

# Multivariate Analysis of Pension Plan Mortality Data

**I**N DECEMBER OF 1994, THE U.S. CONGRESS ENACTED THE RETIREMENT PROTECTION ACT OF 1994 as part of the General Agreement on Tariffs and Trade. This legislation included restrictions on actuarial assumptions used in pension funding calculations. In particular, it mandates the use of the 1983 Group Annuity Mortality table at least through the year 1999. The Treasury Department will specify an updated mortality table for use beginning in the year 2001.

To provide the Treasury Department with a current and thorough study of uninsured pensioner mortality for use in updating the mandatory group annuity mortality table, the Retirement Plans Experience Committee (RPEC) of the Society of Actuaries collected private pension plan experience data for the years 1989-1995. The data includes age, gender, and status (active, disabled, retiree, or beneficiary) of each participant, along with information on whether workers were salaried or paid hourly and whether the plan was collectively bargained. Plans were classified as white collar if at least 70 percent of the participants were salaried and nonunion, blue collar if at least 70 percent of the participants were paid hourly or belonged to a union, or mixed if the plan could not be classified as white or blue collar. Some of the data also contains information relating to the amount of annuities being paid to annuitants and income earned by active employees.

The goal of our analysis is to investigate the extent to which these “independent or explanatory variables” explain differences in mortality. This is done by fitting a suitable statistical model to the available data that relates a response variable (dependent variable) to the set of covariates (independent or explanatory variables).

In our preliminary analysis, we actually investigated the use of several alternate models before we selected a final model that best fits the data. This model is tested for adequacy, both in terms of fit and in terms of pre-

dicted death rate under different covariate inclusions.

## Preliminary Data Analysis

The data consist of 113 pension plans with a total of more than 14.5 million life-years of exposure. It was assembled into a database by Kathleen Elder, FSA, and Laxman Hegde, Ph.D., of Frostburg State University. The RPEC has also provided us with base tables that were categorized by age (15 to 113 years), gender (male, or female), and participant status (employees, retirees and beneficiaries, combined healthy, or disabled). We find it useful to compare the results we get from our fitted models against numbers from this base table, in terms of matching overall patterns in age, gender, and participant status.

Subsequent to checking data accuracy, our next step was to create basic tables for this data set, by the categories age, gender, and participant status. In particular, we computed the overall mortality rates for each age (deaths/exposure) as well as the death rates for those lives with an available amount exposed. As defined by the SOA, death rate is:

$$\frac{[\text{deaths/life years exposed}] \times [\text{average death amount/average amount exposed}]}{}$$

These results, along with all other calculations and output from this project, may be accessed online at [www.soa.org/research/mappmd.html](http://www.soa.org/research/mappmd.html). We will refer to tables from this site as “online tables.” The basic mortality tables, for example, are presented in online Table 1a.

Following discussions with RPEC members, we chose the following variables for investigation in our model:

- **ANNUITY SIZE CLASS:** small, medium, large, unknown
- **COLLAR TYPE:** blue collar, white collar, mixed
- **SIC CODE:** auto, machinery, communications, metals, chemicals, petroleum, transportation, government, others.

Note that each of these determining variables is a



**CHARLES VINSONHALER AND NALINI RAVISHANKER** ARE ON THE FACULTY AND **GUY RASOANAIVO** IS A GRADUATE STUDENT IN MATHEMATICS AT THE UNIVERSITY OF CONNECTICUT, STORR, CT.; **JEYARAJ VADIVELOO** IS VICE PRESIDENT AND APPOINTED ACTUARY AT ING, HARTFORD, CT. A MORE COMPLETE VERSION OF THIS PAPER, INCLUDING TABLES, IS AVAILABLE AT [WWW.SOA.ORG/RESEARCH/MAPPMD.HTML](http://WWW.SOA.ORG/RESEARCH/MAPPMD.HTML).

categorical variable. In general, if a variable has  $k$  categories, we generate  $k$  indicator variables and incorporate these as  $k$  “independent variables” in a regression model (but, without an “intercept” term). More specifically, we assume that our dependent variable is a linear combination of the independent variables with no constant term involved. For example, the annuity size class has four categories; we create four indicator variables as follows:

ANNSZ1 = 1 if Annuity Size Class is small and 0 otherwise;  
 ANNSZ2 = 1 if Annuity Size Class is medium and 0 otherwise;  
 ANNSZ3 = 1 if Annuity Size Class is large and 0 otherwise;  
 ANNSZ4 = 1 if Annuity Size Class is unknown and 0 otherwise.

Similarly, we create three indicator variables for Collar Type (COL1, COL2, COL3) and nine indicators for SIC Code (SIC1, SIC2, ..., SIC9). These will enter as 16 independent variables in a multiple regression model, without an intercept term (to avoid linear dependence).

The *dependent variable* or *response variable* in the logistic regression model is a function of the mortality rate, deaths/life-years exposed.

### Linear Logistic Regression Model

Logistic regression models the relationship between a binary dependent variable and one or more explanatory variables. (See, for example, Hosmer and Lemeshow, 1989.) In general, for each subject a binary variable is defined, which assumes the value 1 for an event and value 0 for a non-event. We associate a binary dependent variable  $y_x$  to an insured age  $x$ , by setting  $y_x = 1$  for a death and  $y_x = 0$ , otherwise. We may write

$$\text{Prob}(y_x = 1) = \pi_x; \text{Prob}(y_x = 0) = 1 - \pi_x$$

for the true probabilities of death (event) and non-death (nonevent) respectively. (Note that we denote by  $q_x$  the corresponding observed proportion obtained from our data.) Associated with each individual (or group of individuals), we have a  $K$ -dimensional vector of explanatory variables  $Z_x$ , which are indicator variables.

The principal objective of a statistical analysis is to investigate the relationship between the probability of response and the explanatory variables. To investigate this relationship, we construct a linear logistic regression model. Specifically, we assume that the dependence of  $\pi_x$  on  $Z_x$  occurs through the linear combination

$$\eta = \beta^T Z_x = \beta_1 Z_{x,1} + \dots + \beta_K Z_{x,K}$$

where  $\beta = (\beta_1, \dots, \beta_K)$  is a vector of unknown model coefficients (parameters). Since  $\eta$  is allowed to assume values on the entire real line, it would be inconsistent with probability laws to express  $\pi_x$  as this linear combination (since  $\pi_x$  lies between 0 and 1). A simple way of avoiding this difficulty is to transform  $\pi_x$  from the unit interval to the entire real line  $(-\infty, \infty)$  using the logit or logistic function. (For details, see McCullagh and Nelder, 1983.)

$$\text{logit}(\pi_x) = \log \frac{\pi_x}{(1 - \pi_x)}$$

We refer to the ratio  $\pi_x/(1 - \pi_x)$  as the *odds* or *odds ratio*. Then, the linear logistic regression model has the form

$$\text{logit}(\pi_x) = \log \frac{\pi_x}{(1 - \pi_x)} = \beta^T Z_x \quad (1)$$

Equation (1) models the logit transformation of the true unknown event probability  $\pi_x$  for an individual age  $x$ , as a linear function of the explanatory variables  $Z_x$ . Thus the logit function is a *link function* through which the mean of the dependent variable is linearly related to the explanatory variables. Although there are other commonly used link functions, such as the probit link, log-log link, and the complementary log-log link, the logit link function has the advantage of being more easily interpreted.

For instance, if  $K=2$ , we have the model written for individual age  $x$  as

$$\log \frac{\pi_x}{(1 - \pi_x)} = \beta_1 Z_{x,1} + \beta_2 Z_{x,2} \quad (2)$$

for the log odds of a positive response. From this, the probability of a positive response is obtained as

$$\pi_x = \frac{\exp(\beta_1 Z_{x,1} + \beta_2 Z_{x,2})}{1 + \exp(\beta_1 Z_{x,1} + \beta_2 Z_{x,2})} \quad (3)$$

Assuming that  $Z_{x,1}$  and  $Z_{x,2}$  are functionally unrelated, we may interpret this model as follows: The effect of a unit change in  $Z_{x,2}$  is to increase the odds ratio  $\pi_x/(1 - \pi_x)$  of a positive response *multiplicatively* by the factor  $\exp(\beta_2)$ , with  $Z_{x,1}$  held fixed. Data for the linear logistic regression analysis may be given either as a binary response  $y_x$  on each individual or in the form of count data from a binomial experiment where we are given number of trials and the number of events. The pension plan mortality data are available in the latter form with life-years exposed corresponding to the number of trials and deaths corresponding to the number of events.

We set up the logistic regression model for our situation as follows: For smoothness, we grouped the ages in sets of five, eg., 20-24, 25-29, ..., 95-100. For each of these age groups, we defined sub-categories by gender (female, male) and participant status (retirees and beneficiaries, disabled, active employees, or all annuitants combined).

For each of these eight subcategories, we considered *lyr* (life years exposed) as the number of independent trials (of a binomial experiment) and *dth* as the number of events (deaths). Again, we only consider the *dth* and *lyr* corresponding to the particular subcategory for plans with integer data. The explanatory variables were described earlier, 16 indicators corresponding to annuity size class, collar type, and SIC code. We fit five different models, incorporating subsets of these explanatory variables, and either including interactions or not. Our computer runs generated a large number of tables that can-

not be included here due to space restraints, but are available online.

As we mentioned earlier, we do not include a constant term in our model. For instance, to fit the full model with all the explanatory variables, the model function is:

$$\text{logit}(\pi_x) = \beta_1 \text{ANNSZ1} + \beta_2 \text{ANNSZ2} + \dots + \beta_{16} \text{SIC9} \quad (4)$$

The method of maximum likelihood is used to estimate the model parameters by finding the values of parameters that maximize the likelihood function  $L(\beta; \text{data})$ . (For details, see McCullagh and Nelder, 1983.) We used SAS to fit the model using PROC LOGISTIC with the *events/trials* syntax and with the *Stepwise Model Selection* option to select a statistically optimal set of explanatory variables. The results are summarized in the online tables.

The regression estimates corresponding to the significant explanatory variables are presented in the online tables for the five different models involving different sets of explanatory variables. Based on these coefficients, the estimated odds ratio for a particular explanatory variable may be calculated. We present our results as multipliers of the base table odds ratios. Based on the estimated  $\beta$ , we computed *multipliers* corresponding to the significant explanatory variables, and these are presented in the online Tables 2b, 2d, 3b, 4b, 5b, and 6b.

Following is a set of instructions and an example that should help a practicing actuary use these multipliers in order to obtain a final probability of mortality for a specific insured. The multipliers are obtained as follows for the model with all three variables, Annuity Size Class, Collar Type and SIC Code, as explanatory variables. Let  $\beta_j, j = 1, 2, 3, 4$  denote the regression coefficients corresponding to ANNSZ1, ..., ANNSZ4; let  $\beta_j, j = 5, 6, 7$  denote the coefficients corresponding to COL1, COL2, and COL3; and let  $\beta_j, j = 8, \dots, 16$  denote the regression coefficients corresponding to SIC1, ..., SIC9.

Note that if a particular variable is present in the final selected model, the corresponding  $\beta$  is non-zero—otherwise it is zero. We denote by  $P$  the number of explanatory variables in the final model; for the above situation,  $P = 3$  (we use  $P = 3$  in the online Table 6b). We also use  $P = 3$  for the model with interaction between Annuity Size Class and Collar Type (see the online Table 5b). If only two of these three variables is present, for instance, Annuity Size Class and Collar Type, then  $P = 2$  (we use  $P = 2$  in the online Table 4b; also see Table 1 of this paper). If only one of these variables is present in the final model,  $P = 1$  (we use  $P = 1$  in the online Tables 2b, 2d and 3b).

For each of the eight sub-categories and age groups, compute  $q_x^0 = \text{dth/lyr}$ , which is the overall mortality rate, and set  $r_x^0 = q_x^0 / (1 - q_x^0)$ . For each age group, the multiplier  $m_j$  is defined as

$$m_j = \frac{\exp\{\beta_j\}}{(r_x^0)^{1/P}} \quad (5)$$

The relevant multipliers are then applied to the base table

odds ratios. An example will best illustrate how the procedure works. Suppose a plan has Annuity Size Class=Large, Collar Type=White Collar and the SIC Code is 37- -, corresponding to the auto industry. The variables present in the model are therefore ANNSZ3, COL1, SIC1 with corresponding coefficients  $\beta_3, \beta_5$ , and  $\beta_8$  and multipliers  $m_3, m_5$ , and  $m_8$ . The predicted mortality rate is then obtained as follows:

$$\text{Step 1: Set } r_x = m_1 m_5 m_8 r_x^0$$

$$\text{Step 2: Set } q_x = \frac{r_x}{1+r_x}$$

The value of  $q_x$  is the mortality rate predicted by the model for a Large Annuity Size, White Collar worker in the auto industry, whose age falls in the interval  $[x - 2, x + 2]$ .

Sometimes, one or more of the variables associated with a plan are excluded from the model (based on a statistical variable selection criterion). An excluded variable is indicated by a blank in the multiplier table. For example, in Table 1, the table for Female Retirees and Beneficiaries, there are three blanks in the row for ages 40-44.

If all variables associated with a plan are excluded from the model, then the corresponding multipliers are set equal to 1. This would be the case in Table 1 for a Female Beneficiary aged 40-44, who is White Collar, with a Large Size Annuity.

Otherwise, the blank multiplier  $m$  is obtained from the  $(r_x^0)^{-1/P}$  column, where the value of  $P$  that we use has been defined above. To illustrate these cases, look in Table 1 at a Female Beneficiary aged 40-44, who is White Collar, with a Medium Size Annuity. The overall multiplier is

$$0.084 \times 17.616 = 1.480$$

which is close to 1. Now, look at a Female Beneficiary aged 40-44, who is White Collar, with a Large Size Annuity; in this case, the overall multiplier will be 1. We did observe that in some cases, the overall multiplier that we computed differed significantly from 1.0, casting doubt on the accuracy of the numbers obtained from modeling this subgroup. This is the case when we consider a Female Beneficiary aged 40-44, who is Blue Collar, with a Large Size Annuity. The blank in the Annuity Size Large column must be replaced by the number 17.616 from the second column. The overall multiplier is then

$$17.616 \times 2.4381 = 42.950.$$

To assess the goodness-of-fit of the regression model, we considered four measures that are output from the SAS procedure: (i) a *coefficient of determination*, max-scaled  $R^2$ ; (ii) the *Akaike Information Criterion*; (iii) the *Schwarz Criterion*; (iv) the *area under the receiver operating characteristic (ROC) curve*.

These four measures, showing satisfactory fit, are presented in the online Tables 2c, 2e, 3c, 4c, 5c, and 6c for selected sub-categories (Female Retirees and Beneficiaries and Male Retirees and Beneficiaries) for each of the fitted models. The results from other sub-categories are available and are similar. In addition,

we looked at two other criteria, not presented in the tables, which indicated adequate fit for all models considered. A more complete explanation of the goodness-of-fit criteria can be found in the original report.

### Conclusion

The models that we have constructed provide guidelines on how to adjust an existing “base rate” mortality table if certain criteria hold. These adjustments are in the form of multipliers pertaining to each of the criteria that act on the base mortality “odds ratio.” The corresponding adjusted mortality rate is then easily determined. When only one explanatory variable is used, the modeling is quite satisfactory. For example, when Annuity Size Class is used as the only explanatory variable, reasonably smooth tables of multipliers are obtained for core age groups in certain Gender × Status categories (see the online Table 2b).

These multipliers could be used to judiciously adjust base mortality rates, particularly if the multipliers were first smoothed via graduation. A similar result holds when only Collar Type is used as an explanatory variable (see the online Table 3b). On the other hand, if multiple explanatory variables are introduced, such as both Annuity Size Class and Collar Type, then results from the models tend to exhibit less “smoothness.” Clearly in



these instances, actuarial judgment needs to be exercised over any adjustment of mortality rates.

### References

- Hosmer, D.W. and Lemeshow, S. (1989). Applied Logistic Regression. Wiley: New York.
- McCullagh, P. and Nelder, J.A. (1983). Generalized Linear Models. Chapman and Hall: New York.

**TABLE 1: Multipliers with Annuity Size Class and Collar Type as Covariates (No Interaction)**

Age Group	$(r^2)_{(1/P)}$	Annuity Size Small	Annuity Size Medium	Annuity Size Large	Annuity Size Unknown	Collar Type BC	Collar Type WC	Collar Type MC
25-29	9.695				0.169			
30-34	11.453	0.146			0.077			
35-39	24.972	0.099			0.032			
40-44	17.616	0.291	0.084		0.055	2.438		
45-49	17.410	0.062	0.029	0.079	0.059			
50-54	15.236	35.297			23.963	0.027	0.038	0.081
55-59	13.276	0.090	0.059	0.056	0.066			18.586
60-64	10.217	0.112	0.089	0.083	0.099	11.239	8.394	
65-69	8.146	0.116	0.088	0.086	0.115	10.031		
70-74	6.501	0.147	0.124	0.097	0.142	7.768		
75-79	5.134	0.188	0.159	0.187	0.195	5.509	4.778	
80-84	3.983	0.247	0.242	0.193	0.260		3.643	
85-89	3.033	0.308	0.298	0.244	0.351		2.705	
90-94	2.312	0.340	0.345	0.379	0.440			2.716
95-99	1.870	0.360	0.407		0.519			2.519