NSF report

Robert A. Yaffee, Ph.D.

June 16, 2010

Mixed Effects Panel Regression analysis on 3 wave panel

Our dataset has period, age, and cohort effects. We examine the age by period effects for the whole sample subdivided by gender within our collected sample. For reference, see Figure 1 showing the age by period by gender characteristics of our interim sample. Because age and gender effects

AgePeriodGender.emf

Figure Age by Period by Gender

often are significant in medical analysis, we retain these variables to control for such effects in our endeavor to model the health beliefs of the 281 respondents included in our survey so far. Because we are still interviewing respondents to acquire our target sample size of 700 respondents, we refer to the sample input so far as our interim sample. If age and gender are related to what is expected of our health, it is often this standard against which we compare our health status as we perceive it. Our evaluation of the health of ourselves and our family is based on what we expect of people of our age and gender, for which reason, we will have to include this in each of our models, regardless of its significance. We begin our analysis of the health beliefs of those living in the Kiev and Zhitomer Oblasts. Chornobyl is located in the Kiev Oblast and Zhitomer is the nearest Oblast downwind of it. In doing so, we hope to test hypotheses about the changes in the percent of conviction in the belief that the health of the respondent has been affected by the Chornobyl radiation. We test the hypothesis that most people believe that the radiation from Chornobyl has significantly affected their health and that this level is increasing over time. At a later time, we will also examine the growth of the respondent’s belief in the percent of effect that such radiation has had on the health of his or her family.

At first glance, the pattern of this belief change may appear to be unclear (Figure 2). No clear pattern emerges from a quick visual examination of the changing individual paths of this conviction. Even when decompose these paths of change by gender, shown in Figure three, it may be difficult to discern distinctive gender-based effects. Nonetheless, we observe considerable change from one level to another. Many changes seem to cancel out others. By more elaborate panel regression, we hope to begin to understand the nature of any variables associated with the changing patterns in these beliefs. By further analysis of these data, we will endeavor to detect, identify, clarify, and explain the nature of this relationship.

To do so, we examine the evolution of this belief for the whole sample over each wave.radhlbyWaveGend.emf

Figure 2 Belief that own health has been affected

by Chornobyl radiation

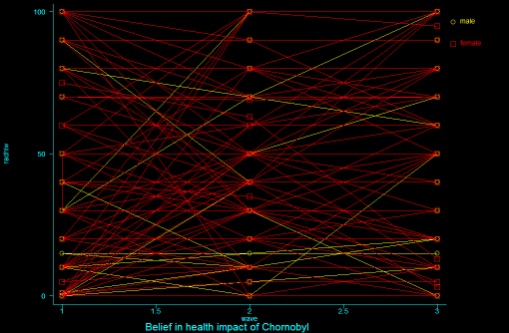


Figure Changes in belief of amount of health

affected by Chornobyl radiation.

We employ a panel data analysis on our dataset. From the vast array of panel data analysis, we are constrained in our choice of a statistical algorithm by a number of factors. Because our data follow a panel of respondents from one wave to another, our model residuals are autocorrelated and therefore in violation of the assumption of ordinary least squares (OLS) that observations be independently and identically distributed (i.i.d). Our choice of algorithm must adjust for this effect, lest our significance estimates be biased upward and the hypothesis testing entailed in our model-building is corrupted.

To test our hypotheses, we wish to perform a robust panel regression analysis using for a dependent variable, the percent of health believed by the respondent to have been affected by Chornobyl radiation. We examine the nature of this variable as its distribution changes from one wave to the next (see Figure 4) to facilitate our understanding of what confronts us. Autocorrelation



Figure 4 Distributional change of percent of radiation impact on own health by wave

among panels ranges from 0.54 at two lags to 0.91 at one lag. Because autocorrelation biases significance tests, this can wreak havoc on the model-building process unless proper adjustments are made.

The male distribution in the last two waves is almost bimodal. We risk getting stuck in a local maximum rather than attaining global maximization. Without a simulated annealing option, we may wish to try random starting for any maximum likelihood estimation we undertake. Alternatively, we might perform the simulated annealing estimation after testing an algorithm using R or Ox at a later time. Owing to sparseness of our data at this point, we refrain from separate models for each gender and at this point merely employ a gender dummy to obtain a notion as to whether there may be a significant difference between gender based models.

# After testing several models, the model that we among the best fitting the data, given our constraints as that of a random intercepts mixed effects panel regression model, the results of which we bootstrapped 400 times while controlling for the cluster of observations within a person over the three waves. The cluster controlled bootstrap with bias correction and acceleration provided empirical standard errors which are used to test the significance of the explanatory variables in the model contained shown in Table one below.

The model consists of significant explanatory variables made up of sociodemographic characteristics, physical symptoms and ailments, as well as the portions of his life that were primarily affected. Among the explanatory variables making up the model are those of gender, current Oblast of residence, level of general physical health, count of the total number of illness as well as the number of terminations of pregnancy to be sociodemographic variables that significantly explained the percent of the health of the respondent that the Chornobyl radiation had affected. It contains a global positive symptom score of the Basic Symptom Inventory along with two variables that measure the areas of work and homecare that are affected.

Even though the age of the respondent at the end of each wave is not significantly related, it remains in the model as a covariate to control for any age. Homecare is a variable of borderline significance, but it remains in the model for theoretical reasons. Gender is very significant in this model, suggesting that men report a belief that a larger percentage of their own health has been affected than do women. People who live in Kiev hold this belief more than those who live now in Zhitomer. The better the general physical health of the respondent, the smaller the proportion of his health he or she believes has been affected by the Chornobyl radiation. The higher the count of illnesses sustained by the respondent, the larger the percent of health he or she believes has been affected b y the Chornobyl radiation. The higher the percentage of health reported affected, the higher the abortion count. Similarly, the higher the percentage of health reportedly affected, the higher the basic positive symptom inventory score. According to this model, the area of life most affected was that of work. Perhaps in some cases, there appears to be a negative relationship between home cooking and repairs and the percentage of health affected by the radiation.

A mixed model was used because the explanatory variables, though considered fixed in most cases, included repeated measurements over time as a as a random effect. Therefore, the intercept in this model is a random effect the estimated value of which changes over the three waves. It is indicated by the asymptotic standard error referred to as the identity. We use maximum likelihood to estimate this model with a constant in it. If we had mean centered our variables, we would have not used a constant and then selected restricted maximum likelihood for estimation instead.

The model residuals, shown in Figure five, are fairly well-behaved but reveal the existence of only one outlier, with some leverage of concern. The standardized residual for this outlier is 3.5 in value. As we have only one negative outlier (t=-3.35, with problematic leverage (Cook’s D=.133, where Kenneth Bollen and Robert Jackman (1990) suggest that 3\*p/n would be the threshold of concern.

The residuals appear to be leptokurtotic and largely symmetric. They are quite nonnormal, so it would not be wise to use statistics requiring a normal distribution for estimation of significance tests. It is better to bootstrap properly or to resort to Bayesian methods, though they take longer. With residuals this symmetric, it does not appear that maximum likelihood would have a problem estimating the parameters, as compared to a situation where the distribution were more bimodal. Therefore, the bootstrapping is performed in such a way that is robust with respect to cluster-correlation that follows from the random repeated measures across the waves. Not only are the residuals empirical, but they are more precise than would be the case if there were a lot more outliers and nonnormality of the residuals. Hence, the option of bias correction is used as well as acceleration to adjust for any skewness in the bootstrapped distribution.

# 

. Table Cluster-controlled bootstrap of Mixed Effects panel regression model

# 

Figure 5 Model residuals

# Project work since last report:

Item analysis

Scale construction

Scale reliability analysis

Pilot study analysis

Poster preparation

Large percent of population believes health has been affected by the radiation

Interim data analysis

Item analysis

Scale construction

Scale reliability analysis

Interim tentative findings

70% of population believes their health and their families’ health has been affected by the Chornobyl radiation

Candidate variable screening and testing

Preliminary model building

Panel analysis data review

Testing of alternative models for panel analysis

Tested data with software developed at Cambridge by Melvin Weeks to do panel data analysis.

Selected a mixed effects panel regression model to test main effects with cluster-controlled bootstrap developed

**Work in progress:**

Will test for linear and nonlinear interaction effects of the regression model

Running a model for family health as well.

Including specific illness counts in the more prevalent: Cardiovascular and GI tracts in models.

Considering Dynamic models that entail feedback and regulation.

Considering other block and wild bootstrap applications for these data.

**Event history models. D**iscrete time logistic and accelerated failure time models will be worked on soon.

# Space-time Bayesian analysis

Writing the Bayesian programs to estimate our threshold spatial autologistic models in R and WinBUGS. A beta-binomial program has already been started in R. More work on conditional spatial autoregressive is anticipated before we begin working on our map evolutions of relative risk. A more complete list can be found in the last grant report. Applications of the Kalman filter and its diffuse versions are being considered for our space-time models.

# I am working on a program to test alternative parameters of each model to perform extreme bounds analysis and a sensitivity analysis to assess robustness of the Bayesian Models.

Work to do:

Space-time modeling. Beginning the programming of the geographical information provided by the respondents.

Related work not part of this grant: Risk analysis in finance: stochastic volatility modeling, realized volatility analysis. Data mining.

References:

Bollen, K. and Jackman, R. (1990). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases, quote Vellman and Welch(1981) in maintaining that the threshold of concern for a Cooks’ D statistic = 3\*p/n where p = number of parameters and n = sample size, in Fox, J. (ed.) Modern Methods of Data Analysis, Sage Publications, Inc.: Newberry Park, CA, 262.

Cameron, A.C. and Trivedi, P.K. (2009) Microeconomics Using Stata, Stata Press: College Station, Tx. Chapters 8, 9.

Doornik, J.A. and Hendry, D.F. with Arellano, M. and Bond, S., Boswijk, HP and Ooms, M. (2009). Econometric Modeling with PcGive, Vol. III, Timberlake Consultants, Ltd, 33-66.

Greene, W.H. (2008) Econometric Analysis. 6th ed. Prentice Hall: Upper Saddle River, NJ, Chapter 23.

Lawson, A. (2009) Bayesian Disease Mapping: Hierarchical Modeling in spatial epidemiology. Chapman Hall, Boca Raton, Fl.

Petris, G., Patrone, S., and Camponelli, P. (2009) Dynamic Models in R., Springer: New York.

Rabe-Hesketh, S. and Skrondal, A. (2005) Multilevel and Longitudinal Modeling Using Stata, 2nd ed., Stata Press: College Station, TX.

Stata Release 11 (2009) Longitudinal and Panel Analysis Users Guide, Stata Press: College Station, Tx.

Wooldridge, Jeffrey M. (2009) Econometric Analysis of Cross-section and Panel Data, MIT Press: Cambridge, MA. 247-291.