Predictive Modeling with Longitudinal Data: A Case Study of Wisconsin Nursing Homes

Marjorie A. Rosenberg,* Edward W. Frees,† Jiafeng Sun,‡
Paul H. Johnson, Jr.,‖ and James M. Robinson¶

Abstract

The recent development and availability of sophisticated computer software has facilitated the use of predictive modeling by actuaries and other financial analysts. Predictive modeling has been used for several applications in both the health and property and casualty sectors. Often these applications employ extensions of industry-specific techniques and do not make full use of information contained in the data. In contrast, we employ fundamental statistical methods for predictive modeling that can be used in a variety of disciplines. As demonstrated in this article, this methodology permits a disciplined approach to model building, including model development and validation phases. This article is intended as a tutorial for the analyst interested in using predictive modeling by making the process more transparent.

This article illustrates the predictive modeling process using State of Wisconsin nursing home cost reports. We examine utilization of approximately 400 nursing homes from 1989 to 2001. Because the data vary both in the cross section and over time, we employ longitudinal models. This article demonstrates many of the common difficulties that analysts face in analyzing longitudinal health care data, as well as techniques for addressing these difficulties. We find that longitudinal methods, which use historical trend information, significantly outperform regression models that do not take advantage of historical trends.

1. Introduction

The recent development and availability of sophisticated computer software has facilitated the use of predictive modeling by actuaries and other financial analysts. Predictive modeling refers to a statistical process of analyzing data related to some problem of interest. This process can be described as (1) defining the problem to be studied, (2) collecting sufficient knowledge about the problem and obtaining appropriate data, (3) examining trends in the data to aid in developing candidate models (sometimes referred to as data mining), (4) estimating the candidate models via reasonable methods, and (5) using diagnostic analyses and selection criteria to decide which of the candidate models is best for

* Marjorie A. Rosenberg, FSA, MAAA, Ph.D. is an Associate Professor with a joint appointment in the Department of Actuarial Science, Risk Management and Insurance and the Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53706, mrosenberg@bus.wisc.edu.

† Edward W. Frees, FSA, MAAA, Ph.D. is a Professor in the Department of Actuarial Science, Risk Management and Insurance, University of Wisconsin–Madison, Madison, WI 53706, jfrees@bus.wisc.edu.

‡ Jiafeng Sun is a Ph.D. candidate in the Department of Actuarial Science, Risk Management and Insurance, University of Wisconsin–Madison, Madison, WI 53706, jiafengsun@wisc.edu.

‖ Paul H. Johnson, Jr. is a Ph.D. candidate in the Department of Actuarial Science, Risk Management and Insurance, School Of Business, University of Wisconsin–Madison, Madison, WI 53706, pauljohnson@wisc.edu.

¶ Jim Robinson, FSA, MAAA, Ph.D., is a Senior Scientist and Deputy Director of the Center for Health Systems Research and Analysis, University of Wisconsin–Madison, Madison, WI 53706-2397, jim@chsra.wisc.edu.

analyzing the problem (Rosenberg and Johnson 2007). This article is intended as a tutorial for the analyst interested in using longitudinal data for predictive modeling and focuses on the development, estimation, and validation of predictive models.

Predictive modeling has been used extensively in the health care sector; see, for example, Cumming and Cameron (2002) for an evaluation of several commercial health-care-related predictive modeling software packages. Many health-sector applications require that predictive models account for variables related to the health status of individuals or groups by including covariates like age, gender, race/ethnicity, diagnosis codes, cultural and socioeconomic attributes, and health assessment measures such as quality of life and self-reported health. Accounting for health-status-related covariates is known as risk adjustment (Iezzoni 1997). There are four major health-sector applications of predictive modeling. One application is provider profiling, where provider performances are ordered based on the quality of treatment, number of tests, and disease severity of their case mix (Christiansen and Morris 1997; Delong et al. 1997; Hu and Lesneski 2004). A second application is provider reimbursement, where providers who treat Medicare or Medicaid insured patients receive payment as determined by a statistical model (Ash and Byrne-Logan 1998; Ash et al. 2000; Pope et al. 2000; Kronick et al. 2000). A third application is the identification of individual or groups of patients that are likely to be high-cost users of medical services in future periods, for the purpose of targeting them with interventions to reduce future costs (Cousins, Shickle, and Bander 2002; Passwater and Seiler 2004; Dove, Duncan, and Robb 2003; Meenan et al. 1999; Zhao et al. 2003). Finally, predictive modeling is used to supplement the underwriting and pricing of small group health insurance (Cumming et al. 2002; Ellis et al. 2003; Hu and Lesneski 2004).

Predictive modeling also has been utilized in the property-casualty sector. Derrig (2002) provided a general overview of insurance fraud detection and deterrence strategies, and other analysts have developed predictive models for detecting and classifying types of insurance fraud (Brockett et al. 2002; Tennyson and Slasas-Forn 2002; Viaene et al. 2002). Predictive modeling also has been used to relate credit scores to personal automobile or homeowners' profitability (Monaghan 2000; Wu and Guszcza 2003). The American Academy of Actuaries summarized four studies that discussed the use of credit history for personal lines of insurance (American Academy of Actuaries Risk Classification Subcommittee of the Property/Casualty Products and Committee 2002). Medical malpractice claims and litigation have been analyzed with predictive models (Cooil 1991; Weycker and Jensen 2000). Claims reserving naturally lends itself to predictive modeling (Guszcza and Lommele 2006). Finally, predictive modeling can assist in predicting claimant behavior in workers' compensation (Biddle and Roberts 2003; Speights, Brodsky, and Chudova 1999).

This article illustrates the predictive modeling process using data from State of Wisconsin nursing home cost reports. We examine utilization of approximately 400 nursing homes from 1989 to 2001 and are interested in forecasting total patient years (number of total patient days in the cost reporting period divided by number of facility operating days in the cost reporting period) by individual nursing home. The Wisconsin nursing home data vary both in the cross-section and over time; therefore, we use longitudinal (panel data) models. Covariates are included as with ordinary regression, yet differences between subjects and changes over time can be incorporated. See Baltagi (2005), Diggle et al. (2002), or Frees (2004) for a general discussion of longitudinal data modeling. This article demonstrates how poorly cross-sectional methods perform because they do not take advantage of historical trends when predicting future outcomes, as compared to longitudinal methods that use historical trend information. This article also demonstrates many of the common difficulties that analysts face in analyzing longitudinal health care data, as well as techniques for addressing these difficulties. The longitudinal data approach used in this article is only one of several predictive modeling approaches; other approaches are discussed in the Summary section.

The data are used to illustrate the modeling techniques. The empirical results obtained provide projections about future nursing home utilization. However, any general conclusions regarding nursing homes/custodial care would consider the economic, financial, and legal environments in which the

nursing homes operate and would incorporate related variables (such as the demand and supply of beds, the degree of competition within the market, Wisconsin poverty levels, and Wisconsin Medicaid legislation).

The remainder of the article is organized as follows. Section 2 describes the nursing home data. Section 3 provides summary statistics and develops and estimates candidate longitudinal models. Section 4 validates these candidate models by predicting future outcomes. Section 5 provides a summary and concluding remarks. An appendix provides details of the model selection and sample statistical code.

2. CASE STUDY ON NURSING HOME DATA

The State of Wisconsin Medicaid program funds nursing home care for individuals qualifying on the basis of need and financial status. Nursing home care providers are categorized into three types: nursing facilities providing skilled care to disabled adults and frail elderly, facilities serving people with developmental disabilities (FDDs), and facilities serving both populations in separate units. Most, but not all, nursing homes in Wisconsin are certified to provide Medicaid-funded care. Those that do not accept Medicaid are generally paid directly by the resident or the resident's insurer. Most, but not all, nursing facilities are certified to provide Medicare-funded care. Medicare provides postacute care for 100 days following a related hospitalization. Medicare does not fund care provided by FDDs.

Nursing homes are owned and operated by a variety of entities, including the state, counties, municipalities, for-profit businesses, and tax-exempt organizations. Private firms often own several nursing homes. Periodically facilities may change ownership and, less frequently, ownership type. Some nursing homes opt not to purchase private insurance coverage for their employees. Instead, these facilities directly provide insurance and pension benefits to their employees; this is referred to as ''self-funding of insurance.''

As part of the conditions for participation, Medicaid-certified nursing homes must file an annual cost report to the Wisconsin Department of Health and Family Services (DHFS) summarizing the volume and cost of care provided to all of its residents, Medicaid-funded and otherwise. These cost reports are audited by DHFS staff and form the basis for facility-specific Medicaid daily payment rates for subsequent periods. Medicaid daily payment rate schedules vary annually by facility and by resident classification of level of care within each facility. The data are publicly available. Interested parties can contact the DHFS to request the data; see http://dhfs.wisconsin.gov/contact.htm for contact information.

Utilization of nursing home care is measured in patient days. Medicaid facilities bill the Medicaid fiscal intermediary at the end of each month for total Medicaid patient days incurred in the month, itemized by resident and level of care. Projections of Medicaid patient days by facility and level of care play a key role in the annual process of updating facility Medicaid rate schedules. The projected total patient days are applied to estimated rate schedules to provide an estimate of aggregate Medicaid payments to nursing homes. The rate formula, which translates historical reported costs per patient day for a facility into a proposed rate schedule, is iteratively adjusted until estimated aggregate payments satisfy budget constraints while providing adequate and equitable reimbursement to providers across ownership type, geographic, and other dimensions. Typically DHFS obtains short-term forecasts of one or two fiscal years of aggregate patient days, by separately trending historical Medicaid patient days by level of care. These level-of-care forecasts are stratified by facility roughly in proportion to reported Medicaid total patient days by facility and level of care for the most recently available audited cost report period.

DHFS is interested in determining whether more sophisticated projection techniques exist that might provide more reliable total patient days forecasts by level of care and facility. Because the factors influencing trends in FDD and nursing facility service utilization differ, and nursing facility payments dominate FDD care payments, we focus our analysis initially on nursing facility total patient days.

3. MODEL DEVELOPMENT AND ESTIMATION

In this section we use longitudinal data techniques to analyze the nursing home data from State of Wisconsin nursing home cost reports between 1989 and 2001. There are three general classes of longitudinal models that have been discussed in the literature: subject-specific models, marginal models, and transition models (Diggle et al. 2002). In this article, we focus on subject-specific models, which are appropriate for forecasting outcomes at the subject level.

We consider commonly used longitudinal models that assume that the outcome variable follows a multivariate normal distribution. These models are easily implemented and interpreted. We begin in Section 3.1 with a description of the key variables for the entire data set. We then partition the data into an in-sample portion to develop models, and an out-of-sample portion to assess the forecasts of competing models. To calibrate these models and select the most appropriate representation, 4076 observations from 398 nursing facilities, from 1989 to 1999, are included in the in-sample data. Section 3.2 discusses the in-sample model fitting. Section 4 examines the prediction accuracy of different models for the out-of-sample data.

3.1 Summary Statistics

Table 1 describes the variables considered in this analysis. The outcome variable is total patient years (TPY) defined to be number of total patient days in the cost reporting period divided by number of facility operating days in the cost reporting period. The median of total patient years per facility was 86.7 per year. Appendix A describes the statistical decision-making process for choosing to analyze patient years in lieu of patient days. The number of beds and square footage of the nursing home both measure the size of the facility. Not surprisingly, these continuous covariates turn out to be important predictors of TPY; larger facilities have a higher capacity and are likely to have more patients.

Table 1 also describes several categorical covariates. About half of the facilities have self-funding of insurance; these facilities have a higher median TPY than nursing homes that do not self-fund. Approximately 70% of the facilities are Medicare-certified; Medicare-certified facilities are also larger in terms of the median TPY. Regarding the organizational structure, about half (52.4%) are run on a for-profit basis, about one-third (36.8%) are organized as tax exempt, and the remainder are governmental organizations. The government facilities have the highest median TPY. Slightly more than half of the

Table 1
Variable Descriptions with Percentages and Median TPY by Subgroup

Variable	Description	Percentage	Median TPY
TPY	Total patient years (median 86.7)		
Continuous covariates:			
YEAR	Cost report year minus 1988 (1 to 13)		
NumBed	No. of beds (median 95)		
SqrFoot	Nursing home net square footage (in thousands of square feet, median 37.27)		
Categorical covariates:			
POPID	Nursing home identification number		
SelfFundIns	Self-funding of insurance		
	Yes	50.4	97.1
	No	49.6	74.3
MCert	Medicare-certified		
	Yes	70.2	92.0
	No	29.8	70.0
Organizational	Pro (nursing home is for-profit)	52.4	82.7
Structure	TaxExempt (nursing home is tax exempt)	36.8	86.2
	Govt (nursing home is a governmental unit)	10.8	111.8
Location	Urban	53.8	95.5
	Rural	46.2	78.7
MSA	Metropolitan Statistical Area code (1 to 13, 0 rural)		

Table 2
Summary Statistics of Total Patient Years (TPY), by Year

Year	No. of Facilities	Mean	Median	Standard Deviation	Minimum	Maximum	Coefficient of Variation
1989	376	102.0	88.3	67.0	15.1	645.0	65.7
1990	376	103.4	88.5	66.9	14.2	650.6	64.7
1991	368	104.1	90.1	66.8	13.2	655.1	64.2
1992	372	105.2	89.3	67.5	16.0	653.0	64.2
1993	367	105.5	89.7	68.0	16.5	657.3	64.5
1994	370	104.4	89.4	67.2	14.7	658.6	64.4
1995	370	104.6	90.4	67.5	16.6	669.3	64.5
1996	372	101.8	87.8	65.3	15.7	657.3	64.1
1997	370	99.7	85.3	62.3	15.7	614.4	62.5
1998	365	97.2	85.1	59.5	11.9	551.5	61.2
1999	370	92.4	82.7	52.3	11.6	439.6	56.6
2000	362	88.8	80.5	46.1	11.6	314.7	51.9
2001	355	89.7	81.1	49.0	12.3	440.7	54.7
1989–2001	4793	100.0	86.7	62.7	11.6	669.3	62.7

Note: Years 1989–99 correspond to in-sample data; years 2000-2001 correspond to out-of-sample data.

facilities are located in an urban environment (53.8%); these facilities have a higher median TPY than those located in rural environments.

Table 2 shows the distribution of TPY over time. The number of facilities varies from year to year, with an average of 368.7 per year (4,793/13). The number of facilities is relatively stable over time, whereas there has been a small decrease in the typical (mean or median) TPY over time.

Because the standard deviation is large relative to the mean, Table 2 also suggests that the distribution of TPY is skewed. This is reinforced by Figure 1, which provides a histogram of TPY. We see that the distribution is right skewed, with large values of TPY relatively common. The histogram of TPY values on a natural logarithmic scale exhibits a more symmetric, and thinner tailed, distribution.

Figure 1
Histogram of TPY and Logarithmic TPY

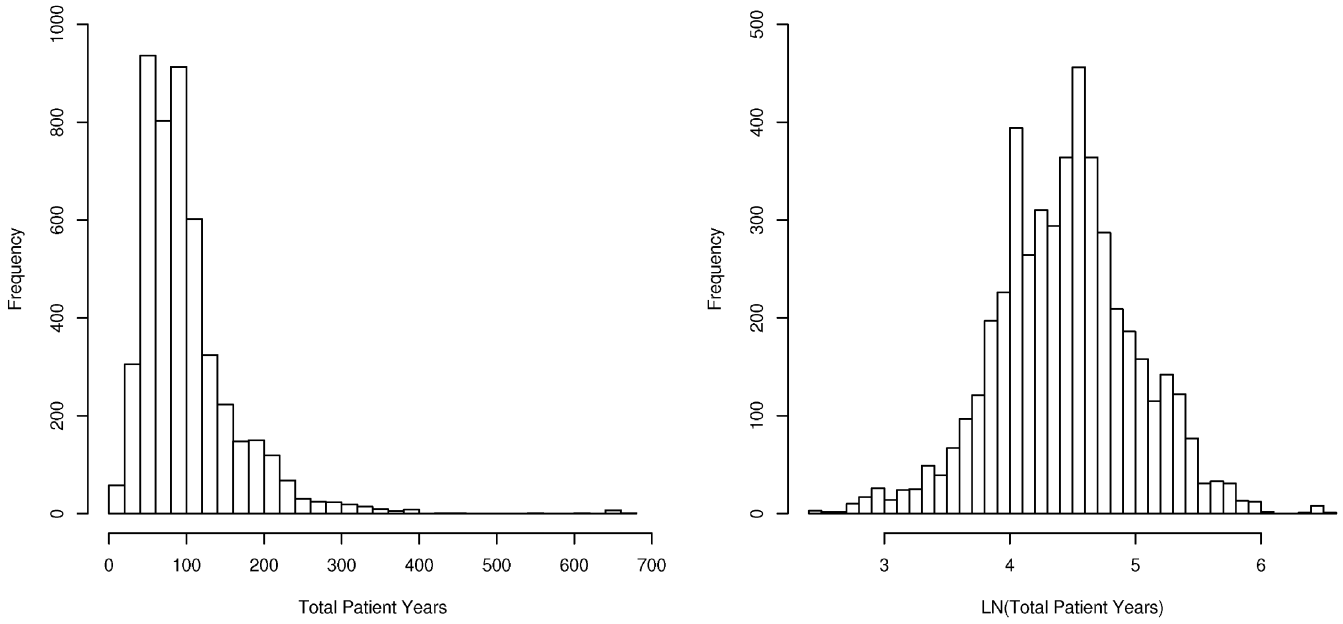


Table 3
Correlations among Continuous Covariates

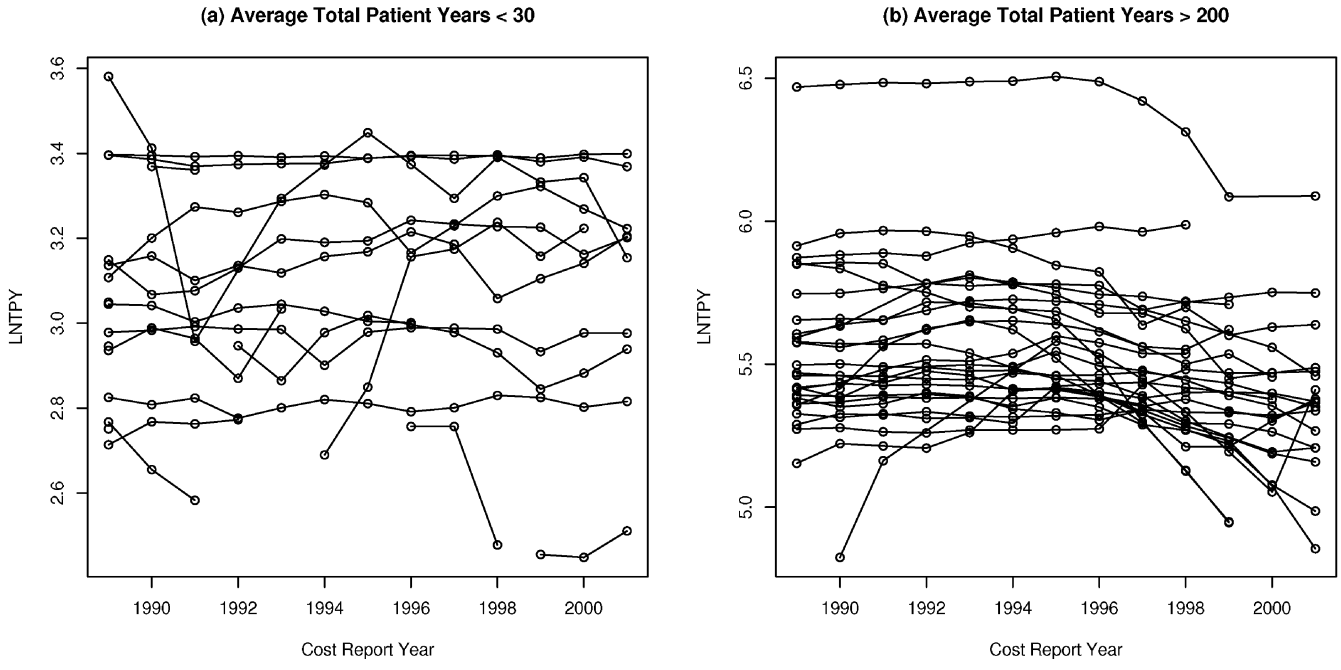
	LN(TPY)	TPY	NumBed	LN(NumBed)	SqrFoot	LN(SqrFoot)
LN(TPY)	1.000					
TPY	0.905	1.000				
NumBed	0.893	0.983	1.000			
LN(NumBed)	0.983	0.895	0.911	1.000		
SqrFoot	0.749	0.825	0.819	0.749	1.000	
LN(SqrFoot)	0.848	0.783	0.781	0.849	0.900	1.000

Motivated by this, we consider the natural logarithm of TPY, LN(TPY), as a potential outcome variable and then convert back to the original units for predicting TPY. Table 3 exhibits correlations among TPY, LN(TPY), and the continuous covariates. This table presents the natural logarithm of number of beds, LN(NumBed), and the natural logarithm of square footage, LN(SqrFoot), and suggests that covariates on a logarithmic scale are more closely related to LN(TPY) than those on the original scale. Therefore, we present our models in terms of continuous covariates on a logarithmic scale.

Knowledge of the size, organizational structure, and other variables will aid in making predictions of TPY. However, for predicting future values of a facility's TPY, the most important piece of information is past behavior. To underscore the importance of the dynamics, Figure 2 presents multiple times series plots for subsets of facilities. Here LN(TPY) is graphed versus year, with each line connecting annual observations of LN(TPY) for a facility. Only subsets of the facilities have been displayed in these graphs so that they do not appear too cluttered.

Figure 2 underscores the fact that prior values of LN(TPY) are important predictors of future values. Moreover, Figure 2 exhibits the so-called heterogeneity among facilities; observations within a facility

Figure 2
Multiple Time Series Plots of Logarithmic TPY



Notes: The left-hand panel (*a*) summarizes a subset of facilities with small average TPY; the right-hand panel (*b*) corresponds to facilities with large average TPY.

tend to have the same value compared to observations across facilities. Values of TPY are quite stable over time for many facilities; however, some facilities have a substantial amount of variability. We interpret these differences in variability as an aspect of heteroscedasticity. Finally, Figure 2 shows the unbalanced nature of our data; some facilities stopped reporting information within our sampling period, whereas other facilities entered.

3.2 In-Sample Model Fitting

The model specification criteria consist of two parts, in-sample selection criteria and out-of-sample assessment criteria. This section focuses on in-sample measures. Section 4 discusses out-of-sample measures.

We include the ordinary regression model as a baseline model to compare with the longitudinal models:

$$\text{LN(TPY)}_{it} = \beta_0 + \beta_1 \text{LN(NumBed)}_{it} + \text{S}_{it} + \varepsilon_{it}, \tag{3.1}$$

where

$$\begin{aligned}\text{S}_{it} = {} & \beta_2 \text{LN(SqrFoot)}_{it} + \beta_3 \text{Pro}_{it} + \beta_4 \text{TaxExempt}_{it} \\ & + \beta_5 \text{SelfFundIns}_{it} + \beta_6 \text{MCert}_{it} + \beta_7 \text{YEAR}_t + \beta_8 \text{YEAR}_t^2\end{aligned} \tag{3.2}$$

is a systematic component that is common to each model. This model is also known as a *pooled cross-sectional* (PCS) model in longitudinal data analysis in that it does not use historical facility-specific information to model the outcome variable.

As suggested by Figure 2, there are strong heterogeneities among the nursing facilities, and strong correlations within each facility over time. This suggests that the PCS model is not appropriate, and longitudinal models are in order. We follow the notation of Frees (2004); for nursing facility i in year t, the subject-specific models are of the form

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\alpha}_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \tag{3.3}$$

where $\mathbf{z}_{it} = (z_{it,1} \ldots z_{it,q})'$ and $\mathbf{x}_{it} = (x_{it,1} \ldots x_{it,K})'$ are nonstochastic covariates. The term $\boldsymbol{\alpha}_i$ is a vector of facility-specific parameters corresponding to $\mathbf{z}_{it}$, and $\boldsymbol{\beta}$ is a vector of parameters common to all facilities corresponding to $\mathbf{x}_{it}$. The vector of error terms is denoted by $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i11})'$. It has mean $(0, \ldots, 0)'$ and 11×11 temporal variance-covariance matrix $\mathbf{R}_i$, where the element in the rth row and sth column of $\mathbf{R}_i$ is $\text{Cov}(y_{ir}, y_{is})$. For facilities with fewer than 11 years of observations, the variance-covariance matrix can be obtained by removing the rows and columns of $\mathbf{R}_i$ corresponding to years that are not available.

To account for the heterogeneity among facilities, we use models with facility-specific parameters in the form of equation (3.3). To account for the serial correlation of the outcome variable TPY over time, we assume an autoregressive of order 1 (AR(1)) correlation structure. For an AR(1) correlation structure, we have $\mathbf{R}_{rs} = \sigma^2 \rho^{|r-s|}$. More complex structures than AR(1) can also be readily fitted; however, as we will see with our data set of only 11 years, AR(1) provides sufficient flexibility. The AR(1) correlation structure makes implicit use of the transition model approach to longitudinal modeling mentioned previously, in that past and current outcomes are assumed to be correlated.

Our first longitudinal representation is a *fixed effects* (FE) model with facility-specific intercepts that are fixed, unknown parameters. Compared to the PCS model in equation (3.1), this model separates out facility-specific effects that capture much of the time-constant information in the outcome variables; therefore, the estimates of the variability are more precise. Let α_{i0} denote the fixed facility-specific intercept to be estimated. The model is

$$\text{LN(TPY)}_{it} = \alpha_{i0} + \beta_1 \text{LN(NumBed)}_{it} + \text{S}_{it} + \varepsilon_{it}, \tag{3.4}$$

which we refer to as model FE1.

In addition to allowing intercepts to vary by facility, we can also vary the slope coefficient associated with LN(NumBed) by facility: that is, we anticipate that the effect of changes of number of beds on

TPY to be facility-specific. We refer to this as model FE2. Let α_{i1} denote the slope coefficient of LN(NumBed) for facility i and use

$$\text{LN(TPY)}_{it} = \alpha_{i0} + \alpha_{i1}\text{LN(NumBed)}_{it} + \text{S}_{it} + \varepsilon_{it}. \tag{3.5}$$

We also consider a situation in which $\boldsymbol{\alpha}_i$ is a vector of random variables instead of fixed, unknown parameters, known as ''random effects.'' Representations that include both fixed and random effects are known as *mixed effects* (ME) models; our first such model (model ME1) is given in equation (3.6). The form of the model is the same as the FE model with variable intercept in equation (3.4), only the intercept of the ith nursing facility is a function of the population intercept (β_0) plus a unique contribution from that facility, α_{i0}. We assume that α_{i0} are independent of ε_i and are identically and independently distributed with mean 0 and variance $\sigma^2_{\alpha 0}$:

$$\text{LN(TPY)}_{it} = (\beta_0 + \alpha_{i0}) + \beta_1\text{LN(NumBed)}_{it} + \text{S}_{it} + \mathbf{MSA}_i'\boldsymbol{\gamma} + \varepsilon_{it}. \tag{3.6}$$

In ME1 we also include the categorical variable MSA^i representing the metropolitan statistical area. As noted in Table 1, there are 14 different categories represented in this factor, so that 13 covariates ($\mathbf{MSA}_i = (\text{MSA}_{i1}, \ldots, \text{MSA}_{i13})'$) and their corresponding fixed regression coefficients ($\boldsymbol{\gamma} = (\beta_9, \ldots, \beta_{21})'$) are included in ME1. The intuition is that facilities in the same MSA share a similar economic climate that can be represented using a constant (within MSA) regression term.

In ME2 we allow the slope of LN(NumBed) to vary by facility i; that is, the slope for the ith facility is the sum of the population slope and a unique contribution from the facility, α_{i1}. As with intercepts, we assume α_{i1} are independent of the error terms and are identically and independently distributed with mean 0 and variance $\sigma^2_{\alpha 1}$. The covariance of α_{i0} and α_{i1} is denoted by $\sigma_{\alpha 01}$. Model ME2 is given by

$$\text{LN(TPY)}_{it} = (\beta_0 + \alpha_{i0}) + (\alpha_{i1} + \beta_1)\text{LN(NumBed)}_{it} + \text{S}_{it} + \mathbf{MSA}_i'\boldsymbol{\gamma} + \varepsilon_{it}. \tag{3.7}$$

Note that in this tutorial article, we include both FE and ME models for illustrative purposes. In practice, analysts often need to decide between these two types of longitudinal models. When there are time-constant covariates within subjects as in our data, such as MSA, ME models are preferred; coefficients for time-constant covariates cannot be estimated in FE models because they are collinear with the subject-specific intercept terms. The sampling method used to select subjects also can aid in the choice between FE and ME models. For example, if the subjects are randomly selected from a population, it is more reasonable to represent the subject-specific effect α_i as a random variable instead of as a fixed, unknown parameter. If, on the other hand, the subjects themselves constitute the population (for example, states in the United States), a FE model is more appropriate. When the choice is not clear, the Hausman test can be employed to choose between FE or ME models (Hausman 1978).

Several popular free or commercial software packages can be used to fit longitudinal models, such as R, Stata, and SAS. We use SAS PROC MIXED to fit both the FE and ME models; illustrative SAS code appears in Appendix B. The variance components are estimated using the restricted maximum likelihood (REML) method rather than the maximum likelihood method. Maximum likelihood estimators of the variance components are often biased and sometimes can be negative. REML produces unbiased, nonnegative estimators (Frees 2004, Chapter 3). Once we estimate the variance components, we can obtain the regression coefficient estimates for the FE or ME models. The estimated variance-covariance matrix of the regression parameters is computed by using the so-called sandwich estimator, which is asymptotically consistent and robust to unsuspected serial correlation and heteroscedasticity (Frees 2004, Chapter 3).

As in ordinary regression models, the statistical significance of individual regression coefficients is determined using t-statistics. Analysts can use penalized likelihood criterion such as Akaike's Information Criterion (AIC) or Schwarz's Bayesian Information Criterion (BIC) to compare alternative models (Brockett 1991; Burnham and Anderson 2004). Both statistics include a penalty that is an increasing function of the number of estimated parameters, but do not always lead to the same choice of models. AIC penalizes the number of free parameters less strongly than BIC. (AIC = -2 ln(maximum likelihood) + 2(number of estimable parameters); BIC = -2 ln(maximum likelihood) + ln(number of

subjects) × (number of estimable parameters).) A smaller value of AIC and BIC is associated with a better model fit.

Table 4 summarizes the coefficient estimates. The fit of the PCS model is much poorer than the other models, based on the large value of the AIC and BIC statistics. This result confirms our observation based on the multiple time series plots that the PCS model is not appropriate.

In the other models, the coefficients of the variables LN(SqrFoot), MCert, and YEAR are positively significant; Pro, TaxExempt, and $YEAR^2$ are negatively significant. Holding the other covariates constant, we interpret the positive coefficient for MCert to mean that Medicare-certified facilities tend to contribute more TPYs. The significant nonzero coefficients on YEAR and $YEAR^2$ indicate a quadratic trend. TPYs are increasing in the early years of the sample and declining in later years. To illustrate, for the FE1 model, the partial impact of going from year 11 ($0.013 \times 11 - 0.001 \times 11^2 = 0.022$) to year 12 ($0.013 \times 12 - 0.001 \times 12^2 = 0.012$) represents an estimated change of -0.010 of logarithmic TPY. One can interpret this as a 1.0% decline going from year 11 (1999) to year 12 (2000).

Comparing models FE1 and FE2, AIC indicates that FE2 provides a better fit, while BIC indicates that FE1 provides a better fit. We note that, for many of the facilities, the number of beds did not change over time. For these facilities, the number of beds is constant and hence perfectly collinear with the constant associated with the intercept term. Thus, slope coefficients are not estimable, and forecasting is not possible for these facilities. Therefore, we drop FE2 from consideration and use the simpler FE1.

Comparing models ME1 and ME2, both AIC and BIC indicate that ME2 provides a better fit. However, one can interpret the covariance terms between the random effects as a correlation; the estimated correlation between α_{i0} and α_{i1} is $-0.165/\sqrt{(0.730) \times (0.037)} = -0.998$. This indicates that the two random effects are highly negatively correlated, and that the slope coefficient is not informative. Therefore, we drop ME2 from consideration and use the simpler ME1.

In FE1 and ME1, the coefficients for both models are largely in agreement in that they provide similar information. Both indicate that our size measures, LN(NumBed) and LN(SqrFoot), are positively statistically significant; the coefficients of Medicare-certified and YEAR are also positively significant; the coefficients of TaxExempt and $YEAR^2$ are negatively significant; and the coefficient of SelfFundIns is not significant.

Table 4
Logarithmic TPY Model Coefficient Estimates Based on In-Sample Data

	Model PCS		Model FE1		Model FE2		Model ME1[a]		Model ME2*	
	Estimate	*t*-stat	Estimate	*t*-stat	Estimate	*t*-stat	Estimate	*t*-stat	Estimate	*t*-stat
Regression variables:										
LN(NumBed)	0.950	150.33	0.535	7.82			0.838	23.11	0.862	38.56
LN(SqrFoot)	0.040	7.05	0.059	2.82	0.027	1.65	0.068	3.25	0.052	3.21
Pro	0.021	3.46	−0.104	−2.02	−0.079	−2.39	−0.035	−2.10	−0.046	−2.26
TaxExempt	0.039	6.83	−0.091	−1.85	−0.081	−2.33	−0.033	−1.82	−0.040	−1.76
SelfFundIns	0.002	0.66	0.002	0.51	0.003	0.74	0.003	0.73	0.005	1.06
MCert	−0.018	−4.42	0.014	2.95	0.014	2.93	0.012	2.55	0.010	2.42
YEAR	0.015	6.93	0.013	4.78	0.009	4.61	0.010	4.43	0.009	4.59
$YEAR^2$	−0.001	−6.76	−0.001	−6.79	−0.001	−6.46	−0.001	−5.69	−0.001	−5.93
MSA			—	—	—	—				
Variance Components:										
Intercept variance ($\sigma^2_{\alpha 0}$)	—		—		—		0.005		0.730	
Slope variance ($\sigma^2_{\alpha 1}$)	—		—		—		—		0.037	
Intercept-slope covariance ($\sigma_{\alpha 01}$)	—		—		—		—		−0.165	
Disturbance variance (σ^2)	0.010		0.012		0.003		0.008		0.008	
Correlation (ρ)	—		0.838	38.00	0.460	22.98	0.740	38.00	0.778	34.52
Goodness-of-fit statistics:										
−2 residual log-likelihood	−6845.3		−9721.1		−12,440.0		−9707.3		−10,479.4	
AIC	−6797.3		−8905.1		−9,952.0		−9651.3		−10,415.4	
BIC	−6705.6		−7286.6		−6,754.9		−9551.6		−10,307.8	

* ME models were estimated including the categorical factor metropolitan statistical area (MSA), although the coefficients are not reported.

4. MODEL VALIDATION

We compare the predictive abilities of our candidate models using the out-of-sample data from the 12th and 13th years, 2000 and 2001. There are 717 observations from 366 nursing facilities in these years. An additional four facilities started after 1999; they are not included in the out-of-sample data.

For each model we used the in-sample data to estimate coefficients and used the estimated coefficients and covariates for the out-of-sample data to compute predicted LN(TPY) (Frees 2004 Chapter 4). Then we exponentiated to get $\widehat{TPY}_{i,11+L}$ for each facility i and forecast leads $L = 1, 2$ (corresponding to $t = 12, 13$, or years 2000, 2001).

We summarize the model accuracy of the forecasts through two statistics, the Mean Absolute Error (MAE),

$$\text{MAE} = \frac{1}{n_{12} + n_{13}} \sum_{L=1}^{2} \sum_{i=1}^{n_{11+L}} |\widehat{TPY}_{i,11+L} - TPY_{i,11+L}|, \tag{4.1}$$

and the Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \frac{100}{n_{12} + n_{13}} \sum_{L=1}^{2} \sum_{i=1}^{n_{11+L}} \frac{|\widehat{TPY}_{i,11+L} - TPY_{i,11+L}|}{TPY_{i,11+L}}. \tag{4.2}$$

Here n_t denotes the number of facilities in year t.

Table 5 summarizes the comparison of the models based on the out-of-sample criteria. Table 5 also summarizes forecasts based on the in-sample average of each facility's TPY: this naive estimator performed poorly. However, an alternative naive estimator, the most recent observation, performed even better than the PCS model on the out-of-sample fit, without any adjustments for trend. The most recent observation is included in our comparisons as a baseline measure; analysts should always consider this naive measure. Although not reported in detail here, this estimator performed poorly in the in-sample assessment. Moreover, the performance of the most recent observation deteriorates rapidly as the forecast horizon (L) increases.

As expected, the PCS model performed poorly compared to the other candidate models. Comparing FE1 and ME1, we see that both models performed better than the most recent observation, with the slight edge to ME1.

To underscore the appropriateness of the natural logarithm transformations, we also fit the PCS, FE1, and ME1 models using the original units of the continuous variables TPY, NumBed, and SqrFoot. The forecast results are reported in Table 5. Compared to the models with LN(TPY) as the outcome variable, the FE1 and ME1 models for TPY have larger MAE and MAPE values, indicating that the LN(TPY) models fit the data better.

Table 5
Using 1989–1999 In-Sample Data to Predict 2000–2001 Out-of-Sample Data

Model	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)
Most recent observation	4.62	5.47
Outcome variable LNTPY:		
In-sample average	12.12	13.36
PCS	6.05	7.16
FE1	4.51	5.28
ME1	4.18	4.84
Outcome variable TPY:		
PCS	5.96	7.71
FE1	5.34	7.02
ME1	4.45	5.51

When developing forecasting models, it is prudent to test the ability of the model to predict over a range of time periods to establish robustness. To assess model performance at different points in time, Table 6 reports the forecast results on the logarithm transformed TPY of another in-sample period, beginning in 1989 and ending in 1997. Similar to the results in Table 5, both the FE1 and ME1 models outperform the other candidates. Interestingly, the PCS model forecasts better than the most recent observation. Because the most recent observation is based on only observation per facility, it is highly variable. Thus, it is not surprising that its forecasting performance fluctuates wildly over different time periods.

5. SUMMARY AND CONCLUDING REMARKS

This article illustrated predictive modeling using State of Wisconsin nursing home data. We emphasized that predictive modeling is a process that involves problem identification, data analysis, and candidate model development, estimation, and validation. Our predictive modeling approach involved longitudinal modeling and compared three types of models, the PCS, FE, and ME models, with two simple approaches, the in-sample average and most recent observation. Introducing covariates and allowing for either a fixed or random intercept term by facility improved the prediction, compared to the simpler approaches or an ordinary regression model. Although not surprising, this result is significant given the common industry practice of using cross-sectional algorithms for predictive modeling. These models are easily computed with appropriate software packages such as Stata, R, or in our case, SAS.

As mentioned in the Introduction, predictive modeling has been used for provider profiling, provider reimbursement, and identification of high-cost users. In all of these applications, costs can and should be linked over time, whether they be by physician, by organization, or by individual. Longitudinal modeling of costs over time accounts for the heterogeneity of individuals, through inclusion of individual-specific intercept and slope coefficients.

The longitudinal data approach used in this article is only one of several predictive modeling approaches. Other approaches include continuance tables, multiple regression analysis, generalized linear models (GLMs), two-part models, Bayesian analysis, finite mixtures of distributions, and statistical algorithms such as clustering, principal components, classification and regression trees, multivariate adaptive regression splines, and neural networks.

Continuance tables traditionally have been used in the health insurance industry for the predictive modeling of frequency distributions for outcomes such as length of stay or claim duration. For example, our Wisconsin nursing home data could have been analyzed by developing a continuance table for total patient days or years, using historical data and actuarial assumptions to estimate the probability that a resident would spend N days in a nursing home, where N is random count variable. Such an approach could be useful for analyzing discrete outcome variables. However, continuance tables cannot be easily generalized to continuous outcome variables, such as the expected cost of treating a nursing home resident or claim severities. There are many articles in the literature that discuss continuance tables; some references are provided in Waters and Phil (1989).

Multiple regression analysis is one of the most widely used predictive modeling techniques in health care. Some widely used multiple regression predictive models for determining provider reimbursement

Table 6
Using 1989–1997 In-Sample Data to Predict 1998–1999 Out-of-Sample Data

Model	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)
In-sample average	11.64	12.27
Most recent observation	6.72	7.25
PCS	5.62	6.35
FE1	4.85	5.31
ME1	4.86	5.16

are the Diagnostic Cost Group Hierarchical Condition Category models (DCG/HCC) (Ash et al. 2000). These models incorporate patient diagnosis codes as covariates to aid in the prediction of provider payments. Other papers contain discussions of health-care-related multiple regression models, including Ash and Byrne-Logan (1998), Fowles et al. (1996), Pope et al. (2000), Pope et al. (2004), Sales et al. (2003), and Zhao et al. (2005).

The GLM approach is a popular method for modeling skewed data. GLMs are flexible in that the link and variance functions used in this method can account for data issues such as overdispersion (Frees 2004, Chapter 10). Papers that have discussed GLMs with regard to predicting health care costs and utilization include Blough, Madden, and Hornbrook (1999), Diehr et al. (1999), Daniels and Gatsonis (1999), and Manning and Mullahy (2001).

Two-part models estimate the utilization of health care separately from the cost. The first component models the likelihood that a patient will use any medical services using either a logistic or probit regression model. The second component models the amount of medical services utilized by the patient, conditional on the patient having used any services. If multiple regression analysis is used, then the outcome variable is typically the logarithm of health care cost, whereas if a GLM is used to model this component, then no transformation is needed. The product of the expected value of each component provides the expected health care costs for an individual (Blough et al. 1999; Deb and Trivedi 2002; Diehr et al. 1999).

Other possible approaches to model health care costs include Bayesian predictive modeling (De Alba 2002; Fellingham, Tolley, and Herzog 2005; Verrall 2004) or the use of finite mixture distributions (Deb and Burgess 2003; Deb and Holmes 2000; Deb and Trivedi 1997). Clustering, principal components, classification and regression trees, multivariate adaptive regression splines, and neural networks are more complex models that can be readily implemented using specialized algorithms embedded in statistical computer software. These techniques allow for high flexibility in the specification of the regression function and easily allow for complex transformations of and interactions between covariates (Berry and Linoff 2004; Hastie, Tibshirani, and Friedman 2001).

Health care modeling using nonnormal distributions can be effective in reducing prediction error. In the future we will extend this work to examine the data using nonnormal distributions. Also, the predictive power of models determined by the predictive modeling process depends on the quality of the data used to generate the models. Analysts often face problems with missing data, which can arise in longitudinal data as attrition, where subjects that contributed at least one subject-time observation fail to provide observations in other time periods. Depending on the type of nonresponse, parameters of predictive models may be improperly estimated because of potential selection bias. A related concern pertains to potential endogeneity of model covariates. For example, the analyst may be unable to include all covariates that are related to the outcome variable in the theoretical model because of data limitations; this is commonly known in the statistical literature as omitted variable bias, which can also cause model parameters to be biased. Future work will develop methods to allow analysts to account for these potential issues in predictive modeling.

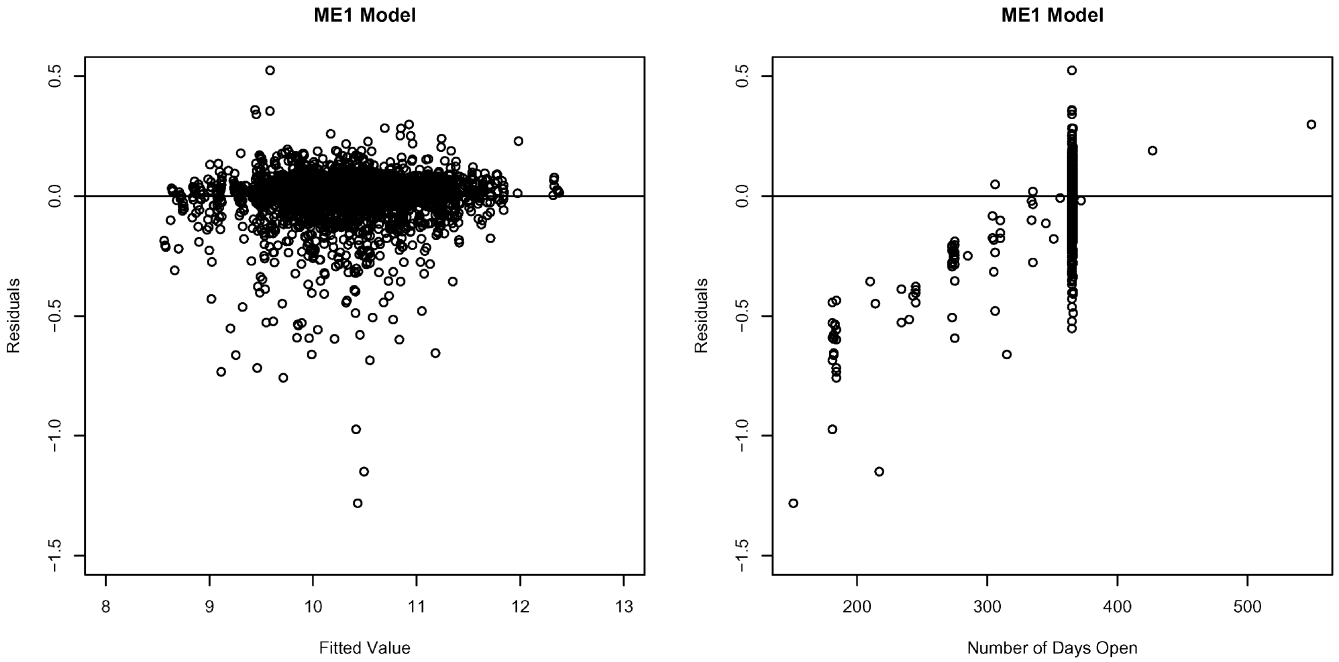
APPENDIX A

DIAGNOSTIC ANALYSES

As with all statistical modeling exercises, the models introduced in Section 3 were the result of an intense study of the data and the literature supporting health care utilization. Although this article focuses on the steps ending with the in-sample estimation and the out-of-sample validation, in any data analysis there are many complex decisions that need to be made prior to recommendation of these candidate models.

In regression and longitudinal data analysis, the process of examining a preliminary model fit and using information about any lack of fit to improve the model specification is known as ''diagnostic analysis.'' Residual analysis is an important type of diagnostic analysis. Residuals are defined as the

Figure 3
Residual Analysis of the Mixed Effects Model (ME1) for Total Patient Days



observed value minus the fitted value, and the analysis involves both the numeric and graphical inspection of the estimated models (Frees 2004).

To illustrate this process, it is interesting to consider the outcome variable total patient years (TPY). When first examining the data, policymakers most interested in the cost reports worked with the total of patient days that a nursing home facility utilized in a cost report period without adjustments to the length of the period. For this reason, we originally used LN(Total Patient Days) as the outcome variable without adjustment of the number of facility operating days in the cost reporting period (DaysOpen). We had already determined that this variable, like TPY, was quite skewed and required a natural logarithmic transform. To get a sense of the data, about 2% of observations were not operating for the whole cost report year. Four observations' cost report periods were longer than one calendar year; this can occur if a facility changes its fiscal year or some other significant change in a facility's characteristics occurs, such as a change in ownership.

Residual analysis uncovered some important patterns that we had not adequately addressed in our models LN(Total Patient Days) as the outcome variable. Figure 3 shows the residual analysis of the mixed effects (ME1) model, using the same covariates as reported in Table 4. (The results of the pooled cross-sectional (PCS) and fixed effects (FE1) models are similar, thus not shown here.) The left panel in Figure 3 exhibits the residuals versus the fitted value of the ME1 model, and the right panel shows the relationship between the residuals and length of cost report period (DaysOpen). It is clear that the residuals decrease as the number of facility operating days decreases, indicated by a positive relation between these variables. One possibility is to include DaysOpen as a covariate in the regression. However, we chose to define a new variable, TPY, as number of total patient days in the cost reporting period divided by number of facility operating days in the cost report period to take the length of cost report period into consideration. This selection also accounts for some heteroscedasticity that is not evident in Figure 3.

APPENDIX B

ILLUSTRATIVE SAS CODE FOR MODEL ESTIMATION

```
proc format;
  value yesno 0=''1:no'' 1=''0:yes'';
run;

proc mixed data=INSAMPLE noclprint order=formatted;
  title 'Pooled Cross-Sectional Model (Model PCS)';
  class POPID MSA;
  model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
  /s outp=cs noint;
  format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
  title 'Fixed Variable Intercept Model (Model FE1)';
  class POPID;
  model LNTPY=POPID LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
  /s noint outp=fe1;
  repeated /type=ar(1) sub=POPID r;
  format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
  title 'Fixed Variable Intercept and Slope Model (Model FE2)';
  class POPID;
  model LNTPY=POPID POPID*LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
  /s noint outp=fe2;
  repeated /type=ar(1) sub=POPID r rcorr;
  format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
  title 'Error Components Model (Model ME1)';
  class POPID MSA;
  model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
  /s outp=me1 noint;
  random intercept /sub=POPID g;
  repeated /type =ar(1) sub=POPID r;
  format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
  title 'Random Variable Intercept and Slope Model (Model ME2)';
  class POPID MSA;
  model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
  /s outp=me2 noint;
  random intercept LNNumBed/ sub=POPID type=UN g;
  repeated / type=ar(1) sub=POPID r;
  format PRO TaxExempt SelfFundIns MCert yesno.;
run;
```

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation, Grant Number SES-0436274, and the Agency for Healthcare Research and Quality, Grant Number R03 HS16519, and the Assurant Health Professorship in Actuarial Science.

REFERENCES

AMERICAN ACADEMY OF ACTUARIES RISK CLASSIFICATION SUBCOMMITTEE OF THE PROPERTY/CASUALTY PRODUCTS, PRICING, AND MARKET COMMITTEE. 2002. The Use of Credit History for Personal Lines of Insurance: Report to the National Association of Insurance Commissioners. *American Academy of Actuaries: Reports to the NAIC,* November 15.

ASH, ARLENE S., AND SUSAN BYRNE-LOGAN. 1998. How Well Do Models Work? Predicting Health Care Costs. *Proceedings of the Section on Statistics in Epidemiology, American Statistical Association,* pp. 42–49.

ASH, ARLENE S., RANDALL P. ELLIS, GREGORY C. POPE, JOHN Z. AYANIAN, DAVID W. BATES, HELEN BURSTIN, LISA I. IEZZONI, ELIZABETH MACKAY, AND WEI YU. 2000. Using Diagnoses to Describe Populations and Predict Costs. *Health Care Financing Review* 21(3): 7–28.

BALTAGI, BADI H. 2005. *Econometric Analysis of Panel Data.* 3rd edition. New York: Wiley.

BERRY, MICHAEL J. A., AND GORDON S. LINOFF. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* New York: Wiley.

BIDDLE, JEFF, AND KAREN ROBERTS. 2003. Claiming Behavior in Workers' Compensation. *Journal of Risk and Insurance* 70(4): 759–80.

BLOUGH, DAVID K., CAROLYN W. MADDEN, AND MARK C. HORNBROOK. 1999. Modeling Risk Using Generalized Linear Models. *Journal of Health Economics* 18: 153–71.

BROCKETT, PATRICK L. 1991. Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications. *Transactions of the Society of Actuaries* 43: 73–135.

BROCKETT, PATRICK L., RICHARD A. DERRIG, LINDA L. GOLDEN, ARNOLD LEVINE, AND MARK ALPERT. 2002. Fraud Classification Using Principal Component Analysis of RIDITs. *Journal of Risk and Insurance* 69(3): 341–71.

BURNHAM, KENNETH P., AND DAVID R. ANDERSON. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Technical report.

CHRISTIANSEN, CINDY L., AND CARL N. MORRIS. 1997. Improving the Statistical Approach to Health Care Provider Profiling. *Annals of Internal Medicine* 127(8), Part 2: 764–68.

COOIL, BRUCE. 1991. Using Medical Malpractice Data to Predict the Frequency of Claims: A Study of Poisson Process Models with Random Effects. *Journal of the American Statistical Association* 86(414): 285–95.

COUSINS, MICHAEL S., LISA M. SHICKLE, AND JOHN A. BANDER. 2002. An Introduction to Predictive Modeling for Disease Management Risk Stratification. *Disease Management* 5(3): 157–67.

CUMMING, ROBERT B., AND BRIAN A. CAMERON. 2002. A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations. Research study sponsored for the Society of Actuaries by Milliman, Inc. www.symmetry-health.com/SOAStudy.pdf.

CUMMING, ROBERT B., IAN G. DUNCAN, ELIZABETH V. LEWIS, AND MARK D. WERNICKE. 2002. SOA Session 99OF: Predictive Modeling. Record of the Society of Actuaries.

DANIELS, MICHAEL J., AND CONSTANTINE GATSONIS. 1999. Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization. *Journal of the American Statistical Association* 94(445): 29–42.

DE ALBA, ENRIQUE. 2002. Bayesian Estimation of Outstanding Claim Reserves. *North American Actuarial Journal* 6(4): 1–20.

DEB, PARTHA, AND JAMES. F. BURGESS, JR. 2003. A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures. Working paper. Hunter College Department of Economics Working Papers 212, Hunter College Department of Economics.

DEB, PARTHA, AND ANN M. HOLMES. 2000. Estimates of Use and Costs of Behavioural Health Care: A Comparison of Standard and Finite Mixture Models. *Health Economics* 9(6): 475–89.

DEB, PARTHA, AND PRAVIN K. TRIVEDI. 1997. Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics* 12: 313–36.

———. 2002. The Structure of Demand for Health Care: Latent Class versus Two-Part Models. *Journal of Health Economics* 21: 601–25.

DELONG, ELIZABETH R., ERIC. D. PETERSON, DAVID M. DELONG, LAWRENCE H. MUHLBAIER, SUZANNE HACKETT, AND DANIEL B. MARK. 1997. Comparing Risk-Adjustment Methods for Provider Profiling. *Statistics in Medicine* 16(23): 2645–64.

DERRIG, RICHARD A. 2002. Insurance Fraud. *Journal of Risk and Insurance* 69(3): 271–87.

DIEHR, P., D. YANEZ, A. ASH, M. HORNBROOK, AND D. Y. LIN. 1999. Methods for Analyzing Health Care Utilization and Costs. *Annual Review of Public Health* 20: 125–44.

DIGGLE, PETER, PATRICK HEAGERTY, KUNG-YEE LIANG, AND SCOTT L. ZEGER. 2002. *Analysis of Longitudinal Data.* 2nd edition. Oxford: Oxford University Press.

DOVE, HENRY G., IAN DUNCAN, AND ARTHUR ROBB. 2003. A Prediction Model for Targeting Low-Cost, High-Risk Members of Managed Care Organizations. *American Journal of Managed Care* 9(5): 381–89.

ELLIS, RANDALL J., MARILYN S. KRAMER, JOSEPH F. ROMANO, AND RONG YI. 2003. Applying Diagnosis-Based Predictive Models to Group Underwriting. *Health Section News* 46(1): 1, 4–7.

FELLINGHAM, GILBERT W., H. DENNIS TOLLEY, AND THOMAS N. HERZOG. 2005. Comparing Credibility Estimates of Health Insurance Claims Costs. *North American Actuarial Journal* 9(1): 1–12.

FOWLES, JINNET B., JONATHAN P. WEINER, DAVID KNUTSON, ELIZABETH FOWLER, ANTHONY M. TUCKER, AND MARJORIE IRELAND. 1996. Taking Health Status into Account When Setting Capitation Rates: A Comparison of Risk-Adjustment Methods. *Journal of the American Medical Association* 276(16): 1316–21.

FREES, EDWARD W. 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

GUSZCZA, JAMES, AND JAN LOMMELE. 2006. Loss Reserving Using Claim-Level Data. *Casualty Actuarial Forum*.

HASTIE, TREVOR, ROBERT TIBSHIRANI, AND JEROME FRIEDMAN. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

HAUSMAN, J. A. 1978. Specification Tests in Econometrics (STMA V21 827). *Econometrica* 46: 1251–71.

HU, GUIZHOU, AND ERIK LESNESKI. 2004. The Differences between Claim-Based Health Risk Adjustment Models and Cost Prediction Models. *Disease Management* 7(2): 153–58.

IEZZONI, LISA I. 1997. *Risk Adjustment for Measuring Healthcare Outcomes*. 2nd edition. Chicago: Health Administration Press.

KRONICK, RICHARD, TODD GILMER, TONY DREYFUS, AND LORA LEE. 2000. Improving Health-Based Payment for Medicaid Beneficiaries: CDPS. *Health Care Financing Review* 21(3): 29–64.

MANNING, WILLARD G., AND JOHN MULLAHY. 2001. Estimating Log Models: To Transform or Not to Transform? *Journal of Health Economics* 20(4): 461–94.

MEENAN, RICHARD T., MAUREEN C. O'KEEFFE-ROSETTI, MARK C. HORNBROOK, DONALD J. BACHMAN, MICHAEL J. GOODMAN, PAUL A. FISHMAN, AND ARNOLD V. HURTADO. 1999. The Sensitivity and Specificity of Forecasting High-Cost Users of Medical Care. *Medical Care* 37(8): 815–23.

MONAGHAN, JAMES E. 2000. The Impact of Personal Credit History on Loss Performance in Personal Lines. Working paper. Casualty Actuarial Society Forum.

PASSWATER, KEITH, AND BRENT SEILER. 2004. Predictive Modeling: Considerations for Care Management Applications. *Health Section News* 47: 13–15.

POPE, GREGORY C., RANDALL P. ELLIS, ARLENE S. ASH, JOHN Z. AYANIAN, DAVID W. BATES, HELEN BURSTIN, LISA I. IEZZONI, EDWARD MARCANTONIO, AND BEI WU. 2000. Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment: Final Report. Health Care Financing Administration Report, December 21.

POPE, GREGORY C., JOHN KAUTTER, RANDALL P. ELLIS, ARLENE S. ASH, JOHN Z. AYANIAN, LISA I. IEZZONI, MELVIN J. INGBER, JESSE M. LEVY, AND JOHN ROBST. 2004. Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model. *Health Care Financing Review* 25(4): 119–41.

ROSENBERG, MARJORIE A., AND PAUL H. JOHNSON, JR. 2007. Health Care Predictive Modeling Tools. *Health Watch* 54: 24–27.

SALES, ANNE E., CHUAN-FEN LIU, KEVIN L. SLOAN, JESSE MALKIN, PAUL A. FISHMAN, AMY K. ROSEN, SUSAN LOVELAND, W. PAUL NICHOL, NORMAN T. SUZUKI, EDWARD PERRIN, NANCY D. SHARP, AND JEFFREY TODD-STENBERG. 2003. Predicting Costs of Care Using a Pharmacy-Based Measure Risk Adjustment in a Veteran Population. *Medical Care* 41(6): 753–60.

SPEIGHTS, DAVID B., JOEL B. BRODSKY, AND DARYA L. CHUDOVA. 1999. Using Neural Networks to Predict Claim Duration in the Presence of Right Censoring and Covariates. *Casualty Actuarial Forum*.

TENNYSON, SHARON, AND PAU SLASAS-FORN. 2002. Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives. *Journal of Risk and Insurance* 69(3): 289–308.

VERRALL, R. J. 2004. A Bayesian Generalized Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. *North American Actuarial Journal* 8(3): 67–89.

VIAENE, STIJN, RICHARD A. DERRIG, BART BAESENS, AND GUIDO DEDENE. 2002. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance* 69(3): 373–421.

WATERS, H. R., AND D. PHIL. 1989. Some Aspects of the Modelling of Permanent Health Insurance. Presented at the seminar ''Applications of Mathematics in Insurance, Finance, and Actuarial Work'' sponsored by the Institute of Mathematics and Its Applications, the Institute of Actuaries, and the Faculty of Actuaries, held at the Institute of Actuaries, July 6–7.

WEYCKER, DEREK A., AND GAIL A. JENSEN. 2000. Medical Malpractice among Physicians: Who Will Be Sued and Who Will Pay? *Health Care Management Science* 3: 269–77.

WU, CHENG-SHENG PETER, AND JAMES C. GUSZCZA. 2003. Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View. *Casualty Actuarial Society Forum*.

ZHAO, YANG, ARLENE S. ASH, RANDALL P. ELLIS, JOHN Z. AYANIAN, GREGORY C. POPE, BRUCE BOWEN, AND LORI WEYUKER. 2005. Predicting Pharmacy Costs and Other Medical Costs Using Diagnoses and Drug Claims. *Medical Care* 43(1): 34–43.

ZHAO, YANG, ARLENE S. ASH, JOHN HAUGHTON, AND BENJAMIN MCMILLAN. 2003. Identifying Future High-Cost Cases through Predictive Modeling. *Disease Management and Health Outcomes* 11(6): 389–97.

Discussions on this paper can be submitted until January 1, 2008. 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.