To: Rosemarie Perez-Foster

And Thomas Borak

From: Robert A. Yaffee

Date: 3 September 2010

Subject: Multilevel Mixed Effects Random Coefficients Models for the Analysis of spatial and temporal parameters.

Acknowledgements: I have to thank Robert G. Gutierrez at Stata Corp for his gracious advice with respect to the Stata programming of such models on this matter. I also have to thank Professors Steve Raudenbush and Aline Sayer for a series of excellent lectures on this subject at ICPSR in 1994.

Multilevel mixed effects models with spatial and temporal dimensions

Research Strategy After excluding endogenous variables from the information set of the model, a multilevel mixed effects regression model is fit to explain respondent perceived Chornobyl health risk over space and time. We follow the Stata parameterization as

$$Y_{itk} = X_{itk} B_{itk} + Z_{itk} U_{itk} + E_{itk} \qquad E \sim N(\mu, \sigma 2), \qquad (1)$$

As represent fixed effects whereas Us denote random effects. The level structure is determined by separate clustering configurations of spatial and temporal autocorrelation within the spatial and temporal model levels. Level one contains the temporal (subscript t) or evolutionary process of repeated measurements, nesting time in waves within the subject level (Laird and Ware, 1982; Diggle, Liang, and Zeger, 1994; Bryk and Raudenbush, 2002). Personal expectations stem from the average of repeated measurements over time from the respondent. Each respondent (subscript i) is spatially nested within an environmental region or raion (as subscript k). A block diagonal partitioning of the covariance matrix permits fixed (time-invariant) and random effects (including time-varying, measurement error, and random samples of target populations) to be separately embedded and managed within the model. By specifying the evolutionary process within the individual as a pattern of residual autocorrelation as nested within the respondent who in turn is nested within the raion, this multilevel mixed effects model can be used for testing hypotheses and modeling psychological change and social, economic and/or environmental influences on the individual. Because this model can control for spatial autocorrelations and temporal autocorrelation, parameter estimation bias that can be greatly attenuated or eliminated when used to study the relationship between the actual and perceived risk on the part of the individual as and the mass public.

Preliminary analysis was undertaken to screen candidate explanatory variables. With exploratory graphical analysis, we studied the functional form of the relationships of the dependent with the candidate explanatory variables. Bivariate tests of the relationships were used to screen out candidate explanatory variables. A battery of endogeneity tests were performed to be sure that no candidate variable would generate endogeneity or simultaneity bias. Any variable that might have done so was excluded from the right – hand side of the equation. Multicollinearity matrices were used to double check this screen.). Explanatory variables were removed when they were found to retain endogeneity tested with a robust Hausman test and two- and three-stage least squares estimation, with two and three stage least squares simultaneous equation models to check upon simultaneity, underidentification and overidentification of the model were tested in a variety of tests. Functional form plots were used to test for linearity. An index test was also used for that purpose, available in PcGive 13 (Hendry, D. F. and Doornik, J.A., 2009.)

Objectives and tests: We test whether there may be environmental or societal phenomena associated with self-perceived Chornobyl health risk (radhlw). We are looking for evidence relating a sense of perceived distance of the residence, workplace, or general location of the respondent to the accident site.

To render the natural log of cumulative external ^{137}CS dose linear and amenable to analysis, we apply a simple natural log transformation to it. We use the nesting for the multilevel mixed effect regression equations to test distance related effects of a contextual type. The raion means of these variables provide expected contextual or environmental impacts on the intercepts and slopes of our model. In addition to finding significant effects on the intercept and/or rate of change of the model, we could use cross-level interactions to test and analyze contextual effects.

Preliminary findings: From the output on the following page, we observe significant macro-level environmental associations with the amount of air and water polluted by Chornobyl and the natural log of cumulative external dose of ^{137}Caesium. Among the significant macro-level effects we note the natural log of cumulative reconstructed external dose of ^{137}CS absorbed by the respondent along with the amount of air and water believed to have been polluted by Chornobyl. The effects of age, gender, external dose of ^{137}CS, any injury the respondent reports that he or she suffered as a result of Chornobyl and the interference with his or her interests and hobbies on account of this tragedy emerge as significant. The growth rate of self-perceived Chornobyl health risk is significant from wave one to wave two, but levels off into nonsignificance after wave three begins. The net effect over time is therefore best estimated by a linear growth trend over the 31 year period of less than a 10 percentage point increase. If we formulate the relationship as a regression line, shown below, the intercept of the vertical axis, α, begins at

$$\text{Radhlw}_{itk} = \alpha_{itk\,+}\ \ \beta\ \text{wave}_{itk} + \text{(other fixed effects that need to be controlled)} + \ \ e_{itk} \quad (2)$$

about 52 percent but does not exceed 60% for thirty-one years. Whether or not, β, the slope coefficient for wave, is statistically significant, we retain that as a temporal reference for the rate of change to keep track of the magnitude of the rate of change. To track the trajectory of change for an individual, we examine the within-subject effect of nesting time within subject among the residuals from our level one model. After observing a high first order residual autocorrelation of more than 0.70, we apply an autocorrelation structure to help control for bias this condition could engender. It is possible that there was second order autocorrelation that was not accounted for, but we did not have enough waves to control for more than first-order autocorrelation. Therefore, borderline significances need further testing and possibly sensitivity analysis to accompany them. By using equation two as basis for a first order time trend for our subjects, we are able to nest both fixed and random effects on the individual into the raion in which he works and resides. From our raion centered covariates we avoid the need for an intercept that could have collinearity with the random intercept and/or the random slope as much as possible. If we perform this analysis for each raion, we get a raion-specific regression equation, which provides the data for a between-raion analysis. Thus, we can observe what variation exists among raions.

Stata combines the level one and level two by allowing the analyst some control in selecting the error covariance structure. From the temporal nesting of multiple waves within the subject level, we find a level one subject effect. To accommodate the contextual level of impact, the subject is nested within his geographic locale, the raion. Thus, the lower level equation five contains a random intercept, γ, to handle impacts of fixed and some random effects, as well as random slope, to accommodate interactions between random effects and the wave variable as shown in equation four.

When we substitute the higher level into the lower level equations, we obtain one equation for the multilevel mixed effects model.

$$\begin{aligned} Radhlw_{itk} = {} & \gamma_{00} + \beta_1 wave_t + b_2 age + b_3 sex + b_4 injselfr + b_5 hp2hobint \\ & + \gamma_6 lcumdosew_t \\ & + \pi_{00}\, wave_t \qquad\qquad \textit{Eq.3} \\ & + \omega_{11} wave_t * rmc\_airw_t + \omega_{12} wave * lcumdosew_t + \zeta_{00} + \zeta_{01} + e_{itk} \end{aligned}$$

The upper panel of Table 1 reveals the nesting structure of the panels in the multilevel model. The spatial autocorrelation clustering comprises 22 raions within which reside 266 respondents who answered all questions used to build the model. It also indicates the number of units of analysis within each cluster. The nesting order proceeds upward in the table.

The repeated waves in the temporal dimension are accommodated by the ar (1) error covariance structure. From Table 1, reveals in the upper panel the level structure of the model. There are 22 raions within which are nested 266 persons with complete answers on all variables included within the model.

In the second panel, the fixed effects (X'B), which define the mean structure of the marginal model, are listed. The overall significance is tested with a multivariate Wald test above the second panel on the right. The output can be interpreted as a standard regression output, with statistical significance of the individual parameters indicated by the p-values. When the fixed effects are graphed against the waves, there is a slight but not significant increase in slope throughout the study, shown in Figure one

Equation 4: Components of the model

Subject level: Fixed effects for the mean structure

$$Radhlw_{itk} = 48.7 + 0.99wave_t + 7.19rmcsex + 0.27rmcage + 25.24injselfr + 18.10rmchp2inthob$$

$$+ \ 14.65rmcLcumdosew_t + e_{itk}$$

Raion level: Random environmental effects of the variance-covariance structure (raion mean centered effects)

$$+ 200.16(\text{constant at raion level}) + 106.92Lcumdosew\ (\text{raion lev.}) + 0.10\ airw\ (\text{raion lev.}) + \zeta_{itk}$$

Subject level: Repeated temporal effects within the subject

$$\mathbf{+} \ 2.95e\text{-}15(\text{constant at the subject level}) + 3.38\ wave_t + \zeta_{itk}$$

Although the wave parameter is not statistically significant, it is retained because it forms the basis of the temporal dimension of the model. The age parameter is not statistically significant either, but it is the basis for many biological differences between the sexes for which we may need a control later. Therefore, we retain this variable as well.

In the third panel of Table 1, the random effects are displayed a subpanel for each layer. The clustering variable is listed in the upper left. At each level, the random intercept is called the constant. This represents the average level or the random intercept at that level. Below the layer panels is the model residual.

When covariates are listed at that level they are interacting with the clustering variable at that level to impact the self-perceived Chornobyl health risk. When the residual structure at the upper levels is identity, the variables are organized in block diagonal form so they would be independent of one another. When covariates are listed within these levels, they are random effects of the parameters. With repeated measures, there is considerable autocorrelation from one wave to the other, the bias for which is attenuated by the fitting of a first-order autoregression process for the residuals.

The raion level effects following raion mean centering can be found in Figure three.

Model estimation with maximum likelihood was performed with two algorithms. A combination of Broyden-Fletcher-Goldfarb-Shanno and Newton-Raphson algorithms improved convergence and model estimation. We tested the likelihood ratio of sets of variables of nested models, which included the unconditional models with unstructured covariances, models with fixed effects, models with mixed effects, models with mixed spatial effects and autocorrelated structures.

Residual diagnosis: Model robustness was tested with the help of residual diagnosis. The residuals were tested for multicollinearity (correlation matrix), endogeneity(with a C test, Kleibergen-Paap (2006) rank statistic, and a Durbin-Hausman-Wu test, robust Durbin-Hausman-Wu test for endogeneity (Cameron and Trivedi, 2009) along with two- and three-stage least squares models to confirm possible endogeneity, normality (Shapiro Wilk test, Kolmogorov Smirnov test) , heteroskedasticity (Breusch-Pagan test), and specification error (Ramsey reset test. We found that the residuals were less than well-behaved. They included six negative outliers and one positive one. Nor were they normally distributed, according to a Shapiro Wilk test ($p < .003$). However they did not harbor heteroskedasticity, according to a Breusch-Pagan test (no $p < .05$). We expect that as more data comes in the outliers will have less effect and that the residuals may become somewhat better behaved.

Empirical Bayes optimization

However, we use the intraclass correlation coefficient as a reliability coefficient with which to tune our estimate/prediction.   Let the intraclasss correlation, the proportion of the variance explained (between subjects variance over the total variance), be λ, with one adjustment---

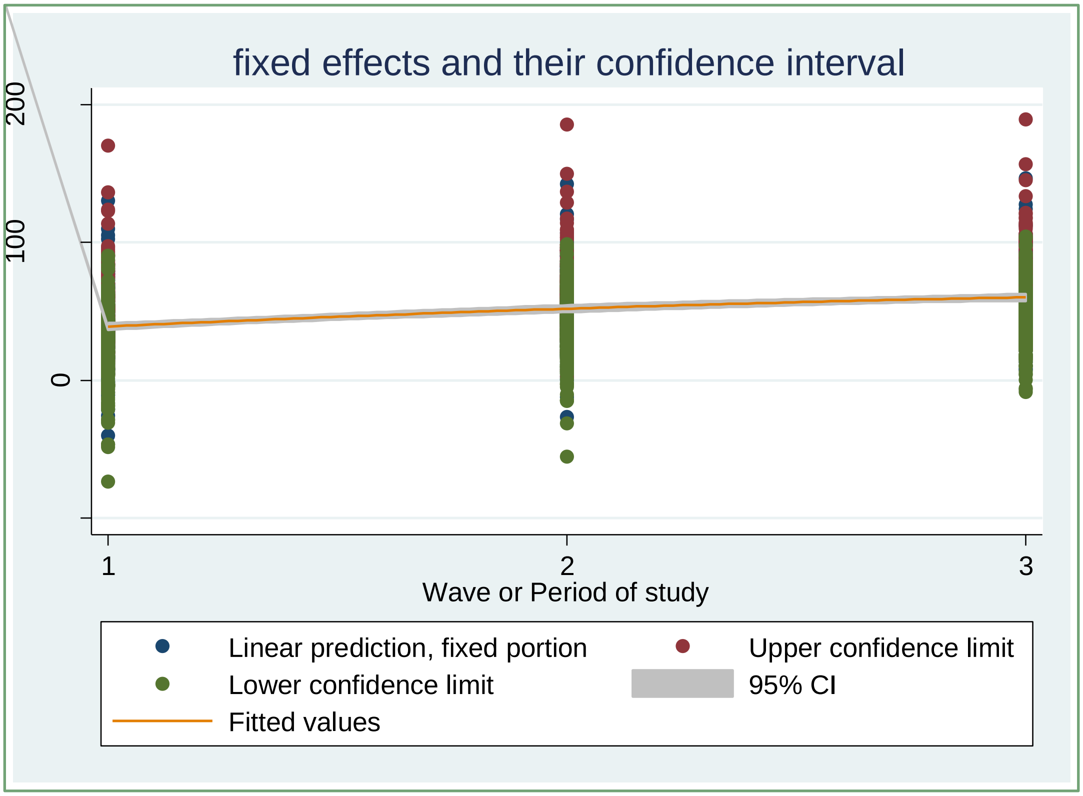
$$\lambda = \frac{Var(U)}{Var(Y)} = \frac{\sigma_u^2}{\sigma_y^2} = \frac{\sigma_u^2}{\sigma_u^2 + \left(\sigma_e^2 / n \right)}$$

then, according to Bryk and Raudenbush, 2002) the best unbiased linear predictor (BLUP) is

$$\text{BLUP} = \lambda_j \bar{Y}_{.j} - (1 - \lambda)(w_{00} + w_{01} X_j). \qquad \text{Eq. (5)}$$

The BLUP is a weighted (according to the inverse of the parameter variance) average of the sample data and the fitted values from the model. When reliability is high, the model estimate is given relatively more weight but with reliability low, the estimate is given less weight by this weighted average. In this configuration, the λ, sometimes called the empirical Bayesian shrinkage factor, optimizes our model estimate.   The BLUPS are easily generated with Stata and these generated to attenuate the slopes in the model inherent in the fixed or fitted effects (Figures 1 and 2).   After empirical Bayes estimation of the random effects at the Raion level, the random effects are graphed against the waves to show the effects of those particular parameters (Figure 3). In this way, we endeavored to robustify our model.

Figure 1: Fixed effects v. self-perceived Chornobyl health risk



Tentative inferences:

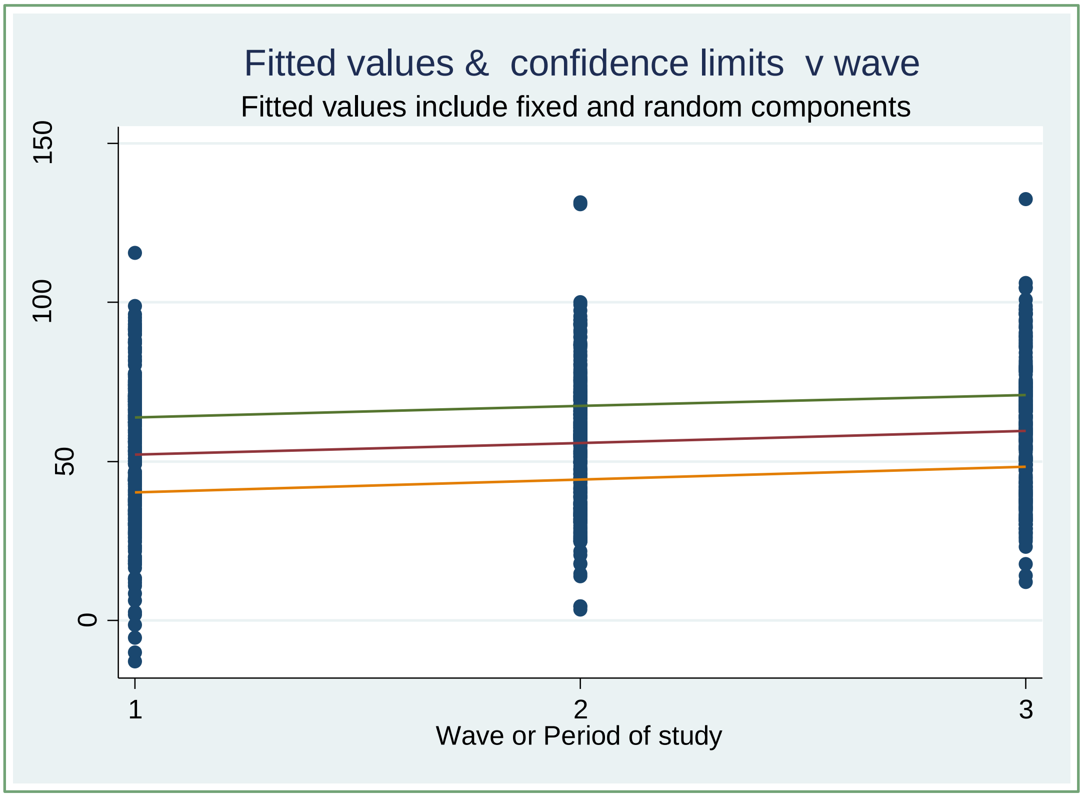
The multilevel mixed effects model added value to our analysis. Compared to a regression analysis, this model explained more of the variance of self-perceived Chornobyl health risk (χ2 = 455.22 with prob > χ^2 = 0.0000). It did so by opening spatial and temporal dimensions for simultaneous consideration. This demonstrates that there are spatial aspects to this problem that merit further exploration. It also shows that the level of this self-perceived self-health Chornobyl, on an environmental level, appears to be statistically related to external cumulative exposure as well as to the percent of air and water polluted by Chornobyl radiation directly or indirectly.

Notwithstanding covariation between perceived health risk and reconstructed cumulative external dose, the absolute level of cumulative dose seems tiny. When the natural log is retransformed back into µGrays, the effects seems to be minute, with the highest reconstructed cumulative external dose from the radioactive aftermath of Chornobyl, calculated at this time of data collection, at only 26.6 µGrays over a 23 year period. According to the Health Physics website,

> For a dental panoramic radiograph, the effective dose is 26 microSv (microsevert), which is the equivalent of about 3.3 days of natural background radiation. A series of four intraoral radiographs is 38 microSv, which is the equivalent of 4.8 days of background radiation. To put this in perspective, the effective dose from a chest radiograph is 80 microSv (10 days) and from a lower gastrointestinal (GI) series 4,060 microSv (507 days).

We hope to run this model again with more data and to perform out-of-sample predictive validation and perhaps pattern-mixture modeling to ascertain its robustness. In the meantime, it provides some evidence, albeit not perfect, that we should pursue our space-time analysis of the ecological impact of the perception of health risk from the Chornobyl radiation.

Figure 2: Model estimates of the self-perceived Chornobyl health risks across the three waves



For the time being, this is what we hoped to find. We shall proceed with our exploration of time and space with our models and reassess the data as it is collected and cleaned. We will begin to explore a Markov Chain Monte Carlo approach to Bayesian Hierarchical Linear and NonLinear Models.

Figure 3: Empirical Bayes Best Linear Unbiased Predictors

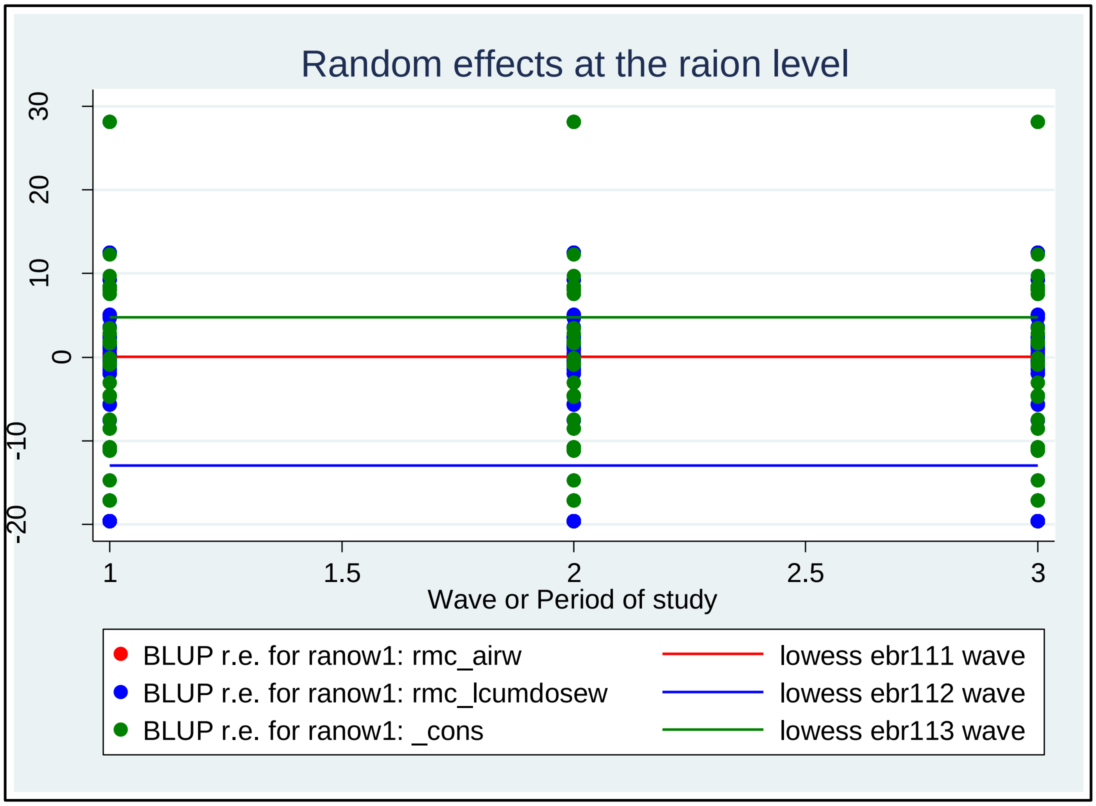
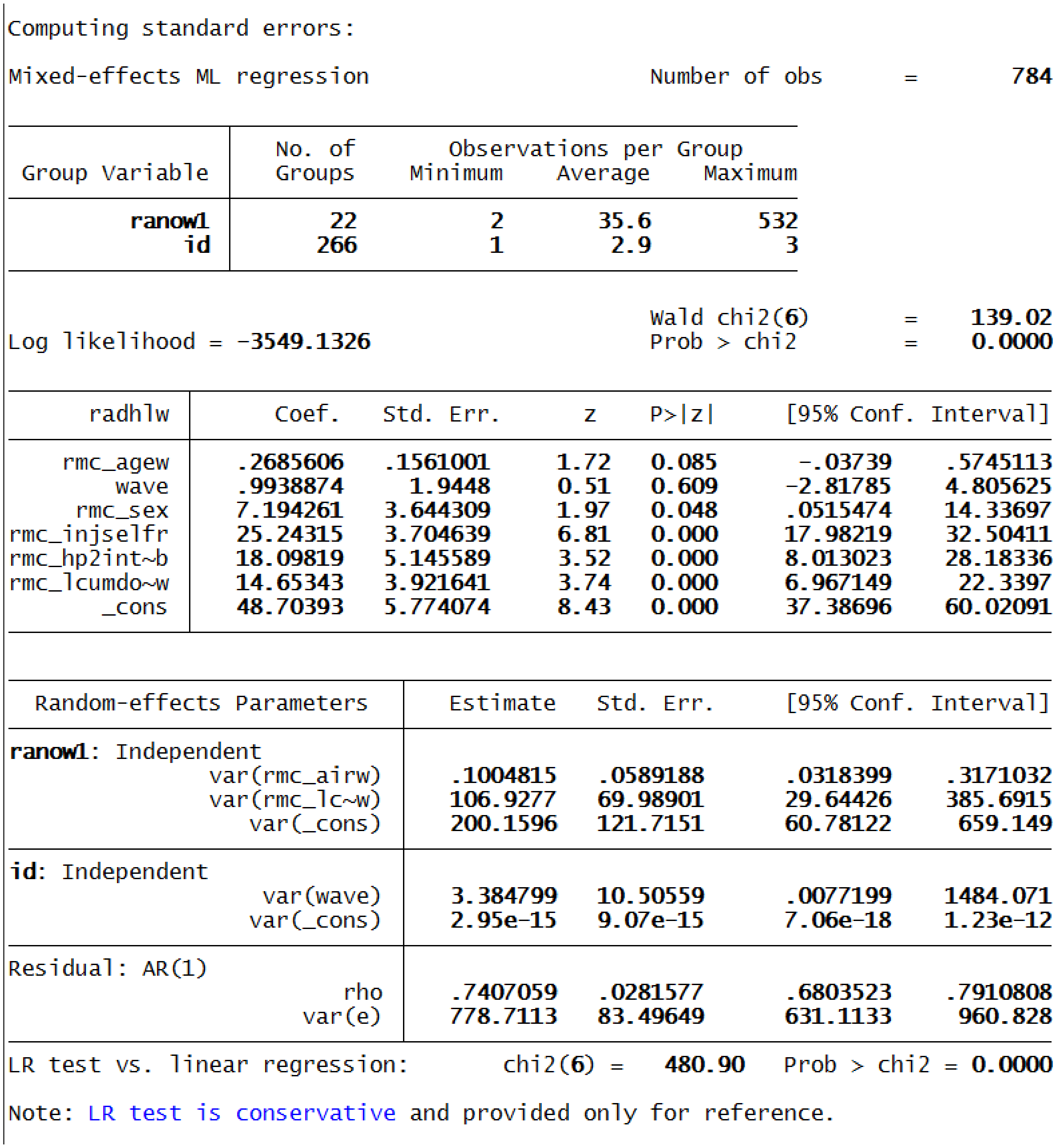


Table 1: Multilevel mixed effects regression analysis



```
Computing standard errors:

Mixed-effects ML regression                     Number of obs      =       784

-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
         ranow1 |       22          2       35.6        532
             id |      266          1        2.9          3
-----------------------------------------------------------

                                                Wald chi2(6)       =    139.02
Log likelihood = -3549.1326                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      radhlw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rmc_agew |   .2685606   .1561001     1.72   0.085     -.03739     .5745113
        wave |   .9938874     1.9448     0.51   0.609    -2.81785     4.805625
     rmc_sex |   7.194261   3.644309     1.97   0.048    .0515474     14.33697
rmc_injselfr |   25.24315   3.704639     6.81   0.000    17.98219     32.50411
rmc_hp2int~b |   18.09819   5.145589     3.52   0.000    8.013023     28.18336
rmc_lcumdo~w |   14.65343   3.921641     3.74   0.000    6.967149      22.3397
       _cons |   48.70393   5.774074     8.43   0.000    37.38696     60.02091
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
ranow1: Independent          |
               var(rmc_airw) |   .1004815   .0589188      .0318399     .3171032
               var(rmc_lc~w) |   106.9277   69.98901      29.64426     385.6915
                   var(_cons) |   200.1596   121.7151      60.78122      659.149
-----------------------------+------------------------------------------------
id: Independent              |
                   var(wave) |   3.384799   10.50559      .0077199     1484.071
                  var(_cons) |   2.95e-15   9.07e-15      7.06e-18     1.23e-12
-----------------------------+------------------------------------------------
Residual: AR(1)              |
                         rho |   .7407059   .0281577      .6803523     .7910808
                      var(e) |   778.7113   83.49649      631.1133      960.828
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(6) =   480.90   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

References:

Arrelano, M. and Bond, S. with Hendry, DF and Doornik, JA. (2009) “Static and Dynamic Panel Data Models” in Doornik and Hendry, Econometric Modeling with PcGive, vol III. London, UK: Timberlake Consultants, Ltd: Part IV, chapters 7 through 12.

Bryk, A. and Raudenbush, S. (2002). Hierarchical Linear Models. Newberry Park: CA, 46.

Cameron, A.C. and Trivedi, P. K. (2009) Microeconometrics using Stata. College Station, TX: Stata Press, 184.

Congdon, Peter. Applied Bayesian Hierarchical Methods. Chapman Hall/CRC Press. Chapter 8.

Doornik, J. and Hendry, D. (2007). Econometric Modeling with PcGive. London, UK: Timberlake Consultants, Ltd: Vols I-III.

Diggle, Liang, and Zeger, 1994. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press.

Gelman, A. and Hill, J. (2007). Data Analysis using Regression and Multilevel / Hierarchical Models. New York: Cambridge University Press, chapters 12 and 13.

Gutierrez, R. G. (2010) Multilevel and Mixed models in Stata. American Statistical Association lecture in Vancouver, Canada. Aug 4, 2010.

Laird, N. M. and Ware, J. H. (1982) Random-Effects Models for Longitudinal Data. Nan M. Biometrics, Vol. 38, No. 4. , 963-974.

Pinheiro, J. C. and Bates, D.M. (2000) Mixed-effects models in S and S-Plus. New York: Springer, Chapters 1 through 4.

Rabe-Hasketh, S. and Skrondal, A. Generalized Latent Variable Modeling. Chapman Hall 2004.

Rabe-Hasketh, S. and Skrondal, A. Multilevel and Longitudinal Modeling using Stata. College Station, TX: Stata Press, Chapter five, 83-86.

Singer, J. and Willett, J. (2003). Applied Longitudinal Data Analysis: Modeling change and event occurrence. Oxford, UK: Oxford University Press, 43.

Stata Release 11 (2009) Longitudinal data/ Panel data reference manual. College Station, TX: Stata Press, 306-356.

Verbecke, G. and Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. New York: Springer, 85-86.

UNSCEAR report http://iopscience.iop.org/0952-4746/21/1/609, 3 September 2001.

White SC, Pharoah MJ. Oral radiology: Principles and interpretation. St. Louis, Mosby, 2000, p 49.)
Sharon L. Brooks, DDS, MS. (Health Physics Society, 2010). Health Physics Society web site. 2 Sept 2010. http://www.hps.org/publicinformation/ate/q1193.html.