Towards Multivariate Ratemaking: Claim Frequency Analysis Examples

Hernán L. Medina, CPCU, API, AU, AIM, ARC

Abstract

Motivation. Test how changes in level and distribution of exposures affect different ratemaking models. Actuaries are well aware that loss trend can be distorted by changes in exposure level and business mix. They are trained to recognize situations in which these distortions may arise, and how to adjust for them. Multivariate models are another way of handling these distortions. Using claim frequency as an example, the paper illustrates the design of multivariate analyses resistant to changes in exposure level and business mix.

Method. Simulate data in which the predominant sources of variation are changing exposure levels and changes in the distribution of exposures. Determine indicated trend, development, and classification factors using multivariate and univariate models. Compare the results.

Results. Trend, development factors, and relativity indications from 30 samples having different levels of variation in exposure levels and distribution are obtained by different methods.

Conclusions. Multivariate analyses that incorporate all available information are more robust than other analyses when data have significant changes in exposure levels or changes in mix of business.

Availability.

Input data sets and model outputs are available at www.casact.org.

Keywords. Ratemaking, Trend and Loss Development, Rating Class Relativities, Generalized Linear Models

1. INTRODUCTION

Actuaries began to develop the art and science of property and casualty insurance ratemaking long before computers were invented. At a time when calculations were done with pencil and paper, it was natural to use methods that relied on total sums and averages. When computers were first introduced, storage media were very expensive and processing speeds were relatively slow by today's standards. Thus ratemaking databases were designed to contain totals and averages, and rating systems continued to rely for the most part on univariate analyses based on aggregate data.

Actuaries are well aware of the pitfalls one might encounter using methods that rely on aggregate data. Part of actuarial training is learning to recognize the distortions that might arise, and how these might be corrected or minimized. For example, the CAS' *Basic Ratemaking* textbook indicates that if calendar year data is used to measure loss trend and the book of business is changing significantly in size, the trend can be over or underestimated.[1] An illustration of this situation follows in Table 1.1.

[1] Werner, Geoff and Claudine Modlin, *Basic Ratemaking,* 3rd ed., Arlington, VA: Casualty Actuarial Society, January 2010, p. 113.

Table 1.1

Calendar Year	Earned Car Years	Calendar Year Claims Closed With Payment	Claim Frequency	Years	Trend
2004	198,017	12,504	6.31461	6	3.0%
2005	215,837	13,770	6.37981	5	3.4%
2006	232,026	14,972	6.45273	4	3.8%
2007	225,064	15,304	6.79984	3	3.1%
2008	211,559	14,928	7.05619		
2009	192,520	13,911	7.22574		

	Claims Closed With Payment				Development Factors			
Accident Year	Age 12	Age 24	Age 36	Age 48	12 to 24	24 to 36	36 to 48	Ultimate Claims
2002	5,435	8,696	10,870	10,870	1.600	1.250	1.000	10,870
2003	6,007	9,611	12,013	12,013	1.600	1.250	1.000	12,013
2004	6,726	10,762	13,452	13,452	1.600	1.250	1.000	13,452
2005	7,332	11,733	14,665	14,665	1.600	1.250	1.000	14,665
2006	7,881	12,609	15,764	15,764	1.600	1.250	1.000	15,764
2007	7,644	12,230	15,288	15,288	1.600	1.250		15,288
2008	7,187	11,500			1.600			14,375
2009	6,540							13,080
Selected Age to Age					1.600	1.250	1.000	
Selected Age to Ultimate					2.000	1.250	1.000	

Accident Year	Earned Car Years	Ultimate Accident Year Claim Count	Claim Frequency	Years	Trend
2004	198,017	13,452	6.79336	6	0.0%
2005	215,837	14,665	6.79448	5	0.0%
2006	232,026	15,764	6.79407	4	0.0%
2007	225,064	15,288	6.79273	3	0.0%
2008	211,559	14,375	6.79479		
2009	192,520	13,080	6.79410		

In the example above, development factors are constant across accident year, so we can be reasonably certain of the estimated ultimate claim counts and the trend based on accident year data. Therefore we can conclude the trend based on calendar year data is overstated. In a real-world

situation, however, development factors may be more volatile, and the selection of loss development factors, "introduces some subjectivity into the [accident year] trend analysis."[2]

The CAS' *Basic Ratemaking* electronic textbook explains that the reason for the distortion in the calendar year trend is that as exposure levels change, the distribution of calendar year claims by accident year changes.[3] In fact, the effect of exposure level changes on calendar year trend is a special case of a more general phenomenon: the effect of changes in business mix on frequency and severity, which can affect both calendar year trend as well as accident year trend. Consider the example in Table 1.2.

Table 1.2

Area	Accident year	Earned Car Years	Ultimate Claims With Payment	Claim Frequency	Years	Trend
Territory A	2004	220,500	18,820	8.53515	6	3.0%
	2005	231,527	20,353	8.79077	5	3.0%
	2006	243,100	22,011	9.05430	4	3.0%
	2007	255,256	23,803	9.32515	3	3.0%
	2008	268,019	25,745	9.60566		
	2009	281,420	27,844	9.89411		
Territory B	2004	179,500	5,105	2.84401	6	3.0%
	2005	168,476	4,936	2.92979	5	3.0%
	2006	156,900	4,735	3.01785	4	3.0%
	2007	144,744	4,501	3.10963	3	3.0%
	2008	131,981	4,225	3.20122		
	2009	118,580	3,911	3.29820		
Statewide	2004	400,000	23,925	5.98125	6	5.8%
	2005	400,003	25,289	6.32220	5	5.9%
	2006	400,000	26,746	6.68650	4	5.9%
	2007	400,000	28,304	7.07600	3	5.9%
	2008	400,000	29,970	7.49250		
	2009	400,000	31,755	7.93875		

[2] Werner and Modlin, p. 113.
[3] Werner and Modlin, *Basic Ratemaking*, p. 113.

In the example above in Table 1.2, each territory has a 3% trend, but the statewide data shows a trend that is almost twice as high, close to 6%. The reason for this is that the distribution of exposures in the state has been changing. "Distributional changes in a book of business also affect frequencies and severities. If the proportion of risky policies is growing, loss costs will be expected to increase."[4]

Although the issues above are well known, they are generally handled on an ad hoc basis, and not much has changed in the basic rate review process. Generally, it involves two major steps: (1) determination of the overall indicated rate level change, and (2) determination of indicated classification relativities. Loss development and trend are two of the processes involved in determining the overall indicated rate level change. Thus, a basic rate review often involves at least three databases and systems: (1) loss development database and system, (2) loss trend database and system, and (3) and classification review database and system.

When accident year trends are used in the rate review, the loss development and loss trend processes are intertwined. For example, determining the accident year claim frequency trend typically involves the following steps: developing claim counts to ultimate, calculating ultimate claim frequency for each accident year, and analyzing the trend using a linear or exponential regression model. So the same database could be used for loss development and trend for rate reviews using accident year trend. In most cases, however, the database has been summarized in such a way that it cannot be used to review classification relativities.

From a data management perspective, as well as a business point of view, it is desirable to have a single database as the source for the analyses involved in the rate review process. This helps simplify data quality reviews and helps ensure that the data used in the different analyses balances. This could easily be accomplished. Appendix G of *A Practitioner's Guide to Generalized Linear Models*[5] presents several forms of data organization that can be used for generalized linear model (GLM) analysis, as well as their advantages and disadvantages. Using personal auto property damage liability as an example, we will expand one of those forms of data organization into a database that can be used as the source for development, accident year trend, and indicated classification relativity analyses. Furthermore, we will see how to integrate all of these processes into one single model using generalized linear models (GLM) and generalized estimating equations (GEE).

---

[4] Werner and Modlin, p. 109.

[5] Anderson, D., et al., "A Practitioner's Guide to Generalized Linear Models," 3rd Ed., CAS Study Note, 2007.

Using a single database for loss trend, loss development, and risk classification requires thoughtful consideration. A company may have some exclusion or adjustments currently used for trend analyses that are not used for loss development or classification analyses, and so on. In a multivariate model, however, you must consider whether it is preferable to adjust the data a priori, or to introduce variables that would control for the factor that would have made an adjustment or exclusion necessary in a univariate analysis. For example, certain vehicle models were recalled in 2010 because of problems involving sudden uncontrollable acceleration.[6] If a large number of such claims are in the data, one option would be to exclude them from the analysis. Another option would be to leave them in the data and add a control variable to identify these claims in the multivariate model. The coefficient of the control variable would provide the actuary with a way to estimate the effect this unusual event had on the experience. The control variable would be equal to 1 for claims related to the recalled vehicles, and 0 for all other vehicles. If all affected vehicles have been recalled and repaired and no further losses related to this event are expected in the future, then the control variable is set to zero when the model is used to project expected claim counts or losses.

Differences in exclusions or adjustments may arise because different types of data are used for different types of analyses. For example, it is quite common for companies to use calendar year paid claim data for trend analysis, and accident year reported claim data for loss development in personal auto property damage liability rate reviews. Presumably, since different types of data are used for the univariate analyses of trend and loss development, some situation might arise that would make it necessary to adjust the trend data while the loss development data needs no adjustment (or vice versa). If this situation arises in a multivariate context in which loss trend, development and classification factors are estimated simultaneously, an adjustment or control variable would be needed for a model based on paid claim data, but no adjustment would be needed for a model based on reported claim data (or vice versa). As will be shown later, the database can be designed in such a way that it contains both paid and reported claim data. Consequently, it would be easy and advisable to perform two multivariate analyses: one using paid claim data and the other using reported claim data.

1.1 Research Context

We focus on three elements of a basic rate review: loss trend, loss development, and rating class relativities. The actuarial literature on loss trend and loss development generally considers these elements in isolation. An exception involves accident year trends, since the latter require that data

[6] Bunkley, Nick, and Bill Vlasic, "Carmakers Initiating More Recalls Voluntarily," *The New York Times*, August 24, 2010.

are developed to ultimate. Even in this case, however, the loss development and loss trend analyses are performed sequentially instead of simultaneously.

Similarly, papers on risk classification tend to consider their subject in isolation. For example, in "A Practitioner's Guide to Generalized Linear Models" the discussion of loss trend and loss development occurs in Appendix F. The *Guide* suggests using a calendar/accident year method of organization and a dummy calendar year variable as a way to "absorb trends in claims experience that purely relate to time."[7] The *Guide* also suggests three options for dealing with loss development:

- Ignoring it — assuming it does not affect the classification factors.
- Including a dummy variable in the model to absorb time-related influences, removing it once the model is finalized, and adjusting the modeled results based on a separate calculation.
- Performing a series of GLM analyses, and comparing GLM relativities based on data at different development periods in order to obtain multivariate development factors.[8]

Styrsky noted that loss trend can be underestimated or overestimated when calendar year data are used in the analysis if the size of the portfolio increases or decreases significantly. He proposed an approach for dealing with this effect by matching each calendar year's claims by accident year to the exposures that produced them.[9] Werner and Modlin propose additional solutions to this problem: (1) using econometric models or generalized linear models to measure trend or (2) using accident year data (developed to ultimate) for trend analysis. They note that the loss development process "may introduce some subjectivity" in trend analyses, and state that the use of econometric models and generalized linear models for quantifying loss trends is beyond the scope of the text.[10]

Werner and Modlin point out a number of factors that can influence loss trends, such as inflation, technological advances, societal changes, and distributional changes. They suggest we can estimate the effect of distributional changes by looking at the trend in average premium at present rate level (PPR).[11] Why do that? The reason is that distributional changes affect both premiums and losses. For example, youthful drivers generally have higher loss costs than adult drivers, and insurers

[7] Anderson et al., p. 107.
[8] Anderson et al., p. 108.
[9] Styrsky, Chris, "The Effect of Changing Exposure Levels on Calendar Year Loss Trends," Casualty Actuarial Society Winter *Forum*, 2005, pp. 125-151.
[10] Werner and Modlin, pp. 111-114.
[11] Werner and Modlin, pp. 8, 81.

generally charge them higher premiums than adult drivers. Thus, if the proportion of youthful drivers in an insurance portfolio increases, both losses and premiums will increase.

As can be seen by the examples and citations above, the effect of changes in exposure level and distribution of exposures on commonly used univariate analysis of loss trend has been well studied and documented. The remaining question is what effect, if any, these exposure changes have on multivariate models.

1.2 Objectives

The objectives of this paper are (1) to illustrate how to expand a driver classification analysis database into a database that can also be used for univariate loss trend and loss development analyses as well as multivariate analyses involving all of these factors, (2) to compare results of using univariate and multivariate models for analyzing ratemaking parameters, (3) to show that multivariate analyses that account for all ratemaking parameters are robust to changes in exposure level and exposure distribution, and (4) to propose a framework for a rate review process completely based on multivariate analyses.

We begin by considering a line of insurance such, as property damage liability, with a relatively simple rating plan involving only territory and driver class. For simplicity, we assume any other rating factors such as anti-lock brake discounts or vehicle symbols are not applicable. We define subjects identified by policy ID and accident year, assuming that each policy insures one driver and one vehicle. Depending on rating manual rules, policies may insure multiple drivers and multiple vehicles. Some rating manuals specify rules for assigning a single driver classification to each vehicle. Other rating manuals assign a weighted average class factor, based on all drivers in the household, to each vehicle. When reviewing the rates and rating factors for a rating manual, the definition of a subject ID should be selected based on the entity to which manual rates and rating factors apply. SAS uses the keyword SUBJECT, but it can handle subjects as well as panels. A panel is a closely related group of subjects such as a household, or all vehicles and drivers insured by one policy, for which observations are expected to be correlated.

We will observe subjects across accident year evaluations, with cumulative claim counts per policy ID and accident year recorded at successive evaluation dates. For example, subject A, identified by policy ID 110000020 and accident year 2004, may have 0 claims as of 12 months, 1 claim as of 24 months, and 1 claim as of 36 months. In contrast, subject B, identified by policy ID 110000020 and accident year 2005, may have 0 claims at all evaluations (12, 24, and 36 months). This form of data organization is an example of longitudinal data, which Molenberghs and Verbeke

describe as the case where "the same characteristic is measured repeatedly over time, and time itself is, at least in part, a subject of scientific investigation."[12] Please note that we are considering observations of the same policy on two different accident years as two different subjects. We could have considered the subject as identified only by policy ID and tracked the claim counts across both accident years and evaluation dates. However, the evaluations for 2005 as of 12 months and 2004 as of 24 months occur at the same time. Similarly, the 24-month evaluation of 2005 and the 36-month evaluation of 2004 are simultaneous. Having some observations precede each other in time while others are simultaneous makes model parameterization more complicated and beyond the scope of this paper.

Methods for analyzing longitudinal data include generalized estimating equations (GEE) and generalized linear mixed-effects models. We focus on population averaged GEE (PA GEE), which are closely related to generalized linear models (GLM). PA GEE can be thought of as "GLM" in which the variance function includes a covariance matrix that represents the correlation between repeated observations of the same subject or panel. Another difference is that the estimating equations for GLM involve likelihood functions, while GEE use quasilikelihood functions. GLM have become standard tools in property and casualty insurance ratemaking. Thus, as we begin to think of insurance data as longitudinal data, it seems natural to use GEE as a tool for analyzing risk classification and time-related effects simultaneously. We will analyze claim frequency trend, claim count development, and claim frequency risk classification factors using SAS PROC GENMOD. We use PA GEE that model the marginal expectation for observations having the same covariate values (time index, evaluation age, territory, and driver class codes). Consequently, even though the inputs are observations from specific policyholders, the model provides information about "average" policyholders.

1.3 Outline

Section 2 of this paper outlines the theoretical background of population averaged generalized estimating equations (PA GEE), and introduces the database organization used as the common starting point for the analysis techniques discussed in this paper. Section 3 presents the results of applying several analysis techniques to simulated data to estimate classification effects (territory and driver class factors) and time-related effects (trend and loss development). The results compared and discussed include accident year claim frequency trend, percentage of cumulative claims closed with payment, and claim frequency relativities by risk classification.

---

[12] Molenberghs, G., and G. Verbeke, "Models for Discrete Longitudinal Data," New York: Springer Series in Statistics, Springer Science+Business Media, Inc., 2005, p. 3.

PA GEE analyses involve making initial assumptions about the correlation structure of measurements taken on the same subject at different times. Therefore, two GEE analyses are presented and discussed: one assuming autoregressive correlation AR(1), and the other assuming all measurements related to the same subject are equally correlated — exchangeable correlation. The output of the model is a set of coefficients for the variables in the estimating equation, and a correlation matrix. For examples of correlation matrices output by these models see Section 2.2.1. Since the models use a log link, the exponential of the coefficients of the fully specified models correspond to the annual trend factor, territory and classification relativity factors, and percentage paid (closed with payment) factors.

2. THEORETICAL BACKGROUND AND DATA ORGANIZATION

This section provides a brief description of the mathematical structure of generalized linear models (GLM) and population averaged generalized estimating equations (PA GEE), describes the method of data organization used as the starting point for the analyses described in this paper, and shows how to prepare the data for application of the analysis techniques discussed in the paper. This paper uses only one type of GEE models: PA GEE. There are other types of GEE models, which are beyond of the scope of this paper. For more information, see Hardin and Hilbe.[13]

2.2 Generalized Linear Models

A Practitioner's Guide to Generalized Linear Models defines a GLM in terms of three components:[14]

- A random component $\mathbf{Y}$ in which each element y_i is assumed to be independent and a member of the exponential family of distributions, for which the variance is a function of the expected value of Y, a scale parameter, and a weight assigned to each observation.
- A systematic component consisting of a set of explanatory or predictive variables, such as territory and driver classification, represented by a vector $\mathbf{X}$ and a set of coefficients represented by a vector $\boldsymbol{\beta}$.
- A link function g such that

$$g(E[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \ldots + x_n\beta_n. \tag{2.1}$$

[13] Hardin, James W. and Joseph M. Hilbe, *Generalized Estimation Equations*, Boca Raton, FL: Chapman & Hall/CRC, 2003.
[14] Anderson et al., pp. 13, 14.

For example, if we used the natural logarithm as the link function g, then $E(\mathbf{Y}) = exp(\mathbf{X\beta})$. Thus, if x_1 represents whether or not a policyholder resides in territory 1, the relativity for that territory would be given by $exp(\beta_1)$.

2.2.1 The Independence Assumption

Suppose we are using the latest three accident years (e.g., 2007 to 2009) to evaluate driver classification factors for an insurance portfolio, and each policy insures one driver and one vehicle. Then, if a policyholder has been insured for three years the vector **Y** has three entries for this policyholder corresponding to 2007, 2008, and 2009. For purposes of the GLM, it is customary to treat these observations as independent. There is no way to do otherwise, because this is one of the fundamental assumptions of GLM. However, they are likely to be correlated because they are observations of the same subject. Furthermore, the prevalence of safe driver insurance plans, accident and violation surcharges, and merit rating plans suggests that actuaries believe these observations are not really independent. In fact, most actuaries believe a policyholder who has had a claim is more likely to have claim in the future than a policyholder who has had no claims.

2.2 Population Averaged Generalized Estimating Equations

Suppose we observe cumulative claims closed with payment by policyholder by accident year from 2004 to 2009, at 12, 24, and 36 months. Further, assume all claims are closed by 36 months. Then we have 15 observations for each policyholder: three for each of the first four years, two for 2008 and one for 2009. Conversely, we have three missing observations: two for 2009 (the 24- and 36-month evaluations), and one for 2008 (the 36-month evaluation). The 18 total missing and non-missing observations correspond to six years and three evaluation dates for a policy in-force throughout the entire experience period. In this way we can see a policyholder's experience as longitudinal data in which the number of claims is observed at different points in time. We are interested in the relationship between time (accident year and evaluation date) and claim count, as well as the relationship between classification variables (territory and driver class) and claim count.

For the purposes of this paper, we will continue to assume independence across accident years, as is generally assumed when using GLM. Therefore we will define our subjects by policy ID and accident year, once again assuming each policy insures only one driver and one vehicle. It is easy to see that claims closed at different evaluation dates are correlated. For example a policyholder with one claim closed with payment as of 12 months for accident year 2007 will generally have at least one claim closed with payment at each successive evaluation date for that year. The way in which a company codes reopened claims can make this relationship more complicated. An actuary pricing a

book of business would have to understand how reopened claims are coded, and whether or not there has been a change in claim reopening patterns during the experience period. For the purposes of this paper, we assume there are no reopened claims. We will seek only to model the correlation among claims closed with payment at different evaluation dates. A more general correlation structure incorporating correlation across accident years could be formulated, but it is beyond the scope of this paper.

The data for an insurance portfolio observed at subsequent accident years and evaluation dates can be seen as having the random and systematic components of a GLM as well as a correlated component for which the GLM does not account. SAS PROC GENMOD can model the systematic component for data with both independent and correlated observations using the same linear predictor, variance function, and link function as the independent case, but it can also model the correlation structure of the correlated observations.[15]

Let $\mathbf{Y}_i$ represent the vector of n_i observations for policyholder i, $\mathbf{X}_{ij}$ the vector of covariates (explanatory or predictive variables) for the j^{th} observation of the i^{th} policyholder, $\boldsymbol{\beta}$ the vector of coefficients, and $\mathbf{V}_i$ the covariance matrix of the n_i correlated observations in $\mathbf{Y}_i$. Then the GEE model can be specified by the following equations:

$$g(E[\mathbf{Y}_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + \ldots + x_{ijp}\beta_p. \tag{2.2}$$

$$\mathbf{V}_i = \varphi \mathbf{A}_i^{1/2}\, \mathbf{W}_i^{-1/2} \mathbf{R}(\alpha)\, \mathbf{W}_i^{-1/2} \mathbf{A}_i^{1/2}. \tag{2.3}$$

Where φ is a dispersion parameter, $\mathbf{A}$ is a diagonal matrix of variance functions $v(\mu_{ij})$, $\mathbf{W}$ is a diagonal matrix of weights, and $\mathbf{R}(\alpha)$ is a working correlation matrix. When no weights are specified by the user, $\mathbf{W}$ defaults to a matrix of 1s, and all observations receive equal weight. When $\mathbf{R}(\alpha)$ is the identity matrix, equation 2.3 reduces to the variance function of the independent case.

2.2.1 Working Correlation Matrix

Six working correlation structures are available in SAS PROC GENMOD: fixed, identity, *m*-dependent, exchangeable, unstructured, and auto regressive AR(1). In the fixed case, the correlation matrix is specified by the user. The identity is the special case with 1s in the diagonal and 0s elsewhere, and it is equivalent to the independence case. *M*-dependent means that only *m* of the observations for a given subject are correlated, and the rest are not. Exchangeable is the case where all observations for a given subject are equally correlated. Unstructured implies that the correlation between any pair of observations is different from, and unrelated to, the correlation between any other pair of observations. The autoregressive structure is more appropriate for observations where

[15] *SAS/STAT® 9.2 User's Guide*, SAS Institute Inc., 2008, pp. 1984, 1985.

the correlation decays as time elapses. Following are illustrations of the autoregressive and exchangeable correlation structures for a PA GEE model with simulated counts of personal auto property damage liability claims closed with payment observed at 12, 24, and 36 months per policy ID and accident year.

Autoregressive

$$\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.8323 & 0.6927 \\ 0.8323 & 1 & 0.8323 \\ 0.6927 & 0.8323 & 1 \end{pmatrix}.$$

The autoregressive correlation matrix above would indicate that the correlation between two successive evaluations (12 months and 24 months or 24 months and 36 months) is roughly 83%, while the correlation between the 12-month and 36-month evaluations is about 69%. In contrast, the exchangeable correlation matrix below would indicate that all evaluations have a correlation of roughly 79%.

Exchangeable

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.7916 & 0.7916 \\ 0.7916 & 1 & 0.7916 \\ 0.7916 & 0.7916 & 1 \end{pmatrix}.$$

For a given accident year, the claim count at 36 months is theoretically more correlated with the claim count at 24 months than with the count at 12 months. This would support using an autoregressive correlation structure. Nevertheless, both correlation structures shown above are tested in this paper for comparison purposes.

2.2.2 Missing Values

As mentioned previously, when one observes claim counts for a policyholder by accident year at different evaluation dates, some evaluation dates are missing. In the examples used in this paper, the latest year only has the 12-month evaluation, and the previous year has the 12- and 24-month evaluations. In cases such as this, the GENMOD procedure uses the "all available pairs" method to estimate the moments for the working correlation parameters. This method depends on the "missing completely at random (MCAR)" assumption.[16]

The pattern of missing values for a policyholder's claim counts by accident year and evaluation date is somewhat systematic. For each accident year, either all evaluation dates are present, or they are all missing after some point. This is called a "dropout" missing pattern. It is similar to that of a

[16] *SAS/STAT® 9.2 User's Guide*, p. 1987.

patient who stops participating in a medical study. Additionally, policies may be written or non-renewed in the middle of the experience period, which creates another source of missing values. A non-renewed policy acts as a dropout. A new policy, on the other hand acts as a drop-in, where the earlier values are missing. Also, as some individuals become older or move, they may become part of another driver class or territory. Finally, insurance companies may not renew policies of people who have had many claims, or they may re-underwrite an insurance portfolio switching policyholders from one class to another if they are found to have been misclassified. Therefore, some of the factors causing missing values are systematic, while others are random. Determining whether or not the pattern of missing values for an insurance book meets the MCAR assumption is beyond the scope of this paper. Readers may refer to section 4.6 of Hardin and Hilbe's *Generalized Estimating Equations*.[17]

The data simulations run for this paper contain no missing values other than the ones that would correspond to future evaluation dates for the most recent accident years. The modeling results indicate that the inclusion of evaluation age parameters adequately accounts for the missing evaluation dates. Furthermore, movement of policyholders from one class or territory to another as a result of aging, moving, re-underwriting or non renewal can be seen as a distributional shift in exposures rather than a source of missing values. The data simulations do include samples with significant distributional changes, and the modeling results show that the claim frequency PA GEE models are not affected by distributional shifts in exposure. Therefore we can conclude that the missing values encountered when fitting a claim frequency PA GEE model to an insurance portfolio are not likely to adversely affect the modeling results.

2.3 Data Organization

Many companies have begun to build data warehouses or ratemaking databases with very detailed information including policy effective and expiration dates, driver attributes, vehicle attributes, date of accident, date of report, date closed, amounts paid, amounts in reserve, etc. At the start of a basic rate review, however, separate summarizations are extracted from this database for the trend system, loss development system, statewide indication, and territory and classification analysis review. Appendix G of *A Practitioner's Guide to Generalized Linear Models*[18] presents several forms of data organization that can be used for generalized linear model (GLM) analysis, as well as their advantages and disadvantages. One of these is the calendar/accident year method in which each record has claim counts and loss amounts as of the latest evaluation. A simple expansion of this

[17] Hardin and Hilbee.
[18] Anderson et al.

database is to include separate columns for each evaluation age. This kind of setup can be used as a starting point from which, with very little manipulation, several univariate and multivariate analyses can be performed. The following table illustrates this method of organization for a hypothetical rating plan using only territory and driver classification and insuring only one driver and vehicle per policy. A real database would contain many other attributes identifying the policies, drivers and vehicles insured as well as rating characteristics associated with them.

Table 2.3.1

Policy Id	Accident Year	Territory	Driver Class	Earned Exposure	Claims With Payment Age 12	Claims With Payment Age 24	Claims With Payment Age 36
110000020	2004	1	1	1	0	1	1
110000020	2005	1	1	1	0	0	0
110000020	2006	1	1	1	0	0	0
110000020	2007	1	1	1	0	0	0
110000020	2008	1	1	1	1	1	
110000020	2009	1	1	1	0		

Paid Loss Amount Age 12	Paid Loss Amount Age 24	Paid Loss Amount Age 36	Claims Reported Age 12	Claims Reported Age 24	Claims Reported Age 36	Reported Losses Age12	Reported Losses Age24	Reported Losses Age36
0	25,000	25,000	1	1	1	20,000	25,000	25,000
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
15,000	15,000		1	1		15,000	15,000	
0			0			0		

The example above illustrates a policy for which the territory and driver class have not changed during the experience period. The only missing values are future evaluation dates for the latest two accident years, assuming that all claims have been reported by age 36. Suppose another policy had been written in 7/1/2005, then the earned exposure for that policy in 2005 would be 0.5 and the exposure and claim counts for 2004 would be missing.

Throughout this paper we assume each policy insures only one driver and one vehicle. Therefore, we use policy ID and year as the subject for our PA GEE models. In reality, most policies actually insure more than one driver and one vehicle. Some companies assign a specific driver to each vehicle on the policy, while others use an average driver factor for all vehicles in the policy. Actuaries wishing to use PA GEE models need to be mindful of the driver and vehicle assignment

procedure use in their specific book of business, and they may need to add driver ID or vehicle ID or both to the subject definition. An alternative is to analyze the data in terms of panels including all drivers and vehicles in each insured household.

Another issue that may arise involves claims not related to a specific driver or vehicle. For example, a minor child who is not a driver may be injured as a pedestrian and covered by medical payments. This could be handled in a number of ways: the claims may be excluded, the claims may be coded with a dummy policy ID and the driver and vehicle attributes of the at-fault driver, or they may be coded with a dummy policy ID and the base driver and vehicle attributes. The best course of action would have to be determined by the actuary working on a particular book of business, based on the available information.

Starting with a database structure such as the one above, it is very easy to summarize claim counts for different evaluation dates by accident year to obtain a claim count triangle for chain-ladder development. For details see Appendix C.

For a traditional classification analysis using a GLM with accident year as a dummy variable, the data can be summarized by keeping only the cumulative claims reported as of the latest evaluation date, as illustrated in the Table 2.3.2 below. This leads to one of the types of data organization in Appendix G of the *Practitioners Guide to GLM* in which there is some loss of some information for policies with multiple claims in the same accident year, but this is generally not material.[19] Data such as the one illustrated below will be used for two types of models investigated in this paper: (1) GLM for claims closed with payment as the dependent variable and territory and classification as independent variables, and (2) GLM for claims closed with payment as the dependent variable and territory, classification, and dummy year as independent variables. As will be shown in Section 3.4, the dummy year parameter captures both trend and development effects. Immature year claim counts can be drastically lower than fully mature year claim counts. Trend, on the other hand, tends to be a gradual change. Therefore, modeling the combined effect of trend and development with a single continuous variable can be difficult. For this reason, it is better to use dummy year as a categorical variable.

[19] Anderson et al., p. 109.

Table 2.3.2

Policy Id	Accident Year	Territory	Driver Class	Earned Exposure	Cumulative Claims Closed With Payment As of 12/31/2009
110000020	2004	1	1	1	1
110000020	2005	1	1	1	0
110000020	2006	1	1	1	0
110000020	2007	1	1	1	0
110000020	2008	1	1	1	1
110000020	2009	1	1	1	0

For GLM and PA GEE analyses involving loss development parameters in addition to trend and classification factors, we are interested in the repeated observations across evaluation dates, so we would stack the evaluation dates into one column in order to get one observed claim count per record as illustrated below in Table 2.3.3. Note that the earned exposure needs to be repeated so that the cumulative claims at each age can be associated with the corresponding accident year's earned exposure for the policy.

Table 2.3.3

Policy ID	Accident Year	Evaluation Date	Territory	Driver Class	Earned Exposure	Closed With Payment
110000020	2004	12	1	1	1	0
110000020	2004	24	1	1	1	1
110000020	2004	36	1	1	1	1
110000020	2005	12	1	1	1	0
110000020	2005	24	1	1	1	0
110000020	2005	36	1	1	1	0
110000020	2006	12	1	1	1	0
110000020	2006	24	1	1	1	0
110000020	2006	36	1	1	1	0
110000020	2007	12	1	1	1	0
110000020	2007	24	1	1	1	0
110000020	2007	36	1	1	1	0
110000020	2008	12	1	1	1	1
110000020	2008	24	1	1	1	1
110000020	2009	12	1	1	1	0

The reason for converting from the triangular format in Table 2.3.1 to a stacked format is that the software expects only one dependent variable. Three multivariate models explored in this paper involve the count of claims closed with payment as the dependent variable, and time index (for trend), territory, driver class, and evaluation date as independent variables. The first one is a GLM. The second one is a PA GEE with policy ID and accident year as subject identifiers and

autoregressive working correlation. The third one is a PA GEE with policy ID and accident year as subject identifiers and exchangeable working correlation.

3. CLAIM SIMULATIONS AND RESULTS

This section presents and compares the results of applying different techniques to 30 synthetic personal auto property damage portfolios: 10 scenarios for each of three hypothetical states X, Y, and Z. These portfolios have a very simple classification plan with only three territories and three driver classes. Details of the procedure used to create the portfolios and simulate the claim counts are provided in Appendix A.

The reasons for using synthetic portfolios are: (1) to generate claim databases with parameters known a priori, (2) to eliminate as much as possible random variation of the expected claim frequencies, and (3) to make change in exposure levels and distribution the predominant source of variation. The objective is to gauge the effect of changes in exposure level and exposure distribution on different analysis methods while holding everything else as constant as possible. To achieve this end, we select the following parameters: base claim frequency, annual frequency trend, percentage of claims closed with payment as of each evaluation age, territory relativity, and driver-class relativity. We then use these selected parameters to determine the expected claim frequency for each territory, driver class, accident year, and evaluation age. For example, given the following parameters:

- base frequency = 0.05.
- territory 1 relativity = 1.50.
- driver class 1 relativity = 1.00.
- percentage of claims paid (closed with payment) as of 12 months = 0.50.

We calculate the 2002 expected claim frequency for State X, Scenario 1, territory 1 and driver class 1, at age 12 as: $0.05 \times 1.50 \times 1.00 \times 0.50 = 0.0375$. With a 3% annual trend, the 2003 claim frequency at 12 months would be $0.0375 \times 1.03 = 0.038625$, and the 2004 claim frequency at 12 months would be $0.0375 \times 1.03^2 = 0.03978375$.

Next, we multiply the expected claim frequencies times the corresponding earned exposures in the portfolio to determine expected claim counts for each accident year, territory, and driver class combination. Once we have used the selected parameters to determine the expected claim count for each territory, driver class, accident year, and evaluation age, we select policies at random with replacement up to the number of expected claim counts. We consider a policy not selected to have

zero claims, a policy selected once to have one claim, a policy selected twice to have two claims, and so on. The synthetic portfolios with claim emergence simulations are available for downloading from the CAS Web Site.

State X scenarios assume the annual trend in claim frequency is zero. The main source of variation is changing exposure level from one year to the next. Changes in claim frequency distribution between territories and driver classes are limited to roughly one-tenth of a percentage point. Base frequency is 0.05 (for territory 2 and driver class 1); claim frequency relativities are constant—1.50 for territory 1, 0.80 for territory 3, 2.00 for driver class 2, and 0.75 for driver class 3. Cumulative percentages of claims paid are 50% at 12 months, 80% at 24 and 100% at 36.

State Y scenarios assume a 3% annual trend in claim frequency and increasing exposure from one year to the next. State Y Scenario 1 has essentially the same distribution of exposures across territories and driver class for each accident year as State X scenario 1. The rest of the State Y scenarios show more random variation than State X scenarios in the distribution of exposures among territories and driver classes from one accident year to the next. Base frequency is 0.06 (for territory 2 and driver class 1); claim frequency relativities and cumulative percentages of claims paid are the same as State X Scenario 1.

State Z scenarios assume a 3% annual trend in claim frequency and decreasing exposure from one year to the next. State Z Scenario 1 has essentially the same distribution of exposures across territories and driver class for each accident year as State X scenario 1. The rest of the State Z scenarios have increasing systematic variation in the distribution of exposures among territories and driver classes across accident years. Each accident year, the territory 1 class 1 earned car years decrease while the territory 3 class 3 earned car years increase, and the magnitude of this changes increases from scenario 2 to scenario 10. Base frequency is 0.02 (for territory 2 and driver class 1); claim frequency relativities and cumulative percentages of claims paid are the same as State X Scenario 1.

The following sections compare parameter estimates obtained by different methods for percentage of claims paid (closed with payment) by evaluation age, accident year trend, claim frequency relativities, quasi-likelihood information criterion, correlation matrices, and covariance matrices, where applicable.

3.1 Percentage of Claims Paid (Closed With Payment)

As mentioned earlier, the claim count simulation parameters were selected so the payment pattern would be approximately 50%, 80%, 100% of claims paid by 12, 24, and 36 months,

respectively. Since the number of claims must be a whole number, some deviation from those percentages is to be expected. For example, if the expected claim frequency for a given accident year, evaluation age, territory and driver class is 0.0375 and there are 1,000 earned car years, we can simulate either 37 or 38 claims, not 37.5.

The percentage paid estimate for the chain ladder method is the reciprocal of the age-to-ultimate development factor. Details of the calculation are shown in Appendix C. The percentage paid estimates for the GLM and GEE models are based on the parameter estimates for the levels of the evaluation age. Details are provided in Appendices D and E. The simulations used in this paper assumed stable development patterns. In a real-world situation, changes in claim adjustment patterns, system changes, etc., may cause development factors to change between years. If the change is gradual over several years, a marginal interaction term (based on time index and evaluation age) can be added to the model to account for these changes. If the change is more abrupt, so that accident years after a certain point are different from earlier accident years, a (0, 1) control variable could be introduced to account for the change. An actuary pricing a specific book of business would have to determine an appropriate course of action based on the available information.

The following Table 3.1.1 presents the resulting estimates of percentage of claims paid (closed with payment) by evaluation date using the chain ladder method (CLM), generalized linear model (GLM Full), generalized estimating equations with autoregressive correlation (GEE AR), and generalized estimating equations with exchangeable correlation (GEE Ex). All four methods produced estimates that are close to each other and close to the percentages used to set up the claim payment simulations.

Table 3.1.1

	Percentage of Claims Paid by Age 12				Percentage of Claims Paid by Age 24			
Sample	CLM	GLM Full	GEE AR	GEE Ex	CLM	GLM Full	GEE AR	GEE Ex
X-01	0.49997	0.50000	0.49998	0.49998	0.79998	0.80001	0.79999	0.79999
X-02	0.49999	0.50001	0.50000	0.50000	0.79999	0.80002	0.79999	0.80000
X-03	0.49999	0.50000	0.49999	0.50000	0.80000	0.80001	0.80000	0.80000
X-04	0.49996	0.49996	0.49996	0.49996	0.79997	0.79998	0.79998	0.79997
X-05	0.49999	0.50001	0.50000	0.50000	0.79999	0.80001	0.80000	0.80000
X-06	0.50001	0.50002	0.50001	0.50001	0.79999	0.79999	0.79999	0.79999
X-07	0.50003	0.50001	0.50002	0.50001	0.79999	0.79998	0.79999	0.79998
X-08	0.50002	0.50003	0.50002	0.50002	0.80002	0.80002	0.80002	0.80002
X-09	0.49999	0.49998	0.49999	0.49999	0.80002	0.80000	0.80001	0.80001
X-10	0.50000	0.49998	0.49999	0.49999	0.79999	0.79998	0.79998	0.79997
Y-01	0.50001	0.50002	0.50001	0.50002	0.79999	0.80000	0.80000	0.80000
Y-02	0.50000	0.49999	0.50000	0.50000	0.79999	0.79999	0.79999	0.80000
Y-03	0.50003	0.50003	0.50002	0.50003	0.80003	0.80004	0.80002	0.80003
Y-04	0.50002	0.50004	0.50002	0.50002	0.79999	0.80003	0.80000	0.80001
Y-05	0.50000	0.50000	0.50000	0.50000	0.79999	0.79998	0.79999	0.79999
Y-06	0.50003	0.50000	0.50002	0.50002	0.80001	0.79999	0.80000	0.80000
Y-07	0.50000	0.50002	0.50000	0.50000	0.80003	0.80003	0.80002	0.80002
Y-08	0.49997	0.49996	0.49997	0.49997	0.79999	0.79996	0.79998	0.79998
Y-09	0.50000	0.50002	0.50000	0.50000	0.79998	0.79999	0.79998	0.79998
Y-10	0.50000	0.50002	0.50001	0.50001	0.80000	0.80000	0.80000	0.80000
Z-01	0.49992	0.49988	0.49990	0.49990	0.80003	0.79997	0.80001	0.80001
Z-02	0.49993	0.49986	0.49991	0.49992	0.80001	0.79994	0.80000	0.80001
Z-03	0.49997	0.49997	0.49997	0.49996	0.80014	0.80010	0.80014	0.80012
Z-04	0.50005	0.50000	0.50004	0.50003	0.79999	0.79993	0.79998	0.79997
Z-05	0.50003	0.50009	0.50004	0.50004	0.80013	0.80019	0.80014	0.80015
Z-06	0.50007	0.50003	0.50006	0.50007	0.79996	0.79991	0.79995	0.79996
Z-07	0.49992	0.49983	0.49990	0.49991	0.80000	0.79992	0.79998	0.80000
Z-08	0.49990	0.49992	0.49991	0.49991	0.79997	0.79995	0.79997	0.79997
Z-09	0.50011	0.50007	0.50008	0.50009	0.80018	0.80026	0.80019	0.80021
Z-10	0.50004	0.50005	0.50004	0.50003	0.79997	0.79998	0.79998	0.79997
Average	0.50000	0.49999	0.50000	0.50000	0.80001	0.80001	0.80001	0.80001
Std Dev	0.00004	0.00006	0.00004	0.00004	0.00005	0.00007	0.00005	0.00005
Min	0.49990	0.49983	0.49990	0.49990	0.79996	0.79991	0.79995	0.79996
Max	0.50011	0.50009	0.50008	0.50009	0.80018	0.80026	0.80019	0.80021
Range	0.00021	0.00026	0.00018	0.00019	0.00022	0.00035	0.00024	0.00025

3.2 Claim Frequency Trend

The calendar year trend analysis is based on calendar year data — claim counts are assigned to the year in which the claim was paid. Details are provided in Appendix B. The accident year trend analysis is based on an exponential regression on ultimate claim frequencies, so it depends on the results of the chain ladder method. Details are shown in Appendix C. The accident year trend

estimates for the GLM Full, GEE AR, and GEE Ex models are based on the coefficient for a time index. Details of GLM and GEE models are provided in Appendices D and E.

The samples were intended to simulate changes in exposure level and changes in the distribution of exposures — two issues the basic ratemaking textbook mentions among the ones that can affect the results of univariate trend analyses, and illustrated in the introduction to this paper. Therefore, it is not surprising that the calendar year (Cal Yr) and accident year (Acc Yr) trend estimates deviate from the actual annual trend used to generate the simulated data: 0% for State X and 3% for States Y and Z.

The multivariate methods include a generalized linear model (GLM Full), generalized estimating equations with autoregressive correlation (GEE AR), and GEE with exchangeable correlation (GEE Ex). These methods are resistant to the changes in exposure level and distribution simulated in these samples. The following Table 3.2.1 summarizes the results.

Table 3.2.1

	6-Point Annual Trend Estimates				
Sample	Cal Yr	Acc Yr	GLM Full	GEE AR	GEE Ex
X-01	2.98%	0.00%	0.00%	0.00%	0.00%
X-02	2.97%	-0.04%	0.00%	0.00%	0.00%
X-03	2.96%	-0.02%	0.00%	0.00%	0.00%
X-04	2.96%	0.00%	0.00%	0.00%	0.00%
X-05	3.02%	-0.04%	0.00%	0.00%	0.00%
X-06	3.10%	0.10%	0.00%	0.00%	0.00%
X-07	2.86%	0.05%	0.00%	0.00%	0.00%
X-08	3.02%	0.02%	0.00%	0.00%	0.00%
X-09	2.97%	0.01%	0.00%	0.00%	0.00%
X-10	2.85%	-0.05%	0.00%	0.00%	0.00%
Average	2.97%	0.00%	0.00%	0.00%	0.00%
Std Dev	0.07%	0.05%	0.00%	0.00%	0.00%
Min	2.85%	-0.05%	0.00%	0.00%	0.00%
Max	3.10%	0.10%	0.00%	0.00%	0.00%
Range	0.25%	0.15%	0.00%	0.00%	0.00%

Table 3.2.1, continued

	6-Point Annual Trend Estimates				
Sample	Cal Yr	Acc Yr	GLM Full	GEE AR	GEE Ex
Y-01	3.00%	3.00%	3.00%	3.00%	3.00%
Y-02	3.09%	3.42%	3.00%	3.00%	3.00%
Y-03	2.80%	3.58%	3.00%	3.00%	3.00%
Y-04	3.26%	2.97%	3.00%	3.00%	3.00%
Y-05	3.24%	3.54%	3.00%	3.00%	3.00%
Y-06	3.97%	4.09%	3.00%	3.00%	3.00%
Y-07	2.64%	2.83%	3.00%	3.00%	3.00%
Y-08	2.76%	3.04%	3.00%	3.00%	3.00%
Y-09	3.52%	4.42%	3.00%	3.00%	3.00%
Y-10	2.76%	3.46%	3.00%	3.00%	3.00%
Average	3.10%	3.44%	3.00%	3.00%	3.00%
Std Dev	0.41%	0.51%	0.00%	0.00%	0.00%
Min	2.64%	2.83%	3.00%	3.00%	3.00%
Max	3.97%	4.42%	3.00%	3.00%	3.00%
Range	1.33%	1.59%	0.00%	0.00%	0.00%
Z-01	3.71%	3.01%	3.01%	3.01%	3.01%
Z-02	2.88%	2.47%	2.99%	2.98%	2.98%
Z-03	2.39%	2.28%	2.98%	2.99%	2.99%
Z-04	3.13%	2.39%	2.99%	2.99%	2.99%
Z-05	3.80%	3.33%	3.00%	3.01%	3.01%
Z-06	2.82%	3.00%	3.01%	3.01%	3.01%
Z-07	3.11%	2.07%	3.01%	3.00%	3.00%
Z-08	3.68%	3.04%	3.00%	3.00%	3.00%
Z-09	2.03%	1.62%	3.02%	3.02%	3.02%
Z-10	3.49%	2.29%	3.01%	3.01%	3.01%
Average	3.10%	2.55%	3.00%	3.00%	3.00%
Std Dev	0.59%	0.53%	0.01%	0.01%	0.01%
Min	2.03%	1.62%	2.98%	2.98%	2.98%
Max	3.80%	3.33%	3.02%	3.02%	3.02%
Range	1.77%	1.71%	0.04%	0.04%	0.04%

3.3 Claim Frequency Relativities

This section compares the results of six different multivariate models for claim frequency relativities — four generalized linear models and two generalized estimating equation (GEE) models. The autoregressive correlation model (GEE AR) makes more sense intuitively than the exchangeable correlation model (GEE Ex), since we would expect the correlation between 36-month and 24-month claim counts to be larger than the correlation between 36-month and 12-month claims counts.

Table 3.3.1

Model	Description
GLM 6Yr	Generalized linear model with latest 6 years of data and territory and driver class as independent variables
GLM 3Yr	Generalized linear model with latest 3 years of data and territory and driver class as independent variables
GLM AYC	Generalized linear model with latest 6 years of data, territory and driver class as independent variables, and accident year as control variable
GLM Full	Generalized linear model with latest 6 years of data, territory, driver class, time index, and evaluation age as independent variables
GEE AR	Generalized estimating equation model with latest 6 years of data, territory, driver class, time index, and evaluation age as independent variables and autoregressive working correlation
GEE Ex	Generalized estimating equation model with latest 6 years of data, territory, driver class, time index, and evaluation age as independent variables and exchangeable working correlation

The first two models, which ignore differences across accident year, are less reliable than the last four models — the range of expected values they produce is wider. Additionally, the base frequency (intercept) estimated by these models, which is the expected value across the 6-year or 3-year period, respectively, is understated because these models ignore the fact that the latest two years are not fully developed The understatement is more pronounced for the 3-year model because two out of three years are not fully developed.

The model with accident year as a control variable (GLM AYC) and the fully specified models (GLM full, GEE AR, and GEE Ex) quite accurately predict the base frequency of 0.05 for State X, 0.06×1.03^2 for State Y, and 0.02×1.03^2 for State Z. The reason for the factor of 1.03 squared in States Y and Z is that a 3% annual trend was assumed in the simulation. Eight accident years were simulated starting with 2002, so to get the base frequency for 2004 we must multiply times 1.03 squared.

Following are the indicated base frequencies and the indicated factors for territories 1 and 3 as well as driver classes 2 and 3. Since the states have different base frequencies, the statistics (average, standard deviation, minimum, maximum, and range) for the intercept are by state. On the other hand, territory and driver class relativities are the same for all states so the statistics are across all 30 samples.

Table 3.3.2

	Indicated Base Frequency (Intercept)					
Sample	**GLM 6Yr**	**GLM 3Yr**	**GLM AYC**	**GLM full**	**GEE AR**	**GEE Ex**
X-01	0.04456	0.03899	0.04999	0.05000	0.05000	0.05000
X-02	0.04458	0.03900	0.05000	0.05000	0.05000	0.05000
X-03	0.04458	0.03900	0.04999	0.05000	0.04999	0.05000
X-04	0.04457	0.03902	0.05000	0.05000	0.05000	0.05000
X-05	0.04454	0.03898	0.05000	0.05000	0.05000	0.05000
X-06	0.04455	0.03894	0.05000	0.05000	0.05000	0.05000
X-07	0.04459	0.03900	0.04999	0.05000	0.05000	0.05000
X-08	0.04461	0.03904	0.05000	0.05000	0.05000	0.05000
X-09	0.04458	0.03904	0.05000	0.05001	0.05001	0.05001
X-10	0.04455	0.03897	0.04999	0.05000	0.05000	0.05000
Average	0.04457	0.03900	0.05000	0.05000	0.05000	0.05000
Std Dev	0.00002	0.00003	0.00001	0.00000	0.00000	0.00000
Min	0.04454	0.03894	0.04999	0.05000	0.04999	0.05000
Max	0.04461	0.03904	0.05000	0.05001	0.05001	0.05001
Range	0.00007	0.00010	0.00001	0.00001	0.00002	0.00001
Y-01	0.05942	0.05405	0.06365	0.06365	0.06365	0.06365
Y-02	0.05939	0.05395	0.06364	0.06364	0.06365	0.06365
Y-03	0.05952	0.05394	0.06365	0.06365	0.06365	0.06365
Y-04	0.05915	0.05322	0.06365	0.06365	0.06365	0.06365
Y-05	0.05987	0.05428	0.06366	0.06366	0.06366	0.06366
Y-06	0.05946	0.05418	0.06365	0.06365	0.06366	0.06366
Y-07	0.05919	0.05355	0.06365	0.06366	0.06366	0.06366
Y-08	0.05942	0.05398	0.06365	0.06365	0.06365	0.06365
Y-09	0.05979	0.05438	0.06366	0.06366	0.06366	0.06366
Y-10	0.06004	0.05444	0.06364	0.06364	0.06364	0.06364
Average	0.05953	0.05400	0.06365	0.06365	0.06365	0.06365
Std Dev	0.00029	0.00038	0.00001	0.00001	0.00001	0.00001
Min	0.05915	0.05322	0.06364	0.06364	0.06364	0.06364
Max	0.06004	0.05444	0.06366	0.06366	0.06366	0.06366
Range	0.00089	0.00122	0.00002	0.00002	0.00002	0.00002
Z-01	0.02039	0.01855	0.02120	0.02121	0.02121	0.02121
Z-02	0.02048	0.01878	0.02123	0.02123	0.02123	0.02123
Z-03	0.02033	0.01845	0.02123	0.02123	0.02123	0.02123
Z-04	0.02049	0.01877	0.02123	0.02122	0.02123	0.02123
Z-05	0.02058	0.01882	0.02122	0.02122	0.02122	0.02122
Z-06	0.02054	0.01887	0.02122	0.02122	0.02122	0.02122
Z-07	0.02048	0.01877	0.02122	0.02122	0.02122	0.02122
Z-08	0.02057	0.01868	0.02123	0.02122	0.02122	0.02122
Z-09	0.02048	0.01870	0.02122	0.02121	0.02121	0.02121
Z-10	0.02055	0.01871	0.02121	0.02121	0.02121	0.02121
Average	0.02049	0.01871	0.02122	0.02122	0.02122	0.02122
Std Dev	0.00008	0.00013	0.00001	0.00001	0.00001	0.00001
Min	0.02033	0.01845	0.02120	0.02121	0.02121	0.02121
Max	0.02058	0.01887	0.02123	0.02123	0.02123	0.02123
Range	0.00025	0.00042	0.00003	0.00002	0.00002	0.00002

Table 3.3.3

	Indicated Rating Factors for Territory 1					
Sample	**GLM 6Yr**	**GLM 3Yr**	**GLM AYC**	**GLM full**	**GEE AR**	**GEE Ex**
X-01	1.49993	1.49999	1.49993	1.49995	1.49992	1.49991
X-02	1.50005	1.50007	1.50019	1.50010	1.50018	1.50024
X-03	1.49988	1.49958	1.50008	1.50005	1.50007	1.50007
X-04	1.49945	1.49933	1.49979	1.49987	1.49979	1.49976
X-05	1.50103	1.50050	1.50011	1.50012	1.50010	1.50009
X-06	1.49981	1.50208	1.50010	1.50010	1.50011	1.50012
X-07	1.49841	1.49927	1.50006	1.49999	1.50005	1.50004
X-08	1.49819	1.49945	1.49982	1.49982	1.49981	1.49978
X-09	1.49848	1.49521	1.49985	1.49982	1.49982	1.49982
X-10	1.49993	1.49660	1.49996	1.50001	1.49995	1.49995
Y-01	1.50014	1.50002	1.50014	1.50010	1.50014	1.50015
Y-02	1.50996	1.52085	1.50002	1.50004	1.50002	1.50003
Y-03	1.48432	1.48797	1.49996	1.49997	1.49996	1.49998
Y-04	1.51082	1.53509	1.50015	1.50015	1.50014	1.50013
Y-05	1.49118	1.49591	1.49992	1.49987	1.49993	1.49994
Y-06	1.49503	1.48626	1.50006	1.50004	1.50005	1.50005
Y-07	1.49821	1.48654	1.50004	1.50005	1.50005	1.50008
Y-08	1.50404	1.50523	1.50006	1.50009	1.50005	1.50004
Y-09	1.49080	1.50031	1.49992	1.49991	1.49992	1.49991
Y-10	1.48157	1.47763	1.50024	1.50023	1.50024	1.50027
Z-01	1.50076	1.50097	1.50076	1.50048	1.50060	1.50054
Z-02	1.49198	1.47865	1.49935	1.49905	1.49924	1.49932
Z-03	1.51716	1.52657	1.49950	1.49934	1.49944	1.49943
Z-04	1.48905	1.46689	1.49996	1.49987	1.49994	1.50001
Z-05	1.48917	1.47332	1.50004	1.50002	1.50005	1.50006
Z-06	1.49382	1.46522	1.49948	1.49946	1.49952	1.49951
Z-07	1.50011	1.47678	1.50008	1.49989	1.50009	1.50018
Z-08	1.50212	1.50813	1.49975	1.49973	1.49977	1.49988
Z-09	1.50589	1.49210	1.50026	1.50002	1.50020	1.50025
Z-10	1.50486	1.50288	1.49970	1.49987	1.49965	1.49964
Average	1.49854	1.49598	1.49998	1.49993	1.49996	1.49997
Std Dev	0.00755	0.01569	0.00026	0.00027	0.00026	0.00026
Min	1.48157	1.46522	1.49935	1.49905	1.49924	1.49932
Max	1.51716	1.53509	1.50076	1.50048	1.50060	1.50054
Range	0.03559	0.06987	0.00141	0.00143	0.00136	0.00122

Table 3.3.4

	Indicated Rating Factors for Territory 3					
Sample	**GLM 6Yr**	**GLM 3Yr**	**GLM AYC**	**GLM full**	**GEE AR**	**GEE Ex**
X-01	0.80010	0.80006	0.80010	0.80002	0.80008	0.80009
X-02	0.79978	0.80021	0.80012	0.80011	0.80013	0.80016
X-03	0.80027	0.80159	0.80020	0.80014	0.80019	0.80017
X-04	0.79964	0.79893	0.79978	0.79982	0.79981	0.79983
X-05	0.79959	0.80026	0.80022	0.80020	0.80024	0.80026
X-06	0.80058	0.79988	0.80002	0.79997	0.80001	0.80002
X-07	0.79895	0.79799	0.80004	0.79988	0.80001	0.80004
X-08	0.80119	0.80236	0.80008	0.80005	0.80011	0.80010
X-09	0.80050	0.79748	0.80002	0.80003	0.80001	0.79999
X-10	0.80284	0.80567	0.80004	0.80007	0.80004	0.80002
Y-01	0.80008	0.80014	0.80008	0.80009	0.80008	0.80007
Y-02	0.80474	0.81198	0.80020	0.80025	0.80022	0.80022
Y-03	0.79996	0.79703	0.80004	0.80004	0.80005	0.80005
Y-04	0.80056	0.81225	0.79999	0.79999	0.79999	0.79998
Y-05	0.79223	0.78969	0.79993	0.79991	0.79992	0.79992
Y-06	0.80515	0.80170	0.80005	0.80013	0.80005	0.80003
Y-07	0.79247	0.79569	0.79997	0.79992	0.79996	0.79995
Y-08	0.81657	0.82273	0.80005	0.80004	0.80006	0.80009
Y-09	0.80232	0.80505	0.80002	0.79998	0.80003	0.80006
Y-10	0.80437	0.81305	0.80009	0.80014	0.80009	0.80007
Z-01	0.80078	0.80073	0.80078	0.80047	0.80073	0.80071
Z-02	0.79741	0.78948	0.79980	0.79989	0.79974	0.79977
Z-03	0.79353	0.78907	0.79999	0.79989	0.80000	0.80003
Z-04	0.80211	0.80804	0.80009	0.80003	0.80003	0.80003
Z-05	0.80568	0.81250	0.79969	0.79969	0.79972	0.79977
Z-06	0.79660	0.79430	0.79915	0.79923	0.79908	0.79895
Z-07	0.79043	0.77850	0.80080	0.80080	0.80080	0.80082
Z-08	0.79660	0.80615	0.79973	0.79970	0.79974	0.79977
Z-09	0.79301	0.79431	0.79983	0.80021	0.79987	0.79988
Z-10	0.80309	0.81325	0.79956	0.79970	0.79949	0.79945
Average	0.80004	0.80134	0.80002	0.80001	0.80001	0.80001
Std Dev	0.00504	0.00897	0.00030	0.00027	0.00031	0.00032
Min	0.79043	0.77850	0.79915	0.79923	0.79908	0.79895
Max	0.81657	0.82273	0.80080	0.80080	0.80080	0.80082
Range	0.02614	0.04423	0.00165	0.00157	0.00172	0.00187

Table 3.3.5

	Indicated Rating Factors for Driver Class 2					
Sample	GLM 6Yr	GLM 3Yr	GLM AYC	GLM full	GEE AR	GEE Ex
X-01	2.00009	1.99986	2.00009	2.00004	2.00012	2.00014
X-02	2.00066	2.00230	1.99982	1.99989	1.99986	1.99985
X-03	1.99854	1.99650	2.00006	2.00007	2.00006	2.00005
X-04	2.00077	1.99887	2.00003	1.99996	2.00005	2.00009
X-05	2.00279	2.00338	1.99985	1.99990	1.99989	1.99991
X-06	1.99924	2.00293	1.99977	1.99975	1.99979	1.99979
X-07	1.99568	1.99215	2.00015	2.00013	2.00014	2.00016
X-08	1.99883	1.99855	2.00006	2.00004	2.00005	2.00007
X-09	2.00018	1.99959	1.99974	1.99969	1.99974	1.99975
X-10	2.00015	1.99842	2.00013	2.00013	2.00017	2.00020
Y-01	2.00006	2.00005	2.00007	2.00002	2.00007	2.00009
Y-02	1.97740	1.97396	1.99992	1.99996	1.99993	1.99993
Y-03	1.98442	1.99056	2.00009	2.00014	2.00011	2.00014
Y-04	2.01840	2.04209	2.00028	2.00028	2.00027	2.00028
Y-05	1.98411	1.99068	1.99987	1.99986	1.99987	1.99987
Y-06	1.99856	1.99691	1.99970	1.99975	1.99971	1.99968
Y-07	1.99981	1.99845	1.99983	1.99983	1.99983	1.99984
Y-08	1.97674	1.96067	1.99988	2.00000	1.99988	1.99985
Y-09	1.97072	1.97999	1.99995	2.00000	1.99995	1.99996
Y-10	1.98134	1.98471	2.00005	2.00008	2.00005	2.00007
Z-01	2.00018	2.00082	2.00018	1.99964	2.00001	1.99996
Z-02	1.99715	1.97901	2.00017	1.99982	1.99999	2.00000
Z-03	2.00068	1.99935	2.00027	2.00033	2.00028	2.00029
Z-04	2.00443	1.99737	1.99930	1.99991	1.99926	1.99908
Z-05	1.97562	1.96000	1.99954	1.99945	1.99958	1.99958
Z-06	1.97755	1.95312	1.99959	1.99968	1.99968	1.99964
Z-07	2.00170	1.99607	1.99935	1.99957	1.99943	1.99957
Z-08	1.97254	1.96864	1.99958	1.99958	1.99950	1.99936
Z-09	1.98454	1.94168	1.99926	1.99915	1.99917	1.99920
Z-10	1.98520	1.97520	2.00101	2.00086	2.00103	2.00110
Average	1.99294	1.98940	1.99992	1.99992	1.99992	1.99992
Std Dev	0.01170	0.01931	0.00035	0.00031	0.00035	0.00037
Min	1.97072	1.94168	1.99926	1.99915	1.99917	1.99908
Max	2.01840	2.04209	2.00101	2.00086	2.00103	2.00110
Range	0.04768	0.10041	0.00175	0.00171	0.00186	0.00202

Table 3.3.6

	Indicated Rating Factors for Driver Class 3					
Sample	GLM 6Yr	GLM 3Yr	GLM AYC	GLM full	GEE AR	GEE Ex
X-01	0.75033	0.75020	0.75033	0.75022	0.75031	0.75035
X-02	0.74927	0.74938	0.74990	0.74997	0.74989	0.74987
X-03	0.74992	0.74999	0.75041	0.75039	0.75040	0.75042
X-04	0.75011	0.74824	0.75026	0.75015	0.75026	0.75031
X-05	0.75116	0.75143	0.75005	0.75002	0.75003	0.75006
X-06	0.75119	0.75154	0.74996	0.74991	0.74993	0.74992
X-07	0.74900	0.74611	0.75025	0.75023	0.75023	0.75024
X-08	0.75114	0.75338	0.75010	0.75009	0.75010	0.75010
X-09	0.75032	0.75210	0.75000	0.75003	0.75001	0.75004
X-10	0.74825	0.74701	0.74994	0.74999	0.74994	0.74994
Y-01	0.75017	0.75013	0.75017	0.75015	0.75017	0.75018
Y-02	0.75196	0.75315	0.75008	0.75004	0.75006	0.75005
Y-03	0.75468	0.76405	0.74997	0.74997	0.74996	0.74994
Y-04	0.75830	0.76694	0.75007	0.75005	0.75006	0.75004
Y-05	0.73462	0.72896	0.75014	0.75010	0.75015	0.75020
Y-06	0.76605	0.77589	0.74993	0.74997	0.74994	0.74993
Y-07	0.75268	0.75035	0.75004	0.75009	0.75004	0.75002
Y-08	0.73668	0.72737	0.74982	0.74983	0.74982	0.74982
Y-09	0.74774	0.73586	0.74995	0.74992	0.74994	0.74994
Y-10	0.74138	0.74617	0.75010	0.75009	0.75011	0.75011
Z-01	0.74969	0.74946	0.74969	0.74972	0.74964	0.74959
Z-02	0.75050	0.75676	0.74997	0.75015	0.75003	0.74995
Z-03	0.74261	0.73244	0.74954	0.74963	0.74956	0.74957
Z-04	0.74387	0.73412	0.74948	0.74961	0.74953	0.74961
Z-05	0.73876	0.74166	0.74960	0.74985	0.74964	0.74954
Z-06	0.73622	0.72883	0.75042	0.75055	0.75044	0.75038
Z-07	0.74471	0.74383	0.75039	0.75055	0.75041	0.75044
Z-08	0.74853	0.74874	0.75010	0.75020	0.75006	0.74996
Z-09	0.74277	0.74448	0.74966	0.74974	0.74971	0.74971
Z-10	0.74335	0.74649	0.75037	0.75014	0.75043	0.75053
Average	0.74787	0.74750	0.75002	0.75005	0.75003	0.75003
Std Dev	0.00659	0.01083	0.00025	0.00023	0.00025	0.00026
Min	0.73462	0.72737	0.74948	0.74961	0.74953	0.74954
Max	0.76605	0.77589	0.75042	0.75055	0.75044	0.75053
Range	0.03143	0.04852	0.00094	0.00094	0.00091	0.00099

3.4 Accident Year as a Control (Dummy or Nuisance) Variable

What happens when we use accident year as a control variable? We get an indicated factor for each accident year that combines trend and development effects. For State X, which has 0% trend, the factors are 1.00 for accident years 2005 through 2007, 0.80 for 2008 (80% of claims paid or 1.25 development factor) and 0.50 for 2009 (50% of claims paid or 2.00 development factor). For States Y and Z, which have 3% trend, the factors are 1.03 for 2005, 1.03^2 for 2006, 1.03^3 for 2007, 1.03^4 ×

0.80 for 2008, and $1.03^5 \times 0.50$ for 2009. Thus, the trend and development factors could be derived from the accident year parameters. Nevertheless, it would be preferable to model them explicitly as shown in Section 3.3 for the GLM full, GEE AR, and GEE Ex models. Following are the accident year parameters for the GLM AYC model with accident year as control variable.

Table 3.4.1

	Indicated Accident Year Control Factors				
Sample	2005	2006	2007	2008	2009
X-01	1.00016	1.00010	0.99991	0.80016	0.50005
X-02	0.99983	0.99995	0.99983	0.80006	0.49987
X-03	1.00018	1.00013	1.00012	0.80018	0.50005
X-04	1.00032	1.00020	1.00014	0.80019	0.50011
X-05	1.00018	1.00018	0.99977	0.80007	0.49998
X-06	1.00001	1.00009	0.99995	0.79997	0.50011
X-07	1.00022	1.00007	1.00009	0.80001	0.50004
X-08	1.00005	1.00002	0.99983	0.80000	0.50001
X-09	0.99999	1.00018	1.00004	0.79992	0.49997
X-10	1.00032	1.00021	1.00018	0.80012	0.49999
Average	1.00013	1.00011	0.99999	0.80007	0.50002
Std Dev	0.00016	0.00008	0.00015	0.00009	0.00007
Min	0.99983	0.99995	0.99977	0.79992	0.49987
Max	1.00032	1.00021	1.00018	0.80019	0.50011
Range	0.00049	0.00026	0.00041	0.00027	0.00024
Y-01	1.02995	1.06083	1.09256	0.90034	0.57961
Y-02	1.03005	1.06105	1.09284	0.90055	0.57973
Y-03	1.03008	1.06092	1.09273	0.90055	0.57965
Y-04	1.03000	1.06084	1.09261	0.90049	0.57963
Y-05	1.03005	1.06085	1.09274	0.90027	0.57961
Y-06	1.03019	1.06108	1.09301	0.90051	0.57969
Y-07	1.03016	1.06102	1.09268	0.90051	0.57969
Y-08	1.03000	1.06105	1.09291	0.90042	0.57975
Y-09	1.02998	1.06085	1.09248	0.90031	0.57962
Y-10	1.03006	1.06093	1.09275	0.90044	0.57972
Average	1.03005	1.06094	1.09273	0.90044	0.57967
Std Dev	0.00008	0.00010	0.00016	0.00010	0.00005
Min	1.02995	1.06083	1.09248	0.90027	0.57961
Max	1.03019	1.06108	1.09301	0.90055	0.57975
Range	0.00024	0.00025	0.00053	0.00028	0.00014

	Indicated Accident Year Control Factors				
Sample	2005	2006	2007	2008	2009
Z-01	1.03016	1.06112	1.09339	0.90049	0.57971
Z-02	1.03020	1.06086	1.09261	0.89975	0.57907
Z-03	1.02959	1.05993	1.09240	0.89959	0.57929
Z-04	1.02926	1.06157	1.09253	0.89975	0.57948
Z-05	1.03017	1.06099	1.09287	0.90103	0.57993
Z-06	1.02962	1.06147	1.09307	0.90042	0.57999
Z-07	1.03019	1.06115	1.09327	0.90025	0.57953
Z-08	1.02930	1.06042	1.09280	0.90002	0.57956
Z-09	1.02966	1.06144	1.09347	0.90176	0.57975
Z-10	1.03018	1.06104	1.09301	0.90066	0.58002
Average	1.02983	1.06100	1.09294	0.90037	0.57963
Std Dev	0.00039	0.00050	0.00037	0.00067	0.00031
Min	1.02926	1.05993	1.09240	0.89959	0.57907
Max	1.03020	1.06157	1.09347	0.90176	0.58002
Range	0.00094	0.00164	0.00107	0.00217	0.00095

3.5 Quasi-Likelihood Information Criterion

The quasi-likelihood information criterion (QIC) provides a means for choosing between working correlation assumptions for GEE models. A model with a lower QIC is preferable. Based on the QIC results, the GEE with autoregressive correlation fits the synthetic data slightly better than the GEE with exchangeable correlation. As mentioned at the beginning of section 3.3, arguments can be made for using an autoregressive working correlation when the repeated measures are cumulative claim counts at different evaluation ages, so the results are not surprising.

Table 3.5.1

State	Scenario	GEE AR	GEE Ex	Smaller
X	01	368,314.52	368,314.76	AR
X	02	368,739.38	368,739.62	AR
X	03	369,161.31	369,161.55	AR
X	04	369,083.32	369,083.56	AR
X	05	368,673.52	368,673.76	AR
X	06	367,446.39	367,446.62	AR
X	07	367,186.50	367,186.74	AR
X	08	366,262.28	366,262.52	AR
X	09	367,637.24	367,637.47	AR
X	10	368,568.34	368,568.58	AR
Y	01	538,890.02	538,890.25	AR
Y	02	530,240.61	530,240.84	AR
Y	03	531,000.33	531,000.55	AR
Y	04	538,734.71	538,734.93	AR
Y	05	534,610.13	534,610.36	AR
Y	06	536,405.47	536,405.70	AR
Y	07	540,098.23	540,098.45	AR
Y	08	530,921.97	530,922.20	AR
Y	09	540,988.29	540,988.52	AR
Y	10	531,617.72	531,617.95	AR
Z	01	130,980.23	130,980.46	AR
Z	02	125,526.50	125,526.72	AR
Z	03	125,799.49	125,799.72	AR
Z	04	126,938.68	126,938.91	AR
Z	05	125,043.45	125,043.68	AR
Z	06	125,115.15	125,115.37	AR
Z	07	125,787.05	125,787.27	AR
Z	08	126,253.36	126,253.59	AR
Z	09	127,081.81	127,082.04	AR
Z	10	127,521.55	127,521.78	AR

SAS PROC GENMOD also calculates the QICu. This is an approximation to the QIC that can be used to choose between models, but it is not appropriate for choosing between working correlations. The theory of quasi-likelihood functions and the details of the QIC are beyond the scope of this paper. Interested readers are encouraged to consult McCullagh and Nelder's *Generalized Linear Models*, Hardin and Hilbe's *Generalized Estimating Equations*, or Pan's *Akaike's Information Criterion in Generalized Estimating Equations*. For complete bibliographical information see the references section.

3.6 Covariance Matrices

The REPEATED statement in SAS PROC GENMOD has the options MCOVB and ECOVB. When these options are used, the procedure outputs both the model-based (also called naïve) covariance matrix and the empirical (also called robust) covariance matrix for the model's parameters. If these two matrices are similar, it is a sign that the choice of working correlation matrix is adequate. If they are substantially different, then a different working correlation structure would be more appropriate. The GEE models used in this paper had eight parameters, corresponding to the variables and classification levels listed below. Therefore, each covariance matrix is an 8x8 matrix, and with thirty simulations and two models there are thirty pairs of matrices to compare. Unfortunately, there is no automated way of doing this. It requires visual inspection and judgment.

Table 3.6.1

Parameter	Effect	class	Territory	eval_date
Prm1	Intercept			
Prm2	time_index			
Prm3	territory		1	
Prm4	territory		3	
Prm5	driver_class	2		
Prm6	driver_class	3		
Prm7	eval_date			12
Prm8	eval_date			24

Both the autoregressive and exchangeable working correlations resulted in models where the model-based and empirical correlation matrices were similar. This implies that both the autoregressive and exchangeable working correlations result in models that fit the simulated data reasonably well. As mentioned in section 3.5, however, the autoregressive correlation is preferable both in terms of the QIC results as well as from an intuitive understanding of development factors. The covariance matrices are not listed in this paper, but they can be downloaded from the CAS web site as Excel files.

3.7 Confidence Intervals

SAS PROC GENMOD provides confidence intervals for the parameter estimates for the models discussed in this paper. This output will be available for downloading from the CAS web site. A potential use of these confidence intervals would be to develop risk loads to take into consideration when developing final rates, or as input to an enterprise risk management model, but such topics are beyond the scope of this paper.

3.8 Quarterly Data

For simplicity, the examples and database used in this paper used accident year data. A company may decide to design a database containing accident quarters instead. When working with quarterly data one of the considerations is the effect of seasonality. A common way to deal with seasonality is to work with 12-month rolling averages. This would tend to complicate the calculations needed to produce the input files for GLM or GEE models taking into account trend, development and classification. A more simple solution would be to use control variables to account for seasonal differences in accident quarters. Dickmann and Merz did so in a paper about loss trend.[20]

3.9 Why Go Back When We Can Go Forward?

This paper has shown that multivariate frequency models incorporating all available information are resistant to changes in exposure level and changes in distribution of exposures. A next step would be to examine the resistance of different models to things that can affect estimates of ultimate claim severity and ultimate losses such as changes in loss payment patterns or changes in reserving practices.

Once we have multivariate estimates of trend and loss development, should we go back and apply them to total losses by accident year to perform a statewide indication? Why not use the estimates of trend, development, territory relativities and driver class relativities to calculate prospective loss costs directly? For example, the parameter estimates (coefficients) for the State Y, Scenario 1 GEE AR frequency model result in the equation:

$$\ln(E[f]) = -2.75436 + 0.4056T_1 - 0.2230T_3 + 0.6932C_2 - 0.2875C_3 + 0.0295t.$$

From which it follows that the expected frequency is:

$$E[f] = 0.06365 \times 1.500^{T_1} \times 0.800^{T_3} \times 2.000^{C_2} \times 0.750^{C_3} \times 1.030^{t}.$$

Where T_1 and T_3 are variables that take the value 0 or 1 depending on whether or not a policy is from Territory 1 or Territory 3. Similarly, C_2 and C_3 are 0 or 1 depending on whether or not the

[20] Dickmann, Kurt S., and James R. Merz, "Consideration in Estimating Loss Cost Trends," Casualty Actuarial Society *Forum*, Winter 2001, pp. 21-60.

driver classification code is 1 or 3, and t is a time index that increases by 1 every year. The parameters corresponding to the percentage of claims paid by 12 months and 24 months have been omitted since we are only interested in the ultimate claim frequency. By picking an appropriate value for the time index we can project the expected frequency appropriate for the future period in which rates will be in effect. A similar equation can be determined for claim severity as well as pure premium. Thus it is possible to obtain four estimates of prospective loss costs:

- Prospective Frequency × Prospective Severity based on paid data
- Prospective Loss Costs based on paid data
- Prospective Frequency × Prospective Severity based on reported data
- Prospective Loss Costs based on reported data

4. CONCLUSIONS

By organizing data as illustrated in Section 2.3 we can easily fit univariate and multivariate models for both time-dependent effects, such as loss trend and loss development, as well as classification effects such as territory and driver class.

Univariate models of loss trend can over- or underestimate the trend when there are significant changes in the level or in the distribution of exposures.

Modeling trend and development explicitly is preferable to using accident year as a control or dummy variable.

Multivariate models that incorporate all the available information — differences across accident years such as trend and loss development, and differences among classification groups — are resistant to changes in exposure level and changes in exposure distribution.

Acknowledgments

I would like to thank Kurt Dickmann and Phil Baum for introducing me to basic ratemaking techniques. Thanks are due to Lisa Monard, who allowed me to work on driver classification relativity reviews and vehicle rating factors for liability, medical payments and no-fault, and to Frank Gribbon who recommended reading Stephen Mildenhall's paper on "Minimum Bias and Generalized Linear Models" back in the late 1990s. I am also grateful to Rob Curry for the opportunity to work on predictive modeling, and to Fred Klinker, John Baldan, and members of the CAS Ratemaking Committee for their helpful comments. I am also grateful to my wife, Emelda, for her patience and support as I labored on the simulations and programs that formed the basis for this paper.

Supplementary Material

The synthetic datasets used as inputs for the models in this paper, and the covariance matrices output by the generalized estimating equation models are stored electronically on the CAS Web Site and available for downloading. SAS code is provided in the appendices.

Appendix A – Claim Emergence Simulations

The synthetic data sets used in this paper were created using a two step process: (1) generate an exposure scenario, and (2) generate paid claim counts. The synthetic data consist of 30 scenarios divided into three hypothetical states (X, Y, and Z) with 10 scenarios per state. They are meant to approximate what one might see for a short tail line of business such as personal auto property damage liability.

The objective of the States X, Y, and Z simulations was to test the sensitivity of models to changes in exposure level and changes in exposure distribution. In order to achieve that goal it was necessary to find a claim count generation process that approximated as much as possible the expected claim counts, leaving changes in exposure level and exposure distribution as the predominant sources of variation.

The first step was to generate an exposure scenario. This was accomplished by preparing an input file with total exposures per year, and the percentage of exposures corresponding to each combination of territory and driver class, as shown below for State X, Scenario 01.

Table A.1

Calendar Accident Year	Earned Car Years	Terr 1 Class 1	Terr 1 Class 2	Terr 1 Class 3	Terr 2 Class 1	Terr 2 Class 2	Terr 2 Class 3	Terr 3 Class 1	Terr 3 Class 2	Terr 3 Class 3
2002	160000	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2003	176800	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2004	198016	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2005	215837	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2006	232025	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2007	225064	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2008	211560	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04
2009	192520	0.17	0.08	0.07	0.23	0.14	0.10	0.12	0.05	0.04

In the example above, the objective was to generate claim counts in which the predominant source of variation was the change in exposure level. Hence the distribution of exposures was kept constant across accident years. For other State X scenarios the percentage was allowed to change across years by roughly one-tenth of one percent. So for example, in some years the Territory 1

Class 1 percentage may have been 17.1% while in others it might have been 16.9%. The reason for the small allowed change in exposure distribution was that the main objective of the State X scenarios was to test the effect of changes in overall exposure level. On the other hand, for State Y Scenario 02, the percentage of exposures for Territory 1 Class 1 was allowed to decrease from 15.9% in 2002 to 13.0% in 2009. For States Y and Z, both the level of exposures and the distribution of exposures were allowed to change.

The percentage for a territory and driver class combination was multiplied times the total exposures, to get the subtotal corresponding to that combination. For example, in Table A.1 above, 10% of exposures correspond to Territory 2 Class 3, and the total accident year 2002 earned exposures are 160,000. Therefore, 16,000 exposures correspond to Territory 2 Class 3. The process generated 16,000 records with one earned car-year each. This is not entirely realistic, since for most companies some policies are cancelled midyear. However, midyear cancellations are a small proportion of the book for most companies. The following SAS code excerpt illustrates the process of generating the exposure records.

```
do territory = 1 to 3;
  do class = 1 to 3;
   exposure_percentage = exposure_portion{ territory, class };
   car_years = round( exposure * exposure_percentage , 1 );
   do k = 1 to car_years;
    policy_id = put( territory, z1. ) || put( class, z1. ) || put( k, z7. );
    earned_exposure = 1;
    output;
   end;
  end;
 end;
```

For states X, Y, and Z, the next step was to calculate the expected claim counts for each accident year, territory, driver class and evaluation age based on the parameters selected for base frequency, territory relativities, driver class relativities, and percentage of claims paid (closed with payment) at each evaluation age for each accident year. Table A.2 shows the parameters selected for State X, Scenario 01. For example, for Territory 2 Class 3 as of 12 months the expected claim count is 16,000 exposures × 0.05 base frequency × 1.00 Territory 2 factor × 0.75 Class 3 factor × 0.50 percentage reported as of 12 month evaluation age = 300.

Table A.2

Calendar Accident Year	Base Frequency	Terr 1	Terr 2	Terr 3	Class 1	Class 2	Class 3	Age 12	Age 24	Age 36
2002	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2003	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2004	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2005	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2006	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2007	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2008	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20
2009	0.05	1.50	1.00	0.80	1.00	2.00	0.75	0.50	0.30	0.20

The expected claim counts were stored in a variable called _NSIZE_. This is a special variable used by PROC SURVEYSELECT to determine how many records to select from each stratum (accident year, territory, driver class, and evaluation age combination). The amount is rounded to the nearest whole number because claim counts are whole numbers.

```
_NSIZE_ = round( earned_exposure * base_frequency * terr_factor * class_factor
  * age_percentage , 1 );
```

The expected claim counts are used to randomly select policies with replacement from each stratum using SAS PROC SURVEYSELECT. Policies not selected are considered to have zero claims, those selected one or more times are considered to have one or more claims. The number of hits determines the number of claims.

```
proc surveyselect data=for_selection ( drop = earned_exposure )
 out=work.policies_with_claim
 method=urs
 sampsize=work.expected_claim_counts
  ( index = ( ytcpa = ( year territory class age policy_id ) ) )
;
 strata year territory class age;
 id year territory class age policy_id ;
run;
```

The 30 synthetic data sets in policy detail, as well as summaries by territory, driver class, accident year, and evaluation age, are available for downloading from the Casualty Actuarial Society's Web Site.

Appendix B – Calendar Year Trend

Calendar year calculations were used in this paper only to illustrate the distorting effect of changing exposure levels on calendar year trend analysis. Only claim frequency trend examples were provided, but the phenomenon occurs for claim severity trend and pure premium trend as well. For more details, the reader should refer to the CAS "Basic Ratemaking" electronic textbook or to Chris Styrsky's paper "The Effect of Changing Exposure Levels on Calendar Year Trend," which are listed in the references. To prepare for calendar year trend analysis, diagonals must be subtracted to determine the claims paid during the year. It is also necessary to include only complete calendar years in the analysis. Since calendar year claims may relate to prior accident years, the database may not have enough prior accident years to get a complete calendar year. For example, if the database includes accident years 2002 through 2009 and it takes three years for an accident year to develop to ultimate, then the first calendar year for which complete claim counts can be calculated is 2004. This means that only six complete calendar years can be calculated from the database. This number can be specified in a macro variable to ensure that only complete calendar years are used. See SAS code below.

```
%let state = X;
%let scenario = 01;
%let complete_cal_years = 6;

* calculate difference in diagonals ;
data for_cal_yr_trend;
 set mylib.state&state.&scenario.d;
 retain last_cal_year 0;
 cal_year = year;
 paid_count = paid_count12;
 output;
 if paid_count24 not = . then do;
  earned_exposure = 0;
  paid_count = paid_count24 - paid_count12;
  cal_year = year + 1;
  output;
 end;
 if paid_count36 not = . then do;
  earned_exposure = 0;
  paid_count = paid_count36 - paid_count24;
  cal_year = year + 2;
  output;
 end;
 if last_cal_year < cal_year then do;
  last_cal_year = cal_year;
  call symput('last_cal_year',put(last_cal_year,4.));
 end;
run;

* sum diagonal differences corresponding to each calendar years ;
* include only calendar years with a comlete set of differences ;
%let first_year = %eval(&last_cal_year. - &complete_cal_years. + 1);
proc summary nway missing data=for_cal_yr_trend
```

```
 ( where = ( cal_year not < &first_year. ) );
 class cal_year;
 var earned_exposure paid_count;
 output out=cal_yr sum=;
run;

* calculate claim frequency for each calendar year ;
data cal_yr_freq;
 set cal_yr ( drop = _type_ _freq_ );
 claim_frequency = paid_count / earned_exposure;
 log_claim_frequency = log ( claim_frequency );
 time_index = cal_year - &last_cal_year.;
run;

title "Calendar Year Frequency Trend Data, State &state., Scenario &scenario.";
proc print label noobs data=cal_yr_freq split='_';
 var cal_year earned_exposure paid_count
  claim_frequency log_claim_frequency;
 format earned_exposure paid_count comma9.
  claim_frequency log_claim_frequency 7.5;
 label cal_year = 'Calendar Year'
  earned_exposure = 'Earned Car Years'
  paid_count = 'Paid Claim Count'
  claim_frequency = 'Claim Frequency'
  log_claim_frequency = 'Log of Claim Frequency';
run;

* fit exponential regression model;
proc reg data=cal_yr_freq outest=cy_trend;
 trend_model: model log_claim_frequency = time_index / noprint; run;
quit;

* calculate annual trend based on model output ;
data freq_factor;
 set cy_trend;
 time_index = round(time_index, 0.0000001 );
 trend_factor = round( exp( time_index ), 0.0000001 );
 annual_trend = trend_factor - 1;
 format time_index trend_factor 10.7 annual_trend percentn7.2;;
 label trend_factor = 'Annual Trend Factor' time_index = 'Time Index Parameter'
  annual_trend = 'Annual Trend';
 keep time_index trend_factor annual_trend;
run;

title "Calendar Year Frequency Trend, State &state., Scenario &scenario.";
proc print noobs label data=freq_factor;
run;
```

Appendix C – Chain Ladder Development and Accident Year Trend

To prepare for chain ladder claim count development, the claim counts were summarized by accident year. Next, age-to-age factors were calculated based on the claim count triangle. The average of all years was calculated and used as the selected link ratio. In practice, an actuary might select a different loss development factor based on knowledge of the book of business, changes in claim practices, or other information. However, this is just simulated data, so the only factor affecting the data is random variation.

The age-to-ultimate link ratios are the cumulative product of the selected link ratios, and the percentage reported estimates are the reciprocal of the age to ultimate factors. The SAS code below illustrates this process.

```
%let state = S;
%let scenario = 01;

* calculate claim triangle ;
proc summary nway missing data=mylib.state&state.&scenario.d;
 class year;
 var earned_exposure paid_count12 paid_count24 paid_count36 paid_count48;
 output out=claim_triangle sum=;
run;

* calculate age to age factors ;
data age_to_age;
 set claim_triangle;
 call symput('last_year',put(year,4.)); * macro var to be used for evaluation date ;
 if paid_count24 > 0 then f12 = round( paid_count24 / paid_count12, 0.0000001 );
 else delete;
 if paid_count36 > 0 then f24 = round( paid_count36 / paid_count24, 0.0000001 );
 if paid_count48 > 0 then do;
  f36 = round( paid_count48 / paid_count36, 0.0000001 );
  f48 = 1;
 end;
 format f12 f24 f36 f48 10.7;
 length factor_type $ 16;
 factor_type = 'Age to Age';
 label factor_type = 'Factor Type' f12 = 'Age 12' f24 = 'Age 24'
  f36 = 'Age 36' f48 = 'Age 48';
run;

* calculate average of all available years ;
proc summary nway missing data=age_to_age;
 var f12 f24 f36 f48;
 output out=z_averages mean=;
run;

* calculate age to ultimate link ratios and percentage reported ;
data link_ratios;
 set age_to_age( in = a keep = factor_type year f12 f24 f36 f48 )
  z_averages ( keep = f12 f24 f36 f48 );
 if a then output;
 else do;
  factor_type = 'All-Year Average';
  f12 = round( f12, 0.0000001 );
  f24 = round( f24, 0.0000001 );
  f36 = round( f36, 0.0000001 );
  f48 = round( f48, 0.0000001 );
  output;
  factor_type = 'Age to Ultimate';
  f12 = round( f12 * f24 * f36 * f48 , 0.0000001 );
  f24 = round( f24 * f36 * f48 , 0.0000001 );
  f36 = round( f36 * f48 , 0.0000001 );
  output;
  factor_type = 'Percent Reported';
  f12 = round( 1 / f12, 0.0000001 );
  f24 = round( 1 / f24, 0.0000001 );
  f36 = round( 1 / f36, 0.0000001 );
  f48 = round( 1 / f48, 0.0000001 );
  output;
 end;
run;
```

Ultimate claim counts are the product of claim counts reported as of the last evaluation times the age-to-ultimate factor. The estimated ultimate claim counts are used to calculate the claim frequency for each year. Then an exponential regression is fit to these claim frequencies to determine the claim frequency trend. See SAS code below.

```
data developed_claims;
 set claim_triangle;
 if _n_ = 1 then set link_ratios ( drop = year where = ( factor_type = 'Age to Ultimate' ) ) ;
 time_index = year - &last_year.;
 eval_year = &last_year.;
 eval_date = '12/31/'||put(eval_year,4.);
 select ( year );
  when ( &last_year.     ) do; age_to_ult = f12; cumulative_claims = paid_count12; end;
  when ( &last_year. - 1 ) do; age_to_ult = f24; cumulative_claims = paid_count24; end;
  when ( &last_year. - 2 ) do; age_to_ult = f36; cumulative_claims = paid_count36; end;
  otherwise                do; age_to_ult = f48; cumulative_claims = paid_count48; end;
 end;
 ultimate_claims = round( cumulative_claims * age_to_ult, 1 );
 claim_frequency = round( ultimate_claims / earned_exposure, 0.0000001 );
 log_claim_frequency = round( log( claim_frequency ), 0.0000001 );;
 label
  time_index = 'Time Index'
  eval_date = 'Evaluation Date'
  claim_frequency = 'Claim Frequency'
  cumulative_claims = "Reported Claim Count"
  age_to_ult = 'Age to Ultimate Development Factor'
  ultimate_claims = 'Ultimate Claim Count'
 ;
 format earned_exposure comma9. cumulative_claims ultimate_claims comma7.0
  claim_frequency log_claim_frequency age_to_ult 10.7;
 keep eval_date year time_index earned_exposure cumulative_claims age_to_ult
  ultimate_claims claim_frequency log_claim_frequency;
run;

title2 'Claim Frequency Trend Analysis';
proc reg data=developed_claims outest=cf_parms;
 model log_claim_frequency = time_index;
 ods select ParameterEstimates;
 run;
quit;

data freq_factor;
 set cf_parms;
 time_index = round(time_index, 0.0000001 );
 trend_factor = round( exp( time_index ), 0.0000001 );
 annual_trend = trend_factor - 1;
 format time_index trend_factor 10.7 annual_trend percentn7.2;;
 label trend_factor = 'Annual Trend Factor' time_index = 'Time Index Parameter'
  annual_trend = 'Annual Trend';
 keep time_index trend_factor annual_trend;
run;

data for_exhibit;
 merge claim_triangle ( keep = year earned_exposure paid_count12 paid_count24 paid_count36 )
  link_ratios ( keep = factor_type f12 f24 f36 )
  developed_claims ( keep = time_index ultimate_claims );
run;

title "Claims Closed With Payment, State &state., Scenario &scenario.";
proc print data=for_exhibit noobs label;
 var year paid_count12 paid_count24 paid_count36 factor_type f12 f24 f36 ultimate_claims;
 format paid_count12 paid_count24 paid_count36 ultimate_claims comma8.0 f12 f24 f36 8.5;
```

```
 label year = 'Accident Year' paid_count12 = 'Paid Count Age 12'
  paid_count24 = 'Paid Count Age 24' paid_count36 = 'Paid Count Age 36'
 ;
run;

proc print noobs label data=freq_factor;
run;
```

Appendix D – Generalized Linear Models

Four generalized linear models (GLM) were tested for this paper as shown in Table D.1.

Table D.1

GLM 6Yr	Generalized linear model with latest 6 years of data and territory and driver class as independent variables
GLM 3Yr	Generalized linear model with latest 3 years of data and territory and driver class as independent variables
GLM AYC	Generalized linear model with latest 6 years of data, territory and driver class as independent variables, and accident year as control variable
GLM Full	Generalized linear model with latest 6 years of data, territory, driver class, time index, and evaluation age as independent variables

The first three models use only the latest evaluation for each calendar/accident year. The fourth model (GLM Full) uses territory, driver class, a time index (for trend), and evaluation age (for development) as independent variables and claim count as the dependent variable. In order to include evaluation age in the model, it is necessary to transpose the evaluation age columns into rows, and to create a variable to identify the evaluation age. The data are assumed to reach ultimate value at 36 months. Therefore, a policy that has been in force for the entire experience period has three observations for each mature accident year, two for the penultimate accident year, and one for the latest accident year. The 36-month evaluation is the reference level, so the 12-month and 24-month parameters are relativities to the 36-month or ultimate claim count. This means that they correspond to the percentage paid as of 12 or 24 months respectively. The age-to-ultimate development factor is the reciprocal of the percentage reported. The SAS code below performs the data preparation and model fitting.

```
%let state = X;
%let scenario = 01;
%let first_year = 2004;
%let ref_year = 2007;
%let last_year = 2009;
```

```
%let evals = 3;

* prepare data for claim frequency Generalized Linear Models ;
* for territory and driver class only ;
* for territory and driver class with accident year as control variable ;
data sample_glm;
 set mylib.state&state.&scenario.d;
 if year < &first_year. or year > &last_year. then delete;
 select;
  when ( year = &last_year.     ) paid_claim_count = paid_count12;
  when ( year = &last_year. - 1 ) paid_claim_count = paid_count24;
  when ( year = &last_year. - 2 ) paid_claim_count = paid_count36;
  otherwise paid_claim_count = paid_count48;
 end;
 Driver_Class = class;
 log_exposure = log( earned_exposure );
 keep year territory driver_class policy_id earned_exposure
  paid_claim_count log_exposure
 ;
run;

title1 "Paid Claim Frequency GLM for Territory and Driver Class Only";
title2 "State &state, Scenario &scenario., Six Accident Years";
proc genmod data=sample_glm;
 class driver_class ( ref = '1' ) territory ( ref = '2' )
  / param = ref;
 model paid_claim_count = territory driver_class
  / link=log dist=Poisson offset=log_exposure scale=p;
run;

title1 "Paid Claim Frequency GLM for Territory and Driver Class Only";
title2 "State &state, Scenario &scenario., Three Accident Years";
%let starting_year = &Last_Year. - 2 ;
proc genmod data=sample_glm ( where = ( year >= &starting_year. ) );
 class driver_class ( ref = '1' ) territory ( ref = '2' )
  / param = ref;
 model paid_claim_count = territory driver_class
  / link=log dist=Poisson offset=log_exposure scale=p;
run;

title1 "Paid Claim Frequency GLM for Territory and Driver Class";
title2 "With Accident Year as Control Variable";
title3 "State &state, Scenario = &scenario.";
proc genmod data=sample_glm;
 class driver_class ( ref = '1' ) territory ( ref = '2' ) year ( ref = "&ref_year." )
  / param = ref;
 model paid_claim_count = territory driver_class year
  / link=log dist=Poisson offset=log_exposure scale=p;
run;

* prepare data for Generalized Linear Model  ;
* for territory factors, driver class factors ;
* trend factor and loss development factors   ;
data sample_glm2;
 set mylib.state&state.&scenario.d ( where = ( year not < &first_year. ) );
 array paid_cnt {4} paid_count12 paid_count24 paid_count36 paid_count48;
 array eval_dates {4} $ ('12' '24' '36' '48');
 time_index = year - &ref_year.;
 Years = "&First_Year. to &Last_Year.";
 driver_class = class;
 label
  time_index = 'Time Index'
  log_exposure = 'Natural Log of Exposure'
  driver_class = 'Driver Class'
  paid_claim_count = "Paid Claim Count"
 ;
 do k = 1 to &evals.;
  eval_date = eval_dates{k};
```

```
  if k = &evals. then call symput('last_eval',eval_date);
  if paid_cnt{k} > . then do;
   paid_claim_count = paid_cnt{k};
   log_exposure = log ( earned_exposure );
   output;
  end;
 end;
 keep Years year territory driver_class earned_exposure
  log_exposure time_index eval_date paid_claim_count
 ;
run;

title1 "Paid Claim Frequency GLM for State &state, Scenario = &scenario.";
title2 "Territory, Driver Class, Trend and Loss Development";
proc genmod data=sample_glm2;
 class driver_class ( ref = '1' ) territory ( ref = '2' ) eval_date ( ref = "&last_eval." )
  / param = ref;
 model paid_claim_count= territory driver_class time_index eval_date
  / link=log dist=Poisson offset=log_exposure scale=p;
run;
```

Appendix E – Generalized Estimating Equations

The first step in fitting generalized estimating equations (GEE) was to create a modeling sample. The GEE algorithm depends on the definition of a subject or panel. A database with a large number of policies, each treated as a subject or panel, can cause the program to run out of memory on a desktop personal computer. The sampling algorithm shown below first classifies policies depending on whether or not they had a claim reported, even if it was closed without payment. Then it selects the entire six accident year history for the policy. All policies with a reported claim are included in the modeling sample, but only ten percent of the claim-free policies are selected for each territory and driver class combination.

The hypothetical ratemaking database used in this paper has only territory and driver class as rating factors. A real database would have other rating variables. If other variables are used in the rating plan, they should be included in the definition of the strata from which the ten percent samples are taken. They should also be included in the generalized estimating equation, either as predictors or as part of the offset term.

Additionally, sampling weights must be calculated to reflect the original number of observations in the database. The procedure uses the ratio of the original number of observations to the number of observations in the sample. So for policies that had at least one reported claim in the six-year accident history, the weight is 1, and for policies that were claim-free the weight is close to 10. The reason the weight is not always exactly equal to 10 for the claim-free policies is that the original number of observations may not have been a multiple of 10, so ten percent would not have been a whole number. Therefore the nearest whole number of policies had to be selected. SAS procedure SURVEYSELECT was used to select the 10% of claim-free policies. The SURVEYSELECT

procedure calculates sampling weights, but those weights were not used because the input to the procedure was just the list of claim-free policies. Some of them may have had 15 observations (all six-accident years), while other policies may have had less. Therefore, the percentage of policies does not exactly equal the percentage of observations selected, and the weights had to be calculated manually once the entire history for each policy had been selected. The SAS code below performs the sampling procedure and calculates the sampling weights.

```
%let state = X;
%let scenario = 01;
%let first_year = 2004;
%let ref_year = 2007;
%let last_year = 2009;
%let evals = 3;

* prepare data sample containing ;
* * all policies with at least one reported claim ;
* * 1 out of every 10 policies with no reported claim ;

 proc sql noprint;
  * find all policies with at least one reported claim in experience period ;
  create table id_with_claim as select unique policy_id
  from mylib.state&state.&scenario.d
  where ( incurred_loss12 > 0 or incurred_loss24 > 0
   or incurred_loss36 > 0 or incurred_loss48 > 0 )
  order by policy_id;
  * sort data by policy id ;
  create table scenario as select *
  from mylib.state&state.&scenario.d
  where ( year not < &first_year. )
  order by policy_id;
 quit;

 * split data into policies with at least one reported claim and those claim free ;
 data with_claim claim_free;
  merge scenario ( in = a ) id_with_claim ( in = b );
  by policy_id;
  driver_class = class;
  if a and b
  then output with_claim;
  else if a then output claim_free;
 run;

 * determine all combinations of territory, driver ;
 * class, and policy id for claim free policies    ;
 proc sql noprint;
  create table id_claim_free as select unique territory, driver_class, policy_id
  from claim_free
  order by territory, driver_class, policy_id;
 quit;

 * select 10% of the claim-free policies for each territory and driver class ;
 proc surveyselect data=id_claim_free
  out=claim_free_sampled method=SRS rate=0.10;
  STRATA territory driver_class;
 run;

 * now that we have a list of selected policies for each territory ;
 * and driver class, select all the data for those policies ;
 proc sort data=claim_free_sampled ( drop = SelectionProb SamplingWeight );
  by policy_id;
```

```
run;
data claim_free_selected;
 merge claim_free(in = a ) claim_free_sampled( in = b );
 by policy_id;
 if a and b;
run;

* combined the selected claim free data and the data with ;
* at least one reported claim to create modeling sample ;
data selected;
 set claim_free_selected( in = a ) with_claim( in = b );
 if a then with_claim = 0;
 if b then with_claim = 1;
 unity = 1;
run;

* count the number of observations in each stratum in the sample ;
proc summary nway missing data=selected;
 class year territory driver_class with_claim;
 var unity;
 output out=sample_counts sum(unity)=sample_record_count;
run;

* classify all the original data depending on whether
* or not the policy had a reported claim ;
data original;
 set claim_free( in = a ) with_claim( in = b );
 if a then with_claim = 0;
 if b then with_claim = 1;
 unity = 1;
run;

* count the number of observations in each stratum in the original data ;
proc summary nway missing data=original;
 class year territory driver_class with_claim;
 var unity;
 output out=original_counts sum(unity)=original_record_count;
run;

* calculate weights equal to the ratio of the number of observations ;
* in the original data to the number of observations in the sample ;
* for each stratum ;
data sample_weights;
 merge original_counts ( drop = _type_ _freq_ )
  sample_counts ( drop = _type_ _freq_ );
 by year territory driver_class with_claim;
 sampling_weight = original_record_count / sample_record_count;
run;

* merge sampling weights with modeling sample ;
proc sort
 data=selected( drop = unity )
 out=selected_for_merge;
 by year territory driver_class with_claim;
run;
data modeling_sample;
 merge selected_for_merge sample_weights;
 by year territory driver_class with_claim;
run;
```

The next step in preparing the data for GEE is to transpose the accident year evaluation dates into rows. A new variable, eval_date, identifies the evaluation date for each record. Furthermore, the policy ID and year date are concatenated into one variable, policy_id_year, which will be used to

identify each subject; the evaluation age identifies the repeated claim count observations for each policy-id-year subject. For the latest three years, some of the claim counts are missing because they correspond to evaluation dates that will occur in the future. These records are omitted.

```
* prepare data for Generalized Estimating Equations ;
data sample_gee;
 set modeling_sample;
 array paid_cnt {4} paid_count12 paid_count24 paid_count36 paid_count48;
 array eval_dates {4} $ ('12' '24' '36' '48');
 time_index = year - &ref_year.;
 Years = "&First_Year. to &Last_Year.";
 driver_class = class;
 policy_id_year = policy_id || put( year, 4. );
 label
  time_index = 'Time Index'
  log_exposure = 'Natural Log of Exposure'
  driver_class = 'Driver Class'
  paid_claim_count = "Paid Claim Count"
  eval_date = 'Evaluation Date'
  policy_id_year = 'Policy Id and Year'
 ;
 do k = 1 to &evals.;
  eval_date = eval_dates{k};
  if k = &evals. then call symput('last_eval',eval_date);
  if paid_cnt{k} > . then do;
   paid_claim_count = paid_cnt{k};
   log_exposure = log ( earned_exposure );
   output;
  end;
 end;
 keep Years year territory driver_class policy_id_year eval_date
  earned_exposure log_exposure sampling_weight time_index paid_claim_count
 ;
run;
```

Two GEE models are tested in this paper. The first one uses an autoregressive working correlation structure, and the second one uses exchangeable working correlation. The autoregressive correlation structure assumes the correlation between successive evaluation ages is stronger than the correlation between evaluation ages that are further apart. The exchangeable working correlation assumes the correlation between any two of evaluation ages is the same. Following is the SAS code for these two models. Proc TEMPLATE is used to increase the number of decimal places output for the parameter estimate.

```
proc template;
 edit Stat.GENMOD.GEEEst;
  define Estimate;
   header = "Estimate";
   format = 10.6;
  end;
 end;
run;
```

```
title1 "Paid Claim Frequency GEE with Autoregressive Working Correlation";
title2 "For Territory, Driver Class, Trend and Loss Development";
title3 "State &state, Scenario &scenario.";
proc genmod data=sample_gee;
 weight sampling_weight;
 class policy_id_year eval_date
  driver_class ( ref = '1' ) territory ( ref = '2' ) eval_date ( ref = "&last_eval." )
  / param = ref;
 model paid_claim_count= territory driver_class time_index eval_date
  / link=log dist=Poisson offset=log_exposure scale=p;
 repeated subject = policy_id_year
  / withinsubject = eval_date corr=AR corrw mcovb ecovb;
run;

title1 "Paid Claim Frequency GEE with Exchangeable Working Correlation";
title2 "For Territory, Driver Class, Trend and Loss Development";
title3 "State &state, Scenario &scenario.";
proc genmod data=sample_gee;
 weight sampling_weight;
 class policy_id_year eval_date
  driver_class ( ref = '1' ) territory ( ref = '2' ) eval_date ( ref = "&last_eval." )
  / param = ref;
 model paid_claim_count = territory driver_class time_index eval_date
  / link=log dist=Poisson offset=log_exposure scale=p;
 repeated subject = policy_id_year
  / withinsubject = eval_date corr=exch corrw mcovb ecovb;
run;
```

5. REFERENCES

[1.] Anderson, Duncan, and Sholom Feldblum, and Claudine Modlin, and Doris Schirmacher, and Ernesto Schirmacher, and Neeza Thandi, "A Practitioner's Guide to Generalized Linear Models," 3rd ed., CAS Study Note, 2007.

[2.] Hardin, J. W., and J.M. Hilbe, *Generalized Estimating Equations*, New York: Chapman & Hall/CRC, 2003.

[3.] Horton, Nicholas J., and Stuart R. Lipsitz, "Review of Software to Fit Generalized Estimating Equation Regression Models,", *The American Statistician* 53, 1999, pp. 160-169.

[4.] Lo, Chi Ho, and Wing Kam Fung, and Zhong Yi Zhu, "Structural Parameter Estimation Using Generalized Estimating Equations for Regression Credibility Models," *ASTIN Bulletin* 37(2), 2007, pp. 323-343.

[5.] McCullagh, P., "Quasi-likelihood Functions," *Annals of Statistics* 11, 1983, pp. 59–67.

[6.] McCullagh, P., and John A. Nelder, *Generalized Linear Models*, New York: Chapman and Hall,. 1989.

[7.] Molenberghs, Geert, and Geert Verbeke, "Models for Discrete Longitudinal Data," New York: Springer Series in Statistics, Springer Science+Business Media, Inc., 2005.

[8.] Pan, W., "Akaike's Information Criterion in Generalized Estimating Equations," *Biometrics*, 57, 2001, pp. 120-125.

[9.] Mildenhall, S., "Minimum Bias and Generalized Linear Models," *Proceedings of the Casualty Actuarial Society* LXXVI, 1999, pp. 393-487.

[10.] "SAS/STAT® 9.2 User's Guide," 2008, SAS Institute Inc., 2008.

[11.] Werner, Geoff and Claudine Modlin, *Basic Ratemaking,* 3rd ed., Arlington, VA: Casualty Actuarial Society, January 2010.

[12.] Zeger, S. L., K.Y. Liang, K. Y., and P.S. Albert, P. S., "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics* 44, 1988, Vol. 44, pp. 1049–1060.

Abbreviations and notations

GEE, generalized estimating equations
GLM, generalized linear models
QIC, quasi-likelihood information criterion
QICu, Pan's approximation to QIC

Biography of the Author

Hernán L. Medina is senior research associate at the Insurance Services Office, Inc., a Verisk Analytics Company in Jersey City, New Jersey. He graduated magna cum laude from Saint Peter's College with a Bachelor of Science degree, majoring in Mathematics with a minor in Physics. He also has a Master of Science degree in Mathematics from New York University. He is a Charter Property Casualty Underwriter and a member of the CPCU Society.