

Panel generalized linear models with panglm

Introduction

Longitudinal and panel data contain repeated observations for each sampling unit. Dependence within units must therefore be represented either through unit-specific effects or through an appropriate covariance estimator. The **panglm** package provides a common interface for pooled, fixed-effects, and random-effects generalized linear models for continuous, binary, and count outcomes.

This vignette defines the models fitted by the package, describes their estimating equations, and illustrates the associated inferential methods. The emphasis is on the statistical specification of each estimator. The numerical routines use Rcpp, RcppArmadillo, and RcppParallel.

```
library(panglm)
data(Grunfeld, package = "p1m")
```

Notation and data organization

Let y_{it} denote the response for unit i at occasion t , and let x_{it} be a vector of observed covariates. The conditional mean is linked to a linear predictor by

$$g\{E(y_{it} \mid x_{it}, \alpha_i, \gamma_t)\} = x_{it}^T \beta + \alpha_i + \gamma_t,$$

where α_i is an individual effect and γ_t is an optional time effect. The **index** argument identifies the individual and time variables. The **model** argument determines how the individual effect is treated:

- "pooling" omits α_i from the mean model;
- "within" treats α_i as a fixed nuisance parameter;
- "random" assigns a distribution to α_i and integrates it out.

The optional **effect** = "twoways" includes both individual and time effects for supported fixed-effects models.

Model matrices follow the usual R formula rules, including factors and interactions. Before estimation, **panglm** checks rank after applying the transformation associated with the requested estimator. Unidentified columns, including time-invariant regressors in a within model, are reported and omitted. Missing responses or covariates follow **na.action**; missing or duplicate panel index pairs are errors. Nontrivial observation weights and nonzero offsets are outside the current estimator definitions and are rejected explicitly.

Pooled models

For Gaussian, Poisson, and binomial responses, pooled models maximize the ordinary generalized linear model likelihood by iteratively reweighted least squares. The negative binomial model uses the NB2 variance function

$$\text{Var}(y_{it} \mid x_{it}) = \mu_{it} + \frac{\mu_{it}^2}{\theta},$$

and alternates coefficient updates with estimation of the shape parameter θ .

```
fit_pool <- panglm(
  inv ~ value + capital,
  data = Grunfeld,
  index = c("firm", "year"),
  model = "pooling",
  family = "gaussian"
)
coef(fit_pool)
#> (Intercept)      value      capital
#> -42.7143694   0.1155622   0.2306785
```

Because pooling does not model within-unit dependence, inference should usually use a covariance estimator that is robust to clustering by unit.

```
sqrt(diag(vcov(fit_pool, type = "cluster"))))
#> (Intercept)      value      capital
#> 20.42520293   0.01589434   0.08496711
```

Fixed-effects models

Gaussian response

The one-way Gaussian estimator applies the within transformation

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i,$$

and estimates β by least squares using the transformed variables. Time-invariant regressors are removed by this transformation and cannot be identified. With `effect = "twoways"`, alternating projections remove both individual and time means. This procedure also accommodates unbalanced panels.

```
fit_fe <- panglm(
  inv ~ value + capital,
  data = Grunfeld,
  index = c("firm", "year"),
  model = "within",
  family = "gaussian"
)
coef(fit_fe)
#>      value      capital
#> 0.1101238 0.3100653
```

Poisson response

For a one-way Poisson model, the individual intercept has a closed-form profile solution. If $Y_i = \sum_t y_{it}$ and $L_i(\beta) = \sum_t \exp(x_{it}^T \beta)$, then

$$\hat{\alpha}_i(\beta) = \log\{Y_i / L_i(\beta)\}.$$

Substitution into the likelihood gives the conditional Poisson estimator. For two-way Poisson models, `panglm` uses an outer iteratively reweighted least-squares loop and weighted alternating projections to absorb the two sets of fixed effects.

Negative binomial response

The one-way fixed-effects NB2 estimator jointly estimates the covariate coefficients, one intercept for each individual, and the shape parameter. Groups for which every count is zero have an unbounded intercept and are excluded with an informative message. The two-way model absorbs individual and time effects through weighted alternating projections while updating the NB2 shape parameter.

The one-way specification is an unconditional nonlinear fixed-effects estimator. It is distinct from the Hausman-Hall-Griliches conditional negative binomial model and avoids interpreting that model's scale restriction as an individual intercept. Explicit dummy intercepts do not, however, remove the general finite-T incidental-parameter concern. Results from short panels should therefore be interpreted with appropriate caution.

Binomial response

The one-way binomial estimator is the exact conditional logistic model. For unit i , inference is conditional on the observed number of successes $S_i = \sum_t y_{it}$. The contribution to the conditional likelihood is

$$L_i(\beta | S_i) = \frac{\exp(\sum_t y_{it} x_{it}^T \beta)}{\sum_{a: \sum_t a_t = S_i} \exp(\sum_t a_t x_{it}^T \beta)}.$$

The denominator and score are evaluated by a forward and backward dynamic program. Units with all-zero or all-one responses do not contribute to this conditional likelihood and are excluded.

Random-effects models

Random-effects models retain time-invariant regressors and account for within-unit dependence by integrating over a latent individual effect. The implemented specifications are:

- Gaussian responses use the Swamy-Arora feasible generalized least-squares estimator;
- Poisson responses use a Poisson-Gamma mixture, which yields a closed-form marginal likelihood;
- negative binomial responses use a two-parameter beta-negative-binomial marginal likelihood;
- binomial responses use a Gaussian random intercept integrated by Gauss-Hermite quadrature.

The number of quadrature nodes for the binomial model is controlled by `R`. Random-effects covariance matrices are based on the observed information.

```
fit_re <- panglm(  
  inv ~ value + capital,  
  data = Grunfeld,  
  index = c("firm", "year"),  
  model = "random",  
  family = "gaussian"  
)  
coef(fit_re)  
#> (Intercept)      value      capital  
#> -57.8344149    0.1097812    0.3081130
```

Covariance estimation and confidence intervals

The default covariance matrix is model based. Pooled and fixed-effects models also support `type = "HC1"` and `type = "cluster"`. Let $s_j(\beta)$ be a score contribution and let B be the inverse information matrix. The sandwich estimator has the form

$$\hat{V}(\hat{\beta}) = B \left(\sum_j s_j(\hat{\beta}) s_j(\hat{\beta})^\top \right) B.$$

For cluster-robust inference, observation scores are first summed within clusters. The panel individual is used as the default cluster. For exact conditional binomial models, the independent likelihood contribution is a complete panel stratum. HC1 therefore uses stratum-level scores, and a custom clustering variable may combine but cannot split strata. Gaussian, Poisson, negative binomial, and conditional binomial fixed-effects fits all provide the score information required by these estimators.

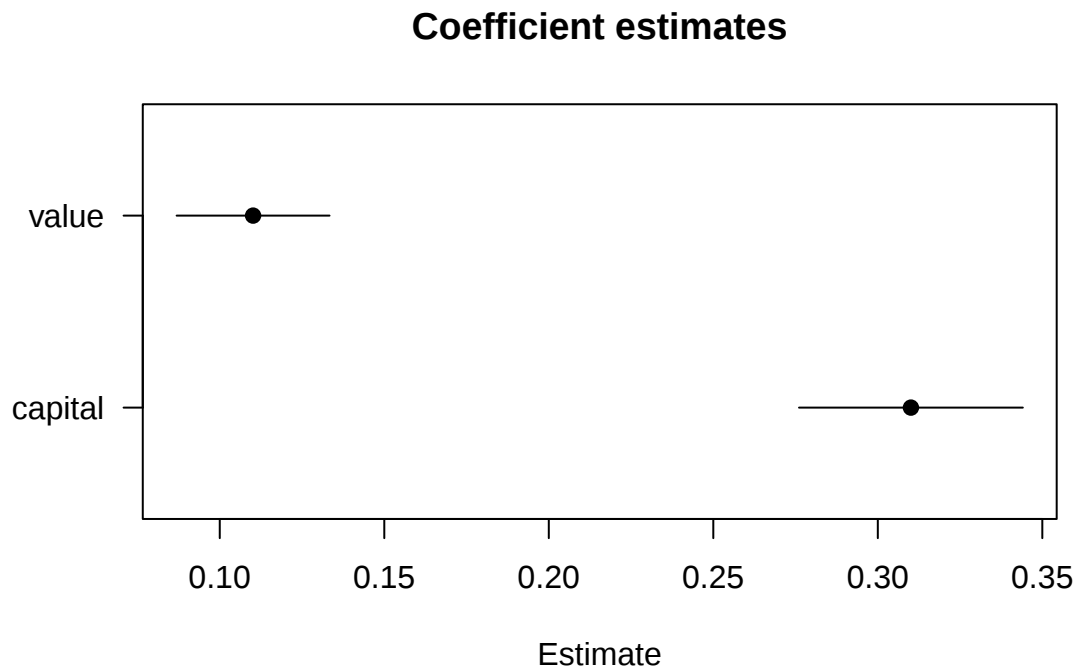
```
vcov(fit_fe, type = "cluster")
#>               value      capital
#> value  0.0002297066 0.0004661857
#> capital 0.0004661857 0.0027686951
confint(fit_fe)
#>           2.5 %    97.5 %
#> value  0.08688511 0.1333625
#> capital 0.27605114 0.3440795
```

Model assessment

For models that store fitted values, `plot()` provides coefficient, residual, and observed-versus-fitted displays. Coefficient plots are available for every `panglm` fit.

`fitted()` and in-sample `predict()` return values in the original data row order. Random-effects response predictions are marginal over the fitted latent-effect distribution. Fixed-effects predictions for new data include estimated effects for known panel levels; unseen levels are rejected unless `allow.new.levels = TRUE`. For exact conditional logistic regression, in-sample fitted values are probabilities conditional on each unit's observed success total. Arbitrary new-data probabilities are not identified because the individual intercept has been conditioned out.

```
plot(fit_fe, which = "coefficients")
```



`panglm_dispersiontest()` reports the Pearson statistic divided by its residual degrees of freedom for Poisson and negative binomial fits. `panglm_hausman()` compares compatible fixed-effects and random-effects estimates using their covariance difference.

For likelihood-based estimators, `logLik()`, `AIC()`, and `BIC()` include estimated shape, dispersion, variance, and absorbed fixed-effect parameters when these parameters belong to the reported likelihood. Conditional Poisson and conditional logistic likelihoods count only the regression parameters that remain after conditioning. Gaussian within and Swamy-Arora random-effects estimators do not report a likelihood, so their information criteria are undefined.

```
panglm_hausman(fit_fe, fit_re)
#>
#> Hausman specification test (fixed vs. random effects)
#>
#> chisq = 2.33 , df = 2 , p-value = 0.3119
#> alternative hypothesis: one model is inconsistent
```

Hurdle models for panel counts

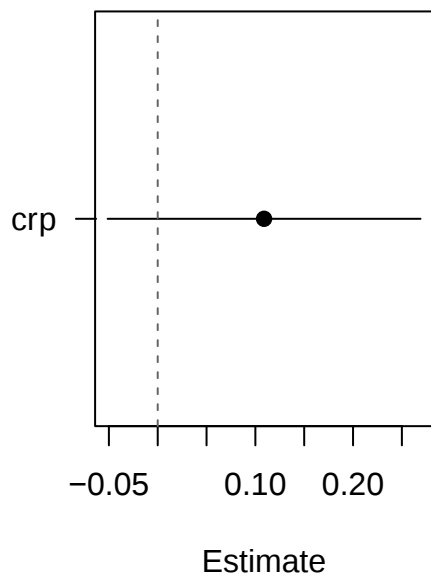
For outcomes with a distinct zero-generating process, `panglm_hurdle()` fits two fixed-effects components. The first is an exact conditional logistic model for $I(y_{it} > 0)$. The second is a zero-truncated Poisson or NB2 model for positive counts. Its density is

$$f_+(y \mid \mu, \theta) = \frac{f(y \mid \mu, \theta)}{1 - f(0 \mid \mu, \theta)}, \quad y > 0.$$

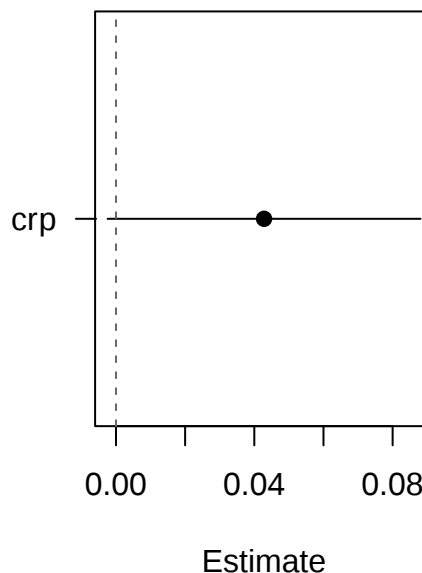
The count distribution is selected with `count_family`. Both positive-count models jointly estimate covariate coefficients and individual intercepts; the NB2 model also estimates θ .

```
data(copd)
fit_hurdle <- panglm_hurdle(
  exacerbations ~ crp,
  data = copd,
  index = c("id", "visit"),
  count_family = "negbin"
)
#> 26 group(s) with no within-group outcome variation dropped (all-0 or all-1) -- they carry no informa
plot(fit_hurdle)
```

Positive-outcome componen



Positive count: NB2



Computational validation

The test suite compares estimators with established implementations where the statistical models coincide. Comparisons include `stats::glm()` for pooled generalized linear models, `MASS::glm.nb()` for pooled NB2, `plm::plm()` for Gaussian panel models, `survival::clogit()` for exact conditional logistic regression, `fixest` for absorbed fixed-effects count models, `pglm` for compatible random-effects likelihoods, and `pscl::hurdle()` for zero-truncated count likelihoods.

Model scope

The package focuses on Gaussian, Poisson, binomial, and NB2 panel models. It does not provide ordinal, Tobit, or between estimators. A general two-way exact conditional logistic likelihood is not available, so the binomial fixed-effects model is restricted to individual effects. These restrictions define the current statistical scope and prevent unlike estimators from being presented under a common option.

References

- Allison, P. D., and Waterman, R. P. (2002). Fixed-effects negative binomial regression models. *Sociological Methodology*, 32, 247-265.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47, 225-238.
- Guimaraes, P., and Portugal, P. (2010). A simple feasible procedure to fit models with high-dimensional fixed effects. *Stata Journal*, 10, 628-649.
- Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, 52, 909-938.