

MULTIVARIATE ANALYSIS OF PENSION PLAN MORTALITY DATA

Charles Vinsonhaler, Nalini Ravishanker, Jeyaraj Vadiveloo, and Guy Rasoanaivo*

ABSTRACT

This paper uses the logistic regression model to examine private pension plan data for 1989–95 collected by the Retirement Plans Experience Committee of the Society of Actuaries. When only one explanatory variable, such as annuity class size, is used in modeling mortality rates, the model provides a reasonable fit to the data. Multiple explanatory variables give less satisfactory results.

1. INTRODUCTION

In December 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 mandated the use of the 1983 Group Annuity Mortality table at least through 1999. The Treasury Department will specify an updated mortality table for use beginning in 2001.

To provide the Treasury Department with a current and thorough study of uninsured pensioner mortality for the purpose of updating the mandatory mortality table, the Retirement Plans Experience Committee (RPEC) of the Society of Actuaries (SOA) collected private pension plan experience data for 1989–95 that have been used to prepare new mortality tables: the RP-2000 Mortality Tables. The data include 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 or not the plan was collectively bargained. Plans were classified as white collar if at least 70% of the participants were salaried and

nonunion, blue collar if at least 70% 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 contain information relating to the amount of annuities 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 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 investigated the use of several alternate models before selecting a final model that best fits the data. We tested the model for adequacy, in terms of both fit and predicted death rate under different covariate inclusions.

We present a summary of our analysis of the pension plan mortality data. We have worked closely with SOA team members in arriving at the final model and interpreting results. The organization of the report is as follows. In Section 2 we discuss how we verified data accuracy through several checks. In Section 3 we describe the use of a linear logistic regression model for the pension plan mortality data. We fit this model for data categorized by age groups (in groups of five), gender, and participant status, for plans for which life years exposed and deaths are integers. Thus, we excluded data from six plans that contained noninteger exposures. The response (dependent) variable is the mortality rate deaths/life years ex-

*Charles Vinsonhaler is a Professor in the Department of Mathematics and Nalini Ravishanker is an Associate Professor in the Department of Statistics at the University of Connecticut, 196 Auditorium Rd., Storrs, CT 06269 e-mail: vinson@uconnvm.uconn.edu and nalini@stat.uconn.edu. Jeyaraj Vadiveloo is Vice President and Appointed Actuary with Aetna Financial Services/ING, 151 Farmington Ave, TN41, Hartford, CT 06156 e-mail: vddivelooj@aetna.com. Guy Rasoanaivo is a graduate student in the Department of Mathematics at the University of Connecticut.

posed, and the categorical independent variables are the following:

- Annuity size class: S (small), M (medium), L (large), or U (unknown)
- Collar type: BC (blue collar), WC (white collar), or MC (mixed), and
- SIC code with first two digits: 37 (auto), 35 or 36 (machinery), 48 (communications), 33 (metals), 28 (chemicals), 29 (petroleum), 42 (transportation), 99 (government).

Five different logistic regression models were fitted, with (1) only annuity size class as an explanatory variable, (2) only collar type as an explanatory variable, (3) annuity size class and collar type as explanatory variables (without interaction), (4) annuity size class and collar type as explanatory variables (with interaction), and (5) annuity size class, collar type, and SIC code as explanatory variables (without interaction). Multipliers corresponding to each significant explanatory variable are provided in tables that can be accessed at www.soa.org/research/mappmd.html. We refer to these as “online tables.” A few sample tables are provided in this paper. A set of instructions on how a practicing actuary should use these multipliers is also provided.

For completeness, we include in Section 4 a summary of other statistical modeling approaches that we tried on these data. We defined explanatory variables and a suitable dependent variable and fitted and tested the normal linear regression (see Section 4.1) and the tobit regression model (see Section 4.2). Section 5 presents a summary and conclusion.

2. PRELIMINARY DATA ANALYSIS

The data, which comprise 113 pension plans with a total of more than 14.5 million life years of exposure, were assembled into a database by Kathleen Elder, F.S.A., and Laxman Hegde, Ph.D., of Frostburg State University. The RPEC has also provided us with base mortality tables that were categorized by age (15–113 years), gender (male or female), and participant status (employees, retirees, and beneficiaries, combined healthy or disabled). We have found 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.

We began our analysis by checking the general description of the data, which consist of 115,415 rows and 18 columns. Approximately 20% of these rows have “unavailable” for the amount exposed field. These records comprise about 57% of the total life years exposed. The data set was consistent, as our checks verified. For instance, when categorized by the various determining variables, the records always added up to 115,415.

As noted, we identified six plans for which either life years exposed or deaths had noninteger values. Since the logistic model that we fit to the data requires an assumption of integer values for these variables, we excluded data from these plans from our model fitting (about 5% of the data).

Subsequent to checking data accuracy, our next step was to create basic tables for this data set, by 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; here we computed the death rate as

$$\frac{\text{Deaths}}{\text{Life Years Exposed}} \times \frac{\text{Average Death Amount}}{\text{Average Amount Exposed}}$$

as defined by SOA. These results are presented in the online Table 1a. Table 1 below gives a sample.

Our next step was to set up the independent

Table 1
Base Mortality Table Based on All Data

Age Group	Death/Exposure (q_x)	$q_x/(1 - q_x)$	Death Rate
25–29	0.01053	0.01064	
30–34	0.00757	0.00762	0.00399
35–39	0.00160	0.00160	0.00038
40–44	0.00321	0.00322	0.00228
45–49	0.00329	0.00330	0.00318
50–54	0.00429	0.00431	0.00332
55–59	0.00564	0.00567	0.00469
60–64	0.00949	0.00958	0.00843
65–69	0.01485	0.01507	0.01289
70–74	0.02311	0.02366	0.02066
75–79	0.03655	0.03794	0.03479
80–84	0.05929	0.06303	0.05693
85–89	0.09806	0.10872	0.09776
90–94	0.15761	0.18710	0.15955
95–99	0.22239	0.28600	0.21250

variables in order to study their effect on mortality. Following a discussion with SOA team members in July 1999, we chose the following independent variables for investigation in our model:

- Annuity size class: small, medium, large, and unknown
- Collar type: blue, white, and mixed collar
- SIC code: auto, machinery, communications, metals, chemicals, petroleum, transportation, government, others.

Note that each of these determining variables is a 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, 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). If the columns of the matrix of explanatory variables are linearly dependent, then some of these columns can be expressed as linear combinations of others. In statistical model fitting, a consequence of this is to create instability in the estimation of parameters.

The dependent variable or response variable in the logistic regression model is deaths/life years exposed. In all our model fitting, we check whether (1) the underlying model assumptions are satisfied, (2) the fitted model is adequate, and (3) the predicted probabilities are consistent and reasonable.

3. LINEAR LOGISTIC REGRESSION MODEL FOR PENSION PLAN MORTALITY DATA

Logistic regression models the relationship between a binary dependent variable and one or more explanatory variables. In general, for each subject, a binary variable is defined, which assumes the value 1 for an event and value 0 for a nonevent. In our case, for a binary dependent variable y_x associated with an insured age x , we assume that $y_x = 1$ for a death and $y_x = 0$, otherwise. We write

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

for the true probabilities of death (event) and nondeath (nonevent), respectively. (Note that we denote by q_x the corresponding *observed* proportion obtained from our data.) Associated with each individual (or groups of individuals), we have a K -dimensional vector of explanatory variables \mathbf{Z}_x , which are indicator variables. The principal objective of the statistical analysis is to investigate the relationship between the probability of response and the explanatory variables. We construct a formal statistical model, which is a linear logistic regression model. Suppose we say that the dependence of π_x on \mathbf{Z}_x occurs through the linear combination

$$\eta = \boldsymbol{\beta}^T \mathbf{Z}_x = \beta_1 Z_{x,1} + \cdots + \beta_K Z_{x,K},$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ is a vector of unknown model coefficients (parameters). Since η assumes values on the entire real line, it would be inconsistent with probability laws to express π_x (since π_x lies between 0 and 1) as this linear combination. A simple way of avoiding this difficulty is to transform π_x from the unit interval to the entire real line $(-\infty, \infty)$ (for details, see McCullagh and Nelder 1983). The logit or logistic function is useful to effect such a transformation:

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

We refer to the ratio $\pi_x/1 - \pi_x$ as the odds or odds ratio. The linear logistic regression model has the form

$$\text{logit}(\pi_x) = \log \frac{\pi_x}{(1 - \pi_x)} = \boldsymbol{\beta}^T \mathbf{Z}_x. \quad (1)$$

Equation (1) models the logit transformation of

the true unknown event probability π_x of 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 related linearly 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 and will be used in our analysis.

For instance, if $K = 2$, we write the model 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 can 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 can 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 the number of trials and 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, for example, 20–24, 25–29, . . . , 95–100. For each of these age groups, we defined subcategories 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 consider only the *dth* and *lyr* corresponding to the particular subcategory for plans with integer data. The explanatory variables were described earlier; 16 indicators correspond to annuity size class, collar type, and SIC code. We fitted 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 cannot be included here because of space constraints. As noted previously, they are available online.

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 $L(\boldsymbol{\beta}; \text{data})$ (for details see McCullagh and Nelder 1983). We used the software package 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.

For each age group, the online Table 2a summarizes the response profile of individuals by annuity size class. For a sample, see Table 2 in this paper. 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 for a particular explanatory variable can be calculated. We present our results as multipliers of the base table rates, as discussed below.

Before we use these results with confidence, we must assess the goodness-of-fit of the regression model through objective measures of how well the model fits the data. We use four measures in the tables that are output from the SAS procedure.

1. A coefficient of determination, max-scaled R^2 (Cox and Snell 1989; Maddala 1983; Nagelkerke 1991):

$$R_{\text{ms}}^2 = \frac{R^2}{R_{\text{max}}^2}, \quad (5)$$

where

$$R^2 = 1 - \left[\frac{L(0)}{L(\hat{\boldsymbol{\beta}})} \right]^{2/n}, \quad (6)$$

where n is the sample size, $L(0)$ is the maximized likelihood of a model without explanatory vari-

Table 2
Deaths and Life Years Exposed by Levels of Annuity Size Class

Age Group	Annuity Size	Deaths	Exposure	Deaths/Exposure (q_x)	$q_x/(1 - q_x)$
25-29	Small	0	64	0.00000	0.00000
	Medium	0	7	0.00000	0.00000
	Large	0	2	0.00000	0.00000
	Unknown	2	117	0.01709	0.01739
30-34	Small	3	239	0.01255	0.01271
	Medium	0	15	0.00000	0.00000
	Large	0	2	0.00000	0.00000
	Unknown	8	1,198	0.00668	0.00672
35-39	Small	3	756	0.00397	0.00398
	Medium	0	60	0.00000	0.00000
	Large	0	8	0.00000	0.00000
	Unknown	7	5,422	0.00129	0.00129
40-44	Small	11	1,778	0.00619	0.00623
	Medium	1	230	0.00435	0.00437
	Large	0	14	0.00000	0.00000
	Unknown	27	10,120	0.00267	0.00268
45-49	Small	13	3,657	0.00355	0.00357
	Medium	3	1,807	0.00166	0.00166
	Large	4	888	0.00450	0.00452
	Unknown	40	11,894	0.00336	0.00337
50-54	Small	62	11,107	0.00558	0.00561
	Medium	41	10,940	0.00375	0.00376
	Large	10	5,437	0.00184	0.00184
	Unknown	109	24,270	0.00449	0.00451
55-59	Small	210	29,661	0.00708	0.00713
	Medium	165	32,805	0.00503	0.00506
	Large	45	10,207	0.00441	0.00443
	Unknown	338	61,677	0.00548	0.00551
60-64	Small	625	57,753	0.01082	0.01094
	Medium	456	54,433	0.00838	0.00845
	Large	85	12,094	0.00703	0.00708
	Unknown	1,418	148,035	0.00958	0.00967
65-69	Small	1,382	87,261	0.01584	0.01609
	Medium	454	38,257	0.01187	0.01201
	Large	65	5,968	0.01089	0.01101
	Unknown	3,403	225,754	0.01507	0.01530
70-74	Small	2,296	94,096	0.02440	0.02501
	Medium	491	24,370	0.02015	0.02056
	Large	44	2,904	0.01515	0.01538
	Unknown	5,561	241,717	0.02301	0.02355
75-79	Small	2,662	74,420	0.03577	0.03710
	Medium	327	10,958	0.02984	0.03076
	Large	29	858	0.03380	0.03498
	Unknown	7,587	203,904	0.03721	0.03865
80-84	Small	2,616	45,893	0.05700	0.06045
	Medium	271	4,929	0.05498	0.05818
	Large	12	277	0.04332	0.04528
	Unknown	7,814	129,581	0.06030	0.06417
85-89	Small	1,908	21,299	0.08958	0.09840
	Medium	153	1,815	0.08430	0.09206
	Large	7	101	0.06931	0.07447
	Unknown	6,336	62,487	0.10140	0.11284
90-94	Small	931	6,949	0.13398	0.15470
	Medium	77	578	0.13322	0.15369
	Large	3	20	0.15000	0.17647
	Unknown	3,379	20,306	0.16640	0.19962
95-99	Small	202	1,118	0.18068	0.22052
	Medium	20	107	0.18692	0.22989
	Large	1	7	0.14286	0.16667
	Unknown	941	4,002	0.23513	0.30742

able, and $L(\beta)$ is the maximized likelihood of the specified model with explanatory variables that are selected using the stepwise selection procedure, and

$$R^2_{\max} = 1 - [L(0)]^{2/n}. \tag{7}$$

The coefficient R^2 is greater than 0 and achieves a maximum of less than one for discrete models, namely R^2_{\max} . The max-scaled R^2 is a rescaled measure.

2. The Akaike Information Criterion (AIC) is an adjustment to $-2 \log L(\hat{\beta})$ based on the number of explanatory variables in the model and the number of observations used. For a given set of data, the AIC is a goodness-of-fit measure that can be used to compare one model to another. Lower values of AIC indicate a more desirable model.

3. The Schwarz Criterion (SC) is another way of adjusting $-2 \log L(\hat{\beta})$ based on the number of explanatory variables and the number of observations. It is also a goodness-of-fit measure, with lower values indicating a more desirable model.

4. The area under the receiver operating characteristic (ROC) curve lies between 0 and 1 and is large for a model with high predictive accuracy. Although this is not an extremely sensitive measure, it is nevertheless useful as one indicator of model adequacy.

These four measures are presented in the online Tables 2c, 2e, 3c, 4c, 5c, and 6c for selected subcategories (female retirees and beneficiaries,

and male retirees and beneficiaries) for each of the fitted models. The results from other subcategories are available and are similar. An example is shown in Table 3 below.

In addition, we looked at two other criteria, not used in the tables, which indicated adequate fit for all models considered:

5. The log likelihood ratio test statistic, $-2 \log L(\hat{\beta})$, has a χ^2 distribution under the null hypothesis that all the regression coefficients in the model are zero (i.e., none of the explanatory variables in that model has any explanatory power on the response). The SAS procedure prints this value and a p -value for this statistic. A significant p -value is one that is less than 0.05 (usually) and provides evidence that at least one of the regression parameters for an explanatory variable is nonzero.

6. The Hosmer and Lemeshow (1989) goodness-of-fit test for logistic regression models is a conservative test that involves dividing the data into g groups of roughly equal size based on the percentiles of the estimated probabilities (usually $g = 10$). The observations are sorted in increasing order of their estimated probability of having an event outcome (deaths). The discrepancies between the observed and expected number of observations in these groups are summarized using the Pearson χ^2 statistic, which is then compared to a χ^2 distribution with $g - 2$ degrees of freedom. The statistic has the form

Table 3

Goodness-of-Fit Criteria for the Model with Annuity Size Class as Covariate

Age Group	Number of Observations	Events	Nonevents	R^2	AIC	SC	Area under ROC Curve
25-29	92	2	188	0.70	123.4	126.7	0.31
30-34	225	11	1,443	0.96	155.9	148.0	0.64
35-39	471	10	6,236	0.99	244.6	167.3	0.72
40-44	713	39	12,103	0.98	546.0	538.5	0.65
45-49	1,140	60	18,186	0.99	811.6	833.2	0.54
50-54	1,971	222	51,532	0.98	2,831.9	2,858.5	0.59
55-59	3,376	758	133,592	0.98	9,354.2	9,383.7	0.53
60-64	3,849	2,584	269,731	0.96	29,178.4	29,209.9	0.53
65-69	3,758	5,304	351,936	0.94	55,135.6	55,168.0	0.53
70-74	3,571	8,392	354,695	0.92	79,753.5	79,785.9	0.52
75-79	3,088	10,605	279,535	0.88	90,971.3	91,003.0	0.52
80-84	2,538	10,713	169,967	0.81	81,299.4	81,329.8	0.51
85-89	2,031	8,404	77,298	0.70	54,970.2	54,998.3	0.51
90-94	1,413	4,390	23,463	0.54	24,253.3	24,278.0	0.52
95-99	608	1,164	4,070	0.37	5,534.8	5,554.5	0.54

$$X_{\text{HL}}^2 = \sum_{i=1}^g \frac{(O_i - N_i \bar{\pi}_i)^2}{N_i \bar{\pi}_i (1 - \bar{\pi}_i)}, \quad (8)$$

where N_i is the number of observations in the i th group, O_i is the number of event outcomes in the i th group, and $\bar{\pi}_i$ is the average estimated probability of an event outcome for the i th group. The statistic is given as HL with the p -value below it in parentheses. For a p -value greater than 0.05, we cannot reject the null hypothesis H_0 . The model provides a good fit to the data, so the fitted model is acceptable.

Based on the estimated β , we computed multipliers corresponding to the significant explanatory variables (see online Tables 2b, 2d, 3b, 4b, 5b, and 6b). Here we give a set of instructions and an example that should help actuaries use these multipliers to obtain a final probability of mortality for a specific insured. The multipliers are obtained as follows for the model with annuity size class, collar type, and SIC code as explanatory variables. Let $\beta_j, j = 1, \dots, 4$ denote the regression coefficients corresponding to ANNSZ1, \dots , ANNSZ4; let $\beta_j, j = 5, \dots, 7$ denote the coefficients corresponding to COL1, COL2, and COL3; and let $\beta_j, j = 8, \dots, 16$ denote the regression coefficients corresponding to SIC1, \dots , SIC9. Note that if a particular variable is present in the final selected model, the corresponding β 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 4 in 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 subcategories and age groups, compute $q_x^0 = dtl/lyr$, which is the overall mortality rate (see Table 1), and $r_x^0 = q_x^0 / (1 - q_x^0)$. For each age group, the multiplier m_j is defined as

$$m_j = \frac{\exp\{\beta_j\}}{(r_x^0)^{1/P}}. \quad (9)$$

An example will illustrate how the procedure

works. Suppose a plan has a large annuity size class, white collar type, and SIC code of 37--, corresponding to the auto industry. The variables present in the model are, therefore, ANNSZ3, COL1, and SIC1 with corresponding coefficients β_3, β_5 , and β_8 and multipliers m_3, m_5 , and m_8 . The predicted mortality rate is then obtained as follows:

- Step 1: Set $r_x = m_1 m_5 m_8 r_x^0$.
- Step 2: Set $q_x = r_x / (1 + r_x)$.

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 4, 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 4 for a female beneficiary age 40–44, who is white collar, with a large-size annuity.

Otherwise, the blank multiplier m is obtained from the $1/r_0^{1/P}$ column, where the value of P that we use has been defined above. To illustrate these cases, look in Table 4 at a female beneficiary age 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. Next, look at a female beneficiary age 40–44, who is white collar, with a large-size annuity; in this case the overall multiplier will be 1.0. 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

$$17.616 \times 2.4381 = 42.950.$$

Table 4
Female Retirees and Beneficiaries Multipliers with Annuity Size Class and Collar Type as Covariates (No Interaction)

Age Group	$(r_x^0)^{-(1/p)}$	Annuity Size				Collar Type		
		Small	Medium	Large	Unknown	BC	WC	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

4. OTHER STATISTICAL ANALYSES

Some aspects of the underlying nature of the pension plan mortality data were not available to us until a later phase of the project. One important aspect was whether *lyr* assumed positive integer values and whether *dth* assumed non-negative integer values. This assumption of integer values is crucial for the use of a logistic regression model that has the binomial distribution as its basis. Our interim investigation using logistic regression models revealed the presence of fractional values on some records for both *lyr* and *dth*. This was disturbing because we realized that, if the data were collected in such a manner that *lyr* and *dth* cannot be assumed to be integer valued, the logistic regression model is no longer the appropriate model to use. Although our interim results showed promise with the logistic regression model, we reluctantly looked in other directions for possible alternative approaches. Subsequently, the data with fractional values were identified as belonging to one of six plans and excluded, and the logistic regression model was taken to completion. However, for completeness, we report briefly the other analyses that we attempted. We mention that the empirical results were similar even when the noninteger valued plans were included in the study.

In our SOA proposal we had suggested the use of the Cox proportional hazards models and ex-

pected survival curves or the Kaplan-Meier estimates, which are standard tools in survival or mortality data analysis. However, our first look at the current pension mortality data made it clear that the proposed analyses such as the Cox proportional hazards model were not suitable, because the information on deaths and life years exposed has been aggregated over time. No information on the mortality of an insured is available at different points in time. This required that we shift our mode of investigation to the area of regression analyses. We considered normal linear regression, logistic regression, and tobit regression analyses as possible approaches. The linear logistic regression model was described in detail in Section 3. We discuss the other two analyses briefly here. In each case, we let *dth* denote deaths and *lyr* denote life years exposed.

4.1 Normal Linear Regression Model

It is possible to carry out the “usual” normal linear regression analysis as follows (Seber 1977). For an initial study, we selected a few ages between 25 and 95, and used the initial set of explanatory variables defined in Section 2. For each subcategory defined by gender and participant status:

- We computed $y = dth/lyr$. Here we only use the *dth* and *lyr* corresponding to the particular sub-

category. Note that $y \in [0, 1]$, with a large proportion of 0 values. We look for a suitable transformation of y to normality. Usual transformations such as $\log(y)$ or $\text{logit}(y) = \log(y/(1 - y))$ are not suitable here because of the high proportion of 0 values of y . We selected \sqrt{y} as the dependent variable, since this transformation frequently will stabilize the variance of the response with count data and make it more closely resemble a normal random variable.

- We fit the normal linear regression model

$$\sqrt{y} = \beta_0 + \beta_1\text{ANNSZ1} + \beta_2\text{ANNSZ2} + \dots + \beta_{16}\text{SIC9} + \epsilon,$$

where ϵ denotes the i.i.d. error. We used SAS to fit the model using PROC REG with the stepwise model selection option to select a statistically optimal set of explanatory variables.

- We computed the estimated rates from this model based on the estimated β coefficients. However, these estimated rates do not seem very satisfactory in terms of the trends they exhibit. This might be because of the fact that, in each subcategory, the number of cases with zero deaths is very large in comparison with the number of cases with deaths; perhaps the square-root transformation is not adequate. No other “reasonable” transformation seems adequate either. In summary, this is a poor competitor to the linear logistic model for this problem.

4.2 Tobit Regression Model

The use of a tobit regression model (Judge et al. 1985) may also be appropriate for such data, when lyr assumes noninteger values and when dth/lyr assumes a large number of zero values. When the dependent variable is real-valued and further assumes several zero values, we fitted a tobit regression model by relating the dependent variable to a set of explanatory variables. This analysis was carried out using SAS PROC LIFEREG, by setting up an interval of dependent variable values suitably. The results are not as satisfactory as the logistic regression model. One drawback with this model is that it is more complicated than the logistic regression setup. Furthermore, we are unable to obtain

multiplier factors in a simple fashion as in logistic regression.

5. CONCLUSION

Having tried a variety of statistical multiple regression models to fit the given pension mortality data, we are led to conclude that there is not enough “pattern” in the data to allow for a single *model* that can be used on a day-to-day basis by a working actuary. Our models certainly give some indication of how an existing “base rate” mortality table might be adjusted if certain criteria hold. Furthermore, these adjustments have been obtained as multiplicative factors pertaining to each explanatory variable in the fitted model. To this extent, we have fully carried out a study exploring, in a statistical sense, the multiplicative effects of significant explanatory variables on the probability of death.

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 \times status categories (see the online Table 2b or Table 5 below). These multipliers could be used to adjust base mortality rates judiciously, particularly if they 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 ex-

Table 5
Multipliers with Annuity Size Class as Covariate

Age Group	Annuity Size			
	Small	Medium	Large	Unknown
25–29				1.63478
30–34	1.66757			0.88189
35–39	2.48446			0.80613
40–44	1.93190	1.35517		0.83018
45–49	1.08131	0.50405	1.37149	1.02278
50–54	1.30302	0.87322	0.42772	1.04721
55–59	1.25670	0.89093	0.78045	0.97116
60–64	1.14201	0.88185	0.73884	1.00956
65–69	1.06778	0.79687	0.73064	1.01551
70–74	1.05711	0.86907	0.65024	0.99528
75–79	0.97783	0.81077	0.92208	1.01868
80–84	0.95903	0.92305	0.71844	1.01812
85–89	0.90502	0.84673	0.68494	1.03786
90–94	0.82683	0.82143	0.94317	1.06691
95–99	0.77108	0.80381		1.07490

planatory variables are introduced, such as annuity size class and collar type, then results from the models tend to exhibit less “smoothness” (see Table 4). The lack of smoothness is more dramatic when SIC code is also introduced (online Table 6b). This lack of smoothness would be met with extreme skepticism by a working actuary.

One possible direction to go with these data is a “descriptive” analysis, which could be done by calculating mortality rates for different combinations of explanatory variables and trying to find significant patterns. A problem with such an approach is that when too many variables are specified with several levels in each, the data points become sparse. A simple method for adjusting calculations on pension plans is unlikely to emerge.

In conclusion, the multiplier tables that we have constructed based on a statistical model can be used as rough guidelines for how mortality might change from plan to plan but may not be valuable for any “exact” calculation of mortality rates.

ACKNOWLEDGMENTS

The authors are grateful to Tom Edwards of the Society of Actuaries for his valuable comments that improved the accuracy and exposition of this paper.

REFERENCES

- COX, D. R., AND E. J. SNELL. 1989. *The Analysis of Binary Data*. London: Chapman and Hall.
- HOSMER, D. W., AND S. LEMESHOW. 1989. *Applied Logistic Regression*. New York: John Wiley.
- JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LUTKEPOHL, AND T. C. LEE. 1985. *The Theory and Practice of Econometrics*. New York: John Wiley.
- MADDALA, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. London: Cambridge University Press.
- McCULLAGH, P., AND NELDER, J. A. 1983. *Generalized Linear Models*. New York: Chapman and Hall.
- NAGELKERKE, N. J. D. 1991. “A Note on a General Definition of the Coefficient of Determination.” *Biometrika* 78: 691–92.
- SEBER, G. A. F. 1977. *Linear Regression Analysis*. New York: John Wiley.

DISCUSSION

JEFFREY PAI*

The idea of applying the multipliers, suggested by the authors, to predict the mortality rate of a certain group based on the base mortality tables is easy to use for practitioners. Excess of mortality provides another way to compare the mortality experience of a group of policyholders to some known base mortality tables. The relative and additive mortality models are two common models for excess mortality.

Suppose we have a group of size n under study and we wish to compare the mortality rate of this particular group to the base mortality table. Let $\theta_j(t)$ be the force of mortality from the base mortality tables for the j th individual in the group. These known base mortality tables depend on the characteristics of the j th policyholder such as gender, status, age, annuity size, or/and collar type. (For detail see Andersen and Væth 1989 and Klein and Moeschberger 1997).

The relative mortality model assumes that the mortality rate at time t for the j th policyholder is a multiple of the base mortality rate for this policyholder, that is,

$$h_j(t) = \beta(t)\theta_j(t), \quad j = 1, 2, \dots, n.$$

Here, a value of $\beta(t)$ greater than 1 indicates that policyholders in the group are experiencing the death event at a faster rate than comparable policyholders in the standard population. For the j th policyholder, let $Y_j(t)$ be 1 if the policyholder is at risk at time t , and 0 otherwise. The value $B(t) = \int_0^t \beta(s)ds$ is called the cumulative relative excess mortality, which can be estimated by

$$\hat{B}(t) = \sum_{t_i \leq t} \frac{d_i}{Q(t_i)},$$

where d_i is the number of deaths at time t_i and $Q(t) = \sum_{j=1}^n \theta_j(t)Y_j(t)$. The multiple, $\beta(t)$, is called the relative risk function, which can be estimated by the slope of the estimated cumulative relative excess mortality.

The additive mortality model assumes that the

*Jeffrey Pai is an Assistant Professor in the I. H. Asper School of Business, University of Manitoba, Winnipeg, MB, Canada R3T 5V4, e-mail: j-pai@umanitoba.ca.

mortality rate at time t for the j th policyholder in the group is a sum of the base mortality rate and an excess mortality function $\alpha(t)$; that is

$$h_j(t) = \alpha(t) + \theta_j(t), \quad j = 1, 2, \dots, n.$$

Let $Y(t) = \sum_{j=1}^n Y_j(t)$ be the number at risk at time t and

$$\Theta(t) = \sum_{j=1}^n \int_0^t \theta_j(s) \frac{Y_j(s)}{Y(s)} ds$$

be the expected cumulative mortality rate. Let $A(t) = \int_0^t \alpha(s) ds$ be the cumulative excess mortality, which can be estimated by

$$\hat{A}(t) = \sum_{t_i \leq t} \frac{d_i}{Y(t)} - \Theta(t).$$

Again, the excess mortality function, $\alpha(t)$, can be estimated using the slope of the estimated cumulative excess mortality.

A sample data set, with Company = C1 and Plan = P1, was edited to left-truncated and right-censored data in Table 1, which consists of 23 rows with 28 policyholders. For example, the policyholder in the second row entered the plan at age 58 and left the plan three years later without experiencing the death event. Here “+” indicates

right-censored data. The policyholder in the 17th row entered the plan at age 76 and died three years later. In the 18th row, two policyholders entered at age 76 and then withdrew from the plan (censored) at age 79 (= 76 + 3).

The only table to be used for the comparison here is the base mortality table with gender = male, status = retirees and beneficiaries, and annuity size = small. I assume the constant force of mortality to estimate the force of mortality. For example, in the age group 50–54 where $q_x = 0.01514$ (see online Table 2a of the paper, Annuity Size Small) and the force of mortality is estimated as $\mu_x = -\log(1 - q_x) = 0.01526$.

In the relative mortality model,

$$\begin{aligned} Q(1) &= \sum_{j=1}^{28} \theta_j(1) Y_j(1) \\ &= \mu_*(57 + 1) + \mu_*(58 + 1) + \dots \\ &\quad + \mu_*(89 + 1) \\ &= 0.01698 + 0.01698 + \dots + 0.22267 \\ &= 1.50582 \end{aligned}$$

Table 1
Left-Truncated and Right-Censored Data

Observation	Begin Age	Time to Event	Number of Policies	Gender	Status	Annuity Size
1	57	1+	1	Male	Retiree	Small
2	58	3+	1	Male	Retiree	Small
3	63	1+	1	Male	Retiree	Small
4	64	2+	1	Male	Retiree	Small
5	66	1+	1	Male	Retiree	Small
6	66	1+	1	Male	Retiree	Small
7	67	1+	2	Male	Retiree	Small
8	67	3+	1	Male	Retiree	Small
9	67	4+	1	Male	Retiree	Small
10	69	4	1	Male	Retiree	Small
11	71	1+	1	Male	Retiree	Small
12	72	1	1	Male	Retiree	Small
13	73	1+	1	Male	Retiree	Small
14	74	1+	2	Male	Retiree	Small
15	75	1+	2	Male	Retiree	Small
16	75	1+	2	Male	Retiree	Small
17	76	3	1	Male	Retiree	Small
18	76	3+	2	Male	Retiree	Small
19	76	4	1	Male	Retiree	Small
20	76	4+	1	Male	Retiree	Small
21	77	1+	1	Male	Retiree	Small
22	78	4+	1	Male	Retiree	Small
23	89	2+	1	Male	Retiree	Small

$$\begin{aligned}
 Q(3) &= \sum_{j=1}^{28} \theta_j(3)Y_j(3) \\
 &= \mu_*(67 + 3) + \mu_*(67 + 3) + \dots \\
 &\quad + \mu_*(78 + 3) \\
 &= 0.04420 + 0.04420 + \dots + 0.10264 \\
 &= 0.58406
 \end{aligned}$$

$$Q(4) = \sum_{j=1}^{28} \theta_j(4)Y_j(4) = 0.39633$$

In general, $\mu_*(\cdot)$ is the value of the force of mortality depending on the policyholder's gender, status, and annuity size. In this example, only Table 2 is needed because all 28 policyholders are in the same category. The estimated cumulative relative excess mortality rates are

$$\begin{aligned}
 \hat{B}(1) &= \frac{1}{1.50582} = 0.66409, \\
 \hat{B}(3) &= \hat{B}(1) + \frac{1}{0.58406} = 2.37623, \\
 \hat{B}(4) &= \hat{B}(3) + \frac{1}{0.39633} = 7.42251.
 \end{aligned}$$

The estimated relative excess mortality function is

$$\hat{\beta}(t) = \begin{cases} 0.66409 & 0 \leq t < 1 \\ (2.37623 - 0.66409)/(3 - 1) \\ \quad = 0.85607 & 1 \leq t < 3 \\ (7.42251 - 2.37623)/(4 - 3) \\ \quad = 5.04628 & 3 \leq t < 4 \end{cases}$$

In the first year of observation, policyholders from Plan P1 were about 34% ($1 - 0.66409$) less likely, while in the fourth year were five (5.04628) times more likely, to die than comparable policyholders in the standard population.

In the additive model, the expected cumulative mortality rate, $\Theta(t)$, can be obtained recursively as

$$\Theta(t) = \Theta(t - 1) + \sum_t \mu_*(a_j + t - 1)/Y(t),$$

where the sum is over all policyholders in the

Table 2
**Base Mortality with Gender = Male,
 Status = Retirees & Beneficiaries, and
 Annuity Size = Small**

Age Group	q_x	μ_x
50-54	0.01514	0.01526
55-59	0.01684	0.01698
60-64	0.02219	0.02244
65-69	0.02938	0.02982
70-74	0.04324	0.04420
75-79	0.06319	0.06527
80-84	0.09755	0.10264
85-89	0.14207	0.15323
90-94	0.19962	0.22267
95-99	0.24040	0.27496

interval $[t - 1, t)$ and a_j is the begin age. For example,

$$\begin{aligned}
 \Theta(1) &= \sum_{j=1}^{28} \mu_*(a_j)/28 = \mu_*(57) + \mu_*(58) + \dots \\
 &\quad + \mu_*(89) = 0.04928 \\
 \Theta(2) &= \Theta(1) + \sum \mu_*(a_j + 1)/12 = 0.04928 \\
 &\quad + 0.06732 = 0.11660 \\
 \Theta(3) &= \Theta(2) + \sum \mu_*(a_j + 2)/10 = 0.11660 \\
 &\quad + 0.05553 = 0.17213 \\
 \Theta(4) &= \Theta(3) + \sum \mu_*(a_j + 3)/5 = 0.17213 \\
 &\quad + 0.06432 = 0.23645
 \end{aligned}$$

The estimated cumulative excess mortality rates are

$$\begin{aligned}
 \hat{A}(1) &= \frac{1}{28} - 0.04928 = -0.01357, \\
 \hat{A}(3) &= \frac{1}{28} + \frac{1}{10} - 0.17213 = -0.03641, \\
 \hat{A}(4) &= \frac{1}{28} + \frac{1}{10} + \frac{2}{5} - 0.23645 = 0.29926.
 \end{aligned}$$

The estimated excess mortality function is

$$\hat{\alpha}_t = \begin{cases} \frac{-0.03641 - (-0.01357)}{3 - 1} = -0.01142 & 1 \leq t < 3 \\ \frac{0.29927 - (-0.03641)}{4 - 3} = 0.33568 & 3 \leq t < 4 \end{cases}$$

The estimated excess mortality function can be interpreted as follows. In the plan P1, if there are 100 policyholders in the first year, we would see about 1 (-0.01357×100) less death than we would expect to see in a standard population. If there are 100 policyholders in the fourth year, we would see about 34 (0.33568×100) more deaths than we would expect to see in a standard population.

REFERENCES

- ANDERSEN, P. K., AND M. VÆTH. 1989. "Simple Parametric and Nonparametric Models for Excess and Relative Mortality." *Biometrics* 45: 523–35.
- KLEIN, J. P., AND M. L. MOESCHBERGER. 1997. *Survival Analysis*. New York: Springer-Verlag, 1997.

Additional discussions on this paper can be submitted until October 1, 2001. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.