

11-1-2008

# The Multinomial Regression Modeling of the Cause-of-Death Mortality of the Oldest Old in the U.S.


Dudley L. Poston Jr.

Texas A&M University, [d-poston@tamu.edu](mailto:d-poston@tamu.edu)

Hosik Min

University of Hawaii, [hosik@hawaii.edu](mailto:hosik@hawaii.edu)

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Poston, Dudley L. Jr. and Min, Hosik (2008) "The Multinomial Regression Modeling of the Cause-of-Death Mortality of the Oldest Old in the U.S.," *Journal of Modern Applied Statistical Methods*: Vol. 7: Iss. 2, Article 24.

Available at: <http://digitalcommons.wayne.edu/jmasm/vol7/iss2/24>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

## The Multinomial Regression Modeling of the Cause-of-Death Mortality of the Oldest Old in the U.S.

Dudley L. Poston, Jr.  
Texas A&M University

Hosik Min  
University of Hawaii

---

The statistical modeling of the causes of death of the oldest old (persons aged 80 and over) in the U.S. in 2001 was conducted in this article. Data were analyzed using a multinomial logistic regression model (MNLN) because multiple causes of death are coded on death certificates and the codes are nominal. The percentage distribution of the 10 major causes of death among the oldest old was first examined; we next estimated a multinomial logistic regression equation to predict the likelihood of elders dying of one of the causes of death compared to dying of an “other cause.” The independent variables used in the equation were age, sex, race, Hispanic origin, marital status, education, and metropolitan/non-metropolitan residence. Our analysis provides insights into the cause of death structure and dynamics of the oldest old in the U.S., demonstrates that MNLN is an appropriate statistical model when the dependent variable has nominal outcomes, and shows the statistical interpretation for complex results provided by MNLN.

Key words: multinomial regression, nominal outcome, logit, log odds, cause of death, mortality, oldest old, elderly, demography.

---

### Introduction

Demographers use multinomial logistic regression models when a dependent variable has more than two nominal categories. The choice is between a logistic model and a probit model, because the nominal categories of a variable are assumed to be unordered and more than two. If the outcome is dichotomous, logistic models are preferred. If the outcome is ordered, ordered or probit models are most appropriate (Long & Freese, 2003).

Background information about the causes of death of the U.S. is helpful in understanding the logic of the data analysis. The National Center for Health Statistics specifies the causes of death based on ICD-10 system (the 10<sup>th</sup> version of International Classification of Disease System). The causes are numerous

and nominal. Although there is a ranking system for the causes of the death, the rank does not mean a certain cause is superior over another; they are ranked based on incidence alone. One cannot say that a death due to a certain disease is more meaningful than the others. This article examines the top 10 causes of death for persons aged 80 and older in the U.S., as well as the likelihood of dying of a particular cause versus other causes.

The best-fitting statistical model for handling a nominal outcome is the multinomial logistic regression model (MNLN). It is not always easy to use MNLN, because MNLN has many parameters and the dependent variables have more than two categories. In addition, these parameters sometimes lead to complex results, which are often difficult to interpret. Poston & Min (2004) employed multinomial logit models for South Korean and American decedents and found that various sociodemographic factors influenced dying of specific causes of death compared to others.

This article focuses mainly on methodological issues, namely, the appropriateness of multinomial logit models for

---

Dudley L. Poston, Jr. is a Professor in the Department of Sociology. Email: dposton@tamu.edu. Hosik Min is a Faculty Assistant Specialist in the Center on The Family at the Mānoa campus. Email: hosik@hawaii.edu.

## MODELING OF THE CAUSE-OF-DEATH MORTALITY

studying causes of death, and the interpretation of the results of such investigations. Thus, the goals were to examine the likelihood of dying of a certain cause versus other causes for the oldest old (age 80 years and over) in the United States and to offer an easily understandable interpretation of MNLM. This is a particularly important concern, given the expected increases in the numbers of persons aged 80 and over in the U.S. in the next few decades. In the year 2000, the U.S. had a population of over 13 million oldest old people, 1.5% of the total U.S. population (Hetzel & Smith, 2001). Projections are for 24 million oldest old population in 2050, over 6% of the total U.S. population (Census Bureau, 2000a; 2000b).

Given such tremendous increases predicted for the population of the oldest old in the next few decades, an analysis of cause-of-death mortality in the current American oldest old population is particularly relevant. A study of the dynamics of current causes of death should suggest patterns of mortality that may be anticipated in the U.S. as the numbers of oldest old increase by 200% over the next five decades.

### Methodology

The data used in this article were obtained from death certificates filed in the U.S. The data were taken from 963,768 death certificates filed in 2001 for decedents age 80 and over (National Center for Health Statistics, 2003). The top 10 major causes of death for the oldest old Americans were heart disease; malignant neoplasms; cerebrovascular disease; chronic respiratory disease; Alzheimer's; influenza and pneumonia; diabetes; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia.

Estimation from a multinomial logistic regression, which predicts the likelihood of dying among oldest old American decedents of one of the major causes of death, compared to dying of an other cause, provides the main focus of this research. The independent variables used in the multinomial logistic equations were age, sex, race, Hispanic origin, marital status, education, and metropolitan/non-metropolitan residence.

With respect to the statistical method used herein, consider as an example only three major causes of death, and a residual category of all other causes. Thus, think of the multinomial logistic regression equation as providing an estimate for each of the independent variables and a set of four logit coefficients corresponding to each of the four categories of the dependent variable as follows (Stata Corporation, 2003, Vol. 2, p. 506-507):

$$\Pr(Y=1) = \frac{e^{Xb(1)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (1)$$

$$\Pr(Y=2) = \frac{e^{Xb(2)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (2)$$

$$\Pr(Y=3) = \frac{e^{Xb(3)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (3)$$

$$\Pr(Y=4) = \frac{e^{Xb(4)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (4)$$

The multinomial model cannot be identified unless one of the logits in each set is set to zero. Strictly speaking, it does not matter which one is set to zero. If we set  $\mathbf{b}^{(1)}$  to zero, then the remaining logit coefficients,  $\mathbf{b}^{(2)}$ ,  $\mathbf{b}^{(3)}$  and  $\mathbf{b}^{(4)}$ , will represent the change relative to the  $\mathbf{y}=\mathbf{1}$  category. In the example of cause-of-death mortality,  $\mathbf{b}^{(1)}$  will be the logit referring to deaths due to all other causes, and  $\mathbf{b}^{(2)}$ ,  $\mathbf{b}^{(3)}$  and  $\mathbf{b}^{(4)}$  will refer to deaths due to the three main causes being analyzed. Regarding the logit set to zero, its value becomes 1 because  $e^0 = 1$ .

If  $\mathbf{b}^{(1)}$  is set to zero, the equations for the four probabilities become:

$$\Pr(y=1) = \frac{1}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (5)$$

$$\Pr(Y=2) = \frac{e^{Xb(2)}}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (6)$$

$$\Pr(Y=3) = \frac{e^{Xb(3)}}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (7)$$

$$\Pr(Y=4) = \frac{e^{Xb(4)}}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (8)$$

In the actual multinomial logistic regression model, the top 10 causes of death and an 11<sup>th</sup> residual cause, i.e., dying of other causes, were used. Thus for each of the independent variables, 10 logits were formed from the contrasts of 10 non-redundant category pairs of the dependent variable modeling the logarithmic odds of dying of one of the 10 major causes of death versus dying of other causes. The estimated parameters are logit coefficients indicating the independent log odds of each independent variable being in the dependent variable category of interest, versus being in the base (or contrast) category of the dependent variable. The multinomial model was estimated using maximum likelihood procedures.

Separate logit coefficients were estimated for each independent variable for each of the dependent variable categories, excluding the outcome reference category. Thus the total number of parameters to be estimated was  $K \times (J - 1)$ , where  $K$  is the number of independent variables and  $J$  is the number of categories in the dependent variable. As shown below, there were 14 independent variables and the dependent variable consisted of 10 specific causes of death and a residual category of other causes. Thus the multinomial logistic equation estimated  $14 \times (10-1)$  logits, for a total of 126 coefficients. The “biggest challenge in using the multinomial logistic regression model was that the model includes a lot of parameters, and it was easy to be overwhelmed by the complexity of the results” (Long & Freese, 2003, p. 189).

### Results

In 2001, there were 963,768 death certificates filed for Americans age 80 and older in the U.S. As Table 1 shows, around 82% died from 10 main causes of death, as follows: heart disease, 36.7%; malignant neoplasms, 14.9%; cerebrovascular disease, 9.5%; chronic

respiratory disease, 5.0%; Alzheimer’s, 4.2%; influenza and pneumonia, 4.0%; diabetes, 2.5%; nephritis, nephrotic syndrome, and nephrosis, 1.9%; accidents, 1.8%, and septicemia, 1.4%. Around 18% of American oldest old decedents died of other causes.

The U.S. has low mortality levels in the general population as well as among the oldest old. The percentage of elderly decedents in 2001 was 80% of total deaths because America has completed the epidemiological transition (Omran, 1971; 1981). Omran’s epidemiological transition describes and explains variations in countries’ experiences of mortality changes through time. For example, at the first stage, mortality is high and fluctuating, precluding sustained population growth. At the second stage, mortality declines progressively, as epidemics decrease in frequency and magnitude, and life expectancy increases. As the gap between birth and death rates widens, rapid population growth ensues. In the third stage, mortality continues to decline and eventually approaches stability. Thus, mortality is low, life expectancy is high (over 70 years for both males and females), and deaths mainly occur from degenerative and man-made diseases (Olshansky & Ault, 1986).

### Multinomial Logistic Regression Results

We have shown that in 2001 there were 10 principal causes of death responsible for more than 82% of the deaths of U.S. oldest old. The remaining 18% of the oldest old decedents died for some other reason, treated here as a residual category of other causes.

Seven major classes of variables were used to predict cause-of-death mortality. They are age, sex, race, Hispanic origin, marital status, education and metro/non-metropolitan residence. From these seven classes of independent variables, we have developed 14 dummy variables, which were scored 1 if yes, as follows: 1) Age 90-99, and 2) Age 100+ (with Age 80-89 used as the reference variable); 3) Female; 4) Whites; and 5) Blacks (with Other races used as the reference group); 6) Hispanic origin; 7) Married, 8) Divorced, and 9) Widowed (with Never Married used as the reference variable); 10) Elementary School, 11) Junior High School, 12) High School, and 13)

## MODELING OF THE CAUSE-OF-DEATH MORTALITY

Table 1: Top 10 Causes of Death among the Oldest Old (80+): U.S., 2001

Cause of Death	Number of Decedents	Percent
Heart Disease	353,315	36.66
Malignant Neoplasms	143,915	14.93
Cerebrovascular Disease	91,848	9.53
Chronic Respiratory Disease	48,419	5.02
Alzheimer's	40,381	4.19
Influenza & Pneumonia	38,254	3.97
Diabetes	23,679	2.46
Nephritis, Nephrotic Syndrome & Nephrosis	18,200	1.89
Accidents	17,559	1.82
Septicemia	13,054	1.35
Other Causes	175,144	18.17
TOTAL	963,768	100.00

*Note:* National Center for Health Statistics, 2001 *Multiple Cause-of-Death File, NCHS CD-ROM, Series 20, No. 10H*. Hyattsville, Maryland: National Center for Health Statistics, 2003.

College or More (with Illiterate used as the reference variable); and 14) Metropolitan Residence. These 14 dummy variables are the explanatory (X) variables used in the multinomial logistic regression.

In Table 2 frequency distributions for these 14 independent variables for the 963,768 American oldest old who died in 2001 are presented. The majority of these decedents were aged 80-89 (almost 69%). Almost thirty percent were aged 90-99, and over 1% were aged 100 and over.

Almost two thirds were females (62%). Whites were the majority among the American oldest old (92%). Almost 7% were African Americans. Only 1.4% of oldest old Americans were other races. The majority of the oldest old Americans were of non-Hispanic origin (97%). According to Rogers et al. (2000), non-Hispanic whites have lower mortality risks than other groups, except Asian Americans. Asian American mortality is generally lower than that of non-Hispanic whites. Young Hispanic adults also have higher odds of mortality compared to non-Hispanic whites. African Americans suffer from the highest mortality risks compared to the

other groups. Widowed is the leading category in these decedents' marital status (63%), and married is next (27%). Almost 5% of the oldest old American decedents were divorced, and another 5% were never married. Regarding education, less than 1% was illiterate and over two thirds had high school or more education.

Most of these American oldest old lived in metropolitan areas at the time of death (76%). In the multinomial logistic regression model, therefore, 10 logits (one for each of the independent variables) are estimated for each of the 10 causes of death, modeling the log odds of dying of a major cause versus dying of other causes. Each logit coefficient will represent the independent log odds of the independent variable of being in the dependent variable category of interest, versus being in the base (or contrast) category of the dependent variable. In the multinomial logistic equation we will estimate  $14 \times (10-1)$  logits, for a total of 126 coefficients.

Table 3 presents the results of the multinomial logistic regression analysis for America's oldest old who died in 2001. Ten logit coefficients were estimated for each of the

10 principal causes of death. Each logit coefficient represents the independent log odds of the independent variable of being in the

Table 2: Frequency Distributions for Explanatory Variables: The Oldest Old (80+) Decedents, U.S., 2001

Variable		Frequency	Percent
Age	80-89	661,738	68.66
	90-99	285,185	29.59
	100+	16,845	1.75
	TOTAL	963,768	100.00
Sex	Male	362,292	37.59
	Female	601,476	62.41
	TOTAL	963,768	100.00
Race	White	883,639	91.69
	Black	66,393	6.89
	Others	13,736	1.43
	TOTAL	963,768	100.00
Hispanic	Hispanic	27,969	2.90
	Non-Hispanic	935,799	97.10
	TOTAL	963,768	100.00
Marital Status	Never Married	50,273	5.22
	Married	259,311	26.91
	Divorced	44,857	4.65
	Widowed	609,327	63.22
	TOTAL	963,768	100.00
Education	Illiterate	6,384	0.66
	Elementary	74,942	7.78
	Junior High	198,781	20.63
	High	455,744	47.29
	College+	227,917	23.65
	TOTAL	963,768	100.00
Residence	Non-Metro	230,239	23.89
	Metro	733,529	76.11
	TOTAL	963,768	100.00

particular cause-of-death category, versus being in the contrast category of the dependent variable of other causes. If no relationship exists, the coefficient would be 0. Negative coefficients indicate a negative association, that is, negative

chances or log odds of being in the dependent variable category of interest, and positive coefficients indicate positive chances.

The second column of Table 3 presents the results of the log odds of dying of malignant neoplasms versus dying of other causes (the residual category). Malignant neoplasms were the second major cause of death among American oldest old who died in 2001, with 143,915 deaths from this cause. Table 1 also shows that the residual category of other causes was the cause of death of 175,144 American oldest old in 2001.

Ten independent variables were used to estimate the log odds of dying of malignant neoplasms versus dying of other causes. The first logit coefficient shown in the second column of Table 3 is -0.73 for age 90-99. This means that for decedents who were age 90-99 compared to those who were 80-89, there is a decrease of 0.73 in the log odds of dying of malignant neoplasms compared to dying of other causes. The second logit coefficient is -1.65; this means that for American decedents age 100 or more, compared to those who were 80-89, there is a decrease of 1.65 in their log odds of dying of malignant neoplasms compared to dying of other causes. Hence, the older the decedent, the less the log odds that the person died of malignant neoplasms compared to other causes. Each logit coefficient reflects the effect of the particular independent variable on the dependent variable, controlling for the effects on the dependent variable of the other independent variables in the multinomial logistic regression equation. Thus, the effects of age on malignant neoplasms mortality are independent of the effects on malignant neoplasms of sex, marital status, race, Hispanic origin, educational attainment, and residence.

The estimated parameter effects are interpreted straightforwardly when converted into odds ratios, which is done by exponentiating the coefficients. Odds ratios in the multinomial logistic regression equation are typically referred to as relative risk ratios. This is the relative risk, or the odds, of being in the dependent variable category of interest and not being in the contrast category of the dependent variable for the dummy independent variable

## MODELING OF THE CAUSE-OF-DEATH MORTALITY

Table 3: Logit Coefficients from Multinomial Logistic Regression of Dying of 1 of 11 Causes vs. Dying of other causes, on Selected Social and Demographic Factors: Oldest Old Decedents, U.S., 2001

Independent Variables	Cause of Death				
	1	2	3	4	5
	Heart Disease	Malignant Neoplasms	Cerebrovascular Disease	Chronic Respiratory Disease	Alzheimer's
Age 80-89	Reference	Reference	Reference	Reference	Reference
Age 90-99	0.03***	-0.73***	-0.06***	-0.63***	0.26***
Age 100+	-0.00	-1.65***	-0.43***	-1.25***	0.43***
Female	-0.08***	-0.33***	0.22***	-0.38***	-0.23***
Other races	Reference	Reference	Reference	Reference	Reference
White	-0.01	-0.14***	-0.37***	0.10*	-0.50***
Black	0.06*	0.10**	-0.26***	-0.41***	-0.55***
Hispanic	0.10***	0.06*	0.01	-0.24***	0.27***
Never Married	Reference	Reference	Reference	Reference	Reference
Married	0.01	0.31***	0.21***	0.08**	-0.22***
Divorced	0.01	0.17***	0.12***	0.45***	-0.05
Widowed	0.04**	0.14***	0.15***	0.22***	-0.10***
Illiterate	Reference	Reference	Reference	Reference	Reference
Elementary	0.08*	0.17***	-0.00	0.03	-0.06
Junior High	0.07*	0.19***	0.02	-0.04	-0.07
High School	0.05	0.25***	-0.02	-0.00	-0.10
College or More	-0.05	0.24***	0.03	-0.19**	-0.12
Metro Residence	0.03***	0.03**	-0.08***	0.01	-0.03*
Intercept	0.65***	-0.12***	-0.51***	-1.08***	-0.75***
N	353,315	143,925	91,848	48,419	40,381

\*p < .05; \*\* p < .01; \*\*\*p < .001; Reference group is dying of other causes (N = 175,144)

versus the reference category (Stata Corporation, 2003, p. 510-511).

The odds ratio for Age 90-99 is  $e^{-0.73} = 0.48$ , which means that the odds of persons aged 90-99, compared to those age 80-89, dying of malignant neoplasms versus dying of other causes may be multiplied by 0.48, which means they decrease. The percentage amount of change may be determined in the odds by subtracting 1 from the odds ratio and multiplying the difference by 100:  $(0.48 - 1) * 100 = -0.52$ . This indicates that the odds of dying of malignant neoplasms versus dying of other causes are 52% less for persons aged 90-99 compared to those aged 80-89. In contrast, the odds of dying of malignant neoplasms versus dying of other

causes are 81% less for persons aged 100 and over compared to those aged 80-89, that is,

$$(e^{-1.65} - 1) * 100 = -81.$$

### Logits on Each of the 10 Causes of Death

The pattern of the effects shown for malignant neoplasms is one in which the log odds of 90-99 year-old decedents dying of malignant neoplasms versus dying of other causes are less than those of 80-89 year-old decedents, and the log odds of 100+ year-old decedents compared to those of 80-89 year-old decedents are more negative.

This pattern of increasingly negative log odds for decedents 100 years or older compared

POSTON & MIN

Table 3 Continued. Logit Coefficients from Multinomial Logistic Regression of Dying of 1 of 11 Causes vs. Dying of other causes, on Selected Social and Demographic Factors: Oldest Old Decedents, U.S., 2001

Independent Variables	Cause of Death				
	6	7	8	9	10
	Influenza & Pneumonia	Diabetes	Nephritis, Nephrotic Syndrome & Nephrosis	Accidents	Septicemia
Age 80-89	Reference	Reference	Reference	Reference	Reference
Age 90-99	-0.57***	0.14***	-0.11***	-0.10***	-0.21***
Age 100+	-1.33***	-0.20***	-0.40***	-0.27***	-0.67***
Female	0.02	0.47***	-0.36***	-0.28***	-0.08***
Other races	Reference	Reference	Reference	Reference	Reference
White	-0.49***	0.53***	-0.19**	-0.28***	0.20*
Black	0.14**	0.29***	0.38***	-0.53***	0.86***
Hispanic	0.74***	-0.22***	0.11*	-0.06	0.07
Never Married	Reference	Reference	Reference	Reference	Reference
Married	0.23***	0.37***	0.04	-0.05	-0.19***
Divorced	0.18***	0.16***	0.01	-0.07	-0.13*
Widowed	0.25***	0.17***	0.06	-0.07*	-0.14***
Illiterate	Reference	Reference	Reference	Reference	Reference
Elementary	0.00	0.14	0.01	0.21	-0.01
Junior High	-0.03	0.17*	0.02	0.26*	-0.03
High School	-0.18*	0.20*	-0.08	0.27*	-0.06
College or More	-0.39***	0.28**	-0.28**	0.33**	-0.25*
Metro Residence	-0.14***	-0.00	-0.13***	-0.25***	0.10***
Intercept	-1.40***	-2.75***	-1.72***	-1.83***	-2.57***
N	38,254	23,679	18,200	17,559	13,054
Model chi-square (df) = 39905.58*** (140); Pseudo R <sup>2</sup> = .01					
*p < .05; **p < .01; ***p < .001; Reference group is Dying of other causes (N = 175,144)					

to 80-89 year-old decedents over the log odds of 90-99 year-old decedents compared to 80-89 year-old decedents is found in most of the other major causes of death except heart disease, Alzheimer's, and diabetes. For instance, the logit coefficients for 90-99 and 100+ year-old decedents for cerebrovascular disease are -0.06 and -0.43; chronic respiratory disease, -0.63 and 1.25; influenza and pneumonia, -0.57 and -1.33; nephritis, nephrotic syndrome and nephrosis -0.11 and -0.40; accidents -0.10 and -0.27; and for septicemia -0.21 and -0.67, respectively.

However, this association does not hold for heart disease, Alzheimer's, and diabetes. Alzheimer's has a positive association with increasing age.

Sex was a dummy variable labeled female (female = 1, male = 0). The logit coefficient for female for malignant neoplasms is -0.33. Exponentiating the logit coefficient transforms it into an odds ratio; that is,  $e^{-0.33} = 0.72$ . This means that the odds of females are 28% lower than the odds of males of dying of malignant neoplasms compared to dying of other causes; that is,  $(e^{-0.33}-1) \times 100 = -28$ . These negative odds of females, compared to males,



## MODELING OF THE CAUSE-OF-DEATH MORTALITY

are also found for heart disease; chronic respiratory disease; Alzheimer's; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia. For the remaining causes of death, the odds of a female compared to a male dying of the specified cause versus dying of other causes are more. There is no statistically significant relationship for the cause of influenza and pneumonia.

Race was comprised of two dummy variables (White and Black), with the other races category used as the reference category. The logit coefficient for whites for malignant neoplasms is -0.14. Exponentiating the logit coefficient transforms it into an odds ratio; that is,  $e^{-0.14} = 0.87$ . This means that the odds of whites are 13% lower than the odds of other races of dying of malignant neoplasms compared to dying of other causes, that is,  $(e^{-0.14} - 1) \times 100 = -13$ . These negative odds of whites, compared to other races are also found for cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents. For the remaining causes of death, except for heart disease, the odds of whites compared to other races of dying of the specified cause versus dying of other causes are more. There is no statistically significant relationship for this one cause of heart disease. The logit coefficient for blacks for malignant neoplasms is 0.10. Exponentiating the logit coefficient transforms it into an odds ratio, that is,  $e^{0.10} = 1.11$ . This means that the odds of blacks are 11% higher than the odds of other races of dying of malignant neoplasms compared to dying of other causes, that is,  $(e^{0.10} - 1) \times 100 = .11$ . We find these positive odds of blacks, compared to other races for heart disease; influenza and pneumonia; diabetes, nephritis, nephrotic syndrome, and nephrosis; and septicemia. For the remaining causes of death, the odds of blacks compared to other races of dying of the specified cause versus dying of other causes are less.

The Hispanic origin dummy variable was labeled Hispanic (Hispanic = 1, non-Hispanic = 0). The logit coefficient for Hispanic origin for malignant neoplasms is 0.06. Exponentiating the logit coefficient transforms it into an odds ratio, that is,  $e^{0.06} = 1.06$ . This means that the odds for Hispanic origin are 6%

higher than the odds for non-Hispanics dying of malignant neoplasms compared to dying of other causes, that is,  $(e^{0.06} - 1) \times 100 = 6$ . These positive odds for Hispanic origin, compared to non-Hispanic origin, are also found for heart disease; Alzheimer's; influenza and pneumonia; and nephritis, nephrotic syndrome, and nephrosis, accidents. For the remaining causes of death, except for cerebrovascular disease, accidents, and septicemia, the odds of Hispanics compared to non-Hispanics dying of the specified cause versus dying of other causes are less. There is no statistically significant relationship for cerebrovascular disease and accidents.

Regarding marital status, 6 of 10 causes of death with the marital status variable had significant relationships. The logit coefficient for married for malignant neoplasms is 0.31. Exponentiating the logit coefficient to an odds ratio, equals  $e^{0.31} = 1.36$ . This means that the odds of those who were married of dying of malignant neoplasms compared to dying of other causes, are 36% higher than the odds of those who never married; that is,  $(e^{0.31} - 1) \times 100 = 0.36$ .

These positive associations are also found for cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes. Only Alzheimer's and septicemia show negative associations. The other remaining causes of death have no statistical relationships. This positive relationship between having been married and the odds of dying of malignant neoplasms versus dying of other causes are also similar to those who were divorced and widowed. The odds of those who were divorced compared to those who never married of dying of the specified cause versus dying of other causes are positive and significant for malignant neoplasms, cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes.

Only septicemia has a negative relationship. The other remaining causes of death have no significant relationships. The odds of those widowed compared to those never married of dying of the specified cause versus dying of other causes are positive and significant for heart disease, malignant neoplasms, cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes.

Alzheimer's, accidents, and septicemia have negative associations. The other remaining causes of death have no statistical relationships.

Education was comprised of four dummy variables (elementary school, junior high school, high school, college or more), with the illiterate category used as the reference category. Education is the least important variable. Only malignant neoplasms has positive and significant relationships for all education categories. However, the highest education category does not show higher log odds than high school education. All the other causes of death, other than malignant neoplasms, are either not statistically significant, or only one or two are significant.

The final explanatory variable pertains to metropolitan/non-metropolitan residence (scored 1 if the person was a metropolitan resident at the time of death, and 0 if a non-metropolitan resident). The logit coefficient for this variable and the likelihood of dying of malignant neoplasms was 0.03. This means that if the oldest old decedent was residing in a metropolitan area at the time of death, he/she had odds of dying of malignant neoplasms versus dying of other causes that are 3% more than those of a decedent who was living in a non-metropolitan area at the time of death, that is  $(e^{0.03}-1) * 100 = 3$ . This kind of positive association is also found for heart disease and septicemia. Five causes of death also have negative and statistically significant logits (cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents). The other remaining causes of death have no statistically significant relationships.

At the base of Table 3 are two statistics that gauge the degree of fit of the overall model examined. The model chi-square statistic has a value of 39,905.58, with 140 degrees of freedom (one for each of the logits being estimated). These chi-square values are sufficiently large to reject the null hypothesis that the 126 logit coefficients are all zero. This finding is also shown by the fact that the majority of the logit coefficients are statistically significant.

A value of 0.01 for Pseudo  $R^2$  statistic for the U.S. is also shown. Although this statistic does not have anywhere near as straightforward

an interpretation as the explained variance interpretation that  $R^2$  has in ordinary least squares regression, it is nevertheless a rough gauge of the degree of fit of the model used. With a low value of 0.01, these indicate that there are surely other independent variables, in addition to those used in Table 3, that are important in predicting the likelihood of oldest old Americans dying of a major cause of death instead of dying of all other causes.

### Conclusion

In this article, the cause of death structure for the oldest old decedents in the United States in 2001 was examined. The three main causes-of-death for the American decedents, which comprised close to two thirds of all deaths, were heart disease, malignant neoplasms, and cerebrovascular disease. The top 10 causes of death accounted for over 82% of all deaths, and the residual category of other causes accounted for 18% of the deaths.

A multinomial logistic regression equation was estimated to predict the patterns of cause-of-death mortality for the 963,768 oldest old Americans who died in 2001. The primary goal was to ascertain which independent variables best predicted the log odds of dying of one of the major causes of death compared to dying of other causes. The best predictors were age, sex, race, Hispanic origin, and metropolitan residence. Marital status and education did not perform as well. In particular, education was found to be the least important variable in the multinomial equation.

Also, the older the American decedent, the less likely he/she would die of a major cause of death compared to other causes. This relationship was found for most of the 10 main causes of death. Regarding the independent variable of sex, it was found that females in the U.S. were less likely than males to die of one of seven main causes (heart disease; malignant disease; chronic respiratory disease; Alzheimer's; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia).

With respect to race, whites had negative and statistically significant logits for 6 of the 10 causes of death (malignant neoplasms; cerebrovascular disease; Alzheimer's; influenza

## MODELING OF THE CAUSE-OF-DEATH MORTALITY

and pneumonia; nephritis, nephritic syndrome, and nephrosis; and accidents); three causes of death had positive and statistically significant logits (chronic respiratory disease, diabetes, and septicemia), while heart disease had no statistical relationships. It was also found that blacks had positive and statistically significant logits for 6 of the 10 causes of death (heart disease; malignant neoplasms; influenza and pneumonia; diabetes; nephritis, nephrotic syndrome, and nephrosis; and septicemia); the remaining four causes of death had negative and statistically significant logits (cerebrovascular disease, chronic respiratory disease, Alzheimer's, and accidents).

Hispanic origin had positive and statistically significant logits for 5 of the 10 causes of death (heart disease; malignant neoplasms; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents); two causes of death had negative and statistically significant logits (chronic respiratory disease and diabetes), and the remaining other causes of death had no statistical relationships.

Results indicated that metropolitan residence had negative and statistically significant logits for 5 of the 10 causes of death (cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents); three causes of death had positive and statistically significant logits (heart disease, malignant neoplasms, and septicemia), and the remaining causes of death had no statistical relationships.

In the next 50 years, the number of oldest old persons in the U.S. is projected to increase almost two times, from 13 million in the year 2000 to almost 25 million in the year 2050. The analyses and results reported in this paper of the cause-of-death structure of the U.S. oldest old decedents in 2001 could well reflect the cause-of-death structure of the increasingly large numbers of oldest old decedents in future decades. The analyses of the dynamics of the current causes of death could suggest the patterns of mortality that may be anticipated for the growing population of oldest old Americans in the next several decades.

This study also demonstrates the appropriate usage of multinomial logistic

regression models when the dependent variable has more than two nominal categories. It has been found that multinomial logistic modeling exhibits suitable statistical interpretations for complex results, weakening the criticism of using such a model for these types of data.

### References

- Hetzel, R., & Smith, A. (2001). *The 65 years and over population: 2000, Census 2000 brief*. U.S. Department of Commerce. Economics and Statistics Administration. U.S. Census Bureau. Washington D.C.
- Long, J. S., & Freese, J. (2003). *Regression models for categorical dependent variables using Stata*. (Revised ed.). College Station, Texas: Stata Press.
- National Center for Health Statistics. (2003). *1998 Multiple cause-of-death file, NCHS CD-ROM, Series 20, No. 10H*. Hyattsville, Maryland: National Center for Health Statistics.
- Olshansky, S. J., & Ault, A. B. (1986). The fourth stage of the epidemiologic transition: The age of delayed degenerative disease." *Milbank Memorial Fund Quarterly*, 64, 355-391.
- Omran, A. R. (1971). The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly*, 49, 509-538.
- Omran, A. R. (1981). The epidemiologic transition. In Ross, J. A. (ed.), *International encyclopedia of population*, pp. 172-175. New York: The Free Press.
- Poston, D. L., Jr., & Min, H. S. (2004). Cause of death mortality of the oldest old in the Republic of Korea in 2001, with comparison to the United States in 1998. In *International Conference on Current Issues of the Elderly in the 21<sup>st</sup> Century*, 1-19, Institute of Gerontology, Yeungnam University, October 31, 2003.
- Rogers, R. G., Hummer, R. A., & Nam, C. B. (2000). *Living and dying in the USA*. Academic Press. San Diego, CA.
- Stata Corporation. (2003). *Stata base reference manual, release 8 (2)*. College Station, Texas: Stata Corporation.