# 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 | 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**

$$\boxed{\begin{array}{l} f(x) = \dfrac{1}{\sigma^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\sigma}}, \\ \text{for } \alpha, \sigma, x > 0 \end{array}}$$

**The Poisson distribution**

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

**Negative binomial distribution**

$$\boxed{\begin{array}{r} X_i \mid \varsigma_i \sim Poisson(\varsigma_i \mu_i) \\ \varsigma_i \sim \dfrac{1}{\theta} Gamma(\theta) \end{array}}$$

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

# Estimation of Discrete Panel Data

1. 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 $\beta$.

$$L_c = \prod_{i=1}^{N} \mathrm{Prob}\left( y_{i1}, y_{i2}, \cdots y_{iT} \mid \sum_{t=1}^{T} y_{it} \right)$$

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

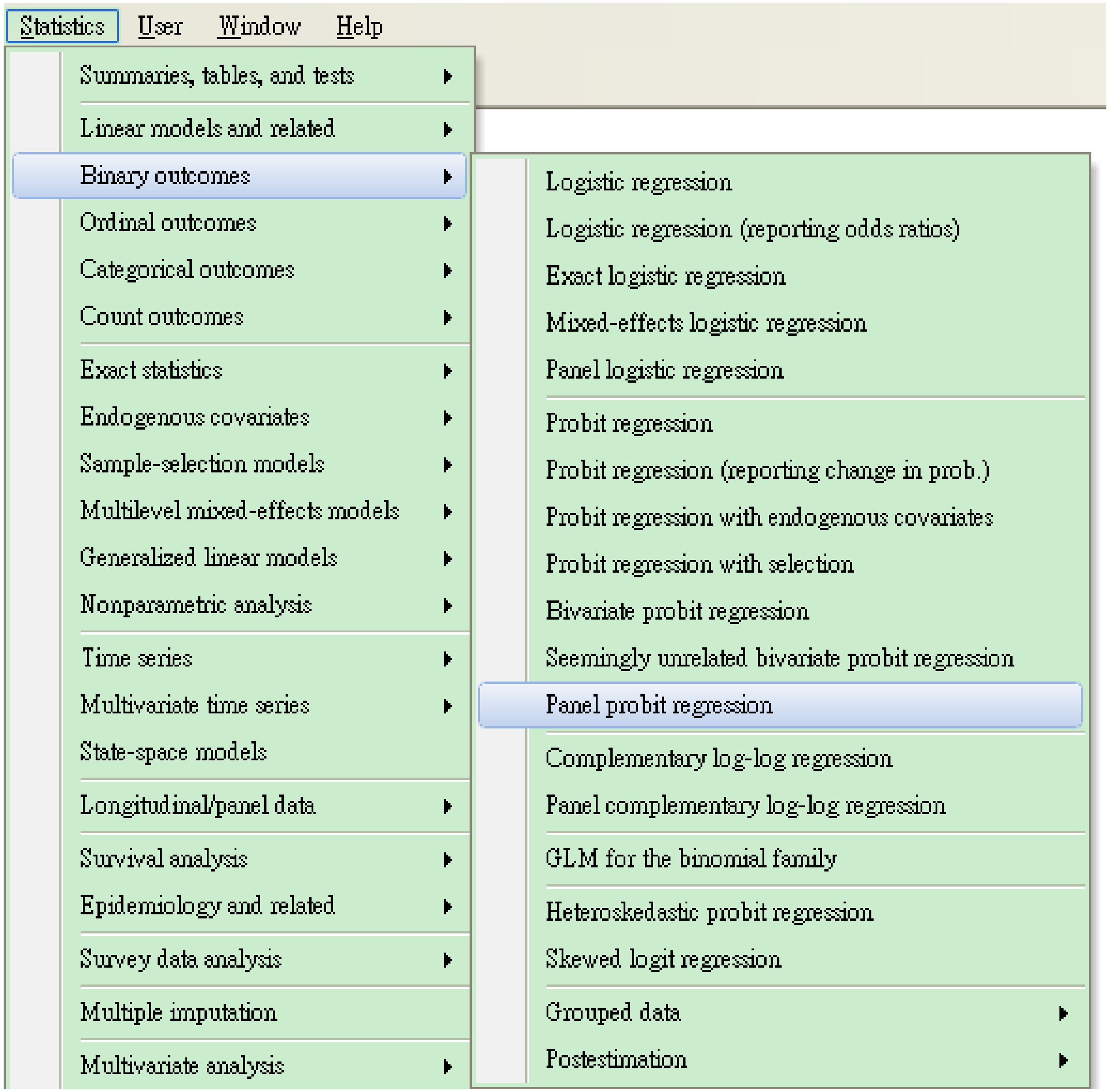
That is, by conditioning on the summation of $y_{it}$, we swept away individual effects $\mu_i$

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

# Case 1. Female union membership

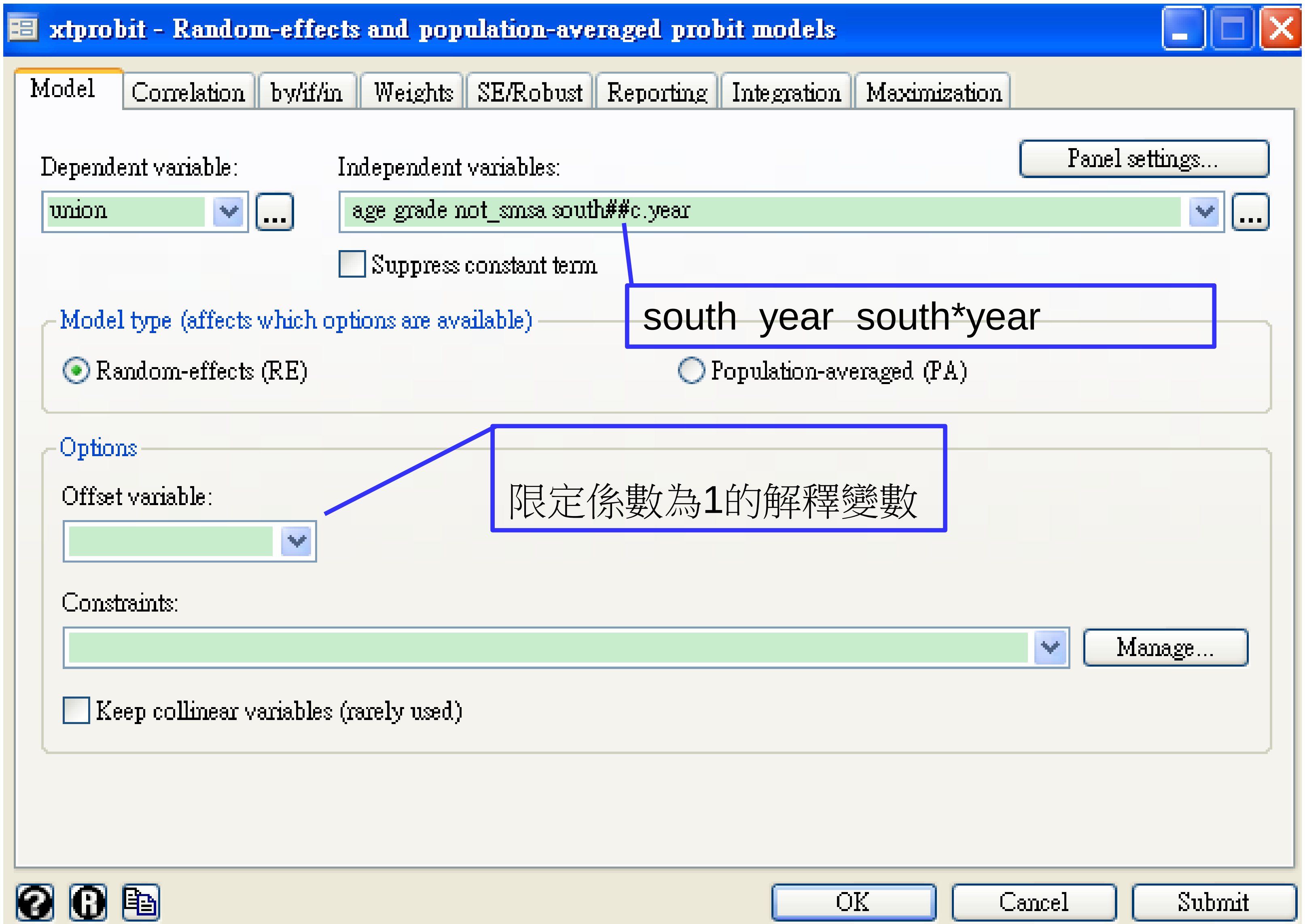
| idcode | year | age | grade | not_smsa | south | union | black |
|---|---|---|---|---|---|---|---|
| 1 | 72 | 20 | 12 | 0 | 0 | 1 | 1 |
| 1 | 77 | 25 | 12 | 0 | 0 | 0 | 1 |
| 1 | 80 | 28 | 12 | 0 | 0 | 1 | 1 |
| 1 | 83 | 31 | 12 | 0 | 0 | 1 | 1 |
| 1 | 85 | 33 | 12 | 0 | 0 | 1 | 1 |
| 1 | 87 | 35 | 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)



# Estimation

Panel probit  by random effect



```
Random-effects probit regression                Number of obs      =     26200
Group variable: idcode                          Number of groups   =      4434

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       5.9
                                                               max =        12

                                                Wald chi2(6)       =    220.91
Log likelihood  = -10552.225                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0082967   .0084599     0.98   0.327    -.0082843    .0248778
       grade |   .0482731   .0099469     4.85   0.000     .0287776    .0677686
    not_smsa |   -.139657   .0460548    -3.03   0.002    -.2299227   -.0493913
     1.south |  -1.584394    .358473    -4.42   0.000    -2.286989   -.8818002
        year |  -.0039854   .0088399    -0.45   0.652    -.0213113    .0133406
             |
south#c.year |
           1 |   .0134017   .0044622     3.00   0.003     .0046559    .0221475
             |
       _cons |  -1.668202   .4751819    -3.51   0.000    -2.599542   -.7368628
-------------+----------------------------------------------------------------
    /lnsig2u |   .6103616   .0458783                      .5204418    .7002814
-------------+----------------------------------------------------------------
     sigma_u |    1.35687   .0311255                      1.297217    1.419267
         rho |   .6480233   .0104643                      .6272511    .6682502
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  5984.32 Prob >= chibar2 = 0.000
```

$$\text{lnsig2u} = \ln(\sigma_\varepsilon^2)$$

$$rho = \rho = \frac{\sigma_\varepsilon^2}{1+\sigma_\varepsilon^2}$$

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.

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.

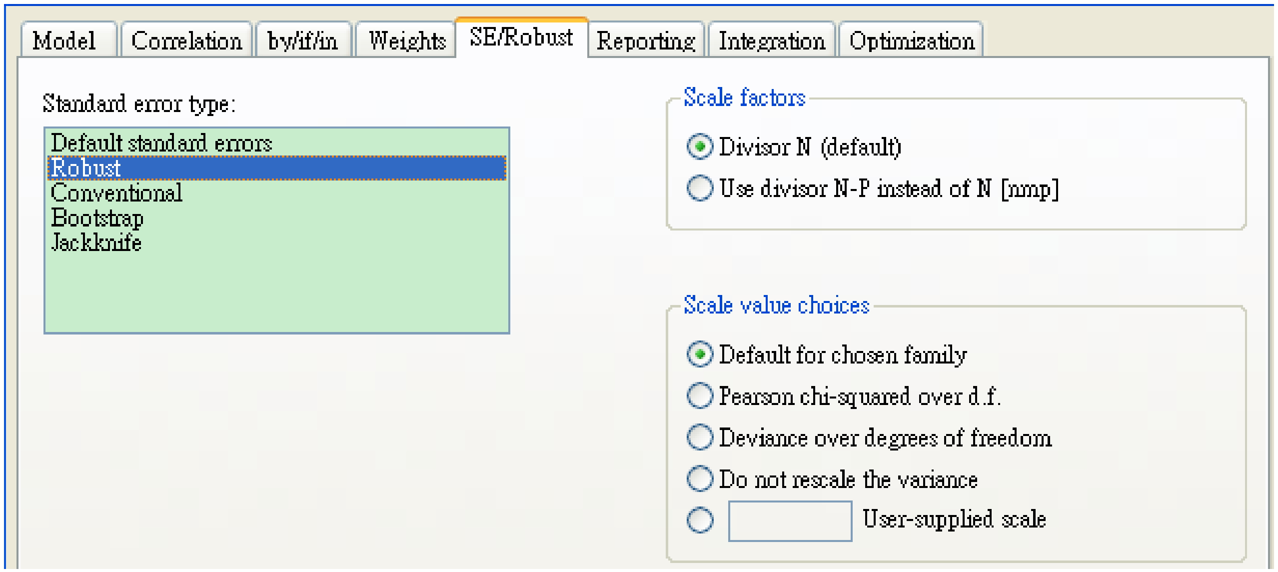
# Estimation

Panel probit by equal-correlation, population -averaged

| | | | | | |
|---|---|---|---|---|---|
| GEE population-averaged model | | | Number of obs | = | 26200 |
| Group variable: | idcode | | Number of groups | = | 4434 |
| Link: | probit | | Obs per group: min | = | 1 |
| Family: | binomial | | avg | = | 5.9 |
| Correlation: | exchangeable | | max | = | 12 |
| | | | Wald chi2(6) | = | 242.57 |
| Scale parameter: | 1 | | Prob > chi2 | = | 0.0000 |

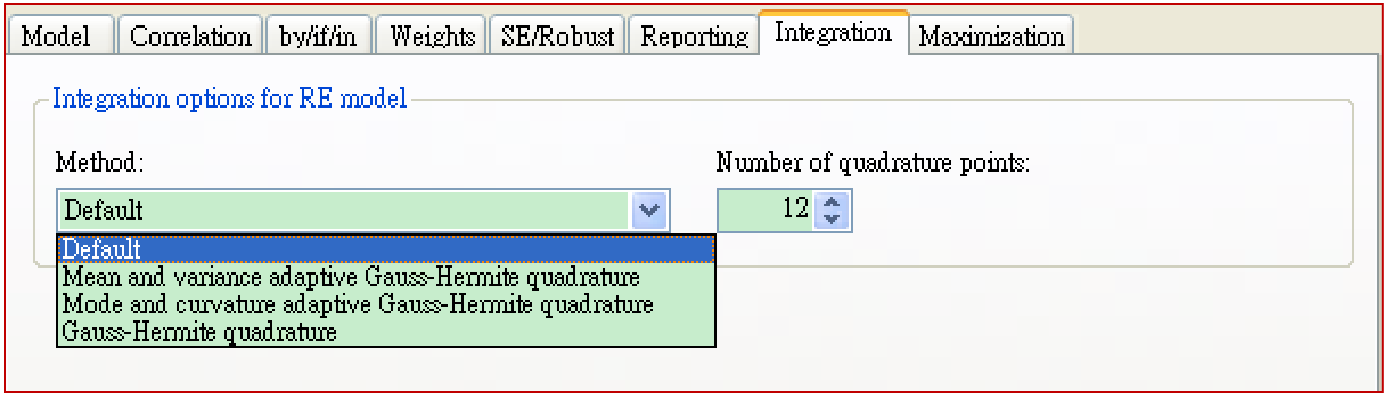
| union | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .0089699 | .0053208 | 1.69 | 0.092 | -.0014586 | .0193985 |
| grade | .0333174 | .0062352 | 5.34 | 0.000 | .0210966 | .0455382 |
| not_smsa | -.0715717 | .027543 | -2.60 | 0.009 | -.1255551 | -.0175884 |
| 1.south | -1.017368 | .207931 | -4.89 | 0.000 | -1.424905 | -.6098308 |
| year | -.0062708 | .0055314 | -1.13 | 0.257 | -.0171122 | .0045706 |
| south#c.year | | | | | | |
| 1 | .0086294 | .00258 | 3.34 | 0.001 | .0035727 | .013686 |
| _cons | -.8670997 | .294771 | -2.94 | 0.003 | -1.44484 | -.2893592 |

# Robust Covariance




```
GEE population-averaged model                   Number of obs      =     26200
Group variable:                     idcode      Number of groups   =      4434
Link:                               probit      Obs per group: min =         1
Family:                           binomial                     avg =       5.9
Correlation:                  exchangeable                     max =        12
                                                Wald chi2(6)       =    156.33
Scale parameter:                         1      Prob > chi2        =    0.0000

                                 (Std. Err. adjusted for clustering on idcode)
------------------------------------------------------------------------------
             |             Semirobust
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0089699   .0051169     1.75   0.080    -.001059    .0189988
       grade |   .0333174   .0076425     4.36   0.000    .0183383    .0482965
    not_smsa |  -.0715717   .0348659    -2.05   0.040   -.1399076   -.0032359
     1.south |  -1.017368   .3026981    -3.36   0.001   -1.610645   -.4240906
        year |  -.0062708   .0055745    -1.12   0.261   -.0171965    .0046549
             |
south#c.year |
           1 |   .0086294   .0037866     2.28   0.023    .0012078    .0160509
             |
       _cons |  -.8670997   .3243959    -2.67   0.008   -1.502904   -.2312955
------------------------------------------------------------------------------
```

# Quadrature Stability

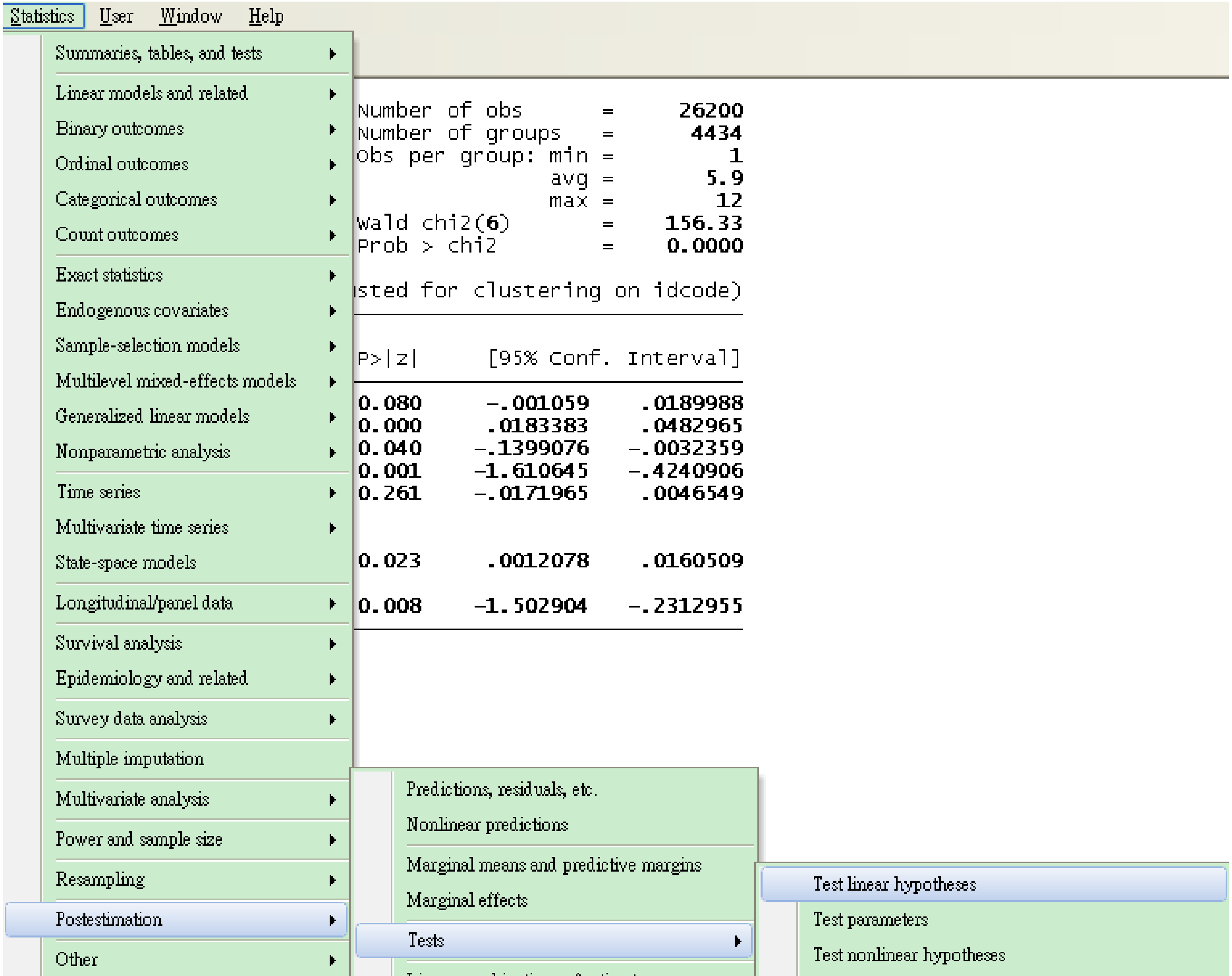


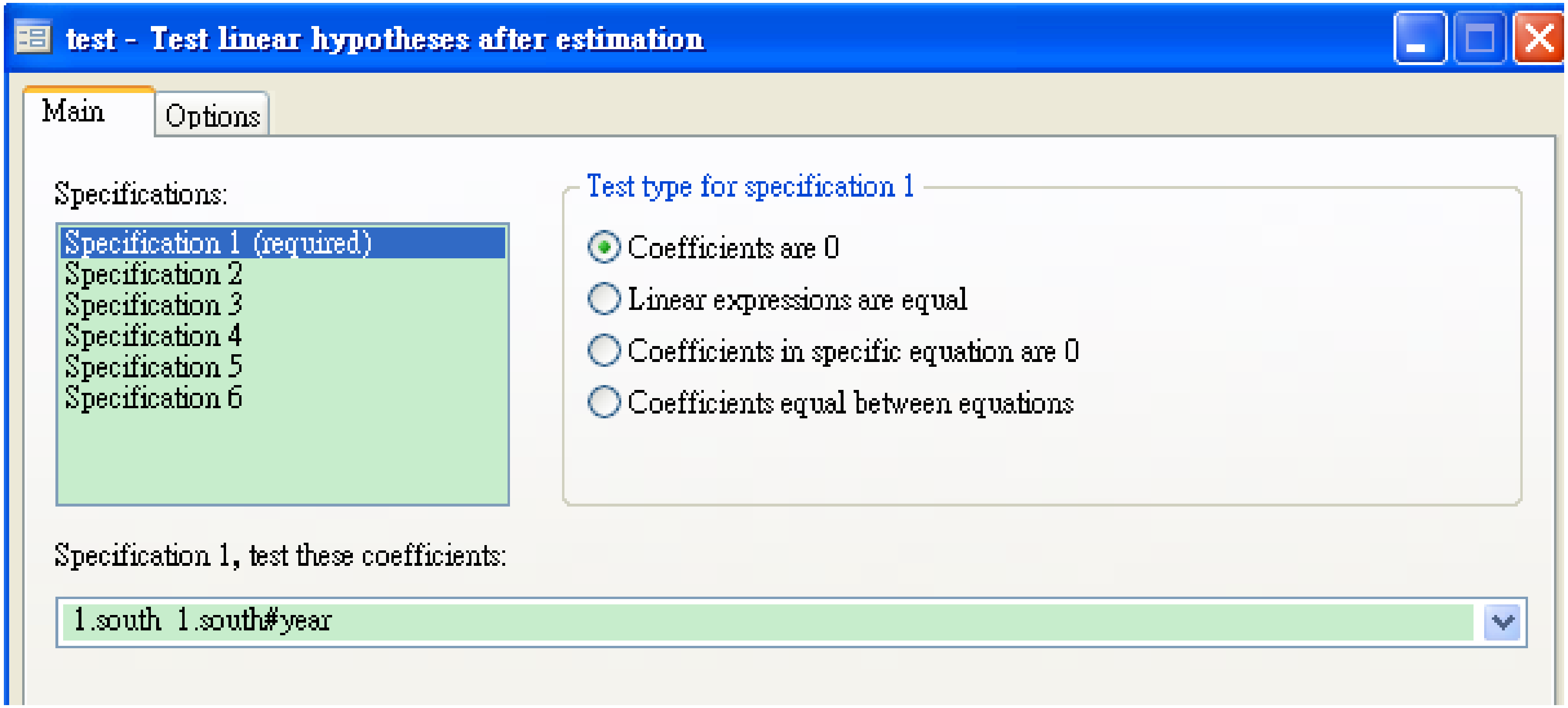

The random-effects model is calculated using quadrature, which is an approximation whose accuracy depends partially on the number of integration points used. We can use the **quadchk** command to see if changing the number of integration points affects the results. If the results change, the quadrature 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






```
. test ( 1.south  1.south#year)

 ( 1)  1.south = 0
 ( 2)  1.south#c.year = 0

           chi2(  2) =  105.22
         Prob > chi2 =    0.0000
```

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

$$\mathbf{H_0: E}(u_{it}|\mathrm{X}_{\mathrm{it}})=0$$

- The null hypothesis implies that:

**The random effects model is better**

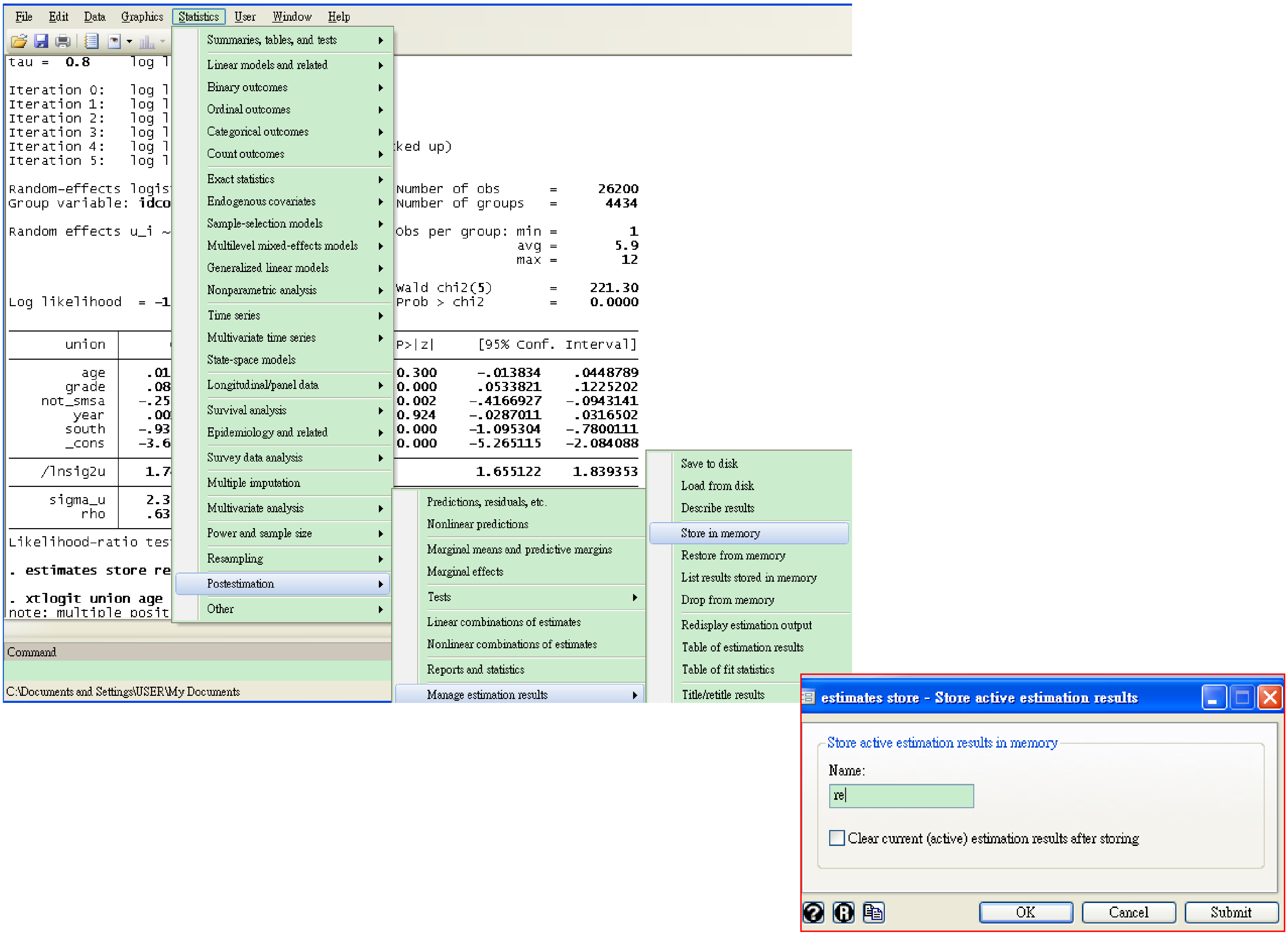
- **If the null if rejected, then the fixed effects model is a better specification.**

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

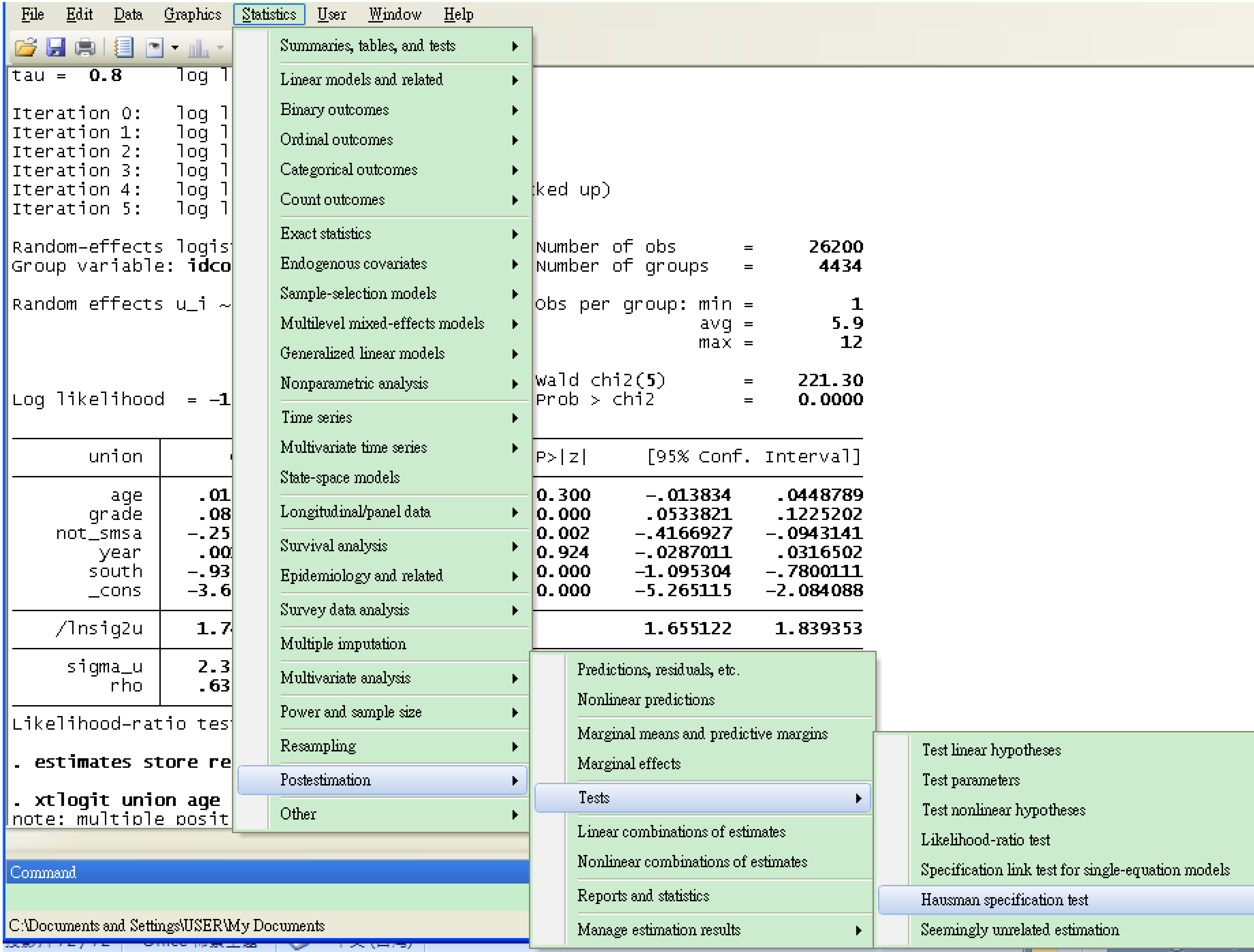
- Under $H_0$,
  - The difference of the estimates of two models are negligibly small.
  - Both LSDV and GLS are *consistent,* but LSDV are *not efficient*. Hence, the random models is the better choice.
- If we reject $H_0$, only LSDV is consistent, but not for GLS, fixed models is more applicable.

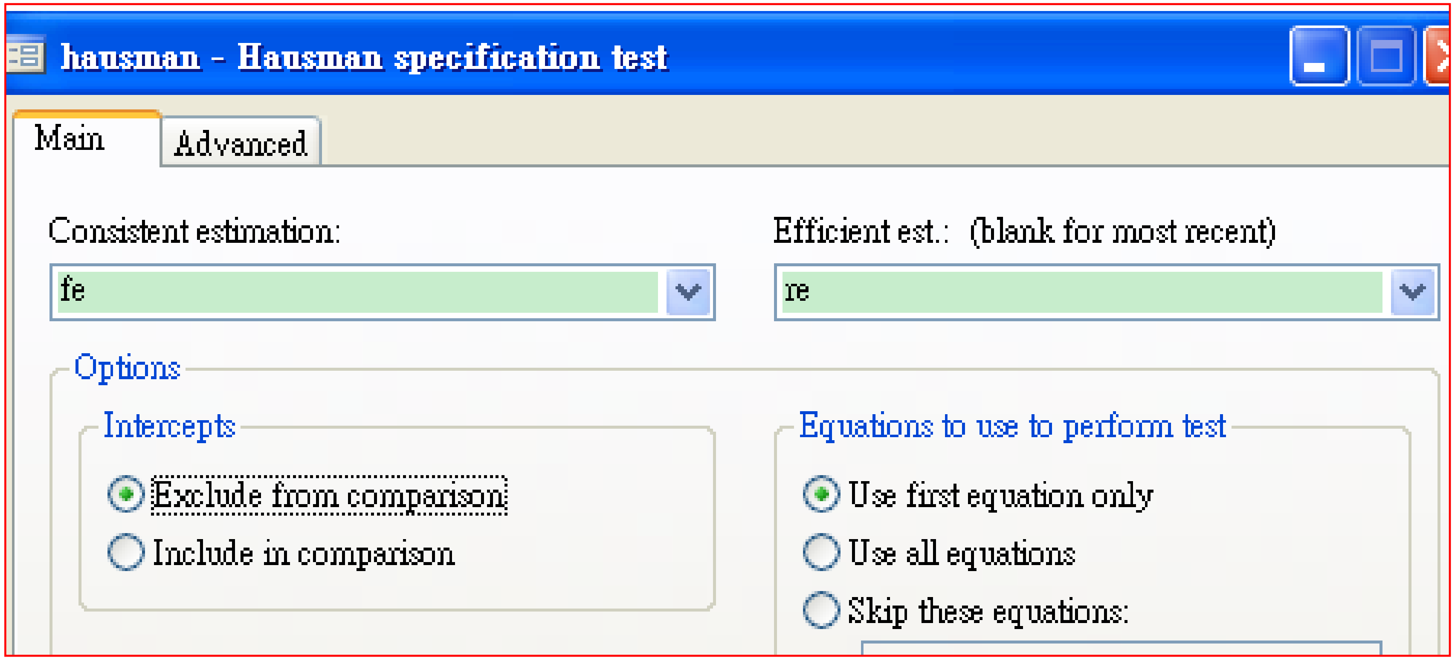
# Hausman test in action

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


File Edit Data Graphics Statistics User Window Help
tau = 0.8 log l
Iteration 0: log l
Iteration 1: log l
Iteration 2: log l
Iteration 3: log l
Iteration 4: log l
Iteration 5: log l
Random-effects logis
Group variable: idco
Random effects u_i ~
Log likelihood = -1
union
age .01
grade .08
not_smsa -.25
year .00
south -.93
_cons -3.6
/lnsig2u 1.7
sigma_u 2.3
rho .63
Likelihood-ratio tes
. estimates store re
. xtlogit union age
note: multiple posit
Command
C:\Documents and Settings\USER\My Documents
ked up)
Number of obs = 26200
Number of groups = 4434
Obs per group: min = 1
avg = 5.9
max = 12
Wald chi2(5) = 221.30
Prob > chi2 = 0.0000
P>|z| [95% Conf. Interval]
0.300 -.013834 .0448789
0.000 .0533821 .1225202
0.002 -.4166927 -.0943141
0.924 -.0287011 .0316502
0.000 -1.095304 -.7800111
0.000 -5.265115 -2.084088
1.655122 1.839353
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Exact statistics
Endogenous covariates
Sample-selection models
Multilevel mixed-effects models
Generalized linear models
Nonparametric analysis
Time series
Multivariate time series
State-space models
Longitudinal/panel data
Survival analysis
Epidemiology and related
Survey data analysis
Multiple imputation
Multivariate analysis
Power and sample size
Resampling
Postestimation
Other
Predictions, residuals, etc.
Nonlinear predictions
Marginal means and predictive margins
Marginal effects
Tests
Linear combinations of estimates
Nonlinear combinations of estimates
Reports and statistics
Manage estimation results
Save to disk
Load from disk
Describe results
Store in memory
Restore from memory
List results stored in memory
Drop from memory
Redisplay estimation output
Table of estimation results
Table of fit statistics
Title/retitle results
estimates store - Store active estimation results
Store active estimation results in memory
Name:
re
Clear current (active) estimation results after storing
OK
Cancel
Submit

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.

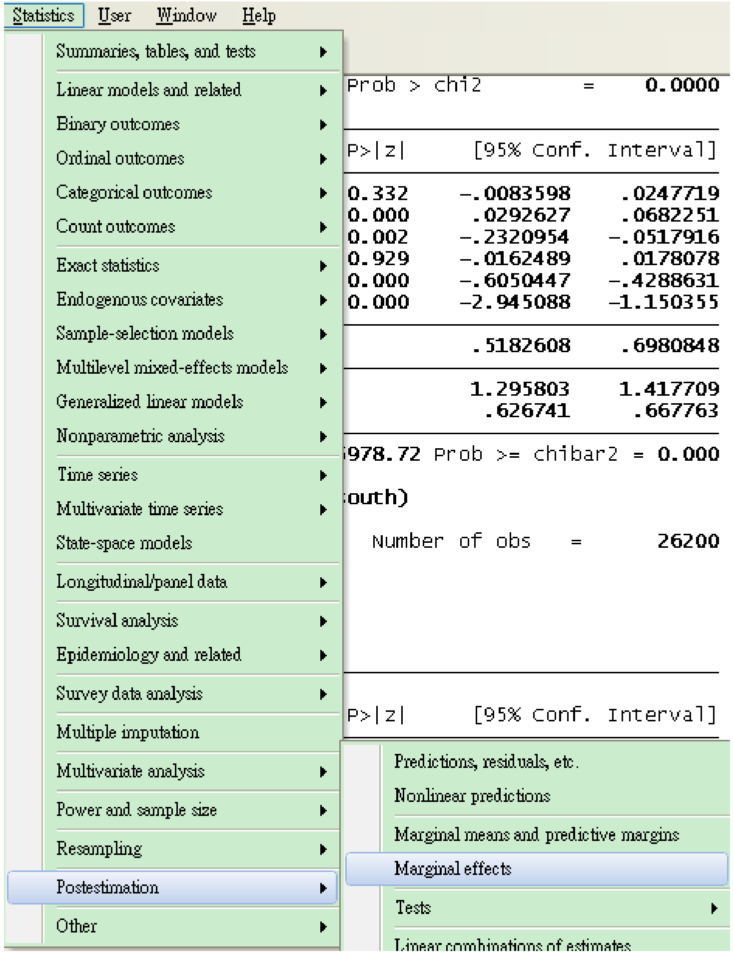

File Edit Data Graphics Statistics User Window Help
tau = 0.8 log l
Iteration 0: log l
Iteration 1: log l
Iteration 2: log l
Iteration 3: log l
Iteration 4: log l
Iteration 5: log l
Random-effects logis
Group variable: idco
Random effects u_i ~
Log likelihood = -1
union
age .01
grade .08
not_smsa -.25
year .00
south -.93
_cons -3.6
/lnsig2u 1.7
sigma_u 2.3
rho .63
Likelihood-ratio tes
. estimates store re
. xtlogit union age
note: multiple posit
Command
C:\Documents and Settings\USER\My Documents
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Exact statistics
Endogenous covariates
Sample-selection models
Multilevel mixed-effects models
Generalized linear models
Nonparametric analysis
Time series
Multivariate time series
State-space models
Longitudinal/panel data
Survival analysis
Epidemiology and related
Survey data analysis
Multiple imputation
Multivariate analysis
Power and sample size
Resampling
Postestimation
Other
ked up)
Number of obs = 26200
Number of groups = 4434
Obs per group: min = 1
avg = 5.9
max = 12
Wald chi2(5) = 221.30
Prob > chi2 = 0.0000
P>|z| [95% Conf. Interval]
0.300 -.013834 .0448789
0.000 .0533821 .1225202
0.002 -.4166927 -.0943141
0.924 -.0287011 .0316502
0.000 -1.095304 -.7800111
0.000 -5.265115 -2.084088
1.655122 1.839353
Predictions, residuals, etc.
Nonlinear predictions
Marginal means and predictive margins
Marginal effects
Tests
Linear combinations of estimates
Nonlinear combinations of estimates
Reports and statistics
Manage estimation results
Test linear hypotheses
Test parameters
Test nonlinear hypotheses
Likelihood-ratio test
Specification link test for single-equation models
Hausman specification test
Seemingly unrelated estimation

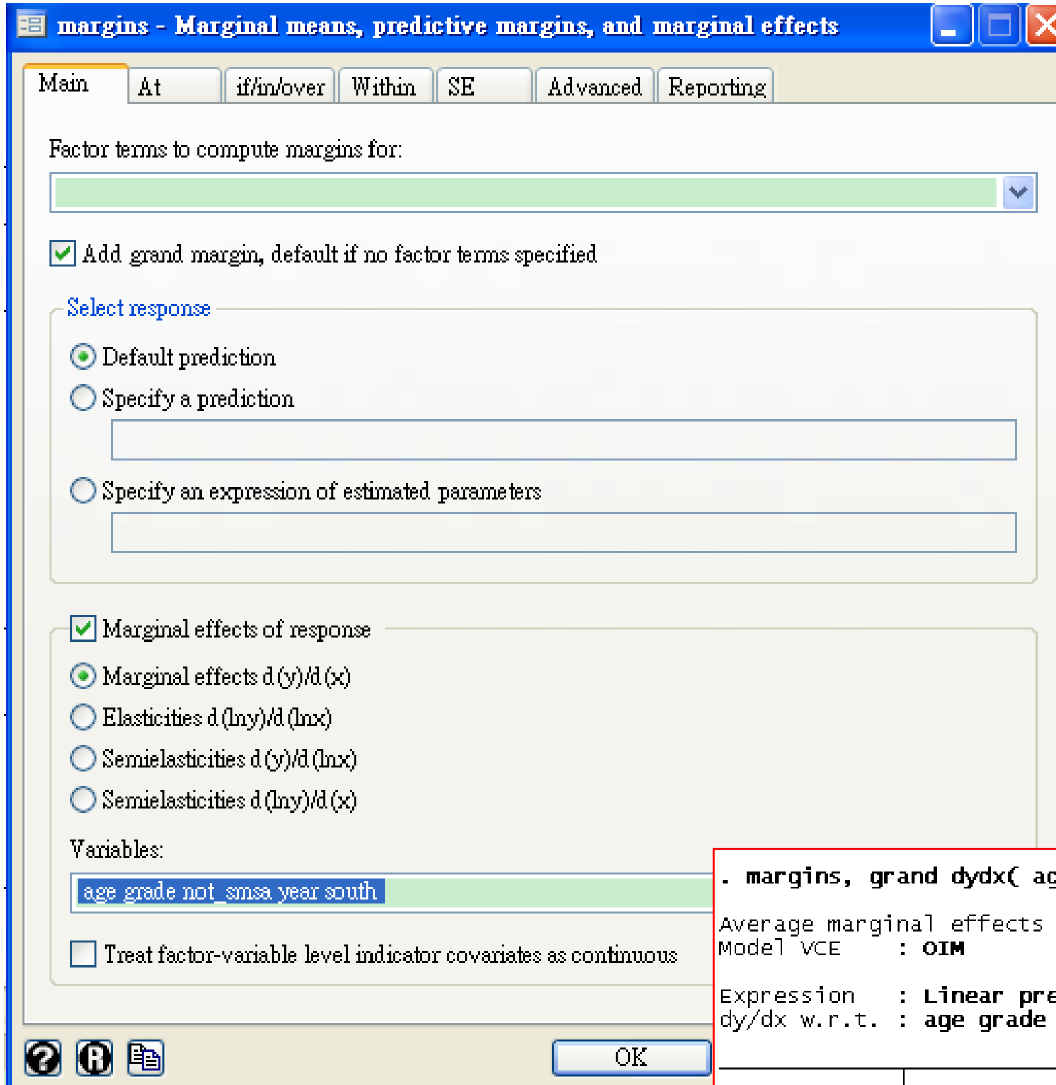



```
. hausman fe re, constant

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fe           re          Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0760241     .0155224        .0605017        .0949138
       grade |    .0857788     .0879511       -.0021723        .0379763
    not_smsa |    .0096844    -.2555034        .2651878        .0772299
        year |   -.0594931     .0014745       -.0609677        .0955826
       south |   -.7476614    -.9376577        .1899964        .0958801
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtlogit
           B = inconsistent under Ha, efficient under Ho; obtained from xtlogit

    Test:  Ho:  difference in coefficients not systematic

                  chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       16.41
                Prob>chi2 =      0.0058
```

# 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.






```
. margins, grand dydx( age grade not_smsa year south)

Average marginal effects                          Number of obs   =      26200
Model VCE    : OIM

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age grade not_smsa year south

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0082061   .0084521     0.97   0.332    -.0083598    .0247719
       grade |   .0487439   .0099396     4.90   0.000     .0292627    .0682251
    not_smsa |  -.1419435   .0459967    -3.09   0.002    -.2320954   -.0517916
        year |   .0007794   .0086881     0.09   0.929    -.0162489    .0178078
       south |  -.5169539   .0449451   -11.50   0.000    -.6050447   -.4288631
------------------------------------------------------------------------------
```

# Count Model in Panel Data

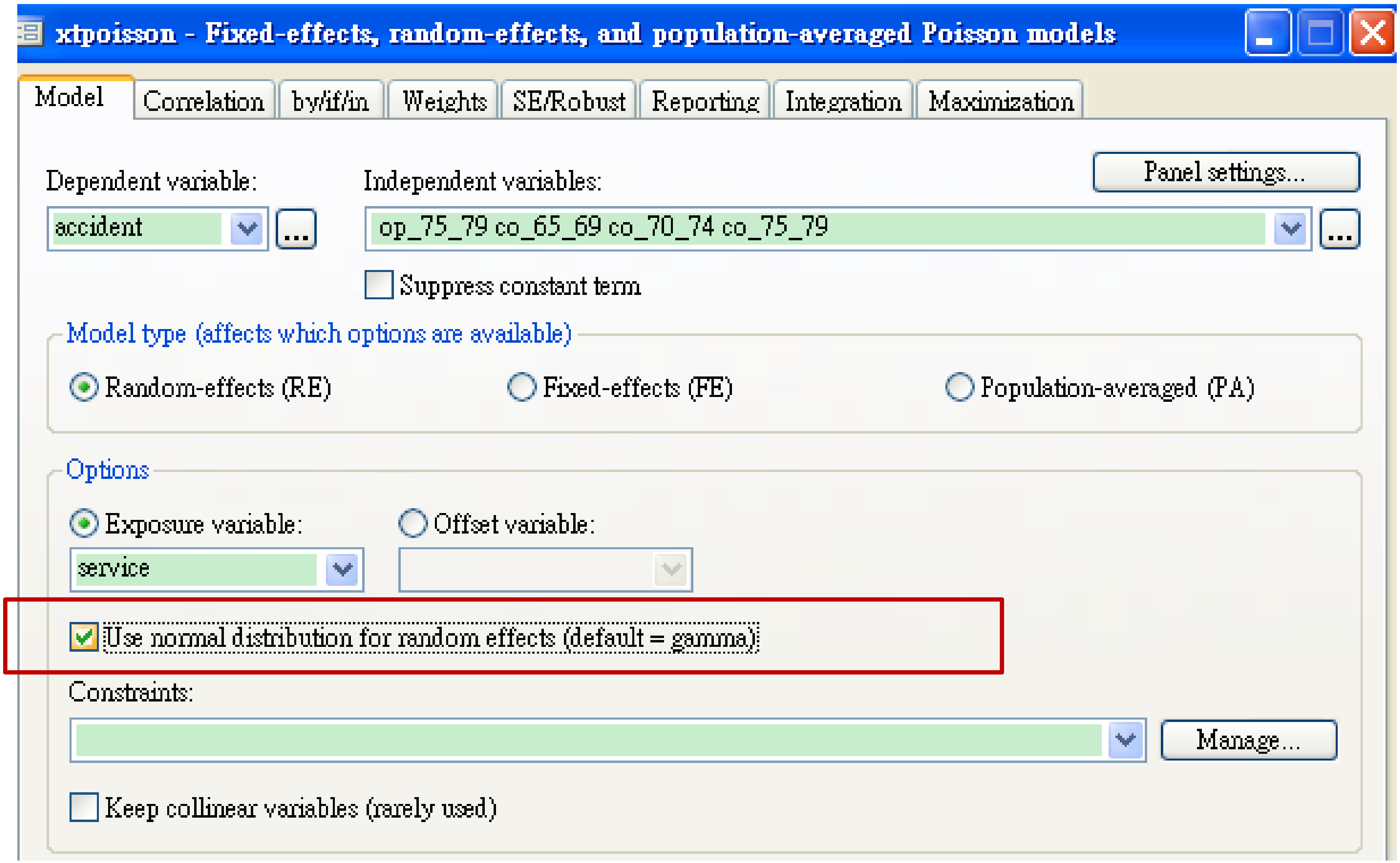
# Count panel data

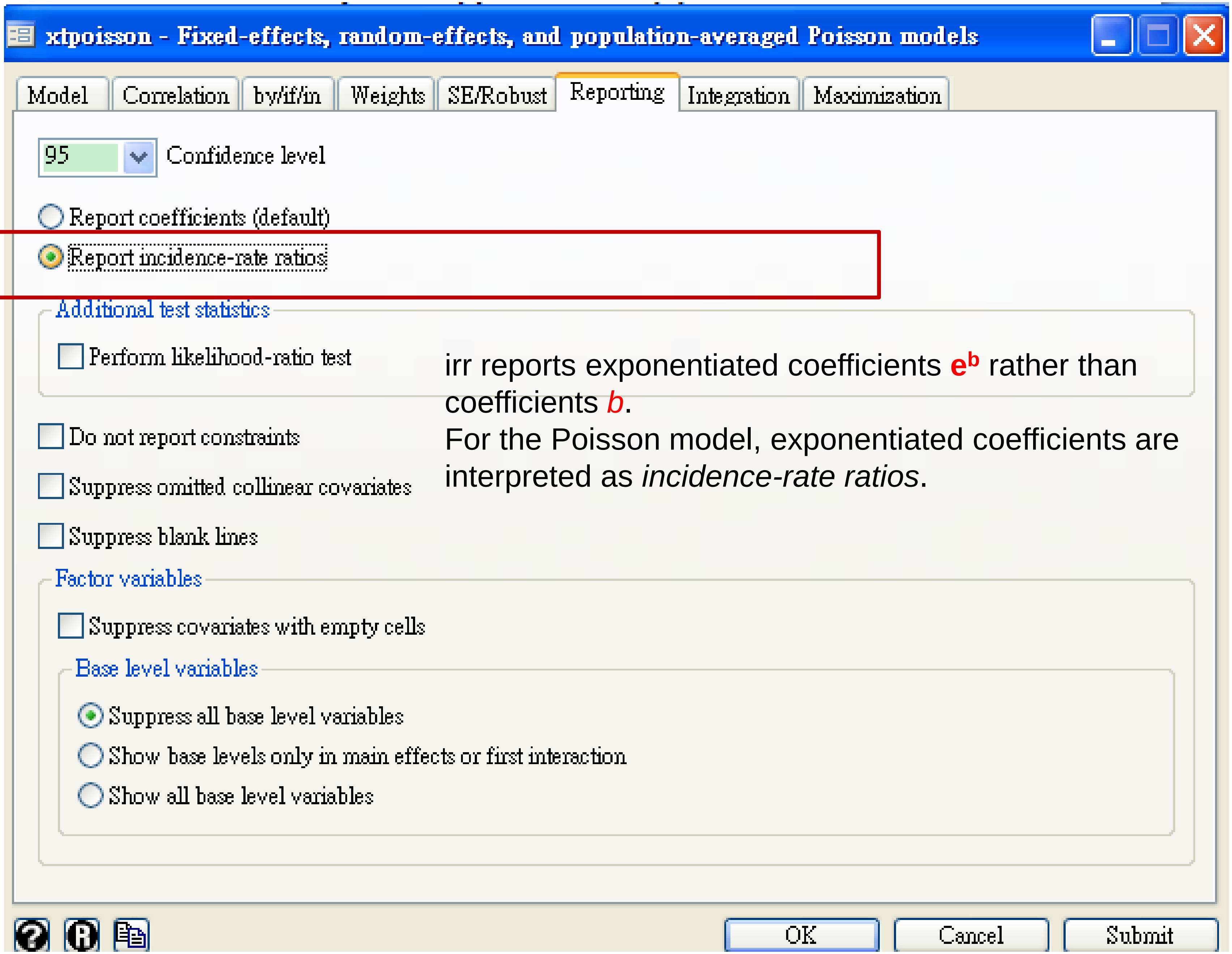
| ship | yr_con | yr_op | service | accident | op_75_79 | co_65_69 | co_70_74 | co_75_79 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 127 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 2 | 63 | 0 | 1 | 0 | 0 | 0 |
| 1 | 2 | 1 | 1095 | 3 | 0 | 1 | 0 | 0 |
| 1 | 2 | 2 | 1095 | 4 | 1 | 1 | 0 | 0 |
| 1 | 3 | 1 | 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→ longitudinal/panel data → count outcomes → **Poisson regression**






irr reports exponentiated coefficients $e^b$ rather than coefficients $b$.
For the Poisson model, exponentiated coefficients are interpreted as *incidence-rate ratios*.

```
Random-effects Poisson regression               Number of obs      =        34
Group variable: ship                            Number of groups   =         5

Random effects u_i ~ Gamma                      Obs per group: min =         6
                                                               avg =       6.8
                                                               max =         7

                                                Wald chi2(4)       =     50.90
Log likelihood  = -74.811217                    Prob > chi2        =    0.0000
```

| accident | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| op_75_79 | 1.466305 | .1734005 | 3.24 | 0.001 | 1.162957 | 1.848777 |
| co_65_69 | 2.032543 | .304083 | 4.74 | 0.000 | 1.515982 | 2.72512 |
| co_70_74 | 2.356853 | .3999259 | 5.05 | 0.000 | 1.690033 | 3.286774 |
| co_75_79 | 1.641913 | .3811398 | 2.14 | 0.033 | 1.04174 | 2.58786 |
| service | (exposure) | | | | | |
| /lnalpha | -2.368406 | .8474597 | | | -4.029397 | -.7074155 |
| alpha | .0936298 | .0793475 | | | .0177851 | .4929165 |

```
Likelihood-ratio test of alpha=0: chibar2(01) =    10.61 Prob>=chibar2 = 0.001
```

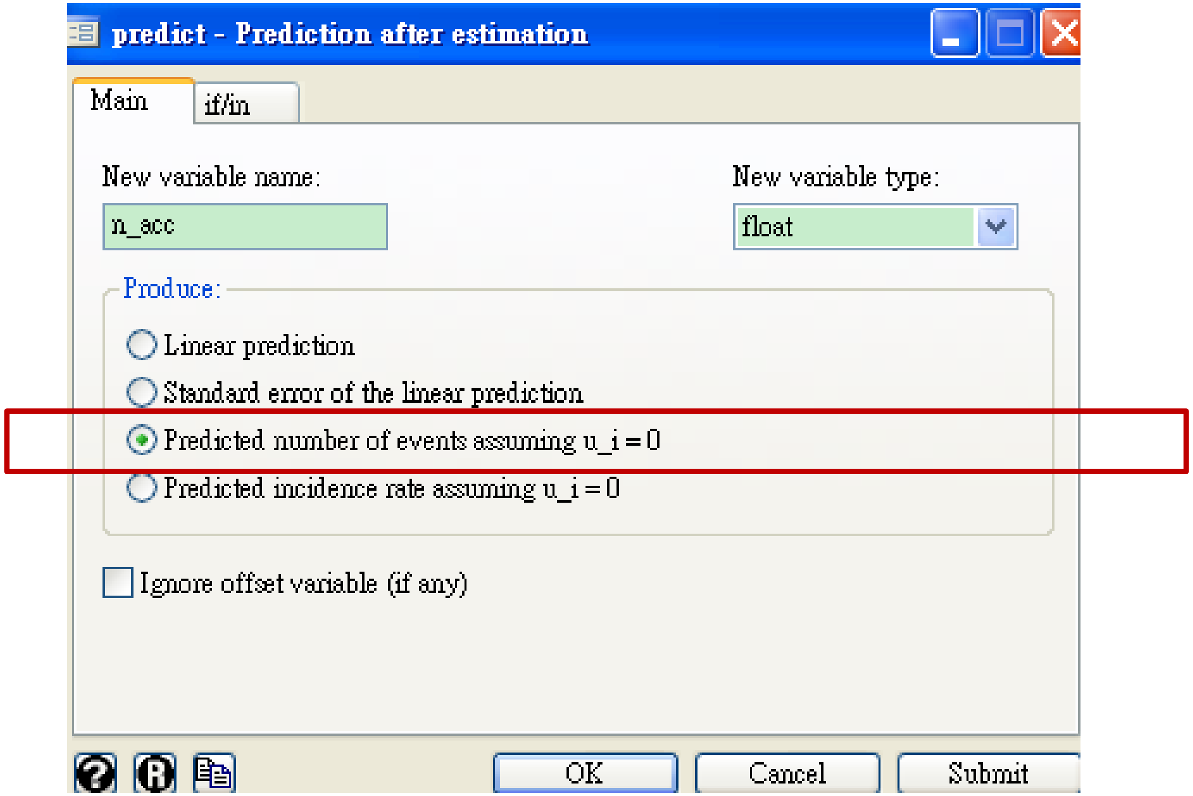
The output also includes a likelihood-ratio test of α = 0, which compares the panel estimator with the pooled (Poisson) estimator.

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

where $\lambda_{it} = \exp(\mathbf{x}_{it}\boldsymbol{\beta})$. 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




```
. predict n_acc, nu0
(6 missing values generated)
```

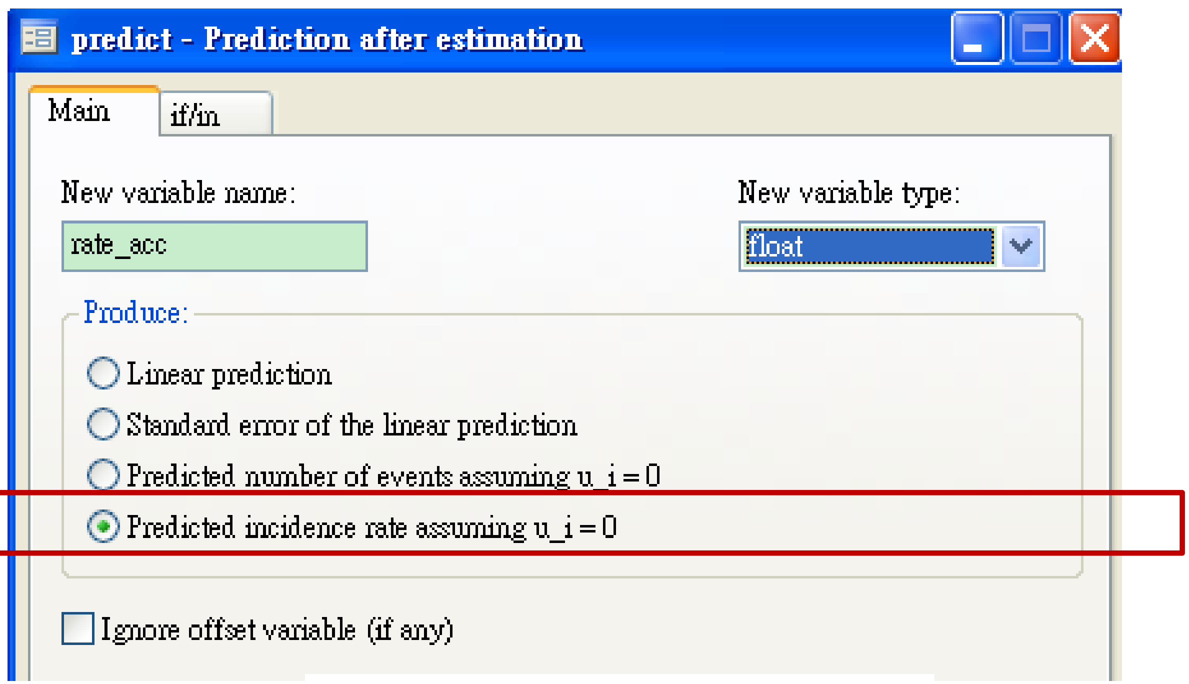
| ship | yr_con | yr_op | service | accident | op_75_79 | co_65_69 | co_70_74 | co_75_79 | n_acc |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 127 | 0 | 0 | 0 | 0 | 0 | .1742982 |
| 1 | 1 | 2 | 63 | 0 | 1 | 0 | 0 | 0 | .1267809 |
| 1 | 2 | 1 | 1095 | 3 | 0 | 1 | 0 | 0 | 3.054522 |
| 1 | 2 | 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 | 83.31905 |
| 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 |

```
. summarize n_acc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       n_acc |        34    13.52307    23.15885   .0617592   83.31905
```

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 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 | .0027895 |
| 1 | 2 | 2 | 1095 | 4 | 1 | 1 | 0 | 0 | 4.478859 | .0040903 |
| 1 | 3 | 1 | 1512 | 6 | 0 | 0 | 1 | 0 | 4.890728 | .0032346 |
| 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 | .0020124 |
| 2 | 2 | 1 | 28609 | 58 | 0 | 1 | 0 | 0 | 79.80531 | .0027895 |
| 2 | 2 | 2 | 20370 | 53 | 1 | 1 | 0 | 0 | 83.31905 | .0040903 |
| 2 | 3 | 1 | 7064 | 12 | 0 | 0 | 1 | 0 | 22.84928 | .0032346 |
| 2 | 3 | 2 | 13099 | 44 | 1 | 0 | 1 | 0 | 62.12753 | .0047429 |
| 2 | 4 | 1 | . | . | 0 | 0 | 0 | 1 | . | .0022534 |
| 2 | 4 | 2 | 7117 | 18 | 1 | 0 | 0 | 1 | 23.51583 | .0033042 |
| 3 | 1 | 1 | 1179 | 1 | 0 | 0 | 0 | 0 | 1.618091 | .0013724 |
| 3 | 1 | 2 | 552 | 1 | 1 | 0 | 0 | 0 | 1.110842 | .0020124 |
| 3 | 2 | 1 | 781 | 0 | 0 | 1 | 0 | 0 | 2.178613 | .0027895 |
| 3 | 2 | 2 | 676 | 1 | 1 | 1 | 0 | 0 | 2.765031 | .0040903 |
| 3 | 3 | 1 | 783 | 6 | 0 | 0 | 1 | 0 | 2.532699 | .0032346 |
| 3 | 3 | 2 | 1948 | 2 | 1 | 0 | 1 | 0 | 9.239211 | .0047429 |
| 3 | 4 | 1 | . | . | 0 | 0 | 0 | 1 | . | .0022534 |

```
. 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*

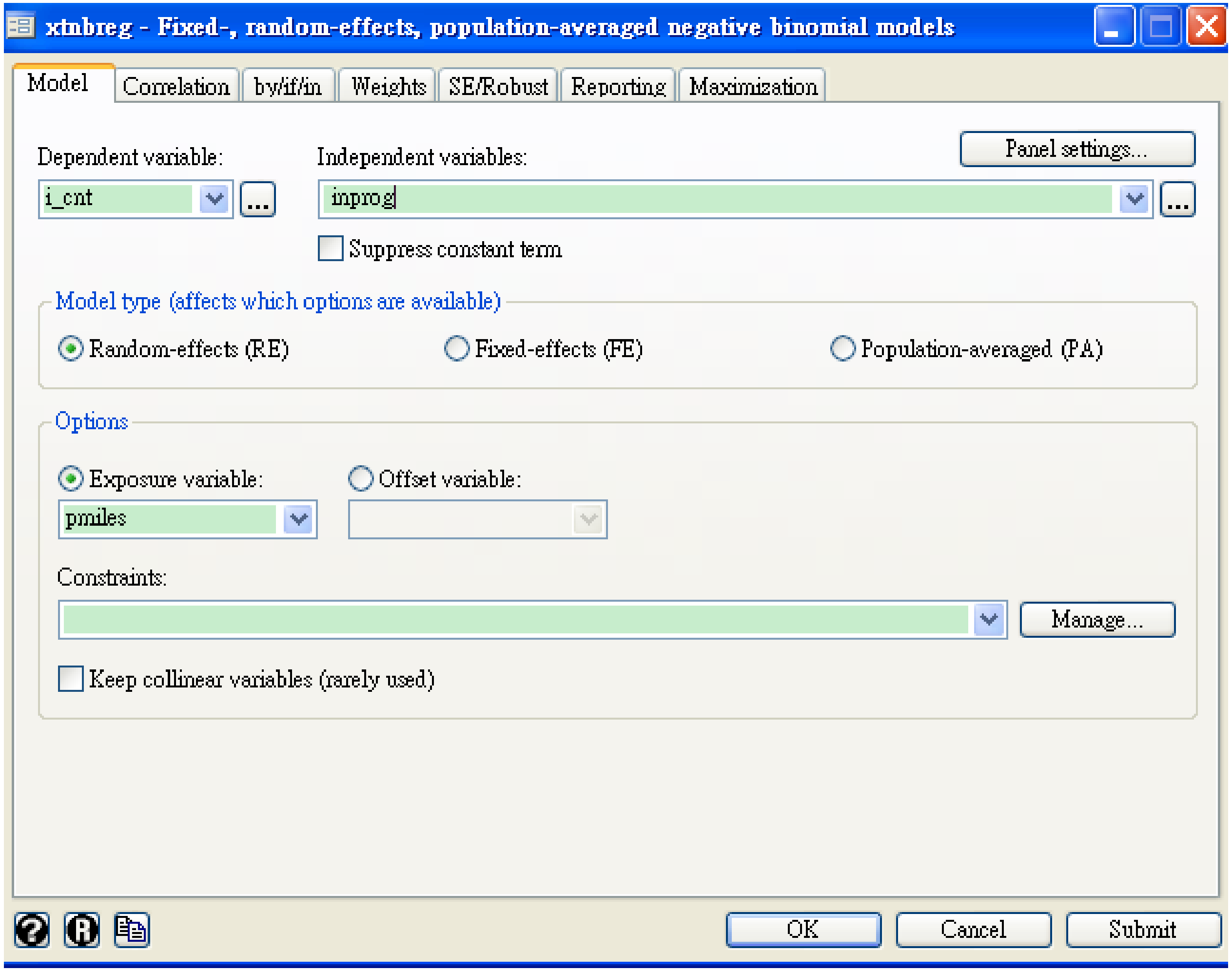
airacc.dta

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

| 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→ longitudinal/panel data → count → ***negative binominal regression***



# logit

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

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

where

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

The panel-level likelihood $l_i$ is given by

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

$$\equiv \int_{-\infty}^{\infty} g(y_{it}, x_{it}, \nu_i) d\nu_i$$

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

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

We now assume that $\epsilon_i$ follows a gamma distribution with mean one and variance $\theta$ so that unconditional on $\epsilon_i$

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

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

# 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 $y_{it} \mid \gamma_{it} \sim \text{Poisson}(\gamma_{it})$, where $\gamma_{it} \mid \delta_i \sim \text{gamma}(\lambda_{it}, \delta_i)$ with $\lambda_{it} = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + \text{offset}_{it})$ and $\delta_i$ 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 $\delta_i$ 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

$$\Pr(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i} | \mathbf{X}_i) = \int_0^\infty \prod_{t=1}^{n_i} \Pr(Y_{it} = y_{it} \mid \mathbf{x}_{it}, \delta_i)\, f(\delta_i)\, d\delta_i$$

$$= \frac{\Gamma(r+s)\Gamma(r + \sum_{t=1}^{n_i} \lambda_{it})\Gamma(s + \sum_{t=1}^{n_i} y_{it})}{\Gamma(r)\Gamma(s)\Gamma(r + s + \sum_{t=1}^{n_i} \lambda_{it} + \sum_{t=1}^{n_i} y_{it})} \prod_{t=1}^{n_i} \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}$$

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

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

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