0.082 | 0.081 | 0.93 |
| 2 | $T = 4$ | +0.003 | 0.053 | +0.003 | 0.046 | 0.047 | 0.95 |
| 2 | $T = 8$ | 0.000 | 0.040 | +0.001 | 0.032 | 0.031 | 0.94 |
| 2 | $T_i = 2$ to 4 | +0.010 | 0.062 | +0.007 | 0.056 | 0.058 | 0.96 |

*Notes:* The Monte Carlo standard error of a coverage rate near 0.95 is about 0.010.

fixed effect is normal with a standard deviation of 0.5 or 1, and Table 13 shows the larger value, for which the level of the mean is highly skewed.

Where share centering holds (the base design and the first five variations), the effects are nearly unbiased: over all cells, including those with 100 panels, eight periods, or the smaller dispersion of the fixed effect, which Table 13 does not show, the largest relative bias is 1.9 percent, and below 1 percent with 1,000 panels. The standard errors are close to the standard deviation of the estimates when the fixed effects are moderately dispersed (with a standard deviation of 0.5 and 1,000 panels, ratios of 0.97 to 1.03 and coverage of 0.931 to 0.961) and somewhat too small when they are widely dispersed. With a standard deviation of 1 the ratio is 0.89 to 1.02 with 1,000 panels and 0.87 to 1.01 with 250, and coverage is 0.905 to 0.954 and 0.857 to 0.950, falling as low as 0.82 with 100 panels.

A ratio below one, and coverage below the nominal rate, can come from the estimation of the standard error, from the error of its linear approximation, from a few influential panels, or from the skewness of the average, which inherits that of the outcome level. A further 500 replications of the gamma, lognormal, weighted, and clustered designs, with 250 and 1,000

panels and the targets of Table 13, are consistent with the last two. The misses were from below: with 250 panels the interval for $x_1$ lay below the population effect in 6 to 12 percent of replications and above it in at most 0.4 percent. The estimate and its standard error were positively correlated (0.80 to 0.85 across replications with 250 panels), as they would be if a sample that lacked the few panels at the top of the level distribution gave an effect and a standard error that were both too small. Substituting other standard errors or critical values scarcely changed coverage. In the five cells where it was run, the delete-one-cluster jackknife standard error averaged 1.000 to 1.006 times that of `estat ape`. Replacing the `estat ape` standard error with the jackknife or bootstrap standard error, or using $t$ critical values on $M - 1$ degrees of freedom, changed coverage by at most 0.008.

**Table 13.** Average partial effects from `estat ape`: bias as a percentage of the population effect, ratio of the mean standard error to the standard deviation of the estimates (SE/SD), and coverage of the nominal 95 percent interval, for the continuous regressor $x_1$ and the indicator $d$. Four periods (two to four in the unbalanced design); standard deviation of the fixed effect 1; 2,000 replications per cell.

| Design | $G$ | $x_1$: bias (%) | $x_1$: SE/SD | $x_1$: coverage | $d$: bias (%) | $d$: SE/SD | $d$: coverage |
|---|---|---|---|---|---|---|---|
| Gamma errors | 250 | +0.7 | 0.90 | 0.914 | +0.3 | 0.95 | 0.936 |
| Gamma errors | 1,000 | +0.2 | 0.97 | 0.944 | −0.1 | 1.02 | 0.945 |
| Gamma errors, 2 to 4 periods | 250 | −0.4 | 0.90 | 0.913 | −0.6 | 0.94 | 0.937 |
| Gamma errors, 2 to 4 periods | 1,000 | 0.0 | 0.98 | 0.939 | −0.3 | 0.95 | 0.934 |
| Lognormal errors | 250 | −0.1 | 0.93 | 0.909 | −0.2 | 0.95 | 0.921 |
| Lognormal errors | 1,000 | 0.0 | 0.94 | 0.919 | −0.2 | 0.97 | 0.939 |
| Polynomial and interaction | 250 | +0.7 | 1.01 | 0.950 | −0.6 | 1.00 | 0.946 |
| Polynomial and interaction | 1,000 | +0.1 | 0.97 | 0.942 | 0.0 | 0.99 | 0.950 |
| Informative sampling, weighted | 250 | −1.6 | 0.87 | 0.857 | −1.8 | 0.91 | 0.903 |
| Informative sampling, weighted | 1,000 | −0.1 | 0.89 | 0.910 | −0.4 | 0.95 | 0.934 |
| Clusters of five panels | 250 | −0.9 | 0.91 | 0.887 | −1.1 | 0.93 | 0.906 |
| Clusters of five panels | 1,000 | +0.8 | 0.96 | 0.937 | +0.8 | 0.98 | 0.942 |
| Dispersion differs by period | 250 | +0.1 | 0.91 | 0.911 | −0.3 | 0.97 | 0.937 |
| Dispersion differs by period | 1,000 | +0.2 | 0.97 | 0.944 | −0.1 | 0.98 | 0.946 |
| Same, with period indicators | 250 | +2.9 | 0.90 | 0.933 | +2.4 | 0.96 | 0.946 |
| Same, with period indicators | 1,000 | +3.0 | 0.97 | 0.957 | +2.7 | 0.98 | 0.953 |

*Notes:* The Monte Carlo standard error of a coverage rate near 0.95 is about 0.005, and that of the simulated population effects is 0.10 to 0.23 percent of the effect in the cells shown. In the weighted design $G$ is the expected number of panels that the informative sampling retains (214 to 294 in the replications with 250); in the clustered design the 250 and 1,000 panels form 50 and 200 independent clusters. The last four rows show one design, fitted without and with period indicators.

Intervals that follow the skewness did better. The interval formed on the log scale, $\operatorname{sign}(\hat{\theta}) \exp\{\ln|\hat{\theta}| \pm 1.96\, \mathrm{se}(\hat{\theta})/|\hat{\theta}|\}$, needs no refitting; it cannot contain zero, and so suits

effects whose sign is not in doubt. On the 2,000 replications of each of the first fourteen rows of Table 13 it raised coverage in every cell, to 0.885 to 0.958 for $x_1$ and 0.934 to 0.961 for $d$, and in the further replications it was at most 4.5 percent longer on average. A panel bootstrap-t with 399 resamples of the panels (of the clusters in the clustered design) studentizes each resampled effect by the standard error that `estat ape` reports for it. In the further replications of the lognormal, weighted, and clustered designs with 250 panels, it raised coverage for $x_1$ to between 0.922 and 0.934, from between 0.878 and 0.900 for the normal interval in the same replications, at the cost of one fit per resample and intervals 11 to 20 percent longer at the median. For $d$ it did no better than the log scale except in the clustered design. Percentile, bias-corrected, and bias-corrected and accelerated bootstrap intervals helped less. No interval restored the nominal rate where the level was most skewed (0.92 at best for $x_1$ in the weighted design with 250 panels), as might be expected if the shortfall comes from panels that a sample lacks, since resampling can only reuse the panels that were drawn. Intervals for effects on the dollar scale therefore need more panels than intervals for the coefficients, most of all in the weighted design (coverage of 0.857 for $x_1$ with 250 panels; there the unweighted fit estimates the average over the sample, 22 to 37 percent below the population effect). The help file of `xtgamfe` gives the code for both intervals.

The last four rows illustrate the fitted-mean bias described in Section 2.7. The regressors have no time pattern in this design, so that the coefficients are centered with or without period indicators, and without them the average partial effects are centered as well. With period indicators, however, the period coefficients absorb the differences in dispersion, and with 1,000 panels the average partial effects are 2.5 to 3.0 percent too large with four periods and 0.4 to 1.0 percent with eight. The product of the coefficient and the mean outcome, by contrast, is within 0.5 percent of the target in each of those cells for the two continuous regressors. `estat ape, ybar`, which computes that product and extends it to the indicator, was checked in 300 further replications with 1,000 panels of four periods and 300 with eight. Its effects were within 1.1 percent of the targets with four periods and within 0.4 percent with eight (Monte Carlo standard errors of 0.5 to 0.7 percent), against 3.0 to 4.0 and 0.5 to 0.8 percent for the fitted-mean effects, and their standard deviation across replications was 1.9 to 3.4 percent smaller. Its intervals covered the targets in 0.947 to 0.960 of the

replications with four periods and in 0.910 to 0.943 with eight, against 0.943 to 0.967 and 0.907 to 0.927 for the fitted-mean effects; with eight periods both show the undercoverage with widely dispersed fixed effects described above. The log-scale and bootstrap-t intervals above were evaluated for fitted-mean effects. In these replications, the log-scale interval for the `ybar` effects covered the targets in 0.950 to 0.960 of replications with four periods and 0.910 to 0.933 with eight. That shift of $-0.010$ to $+0.013$ from the normal interval is at most about 1.4 times the Monte Carlo standard error of the paired difference (0.005 to 0.010), so in this design the log scale neither helped nor hurt.

### 4.9 Run time

Table 14 reports the time taken by `xtgamfe` and by the other routes to the same coefficients, on simulated gamma data with 2, 10, or 50 regressors. The operations timed differ across commands. For `xtgamfe`, the timing covers the fit (which includes both shape estimates and the clustered standard errors) and `predict, mu`; for the auxiliary regression with `xtpoisson`, it covers the fit and the lines that recover the fitted means and the moment estimate of the shape. For `ppmlhdfe`, it covers the fit alone, with standard errors clustered on panels; for `xtspj`, the fit on the outcome multiplied by 1,000, with no fitted means, shape, or clustered standard errors; and for `glm`, the fit with an indicator for every panel and clustered standard errors. All 18 of those fits of `xtspj` agreed with `xtgamfe` on the coefficient on $x_1$, the only coefficient compared, to within $10^{-4}$, whereas on the outcome in its original units it reported convergence at wrong coefficients with two regressors and stopped with a convergence error with 10 and 50 (Section 4.1). The fits of `ppmlhdfe` agreed with `xtgamfe` to within $10^{-9}$ in the coefficient on $x_1$ and to a relative difference below $10^{-5}$ in its clustered standard error. All fits were run on one processor of an AMD Ryzen Threadripper 1950X after an untimed first fit of each command.

On one processor `xtgamfe` is faster than the auxiliary regression with `xtpoisson` in every configuration, by a factor of 1.6 to 5.0 that grows with the number of regressors. It is 3.6 to 13.5 times as fast as `xtspj` with the exponential model when the two are timed in the same run, and `glm` with indicators takes about half a minute with 1,000 panels. The auxiliary regression with `ppmlhdfe`, which returns the coefficients and clustered standard errors but not

the fitted means or the shape, took 0.68 to 2.1 times as long as `xtgamfe` with its fitted means in the same run (0.68 to 1.2 times with two regressors). The advantage over `xtpoisson` does not carry over when Stata/MP uses several processors. On the same machine, in a benchmark with 10,000 panels, four periods, two regressors, and an offset, with Stata/MP free to use 16 processors, the auxiliary regression was the faster in eight of nine repetitions in three separate runs (0.19 to 0.25 seconds, median 0.21, against 0.22 to 0.27, median 0.22, for `xtgamfe`), so that there the auxiliary regression is slightly the faster. Set beside the single-processor times of Table 14, this suggests that `xtpoisson` makes better use of several processors. `xtgamfe` and the two auxiliary regressions all fit a million observations in under a minute, so that run time is not a reason to choose among them.

**Table 14.** Run time in seconds (median of three fits, one processor).

| $G$ | $T$ | Regressors | `xtgamfe` | Aux. `xtpoisson` | Aux. `ppmlhdfe` | `xtspj`, exponential, outcome × 1,000 | `glm` with indicators | FE Poisson | Log-linear FE |
|---|---|---|---|---|---|---|---|---|---|
| 1,000 | 4 | 2 | 0.05 | 0.08 | 0.05 | 0.58 | 25.9 | 0.10 | 0.03 |
| 1,000 | 8 | 2 | 0.07 | 0.13 | 0.08 | 0.63 | 27.2 | 0.13 | 0.04 |
| 10,000 | 4 | 2 | 0.45 | 0.88 | 0.36 | 5.92 | — | 0.87 | 0.28 |
| 10,000 | 8 | 2 | 0.89 | 1.56 | 0.60 | 6.53 | — | 1.48 | 0.50 |
| 10,000 | 8 | 10 | 1.28 | 3.92 | 1.62 | 7.21 | — | 3.90 | 0.65 |
| 10,000 | 8 | 50 | 4.28 | 21.6 | 8.31 | 14.8 | — | 21.3 | 1.69 |
| 125,000 | 8 | 2 | 7.65 | 12.6 | 7.23 | — | — | 12.3 | 3.62 |
| 125,000 | 8 | 10 | 13.0 | 36.4 | 20.7 | — | — | 36.2 | 5.94 |

*Notes:* A dash means that the fit was not attempted (`glm` with more than 1,000 panels; `xtspj` with more than 10,000). `xtspj` (default settings, outcome multiplied by 1,000) and `ppmlhdfe` were timed in separate runs on the same datasets (the same code and seed, one processor, no other Stata job). In those runs `xtgamfe` took 2 to 7 percent less time than shown (with `xtspj`) and up to 8 percent less (with `ppmlhdfe`). Each column times the operations described in Section 4.9. Aux. denotes the auxiliary regression of a constant (Section 2.5). FE Poisson (`xtpoisson, fe` on the outcome) and log-linear FE are different estimators, shown for scale.

# 5. Illustration

## 5.1 Data and design

The command is illustrated on a design taken from the medical cost-offset literature, in which nondrug medical spending is regressed on *lagged* medication use with person fixed effects (Roebuck et al. 2015). The illustration adapts that design to public survey data, with days supplied as the exposure and the positive part of spending as the outcome, and applies the

checks of Sections 2 and 4 to the resulting unbalanced panel; it does not estimate a cost offset.

The data are the public-use files of the Medical Expenditure Panel Survey (MEPS) for 2011–2019 (Agency for Healthcare Research and Quality 2021b), panels 16–23. MEPS interviews households, and the analysis follows persons across the first five regular interview rounds of their panel, which span two calendar years; the unit of analysis is the person-round. The outcome for a round is total paid medical expenditure excluding prescription drugs (in nominal dollars), summed over the seven medical event files (office-based, outpatient, emergency room, inpatient, home health, dental, and other medical expenses such as equipment, supplies, and glasses). The exposure is the days supplied on fills of the condition's medications purchased in the *prior* round, expressed in 30-day units. Each fill with missing days supplied is counted at the person's median valid days supplied in the panel (Section 5.2 examines other codings). A fill is classified by its Multum therapeutic class or by its drug name, with the classes that AHRQ corrected for 2011–2013. Two cohorts of adults are analyzed in parallel, one entering on evidence of diabetes and the other on a condition record coded to schizophrenia or another psychotic disorder or to bipolar disorder, or on an antipsychotic fill. Because 60 percent of the second cohort's estimation sample (1,377 of 2,295 persons) enters through a fill alone, it is called the antipsychotic cohort here (Appendix B gives the entry rules, the codes, and the classification). With $y_{ir}$ the spending of person $i$ in round $r$, $d_{i,r-1}$ the prior-round days supplied in 30-day units, and $L_{ir}$ the length of the round in days, the model is

$$\mathrm{E}[\, y_{ir} \mid \mathbf{X}_i, \alpha_i \,] = \exp\Big(\alpha_i + \beta\, d_{i,r-1} + \mathbf{z}'_{ir}\boldsymbol{\gamma} + \ln L_{ir}\Big),$$

where $\mathbf{X}_i$ collects the regressors and offsets of all of person $i$'s retained rounds, and $\mathbf{z}_{ir}$ contains indicators for prior-round self-rated health, calendar-year indicators (the year in which the round ends), and an indicator for prior-round use of a drug subclass (insulin; atypical antipsychotics). Rounds differ in length by design, and the log of the round's length enters as an offset: the conditional mean of positive spending in a round is then the round's length times the expected spending per day of the round, which fixes at one the elasticity of positive spending with respect to the round's length. Row 10 of Table 15 estimates that

elasticity instead, as row 11 does for fixed-effects Poisson, and the data reject the restriction (Section 5.2). The estimates are unweighted, which is acceptable for an illustration but would not be for population inference. A weight that is constant within person preserves share centering within each person, provided that the condition also holds given the weight, whereas the annual person weights of the full-year files change from year to year, so that they would vary within panel, the case of Section 2.8. AHRQ's panel longitudinal files supply such a weight (`LONGWT`), one value per person for the two years of a panel, for the panel members who took part for as long as they were in the population (Agency for Healthcare Research and Quality 2021a, HC-217 documentation, C-1 and C-6); it is not used here.

Because `xtgamfe` requires a strictly positive outcome, the estimation sample comprises rounds with positive spending, an observed prior round with reported self-rated health, and a usable positive round length, restricted to persons with at least two such rounds. This yields 42,368 rounds for 12,123 persons in the diabetes cohort and 8,230 rounds for 2,295 persons in the antipsychotic cohort, with two to four rounds per person. The requirement of positive spending, with at least two positive rounds per person, removes 20 and 14 percent of the otherwise eligible rounds (Appendix B gives the sample flow). The lags are formed on the full grid of rounds before that restriction, so that a lag can come from a round with no spending and a person's retained rounds need not be consecutive. Every coefficient below therefore describes an association conditional on positive spending, not with total spending. Both outcomes have the usual shape of positive cost data: mean spending per round is \$4,491 against a median of \$892 in the diabetes sample, and \$5,645 against \$1,431 in the antipsychotic sample.

### 5.2 Estimates

In each cohort the main model is fit with

```
. xtset person_id round
. xtgamfe nonrx_round rxdays30_r_lag subrx_r_lag i.rthlth_lag i.round_year ///
>     if gamma_sample, offset(ln_round_days)
```

(`gamma_sample` marks the estimation sample defined in Section 5.1; without that restriction the command stops, because some panels contain rounds with no spending), and Table 15 reports the coefficient on prior-round days supplied from that model and from the checks

that follow.

Under the main coding neither association can be distinguished from zero. In the diabetes cohort the coefficient is 0.0009 (standard error 0.0016) in row 1 and at most 0.8 standard errors from zero in any `xtgamfe` row of Table 15, although fixed-effects Poisson on the same rounds puts it at 0.0055 (0.0025), about 2.2 standard errors above zero. In the antipsychotic cohort it is −0.0013 (0.0075) in row 1, so that, within person, an additional 30 days supplied in the prior round is associated with about 0.13 percent lower positive nondrug spending in the current round (95 percent interval, 1.6 percent lower to 1.3 percent higher). It is negative in every row but those of fixed-effects Poisson (rows 4 and 11), and more than two standard errors from zero only when fills with missing days supplied are counted as zero days (row 7). The checks below matter most for three points: the coding of the exposure, the difference between `xtgamfe` and fixed-effects Poisson, and strict exogeneity, on which the evidence is mixed.

Because calendar-year indicators could absorb a period pattern in the errors (Section 2.7), the dollar figure to report is that of `estat ape, ybar`, which multiplies the coefficient by mean observed spending over the retained positive rounds: for the antipsychotic coefficient of row 1, −$7.49 a round for each 30 days supplied (standard error $42.19), in nominal dollars pooled over 2011–2019. It translates the fitted association into dollars among those rounds, and it is an average partial effect on their conditional mean only if that mean is correctly specified given the pattern of retained rounds (Section 2.7). The default `estat ape` averages the fitted means instead (mean $5,773, against a mean outcome of $5,645) and gives −$7.66 ($43.14), about 2 percent larger in magnitude. With `finite`, a 30-day increase in every retained person-round would lower the average fitted mean by $7.65 ($43.09), or 0.13 percent (Appendix A.5).

*Period indicators (row 3).* Of the indicators for the round of the panel, the contrasts of rounds 4 and 5 with round 3 are jointly significant in the diabetes cohort ($\chi^2(2) = 10.0$, $p = 0.007$) but not in the antipsychotic cohort ($\chi^2(2) = 0.1$, $p = 0.97$; Appendix B explains why only these two contrasts are tested), and the exposure coefficients barely move (0.0009 and −0.0012). Row 3 is a sensitivity check rather than an application of Proposition 2. The panel is unbalanced, dropping zero-spending rounds can leave gaps, and retention can depend

on the spending shocks themselves. Neither the significance of the indicators nor the stability of the exposure coefficient therefore shows that share centering holds.

*Round length (row 10).* The main specification models positive spending per day of the round. Both cohorts reject the unit elasticity that the offset imposes, but neither `xtgamfe` exposure coefficient is distinguishable from zero under either specification. Entered as a regressor instead, the log length has coefficients of 0.752 (0.029) and 0.756 (0.060), and the tests of unit elasticity give $\chi^2(1) = 73.0$ and 16.3. An elasticity below one among rounds with positive spending is to be expected even if spending accrues at a constant daily rate, because the probability that a round has any spending also rises with its length, so that the mean among positive rounds rises less than in proportion. Selection on positive spending is not the whole explanation, however: fit to all eligible rounds, zeros included, with the log length as a regressor, fixed-effects Poisson gives elasticities of 0.871 (0.059) and 0.745 (0.095), also below one, so that, given the model's other terms, expected spending itself rises less than in proportion to the round's length. Because rounds of different length also differ in season and in position within the panel, the coefficient need not be an effect of length alone. In `xtgamfe` on the positive rounds with the log length as a regressor (row 10), the exposure coefficients are −0.0010 (0.0017) and −0.0047 (0.0074), shifts of −0.0019 and −0.0034 from row 1. These are 1.2 and 0.4 times the standard errors of row 1, a scale and not a test, because both fits use the same observations. The offset model serves as the main one for its interpretation, as a model of spending per day of the round that puts rounds of different lengths on one scale. The comparison with fixed-effects Poisson on the same rounds below is reported both under the restriction and with the round's length free in both models (rows 10 and 11); the other checks below keep the offset.

*Comparison estimators (rows 4, 5, and 11).* On the same observations, log-linear fixed effects stays close to `xtgamfe` (0.0005 and −0.0008), whereas fixed-effects Poisson gives 0.0055 (0.0025) in the diabetes cohort, about 2.2 standard errors above zero, and 0.0118 (0.0114) in the antipsychotic cohort, of the opposite sign. With the round's length a regressor in both models (row 11), fixed-effects Poisson gives 0.0041 (0.0025) in the diabetes cohort, 1.7 standard errors above zero, and 0.0076 (0.0115) in the antipsychotic cohort, with elasticities with respect to the round's length of 0.793 (0.063) and 0.661 (0.100), 3.3 and 3.4 standard

**Table 15.** Coefficient on prior-round days supplied (30-day units) in models of positive current-round nondrug spending, MEPS 2011–2019.

| Row | Model | Diabetes | Antipsychotic |
|---|---|---|---|
| 1 | `xtgamfe` | 0.0009 (0.0016) | −0.0013 (0.0075) |
| 2 | Row 1, standard errors clustered on households | 0.0009 (0.0017) | −0.0013 (0.0074) |
| 3 | Row 1 plus round-of-panel indicators | 0.0009 (0.0016) | −0.0012 (0.0075) |
| 4 | Fixed-effects Poisson, same observations | 0.0055 (0.0025) | 0.0118 (0.0114) |
| 5 | Log-linear fixed effects, same observations | 0.0005 (0.0015) | −0.0008 (0.0071) |
| 6 | Row 1 without prior-round self-rated health | 0.0010 (0.0016) | −0.0015 (0.0075) |
| 7 | Row 1, fills with missing days supplied counted as zero days | 0.0014 (0.0018) | −0.0163 (0.0075) |
| 8 | Row 1, such fills counted as 30 days | 0.0011 (0.0018) | −0.0062 (0.0073) |
| 9 | Row 1, persons with no such fill | 0.0008 (0.0024) | −0.0073 (0.0097) |
| 10 | Row 1, log round length as a regressor instead of an offset | −0.0010 (0.0017) | −0.0047 (0.0074) |
| 11 | Row 4, log round length as a regressor instead of an offset | 0.0041 (0.0025) | 0.0076 (0.0115) |
| | Rounds / persons, rows 1–8 and 10–11 | 42,368 / 12,123 | 8,230 / 2,295 |
| | Rounds / persons, row 9 | 23,464 / 6,763 | 5,330 / 1,494 |
| | Households, row 2 | 10,871 | 2,202 |

*Notes:* Standard errors in parentheses, clustered on persons except in row 2 (households: dwelling units within MEPS panels). Every model includes prior-round self-rated health indicators (except row 6), calendar-year indicators, prior-round subclass use, and the log round length, as an offset (subtracted from log spending in row 5) except in rows 10 and 11, where it is a regressor. A fill with missing days supplied is a classified fill with a missing-data code. Rows 1–6 and 10–11 count such a fill as the person's median days supplied over the person's valid fills of the condition's medications (30 days if there are none), row 7 as zero days, and row 8 as 30 days; fills coded as taken as needed or with zero days count as zero days in every row. Row 9 excludes persons with a missing-coded fill in any round on the grid, including rounds that supply no lag. The standard errors treat the imputed days as observed and so do not reflect the uncertainty of the imputation.

errors below one, against 0.752 and 0.756 for `xtgamfe` (row 10). Fit instead to all eligible rounds, zeros included, with the offset (51,019 rounds of 13,124 persons and 9,427 rounds of 2,422 persons, once fixed-effects Poisson drops the persons with no spending in any eligible round), fixed-effects Poisson gives 0.0038 (0.0025), 1.5 standard errors above zero, and 0.0115 (0.0111), estimates of the association with all nondrug spending rather than with positive spending.

With the offset in both models, the `estat fepoisson` panel bootstrap (999 replications, persons resampled) puts the difference between the `xtgamfe` and fixed-effects Poisson exposure coefficients at −0.0046 for diabetes (standard error 0.0019; $p = 0.017$; normal-based 95 percent interval −0.0084 to −0.0008) and −0.0131 for antipsychotics (0.0094; $p = 0.16$; −0.0315 to 0.0053). Referred to the Hotelling-type $F$ approximation, the joint tests of all 14 coefficient differences give $F(14, 985) = 1.85$ ($p = 0.028$) for diabetes and 0.81 ($p = 0.66$) for antipsychotics. With the round's length free in both models, the difference is −0.0051 (standard error 0.0020; $p = 0.009$; −0.0089 to −0.0013) for diabetes and −0.0123 (0.0094; $p = 0.19$; −0.0307 to 0.0060) for antipsychotics, and the joint tests of all 15 coefficient differences give $F(15, 984) = 1.79$ ($p = 0.031$) and 0.84 ($p = 0.64$).

In the diabetes cohort the `xtgamfe` exposure coefficient is thus below the Poisson one whether the round's length enters as an offset or as a regressor, with nominal $p$-values of 0.017 and 0.009 for the difference, so that the restriction that the offset imposes does not account for the difference. Both joint tests give $p$-values of about 0.03. No adjustment is made for testing two cohorts. In the antipsychotic cohort the comparison is too imprecise to say.

The simulations give one reason for caution with these tests. The standard deviation of the persons' estimated log spending levels per day (the fixed effect plus the person's mean of the index) is 1.43 (1.38 in the antipsychotic cohort). In the main simulation designs of Section 4.4 the same statistic, which includes the noise of estimating each level, averages about 1.34, and the tests rejected a true null hypothesis there in 6 to 8.5 percent of replications at the nominal 5 percent level and in 0.9 to 2.8 percent at the nominal 1 percent level. The comparable dispersion makes that finding relevant here, but the rejection rates of those designs, which have two regressors, four periods, and 250 or 1,000 panels, against 12,123 and

2,295 persons here, do not calibrate the MEPS tests.

Persons share primary sampling units in the survey's design. Resampling instead the 621 primary sampling units of the diabetes cohort (each unique within its stratum, from AHRQ's pooled linkage file for variance estimation; Appendix B) as independent clusters from one pool, ignoring the survey's 282 strata, is a check for dependence across persons, not design-based inference. With the same number of replications and the same seed, it raises the bootstrap standard errors of that cohort's exposure differences in the two specifications by 13 and 11 percent, to 0.0022, and gives nominal $p$-values of 0.035 and 0.019, with joint tests at $p = 0.039$ and 0.044. A difference does not show which assumption fails (share centering, the mean specification, or strict exogeneity, which the check below calls into question in the diabetes cohort), and equal coefficients would not establish share centering.

The Poisson standard errors are 1.5 times those of `xtgamfe` in both cohorts, as in the designs of Section 4.4 with heterogeneous levels, and those of log-linear fixed effects slightly smaller: a comparison of precision, not of closeness to the target. The slopes of the Park regression (`estat park`) on the `xtgamfe` residuals are 2.26 and 2.17 (standard errors 0.009 and 0.021); the persons have different numbers of rounds, so `estat park` does not test slope 2. As Section 6 explains, the regression is descriptive, and its slopes can only be compared with those of the simulation designs: slopes above 2 arose there under lognormal errors with a variance proportional to the squared mean (2.06 to 2.13), under which the estimator is consistent, and under a variance proportional to $\mu^{2.5}$ (2.39 to 2.45), under which it was biased downward. Those designs have balanced panels; with persons observed for two to four rounds, slopes above 2 can also arise under the gamma model itself (Section 6). The slopes therefore do not tell these cases apart.

*The shape.* The two estimates of the gamma shape differ: 0.70 by the moment estimator and 0.85 (standard error 0.007) by maximum likelihood in the diabetes sample, and 0.79 and 0.92 (0.017) in the antipsychotic sample. A panel bootstrap with the same seed and replications that also bootstraps the difference between the two shape estimates puts that difference at 41 and 12 standard errors from zero. Under the gamma model the two would agree up to sampling error (Section 2.6), so that the gamma model of Section 2.1, with one shape and errors independent over the periods of a panel, is rejected (persons are taken as

independent, as the bootstrap does), and `vce(oim)` should not be used; the coefficients do not require it.

*The coding of the exposure (rows 1 and 7 to 9).* In MEPS, days supplied is missing for a large share of fills: 30.2 percent of the antidiabetic and 29.1 percent of the antipsychotic fills classified here carry a missing-data code, and 5,360 of the 12,123 persons in the diabetes sample and 801 of the 2,295 in the antipsychotic sample have at least one such fill. These are purchased fills of the condition's medications whose days supplied are unknown, so that counting them as zero days understates exposure, and excluding the persons who have them would drop 44 and 35 percent of the persons. Row 1 counts each such fill at the person's own median over all of the person's valid fills of the condition's medications in the panel (Appendix B gives the details). Because the median draws on fills from later rounds as well, the exposure of a round can depend on fills made after its outcome round, which matters if medication use responds to past spending (see *Strict exogeneity* below). The antipsychotic coefficient depends on the coding. Counting missing-coded fills as zero days gives −0.0163 (0.0075), about 12 times the row 1 estimate. Counting them as 30 days (the median valid value) gives −0.0062 (0.0073), even though rows 1 and 8 assign the same value to 91 percent of missing-coded fills in the antipsychotic estimation sample (and 74 percent in the diabetes sample). Restricting the sample to persons with no missing-coded fill gives −0.0073 (0.0097). Across these codings the estimate lies between 0.2 and 2.2 standard errors below zero. The diabetes coefficient stays near zero under every coding; with the exposure expressed as a rate it is negative but within 1.3 standard errors of zero (Appendix B).

*Other checks (rows 2 and 6).* Clustering on households changes the standard errors by less than 1 percent, and clustering on MEPS primary sampling units changes them by less than 4 percent (Appendix B). Dropping prior-round self-rated health, which may itself respond to earlier spending shocks, changes neither coefficient materially. Restricting either cohort to persons with diagnosis evidence leaves both signs unchanged and neither estimate distinguishable from zero (Appendix B).

*Strict exogeneity.* A lagged exposure need not be strictly exogenous, because a spending shock in one round may change medication use in the next. Adding same-round days supplied and subclass use to row 1 is a check in the spirit of Wooldridge (2010, 325 and 764): when the

next round is in the sample, the same-round exposure is the lagged regressor of that round's equation, and under strict exogeneity it should add nothing to the conditional mean. It checks the lag-only specification as a whole: besides feedback from spending to later medication use, a rejection can come from same-round effects that the model omits or from the selection of rounds by positive spending. The test on all eligible equations is broader than a test of strict exogeneity, since the next round is outside the sample for 32 and 30 percent of rounds. The tests restricted to rounds whose next round is in the sample change precision as well as composition, and they condition on positive spending in the next round, a selection on the outcome that can itself move the estimates. If the rejection reflects feedback from spending to later medication use, generalized method of moments estimators that difference out the fixed effect and require only sequential exogeneity (Chamberlain 1992; Wooldridge 1997) remain consistent for the coefficients of the exponential mean (Section 6) when the rounds are not selected on the outcome, which here would mean modeling all eligible rounds, zeros included; they are not applied here.

In the diabetes cohort the two variables are jointly significant with `xtgamfe` ($\chi^2(2) = 26.2$, $p < 0.001$), through same-round subclass use (coefficient 0.200, standard error 0.048) and same-round days supplied (0.0036, 0.0017). The rejection holds with fixed-effects Poisson ($\chi^2(2) = 11.0$, $p = 0.004$) and when the test is restricted to rounds whose next round is in the sample (26,709 rounds of 9,464 persons, once persons left with a single such round are dropped): $\chi^2(2) = 10.1$ ($p = 0.007$) with `xtgamfe` and 9.2 ($p = 0.010$) with fixed-effects Poisson.

In the antipsychotic cohort, the `xtgamfe` test on all eligible equations also rejects ($\chi^2(2) = 9.8$, $p = 0.008$), driven by same-round subclass use (0.151, 0.059). The other versions do not reject: fixed-effects Poisson gives $\chi^2(2) = 1.9$ ($p = 0.38$), and on the 5,431 rounds (1,914 persons) whose next round is in the sample, `xtgamfe` gives 3.2 ($p = 0.21$) and fixed-effects Poisson gives 1.7 ($p = 0.43$). A nonsignificant result cannot establish exogeneity. An omitted same-round effect could be an illness that raises spending and medication use in the same round or, in the gamma fit, a regressor that marks rounds with unusual expected error shares.

Under the main coding, `xtgamfe` finds no association between prior-round days supplied and positive nondrug spending distinguishable from zero in either cohort, whether round

length enters as an offset or as a regressor (rows 1 and 10). The antipsychotic point estimate is negative under every coding of missing days supplied ($-0.0013$ to $-0.0163$), but exceeds two standard errors from zero only when missing days are counted as zero. In the diabetes cohort fixed-effects Poisson, fit to the same positive rounds, puts the coefficient 2.2 standard errors above zero with the same offset and 1.7 with the round's length free. Its exposure coefficient exceeds that of `xtgamfe` under either specification, with nominal $p$-values for the difference of 0.009 to 0.035 whether persons or primary sampling units are resampled, and the joint tests of all coefficients give $p = 0.028$ and 0.031. These $p$-values are not adjusted for the two cohorts, and in simulations with dispersed panel levels the tests rejected a true null hypothesis somewhat too often. Fit to all eligible rounds, zeros included, with the offset, Poisson puts the coefficient 1.5 standard errors above zero. Same-round medication use also predicts spending in that cohort under every version of the check, with the round's length as an offset or as a regressor in both models ($p \leq 0.011$ with the length free), so that the lag-only specification that both estimators use is in doubt there. In the antipsychotic cohort only the `xtgamfe` test on all eligible equations rejects it, under either treatment of the round's length ($p = 0.008$ with the offset and 0.003 with the length free). None of these results is an estimate of a causal effect of medication or of a cost offset.

## 6. Discussion

The `xtgamfe` command makes fixed-effects gamma regression an ordinary panel estimator in Stata. Each of the other routes supplies part of what `xtgamfe` reports. The auxiliary regression with `ppmlhdfe` gives the coefficients and clustered standard errors; `glm` with indicators gives these together with the fitted means and, through `margins`, the average partial effects with their standard errors, at a cost that grows quickly with the number of panels; and `xtspj` gives the coefficients when its optimizer reaches them, with information-based standard errors. `xtgamfe` provides all of this in one step, with a cost per iteration linear in the number of observations. It also adds two estimates of the shape that, under the gamma model, remain consistent when the number of periods is fixed, effects computed from the observed outcome, and postestimation commands for the fixed effects, fitted means, average partial effects, and information criteria (Sections 2.5, 4.1, and 4.9). A correlated-random-effects gamma

regression did as well as `xtgamfe` in the main simulation design, in which its normal random intercept, given the panel means of the regressors, is correctly specified, and in one in which $x_1$ depended on the fixed effect nonlinearly, although the fixed-effects estimator does not need that assumption (Section 4.2, which also states what was not simulated).

For applied work the choice is among three estimators of the same coefficients, each consistent under its own sufficient condition, and errors that are exchangeable given the regressors satisfy all three (that of log-linear fixed effects when the logarithm of the error has a finite mean). Table 16 collects these conditions for the error structures examined in this paper.

**Table 16.** Consistency, with the number of periods fixed, of the three estimators of the coefficients under the error structures examined in this paper.

| Errors | `xtgamfe` | `xtgamfe`, period indicators | Log-linear FE | Log-linear FE, period indicators | FE Poisson | Evidence |
|---|---|---|---|---|---|---|
| Exchangeable given the regressors, including independent and identically distributed given them | Consistent | Consistent | Consistent[a] | Consistent[a] | Consistent | Proposition 1; Table 3 |
| Stationary serial dependence, independent of the regressors | Can be inconsistent[b] | Consistent[c] | Consistent[a] | Consistent[a] | Consistent | Proposition 2; Table 6 |
| Dispersion that differs by period, independent of the regressors | Can be inconsistent[b] | Consistent[c] | Can be inconsistent[b] | Consistent[d] | Consistent | Proposition 2; Table 7 |
| Dispersion that depends on the regressors, including a variance proportional to a power of the mean other than two | Generally inconsistent; bias falls with the number of periods[e] | Generally inconsistent; bias falls with the number of periods[e] | Generally inconsistent; bias does not fall with the number of periods[e] | Generally inconsistent[f] | Consistent | Tables 4 and 5 (period indicators: Table 5) |
| Observations selected on the error (slopes of the mean before selection) | Can be inconsistent[g] | — | Can be inconsistent[g] | — | Can be inconsistent[g] | Section 4.5 |

*Notes:* Consistency as the number of panels grows with the number of periods fixed, for the coefficients of regressors that vary within panels. Every column assumes a correctly specified conditional mean (in the last row, of the outcome before selection) and strictly exogenous regressors. In unbalanced panels the error structures of the first column, and their independence from the regressors, are understood given the pattern of observation (Section 2.3; Appendix A.2). A dash marks a combination not examined.

[a] When the logarithm of the error has a finite mean.

[b] Whether a coefficient is biased depends on the time pattern of the whole design, not of its own regressor alone (Section 2.3): a regressor with a time pattern within panels, such as the treatment of Tables 6 and 7, which switches on and stays on, can be biased, and so can the coefficients of regressors related to it. In balanced panels whose errors are independent of the regressors, the coefficients are centered when every regressor has the same mean in every period, as in the design of Section 4.8.

[c] For the coefficients other than those of the period indicators, in balanced panels whose error vectors are independent of the regressors (Proposition 2); the error distribution may differ across panels in ways unrelated to the regressors. Section 2.3 and Table 8 treat unbalanced panels.

[d] For the coefficients other than those of the period indicators, which absorb the period differences in the mean of the logarithm of the error, when that mean, given the regressors, is finite and differs across periods only by amounts common to all panels (Section 2.3); under the conditions of Proposition 2, that mean differs across periods only by amounts common to all panels whenever it is finite.
[e] In the designs simulated, whose regressors are independent and identically distributed across periods given the fixed effect (Section 2.3).
[f] Not simulated; it follows from Section 2.3.
[g] Refers to the coefficients of the pre-selection mean. Each estimator recovers them when selection scales the conditional mean of retained observations by a factor constant within each panel and the estimator's consistency condition holds among retained observations given the retention pattern (for example, when i.i.d. errors are retained above a common threshold). In Section 4.5, where retention depends on the regressor, the fixed effect, and the error, all three estimators were biased.

Fixed-effects Poisson regression of the outcome is consistent under a correctly specified conditional mean and strictly exogenous regressors, with no condition on how the errors of a panel relate to one another, and it accommodates zeros. Its score weights each panel by its level, however, so that, besides identification, its consistency needs a law of large numbers for those level-weighted terms, and its clustered standard errors need them to have finite second moments: conditions on the outcome, and so on the level-weighted errors, that the gamma estimator, whose score is bounded given the regressors (Appendix A.1), does not need. Log-linear fixed effects is consistent when the mean of the log error, given the regressors, is the same in every period (Section 2.3), which a stationary error process independent of the regressors satisfies without period indicators. The gamma estimator is consistent under share centering, which such a process need not satisfy, so that under serial dependence it can need period indicators when a regressor has a time pattern within panels. In the designs of Section 4.3 with serially correlated errors and four or eight periods, log-linear fixed effects was centered without them, whereas `xtgamfe` needed them and then had a standard deviation 1.2 to 1.6 times that of log-linear fixed effects. In the independent-error designs of Table 3, in which both were consistent, the root-mean-squared error of log-linear fixed effects was 0.84 to 1.18 times that of `xtgamfe`, larger under gamma errors and smaller under lognormal errors. When the dispersion of the errors depended on a regressor, both were biased. The bias of `xtgamfe` fell as periods were added (Tables 4 and 5), whereas that of log-linear fixed effects, which estimates effects on the mean of the logarithm, did not. Only Poisson was consistent there.

An analyst who wants a single estimate with the fewest assumptions about how the errors of a panel relate to one another should, therefore, report Poisson. In the main designs,

however, its intervals with cluster-robust standard errors were the least reliable of the three (coverage of 0.85 to 0.94 with 250 or fewer panels, and 0.80 to 0.87 with 20 to 50; Sections 4.2 and 4.6), largely because a few panels with large levels dominate its cluster-robust variance. The gamma estimator is worth fitting beside it because it can be considerably more precise. With the variance proportional to the squared mean and share centering holding, the root-mean-squared error of Poisson ranged from about equal to that of `xtgamfe`, when panels had similar levels, to more than twice as large (Section 4.4), and in the MEPS illustration, on the same positive rounds, its standard errors were 1.5 times as large. With variances proportional to $\mu^{1.5}$ and $\mu$ (Table 5), the gamma estimator was biased, by the amounts given above, whereas fixed-effects Poisson, which is consistent in these designs, had much smaller simulated bias and the smaller standard deviation and root-mean-squared error (Section 4.3). With variances proportional to $\mu^{2.5}$ and $\mu^3$, the gamma estimator was biased by about as much in the opposite direction, yet with 250 or 1,000 panels it still had the smaller root-mean-squared error because Poisson's standard deviation was 3.6 to 9.5 times as large. Poisson's bias remained much smaller, at most 0.021 in absolute value in the cells of Table 5 (Monte Carlo standard error 0.009). The gamma estimator's intervals then often missed the true value (with 1,000 panels they covered it in 42 to 60 percent of replications with $\mu^{2.5}$ and in at most 4 percent with $\mu^3$), and its bias does not shrink as panels are added. A smaller standard deviation alone is no guide, because when the dispersion depended on a regressor (Table 4) the gamma estimator had the smaller one in every cell but was biased.

No single diagnostic tells these cases apart. The Park regression of `estat park` runs ordinary least squares of the log squared residuals of `xtgamfe` on the log fitted means, the form of the test of Park (1966) with the fitted mean as the variable, which Manning and Mullahy (2001) extend. It had a slope of 2.00 on average under gamma errors, of 1.3 to 1.7 when the variance was proportional to $\mu^{1.5}$ or $\mu$, and of 2.35 to 2.83 when it was proportional to $\mu^{2.5}$ or $\mu^3$, and so flags departures in either direction.

The population slope is exactly 2 under three conditions: balanced panels (the same number of observations in every panel), a fit that is unweighted or whose weights are constant within panel (with the gamma model holding given the weights), and independent gamma errors with one shape. At the true coefficients $y_{it}/\hat{\mu}_{it} = T_i W_{it}$, so that $\ln(y_{it} - \hat{\mu}_{it})^2 =$

$2 \ln \hat{\mu}_{it} + \ln(T_i W_{it} - 1)^2$, and the shares of independent gamma variables with a common scale are independent of their sum, whereas $\hat{\mu}_{it}$ depends on the errors only through the panel's mean error. With weights that vary within panel the slope can differ from 2 even under the gamma model. With panels of different lengths, the slope can also differ from 2, especially when panel length is related to the level. At the true coefficients, the relative residuals are independent of the fitted mean given $T_i$, but their distribution still depends on $T_i$. For example, when half the panels have two periods at a mean of 1 and half have eight periods at a mean of $e^2$, the population slope is 2.18 with exponential errors and 2.22 with gamma errors of shape 2. `estat park` then reports the slope without the test, as with weights that vary within panel. The standard error that `estat park` reports is that of the least-squares regression itself: it treats the fitted means and the residuals as data and so leaves out the variation from estimating the coefficients, so that the $p$-value of its test of slope 2 is nominal and its calibration depends on the design. In the design of Table 5 with $p = 2$, under which the gamma model holds, with 250 and 1,000 panels and two, four, and eight periods (500 replications per cell), its $t$ test of slope 2 at the nominal 5 percent level rejected in 3.8 to 6.0 percent of replications.

Unlike in cross-sectional models, in which the coefficients are consistent whichever linear-exponential-family pseudo-likelihood, and so whichever variance function, is used (Gouriéroux et al. 1984, sec. 4), a departure here bears on consistency. The regression is descriptive, however: its residuals and fitted means share an estimated fixed effect, and the distribution of the relative residuals depends on the number of periods, so that a slope away from 2 can reflect features of the panel as well as the variance function. It concerns the gamma model rather than share centering: it rejected a slope of 2 in 93 to 98 percent of replications under lognormal errors whose variance was proportional to the squared mean, for which `xtgamfe` is consistent. It addresses the variance function, not serial dependence or dispersion that changes over time. A slope well away from 2 is a warning, but one near 2 does not establish share centering, and the two shape estimates check the gamma model in the same way.

More directly, both estimators should be fit to the same observations and both reported, and `estat fepoisson` tests whether their coefficients differ by a panel (or cluster) bootstrap. Its tests rejected a true null hypothesis somewhat too often when panel levels were very

dispersed: in 6 to 8.5 percent of replications at the nominal 5 percent level (Table 10) and in 0.9 to 2.8 percent at the nominal 1 percent level (Section 4.4). A $p$-value between about 0.01 and 0.05 is therefore weaker evidence than it appears, and with a few hundred such panels so is one somewhat below 0.01. With four periods their power was also limited against variances proportional to $\mu^{1.5}$ and $\mu^{2.5}$ (the joint test rejected in 22 and 59 percent of replications with 250 and 1,000 panels under the first, and in 11 percent under the second), and the absence of a difference does not establish share centering. A material difference calls for checking the mean specification, strict exogeneity, and share centering, and when the first two are credible, Poisson requires the weaker condition on how the errors of a panel relate to one another.

Head and Mayer (2014) advise the same pairing for pseudo-likelihood estimators of gravity equations, on the evidence of simulations of a cross-section, recommending that both be estimated, with least squares on logs, and that a large difference between them in large samples be read as a possible sign of misspecification. Weidner and Zylkin (2021, supplementary appendix, 30n11) note that the comparison of the Poisson and gamma pseudo-likelihood estimators is not straightforward in their three-way setting, because the gamma estimator is likely to be inconsistent there, whereas comparing it with least squares on logs still seems sensible to them. Here, too, the gamma estimator can be inconsistent, when share centering fails, so that this paper reads a difference as evidence against share centering, the mean specification, or strict exogeneity, without identifying which. Santos Silva and Tenreyro (2006), working with trade flows, prefer Poisson as a compromise, because the gamma pseudo-likelihood, efficient when the variance is proportional to the squared mean, may give excessive weight to the observations most prone to measurement error.

Period indicators should be included when the panel has a common time dimension, as additions to the model's own specification of the mean, which must still include any real period effects. In balanced panels with errors independent of the regressors, they absorb the share distortions from serial dependence and period-specific dispersion, at a cost in precision (a standard deviation 1.3 to 1.5 times as large in Section 4.3), but in the design of Table 5 not those from a variance proportional to another power of the mean. Because an analyst can rarely tell whether the indicators are doing that work, the average partial effects to report from a model with period indicators are those of `estat ape, ybar`. These

are consistent either way for regressors whose coefficients are consistent (not for the period contrasts themselves), whereas fitted-mean effects generally are not when the indicators absorb a pattern in the errors (Section 2.7). The difference in precision is small: a standard error of 0.84 against 0.80 in the example of Section 3.4, and, in the design of Section 4.8, standard deviations of the `ybar` effects 1.9 to 3.4 percent smaller than those of the fitted-mean effects. In unbalanced panels the pattern of observation should be described, because the choice of indicators depends on it.

Taken together, these recommendations amount to a short workflow:

1. Fit `xtgamfe`, with period indicators when the panel has a common time dimension, and in unbalanced panels describe the pattern of observation.
2. Fit fixed-effects Poisson regression to the same observations and report both; `estat fepoisson` tests the difference (in an unweighted fit), and a material difference calls for checking the mean specification, strict exogeneity, and share centering.
3. From a model with period indicators, report the average partial effects of `estat ape, ybar`.
4. Read the slope of `estat park` and a difference between the two shape estimates as warnings, not as tests of share centering.
5. With few clusters, use the `small` option or `vce(jackknife)` (Section 4.6); for effects on the dollar scale whose sign is not in doubt, consider the interval on the log scale (Section 4.8).

*Limitations.* First, the command covers one set of fixed effects, a log link, and strictly positive outcomes. It does not treat zeros, for which a two-part model and the Poisson estimator are among the alternatives, and dropping them changes what is estimated. If whether an outcome is positive depends on the same shock that determines its size, the estimates can be biased (Section 4.5), and no estimator run on the positive observations can repair that without further assumptions. Second, Proposition 2 and its extensions assume errors that are independent of the regressors (in unbalanced panels, within each pattern of observation), and weights that vary within panel can bias the coefficients even when the gamma model holds (Section 2.8).

Third, the estimator requires strict exogeneity, as fixed-effects Poisson does, and so does not accommodate lagged outcomes or regressors that respond to past outcomes; the illustration is a case in point. Generalized method of moments estimators that difference out a multiplicative fixed effect require the mean to be correctly specified only given current and past regressors, so they allow such feedback (Chamberlain 1992; Wooldridge 1997). Because they compare outcomes across consecutive periods, however, they assume observations are not selected on the outcome. Stata's `gmm` command fits them via a moment-evaluator program (StataCorp 2025, [R] gmm, Example 13). Fourth, no correction is applied for finite-$G$ bias (Section 2.3). Fifth, $t$ critical values and the jackknife brought coverage close to the nominal rate with 20 to 50 panels in the design of Section 4.6, but that design has well-behaved gamma errors, balanced panels, and independent panels. Sixth, the shape estimators and the model-based covariance matrix are valid under the gamma model or, more generally, when panels are independent and the share likelihood of Proposition 3 is correctly specified; only under the gamma model is the estimated shape that of each outcome's distribution. Seventh, intervals for effects on the level of the outcome, which average a level that can be highly skewed, covered the true value in as few as 86 percent of replications with 250 panels when the fixed effects were widely dispersed (Section 4.8). An interval on the log scale and a panel bootstrap-t narrowed the shortfall without closing it, and more panels remain the remedy.

Eighth, the command offers no direct test of share centering. The condition implies that, at the true coefficients, the relative residuals of every period have mean zero given the regressors, so that a conditional-moment test with functions of the regressors that the model omits would, like the comparison with Poisson, test the mean specification and share centering jointly, but without a second estimator; such a test is left for future work. Finally, the simulations cover particular designs. The test suite of version 1.1.1 has been run in StataNow 19.5 MP on Linux and Windows and in Stata 15.1 MP on Windows, and not in other releases or editions of Stata. In Stata/BE, where matrices are limited to 800 rows, `e(fe)` and `e(fe_id)` are stored for at most 800 panels.

## Acknowledgments

The author is grateful to the late Willard G. Manning for years of generous conversations about modeling health care costs, in person and in regular correspondence; his influence runs throughout this work. The author also thanks an anonymous referee for helpful comments. Any remaining errors are the author's.

## Generative AI disclosure

Generative AI tools (Anthropic's Claude, OpenAI's ChatGPT and Codex, Google's Gemini, and xAI's Grok) were used to develop and check the mathematical arguments, write and test the Stata and Mata code, design and run the simulations, search and check the literature, and draft and edit the text. The author directed the work, made the methodological decisions, and takes full responsibility for the paper and the software.

## Declarations

*Funding and conflicts of interest.* The author received no financial support for this work and has no conflicts of interest to declare.

*Ethics.* The study used publicly available, deidentified MEPS public-use files and involved no interaction with human participants.

*Data and code availability.* Version 1.1.1 of the command, which this paper describes, is available from the Statistical Software Components archive (`ssc install xtgamfe, all`, which also copies the example dataset); `which xtgamfe` shows the version installed. The replication materials (the test suite, the simulation drivers with their output, and the code that rebuilds the MEPS analysis files from the public sources) are available from the author on request. The MEPS source files are available from the Agency for Healthcare Research and Quality; person-level derived files are not redistributed.

## References

Agency for Healthcare Research and Quality. 2018. *MEPS HC-068: Multum Lexicon Addendum Files to MEPS Prescribed Medicines Files 1996–2013*. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-068.

Agency for Healthcare Research and Quality. 2021a. *Medical Expenditure Panel Survey, Household Component Panel Longitudinal Data Files, Panels 16–23 (HC-156, HC-164, HC-172, HC-183, HC-193, HC-202, HC-210, HC-217)*. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-217.

Agency for Healthcare Research and Quality. 2021b. *Medical Expenditure Panel Survey, Household Component Public-Use Files, 2011–2019: Full-Year Consolidated, Medical Conditions, Prescribed Medicines, and Medical Event Files*. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

Agency for Healthcare Research and Quality. 2026. *MEPS HC-036: 1996–2024 Pooled Linkage File for Common Variance Structure*. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-036.

Aitkin, Murray, and David Clayton. 1980. "The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data Using GLIM." *Applied Statistics* 29 (2): 156–63.

Bergé, Laurent R., Kyle Butts, and Grant McDermott. 2026. *fixest: A Fast and Feature-Rich Framework for Econometric Estimations in R*. arXiv:2601.21749v2. https://doi.org/10.48550/arXiv.2601.21749.

Blough, David K., Carolyn W. Madden, and Mark C. Hornbrook. 1999. "Modeling Risk Using Generalized Linear Models." *Journal of Health Economics* 18 (2): 153–71. https://doi.org/10.1016/S0167-6296(98)00032-0.

Boehm, Johannes, and contributors. 2025. *jmboehm/GLFixedEffectModels.jl: v0.5.5*. https://doi.org/10.5281/zenodo.15249924.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–72.

Carroll, Raymond J., and A. H. Welsh. 1988. "A Note on Asymmetry and Robustness in Linear Regression." *The American Statistician* 42 (4): 285–87. https://doi.org/10.2307/2685149.

Chamberlain, Gary. 1985. "Heterogeneity, Omitted Variable Bias, and Duration Dependence." In *Longitudinal Analysis of Labor Market Data*, edited by James J. Heckman and Burton Singer, 3–38. Cambridge University Press. https://doi.org/10.1017/CCOL0521304539.001.

Chamberlain, Gary. 1992. "Comment: Sequential Moment Restrictions in Panel Data." *Journal of Business & Economic Statistics* 10 (1): 20–26. https://doi.org/10.1080/07350015.1992.10509881.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin. 2020. "Fast Poisson Estimation with High-Dimensional Fixed Effects." *Stata Journal* 20 (1): 95–115. https://doi.org/10.1177/1536867X20909691.

Deb, Partha, Edward C. Norton, and Willard G. Manning. 2017. *Health Econometrics Using Stata*. Stata Press.

Dhaene, Geert, and Koen Jochmans. 2015. "Split-Panel Jackknife Estimation of Fixed-Effect Models." *Review of Economic Studies* 82 (3): 991–1030.

Dhaene, Geert, and Koen Jochmans. 2019. *Profile-Score Adjustments for Incidental-Parameter Problems*. Cambridge Working Papers in Economics 1959. Faculty of Economics, University of Cambridge. https://doi.org/10.17863/CAM.41229.

Duan, Naihua. 1983. "Smearing Estimate: A Nonparametric Retransformation Method." *Journal of the American Statistical Association* 78 (383): 605–10. https://doi.org/10.2307/2288126.

Gouriéroux, Christian, Alain Monfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52 (3): 681–700.

Greene, William. 2004. "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects." *Econometrics Journal* 7 (1): 98–119.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents–R&D Relationship." *Econometrica* 52 (4): 909–38. https://doi.org/10.2307/1911191.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse, Toolkit, and Cookbook." In *Handbook of International Economics, Volume 4*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 131–95. Elsevier. https://doi.org/10.1016/b978-0-444-54314-1.00003-3.

Hinz, Julian, Alexander Hudlet, and Joschka Wanner. 2019. *Separating the Wheat from the Chaff: Fast Estimation of GLMs with High-Dimensional Fixed Effects*. Unpublished technical note, dated 21 March 2019. https://github.com/julianhinz/R_glmhdfe/blob/master/resources/glmhdfe_technical_note.pdf.

Holford, Theodore R. 1980. "The Analysis of Rates and of Survivorship Using Log-Linear Models." *Biometrics* 36 (2): 299–305. https://doi.org/10.2307/2529982.

Laird, Nan, and Donald Olivier. 1981. "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques." *Journal of the American Statistical Association* 76 (374): 231–40. https://doi.org/10.2307/2287816.

Lancaster, Tony. 2000. "The Incidental Parameter Problem Since 1948." *Journal of Econometrics* 95 (2): 391–413.

Manning, Willard G. 1998. "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem." *Journal of Health Economics* 17 (3): 283–95. https://doi.org/10.1016/S0167-6296(98)00025-3.

Manning, Willard G., and John Mullahy. 2001. "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics* 20 (4): 461–94. https://doi.org/10.1016/S0167-6296(01)00086-8.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. 2019. *Robust Statistics: Theory and Methods (with R)*. 2nd ed. Wiley.

Martin, Robert S. 2017a. "Estimation of Average Marginal Effects in Multiplicative Unobserved Effects Panel Models." *Economics Letters* 160: 16–19. https://doi.org/10.1016/j.econlet.2017.08.020.

Martin, Robert S. 2017b. *Estimation of Average Marginal Effects in Multiplicative Unobserved Effects Panel Models*. BLS Working Paper 497. U.S. Bureau of Labor Statistics. https://www.bls.gov/osmr/research-papers/2017/pdf/ec170050.pdf.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Chapman and Hall.

Monti, G. S., G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue. 2011. "The Shifted-Scaled Dirichlet Distribution in the Simplex." In *Proceedings of the 4th International Workshop on Compositional Data Analysis (CoDaWork 2011)*, edited by J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego. CIMNE. https://hdl.handle.net/10256/13637.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1): 69–85. https://doi.org/10.2307/1913646.

Newey, Whitney K., and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics, Volume 4*, edited by Robert F. Engle and Daniel L. McFadden, 2111–245. Elsevier. https://doi.org/10.1016/s1573-4412(05)80005-4.

Neyman, Jerzy, and Elizabeth L. Scott. 1948. "Consistent Estimates Based on Partially Consistent Observations." *Econometrica* 16 (1): 1–32. https://doi.org/10.2307/1914288.

Park, R. E. 1966. "Estimation with Heteroscedastic Error Terms." *Econometrica* 34 (4): 888.

Roebuck, M. Christopher, J. Samantha Dougherty, Robert Kaestner, and Laura M. Miller. 2015. "Increased Use of Prescription Drugs Reduces Medical Costs in Medicaid Populations." *Health Affairs* 34 (9): 1586–93. https://doi.org/10.1377/hlthaff.2015.0335.

Santos Silva, J. M. C., and Gordon C. R. Kemp. 2016. *Partial Effects in Fixed-Effects Models*. Presentation at the 22nd London Stata Users Group Meeting, 8 September 2016. https://www.stata.com/meeting/uk16/slides/santos_uk16.pdf.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88 (4): 641–58.

Stammann, Amrei, and Daniel Czarnowske. 2025. *alpaca: Fit GLM's with High-Dimensional k-Way Fixed Effects*. R package version 0.3.5. https://doi.org/10.32614/CRAN.package.alpaca.

StataCorp. 2025. *Stata 19 Base Reference Manual*. Stata Press.

Sun, Yutao, and Geert Dhaene. 2019. "xtspj: A Command for Split-Panel Jackknife Estimation." *Stata Journal* 19 (2): 335–74. https://doi.org/10.1177/1536867X19854016.

Weidner, Martin, and Thomas Zylkin. 2021. "Bias and Consistency in Three-Way Gravity Models." *Journal of International Economics* 132: 103513. https://doi.org/10.1016/j.jinteco.2021.103513.

Wooldridge, Jeffrey M. 1997. "Multiplicative Panel Data Models Without the Strict Exogeneity Assumption." *Econometric Theory* 13 (5): 667–78. https://doi.org/10.1017/S0266466600006125.

Wooldridge, Jeffrey M. 1999. "Distribution-Free Estimation of Some Nonlinear Panel Data Models." *Journal of Econometrics* 90 (1): 77–97. https://doi.org/10.1016/S0304-4076(98)00033-5.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press.

# Appendix A: Proofs and derivations

## A.1 Regularity conditions and the proof of Proposition 2

The population criterion is defined by subtracting each panel's criterion at the true coefficients before taking expectations; convexity and a zero-mean score then identify its minimum. One sufficient set of conditions for the sampling results is the following:

(i) the clusters (the panels, or the groups of panels on which the standard errors are clustered) are independent draws from a common distribution, each with a bounded number of observations;

(ii) the regressors have a finite second moment;

(iii) the expected score is zero at the limit point (the true coefficients under share centering, and under Proposition 2 the true coefficients of the other regressors together with the true period coefficients plus $\boldsymbol{\delta}^*$); and

(iv) the expected Hessian there is nonsingular.

Because every share lies between zero and one, the score of panel $i$ is bounded by $(T_i - 1)\sum_t \|\mathbf{x}_{it}\|$ and its Hessian by $T_i \max_{s,t} \|\mathbf{x}_{it} - \mathbf{x}_{is}\|^2$ at any $\boldsymbol{\beta}$, so that the cluster totals are bounded by sums of a bounded number of such terms. These conditions thus deliver the laws of large numbers, the score moments, and the dominance of the Hessian that consistency and asymptotic normality require, with no moment condition on the errors. With weights constant within panel ($a_{it} = a_i$, so that $A_i = T_i a_i$; Section 2.8), the score and Hessian of panel $i$ are $a_i$ times their unweighted counterparts and are therefore bounded by $A_i \sum_t \|\mathbf{x}_{it}\|$ and $A_i \max_{s,t} \|\mathbf{x}_{it} - \mathbf{x}_{is}\|^2$. Condition (ii) is then replaced by a finite second moment of the first bound and a finite mean of the second, with the weights included in each cluster's draw in (i).

With clusters of unbounded size, one must instead assume a law of large numbers for the cluster Hessian totals and a central limit theorem with finite second moments for the cluster score totals. Valid clustered standard errors also require a law of large numbers, under the same normalization, for the outer products of the score totals; the central limit theorem does not imply this, because a single dominant cluster can leave the estimated covariance random in the limit.

For consistency, average partial effects additionally require the conditions of Section 2.7 on the averaged quantities: a finite mean, a dominance condition in a neighborhood of the limiting coefficients, and errors with mean one on the periods averaged. In a weighted fit, these conditions apply to the weight-multiplied quantities, and the cluster weight total must have a finite positive mean. For valid standard errors, the cluster totals of the influence function (Appendix A.5) must have finite second moments.

*Proof of Proposition 2.* Expectations are over a panel drawn at random from the population of panels; when the sampling units are clusters of panels, the population criterion (the expected cluster total divided by the expected number of panels per cluster) is that same expectation. Independence is therefore required for such a draw, because a single vector of period coefficients absorbs only the period pattern of the errors averaged over panels. Cluster composition is one possible source of dependence: a characteristic that shifts the error distribution, such as a panel's position within its cluster, can make the errors depend on the regressors when it is related to them.

Let $\boldsymbol{\delta}$ denote the period coefficients measured as deviations from their true values (with $\delta_t = 0$ for the omitted period), and let $W_{it}(\boldsymbol{\delta}) = \varepsilon_{it}e^{-\delta_t} / \sum_s \varepsilon_{is}e^{-\delta_s}$ be the shares entering the score when the other coefficients are at their true values. Measured from its value at $\boldsymbol{\delta} = \mathbf{0}$, the population criterion in $\boldsymbol{\delta}$ is $\mathrm{E}[\,T \ln \sum_t W_{it}(\mathbf{0})\, e^{-\delta_t}] + \sum_t \delta_t$. This criterion is finite (the log sum lies between $-\max_t \delta_t$ and $-\min_t \delta_t$) and strictly convex on the free period coefficients. Along any nonzero direction $\boldsymbol{v}$ with zero in the omitted period, its recession slope is $T(\bar{v} - \min_t v_t) > 0$, so it is coercive and has a unique minimizer $\boldsymbol{\delta}^*$.

Because the integrand's derivative with respect to $\delta_t$ is bounded by $T - 1$, differentiation under the expectation gives $\mathrm{E}[1 - T\, W_{it}(\boldsymbol{\delta})]$. At the minimizer, $\mathrm{E}[1 - T\, W_{it}(\boldsymbol{\delta}^*)] = 0$ for every period with an indicator, and for the omitted period as well because these terms sum to zero over $t$. For any other regressor $x_{it,j}$, the score is $\sum_t x_{it,j}\,(1 - T\, W_{it}(\boldsymbol{\delta}^*))$; its integrand derivative is bounded by $(T-1)\sum_t \|\mathbf{x}_{it}\|$, which has a finite mean, again permitting the interchange. Because $W_{it}(\boldsymbol{\delta}^*)$ depends only on the errors, it is independent of $x_{it,j}$, so each term has expectation $\mathrm{E}[x_{it,j}]\,\mathrm{E}[1 - T\, W_{it}(\boldsymbol{\delta}^*)] = 0$. The gradient of the population criterion is therefore zero at this point, and strict convexity under the identification condition rules out any other minimizer. □

### A.2 Unbalanced panels

Let $S_i$ denote the set of periods in which panel $i$ is observed, which takes values in a fixed finite set of patterns, and suppose that, given $S_i$, the observed errors are independent of $\mathbf{X}_i$. The argument uses only this independence within each pattern, so that the distribution of the errors may differ between patterns. For the first extension of Proposition 2, suppose that the period means of the regressors do not depend on the pattern of observation apart from a constant, $\mathrm{E}[\mathbf{x}_{it} \mid S_i] = \mathbf{m}_t + \mathbf{c}(S_i)$. With calendar-period indicators, the first-order condition of period $t$ is then

$$\sum_{S \ni t} \Pr(S)\, \mathrm{E}\Big[1 - |S|\, W_{it}(\boldsymbol{\delta}^*) \mid S\Big] = 0,$$

and the expected score of a regressor is

$$\sum_{S} \Pr(S) \sum_{t \in S} \Big(\mathbf{m}_t + \mathbf{c}(S)\Big)\, \mathrm{E}\Big[1 - |S|\, W_{it}(\boldsymbol{\delta}^*) \mid S\Big],$$

in which the part in $\mathbf{c}(S)$ vanishes because the terms sum to zero within a pattern and the part in $\mathbf{m}_t$ is a weighted sum of the first-order conditions. The first-order condition of the omitted period follows from the others, because the terms sum to zero within each pattern. The calendar-period indicators must themselves be identified, which requires the patterns of observation to connect the periods (with only the patterns $\{1, 2\}$ and $\{3, 4\}$, the fixed effects absorb the difference between the two pairs of periods). The criterion in $\boldsymbol{\delta}$ is then a mixture over patterns of criteria of the form used in Appendix A.1, and $\boldsymbol{\delta}^*$ exists and is unique as there. For the second extension, with an indicator for every period within every pattern, Proposition 2 applies pattern by pattern. Indicators for position by length pool the patterns of each length; the argument then holds length by length, with the first-order conditions of each pattern satisfied separately, when the observed errors have the same distribution in every pattern of that length, as when panels are observed for consecutive periods and the error process is stationary and independent of where the window falls.

### A.3 Proof of Proposition 3

*Proof of Proposition 3.* The mean-scaled outcomes $y_{it}q_{it}$ are independent gamma variates with shape $\kappa$ and a common scale, so their shares $w_{it}$ have the symmetric Dirichlet density $\Gamma(T_i\kappa)\Gamma(\kappa)^{-T_i}\prod_t w_{it}^{\kappa-1}$, which is free of $\alpha_i$. The observed shares are related to them by $w_{it} = s_{it}q_{it}/\sum_r s_{ir}q_{ir}$, a transformation with Jacobian $\prod_t q_{it}/(\sum_t s_{it}q_{it})^{T_i}$. □

The shares of the mean-scaled outcomes are symmetric Dirichlet; the observed shares generally are not, and are symmetric Dirichlet only when the factors $q_{it}$ are equal within the panel. The density of the observed shares is that of normalized independent gamma variates with a common shape and unequal scales, a special case of what the literature on compositional data calls the scaled Dirichlet distribution (Monti et al. 2011, Property 3.7).

### A.4 The shape estimators

At the true coefficients $\rho_{it} = T_iW_{it} - 1$. Under the gamma model each $W_{it}$ has a beta distribution with parameters $\kappa$ and $(T_i - 1)\kappa$, and so $\mathrm{E}[\sum_t \rho_{it}^2] = T_i(T_i - 1)/(\kappa T_i + 1)$. When a finite positive solution exists, the moment estimate `e(kappa)` solves

$$\sum_{i,t} \rho_{it}^2 = \sum_i \frac{T_i(T_i - 1)}{\kappa T_i + 1},$$

whose right-hand side is strictly decreasing in $\kappa$.

By Proposition 3 the log likelihood of the shares at $\hat{\boldsymbol{\beta}}$ is, apart from a constant, $\sum_i[\ln\Gamma(T_i\kappa) - T_i\ln\Gamma(\kappa) + \kappa\Lambda_i(\hat{\boldsymbol{\beta}})]$, which is strictly concave in $\kappa$. When a finite maximizer exists, it is `e(kappa_ml)`; when the fitted shares are all equal, there is none (Section 2.6). Because the score for $\boldsymbol{\beta}$ is zero at $\hat{\boldsymbol{\beta}}$ whatever the value of $\kappa$, the mixed derivative is zero there (its expectation, minus the expected score, is zero at the limit of $\hat{\boldsymbol{\beta}}$). As a result, the linearization of the shape estimate does not involve the estimation of the coefficients, and its standard error follows from the derivatives in $\kappa$ alone, by default as a sandwich of the panels' shape scores (Section 2.6). Under the gamma model with independent panels the coefficient estimates and the maximum likelihood shape estimate are, moreover, uncorrelated to first order.

### A.5 The standard error of average partial effects

Writing $\hat{\theta} = n^{-1}\sum_{it} f_{it}(\hat{\boldsymbol{\beta}})$, the contribution of panel $i$ to the influence function of $\hat{\theta}$ is

$$\frac{1}{n}\sum_{t}\left(f_{it} - \hat{\theta}\right) \;-\; \mathbf{J}\,\mathbf{H}^{-1}\mathbf{g}_i,$$

where $\mathbf{J}$ is the derivative of $\hat{\theta}$ with respect to the coefficients and $\mathbf{H} = \sum_i \mathbf{H}_i$. The variance is the sum over clusters of the squared cluster totals of these contributions, multiplied by $M/(M-1)$, with $M$ the number of clusters used for the effect, counted as in Section 3.2; the factor is that of the sandwich of the coefficients when the same clusters are used there. With weights, the averages are weighted, the sum of the weights replaces $n$, and the contribution of panel $i$ becomes $\sum_t a_{it}(f_{it} - \hat{\theta})/\sum_r A_r - \mathbf{J}\,\mathbf{H}^{-1}\mathbf{g}_i$, with $a_{it}$ and $A_r$ as in Section 2.8 and $\mathbf{g}_i$ and $\mathbf{H}$ the score and Hessian of the weighted criterion there. With frequency weights, write $u_i$ for this contribution of a stored panel and $m_i$ for its weight. In the expanded data the panel is $m_i$ copies, each with the unweighted score $\mathbf{g}_i/m_i$ and the same $\mathbf{H}$, so that each copy contributes $u_i/m_i$. With the standard errors clustered on panels, each copy is a cluster, the sum of squared cluster totals is $\sum_i u_i^2/m_i$, and $M = \sum_i m_i$; with clusters above panels, the copies stay in the cluster of their panel, so that the cluster total is the sum of the $u_i$ of its stored panels and $M$ is the number of those clusters. Three choices of $f_{it}$ are used in this paper.

*Effects from fitted means.* For the average partial effect of Section 2.7, $f_{it} = \hat{\beta}_j\hat{\mu}_{it}$ for a regressor that enters linearly, and $\mathbf{J}$ is computed with the fixed effects re-profiled, so that $\partial\hat{\mu}_{it}/\partial\boldsymbol{\beta} = \hat{\mu}_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)$. For fixed-effects Poisson, Martin (2017b, Theorem 2) gives the corresponding asymptotic variance.

*Effects from the observed outcome.* For `estat ape, ybar`, $f_{it}$ is $y_{it}$ times the derivative of the index, or $y_{it}\,[\exp(\eta_{it}^{\ell} - \eta_{it}) - \exp(\eta_{it}^{0} - \eta_{it})]$ for a factor level (Section 2.7). Because $y_{it}$ does not depend on $\boldsymbol{\beta}$, $\mathbf{J}$ has no re-profiling term. For a regressor that enters linearly, $f_{it} = \hat{\beta}_j y_{it}$ and $\mathbf{J} = \bar{y}\,\mathbf{e}_j'$, where $\mathbf{e}_j$ selects coefficient $j$. These averages are consistent under the conditions of Section 2.7: errors with mean one on the periods averaged and, for changes other than the derivative of a term that enters linearly, a bound on the absolute value of the outcome times the proportional change that has a finite mean in a neighborhood of the

limiting coefficients.

*The finite change.* For the change in the average fitted mean when a regressor that enters linearly rises by $h$ units, write $c = e^{\hat{\beta}_j h} - 1$ and $m = n^{-1}\sum_{it}\hat{\mu}_{it}$. The change is $cm$, with $f_{it} = c\,\hat{\mu}_{it}$, and by the product rule $\mathbf{J} = c\,\partial m/\partial\boldsymbol{\beta}' + h\,e^{\hat{\beta}_j h}\,m\,\mathbf{e}_j'$, with $\partial m/\partial\boldsymbol{\beta}' = n^{-1}\sum_{it}\hat{\mu}_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)'$. `estat ape, finite` reports it, with $h = 1$ unless `delta(#)` gives another value, and evaluates $c$ without cancellation, so that, as $h \to 0$, the change divided by $h$ tends to the derivative, and its standard error divided by $|h|$ to the derivative's standard error; Section 5.2 gives an example in dollars. With `ybar` as well, the observed outcome replaces $\hat{\mu}_{it}$, so that the change is $c\,\bar{y}$ and $\mathbf{J} = h\,e^{\hat{\beta}_j h}\,\bar{y}\,\mathbf{e}_j'$.

### A.6 Information-based standard errors of the exponential model under gamma errors

The exponential model is the gamma model with $\kappa = 1$, and the information for $\boldsymbol{\beta}$ in its log likelihood, profiled over the fixed effects, is $\mathbf{H} = \sum_i \mathbf{H}_i$. Under the gamma model with shape $\kappa$, the part of the log likelihood of the shares of panel $i$ that depends on $\boldsymbol{\beta}$ is $-\kappa\,Q_i^p(\boldsymbol{\beta})$ (Proposition 3), whose score is $-\kappa\,\mathbf{g}_i$ and whose Hessian is $-\kappa\,\mathbf{H}_i$. The information equality of that likelihood therefore gives, at the true coefficients and for any $T_i$,

$$\mathrm{Var}(\mathbf{g}_i \mid \mathbf{X}_i) = \frac{1}{\kappa}\,\mathrm{E}[\mathbf{H}_i \mid \mathbf{X}_i].$$

In large samples the sandwich, which is the correct covariance matrix, is therefore $(\kappa\sum_i \mathrm{E}[\mathbf{H}_i])^{-1}$, whereas the inverse information of the exponential model is $(\sum_i \mathrm{E}[\mathbf{H}_i])^{-1}$. The second is $\kappa$ times the first, so that the standard errors of the exponential model are $\sqrt{\kappa}$ times the correct ones. (In a balanced panel, the moments of the symmetric Dirichlet distribution give $\mathrm{E}[\mathbf{H}_i \mid \mathbf{X}_i] = \{T\kappa/(T\kappa+1)\}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'$, with $\bar{\mathbf{x}}_i$ the unweighted panel mean.) The model-based covariance matrix of `xtgamfe`, $(\hat{\kappa}\mathbf{H})^{-1}$, allows for the shape and is consistent under the gamma model.

### A.7 The second-order approximation

Expectations, variances, and covariances of the errors are conditional on $(\mathbf{X}_i, \alpha_i)$, and the errors are assumed to have finite variances. Let $e_{it} = \varepsilon_{it} - 1$ and $\bar{e}_i = T_i^{-1}\sum_s e_{is}$, so that

$T_i W_{it} = (1 + e_{it})/(1 + \bar{e}_i)$. The identity $1/(1+u) = 1 - u + u^2/(1+u)$ gives, exactly,

$$T_i W_{it} - 1 = (e_{it} - \bar{e}_i) - \bar{e}_i\,(e_{it} - \bar{e}_i) + \bar{e}_i^2\,(T_i W_{it} - 1).$$

The first term has mean zero, and because $\mathrm{E}[\bar{e}_i e_{it}] = \omega_{it}/T_i$ and $\mathrm{E}[\bar{e}_i^2] = \bar{\omega}_i/T_i$, the second has mean $-(\omega_{it} - \bar{\omega}_i)/T_i$. This is (2), exactly, with $r_{it} = \mathrm{E}[\bar{e}_i^2(T_i W_{it} - 1)]$. Because $T_i W_{it} - 1 = (e_{it} - \bar{e}_i)/(1 + \bar{e}_i)$ lies between $-1$ and $T_i - 1$, bounding it by $2|e_{it} - \bar{e}_i|$ when $\bar{e}_i \geq -1/2$ and by $T_i - 1$ otherwise gives $|r_{it}| \leq 2\,\mathrm{E}[\bar{e}_i^2 |e_{it} - \bar{e}_i|] + 2(T_i - 1)\,\mathrm{E}|\bar{e}_i|^3$. If the deviations are $e_{it} = \sigma z_{it}$, with the $z_{it}$ bounded below and having finite third absolute moments, the second term of (2) is therefore of order $\sigma^2$ and $r_{it}$ of order $\sigma^3$. The leading part of $r_{it}$, $\mathrm{E}[\bar{e}_i^2(e_{it} - \bar{e}_i)]$, involves third moments; for errors independent across periods it equals $\{\mathrm{E}[e_{it}^3] - T_i^{-1}\sum_s \mathrm{E}[e_{is}^3]\}/T_i^2$, and, when such errors are also bounded above and away from zero uniformly in $T_i$, $r_{it}$ is of order $1/T_i^2$, against $1/T_i$ for the second term.

Because $\sum_t(\omega_{it} - \bar{\omega}_i)$ and $\sum_t r_{it}$ are zero, (2) gives the conditional mean of the score at the true coefficients as $T_i^{-1}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\,\omega_{it} - \sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\,r_{it}$. The limit $\boldsymbol{\beta}^*$ solves $\mathrm{E}[\mathbf{g}_i(\boldsymbol{\beta}^*)] = \mathbf{0}$, and because the derivative of $\mathbf{g}_i$ is $\mathbf{H}_i$, $\boldsymbol{\beta}^* - \boldsymbol{\beta}_0 = -\bar{\mathbf{H}}^{-1}\,\mathrm{E}[\mathbf{g}_i(\boldsymbol{\beta}_0)]$, where $\bar{\mathbf{H}}$ is the average of $\mathrm{E}[\mathbf{H}_i]$ along the segment from $\boldsymbol{\beta}_0$ to $\boldsymbol{\beta}^*$. At $\boldsymbol{\beta}_0$ the weights of $\mathbf{H}_i$ are the shares $W_{it}$, which equal $1/T_i$ plus terms whose means, and the means of whose products, are of second order, so that $\mathrm{E}[\mathbf{H}_i(\boldsymbol{\beta}_0)] = \mathrm{E}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' + O(\sigma^2)$, and $\boldsymbol{\beta}^* - \boldsymbol{\beta}_0$ is itself of order $\sigma^2$. Under the usual regularity conditions (in a balanced panel, for example, a nonsingular matrix $\mathrm{E}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'$ and a finite expectation of $(1 + \max_t\|\mathbf{x}_{it}\|^3)(1 + \max_t|z_{it}|^3)$ for the family $e_{it} = \sigma z_{it}$ above), therefore,

$$\boldsymbol{\beta}^* - \boldsymbol{\beta}_0 = -\Big\{\mathrm{E}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\Big\}^{-1}\mathrm{E}\Big[\frac{1}{T_i}\sum_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\,\omega_{it}\Big] + O(\sigma^3),$$

which is (3) in a balanced panel.

Three cases have closed forms. First, suppose that $(\alpha_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ is jointly normal with moments that are the same in every period, that there is no offset, and that the errors are independent across periods with $\omega_{it} = \lambda(\ln \mu_{it})$. Stein's identity, $\mathrm{E}[Z\,\lambda(V)] = \mathrm{Cov}(Z, V)\,\mathrm{E}[\lambda'(V)]$ for jointly normal $(Z, V)$ with $\mathrm{E}[Z] = 0$, applied with $Z$ an element of $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$ (which is uncorrelated with $\alpha_i$) and $V = \ln \mu_{it}$, gives $\boldsymbol{\pi} = \mathrm{E}[\lambda'(\ln \mu_{it})]\,\boldsymbol{\beta}_0$: to this

order every coefficient is multiplied by the same factor, $1 - \mathrm{E}[\lambda'(\ln \mu_{it})]/T$. With an outcome variance of $\phi\mu_{it}^{p}$, $\lambda(v) = \phi\, e^{(p-2)v}$ and the factor is $1 - (p-2)\,\phi\,\mathrm{E}[\mu_{it}^{p-2}]/T$. In the design of Table 5, in which $\ln \mu_{it}$ has variance 1.84, $\phi\,\mathrm{E}[\mu_{it}^{p-2}] = 0.5\, e^{0.92(p-2)^2}$, which is 0.63 with $p$ of 1.5 or 2.5 and 1.25 with $p$ of 1 or 3.

Second, suppose that the errors are stationary and independent of the regressors and the fixed effect, with autocovariance $\tau_k$ at lag $k$, so that $\omega_{it} = \sum_s \tau_{|t-s|}$, and that a binary regressor is one only in the last period of treated panels, the other regressors being independent of it and of the errors, with mean zero. Then (3) gives the bias of its coefficient as $\sum_{k=1}^{T-1}(T - 2k)\,\tau_k/\{T(T-1)\}$, which is zero when $T = 2$, positive when the autocovariances are positive and decrease with the lag, and, if the autocovariances are absolutely summable, close to $\sum_{k\geq 1}\tau_k/T$ when $T$ is large. For a regressor that is one from the second period on, the sign reverses, because stationarity makes $\omega_{i1} = \omega_{iT}$.

Third, suppose that the errors are independent across periods and of the regressors and the fixed effect, with variances $\omega_t$ that differ by period, and that a binary regressor is defined as in the second case. Then (3) gives the bias of its coefficient as $(\bar{\omega} - \omega_T)/(T-1)$ when it is one only in the last period and $(\omega_1 - \bar{\omega})/(T-1)$ when it is one from the second period on, with $\bar{\omega}$ the mean of the $\omega_t$. When $\omega_{it} = \omega_t$ for every panel, the panel is balanced, and the model includes period indicators, the within regression in (3) assigns $\omega_t - \omega_1$ to the indicator of period $t$ and nothing to the other regressors, so that the period coefficients are shifted by approximately $-(\omega_t - \omega_1)/T$, the second-order counterpart of $\boldsymbol{\delta}^*$ in Proposition 2.

The direction of the bias in the design of Table 5 (Section 4.3) rests on an exact property rather than on the expansion. Let the errors of a panel be independent gamma variates with mean one and shapes $\kappa_t$, given $(\mathbf{X}_i, \alpha_i)$. Writing $1/\sum_s \varepsilon_{is} = \int_0^\infty \exp(-u\sum_s \varepsilon_{is})\,du$ and using $\mathrm{E}[e^{-u\varepsilon_{is}}] = (1 + u/\kappa_s)^{-\kappa_s}$ and $\mathrm{E}[\varepsilon_{it}e^{-u\varepsilon_{it}}] = (1 + u/\kappa_t)^{-\kappa_t - 1}$ gives

$$\mathrm{E}[\,W_{it} \mid \mathbf{X}_i, \alpha_i\,] = \int_0^\infty \frac{1}{1 + u/\kappa_t} \prod_s \Big(1 + \frac{u}{\kappa_s}\Big)^{-\kappa_s} du,$$

in which only the first factor depends on $t$, and it increases with $\kappa_t$: for any number of periods, the error with the larger shape has the larger expected share.

## Appendix B: Details of the MEPS illustration

*Cohorts and codes.* The diabetes cohort enters on the MEPS diabetes flag, a diabetes condition record, or an antidiabetic fill, and the antipsychotic cohort on a condition record coded to schizophrenia or another psychotic disorder or to bipolar disorder, or on an antipsychotic fill; a person must be an adult in every year observed, and evidence from either year qualifies. The public files give diagnoses at three digits. The diabetes records are those with the ICD-9 code 250, the ICD-10 codes E10 to E13 (leaving out E08 and E09, diabetes due to an underlying condition or to a drug), or the grouped codes CCS 49 and 50 and CCSR END002 and END003. The antipsychotic cohort's condition records are those with the ICD-9 code 295, the ICD-10 codes F20, F25, F30, or F31, or the grouped codes for schizophrenia and other psychotic disorders (CCS 659; CCSR MBD001) and for bipolar and related disorders (CCSR MBD003), which are broader than these ICD codes and may carry information that the public ICD field suppresses. The ICD-9 code 296, which then covers major depression as well as bipolar disorder, is not used, so that entry for bipolar disorder before 2016 runs through a fill. Of the 918 persons with diagnosis evidence in the antipsychotic cohort, 133 qualify through the grouped codes alone.

*Classification of fills.* A fill counts toward a cohort's medications when one of its Multum therapeutic classes is 99, 215, 309, or 314 (diabetes) or 77, 210, 251, or 341 (antipsychotics), or when its drug name contains the generic name of one of the cohort's drugs; supplies and equipment are excluded. It counts toward the insulin or atypical-antipsychotic subclass when, in addition, its class is 215 or 341 or its name is that of an insulin or an atypical antipsychotic. Every drug name that the class route alone admits was reviewed. Lithium (class 77; 1,745 fills) and chlordiazepoxide with amitriptyline (class 251; 2 fills), which are not antipsychotics, are excluded, and so, by a coding decision, are prochlorperazine and the other phenothiazines classified as antiemetics (class 210; 1,248 fills), whereas brand names of antipsychotics and combinations that contain one, such as olanzapine with fluoxetine, which counts toward the atypical subclass, are kept. Insulin syringes, pen needles, pumps, pods, and other delivery devices, which the name insulin would admit and the general exclusion of supplies misses, are excluded (4,975 fills). Fills reported as NovoLog or insulin aspart whose Multum name

is sterile water (421 fills, all in 2013) count as insulin, as their corrected class does, and fills reported as metformin whose Multum name is propranolol (3,648 fills, 2013–2019) are classified by the reported name. For 2011–2013 the classes are AHRQ's corrected Multum classes (file HC-068, Agency for Healthcare Research and Quality 2018), which AHRQ issued because the original files assigned some classes incorrectly (Agency for Healthcare Research and Quality 2021b); in these data the correction changes only the insulin indicator of 928 fills of insulin aspart, and no fill enters or leaves either classification. Of the classified fills with a missing-data code for days supplied, more than 99.7 percent are coded "don't know." A person with such a fill and no valid fill of the condition's medications in the panel (31 and 50 percent of these persons in the two estimation samples) has the fill counted at 30 days in rows 1–6 and 10–11 of Table 15.

*Round length.* The public files give the month and year in which each round begins and ends. A round 1 beginning in January of the panel's first year is dated from 1 January, and a round 5 ending in December of its second year is dated to 31 December, as fixed by the survey design. In the 2011 and 2012 files, which report the exact day, 99.8 percent of such round-1 beginnings (panels 16 and 17) fall on 1 January, and 99.9 percent of such round-5 ends (panel 16, whose second year is 2012) fall on 31 December. Every other date is taken at the 15th of its month, and the length is the difference in days. Round 3, which both annual files of a panel report, takes the second-year file's dates, or the first-year file's where the second-year file lacks one. For a person with no second-year file, round 3 ends on 31 December of the panel's first year when the first-year file puts its end in the second year, because the second-year part of the round is not observed (313 and 62 of the estimation rounds). A round has no computable length and leaves the sample if it lacks a begin or end date, or if it begins and ends in the same month without a fixed date (1 January, 31 December, or the 31 December ending a cut round 3). Among otherwise qualifying rounds, 2 in the diabetes cohort and 1 in the antipsychotic cohort lack a begin or end date, and 16 and 8 have no length for the second reason.

*Sample flow.* Of the 67,438 rounds of the diabetes cohort on the grid of rounds, 53,570 have an observed prior round, 52,883 also have prior-round self-rated health and a round length and belong to persons with at least two such rounds, and 42,368 remain once spending

must be positive in at least two rounds per person (12,225, 9,722, 9,581, and 8,230 in the antipsychotic cohort).

*Clustering.* Row 2 of Table 15 clusters on households (`vce(cluster hh_id)`). Clustering instead on the MEPS primary sampling unit within its stratum gives standard errors of 0.0016 and 0.0072 across 621 and 545 units. These units come from AHRQ's pooled linkage file for variance estimation (Agency for Healthcare Research and Quality 2026), which assigns every person to one such unit, and in these data every dwelling unit lies within one. Most households contribute only one sample member (80 and 92 percent of persons are the only one in theirs), although that share does not by itself determine how much the clustering matters.

*Round indicators.* With the person effects and calendar-year indicators in the model, the indicators for the round of the panel are separated from a linear calendar trend only by the 373 and 71 third rounds that end in a panel's first year. The test reported for row 3 in Section 5.2 uses only the two contrasts that do not rely on those rounds (rounds 4 and 5 against round 3). In the antipsychotic cohort, with 71 such rounds, the coefficients on the indicators of rounds 3, 4, and 5 are −0.0548, −0.0633, and −0.0634, each with a standard error of about 0.21, and the two contrasts are −0.0084 and −0.0086 ($\chi^2(2) = 0.1$).

*Cohort definition.* Among persons with diagnosis evidence (a condition record with the selected ICD codes or the grouped codes or, for diabetes, the MEPS diabetes flag), the main model gives 0.0010 (0.0016) for 11,507 persons with diabetes and −0.0087 (0.0107) for 918 persons in the antipsychotic cohort. For the 785 of these 918 who have one of the named ICD codes it gives −0.0045 (0.0113); both antipsychotic estimates are larger in magnitude and less precise than that of row 1.

*The exposure as a rate.* Dividing days supplied by the length of the prior round gives days supplied per day of that round, summed over fills and medications, so that it can exceed one (as it does in 13,870 and 1,137 of the rounds used) and is not a measure of adherence. The prior length is unknown in 8 of the diabetes rounds and 4 of the antipsychotic rounds. Dropping those rounds and any persons left with a single round leaves 42,358 and 8,224 rounds. On that sample, the rate gives coefficients of −0.011 (0.009) for diabetes and −0.006 (0.042) for antipsychotics per unit, compared with 0.0009 (0.0016) and −0.0016 (0.0075)

for days supplied in 30-day units. The diabetes coefficient changes sign and lies within 1.3 standard errors of zero, while the antipsychotic coefficient keeps its sign and lies about 0.1 standard errors from zero.

## About the author

M. Christopher Roebuck is a health economist and the president of RxEconomics LLC in Miami, Florida. His research focuses on medication adherence, pharmacy benefit design, and the use of real-world data to measure the value of medical and pharmaceutical interventions.