Limited Dependent Variable in Panel Data with Stata

Generalized Linear Model

- Nelder and McCullagh (1972) describe a class of *Generalized* Linear Models (GLMs) that extends linear regression to permit non-normal stochastic and non-linear systematic components.
- GLMs encompass a broad and empirically useful range of specifications that includes linear regression, logistic and probit analysis, and Poisson models.
- GLMs offer a common framework in which we may place all of these specification, facilitating development of broadly applicable tools for estimation and inference.
- In addition, the GLM framework encourages the relaxation of distributional assumptions associated with these models, motivating development of robust *quasi-maximum likelihood (QML) estimators* and robust covariance estimators for use in these settings.

GLM Structure

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$
$$= \mu_i + e_i$$

Model	Family	Link
Linear Regression	Normal	Identity: $g(\mu) = \mu$
Exponential Regression	Normal	$\operatorname{Log:} g(\mu) = \log(\mu)$
Logistic Regression	Binomial	Logit: $g(\mu) = \log(\mu/(1-\mu))$
Probit Regression	Binomial	Probit: $g(\mu) = \Phi^{-1}(\mu)$
Poisson Count	Poisson	$Log: g(\mu) = log(\mu)$

The Gamma distribution

$$f(x) = \frac{1}{\sigma^{\alpha} \Gamma(\alpha)} x^{\alpha - 1} e^{-\frac{x}{\sigma}},$$

for $\alpha, \sigma, x > 0$
$$f(x) = \frac{\lambda^{x} e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \cdots$$

٠

The Poisson distribution

$$X_{i} | \varsigma_{i} \sim Poisson(\varsigma_{i}\mu_{i})$$
$$\varsigma_{i} \sim \frac{1}{\theta} Gamma(\theta)$$

$$X_i \sim \text{NegBinorm}(\mu_i, \theta) = f(x_i) == \frac{\Gamma(\theta + x_i)}{x_i! \Gamma(\theta)} \frac{\mu_i^{x_i} \theta^{\theta}}{(\mu_i + \theta)^{\theta + x_i}}$$

Estimation of Discrete Panel Data

- Binary outcome: probit logit
- 2. Count outcome: poisson
 - negative binominal

Parameter Estimation, given incidental parameter problem

Estimation by conditional likelihood: Searching a minimal sufficient statistic. Chamberlain (1980) indicates that, summation of y_{it} is a minimal sufficient statistic for the individual effects. Therefore, maximizing the conditional likelihood function below yields the conditional logit estimate estimates for β .

$$L_{c} = \prod_{i=1}^{N} \operatorname{Prob}\left(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \cdots \mathbf{y}_{iT} \mid \sum_{t=1}^{T} \mathbf{y}_{it}\right)$$

And, by definition of sufficient statistic, the distribution of the data given this sufficient statistic will not dependent upon individual effects μ_i .

That is, by conditioning on the summation of y_{it} , we swept away individual effects μ_i

Chamberlain G. (1980) Analysis of covariance with qualitative data. Review of Economic Studies, 47, 225-238

Case 1. Female union membership

idcode	year	age	grade	not_smsa	south	union	black
1	72	20	12	0	0	1	1
1	77	25	12	0	0	0	1
1	80	28	12	0	0	1	1
1	83	31	12	0	0	1	1
1	85	33	12	0	0	1	1
1	87	35	12	0	0	1	1
1	88	37	12	0	0	1	1
2	71	19	12	0	0	0	1
2	77	25	12	0	0	1	1
2	78	26	12	0	0	1	1
2	80	28	12	0	0	1	1
2	82	30	12	0	0	1	1
2	83	31	12	0	0	1	1
2	85	33	12	0	0	1	1
2	87	35	12	0	0	1	1
2	88	37	12	0	0	1	1
3	70	24	12	0	0	1	1
3	71	25	12	0	0	0	1
3	72	26	12	0	0	0	1
3	73	27	12	0	0	0	1
3	77	31	12	0	0	0	1
3	78	32	12	0	0	0	1
3	80	34	12	0	0	0	1
3	82	36	12	0	0	0	1
3	83	37	12	0	0	0	1
3	85	39	12	0	0	0	1
3	87	41	12	0	0	0	1

Union=F(age, grade, i.not_smsa)

<u>Statistics U</u> ser <u>W</u> indow <u>H</u> elp	
Summaries, tables, and tests	
Linear models and related	
Binary outcomes	Logistic regression
Ordinal outcomes 🔹 🕨	Logistic regression (reporting odds ratios)
Categorical outcomes	Exact logistic regression
Count outcomes	Mixed-effects logistic regression
Exact statistics	Panel logistic regression
Endogenous covariates	Probit regression
Sample-selection models	Probit regression (reporting change in prob.)
Multilevel mixed-effects models	Probit regression with endogenous covariates
Generalized linear models	Probit regression with selection
Nonparametric analysis	Bivariate probit regression
Time series 🕨	Seemingly unrelated bivariate probit regression
Multivariate time series	Panel probit regression
State-space models	Complementary log-log regression
Longitudinal/panel data 🔶	Panel complementary log-log regression
Survival analysis	GLM for the binomial family
Epidemiology and related	Heteroskedastic probit regression
Survey data analysis	Skewed logit regression
Multiple imputation	Grouped data 🔸
Multivariate analysis	Postestimation •

Estimation

Panel probit by random effect

xtprobit - Random-effects and population-avera	ged probit models
Model Correlation by/if/in Weights SE/Robust R	eporting Integration Maximization
Dependent variable: Independent variables: union v age grade not_smsa south##	Panel settings
Suppress constant term	
Model type (affects which options are available)	south year south*year
⊙ Random-effects (RE)	O Population-averaged (PA)
Options Offset variable: Constraints: Keep collinear variables (rarely used)	數為1的解釋變數 ✓ Manage
0 B B	OK Cancel Submit

Random-effects Group variable:			of obs = of groups =			
Random effects		obs per	r group: min = avg = max =	5.9		
Log likelihood	= - 10 552.22	25		Wald cł Prob >	ni2(6) =	220.91
union	Coef.	Std. Err.	z	P> z	[95% Conf.	Interval]
age grade not_smsa 1.south year	.0082967 .0482731 139657 -1.584394 0039854	.0084599 .0099469 .0460548 .358473 .0088399	0.98 4.85 -3.03 -4.42 -0.45	0.327 0.000 0.002 0.000 0.652	0082843 .0287776 2299227 -2.286989 0213113	.0248778 .0677686 0493913 8818002 .0133406
south#c.year 1	.0134017	.0044622	3.00	0.003	.0046559	.0221475
_cons	-1.668202	.4751819	-3.51	0.000	-2.599542	7368628

Likelihood-ratio test of rho=0: chibar2(01) = 5984.32 Prob >= chibar2 = 0.000

.0458783

.0311255

.0104643

lnsig2u = ln(
$$\sigma_{\varepsilon}^{2}$$
)
 $rho = \rho = \frac{\sigma_{\varepsilon}^{2}}{1 + \sigma_{\varepsilon}^{2}}$

/lnsiq2u

siqma_u

rho

.6103616

1.35687

.6480233

rho is the proportion of the total variance contributed by the panel-level variance component. When rho is zero, the panel-level variance component is unimportant, and the panel estimator is not different from the pooled estimator.

.7002814

1.419267

.6682502

.5204418

1.297217

.6272511

A likelihood-ratio test of rho=0 is shown at the bottom of the output. This test formally compares the pooled estimator (probit) with the panel estimator.

Estimation

Panel probit by equal-correlation, population -averaged

GEE population Group variable Link: Family: Correlation:	<u> </u>	id		Number	of obs = of groups = group: min = avg = max =	4434 1 5.9 12
Scale paramete	2r:		1	Prob >		
union	Coef.	Std. Err.	Z	P> z	[95% Conf.	Interval]
age grade not_smsa 1.south year	.0089699 .0333174 0715717 -1.017368 0062708	.0053208 .0062352 .027543 .207931 .0055314	1.69 5.34 -2.60 -4.89 -1.13	0.092 0.000 0.009 0.000 0.257	0014586 .0210966 1255551 -1.424905 0171122	.0193985 .0455382 0175884 6098308 .0045706
south#c.year 1 _cons	. 0086294 8670997	. 00258 . 294771	3.34 -2.94	0.001 0.003	.0035727 -1.44484	. 013686 2893592

Robust Covariance

fodel Correlation by/if/in Weights SE/Robu	st Reporting Integration Optimization
Standard error type:	-Scale factors
Default standard errors	⊙ Divisor N (default)
Robust Conventional Bootstrap Jackknife	O Use divisor N-P instead of N [nmp]
	-Scale value choices
	 Default for chosen family
	O Pearson chi-squared over d.f.
	O Deviance over degrees of freedom
	O Do not rescale the variance
	O User-supplied scale

GEE population Group variable Link: Family: Correlation: Scale paramete	2:	id pr	lcode obit mial eable 1		of groups = group: min = avg = max = ni2(6) =	4434 1 5.9 12 156.33
		(std.	Err. adj	justed fo	or clustering	on idcode)
union	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf.	Interval]
age grade not_smsa 1.south year	.0089699 .0333174 0715717 -1.017368 0062708	.0051169 .0076425 .0348659 .3026981 .0055745	1.75 4.36 -2.05 -3.36 -1.12	0.080 0.000 0.040 0.001 0.261	001059 .0183383 1399076 -1.610645 0171965	.0189988 .0482965 0032359 4240906 .0046549
south#c.year 1 _cons	. 0086294 8670997	.0037866 .3243959	2.28 -2.67	0.023 0.008	.0012078 -1.502904	.0160509 2312955

Quadrature Stability

Model Correlation by/if/in Weights SE/Robust Rep	orting Integration Maximization							
- Integration options for RE model								
¥-4-1.	New bar of an darkers or inter							
Method:	Number of quadrature points:							
Default 💙	12 🤤							
Default								
Mean and variance adaptive Gauss-Hermite quadrature								
Mode and curvature adaptive Gauss-Hermite quadrature								
Gauss-Hermite quadrature								

The random-effects model is calculated using quadrature, which is an approximation whose accuracy depends partially on the number of integration points used. We can use the **quadchk** command to see if changing the number of integration points affects the results. If the results change, the quadrature approximation is **not** accurate given the number of integration points.

Try increasing the number of integration points using the intpoints() option and run quadchk again.

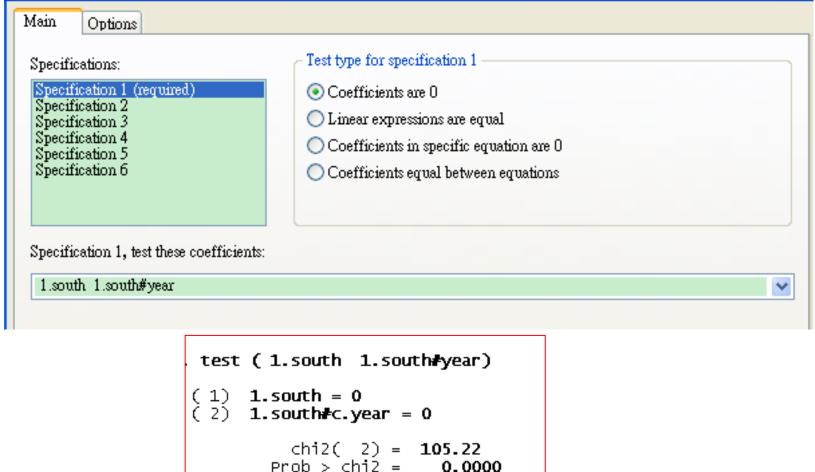
Do not attempt to interpret the results of estimates when the coefficients reported by quadchk differ substantially.

Post-estimation test

<u>Statistics</u> <u>U</u> ser <u>W</u> indow <u>H</u> elp	
Summaries, tables, and tests	
Linear models and related Binary outcomes Ordinal outcomes Categorical outcomes Count outcomes Exact statistics Endogenous covariates Sample selection models Linear models and related Number of obs = 26200 Number of groups = 4434 obs per group: min = 1 avg = 5.9 max = 12 Wald chi2(6) = 156.33 Prob > chi2 = 0.0000 sted for clustering on idcode)	
P>[Z] [95% CONT. Interval]	
Multilevel mixed-effects models • Generalized linear models • Generalized linear models • Nonparametric analysis • Time series • Multivariate time series • Multivariate time series • State-space models • Longitudinal/panel data • Survival analysis • Epidemiology and related • Survey data analysis • Multiple imputation •	
Multivariate analysis Predictions, residuals, etc.	
Power and sample size	
Resampling Marginal means and predictive margins Marginal effects Test linear hypotheses	
Postestimation Test parameters	
Other Tests Test nonlinear hypotheses	

📧 test - Test linear hypotheses after estimation





To test whether residing in the south affects union status, we must determine whether **1**.south and south#c.year are jointly significant.

Post-estimation Tests

1. Hausman Test for random effect

H₀: **E**($u_{it}|X_{it}$)=0

• The null hypothesis implies that:

The random effects model is better

• If the null if rejected, then the fixed effects model is a better specification.

$$m = (\hat{\beta} - \hat{b}_s)'\hat{\Sigma}^{-1}(\hat{\beta} - \hat{b}_s) \qquad \hat{\Sigma} = C\hat{o}v(\hat{\beta}) - C\hat{o}v(\hat{b}_s)$$

- Under H_0 ,
 - The difference of the estimates of two models are negligibly small.
 - Both LSDV and GLS are *consistent*, but LSDV are *not efficient*. Hence, the random models is the better choice.
- If we reject *H*₀, only LSDV is consistent, but not for GLS, fixed models is more applicable.

Hausman test in action

- 1. 從選單選擇logit,估計模型選 random effect。
- 2. 將估計結果存取,命名為re。 存取方法如次頁

<u>File Edit Data Graphics Statistics User Window H</u> elp	
😂 🛃 💼 🗐 💿 🗸 📗 🔹 Summaries, tables, and tests 🔹 🕨	
tau = 0.8 log Linear models and related	
Iteration 0: log Binary outcomes	
Iteration 1: log Ordinal outcomes	
Iteration 3: log Categorical outcomes •	
Iteration 4: log l Iteration 5: log l Countoutcomes	:ked up)
Random-effects logis	Number of obs = 26200
	Number of groups = 4434
Random effects u_i ~ Sample-selection models	Obs per group: min = 1
Multilevel mixed-effects models	avg = 5.9 max = 12
Generalized linear models	
Log likelihood = -1	wald chi2(5) = 221.30 - Prob > chi2 = 0.0000
Time series	
union Multivariate time series	P> z [95% Conf. Interval]
age .01 State-space models	0.300013834 .0448789
grade .08 Longitudinal/panel data 🕨	0.000 .0533821 .1225202
not_smsa25 year .00 Survival analysis	0.00241669270943141 0.9240287011 .0316502
south93 _cons -3.6 Epidemiology and related →	0.000 -1.0953047800111 -0.000 -5.265115 -2.084088
Survey data analysis	Save to diele
/lnsig2u 1.7 Multiple imputation	1.655122 1.839353 Load from disk
sigma_u 2.3 rho .63 Multivariate analysis	Predictions, residuals, etc. Describe results
Power and sample size	Nonlinear predictions Store in memory
Likelihood-ratio tes Resampling	Marginal means and predictive margins Restore from memory
. estimates store re	Marginal effects List results stored in memory
. xtlogit union age	Tests Drop from memory
note: multiple posit Other •	Linear combinations of estimates Redisplay estimation output
Command	Nonlinear combinations of estimates Table of estimation results
	Reports and statistics Table of fit statistics
C:\Documents and Settings\USER\My Documents	Manage estimation results Title/retitle results estimates store - Store active estimation res

r	Store active estimation results in memory									
	Name	:								
	re									
	Clear current (active) estimation results after storing									
0	0	B	0	К		Cancel		Submit		

X

- 3. 重複上面步驟1,估計模型選 fixed-effect。
- 4. 將估計結果存取,命名為fe。
- 5. 執行 Hausman test.

<u>F</u> ile <u>E</u> dit <u>D</u> ata	<u>G</u> raphics <u>S</u> ta	tistics <u>U</u> ser <u>W</u> indow <u>H</u> elp		
		Summaries, tables, and tests		
tau = 0.8	log l	Linear models and related	-	
Iteration 0:	log 1	Binary outcomes +		
Iteration 1:	log 1 log 1	Ordinal outcomes		
Iteration 2: Iteration 3:	log 1	Categorical outcomes		
Iteration 4: Iteration 5:	log 1 log 1	Count outcomes +	:ked up)	
Random-effects	-	Exact statistics	Number of obs = 26200	
Group variable		Endogenous covariates	Number of groups = 4434	
Random effects	u_i ~	Sample-selection models	obs per group: min = 1	
		$Multilevel\ mixed-effects\ models \qquad \bullet$	avg = 5.9	
		Generalized linear models	max = 12	
Log likelihood	= _1	Nonparametric analysis	wald chi2(5) = 221.30 Prob > chi2 = 0.0000	
		Time series +		
union		Multivariate time series	P> z [95% Conf. Interval]	
age	. 01	State-space models	0.300013834 .0448789	
grade	. 08	Longitudinal/panel data	0.000 .0533821 .1225202	
not_smsa year	25 .00	Survival analysis	0.00241669270943141 0.9240287011 .0316502	
south	93 -3.6	Epidemiology and related	0.000 -1.0953047800111 0.000 -5.265115 -2.084088	
cons		Survey data analysis		
/lnsig2u	1.7	Multiple imputation	1.655122 1.839353	
sigma_u rho	2.3	Multivariate analysis	Predictions, residuals, etc.	
		Power and sample size	Nonlinear predictions	
Likelihood-rat		Resampling +	Marginal means and predictive margins	Test linear hypotheses
. estimates st	ore re	Postestimation	Marginal effects	Test parameters
. xtlogit unio note: multiple	n age	Other 🕨	Tests •	- Test nonlinear hypotheses
mote. muitible	DOSTC		Linear combinations of estimates	Likelihood-ratio test
Command			Nonlinear combinations of estimates	Specification link test for single-equation models
			Reports and statistics	Hausman specification test
C:\Documents and Settin			Manage estimation results	Seemingly unrelated estimation
	100 PP 2012 22			

🗉 hausman - Hausman specification test	
Main Advanced	
Consistent estimation:	Efficient est.: (blank for most recent)
fe 🔽	re 🔽
Options	
Intercepts	Equations to use to perform test
• Exclude from comparison	💿 Use first equation only
🔘 Include in comparison	🔘 Use all equations
	◯ Skip these equations:

. hausman fe i	re, constant			
	Coeffi (b) fe	cients —— (B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
age grade not_smsa year south	.0760241 .0857788 .0096844 0594931 7476614	.0155224 .0879511 2555034 .0014745 9376577	.0605017 0021723 .2651878 0609677 .1899964	.0949138 .0379763 .0772299 .0955826 .0958801
	inconsistent u	nder Ha, effic		obtained from xtlogit obtained from xtlogit
	chi2(5) = = Prob>chi2 =	(b-B)'[(V_b-V_ 16.41 0.0058	В)^(-1)](b-B)	

2. Computing Marginal Effects

In this example, we fit a population-averaged model of union status on the woman's age and level of schooling, whether she lived in an urban area, whether she lived in the south, and the year observed. Here we compute the average marginal effects from that fitted model on the probability of being in a union.

<u>S</u> tati:	stics <u>U</u> ser <u>W</u> indow <u>H</u> elp		-
	Summaries, tables, and tests	۲	
	Linear models and related	۲	Prob > chi2 = 0.0000
	Binary outcomes	×	
	Ordinal outcomes	⊬	P> z [95% Conf. Interval]
	Categorical outcomes	¥	0.3320083598 .0247719
	Count outcomes	•	0.000 .0292627 .0682251
	Exact statistics		0.00223209540517916 0.9290162489 .0178078
			0.00060504474288631
	Endogenous covariates	×	0.000 -2.945088 -1.150355
	Sample-selection models	۲	. 5182608 . 6980848
	Multilevel mixed-effects models	×	
	Generalized linear models	۲	1.295803 1.417709 .626741 .667763
	Nonparametric analysis	×	
	Time series		978.72 Prob >= chibar2 = 0.000
		ŗ.	outh)
	Multivariate time series	•	
	State-space models		Number of obs = 26200
	Longitudinal/panel data	۲	
	Survival analysis	۲	
	Epidemiology and related	۲	
	Survey data analysis	۲	
	Multiple imputation		P> z [95% Conf. Interval]
	Multivariate analysis	►	Predictions, residuals, etc.
	Power and sample size	•	Nonlinear predictions
	· · · · · · · · · · · · · · · · · · ·	-	Marginal means and predictive margins
	Resampling	×.	Marginal effects
	Postestimation	×	, ,
	Other	۲	Tests •
_			Linear combinations of estimates

margins - Marginal means, predictive margins, and mar	ginal effects						
Main At if/in/over Within SE Advanced Rey	porting						
Factor terms to compute margins for:							
Pacifor terms to compute margins for.		~					
Add grand margin, default if no factor terms specified							
Select response							
 Default prediction 							
○ Specify a prediction							
O Specify an expression of estimated parameters							
✓ Marginal effects of response							
⊙ Marginal effects d (y)/d (x)							
◯ Elasticities d (lny)/d (lnx)							
🚫 Semielasticities d (y)/d (lnx)							
🔘 Semielasticities d (lny)/d (x)							
Variables:	· ·						
age grade not_smsa year south	. margins, gra		grade not_si	nsa year	_	. .	
Treat factor-variable level indicator covariates as continuous	Average margir Model VCE	nal effects			Numbe	r of obs =	26200
I reat factor-variable level indicator covariates as continuous	Model VCE	OIM					
	Expression	: Linear pred	iction, pred	ict()			
		: Linear pred	iction, pred ot_smsa year	ict() south			
	Expression	: Linear pred : age grade n	ot_smsa year Delta-method	ict() south			
	Expression	: Linear pred : age grade n	ot_smsa year	ict() south	P> z	[95% Conf.	Interval]
	Expression dy/dx w.r.t.	Linear pred age grade n dy/dx .0082061	ot_smsa year Delta-method Std. Err. .0084521	south 2 0.97	0.332	0083598	.0247719
	Expression dy/dx w.r.t.	: Linear pred age grade n dy/dx	ot_smsa year Delta-method Std. Err.	south z		_	_

Count Model in Panel Data

Count panel data

ship	yr_con	yr_op	service	accident	op_75_79	CO_65_69	co_70_74	C0_75_79
1	1	1	127	0	0	0	0	0
1	1	2	63	0	1	0	0	0
1	2	1	1095	3	0	1	0	0
1	2	2	1095	4	1	1	0	0
1	3	1	1512	6	0	0	1	0
1	3	2	3353	18	1	0	1	0
1	4	1			0	0	0	1
1	4	2	2244	11	1	0	0	1
2	1	1	44882	39	0	0	0	0
2	1	2	17176	29	1	0	0	0
2	2	1	28609	58	0	1	0	0
2	2	2	20370	53	1	1	0	0
2	3	1	7064	12	0	0	1	0
2	3	2	13099	44	1	0	1	0
2	4	1	•	•	0	0	0	1
2	4	2	7117	18	1	0	0	1

Ships.dta is data on the number of ship accidents for five different types of ships (McCullagh and Nelder 1989, 205). We wish to analyze whether the "incident" rate is affected by the period in which the ship was constructed and operated. Our measure of exposure is months of service for the ship, and in this model, we assume that the exponentiated random effects are distributed as gamma with mean one and variance alpha.

N=ship

Statistics \rightarrow longitudinal/panel data \rightarrow count outcomes \rightarrow **Poisson regression**

🗧 xtpoisson - Fixed-effec	ts, random-effects, and population-averag	ged Poisson models 👘 🔲 🔁 🚺
Model Correlation by/if/	n Weights SE/Robust Reporting Integratio	on Maximization
Dependent variable: accident variable Model type (affects which o	Independent variables: op_75_79 co_65_69 co_70_74 co_75_79 Suppress constant term options are available)	Panel settings
⊙ Random-effects (RE)	◯ Fixed-effects (FE)	O Population-averaged (PA)
Options • Exposure variable: service	Offset variable:	
Use normal distribution	for random effects (default = gamma)	
Constraints:	: (rarely used)	Manage

Model Correlation by/if/in Weights	SE/Robust Reporting Integration Maximization
95 🛛 🖌 Confidence level	
🔘 Report coefficients (default)	
Report incidence-rate ratios	
Additional test statistics	
Perform likelihood-ratio test	irr reports exponentiated coefficients e ^b rather than coefficients b .
Do not report constraints	For the Poisson model, exponentiated coefficients are
Suppress omitted collinear covariates	interpreted as incidence-rate ratios.
Suppress omitted collinear covariates Suppress blank lines	interpreted as incidence-rate ratios.
	interpreted as incidence-rate ratios.
Suppress blank lines	interpreted as incidence-rate ratios.
Factor variables	interpreted as incidence-rate ratios.
Suppress blank lines Factor variables Suppress covariates with empty cells	interpreted as incidence-rate ratios.
 Suppress blank lines Factor variables Suppress covariates with empty cells Base level variables 	
 Suppress blank lines Factor variables Suppress covariates with empty cells Base level variables Suppress all base level variables 	

Random-effects Group variable	2	Number o Number o	of obs = of groups =	5.		
Random effects	su_i∼Gamma			obs per	group: min = avg = max =	6.8
Log likelihood	d = - 74.81121	.7		Wald chi Prob > <		A AAAA
accident	IRR	Std. Err.	z	P> Z	[95% Conf.	Interval]
op_75_79 co_65_69 co_70_74 co_75_79 service	1.466305 2.032543 2.356853 1.641913 (exposure)	.1734005 .304083 .3999259 .3811398	3.24 4.74 5.05 2.14	0.001 0.000 0.000 0.033	1.162957 1.515982 1.690033 1.04174	1.848777 2.72512 3.286774 2.58786
/lnalpha	-2.368406	.8474597			-4.029397	7074155
alpha	. 0936298	.0793475			.0177851	.4929165
Likelihood-rat	tio test of al	pha=0: chiba	ir2(01) =	10.61	L Prob>=chiba	r2 = 0.001

The output also includes a likelihood-ratio test of $\alpha = 0$, which compares the panel estimator with the pooled (Poisson) estimator.

$$\Pr(y_{i1},\ldots,y_{in_i}|\alpha_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}) = \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right) \exp\left\{-\exp(\alpha_i)\sum_{t=1}^{n_i} \lambda_{it}\right\} \exp\left(\alpha_i\sum_{t=1}^{n_i} y_{it}\right)$$

where $\lambda_{it} = \exp(\mathbf{x}_{it}\beta)$. We may rewrite the above as (defining $\epsilon_i = \exp(\alpha_i)$)

Prediction. Case 1

we fit a random-effects model of the number of accidents experienced by five different types of ships on the basis of when the ships were constructed and operated. Here we obtain the *predicted number of accidents* for each observation, assuming that the random effect for each panel is zero

float 💌
0
Cancel Submit

ship	yr_con	yr_op	service	accident	op_75_79	CO_65_69	C0_70_74	CO_75_79	n_acc
1	1	1	127	0	0	0	0	0	.1742982
1	1	2	63	0	1	0	0	0	.1267809
1	2	1	1095	3	0	1	0	0	3.054522
1	2	2	1095	4	1	1	0	0	4.478859
1	3	1	1512	6	0	0	1	0	4.890728
1	3	2	3353	18	1	0	1	0	15.90302
1	4	1			0	0	0	1	
1	4	2	2244	11	1	0	0	1	7.414576
2	1	1	44882	39	0	0	0	0	61.59727
2	1	2	17176	29	1	0	0	0	34.56491
2	2	1	28609	58	0	1	0	0	79.80531
2	2	2	20370	53	1	1	0	0	83.31909
2	3	1	7064	12	0	0	1	0	22.84928
2	3	2	13099	44	1	0	1	0	62.12753
2	4	1	•		0	0	0	1	
2	4	2	7117	18	1	0	0	1	23.51583
3	1	1	1179	1	0	0	0	0	1.618091
summar	ize n_a	сс							
Vari	able	0	bs	Mean	std.	Dev.	Mir	1	Max
n	_acc		34 1 3	3.52307	23.1	5885	. 0617592	83.3	1905

From these results, you may be tempted to conclude that some types of ships are safe, with a predicted number of accidents close to zero, whereas others are dangerous, because 1 observation is predicted to have more than 83 accidents.

Prediction. Case 2

However, when we fit the model, we specified the exposure(service) option. The variable service records the *total number of months* of operation for each type of ship constructed in and operated during particular years.

Because ships experienced different utilization rates and thus were exposed to different levels of accident risk, we included service as our exposure variable. When comparing different types of ships, we must therefore predict the number of accidents, assuming that all ships faced the same exposure to risk. For this purpose, we do the following:

predict - Prediction after estimation Main if/in Main if/in	
New variable name: rate_acc - Produce:	New variable type:
 Linear prediction Standard error of the linear prediction Predicted number of events assuming u_i = 0 	
Predicted incidence rate assuming u_i = 0	
Ignore offset variable (if any)	

predict rate_acc, iru0

ship	yr_con	yr_op	service	accident	op_75_79	CO_65_69	CO_70_74	CO_75_79	n_acc	rate_acc
1	1	1	127	0	0	0	0	0	.1742982	.0013724
1	1	2	63	0	1	0	0	0	.1267809	.0020124
1	2	1	1095	3	0	1	0	0	3.054522	.0027899
1	2	2	1095	4	1	1	0	0	4.478859	.0040903
1	3	1	1512	6	0	0	1	0	4.890728	.0032340
1	3	2	3353	18	1	0	1	0	15.90302	.0047429
1	4	1			0	0	0	1		.0022534
1	4	2	2244	11	1	0	0	1	7.414576	.0033042
2	1	1	44882	39	0	0	0	0	61.59727	.0013724
2	1	2	17176	29	1	0	0	0	34.56491	.002012
2	2	1	28609	58	0	1	0	0	79.80531	.002789
2	2	2	20370	53	1	1	0	0	83.31905	.004090
2	3	1	7064	12	0	0	1	0	22.84928	.003234
2	3	2	13099	44	1	0	1	0	62.12753	.004742
2	4	1			0	0	0	1		.002253
2	4	2	7117	18	1	0	0	1	23.51583	.0033043
3	1	1	1179	1	0	0	0	0	1.618091	.001372
3	1	2	552	1	1	0	0	0	1.110842	.002012
3	2	1	781	0	0	1	0	0	2.178613	.002789
3	2	2	676	1	1	1	0	0	2.765031	.004090
3	3	1	783	6	0	0	1	0	2.532699	.003234
3	3	2	1948	2	1	0	1	0	9.239211	.004742
3	4	1			0	0	0	1		.002253

summarize rate_acc							
Variable	obs	Mean	Std. Dev.	Min	Max		
rate_acc	40	.002975	.0010497	.0013724	.0047429		

These results show that if each ship were used for 1 month, the expected number of accidents is 0.002975. Depending on the type of ship and years of construction and operation, the incidence rate of accidents ranges from 0.00137 to 0.00474.

Negative binomial distribution if y is over-dispersed or rare event data

airacc.dta

You have (fictional) data on injury "incidents" incurred among 20 airlines in each of 4 years.

(Incidents range from major injuries to exceedingly minor ones.) The government agency in charge of regulating airlines has run an experimental safety training program, and, in each of the years, some airlines have participated and some have not. You now wish to analyze whether the "incident" rate is affected by the program. You choose to estimate using random-effects negative binomial regression, as the dispersion might vary across the airlines for unidentified airline-specific reasons. Your measure of exposure is passenger miles for each airline in each year.

airacc.dta

Г

airline	ai	rec	inprog	ait	uit	relcnt	relsize	pmiles	i_cnt	time
1	.3117136	1	1	.1617136	.7607015	.4845998	3.654406	3654.406	25	1
1	.3117136	21	1	.1617136	.3718922	.2470596	3.931338	3931.338	17	2
1	.3117136	22	0	.3117136	.8252394	.6156702	2.814388	2814.388	22	3
1	.3117136	23	0	.3117136	.9317297	.6807297	3.926849	3926.849	34	4
2	.7895887	2	0	.7895887	.9370923	.9759603	2.229489	2229.489	26	1
2	.7895887	24	0	.7895887	.8567271	.9268618	3.997382	3997.382	45	2
2	.7895887	25	0	.7895887	.9764408	1	2.561755	2561.755	30	3
2	.7895887	26	1	.6395887	.994688	.9195066	2.283699	2283.698	25	4
3	.2410518	3	0	.2410518	.169451	.1718506	2.859919	2859.919	10	1
3	.2410518	27	0	.2410518	.5827496	.4243525	3.724612	3724.612	23	2
3	.2410518	28	1	.0910518	.0963485	.0355478	3.538779	3538.779	8	3
3	.2410518	29	0	.2410518	.9820369	.6682943	2.524797	2524.797	21	4
4	.0325279	4	0	.0325279	.6308963	.3263712	3.367893	3367.893	17	1
4	.0325279	30	1	1174721	.6955879	.2742526	3.966115	3966.115	18	2
4	.0325279	31	0	.0325279	.0966873	0	2.676133	2676.133	5	3
4	.0325279	32	0	.0325279	.8351641	.4511671	3.281538	3281.539	21	4
5	.2406449	5	0	.2406449	.9501256	.6485496	2.124044	2124.044	18	1
5	.2406449	33	0	.2406449	.8795498	.6054319	2.447307	2447.307	19	2
5	.2406449	34	0	.2406449	.4237015	.3269345	2.635407	2635.407	13	3
5	.2406449	35	1	.0906449	.9640464	.5654129	3.554512	3554.512	27	4
6	.8185568	6	0	.8185568	.9408058	.995927	3.036886	3036.886	36	1
6	.8185568	36	0	.8185568	.8469768	.9386028	2.863813	2863.813	32	2
6	.8185568	37	0	.8185568	.1380819	.5055085	3.297026	3297.026	23	3

Statistics \rightarrow longitudinal/panel data \rightarrow count \rightarrow *negative binominal regression*

🔳 xtnbreg - Fixed-, random-effects, population-averaged neg	ative binomial models 📃 🗖 🗙						
Model Correlation by/if/in Weights SE/Robust Reporting M	aximization						
Dependent variable: Independent variables: i_cnt inprog Suppress constant term	Panel settings						
Model type (affects which options are available)							
	O Population-averaged (PA)						
Options Exposure variable: Differ variable: Constraints:							
Keep collinear variables (rarely used)	Manage						
2 B h	OK Cancel Submit						

logit

If we assume a normal distribution, $N(0, \sigma_{\nu}^2)$, for the random effects ν_i ,

$$\Pr(y_{i1},\ldots,y_{in_i}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}) = \int_{-\infty}^{\infty} \frac{e^{-\nu_i^2/2\sigma_\nu^2}}{\sqrt{2\pi}\sigma_\nu} \left\{ \prod_{t=1}^{n_i} F(y_{it},\mathbf{x}_{it}\beta + \nu_i) \right\} d\nu_i$$

where

$$F(y,z) = \begin{cases} \frac{1}{1 + \exp(-z)} & \text{if } y \neq 0\\ \frac{1}{1 + \exp(z)} & \text{otherwise} \end{cases}$$

The panel-level likelihood l_i is given by

$$\begin{split} l_i &= \int_{-\infty}^{\infty} \frac{e^{-\nu_i^2/2\sigma_{\nu}^2}}{\sqrt{2\pi}\sigma_{\nu}} \left\{ \prod_{t=1}^{n_i} F(y_{it}, \mathbf{x}_{it}\beta + \nu_i) \right\} d\nu_i \\ &\equiv \int_{-\infty}^{\infty} g(y_{it}, x_{it}, \nu_i) d\nu_i \end{split}$$

Poisson

For a random-effects specification, we know that

$$\Pr(y_{i1},\ldots,y_{in_i}|\alpha_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}) = \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right) \exp\left\{-\exp(\alpha_i)\sum_{t=1}^{n_i} \lambda_{it}\right\} \exp\left(\alpha_i\sum_{t=1}^{n_i} y_{it}\right)$$

where $\lambda_{it} = \exp(\mathbf{x}_{it}\beta)$. We may rewrite the above as (defining $\epsilon_i = \exp(\alpha_i)$)

$$\Pr(y_{i1}, \dots, y_{in_i} | \epsilon_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}) = \left\{ \prod_{t=1}^{n_i} \frac{(\lambda_{it} \epsilon_i)^{y_{it}}}{y_{it}!} \right\} \exp\left(-\sum_{t=1}^{n_i} \lambda_{it} \epsilon_i\right)$$
$$= \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right) \exp\left(-\epsilon_i \sum_{t=1}^{n_i} \lambda_{it}\right) \epsilon_i^{\sum_{t=1}^{n_i} y_{it}}$$

We now assume that ϵ_i follows a gamma distribution with mean one and variance θ so that unconditional on ϵ_i

$$\begin{aligned} \Pr(y_{i1}, \dots, y_{in_i} | \mathbf{X}_i) &= \frac{\theta^{\theta}}{\Gamma(\theta)} \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!} \right) \int_0^\infty \exp\left(-\epsilon_i \sum_{t=1}^{n_i} \lambda_{it} \right) \epsilon_i^{\sum_{t=1}^{n_i} y_{it}} \epsilon_i^{\theta-1} \exp(-\theta\epsilon_i) d\epsilon_i \\ &= \frac{\theta^{\theta}}{\Gamma(\theta)} \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!} \right) \int_0^\infty \exp\left\{ -\epsilon_i \left(\theta + \sum_{t=1}^{n_i} \lambda_{it} \right) \right\} \epsilon_i^{\theta + \sum_{t=1}^{n_i} y_{it} - 1} d\epsilon_i \\ &= \left(\prod_{t=1}^{n_i} \frac{\lambda_{it}^{y_{it}}}{y_{it}!} \right) \frac{\Gamma\left(\theta + \sum_{t=1}^{n_i} y_{it} \right)}{\Gamma(\theta)} \left(\frac{\theta}{\theta + \sum_{t=1}^{n_i} \lambda_{it}} \right)^{\theta} \left(\frac{1}{\theta + \sum_{t=1}^{n_i} \lambda_{it}} \right)^{\sum_{t=1}^{n_i} y_{it}} \end{aligned}$$

for $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})$.

negative binominal

For the random-effects and fixed-effects overdispersion models, let y_{it} be the count for the *t*th observation in the *i*th group. We begin with the model $y_{it} | \gamma_{it} \sim \text{Poisson}(\gamma_{it})$, where $\gamma_{it} | \delta_i \sim \text{gamma}(\lambda_{it}, \delta_i)$ with $\lambda_{it} = \exp(\mathbf{x}_{it}\beta + \text{offset}_{it})$ and δ_i is the dispersion parameter. This yields the model

$$\Pr(Y_{it} = y_{it} \mid \mathbf{x}_{it}, \delta_i) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{1}{1 + \delta_i}\right)^{\lambda_{it}} \left(\frac{\delta_i}{1 + \delta_i}\right)^{y_{it}}$$

For a random-effects overdispersion model, we allow δ_i to vary randomly across groups; namely, we assume that $1/(1 + \delta_i) \sim \text{Beta}(r, s)$. The joint probability of the counts for the *i*th group is

$$\Pr(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i} | \mathbf{X}_i) = \int_0^\infty \prod_{t=1}^{n_i} \Pr(Y_{it} = y_{it} | \mathbf{x}_{it}, \delta_i) f(\delta_i) d\delta_i$$
$$= \frac{\Gamma(r+s)\Gamma(r+\sum_{t=1}^{n_i} \lambda_{it})\Gamma(s+\sum_{t=1}^{n_i} y_{it})}{\Gamma(r)\Gamma(s)\Gamma(r+s+\sum_{t=1}^{n_i} \lambda_{it}+\sum_{t=1}^{n_i} y_{it})} \prod_{t=1}^{n_i} \frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}$$

for $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})$ and where f is the probability density function for δ_i . The resulting log likelihood is

$$\ln L = \sum_{i=1}^{n} w_i \left[\ln\Gamma(r+s) + \ln\Gamma\left(r + \sum_{k=1}^{n_i} \lambda_{ik}\right) + \ln\Gamma\left(s + \sum_{k=1}^{n_i} y_{ik}\right) - \ln\Gamma(r) - \ln\Gamma(s) - \ln\Gamma(r) - \ln\Gamma(s) \right]$$
$$- \ln\Gamma\left(r + s + \sum_{k=1}^{n_i} \lambda_{ik} + \sum_{k=1}^{n_i} y_{ik}\right) + \sum_{t=1}^{n_i} \left\{ \ln\Gamma(\lambda_{it} + y_{it}) - \ln\Gamma(\lambda_{it}) - \ln\Gamma(y_{it} + 1) \right\}$$

where $\lambda_{it} = \exp(\mathbf{x}_{it}\beta + \text{offset}_{it})$ and w_i is the weight for the *i*th group (Hausman, Hall, and Griliches 1984, equation 3.5, 927).