A quest for latent variables among explanatory variables for Chornobyl related health threat to family

Robert Alan Yaffee
Silver School of Social Work
New York University

robert.yaffee@nyu.edu

9 May 2012

Outline

1. Primary Objective
2. Male models with radfmw as dep var
3. Female Models with Radfmw as dep var
4. Recapitulation
5. Conclusion

Associated files

- Associated IMAC directory: /Users/robertyaffee/Documents/data/research/chwk/ /phase3/sem/LatentVarSearch /Users/robertyaffee/Documents/data/research/chwk/phase3/data/ox/ /Users/robertyaffee/Documents/data/research/chwk/phase3/data/ox/varsel/OxOut /Users/robertyaffee/Documents/data/research/chwk/phase3/data/ox/varsel/bfmodels/ output/FinalModels
  Associated Toshi1 directory: /Users/ry/Desktop/chwk/phase3/varsel
- data files:chwide19apr2012malesold.dta,chwide19apr2012femmesold.dta. chwide28apr2012chicas.dta, chwide28apr2012chicos.dta. These Stata datasets are saved in the old Stata format (Stata 9) so they can be imported directly into OxMetrics 6.3 for AutoMetrics analysis.
- do files and Ox fl files: factorRadfmw.do, radfmwbf.do
- output files:Radfmfw_factorAnalysis_over_3_waves.smcl, FemaleRadfmw2bfNoRaions.out, Radfmfw_factorAnalysis_over_3_waves.pdf

- MARS files: located on Toshi1 in /Users/ry/Desktop/chwk/phase3/varsel
  output files: MalesRadfmw3Apr26_2012.rtf MalesRadmfw3OutputPass3.dat
- SAS Bayesian test runs: radfmw3chicasmcmc.sas
- This file: LatentVarSearchRadfmwV2bf.pdf

Acknowledgments:
We would like to thank Stas Kolenikov for writing Polychoricpca in Stata, which we have used for our factor extraction and rotation. We need to thank Sir David Hendry, Jurgen Dornik, Jenny Castle, David Corbett and Teresa Timberlake, Noelia Germino, and Aiste Luksyte for their assistance on matters relating to this project. Also, I have to thank the National Science Foundation for their support of this project with HSD grant 082 6983.

1 Introduction

1.1 Primary Objectives

The primary objective of this analysis is to search for and assess possible latent variables among the variables selected by AutoMetrics with a view toward using those variables in our analysis. With this version of the paper, we try to find the optimal models to explain two principal aspects of post-nuclear event health risk: the threat to the family and the threat to oneself, both as reported by the respondent.

1.2 Strategy

1.2.1 Premier statistical packages

No one statistical package does everything, and each has its own strengths and weaknesses. Some packages offer features that others do not. If we can make the assumption that the data are not altered significantly by transferring it from one package to another, we can compare the results of one package to those of another. To the extent that the packages permit identical control of missing values and their replacement, we may compare the results of the same statistical procedure with those results produced by others. This permits corroboration of the analysis as well as comparison of algorithms. It also enables us to apply the outstanding features of one package to those of another.

For variable selection, we use AutoMetrics within OxMetrics, Stata, and MARS. If we cannot directly import a dataset from one package to another, we use Stattransfer to convert the dataset.

1.2.2 Statistical methods

AutoMetrics

The first strategy is to employ AutoMetrics as a front end to build an optimal regression model. It uses a general-to-specific (GETS) modeling strategy to minimize violation of the regression assumptions. It begins with a general unrestricted model (GUM) in which all theoretically relevant variables are included. The GUM therefore is a specification of the local data generating process, which is the driving process. As part of this process, the GETS algorithm will perform variable selection out of all plausible explanatory variables, to the fullest extent possible congruent with the statistical assumptions that validate such models while attempting to form an optimally statistically congruent, parsimonious, and encompassing model. By statistical congruence, I refer to fulfillment of the assumptions of the model being developed and by encompassing, I refer to both the scope of theoretical explanation defined by the variables chosen and the amount of variance explained by them.

We have more than 2000 variables in our dataset. We have approximately 700 respondents– consisting of 340 male and 363 females. We are performing, as is conventional, in the psycho-socio-medical sciences, gender-specific analysis. We consider these datasets to be sufficiently large to detect a small to medium effect size in a regression analysis.

We have many more variables than we do observations and to date, AutoMetrics is the only statistical package that can handle a regression analysis with more variables than observations, due to its remarkable dataset blocking technique. By segmenting the data into blocks smaller than the number of observations, AutoMetrics can test blocks for optimal predictors and then combine them into blocks that remain smaller than the sample size. By reducing the number of nonsignificant explanatory variables, AutoMetrics selects those variables which are congruent with statistical theory and optimally explain the dependent variable(s). By sorting the variables by the square of their t-values, only one test is necessary for all predictors, in the variable selection thereby elegantly circumnavigating the experimentwise error rate problem.

In a cross-sectional model, there is no need for a pre-search lag reduction of the GUM. It is also contended that if the general unrestricted model is congruent, any of its mutually encompassing sub-models which cannot dominate the GUM another will also be congruent.

AutoMetrics searches for omitted variables to avoid omitted variable bias in the model, after forming blocks of dataset segments for further analysis. Because specification error can bias parameter estimates of an analysis easily, we use AutoMetrics to employ the general-to-specific algorithm to minimize this potential problem in our variable search for an optimal model. The program repeatedly searches within the dataset for variables that the user might have inadvertently omitted to minimize the specification error.

AutoMetrics proceeds to generate contending models by a tree search for significant explanatory variables. All possible combinations of variables are tested. If the inclusion of a variable violates an essential regression specification require-

ment or if the inclusion of that variable reduces model encompassing, that path is terminated. Successive congruent models must encompass earlier models and this may be represented by decreases in the deviance or the information criteria. Successive simplification proceeds according to the theory of model reduction as long as later models are congruent and are more encompassing than former ones. The gauge, which is the average retention rate of irrelevant variables, is close to α for AutoMetrics, whereas the potency, which is the average retention rate for relevant variables, is close to theory power for AutoMetrics [**?** , 36]. If contending models in the final competition are tied in fit, those models are sorted, sifted, and selected according which model has the smaller tie-breaking Schwartz information criterion.

MARS data mining

We take the all explanatory variables and input them to a program, Multivariate Adaptive Regression Splines (MARS), developed several years ago by Jerry Friedman at the Stanford Linear Accelerator Center. It generates radial basis functions, which sometimes emphasize parts of our existing variables that are linear rather than the parts, particularly when they appear as pieces of local trends linked at knots. When such piecewise planks contribute to the overall goodness of fit, we add them to the pool of variables to be assessed by AutoMetrics. The AutoMetrics program then sifts them out if they are detrimental to the statistical congruency, parsimony, or encompassing scope of the model.

Those basis functions selected by AutoMetrics to be deemed make important contributions to the theory are included in the final models. We expected to find that in some instances, these transformed variables are recentered or restructured to emphasize the part of the variable that contributes to the explanation. This may particularly be the case if they represent a delayed or partial contributor to the endogenous variable under analysis, such that the explanatory variable is transformed into a the shape of a hockey stick. The effect explained by the variable may not appear until after a threshold has been reached. The basis function transformation permits this aspect of the variable to be implemented, whereas if the variable were to be untransformed it might be overlooked as generally not statistically significant.

MARS may also point out which interactions have to be investigated, something that AutoMetrics does not do, except insofar as it tests whether White's specification test generates statistically significant results.

In these respects, MARS may complement our AutoMetrics regression model-building and we will offer some examples later. It will be interesting to observe how much of our dataset MARS can transform so that AutoMetrics will accept it in this manner.

Dimension reduction

In the interest of dimension reduction, which is an important aspect of parsimony, we perform a principal components and factor analysis on the variables selected by AutoMetrics as important explanatory predictor variables of our focal variables of interest–that of self–perceived Chornobyl health threat to oneself

and that of self-perceived Chornobyl health threat to the family.

Because these variables are to some extent synonymous, we tend not to use them in the same model unless we can control for simultaneity, which we do at a later stage when we perform our etiological path analysis. In preparation for that analysis, we perform a cumulative modeling process for wave 1, waves 1 and 2, and waves 1, 2, and 3.

By examining the stability of the extracted rotated factors from one wave to the next, we obtain a sense of factor structure and stability.

Although we are constrained by the assumptions of valid statistical regression models as well as those of of the types of principal components and factor analysis employed, we can determine from the relative and absolute eigenstructure of these model results, whether factors persist long enough for us to consider a dynamic path model with dynamic factors. For each of our three models, we will take note of the coverage, definition, and persistence criteria in order to assess factor structure and stability. By making our factor analytic variable pool dependent upon AutoMetrics variable selection within the model-building process, we retain our principal criterion of focus and dependence upon the core constructs of Chornobyl related health threat to the self and to the family.

1.2.3 Three types of dynamic factor structure: coverage, definition, and persistence

1. the percentage of common variance explained by extracted factors
2. the number of factors with eigenvalues greater than unity.
3. the number of factors with high loaders (greater than .40)
4. the number of factors with more than 3 high loaders.
5. the proportion of factors extracted with more than 3 high loaders
6. the proportion of factors that persist over 2 waves
7. the proportion of factors that persist over 3 waves
8. the proportion of factors that last only for 1 wave

1.2.4 Limitations

If there are missing values, and the sample size therefore is not constant throughout, the program will not generate the correlations. Therefore, items with reduced sample sizes or persistent non-significance had to be dropped for the analysis to be completed. This changes the item selection subject to factor analysis, which could influence the factor structure found. For this reason, the results of this analysis are to be used only as general guidelines as to possible latent variables to be considered for path analysis to be estimated. They are not to be used as a basis for inference.

Because there are autocorrelations in the models between one wave and the other, specialized programs are required to handle these problems properly in wide formatted files, like those being used here. AutoMetrics addresses these problems with its Newey-West robust estimators used in wide format, whereas Stata requires the format to be changed to a time series or panel data long form in order for proper management. The Newey-West standard errors are analogous to White standard errors in cases where there is no autocorrelation between waves, as would be the case in the Wave one analysis, However, at that point the programs cease to be comparable owing to the different data format.

This unaddressed autocorrelation problem lead to inflation of the R^2, F-tests, and t-values as well as a bias in the estimation in the wide formatted files. AutoMetrics uses ordinary least squares and conditional least squares to deal with this problem. The Stata files employ ordinary least squares with robust standard errors which asymptotically control for heteroskedasticity but not for autocorrelation in the wide format. For this reason, the AutoMetrics models are considered the guideline on which the analysis relies.

For the reason that the model changes from beginning general unrestricted version to a more trimmed one, the goodness of fit may change. Also, any autocorrelation remaining might inflate the R^2 and associated statistics, for which reasons AutoMetrics generally drops that goodness of fit measure from its output and relies on information criteria instead. The use of the R^2 statistic as a measure of power of the model is not recommended unless autocorrelation is correctly dealt with in the file being used.

The net result is to add value to the final models by examining all alternative interpretations and transformations of the variables available for this analysis. In Table 1, we examine the variables selected by AutoMetrics at each cumulative wave for their congruency, parsimony, and encompassing character in explaining the perceived threat to one's family.

2 Factor Analysis of explanatory variables taken from AutoMetrics selected variables from Regression analysis of Chornobyl related health threat to the family

At this stage of the investigation, we place the MARS basis functions in the data pool from which AutoMetrics selects the optimal explanatory variables in accordance with those three criteria–statistical congruence, parsimony, and encompassing. The model selected is the model which maximizes these standardizes and minimizes their violations. By placing all of our variables into datasets into the pool from which the program selects variables, we do not prejudice the selection. Rather we enhance its efficiency greatly as a result of the automation of the process.

We do not yet introduce the raion dummy variables because the sample

gathered sparse data from some of the outer-lying rural raions given the time and funding constraints imposed upon us. For this reason we will have to combine some of the adjacent areas to provide for sufficiently large sample sizes in which to perform our kriging and spatial autoregression analysis. We intend to do this soon but have not do it yet. But this is not necessary until we use the panel and multilevel analysis where the raions will become an upper level of analysis.

The only statistical program to date that can handle more variables than observation is AutoMetrics owing to its capability of partitioning the dataset into blocks or segments, and operating on these blocks sequentially and cumulatively. Because our combined dataset consists of more than 2000 variables with approximately 700 respondents, of the large number of variables in the general unrestricted model, we use a tiny significance level of 0.001 to avoid over-inclusion of variables with borderline significance.

We are performing separate analysis with respect to our dependent variables, Chornobyl related health threat to the family (radfmw1 radfmw2 and radfmw3 respectively for waves 1, 1 and 2, along with waves 1, 2 and 3) and the perceived Chornobyl related health threat to the self (radhlw1 radhlw2 and radhlw3 respectively for cumulative waves 1, 2, and 3). These two notions of threat greatly overlap by definitions, although they are not the same.

This method is not atheoretical. Actually, it presumes that the researchers have gathered the variables needed to accurately describe the phenomenon they are investigating. It merely minimizes any inadvertent violation of the fundamental assumptions of multiple linear regression modeling in a linear or simultaneous context.

If we use both variables of the same wave, we will be building an intrinsic reciprocal relationship into our model. Unless we model this with a nonrecursive simultaneous equation, we would do well at first to avoid this problem by preventing the use of both variables in the same wave. The avoidance will preclude an artificial inflation of the R^2, which will enable a more comprehensive variable search for variables that will contribute to the explanation of the chosen dependent variable.

Therefore, at this stage of the analysis, we artificially exclude this variable at the time of variable selection, with the intention of test simultaneity of the relation with a nonrecursive robust path model at a later time.

Therefore, we perform two separate sets of regression models, with only one of these two variables as the dependent and the other is not included except in later models where there may be a relationship, although lags of it may be included as instrumental variables. However, this exclusion is not necessary if we have more than one wave in the analysis, as we do for waves one and two, as well as in the analysis for waves one, two, and three. In those cases, the lagged value of the other variable may be used as an instrument in a partial adjustment model, as long as we use the proper standard error adjustment.

We use Newey-West standard errors only to render this model comparable to those AutoMetrics models that follow it. However, without any other waves in this model those standard errors are consistent with White standard errors, that

are asymptotically robust to violations of heteroskedasticity. Stata requires a time series or panel data model format to handle Newey-West standard errors, but permits White robust standard errors in a wide data format. However, because AutoMetrics permits Newey-West standard errors in the wide data format, we rely on AutoMetrics for the first stage in this analysis.

In Table 2, we present the male regression model for radfmw1 which has selected the variables listed therein as the best explanatory variables for that dependent variable. Although the model assumptions of hetero- skedasticity appear not to be satisfied, those problems are accommodated asymptotically by the use of the Heteroskedastic and Autocorrelation corrected standard errors (HACSE) in the model. Even if there may be a problem with normality of the residuals, that problem might not be fatal if the residual distribution is sufficiently symmetrical and the third and fourth moments are not seriously compromised.

In this model, we note that two of the basis functions, bf1 and bf11, were selected as explanatory variables. In Figure 1, the relationship between the dependent variable and the number of cancer cases in the Kiev and Zhitomyr Oblasts is located in the top panel, whereas the transformed basis function, BF1, can be seen in the lower panel.

Bf1 is a recentering of the variable measuring the belief most of the cancer cases in Kyiv and Zhitomyr Oblasts stem from Chornobyl radiation, shown in Figure 1. This transformation may reflect the hockey stick structure of this variable at lower levels of cancer there seems to be little relationship, and the relationship depicted in a lowess graph shows that line of optimal local fit is more or less level but after the number of cases becomes sufficiently large, perhaps around 40, a positive relationship emerges between the perceived threat to the family and the number of cancer cases in these two Oblasts. After this level a slight slope begins to appear. By the time the number of cancer cases observed reaches 60, the slope become steeper. This gives the relationship the shape of a hockey or lacrosse stick. By recentering the variable, we move the point at which that positive relationship emerges to the left by about 40 points, so that more of the slope appears throughout the graph, the relationship is more uniformly steeper in the window of view and the transformed relationship appears to be now more uniformly a linear relationship, which is consistent with the requirements of parameter constancy by AutoMetrics.

2.1 Male regression models with self-perceived Chornobyl health risk (radfmw1) as dependent variable

2.1.1 Radfmw1 model For wave one

Table 1 Male wave 1 radfmw1 Variable index 89 variable name type format label variable label 79 shhousw1 double Percentage of strains and hassles related to housing in 1986 kmwork double approximately how far away was your w/s from the chornobyl plant (in kilometers) radtlw1 double believed % of cumulative radiation exposed to in a lifetime in 1986 BSIhos double Basic symptom invenstory hostility subscale Havmil double Distance from Chornobyl in miles bf1 float %9.0g bf1 = max(0, kzchorn - 40) bf11 float %9.0g bf11= max(0, 20 - sufamw1) airw1 double consider hazardous (in percent) - air and water pollution in 1986 dafter double how many days lapsed after Chornobyl accident before you heard about the acciden BSIpar double Basic symptom invenstory Paranoia subscale fallasr double I have trouble falling asleep: I do not fall alseep easily at night (reversal of ffallas) icdx1nr9 double icdx1nr==454 chronic t & a dis icdx1nr2 double icdx1nr==goiter nec icdx4nr8 double icdx4nr==angina pectoris icdx5nr10 double icdx5nr==ac bronchitis/brnchial

89

From these variables, AutoMetrics finds that the optimal regression equation explaining the variance in the Chornobyl related health threat to the family reported by the male respondent as

$$\begin{aligned} Eq_{M1} : Radfmw1_t = {} & 0.142 shhousw1_t - 0.0612 kmwork \\ & + 0.286 radtlw1_t + 2.612 BSIhos \\ & + 0.087 Havmil + 0.301 BF1 - .666 BF11 \\ & + 0.255 airw1_t + 0.037 dafter \\ & - 1.297 BSIpar + 3.997 fallasr \\ & + 31.19 icdx1nr9(chronic\ t\ and\ a\ dis\ veins\ lower\ extremities) \\ & - 28.297 icdx2nr2(thyrotoxicosis) - 37.617 icdx4nr8(angina\ pectoris) \\ & - 26.711 icdx5nr10(acute\ bronchitis) + e_t \end{aligned} \quad (1)$$

The parameter estimate details of which may be found in Table 2. The time-varying parameters have a subscript of t, whereas the time-invariant parameters do not. The basis functions are enumerated and begin with a BF or bf. Their definitions are contained within the accompanying legends. Because of the heteroskedasticity and the autocorrelation between waves, we employ the Newey-West robust standard errors for these equations, thereby asymptotically robustifying our equations from deviations from autocorrelation bias and aberrations in the residual homogeneity of variance.

79 Table 2 Male Radfmw1 model for wave 1 including basis functions 79 EQ(21) Modeling radfmw1 by OLS-CS male final model using bf The dataset is: /Users/robertyaffee/Documents/data/research/chwk/phase3 /data/ox/chwide28apr2012males.dta The estimation sample is: 2 - 340 Dropped 27 observation(s) with missing values from the sample

Coefficient Std.Error HACSE t-HACSE t-prob Part.R^2 $shhousw1$0.1424120.040480.040763.490.00050.0395kr
0.06129250.015980.01612−3.800.00020.0464$radtlw1$0.2859300.036260.043426.590.00000.1274$BSIhos$2.612120.5
0.6662390.16630.1793−3.720.00020.0444$airw1$0.2549990.039850.044405.740.00000.0999$dafter$0.03772940.0314(
1.297110.45750.4506−2.880.00430.0271$fallasr$3.996881.0061.1123.590.00040.0417$icdx1nr$931.185810.779.1123.
28.296723.734.571−6.190.00000.1143$icdx4nr$8−37.617112.067.779−4.840.00000.0730$icdx5nr$10−
26.710713.895.938 − 4.500.00000.0638

sigma 23.3991 RSS 162612.383 log-likelihood -1418.66 no. of observations 312 no. of parameters 15 mean(radfmw1) 51.4423 se(radfmw1) 35.8812 When the log-likelihood constant is NOT included: AIC 6.35228 SC 6.53223 HQ 6.42420 FPE 573.839 When the log-likelihood constant is included: AIC 9.19015 SC 9.37010 HQ 9.26207 FPE 9800.87

Normality test: Chi$^2(2) = 7.8760[0.0195]*Heterotest : F(25, 285) = 1.6317[0.0318]*$
$Hetero - Xtest : F(80, 230) = 3.2156[0.0000] * *RESET23test : F(2, 295) =$
2.3091[0.1011]

We should note that a statistical regression model is valid when its assumptions are fulfilled. To the extent that they are violated and measures are not taken to deal successfully with those violations, the ability of the model to explain, predict, or even evaluate existing policy are undermined. For these reasons, AutoMetrics assesses the principal assumptions of the models at the base of each output. One of these fundamental assumptions of regression models are that the residuals are normally distributed, so the normal probability curve can serve as a vehicle by which to assess the model. In this case, the residuals deviated somewhat from perfect theoretical normality, as measured by the Doornik and Hansen test, and therefore there is an asterisk after it's p-value. This test contains a small sample correction factor that others do not possess [**?** , 287].

Another assumption of ordinary least square regression models is that the residuals are homogeneous, with respect to the fitted line. PcGive contains several tests for homogeneity–entailing the use of Halbert White's general specification test which regresses the squared residuals on all of the squares of explanatory variables to ascertain where there was an significant relationship that was omitted in the modeling. Another of White's tests is also used, denoted the "Hetero-X test", which also includes the cross-products of all of the regressors in this model [**?** , 288]. A significance star next to the p-value of this test suggests that there was a violation. However, we have employed Newey-West standard errors, which are based on White's standard errors, and are designed to asymptotically alleviate this problem.

Also, James Ramsey of New York University's economics department developed a RESET test in 1969 to assess the specification error of the model by regressing the squares and cubes of the predicted value of the regression on the

parameter vector to determine whether reasonable specifications of the model have been omitted [**?** , 289].

Multivariate adaptive regression splines (MARS) was developed by Jerry Friedman as a form of machine learning algorithm to help people perform regression analysis. We use it here to try to find transformations and interactions that we may have overlooked. It adds particular value in linearizing effects that appear to be piecewise or spline-like. These effects appear to be some form of hockey stick or lacrosse stick. Hastie, Tibshirani, and Friedman refer to these forms as piecewise polynomials and splines [**?** , 117-138]. Harrell describes them as splines. If they were in a time series framework David Reilly would call them a local trend whereas Andrew Harvey might refer to them as a broken trend. Rather than test each of our several thousand variables in our dataset, we decided to take advantage of what MARS has to offer in machine learning to locate the variables which are candidates for MARS essentially can form most versions of this transformation to render an effect more amenable to a regression analysis. If the effect of the basis transformation is hard to identify in Figure 1, it is easier to observe in Figure 2. MARS identifies those effects that are amenable to basis transformations (BF) and interactions and suggests them to the user. Bf11 entails a re-centering of the variable, sufamw1, family support in 1986. To the extent that this situation appears to threaten the family, the amount of family support could explain the threat to the family. By reversing the sign of the change and by shifting the location of the mean by 20 points, a basis function transformation is implemented, as is shown in Figure 2. By transforming variables that may have originally appeared to exhibit a delayed response, the basis functions appear to be able to linearize the relationship between the dependent and the independent variable, thereby rendering the variable more amenable to analysis in a program that supports parameter constancy.

Now that we reviewed the variables in the male model, we can examine what happens when we examine them with a view those underlying correlations among them, which we call latent variables or factors. Factors in this case are not discrete variables with two or more levels but rather patterns of correlations among the variables that explain the Chornobyl related health threat to the family.

To do so, we submit these explanatory variables to a factor analysis. Because some of the variables are dichotomous, we have to begin with a principal components analysis, which generates the components from polychoric correlations. These matrix output is then read into a factor analysis program which proceeds to orthogonally rotate these factors in hyperspace to optimally approximate a simple structure by which we can define the factors.

We then examine the eigenstructure of the factors as a baseline from which to assess factor definition, integrity, and stability. We store these assessments in a matrix after each analysis and at the end we compare and contrast these salient aspects of the eigenstructure to learn what these factors are, how long they persist, and whether or not they exhibit any evolution.

With these objectives in mind, we assess the factor analytic output of the

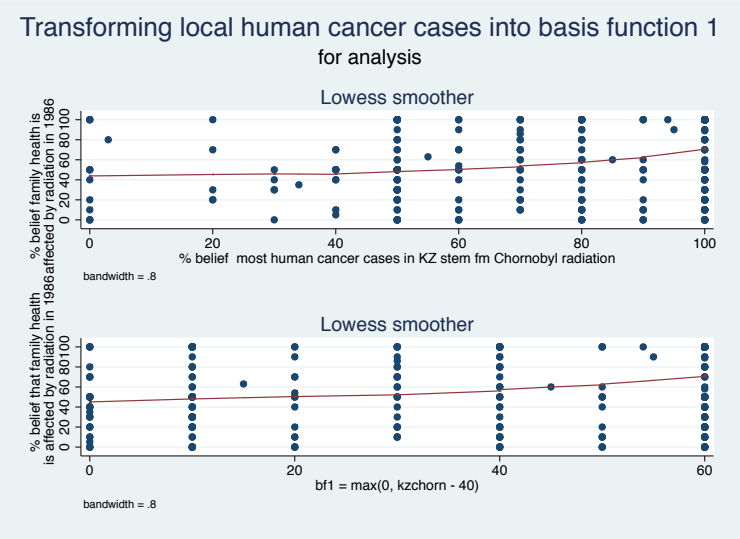

Figure 1: transforming local cancer cases variable into a basis function

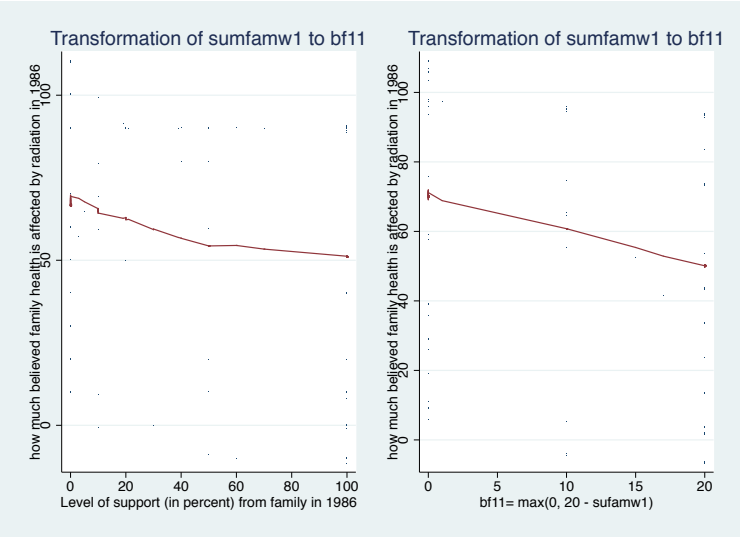

Figure 2: transforming family support into a basis function

radfmw1 male analysis at wave one.

79 Table 3 Principal component analysis for male respondents in 1986 (wave one) 79 k Eigenvalues Proportion explained Cum. explained 5152418 1 2.660461 0.241860 0.241860 2 1.984645 0.180422 0.422282 3 1.216254 0.110569 0.532851 4 1.089010 0.099001 0.631852 5 0.990000 0.090000 0.721852 6 0.835173 0.075925 0.797777 7 0.712497 0.064772 0.862549 8 0.612480 0.055680 0.918229 9 0.550436 0.050040 0.968269 10 0.226380 0.020580 0.988849 11 0.122664 0.011151 1.000000 . matrix define mpcorr1= r(R) . factormat mpcorr1, n(340) mineigen(1) blanks(.36)

From this analysis, we note that there are only four factors with eigenvalues above one, explaining more variance than an individual variable. Because this is a principal components analysis, these four factors are essentially equal to an extraction of all of the variance. When we input this matrix into a factor analysis, shown in Table 4, we note that three of the factors account for about 99% of the variance.

We do not want to be bothered by unique variances, so we focus on the common variance and find int table 4 that the first two factors account for approximately 87.7% of the common variance. Therefore, we extract only two of these factors.

79 Table 4 Unrotated factor structure for male respondents in wave one 79 Factor analysis/correlation Number of obs = 340 Method: principal factors Retained factors = 2 Rotation: (unrotated) Number of params = 21 1360 Factor Eigenvalue Difference Proportion Cumulative 1360 Factor1 2.07694 0.37830 0.4826 0.4826 Factor2 1.69864 1.22264 0.3947 0.8772 Factor3 0.47600 0.08940 0.1106 0.9878 Factor4 0.38660 0.13039 0.0898 1.0776 Factor5 0.25621 0.16818 0.0595 1.1371 Factor6 0.08803 0.08028 0.0205 1.1576 Factor7 0.00775 0.04737 0.0018 1.1594 Factor8 -0.03962 0.06164 -0.0092 1.1502 Factor9 -0.10125 0.08430 -0.0235 1.1267 Factor10 -0.18555 0.17412 -0.0431 1.0836 Factor11 -0.35967 . -0.0836 1.0000 1360 LR test: independent vs. saturated: chi2(55) = 1092.15 Prob¿chi2 = 0.0000

Many analysts are content to try to divine the nature of the factors from the high loadings of the variables without bothering to rotate the factors. However, we have found that an orthogonal rotation in hyperspace optimizes simple structure and enhances the ability of the analyst to define the factors. To further clarify which variables have high loadings on the factors, we cut off the display at 0.36 so that loadings of magnitudes less than that threshold are not shown.

79 Table 5 Factor loadings (pattern matrix) and unique variances 132014 Variable Factor1 Factor2 Uniqueness 132014 shhousw1 0.3668 0.8653 kmwork 0.8747 0.1950 radtlw1 0.9450 BSIhos 0.7601 0.3930 Havmil 0.8829 0.1945 bf1 0.6222 0.6094 bf11 0.8859 airw1 0.3698 0.8607 dafter 0.9979 BSIpar 0.5358 0.6622 fallasr 0.6045 0.6156 132014 (blanks represent abs(loading)¡.36)

With this visual aid of suppressing the display of those variable which do not load highly on these factors, we define the two factors extracted by an assessment of what the high loadings on the factor, after rotation have in common.

79 Table 6 Analysis of the rotated factors for male respondents 79 . rotate, blanks(.36) Factor analysis/correlation Number of obs = 340 Method: principal factors Retained factors = 2 Rotation: orthogonal varimax (Kaiser off) Number of params = 21 1360 Factor Variance Difference Proportion Cumulative 1360 Factor1 2.06121 0.34683 0.4789 0.4789 Factor2 1.71437 . 0.3983 0.8772 1360 LR test: independent vs. saturated: chi2(55) = 1092.15 Prob¿chi2 = 0.0000 Rotated factor loadings (pattern matrix) and unique variances 132014 Variable Factor1 Factor2 Uniqueness 132014 shhousw1 0.8653 kmwork 0.8970 0.1950 radtlw1 0.9450 BSIhos 0.7790 0.3930 Havmil 0.8972 0.1945 bf1 0.6212 0.6094 bf11 0.8859 airw1 0.8607 dafter 0.9979 BSIpar 0.5705 0.6622 fallasr 0.6199 0.6156 132014 (blanks represent abs(loading)¡.36)

The most powerful factor, the first of two orthogonal factors, has four variables that are loading highly onto it. From the lower panel of Table 6, we examine the variable loadings on the factors and note that four variables have high loadings on the first factor. We consider a high loading to be one that has a loading of over .36 on the factor. and try to decide what the Basic symptom inventory (BSI) hostility subscale, the BSI paranoia subscale, trouble falling asleep (fallasr) and a linearized belief that most of the cancer cases in Zhitomyr and Kiev Oblasts stem from Chornobyl radiation exposure. This may be a representation of the trauma sustained by the males after Chornobyl. The full impact may be enhanced by the delay on the part of the authorities fully assessing the gravity of the situation and to admitting to the outside word what the Swedes had already figured out when their measurement equipment began indicating radioactivity in the environment. Moreover, the impact may have been further enhanced by the delay in admission of the nature of the emergency and what was needed in order to protect the public from the hazards that might arise. This first factor, which was an emotional reaction to the nuclear event, accounts for about 47.9% of the common variance (Table 6 upper panel).

The second factor, that accounts for an additional 39.8% of the variance is a fixation on the distance from the event. The two variables loading highly on this factor are the number of kilometers of the workplace to the accident site (kmwork) and the geodesic distance in miles computed from the Haversine formula for spherical distances between two geographical locations between the residential location in 1986 and the location of Chornobyl (Havmil). Together these two factors account for 87.7% of the common variance within the intercorrelation of the selected variables for the male respondents. With only two high

loaders on this factor and with both of these variables being different measures of the same underlying construct, we do not have difficulty defining this factor. It is clearly a distance from the accident site. We will be looking for evidence of these factors on the part of the female respondents at the time of the accident, as well as for evidence of the persistence of these factors in later waves.

At this point in time, we construct a matrix to help us compare the salient aspects of the eigenstructure of this canonical decomposition of the intercorrelations. This list followed the subsection on dimension reduction. In the rows, we have the measures that we discussed there and in the columns of the matrix we have the six different models that we will compare– consisting of the cumulative models for males and females from wave one (the year of the accident–1986) through wave three, the recent years since 1997 to the time of the interview. Therefore, at this juncture the matrix is shown in Table 7.

79 Table 7 Factor structure Comparison 79 MaleW1 FemW1 MaleW12 FemW12 MaleW123 FemW123 1365 PctCmnVarExp .46 0 0 0 0 0 NFex 2 0 0 0 0 0 NFhiloadrs 1 0 0 0 0 0 NF3PlusHi 1 0 0 0 0 0 PropFw3HL .5 0 0 0 0 0 PropFpers3 . 0 0 0 0 0 PropFpers2 . 0 0 0 0 0 PropFpers1 .5 0 0 0 0 0 79 Legend: 1 = percentage of common var explained by extracted factors 2 = number of factors with eigenvalues ¿ 1 3 = number of factors with high loaders ¿ .40 4 = number of factors with more than 3 high loaders 5 = proportion of factors extracted with more than 3 high loaders 6 = proportion of factors that persist over 2 waves 7 = proportion of factors that persist over 3 waves 8 = proportion of factors that persist for only one wave MaleW1 = males in wave one FemW12 = female respondents in waves one and two MalesW123 = male respondents in waves one, two, and three, etc. 79

2.2 Female respondents in 1986 (wave one)

When we examine the variables selected by AutoMetrics to explain the Chornobyl related health threat to the family as perceived and reported by the female respondents, we obtain

79 Table 8 Variables explaining Chornobyl related health threat to the family as reported by the female respondents 79 variable name type format label variable label 79 icdx1nr7 byte %8.0g icdx1nr==454 varicose veins edu7 byte %9.0g Ph.D./Doctor of science aborw1 byte %8.0g number of pregnancy terminations in time period 1976-1986 polprw1 byte %8.0g consider hazardous (in percent) - political problems in 1986 radltw1 byte %8.0g % belief cumulative radiation over lifetime exposure dangerous in 1986 HP2vacatn byte %9.0g hp2fmt Hlth profile Pt2: Hlth probs interfering with vacations icdx4nr9 byte %8.0g icdx4nr==434.91 crbrl art ocl nos w infarc icdx5nr9 byte %8.0g icdx5nr==varicose veins in legs bf10 byte %9.0g bf10 = max(0, sufamw1 - 20) 79

Given these variable definitions, AutoMetrics formulates the optimal explanation of the Chornobyl related health threat to the family in the form of a linear regression estimated by ordinary least squares and conditional least squares as

$$\begin{aligned} Eq_{F1:}Radfmw1_t = &-38.223icdx1nr7(varicose\ veins) + 89.182edu7 + 5.409aborw1 \\ &+0.259polprw1 \\ &+0.331radtlw1 + 22.752HP2vacatn \\ &-57.113icdx4nr9(stroke) - 36.984icdx5nr9(varicose\ veins\ in\ legs) \\ &+1.348BF10(recentered\ family\ support) + e_t \end{aligned} \quad (2)$$

with a more detailed explanation of the parameter estimates found in Table 9 on the next page. If the reader wonders why the signs of the medically diagnosed diseases (icd9 codes) have negative signs, all he or she has to do is examine their tabulations. Instead of being commonplace, these diseases are often very rare

events and are often due to things other than those of exposure to radioactivity. For example, if we examine the tabulations for icdx1nr7 or icdx5nr9, both of which are varicose veins, we observe that only seven of the female respondents report varicose veins and only two of those report them in their legs.

79 Table 9 Tabulations varicose veins reported by female respondents 79

icdx1nr==45 4 varicose veins Freq. Percent Cum. 1235 0 343 98.00 98.00 1 7 2.00 100.00 1235 Total 350 100.00

icdx5nr==va ricose veins in legs Freq. Percent Cum. 1235 0 361 99.45 99.45 1 2 0.55 100.00 1235 Total 363 100.00

79 Table 10 Prevalence of Strokes reported by female respondents 79 icdx4nr==43 4.91 crbrl art ocl nos w infarc Freq. Percent Cum. 1235 0 361 99.45 99.45 1 2 0.55 100.00 1235 Total 363 100.00 The prevalence of stroke,(icdx4nr9),

is also rare, with only two women suffering from it. However, the fact that the signs are negative suggest that the correlation with the Chornobyl related health threat with respect to these problems happens to be negative, with ratio of standard deviations of that of the y to that of the x in the independent variable, contributing to the size of the parameter estimate.

The parameter estimate details for the above model are found in Table 11 on the next page. We will then examine a factor analysis of those explanatory variables selected for the female model. As before, because some of these variables are dichotomous in coding, and therefore not continuous, as is required by conventional factor analysis, we use a polychoric correlation matrix, which assumes an underlying normal distribution which is normalized to a quantity of one. Where the cut is made in that normal distribution allows the proper percentages of probability to reside on each side of the cut, representing the two percentages of area under the curve.

Moreover, we switch from a canonical decomposition of the full variance to one of the common variance when we shift from a principal components analysis to a factor analysis. However, this shift merely separates out the uniquenesses from the total variance, and then proceeds conventionally with the analysis of choice.

As we did before, we noticed residual heterogeneity in the model. By applying the Newey-West standard errors, we asymptotically attenuate the effects of that heterogeneity in addition to control for autocorrelation when more than one wave is included in the model.

Furthermore, AutoMetrics attenuates both of these biases with a wide- formatted file that permits an relatively easy movement of the variables into the principal components and subsequent factor analysis. This transition is not always as seamless as we would like. The varying sample size of the constituent variables in the correlation matrix as well as a row or column of nonsignificant correlations often generates sum of squares and cross -products matrix that is not of full-rank, a non-positive semi- definite correlation matrix, which is conventionally noninvertible, and therefore not amenable to regression based factor score generation. To generate an invertible correlation matrix, some of the variables responsible for generating nonsignificant rows or columns of correlations in the matrix have to be pruned from the model.In short, we often have to rely on the correlation of a trimmed model in order to complete our factor analysis. The full model shown in Table 11 has to be shorn of the icd9 codes and those earning a Ph.D. (edu7) before an input matrix of full rank is attained.

79 Table 11 Full Female model for radfmw1 with basis function bf10 79 EQ(4) Modeling radfmw1 by OLS-CS The dataset is: /Users/robertyaffee/Documents/data/research/chwk/phase3/data/ox/chwide28apr2012femmesold.dta The estimation sample is: 97 - 363 Dropped 5 observation(s) with missing values from the sample

Coefficient Std.Error HACSE t-HACSE t-prob Part.R^2$icdx1nr7-38.223411.9710.18-$
$3.750.00020.0530edu789.181626.532.44436.50.00000.8409aborw15.409191.6481.2224.430.00000.0722polprw10.25$
$57.112526.925.997-9.520.00000.2647icdx5nr9-36.983726.483.210-11.50.00000.3450bf101.347640.45120.13271$

sigma 26.3257 RSS 174646.747 log-likelihood -1223.55 no. of observations 262 no. of parameters 10 mean(radfmw1) 59.4542 se(radfmw1) 34.8081 When the log-likelihood constant is NOT included: AIC 6.57851 SC 6.71471 HQ 6.63325 FPE 719.495 When the log-likelihood constant is included: AIC 9.41639 SC 9.55259 HQ 9.47113 FPE 12288.6

Normality test: Chi^2(2) = 0.41023[0.8146]$Heterotest : F(12, 246) = 2.2322[0.0110]*$
$Hetero - Xtest : F(22, 236) = 1.9921[0.0065] * *RESET23test : F(2, 250) =$
0.79036[0.4548]

To attain an invertible matrix, we are compelled to drop some variables-such as, icdx1nr7, woman, icdx4nr9 icdx5nr9 and edu7 –leaving us with a polychoric correlation matrix of correlations among those variables that we can decompose into an eigenvalue- eigenvector configuration, shown panels of Table 12. The rotated factor structure is shown in the following Table on the next page.

79 Table 12 Polychoric correlation matrix 79 . polychoricpca aborw1 polprw1 HP2vacatn bf10 Polychoric correlation matrix aborw1 polprw1 HP2vacatn bf10 aborw1 1 polprw1 .35603606 1 HP2vacatn -.13932273 -.02433289 1 bf10 -.34468496 -.52405491 .05788011 1 Principal component analysis k Eigenvalues Proportion explained Cum. explained 4152418 1 1.840301 0.460075 0.460075 2 1.008726 0.252182 0.712257 3 0.676746 0.169186 0.881443 4 0.474227 0.118557 1.000000 . matrix define pcorr=r(R) . factormat pcorr, n(340) factors(2) (obs=340) Factor analysis/correlation Number of obs = 340 Method: principal factors Retained factors = 2 Rotation: (unrotated) Number of params = 6 1360 Factor Eigenvalue Difference Proportion Cumulative 1360 Factor1 1.11022 1.04352 1.3692 1.3692 Factor2 0.06670 0.21066 0.0823 1.4515 Factor3 -0.14396 0.07816 -0.1775 1.2739 Factor4 -0.22212 . -0.2739 1.0000 1360 LR test: independent vs. saturated: chi2(6) = 174.97 Prob¿chi2 = 0.0000 Factor loadings (pattern matrix) and unique variances 132014 Variable Factor1 Factor2 Uniqueness 132014 aborw1 0.5039 -0.1049 0.7350 polprw1 0.6518 0.0714 0.5701 HP2vacatn -0.1134 0.2197 0.9389 bf10 -0.6470 -0.0482 0.5790 132014

Table 13 reveals that only one factor emerges from this analysis with clear definition. The variables loading highly onto it are those pertaining to a belief that political problems have lead to a dangerous situation, a belief in family support (bf10) associated with reduced the danger posed to the family, with an amelioration of the health threat to the family, and having an abortion in 1986. A second factor is not well defined and can be disregarded, as its loadings are smaller in magnitude than 0.40.

79 Table 13 Rotated factor analysis for females in wave one (1986) 79 . rotate, blanks(.36) Factor analysis/correlation Number of obs = 340 Method: principal factors Retained factors = 2 Rotation: orthogonal varimax (Kaiser off) Number of params = 6 1360 Factor Variance Difference Proportion Cumulative 1360 Factor1 1.08400 0.99107 1.3369 1.3369 Factor2 0.09293 . 0.1146 1.4515 1360 LR test: independent vs. saturated: chi2(6) = 174.97 Prob¿chi2 = 0.0000 Rotated factor loadings (pattern matrix) and unique variances 132014 Variable Factor1 Factor2 Uniqueness 132014 aborw1 0.4809 0.7350 polprw1 0.6548 0.5701 HP2vacatn 0.9389 bf10 -0.6465 0.5790 132014 (blanks represent abs(loading)¡.36)

Before beginning our comparison of the factor structure of the various models, we should examine the cumulative models for waves 1 and 2, and then for waves 1, 2, and 3.

3 Models for waves 1 and 2

We continue to use that criterion of variable selection. Meanwhile, we retain all of the basis functions relevant to waves 1 and 2 and drop the raions, owing to the sparse data in some and the consequent multicollinearity that arises due to too many with a small sample, we obtain the following AutoMetrics model, until we decide which raions we will aggregate in order to have the proper sample size for an analysis of them.

3.1 Male Radfmw2 model for waves 1 and 2 with basis functions input

We begin this section by examining the cumulative male model for waves 1 and 2 using radfmw2 as a dependent variable. Although basis functions were used in the input to this model, they were not selected for the ultimate model.For the male perception of Chornobyl related family threat over waves 1 and 2, the general focus appears to have been on overexposure to ionizing radiation. Auto-Metrics selected for optimal explanatory variables for the male model including both waves one and two the following variables:

79 Table 14 Male Waves 1 and 2 Variable index for the male reports of Chornobyl related health threat to the family 79 variable name type format label variable label 79 carcin double a person exposed to carcinogen is likely to get cancer (% of agreement) edu4 double some college edu8 double M.D. hospw1 double number of days per year as a patient in a clinic for medical condition in 1976- defnw2 double consider hazardous (in percent) - deficiencies in essential nutrition in 1996 radw1 double believed % of the radioactively contaminated area in 1986 radchw2 double believed % of polution related to chornobyl in 1996 radfmw1 double how much believed family health is affected by radiation in 1986 cloud double radioactive fallout is only harmful when visible (% of agreement) CSprbslv double Coping Problem Solving Subscale BSIhos double Basic symptom invenstory hostility subscale icdx1nr1 double icdx1nr==218.9 uterine leimyoma nec icdx4nr2 double icdx4nr==thyrotoxicosis icdx4nr6 double icdx4nr==hypertension icdx4nr8 double icdx4nr==angina pectoris icdx5nr5 double icdx5nr==hypertension 79

For the male perception of Chornobyl related family threat over waves 1 and 2, the focus seems to have been on the dangers posed by exposure to ionizing radiation. The male model could be expressed as partial adjustment model with a lagged endogenous variable.

$$\begin{aligned} Eq_{M2}: Radfmw2_t = {} & 0.532 radfmw1 + 0.128 carcin - 8.187 edu4 + 30.07 edu8 + 0.089 hospw1 \\ & + 0.219 defnw2 - .115 radw1 + .106 radchw2 + .115 cloud - .379 CSprbslv \\ & + 2.212 BSIhos - 55.282(uterineleimyoma) + 21.112(thyrotoxicosis) \\ & + 48.469(hypertension) - 74.257(otherhypertension) + e_t \end{aligned} \quad (3)$$

79 Table 15 Male radfmw2 model for waves one and two with basis functions