Goodness-of-fit testing by means of sub-population divergences, by Leroy O. Stone

Introduction

This text is technical appendix D2012 to Leroy O. Stone (Ed.), *Key Demographics in Retirement Risk Management* (Springer, 2012), available at http://www.springer.com/social+sciences/population+studies/book/978-94-007-4043-3 . (Below *"Key Demographics …"* refers to this book.) At http://www.retirementresearch.org , the reader will find detailed information about the uses of this book among the following groups: (a) researchers and graduate students studying links between population variables and retirement issues, (b) educators teaching about patterns of preparedness to address retirements' challenges and aspects of related capabilities, and (c) researchers in work groups that support the outreach activities of organizations engaged in population-segment targeting for marketing, service delivery or educational purposes (http://www.retirementresearch.org (click on "Who is this book for?").

This appendix supports Chapter 6 of *"Key Demographics …"* by providing details about the goodness-of-fit testing procedures and their results. Featured among the procedures is an approach that focuses on the degree of divergence between a cross-tabulation (observed versus expected) associated with the null hypothesis and another associated with the fitted model -- a sort of demographic approach to goodness-of-fit testing.

Goodness-of-fit testing strategy

Since prediction modeling is often done in a context where some practical use of the results is envisaged, it is essential to briefly consider the results of the goodness-of-fit tests that have been conducted in the current modeling effort. We begin by reviewing the goodness-of-fit computations that are produced routinely in the output of SAS PROC LOGISTIC. Then will follow some specialized goodness-of-fit computations that we have devised for the case where the focus is upon the identification of key demographics.

The important distinction between these two contexts is that in the former the question is how well the model works in predicting the outcomes for particular cases (respondents), whereas in the latter the focus is on how well the model performs in predicting the *distribution* of outcomes for selected population subgroups. The latter approach is consistent with our focus upon the study of populations, a demographic analysis.

Results

Tables D2012_1 and D2012_2 present the evidence concerning goodness of fit. The first page of Table D2012_1 provides the more traditional coefficients generated by SAS PROC LOGISTIC. These are shown under the heading "Association of predicted probabilities and observed responses". While the measures are modest relative to their theoretical maxima, they are within the range one tends to see in journal articles where the researcher attempts to analyze inter-individual variations by way of a micro-data file.

For example, while the maximum possible values of the measures called "Somers D" and "Gamma" are 1.0, in published studies of inter-individual variation that level is rarely approached closely. The values close to 0.35 shown in Table D2012_1 are within the range of published values for these two measures, in multivariate analysis of inter-individual variation.

Table D2012_1. *Goodness of fit of an ordinal logistic regression model tha generates probabilities of levels of the retirement-related risk management scale for combinations of values of predictor variables, the pre-retired population aged 45 to 69, by household income group, Canada, 2007*

	Household income groups		
	Low income[1]	Middle income	High income
Type of model	Cumulative logit	Cumulative logit	Cumulative logit
levels	4	4	4
Weight variable	NormWeight[2]	NormWeight	NormWeight
Number of observations used	3870	2961	2954
Sum of weights used	3870	2961	2954
Response frequencies			
1	112	205	429
2	522	707	971
3	2471	1768	1443
4	765	281	111
Testing the global null hypothesis: beta=0			
-2 Log L[3] for beta = 0	7508	6289	6564
-2 Log L for model	7053	5964	6145
Chi-Square	455	325	419
Degrees of freedom	24	24	23
Probability of chi-square under the null hypothesis	<0.0001	<0.0001	<0.0001
Association of predicted probabilities and observed responses			
Percent concordant	67.2	66.0	67.3
Somers D	0.35	0.33	0.35
Gamma	0.35	0.33	0.36

In passing to the second page of Table D2012_1 we arrive at the measures of goodness of fit that are focused upon *distributional* comparisons, which is typical of demographic analysis and appropriate where the focus of study is upon the identification of key population subgroups.

Table D2012_1 (continued) . *Goodness of fit of an ordinal logistic regression model that generates probabilities of levels of the retirement-related risk management scale for combinations of values of predictor variables, the pre-retired population aged 45 to 69, by household income group, Canada, 2007*

	Household income groups		
	Low income[1]	Middle income	High income
Cross-tabulation of predicted and observed scale levels in Sub-sample A4			
Probability of chi-square	<0.0001	<0.0001	<0.0001
Contingency coefficient	0.15	0.14	0.19
Cross-tabulation of predicted and observed scale levels in Sub-sample B			
Probability of chi-square	<0.0001	<0.0001	<0.0001
Contingency coefficient	0.12	0.12	0.11
Coefficient of predictive efficiency[5]	0.8	0.9	0.6
Benchmark association of sex with predicted outcome[6]			
Probability of chi-square	<0.0001	<0.0001	<0.0001
Contingency coefficient	0.06	0.06	0.04
Coefficient of predictive effectiveness[7]	2.0	2.0	2.8

The key summary measure for the distributional comparisons is the contingency coefficient that is shown on the second page of Table D2012_1. It is based upon the cross tabulation of the observed and predicted positions of respondents on the scale of retirement-related risk management activities.

Table D2012_1 (concluded) . *Goodness of fit of an ordinal logistic regression model that generates probabilities of levels of the retirement-related risk management scale for combinations of values of predictor variables, the pre-retired population aged 45 to 69, by household income group, Canada, 2007*

1. The model runs within each household income group separately.
The 2007 household income ranges for each group are as follows:
Low income group: the 34th percentile or below ($60,000 or lower)
Middle income group: from 34th to 65th percentile ($60,000 to $100,000)
High income group: over 65th percentile (over $100,000).
The boundaries between income groups are limited by the raw data.
2. The original GSS survey weight for each respondent is divided by the average of all weights in the sample.
3. Minus 2 times the log-likelihood statistic.
4. See Appendix C for related details about sub-sample B.
5. The contingency coefficient for sub-sample A (Pearson's C) will usually be greater than that for sub-sample B. Thus the closer the value for sub-sample B is to that for sub-sample A the more efficient is the model in achieving within B its usual maximum (the value shown for A).
6. This measures how much better the model is than the the practical minimum association, based on the sub-sample that made no contribution to the estimation of the model's parameters.
7. While the coefficient has theoretical minimum and maximum of 0 and 1, in fact it rarely approaches these values. For example, when we cross tabulated two very closely linked measures of class of worker, such that 90% or more of the observations in a row were concentrated in one single cell, the computed contingency coefficient was 0.7. At the lower extreme, we compute a value of C for two variables known to have very low association (one being the dependent variable of the model), to give us an idea of a practical minimum for C, and an improved notion of how much better the model has performed compared to the practical minimum association.
Source: Statistics Canada, 2007 General Social Survey.

To illustrate the point of distributional comparison, let us jump ahead to Table D2012_2 for a moment. The first panel of numbers in the table is exactly the cross tabulation that is associated with the contingency coefficient of 0.12 that is in the first column of Table D2012_1. The first line in Table D2012_2 is for the subpopulation *predicted* to be at the bottom level of the scale, and the fourth line is that for the population *predicted* to be at the top of the scale.

Table D2012_2. *Cross-tabulation of predicted and observed levels of the retirement-related risk management index, within sub-sample B of the pre-retired population aged 45 to 69, by household income group, Canada, 2007*

	Observed				
Predicted	Level 1	Level 2	Level 3	Level 4	Sum
	Low income population				
Level 1	26	64	9	1	100[2]
Level 2	22	61	15	3	100
Level 3	14	69	13	4	100
Level 4	23	50	24	4	100
Contingency coefficient[3]					0.12
	Middle income population				
Level 1	11	65	21	3	100
Level 2	11	61	22	6	100
Level 3	9	57	25	9	100
Level 4	3	56	33	8	100
Contingency coefficient					0.12
	High income population				
Level 1	7	64	24	5	100
Level 2	4	53	29	14	100
Level 3	3	46	34	17	100
Level 4	3	47	33	17	100
Contingency coefficient					0.11
	Observed for all pre-retirees aged 45 to 69				
Total	12	57	23	8	100

Table D2012_2 continued *Cross-tabulation of predicted and observed levels of the retirement-related risk management index, within sub-sample B of the pre-retired population aged 45 to 69, by household income group, Canada, 2007*

	Observed				
Benchmark	Level 1	Level 2	Level 3	Level 4	Sum
	Low income population				
Sex[4]					
Men	21	65	12	3	100
Women	19	69	10	1	100
Contingency coefficient					0.06
	Middle income population				
Sex					
Men	9	57	26	8	100
Women	13	55	24	7	100
Contingency coefficient					0.06
	High income population				
Sex					
Men	3	49	33	15	100
Women	3	53	32	13	100
Contingency coefficient					0.04

1. See Table D2012_1 footnote 4.
2. The values may not add to 100 because of rounding.
3. This is the same coefficient as shown in Table D2012_1, and it is a measure of the degree of association between the row and column variables of the table. It is based on the numbers in the cross-tabulation and is computed from the chi-square shown in Table D2012_1.
4. The distributions for men and women are shown here to serve the benchmarking function described in footnote 7 of Table D2012_1.
Source: Statistics Canada, 2007 General Social Survey.

The model has good fit when the distributions of these two subpopulations over levels of the *observed* scale scores are greatly divergent. By comparing this distributional divergence between the first and fourth lines of Table D2012_2 with that between men and women (shown just below in the same table), we can see indications of strong distributional divergence between the populations predicted to be at the top and bottom levels of the scale. And the pattern of the divergence is meaningful as we will show below.

The key results on the second page of Table D2012_1 are those for sub-sample B. The members of sub-sample A were used to estimate the parameters of the model, and the members of sub-sample B were excluded from this process. The two sub-samples are randomly drawn using the random number generation procedure built into SAS. Using the parameters estimated in sub-sample A, we predicted the scale scores of the members of sub-sample B, and the results are shown in the second line of contingency coefficients in Table D2012_1.

The third set of contingency coefficients shown in Table D2012_1 are benchmark calculations designed to indicate what the coefficient is expected to be when the association is low. This will be called "the low-association benchmark".

We raised two questions concerning the result of using the parameters estimated in sub-sample A to predict scale scores in sub-sample B. First, to what extent does the predictive accuracy in sub-sample B approach its expected maximum level? Second, how much better is that accuracy than that of the low-association benchmark?

The coefficient of predictive efficiency shown in Table D2012_1 answers the first question. It compares the contingency coefficient achieved in sub-sample B with that of sub-sample A (keep in mind that the members of sub-sample A were used to estimate the parameters of the prediction model). The predictive accuracy achieved in sub-sample B is close to its expected maximum value in the Low and Middle income groups only.

The coefficient of predictive effectiveness shown in Table D2012_1 answers the second question. It compares the contingency coefficient achieved in sub-sample B with that of the low-association benchmark. Among the income groups, the contingency coefficient achieved in sub-sample B is roughly twice as large as that of the low-association benchmark. While this difference is not very great, it should be noted that this traditional divergence between men and women is statistically significant.

The meaning of these contingency coefficients is best grasped by looking at the underlying cross classifications of observed versus predicted scale values. It is these cross classifications that determine the levels of the contingency coefficients. The cross classifications for sub-sample B and for the low-association benchmark are shown in Table D2012_2.

Consider, for example, the subpopulation that is at the bottom level of the predicted values. These are shown in the topmost row for each income group in Table D2012_2, and in this subpopulation we would expect the concentration of cases in the lower two quadrants of the scale score would be very much greater than is the case for the population as a whole, as well as in comparison with the subpopulation that is predicted to have high levels of the scale score. This latter sub-population is shown in the fourth row for each income group in Table D2012_2.

The comparison of these two rows within each income group in Table D2012_2 suggests that the model has achieved substantial predictive accuracy. The subpopulation predicted to be at the lowest level of the scale has a substantially greater concentration of cases at the bottom two quadrants than is the case for the subpopulation predicted to be at the highest level.

Moving to the top level of the observed values on the scale (shown in the fourth column of table D2012_2), we see, as expected, that the subpopulation that is predicted to have high level on the scale has a much greater concentration at the topmost level of the observed scale values than is the case for the subpopulation predicted to have the lowest level of the scale.

A similar but much less consistent pattern is shown when predicted Levels 2 and 3 are compared within each income group. Those predicted to be at the higher level scale (Level 3) tend to have lower percentages in the first two quadrants of the observed distribution (the first two columns of the table). However, this pattern is limited to the middle and upper income groups.

Thus while there is low accuracy in case-by-case prediction, when the model predicts a low level on the scale the *observed* distribution has a much greater than average concentration at the two lower levels of the scale. When the model predicts a high level of the scale, the observed distribution has a greater than average concentration at the two upper levels of scale.

In short, in the context of the search for key demographics, the examination of goodness of fit of the model appropriately places a focus upon distributional divergence between subpopulations predicted to have high and low levels of the scale score. Looking at this distributional divergence as shown in Table D2012_2, it is reasonable to conclude that the prediction model has a tolerable level of accuracy and we can move on to the consideration of key demographics, based upon the results of having fitted this model.