A Comparative Analysis of Alternative Pure Premium Models in the Automobile Risk Classification System

Beomha Jee


Stable URL: http://links.jstor.org/sici?sici=0022-4367%28198909%2956%3A3%3C434%3AAACAOAP%3E2.0.CO%3B2-7

The Journal of Risk and Insurance is currently published by American Risk and Insurance Association.
A Comparative Analysis of Alternative Pure Premium Models in the Automobile Risk Classification System

Beomha Jee

Abstract

The primary purpose of the article is to provide empirical evidence to the unanswered questions regarding the modeling approach for estimating pure premiums under the cross-classified rating system. The alternative estimation models are compared in terms of their predictive accuracy and adequacy of the underlying distributional assumptions by using the single- or two-period data. Recently developed estimation approaches are used. They include the Box and Cox heteroskedastic model and the empirical Bayes estimation. It has been found that these models substantially improve predictive accuracy. The role of interaction term is found material especially for the Box-Cox model.

In the insurance market, risks are traditionally divided into various classes whose prices differ according to various characteristics of those risks. The goal of a classification system is to group homogeneous risks and charge each group a premium commensurate with the average expected loss of its members. Failure to achieve this goal may lead to adverse selection and, perhaps, moral hazard with welfare losses to consumers and economic losses to insurers. Though it is important to be able to estimate the pure premiums for a class using data for only that class because the experience for any given class may not be sufficiently credible. As a result, a large number of studies have developed and estimated models of pure premiums that utilize data for multiple classes.

The class rating process may be divided into two broad phases: (a) the optimal design of the system, which involves the selection of risk classification factors and the definition of associated risk classes; and (b) the estimation of pure premiums for the risk classes, which involves the choice of an appropriate functional form and estimation method. This article primarily attempts to answer a number of unanswered questions concerning the second

Beomha Jee is Assistant Professor of Insurance at The Pennsylvania State University. The author gratefully acknowledges the helpful suggestions of Scott E. Harrington at the University of South Carolina. The comments of anonymous reviewers and the Associate Editor are also appreciated.
phase: establishing a modeling approach to estimate pure premiums for risk
classes structured into the cross-classified rating system, focusing on some
recently developed estimation models. In addition, the article reviews the
previous modeling work, identifies important questions that remain unan-
swered, and looks for empirical evidence concerning these unanswered
questions using various methods.

**Background**

An early study on modeling of pure premiums within the risk classification
system was conducted by Almer [4], who suggested a multiplicative model for
use with cross-classified data with the following general form:

\[ P_{ij} = P_0a_ib_j + e_{ij} \]  

where \( P_{ij} \) is the claim proportion for class \( ij \); \( P_0 \) is overall mean; \( a_i \) and \( b_j \) are
the effects of the levels \( i \) and \( j \), for rating factors \( a \) and \( b \) respectively; and \( e_{ij} \)
is the error term. Bailey and Simon [6] analyzed loss ratios by comparing the
multiplicative model and the additive model which may be expressed as follows:

\[ P_{ij} = P_0 + a_i + b_j + e_{ij}. \]  

Since these two works were published, a considerable number of studies
have focused on comparing additive and multiplicative models. In addition,
many researchers have suggested various statistical procedures to estimate
model parameters. A set of parameter values may be estimated so that it gives
a best fit of the observed data by minimizing some function of deviations of
actual from expected results. As Bennett [8] suggested, a weighted sum of
squares of deviations can be considered an appropriate function to minimize
under various assumptions about the value of the weight. Bennett also
discussed some alternative sets of weights.

While the iteration procedure was used to derive the parameters for the
multiplicative model in previous studies, a loglinear form of function has been
employed to estimate the multiplicative model in the studies by Chang and
Fairley [11], Sant [41], Fairley, Tomberlin, and Weisberg [19], Weisberg and
Tomberlin [49], and Samson and Thomas [39]. Seal [42] constructed an
additive model using the logit transformation, and Coutts [12] adopted Seal's
approach in a study of automobile premium rating in the United Kingdom.

Chang and Fairley [11] and Fairley, Tomberlin, and Weisberg [19] noted that
a loglinear model tends to overestimate high risks by double-counting
classification factors. The dispute about whether risk classes for high risks are
charged excessive premiums goes back to the findings of Bailey and Simon [6],
where the results of analyzing Canadian automobile insurance data indicated
that the multiplicative models produced systematic overestimates for the

---

1 The two-factor case is used for simplicity to illustrate alternative models, but unidimensional
cases are not discussed.
highest risk merit rating and driver classes. Holmes [24] and Wilcken [52] later debated the existence of this tendency based on Canadian data [23]. Holmes indicated, as a source of bias, treatment of dependent risk classes as if they were independent. He recommended replacement of the differential complexity by an estimation method avoiding the unwanted assumption of independence among risk classes and avoiding double-counting of risk. It is undesirable to use methods that overestimate pure premiums for the highest risk classes because overestimation may cause increased adverse selection.

Some analysts have suggested more complex models that produce the purely additive and multiplicative models as special cases. Bailey and Simon [6] and Chamberlain [10] suggested additive models with multiplicative interaction terms. Chang and Fairley suggested weighting estimates by the additive and multiplicative models. However, Samson and Thomas [39] showed some empirical evidence that such a linearly combined model failed to improve the predictive accuracy substantially. DuMouchel [18] proposed a hybrid functional form expressed as [50]:

\[ P_{ij} = a_i + b_j T_i + e_{ij}, \] (3)

where \( T_i \) is an additional parameter representing the multiplicative factor for class-ij. He estimated pure premiums under this functional form using a Bayesian credibility scheme\(^2\) [18]. From (3), the pure additive case can be obtained by setting the value of \( T_i \) equal to one, and the multiplicative form can be obtained by allowing \( a_i \) to be zero.

Harrington [23] suggested a flexible functional form applying the transformation technique developed by Box and Cox [9]. This model has the following functional form [9, 23]:

\[ P_{ij}^{(c)} = (P_{ij}^c - 1)/c = a_i + b_j + e_{ij}, \] (4)

where \( c \) is the transformation parameter. The values of \( a, b, \) and \( c \) may be obtained by searching within some reasonable range for the value of \( c \) that maximizes the likelihood function under specified assumptions about the distribution of \( e_{ij} \). This model essentially includes the additive (\( c = 1 \)) and multiplicative (\( c = 0 \)) models as special cases. Harrington showed this model's goodness of fit was impressive using cross-sectional data which were cross-classified by risk classification factors, though the effectiveness of a comparable model with time series data has been questioned by Cummins and Harrington [13].

The choice of functional form has been reviewed in Albrecht [3] and two articles by ter Berg [45, 46], and van Eeghen, Greup, and Nijssen [48] from theoretical perspectives. Ajne [2] developed the hypothesis testing procedure in

\(^{2}\text{Credibility theory implies that an estimate for a certain class should reflect the reliability of observed data for each class. Traditional credibility theory estimates pure premium for each cell based on a weighting of the observed mean and overall mean pure premium by using the class size as a weight. The Bayesian approach attempts to combine information from individual experience and pooled group experience in the context of prior and posterior distributions. For detail, see Kahn [28].}
the multiplicative models. Freifelder [21] compared functional forms by predicting the pattern of estimation errors from misselection of form.

Even though Seal [42] and Baxter, Coutts, and Ross [7] showed that interaction terms were not significant by tests of subsets of parameters, some studies have attempted to construct models with some interaction terms and have provided empirical evidence that their models' goodness of fit is better than that of other models without interaction terms [10, 18, 23, 39, 41]. Chamberlain [10] and Samson and Thomas [39] suggested F-tests or stepwise regressions as appropriate methods to select significant interaction terms.

While studies using U.S. data have focused on estimation of pure premium models, separate estimation of frequency and severity models has not been undertaken for ratemaking purposes. A notable exception is Weisberg, Tomberlin, and Chatterjee [50]. They estimated a multiplicative model of claim frequency using the empirical Bayes method, and estimated severity models by applying a traditional ANOVA for both mean severity (additive) and its log transformation (multiplicative) for each class under the assumption of constant within-class variation [47, 50]. Their empirical Bayes method produced estimates based on a weighting of the observed cell means and smoothed estimates obtained by the multiplicative model. The motivation for this procedure is analogous to that underlying credibility estimates by actuaries. This model can be expressed as:

\[
\log E(P_{ij}) = a_j + b_j + c_{ij}, \tag{5}
\]

where \(c_{ij}\) is normally distributed with zero mean rather than a fixed parameter. The model (5) implies that without data analysis using observed values for \(P_{ij}\), it is impossible to get information regarding departure of the model (5) from its reduced form, that is, the loglinear or multiplicative model without interaction terms expressed as (1).

For small cells, the prior distribution of \(c_{ij}\) in (5) will dominate, leading their posterior estimates close to zero. This implies that the estimated claim rates for these classes will be close to those obtained by the multiplicative model. However, estimates for large classes will be close to observed values [50]. This study also compared the performance of several variants of the additive and multiplicative approaches and pure premium estimation relying on individual cell means. Table 1 provides a summary of the suggested functional form, methodology, and data used in previous work for modeling pure premiums.

**Problem Statement**

Previous work on estimation of pure premiums has highlighted the important issue in modeling and estimating risk classification factors: the choice of an appropriate functional form for pure premium models. Studies comparing functional forms have been conducted using different, usually

---

3 European studies on the estimation of claim frequency include Grimes [22], Johnson and Hey [26], Kahane and Levy [29], and Baxter, Coutts, and Ross [7].
Table 1
Previous Work for Pure Premium Modeling

<table>
<thead>
<tr>
<th>Study</th>
<th>Functional Form</th>
<th>Form of Disturbance Variance*</th>
<th>Data</th>
<th>Estimation Method</th>
<th>Major Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almer [4]</td>
<td>Multiplicative</td>
<td>$P_i/n_{ij}$</td>
<td>Swedish (1947-52) 3rd party</td>
<td>Weighted Least Squares</td>
<td>Fine goodness of fit of residuals were generated not by random choice of factors, but by equalization</td>
</tr>
<tr>
<td>Bailey-Simon [6]</td>
<td>Additive</td>
<td>$P_i/Kn_{ij}$</td>
<td>Canadian Liability</td>
<td>Minimum Chi-Squares</td>
<td>Tendency of double counting in the high risk classes from the multiplicative form</td>
</tr>
<tr>
<td>Bailey [5]</td>
<td>Multiplicative</td>
<td>$P_i/Kn_{ij}$</td>
<td>Canadian Liability</td>
<td>Minimum Bias Method</td>
<td>Multiplicative form is suggested if all factors are percents, additive if they are cents</td>
</tr>
<tr>
<td>Jung [27]</td>
<td>Multiplicative</td>
<td>$P_i/Kn_{ij}$</td>
<td>Canadian Liability</td>
<td>Heuristic Method (Modified chi-square minimum)</td>
<td>Minimum chi-square of multiplicative model will be biased</td>
</tr>
<tr>
<td>Ajne [1]</td>
<td>Multiplicative</td>
<td>$P_i/Kn_{ij}$</td>
<td>Simulation</td>
<td>Method of Moments (Modified chi-square minimum)</td>
<td>The estimator may have a positive finite sample bias</td>
</tr>
<tr>
<td>Chang-Fairley</td>
<td>Additive Log-linear</td>
<td>$s^2/n_{ij}$</td>
<td>Massachusetts (1974-76)</td>
<td>Weighted Least Squares</td>
<td>Predictive accuracy of the additive model was better. Separate estimation for claim frequency and severity was suggested</td>
</tr>
<tr>
<td>Chamberlain [10]</td>
<td>Additive with multiplicative interaction</td>
<td>$s^2/n_{ij}$</td>
<td>Canadian (1957-58) Liability</td>
<td>Least Squares ANOVA</td>
<td>Suggests a hybrid form model using F-tests to choose interactions</td>
</tr>
<tr>
<td>Sant [41]</td>
<td>Multiplicative with interaction</td>
<td>$s^2/n_{ij}$</td>
<td>Massachusetts (1974-75)</td>
<td>Weighted Least Squares</td>
<td>Bias for high risks and low</td>
</tr>
<tr>
<td>DuMouchel [18]</td>
<td>Hybrid</td>
<td>$S^2 + t^2/n_{ij}^{**}$</td>
<td>Massachusetts (1982) Liability</td>
<td>Iterative Weighted Least Squares</td>
<td>Pure premiums are determined by a Bayesian credibility scheme</td>
</tr>
<tr>
<td>Samson-Thomas (1964)</td>
<td>Additive with interactions</td>
<td>$s^2/n_{ij}$</td>
<td>U.K. Collision</td>
<td>Weighted Least Squares</td>
<td>Suggests the step-wise regression or F-test to choose interactions</td>
</tr>
</tbody>
</table>

Notes: * K means a constant and $s^2$ means a variance of individual claim. ** t means an additional parameter for class-. Refer to eq. (3). *** d means a parameter indicating the degree of heteroskedasticity. Refer to eq. (4).
small, data bases. It is not clear which form, if any, may generally be most appropriate for modeling pure premiums. The functional forms proposed so far may be classified as: (a) additive, (b) multiplicative, and (c) flexible.

The flexible category includes the so-called hybrid models, i.e., those that include both additive and multiplicative terms, and the Box-Cox model. Harrington [23] has suggested that the Box-Cox flexible functional form may be more accurate in estimating pure premiums than purely additive and multiplicative forms. It would be especially desirable to analyze this method using additional data to see if these results hold.

The purpose of modeling pure premiums (or rate relatives) using past data is to determine rate levels that will be applied in the future. A limitation of most studies comparing functional forms is that the predictive accuracy of different models has been analyzed only using model residuals for the estimation period, as opposed to analyzing forecast errors for a subsequent period. An exception is Weisberg, Tomberlin and Chatterjee [50].

With the exception of Harrington [23], previous studies have paid little attention to whether the statistical assumptions underlying their methods have been appropriate. As Coutts [12] has pointed out, no investigation of the underlying distribution for the model disturbances has been undertaken and no residual tests comparing actual and fitted values have been published to support the assumptions of the models. If normality of disturbances is assumed in a pure premium model, this assumption should be tested, perhaps using the methods employed by Harrington.

As discussed previously, the existence of significant interaction effects in the risk classification models and whether those interaction terms should be in the model are controversial. In general, predictive errors would be expected to be lower by including interaction terms at best because it may result in using extra parameters. However, the importance of including these additional parameters may be validated by statistical tests. Chamberlain [10] tested the interaction effects using the F-statistic obtained from the traditional Analysis of Variance with a different sample size for each class. Even though statistical procedures may be used to test the hypothesis of no interactions in any functional forms of pure premium, statistical methods cannot be the only guidance or procedures to this controversy as Sant [41] indicated. Thus, the question is how to estimate the pure premium in those classes with few observations when one cannot be comfortable with the hypothesis that no interaction exists [41]. In other words, whether or not to include interaction terms should be decided from a practical perspective as well as a theoretical point of view.

While the Box-Cox method focuses on constructing a model that generates a flexible functional form by transforming the dependent variable according to the rule (4), the Bayes estimation attempts to model the disturbance terms by considering the interaction terms as random effects as well as by incorporating credibility theory. However, there has been no effort to compare these two estimation approaches yet. Thus a comparative analysis of these
models is needed. Thus the purpose of this article is to provide evidence concerning the following major questions:

(a) If the pure premium approach is used, does the Box-Cox flexible functional form estimator or the empirical Bayes estimator consistently improve predictive accuracy relative to the additive, multiplicative, and hybrid estimators?
(b) Is the predictive accuracy of the proposed Box-Cox model and the empirical Bayes model robust to various distributional assumptions?
(c) How significant is the role of interaction terms, and is it possible to improve predictive accuracy through modeling these terms?

Research Methods

This study has compared predictive accuracy of alternative pure premium models to provide empirical evidence for answering the above-mentioned questions. These models are as follows:

Model 1-1: Additive (linear) model
Model 1-2: Multiplicative (loglinear) model
Model 1-3: Massachusetts hybrid model
Model 1-4: Empirical Bayes model
Model 1-5: Box-Cox heteroskedastic model
Model 1-6: Additive model with interaction terms
Model 1-7: Multiplicative model with interaction terms
Model 1-8: Box-Cox heteroskedastic model with interaction terms.

Model Specifications and Estimation Methods

Model-1 and Model-2: For the purpose of estimation, Models (1) and (2) can be described as:

\[ y_{ij} = b_0 + \sum_{i} (b_{1i}X_{1i}) + \sum_{j} (b_{2j}X_{2j}) + e_{ij}, \]  

where \( y_{ij} = P_{ij} \) or \( \log(P_{ij}) \),

\[ X_{1i} = \begin{cases} 1 & \text{for the } i\text{-th level of factor } a,^4 \\ 0 & \text{for all other levels} \end{cases} \]

\[ X_{2j} = \begin{cases} 1 & \text{for the } j\text{-th level of factor } b \\ 0 & \text{for all other levels} \end{cases} \]

and \( b_0, b_{1i}, \) and \( b_{2j} \) are parameter estimates for the constant, \( X_{1i}, \) and \( X_{2j}, \) respectively, and \( e_{ij} \) is the error term. In cases of the linear or loglinear model, equation (6) can be estimated directly through the weighted least squares method using \( n_{ij} \), the number of observations in cell \( i, j \), as the weight.

^4 This study uses regression procedures rather than analysis of variance since the former has some important advantages. However, in order to avoid singularity of the design matrix, one dummy variable for one level of each factor in the model must be deleted. The choice of the level to be deleted is arbitrary and does not affect predicted values. For details, see Morrison [35].
In case data should be transformed into a certain form, it would be theoretically preferable to estimate models using the mean of the transformed values of the individual data rather than using the transformed value of the mean. This may require some caution in parameter estimation under the loglinear model. Otherwise, the estimates of parameters might be affected by the number of exposures in the given class. Aggregate data, that is, average claims amount per exposure for each class are used in the present analysis. However, from a practical perspective, it might be reasonable to treat these parameters as constants for the same classes of different sizes if the number of exposures in each class is considerably large [23]. This process would especially be reasonable when there was a chance some individuals would have no claims.

Massachusetts Hybrid Model (Model-3): As an experimental approach for more complex models that can produce the pure additive and multiplicative models as special cases, the model of the form (3) was proposed for the 1982 Massachusetts Automobile Insurance Classification Scheme. DuMouchel [18] estimated this model by assuming the disturbance variance so that:

\[
\text{Var}(e_{ij}) = s^2 + c_{ij}^2,
\]

where \( c_{ij}^2 = t_i^2/n_{ij} \) and \( t_i^2 \) is the within-cell variation for the class \( i \) of the factor 'a.' DuMouchel's estimation procedure requires two uses of an iterative weighted least squares procedure for estimating \( \{a_i, T_i, b_j\} \). The first preliminary estimates use weights \( w_{ij} = 1/c_{ij}^2 \); the results are used to estimate \( s^2 \). The final estimates are computed by repeating the procedure using \( w_{ij} = (s^2 + c_{ij}^2)^{-1} \).

Empirical Bayes Model (Model-4): A more general model could be employed as the form of (5) applying the empirical Bayes estimation developed by Leonard [33] and Laird [31]. For the estimation of this type of model, Laird [31] described model (5) as:

\[
\log q_{ij} = u_0 + u_{1(i)} + u_{2(j)} + u_{12(ij)},
\]

such that \( q_{ij} = E(P_{ij})/ \sum_{ij} n_{ij} P_{ij} \), \( \sum_{ij} q_{ij} = 1 \), and \( u_{12(ij)} \)'s are independently and identically distributed as \( N(m,s^2) \) with \( m = 0,^5 \) where \( P_{ij} \) is the observed pure premium for class-ij, \( u_{1(i)} \) is the main effect of row variable at level \( i \), \( u_{2(j)} \) is the column effect at level \( j \), \( u_{12(ij)} \) is the random effect of interaction, and \( u_0 \) is a normalizing constant ensuring the \( p_{ij} \)'s term to one.

To solve this problem, Laird [31] modified Leonard's two-stage Bayesian approach [33] by assuming the \( u_{1(i)} \)'s and the \( u_{2(j)} \)'s have flat prior distribution, with the parameters being subject to the constraints \( u_{1(+)} = u_{2(+)} = 0 \), where a ' + ' replacing a subscript indicates summation over the script. Placing a flat prior on these terms is equivalent to specifying a model such that the interaction terms are random variables while the overall main effects remain fixed [47].

---

5The choice of a zero mean is arbitrary for estimation of the \( q_{ij} \), since the posterior distribution of the \( q_{ii} \) given the \( P_{ii}, m \) and \( s^2 \), does not depend upon \( m \) [31].
To estimate the prior variance of the \( u_{12(ij)} \) terms, both Laird [31] and Tomberlin [47] employed a technique described by Dempster, Laird, and Rubin [17], where estimation of \( s^2 \) may be conducted treating the \( u \) terms as missing data. Their approach has been named as the EM algorithm, since each iteration of the algorithm involves both an expectation (E) step and a maximization (M) step. The E-step of a single iteration corresponds to deriving the expected value of the log-likelihood, \( \log m(P, u | s^2) \), given the observed data (\( P \)) and a current value of \( s^2 \). At the M-step, \( s^2 \) is estimated by maximizing the resultant expression over all values of \( s^2 \) [31, 47]. The algorithm then iterates back to the E-step, using the new value of \( s^2 \) for deriving the expected value of the log-likelihood such that [47]

\[
s^2_{t+1} = E_i \sum_{ij} u^2_{12(ij)} / IJ = \sum_{ij} [(u^{*}_{12(ij)}(t))^2 + \text{Var}_r(u_{12(ij)})] / IJ,
\]

where \( I \) and \( J \) are the numbers of class \( i \) and \( j \) respectively; and \( u^{*}_{12(ij)}(t) \) and \( \text{Var}_r(u_{12(ij)}) \) are the mean and variance obtained from the normal approximation at step \( t \) of the EM algorithm. The log-likelihood function for this estimation can be expressed as follows [31]:

\[
\log m^*(P | s^2) = K(P) + \sum_{ij} P_{ij} \log(q^{*}_{ij}) - (IJ/2)\log(s^2)
\]

\[-\sum_{ij} (u^{*}_{12(ij)})^2/(2s^2) - (1/2)\log |\Sigma^*| ,
\]

where \( K(P) \) is the part that does not include \( u \), \( q \), or \( s^2 \), \( q^{*}_{ij} \) is obtained by evaluating \( \log(q_{ij}) \) at \( u^* \), and \( \Sigma^* \) is the variance-covariance matrix of \( u^* \) [31, 47].

**Box-Cox Heteroskedastic Model (Model-5):** In the context of the risk classification scheme discussed in this study, the Box and Cox model makes the functional form of pure premium model described as equation (4). This model has been derived from the idea that as the dependent variables were transformed according to a certain rule, it would be possible to obtain a transformed dependent variable which would be a linear function of independent variables with normally distributed and homoscedastic error terms. Lahiri and Egy [30] extended the Box and Cox model by allowing heteroskedasticity.

Harrington [23] applied the Box and Cox heteroskedasticity model to the insurance pure premium estimation by assuming the \( e_{ij} \)'s to be normally distributed with zero means and variance \( s^2(n_{ij})^{-d} \) with the concentrated log-likelihood function given by [23, 30]:

\[
L(c,d) = K + (d/2) \sum_{ij} \log(n_{ij}) + (c-1) \sum_{ij} \log(P_{ij})
\]

\[-(N/2)\log(s^2(c,d)),
\]

where \( K \) is a constant and \( N \) is the total number of classes.

To estimate parameters \( c \) and \( d \) as well as \( s^2 \) and vector \( b \), Lahiri and Egy [30] and Harrington [23] used the zig-zag iterative search technique that was developed by Oberhofer and Kmenta [36], which proved that the iterative procedure always converges to a solution of the first-order maximizing
conditions. According to this procedure, by fixing \( d \) to a certain level, a maximum likelihood estimate of \( c \) may be obtained by searching within some reasonable range for the value of \( c \) which maximizes \( L \). A range of \( c \) may be specified and the value of \( L \) will be calculated with some magnitude of increments for \( c \) (i.e., 0.05 or 0.1, etc.) within this range. If the maximum value of \( L \) occurs at an end point of range, the range should be modified until an internal maximum is obtained. Then, with the value of \( c \) set to this level, the same procedure will be applied to find the value of \( d \) that maximizes \( L \). Such an iteration (i.e., zig-zag) may be repeated until (11) attains maximum with maximum likelihood estimators \( c^* \) and \( d^* \). The estimators for \( b \) are then obtained from the weighted least squares estimation of the function (4), using \( n_{ij} \) as the weight.

**Models with Interaction Terms (Model-6, 7, and 8):** The effect on the predictive accuracy of including interaction terms in the alternative models are compared in this study by using a stepwise regression with forward selection in both the linear and the loglinear model under the restriction that the parameters for the effects of factor \( a \) at all levels and the effects of factor \( b \) at all levels remain in the model at each step. Then, interaction terms which have been selected in both cases are included in Model-6, Model-7, and Model-8. These models can be described for estimation purposes as follows:

\[
Q_{ij} = b_0 + \sum_{i=1} \left( b_{1i} X_{1i} \right) + \sum_{j=1} \left( b_{2j} X_{2j} \right) + \sum_{i=1} \sum_{j=1} \left( b_{3ij} X_{2j} D_{ij} \right) + e_{ij}, \tag{12}
\]

where

\[
Q_{ij} = \begin{cases} 
P_{ij} & \text{for Model-6,} \\
\log P_{ij} & \text{for Model-7,} \\
P_{ij}(c) & \text{for Model-8}
\end{cases}
\]

\[
X_{1i} = \begin{cases} 
1 & \text{for the } i-\text{th level of factor } a, \\
0 & \text{for all other levels}
\end{cases}
\]

\[
X_{2j} = \begin{cases} 
1 & \text{for the } j-\text{th level of factor } b, \\
0 & \text{for all other levels}
\end{cases}
\]

\[
D_{ij} = \begin{cases} 
1 & \text{if the } ij-\text{th interaction term is included in the model,} \\
0 & \text{otherwise}
\end{cases}
\]

---

6 Since the selection method for interaction terms is heuristic, it is not guaranteed that an optimal solution is reached, especially when the factors are strongly correlated. But other selection methods without constraints for the main effects have not been used, partly because this study primarily attempts to see whether predictive accuracy may be improved by adding some interaction terms with other variables constant as for the model's without interaction terms. For the test for interaction effects and selection of interaction terms, see Chamberlain [10] and Samson and Thomas [39].
Comparison Criteria and Data

Selection of comparison criteria in this study may partly be based on the classic criteria proposed by Bailey and Simon [6], which can be described as follows:

**Balance:** For certain effect (or level), the actual marginal mean of pure premium must equal the predicted marginal mean.

**Credibility:** A rate should reflect the relative sizes of the various groups involved.

**Minimal Departure:** Departure of estimates from the raw data should be minimized.

**Differences caused by chance:** Differences between the raw data and the estimated relatives should be small enough to be cause by chance.

Among these four criteria, criterion of "minimal departure: and criterion of "caused by chance" have been emphasized for the model comparison. These criteria may be equivalent to those of goodness of fit and randomness of prediction errors, respectively. Chang and Fairley [11] expressed these features as "global accuracy" and "the absence of systematic biases."

The relative credibility of the classes is implicitly reflected in all alternative models considered in this section. For example, Model-4 (the empirical Bayes estimation) allows the observed cell experience to be used when it is reliable (or credible) and the smoothed model estimates to be used when only a small amount of data are available. Thus the empirical Bayes estimator is a credibility estimator. Even the other models reflect the credibility of data indirectly by taking the estimation weights as some functions of the number of observations for each cell.

With regard to criterion of balance, only Model-1 will generally produce predicted values that are balanced for each class and in total. But it may be questionable to apply this criterion strictly to the case of estimation under the risk classification system which consists of classes with significantly different numbers of exposures. Harrington [23, p. 538] discussed the validity of the balance criterion as follows:

The motivation for the balance criterion evidently is that it is reasonable to assume full credibility for certain partitions of the data. Within the context of statistical modeling of auto insurance claims, however, there would appear to be no persuasive reason to assume that the deviation of actual experience from the average true pure premium for any particular subset of data or for all exposures is zero.

Therefore, in this study, the alternative pure premium models have been compared based on the last two criteria.

With respect to the criterion of goodness of fit (minimal departure), the common procedure of comparing predicted cell means during the estimation period may be taken by using the mean squared error or mean absolute error as a measure of overall fit. These statistics can be expressed as follows:

\[
\text{Mean Squared Error} = \frac{1}{M} \sum_{i,j} n_{ij}(P_{ij} - \hat{P}_{ij})^2
\]  \hspace{1cm} (13)

\[
\text{Mean Absolute Error} = \frac{1}{M} \sum_{i,j} n_{ij} |P_{ij} - \hat{P}_{ij}|
\]  \hspace{1cm} (14)
where $P_{ij}$ is the observed pure premium, $\hat{P}_{ij}$ is the predicted value of pure premium for class-ij, and $M$ is the total number of observation in all classes.

An ideal procedure might analyze forecast error for a period following the estimation period. Johnson and Hey [26] and Weisberg, Tomberlin, and Chatterjee [50] took such an approach. When two-period data are used, two types of adjustment should be made. One is an adjustment to the trend and the other is to the random variation of pure premium for each class from year to year. A similar adjustment procedure was conducted by Coutts [12].

In order to incorporate trends into the estimates, all cell means for the predicted mean for the year equal to that in the base year. Weisberg, Tomberlin, and Chatterjee [50] argued that a failure to adjust when substantial trends exist can be clearly misleading, since interest centers on errors of classification alone and not on errors resulting from failure to anticipate inflation or other time trends. Thus, to compare estimates of pure premium with the following year's observations of pure premium, the estimates from the preceding year will be adjusted to increase by the growth rate for the overall weighted average of pure premium.

With respect to the adjustment to random variation of pure premium for each class from year to year, however, there is no reason that the randomness from year to year may affect the alternative models in different ways. Such a problem may also be reduced by comparison using various data sets. Thus in this study, first, the goodness of fit has been compared, and for some data sets with two-period data, the predictive accuracy has been checked using those data after adjusting only the trend factor.

Since all concerned pure premium models assume normality of error terms in some way, the normality assumption of the models has been tested with regard to the randomness of predicted errors (caused by chance). In the field of statistics, various tests for normality have been developed. One of the popular and simple tests traditionally used might be that based on the skewness or the kurtosis of the disturbance distribution. These statistics can be expressed as follows:

Skewness: $\sqrt{b_1} = m_3/m_2^{3/2}$  \hspace{1cm} (15)

Kurtosis: $b_2 = m_4/m_2^2$  \hspace{1cm} (16)

where $m_k = (1/N)\Sigma (u_{ij} - \bar{u})^k$, $\bar{u} = \Sigma u_{ij}/N$, $u_{ij}$ is the residual for cell ij, and $N$ is the number of cells.

Shapiro and Wilk's W-statistic has been regarded as one of the most powerful test statistic in the related area [43, 44]. But this statistic was originally designed for the small sample size of less than 50 observations. Saniga and Miles [40] and White and MacDonald [51] show that the Shapiro-Wilk's W-statistic is not as powerful for large sample sizes or against asymmetric stable alternatives. For moderate or large sample sizes, these

---

7 Even Weisberg, Tomberlin and Chatterjee [50] did not use their adjusted measures of mean squared error estimates for the reason that the over-interpretation of small differences must be avoided.
comparative studies support some recently-developed omnibus test for non-normality based on the joint use of $\sqrt{b_1}$ and $b_2$. Of these, the most easily available is the R-test proposed by Pearson, D'Agostino, and Bowman [37], while the test based on the D-statistic proposed by D'Agostino [15] was the best against asymmetric alternatives (e.g., double exponential distribution).

The idea underlying the R-test is that one can construct a rectangle in $\sqrt{b_1}$, $b_2$ space such that a joint observation $(\sqrt{b_1}, b_2)$ will fall outside the rectangle with probability 'a' when sampling from an underlying normal distribution. If $\sqrt{b_1}$ and $b_2$ were independent, the rectangle could be constructed from the upper and lower 100a' percentage points of $\sqrt{b_1}$ and $b_2$ such that $a' = (1 - (1 - a)^{1/2})/2$. Thus if $\sqrt{b_1}$ or $b_2$ lies upper or lower 100a' percentage points, the R test rejects the null hypothesis of normality at the a-level. Even though $\sqrt{b_1}$ and $b_2$ are not independent, the a' can be adjusted to compensate for lack of independence. The value of a' appropriate for $a = .05$ and $a = .10$ given for samples of size 20, 50, and 100 by Pearson, D'Agostino, and Bowman [37] have been used in this study.

The test based on the D-statistic is another omnibus test that compares favorably in power with the Shapiro-Wilk W test for highly skewed distributions. This statistic is given by D'Agostino [15].

Finally, the approximate test for outliers devised by Lund [34] has been conducted. For this test, the maximum or minimum standardized residual for each model is compared to the estimated upper bound tabulated by Lund for this statistic under the normality assumption. This test may provide evidence of whether the error for at least one of the risk classes cannot be attributed to chance.

The predictive accuracy of eight alternative models has been compared using two data sets. The first one is the insurance data collected by the Massachusetts Automobile Rating and Accident Prevention Bureau (MAR B) for years 1979 and 1980. The data analyzed here, which are reported in Tomberlin [47], relate to the property damage liability coverage and these are presented in aggregate form, classified by territory and driver class (five driver classes and 24 territories). The second data set contains 1979 accident data for the United Kingdom. These data are reported in Coutts [12] and classified by vehicle age, vehicle group, and policyholder age, yielding 60 classes.

Empirical Results

Comparison of Predictive Accuracy

The results of estimating Model-1 through Model-8 are given in Table 2. The first conclusion indicated from these results is that both the empirical Bayes models and the Box-Cox heteroskedastic models outperformed traditional linear or loglinear models. But it should come as no surprise that the empirical Bayes models or the Box-Cox heteroskedastic models fit the data better than the traditional models partly because they are based on models

---

8 In reality, $\sqrt{b_1}$ and $b_2$ are not independent in finite samples [51, p. 17].
with extra parameters. That is, the empirical Bayes model has an extra parameter, the prior variance, which may lead to a compromise between the observed rates and estimates from the traditional linear or loglinear model. For example, in Massachusetts' 1979 data set, the mean absolute error of 1.413 for Model 2 has been reduced to 0.132 for Model 4 estimates. An even larger reduction has been achieved in the mean squared error. On the other hand, the Box-Cox heteroskedastic model has additional parameters c and d which may transform the dependent variable and adjust the disturbance terms. Both mean squared error and mean absolute error have substantially been reduced from those of Model 1 or Model 2.

Table 2
Comparison of Alternative Pure Premium Models
(Predictive Accuracy)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE*</td>
<td>MSE**</td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>1. Additive</td>
<td>1.66</td>
<td>17.79</td>
<td>3.02</td>
<td>94.82</td>
</tr>
<tr>
<td>2. Loglinear</td>
<td>1.41</td>
<td>20.75</td>
<td>2.69</td>
<td>87.37</td>
</tr>
<tr>
<td>3. Mass Hybrid</td>
<td>1.41</td>
<td>12.54</td>
<td>2.51</td>
<td>103.85</td>
</tr>
<tr>
<td>4. Empirical Bayes</td>
<td>0.13</td>
<td>2.52</td>
<td>1.85</td>
<td>57.86</td>
</tr>
<tr>
<td>5. Box-Cox Heteros</td>
<td>1.26</td>
<td>14.94</td>
<td>2.42</td>
<td>75.71</td>
</tr>
<tr>
<td>6. Linear w/Interaction</td>
<td>0.79</td>
<td>8.81</td>
<td>2.23</td>
<td>75.48</td>
</tr>
<tr>
<td>7. Loglinear w/Interaction</td>
<td>0.43</td>
<td>3.42</td>
<td>2.41</td>
<td>69.19</td>
</tr>
<tr>
<td>8. Box-Cox Heteros w/Interaction</td>
<td>0.38</td>
<td>3.27</td>
<td>2.06</td>
<td>62.55</td>
</tr>
</tbody>
</table>

* Mean Absolute Error
** Mean Squared Error

Application of the likelihood ration test would result in rejection of the null hypothesis that c = 0 or c = 1. The hypothesis that d = 1 was rejected for the United Kingdom data, but the likelihood ratio test for this hypothesis could not be rejected with Massachusetts data.¹⁰

¹⁰The estimated optimal value of c for the Box-Cox heteroskedastic model without interaction terms (Model 5) are 0.25 for Massachusetts data and 0.3 for the United Kingdom data. The optimal value of d is 1.05 for both data sets. The result of likelihood ratio tests are as follows:

Massachusetts data:

\[ L(c = 0.25) = -418.0227; L(c = 0) = -422.8196; L(c = 1) = -459.1621 \]

\[ H_0: c = 0 \text{ vs } H_1: c = c^*(=0.25) - h^* = 9.594 > h^* \]

\[ H_0: c = 1 \text{ vs } H_1: c = c^*(=0.25) - h^* = 82.28 > h^* \]

[Chi-squares with 0.01 level: \( h^* = 6.635 \)]

United Kingdom data:

\[ L(c = 0.3) = -222.778; L(c = 0) = -229.518; L(c = 1) = -265.275 \]

\[ H_0: c = 0 \text{ vs } H_1: c = c^*(=0.3) - h^* = 13.48 > h^* \]

where \( L(C) \) denote the log-likelihood at \( C \) and \( h \) denote the likelihood ration.
Even though Coutts [12] and Baxter, Coutts and Ross [7] argued for no significant role of interaction terms, Samson and Thomas [39] and Harrington [23] noted that there were clear patterns in the prediction errors for the models without interaction terms and these patterns could be mitigated by including some selected interaction terms. In this study, Models 6 through 8 have been constructed with some interaction terms selected by the procedure of stepwise regression, interaction terms were included if F statistics reflecting the variable's contribution to the model were significant at .15 level. Forty-six interaction terms were included for estimation for Massachusetts data and 13 from the United Kingdom data set. The results shown in Table 2 indicate that estimation in Models 6, 7, and 8 result in a substantially better fit than Models 1, 2, and 5 for both data sets. The mean squared errors as well as the mean absolute errors have been significantly reduced. This implies that discreet inclusion of interaction terms in the pure premium model may make it possible to predict pure premiums in the following period more accurately.11

For the models with interactions, the values of $c_{\text{max}}^*$ are 0.2 for the Massachusetts data and 0.35 for the United Kingdom data with $d^* = 1.25$ and 1.1 respectively. In both cases, the likelihood ratio tests indicate rejection of the hypothesis that $c = 0$ or $c = 1$.12

Among the new models (i.e., excluding the traditional linear and log-linear models), Model 3 (the Massachusetts hybrid model) produced worse predictive accuracy than did the empirical Bayes models and the Box-Cox heteroskedastic models even though its goodness of fit was better than the two traditional models. The result from Coutts data was even worse for Model 3. That might be because this special functional form has been devised for Massachusetts automobile ratemaking. So it might not work in other cases which have characteristics different from Massachusetts'. The results show that the empirical Bayes models produce better predictive accuracy than does the Box-Cox heteroskedastic model does when interaction terms are included (Model 8).

The ultimate power of a model may be its ability to forecast pure premiums in the following period. The right side of Table 2 shows the results of forecasting pure premiums for 1980 by estimation of alternative models using 1979 Massachusetts data. All forecasted pure premiums were multiplied by a factor that makes the overall weighted mean of pure premium in 1980 equal to that in the base year, that is, 1979 in order to incorporate in the estimates actual trends that occurred. Because actual rather than projected time trends were used, the results in the right side of Table 2 should be somewhat more accurate than those that can be expected in practice. There is no significant

---

11 This result shows some practical usefulness of including interaction terms, but it may not necessarily imply that there exist statistically significant interaction effects. Tests for significance of interaction terms have not been attempted because of forced inclusion of all levels of factors $a$ and $b$ in the models, which facilitated comparison of the results to those for other models. For the test of interaction effects, see Chamberlain [10].

12 Tomberlin data: $h^*(\text{under } H_0; c = 0) = 17.93; h^*(\text{under } H_0; c = 1) = 37.75$. Coutts data: $h^*(\text{under } H_0; c = 0) = 14.078; h^*(\text{under } H_0; c = 0) = 76.122$. 
Alternative Pure Premium Models in Automobile Risk Classification

449

difference from the results of goodness of fit using 1979 data, except that Model 3 (the Massachusetts hybrid model) produced the best mean squares error. But this might not be so reliable because it also shows the poor mean absolute error. Model 8 (Box-Cox with interaction terms) consistently improves its predictive accuracy which relatively produced a better result than did empirical Bayes models compared to their relationships in the results from 1979 data.

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison of Alternatives Pure Premiums Models</td>
</tr>
<tr>
<td>(Test for Normality of Residuals)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Massachusetts Data</th>
<th>United Kingdom Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sqrt{b_1}$</td>
<td>$b_2$</td>
</tr>
<tr>
<td>1. Additive</td>
<td>0.52</td>
<td>4.62*</td>
</tr>
<tr>
<td>2. Loglinear</td>
<td>-0.96*</td>
<td>2.78</td>
</tr>
<tr>
<td>3. Mass Hybrid</td>
<td>0.35</td>
<td>3.51</td>
</tr>
<tr>
<td>4. Empirical Bayes</td>
<td>0.57</td>
<td>6.06*</td>
</tr>
<tr>
<td>5. Box-Cox Heteros</td>
<td>-0.79*</td>
<td>2.49</td>
</tr>
<tr>
<td>6. Linear w/Interaction</td>
<td>0.51</td>
<td>6.41*</td>
</tr>
<tr>
<td>7. Loglinear w/Interaction</td>
<td>-1.23*</td>
<td>6.39</td>
</tr>
<tr>
<td>8. Box-Cox Heteros w/Interaction</td>
<td>-0.72</td>
<td>6.26*</td>
</tr>
</tbody>
</table>

* Null hypothesis of normality rejected at 0.01 level  
** Maximum absolute normalized error

Residual Analysis

A certain type of model or functional form may not be recommended as most appropriate simply for the reason that it produces the best predictive accuracy without consideration for its conceptual characteristics including the adequacy of underlying assumptions. Hence disturbance variances and predictive errors of pure premiums have been analyzed.

The first analysis with regard to the disturbance is the test for normality of the disturbance since all alternative pure premium models assume that disturbances are normally distributed even though they may be heteroskedastic. Residuals for each model were normalized by the estimated standard error for each cell\(^{13}\) and various methods for the test of normality described in the previous section have been applied. The results are shown in Table 3. The models which passed all tests by skewness, kurtosis, Pearson's R, D'Agostino's D, and Lund's outliers are Models 2, 3, and 5 for Massachusetts data, and Models 2, 4, 5, and 8 for the United Kingdom data. The linear model has been found highly skewed whether it includes interaction terms or not. The Massachusetts hybrid model shows a fine disturbance distribution for Massachusetts data, but it distribution for the United Kingdom data seems

\(^{13}\)In the empirical work for this study, standardized residuals have been replaced with normalized residuals, partly because of complexity in calculating standardized results and partly because they may not be considered to distort the results.
to seriously violate the normality assumption. This implies Model 3 may not be suggested outside Massachusetts or it may need at least a comprehensive survey of relationships among risk classification factors to construct a model.

The models with interaction terms produced large values of \( b_2 \) and outliers, especially for Massachusetts data, perhaps partly due to the arbitrariness in selection procedures. This indicated that selection of interaction terms may require careful consideration.

Unlike the findings of Harrington [23], the normality assumption of the Box-Cox heteroskedastic model has not been rejected regardless of inclusion of interaction terms, with only a single exception of the Lund's test for outliers with Massachusetts data. This finding may support that the Box-Cox heteroskedastic model could be very powerful from a conceptual standpoint as well as in terms of predictive accuracy, especially if the underlying distribution of the concerned variable is not known.

Simulation Analysis

In the previous section, it has been shown that the Box-Cox type of flexible functional forms or the empirical Bayes estimation enables more accurate estimation of pure premiums. It may be desirable to compare the alternative models with new data sets to supplement the analysis in the preceding models to obtain additional empirical evidence whether these models are powerful enough to predict pure premiums in various circumstances. Because no additional data set was available to the author, simulation has been employed to create data under different distributional assumptions for claim frequency and severity. The main principles of the simulation approach on insurance mathematics have been discussed in Cummins and Wiltbank [14] and Pitacco [38].

For the simulation purposes, the three by two risk classification plan has been assumed. Two driver classes appear in each of three territories. By assuming parameter values for the driver class and territory effects, the number of exposures in each cell, and probability distributions for both claim frequency and severity, a large number of sample data sets can be generated, which include the average number of claims, average claim sizes, and average pure premiums for these risk classes. The alternative functional forms and estimation methods then can be applied to the data to gain insight into the loss of predictive accuracy that arises under various methods by not knowing the true model. A drawback to this approach is the computational cost associated with comparing many underlying models for frequency and severity. As a result, only a few cases have been selected.

Lemaire [32] and Ferreira [20] have shown the negative binomial distribution as the appropriate underlying distribution for frequency of automobile claims. Hence the negative binomial distribution is assumed to generate data for the number of claims for each class. Its probability frequency function can be written as:

\[
p(z \mid r, p) = \binom{r+z-1}{z} p^r (1-p)^z, \quad z = 0, 1, 2, \ldots
\]  
(17)
with mean, \( r(1 - p)/p \), and variance, \( r(r(1 - p)/p^2 \). For selection of the values of parameters \( r \) and \( p \), the value parameter \( r \) is assumed constant for all risk classes in the single data set. The estimate of \( r \) for California drivers during 1961 and 1962 in Ferreira [20] was 1.1571 and that for the Belgian automobile claim data in Lemaire [32] was 1.6222. Hence four values for parameter \( r \) have been used for the simulation such that \( r = 2.0, 1.5, 1.0, \) and \( 0.5 \). A negative binomial distribution with \( r = 1.0 \) is equivalent to a geometric distribution.

With regard to the severity distribution, the Gamma distribution has been assumed since the residual tests in Jee [25] did not reject the hypothesis that the underlying distribution of claim severity is Gamma distributed. Its probability density function can be expressed as:

\[
f(q \mid s,a) = q^{a-1} \exp(-aq/s)a^s s^{-a}/r(a),
\]

with a mean and variance given by \( s \) and \( s^2/a \). It is also assumed, like the frequency distribution, that the scale parameter \( 'a' \) has the same value for all risk classes in a specific data set. The squared root of this parameter, \( \sqrt{a} \), implies the reciprocal of the coefficient of variation. For the simulation, three cases are assumed with respect to the severity distribution such that the coefficients of variance are two, three, and four. This means that the values of parameter \( 'a' \) are \( a = .25, a = .1111, \) and \( a = .0625, \) respectively.\(^{14} \) As the value for \( 'a' \) increases here, the associated data would be less variable.

Based on four distributional assumptions about claim frequency distribution and three assumptions for severity, 12 data sets have been created. For convenience, they have been classified into three groups according to the value of parameter \( 'a' \) for the severity distribution and subgrouped by the value of parameter \( 'r.' \) Group I includes data sets with \( a = .25, \) Group II with \( a = .1111, \) and Group III with \( a = .0625. \)

The values of the parameters for average number of claims, average claim size, and the number of exposures for each of six risk classes have been given referring to the data structure of Massachusetts' automobile claim experiences in 1979 and 1980 used in the previous section. The observations of claim rates and average costs per claim for driver classes 2 and 4 in the three territories with the largest exposures (Territories 3, 8, and 11) were used as parameters of claim frequency and severity for six risk classes. The number of exposures for each class was reduced to one tenth of the original size considering the computer capacity for simulation.

To obtain two-period claim data, two data sets were created with the same parameter values. For convenience, it has been assumed that there is no change between two consecutive periods with respect to inflation, population, and other conditions surrounding automobile accidents. It is believed that these assumptions may not influence the validity of comparison results. As a result, 24 data sets have been created.

\(^{14} \)Selection of values for parameter \( 'a' \) may be somewhat ad hoc, because no empirical work has provided insight into the magnitude of \( 'a' \) for the Gamma distribution.
The first period data of pure premiums for each data set were used to estimate pure premiums for the alternative pure premium models specified in the previous section. To consider the interaction effect, a single dummy variable indicating the class with the largest number of exposures (i.e., Driver class 1 with Territory 2) was arbitrarily included for the models with interaction terms (i.e., Models 6, 7, and 8). Then, mean absolute errors and mean squared errors, as expressed in (13) and (14), were calculated. For the purpose of comparing predictability of the models, the second period observations of pure premiums were used to calculated the deviations from the predicted values of pure premiums.

The comparison results for 12 data sets are illustrated in Table 4 which shows that the flexible type of functional forms or the empirical Bayes estimation (Models 4, 5, and 8) still produced better estimates of pure premiums under various distributional assumptions. The best pure premium models which produced the lowest mean absolute error or mean squared error for individual data sets are underlined. With only on exception for each criterion (MAE of Model 7 with \( r = 0.5 \) in Group II and MSE of Model 3 with \( r = 1.5 \) in Group III), none of them belongs to traditional models, whether additive, multiplicative, or hybrid estimators.

From a theoretical perspective, it may not make sense to test the interaction effect from the data sets created by the simulation, since the data for each risk class were generated independently of those for other classes. However, it might also be hypothesized that predictive accuracy would be improved by including such an additional parameter at least for the cases of flexible functional forms.

Table 4 shows that the predicted accuracy of the additive model with interaction terms (Model 6) was no better than that of the model without interactions (Model 1). It was even worse in some cases. But, for data sets in Groups II and III, the mean absolute errors and mean squared errors from the loglinear model or Box-Cox heteroskedastic model with interaction terms are consistently lower than those without interactions, even though they are not different for data sets in Group I. This may imply that the greater the within-class variabilities of the data set, the more modeling interaction terms could improve the predictive accuracy. This hypothesis may be supported by the fact that most of the best pure premium models in Group III are Box-Cox heteroskedastic models with interaction terms (Model 8) while these models in Group I show no difference from the original Box-Cox heteroskedastic model. It also implies that if interaction terms are included, the estimators from the Box-Cox type of flexible functional form could be better than the Bayes estimators.

Residuals from the alternative models employed for the simulation were investigated to check the validity of the models. In fact, all the models for the

---

15 This study has not attempted to investigate the predictive accuracy simply using the cell means for the first period to predict pure premiums in the second period, since Weisberg, Tomberlin and Chatterjee [50] found very poor results for the simple use of cell means.
### Table 4
Simulation Results: Predictive Accuracy

#### Mean Absolute Error

<table>
<thead>
<tr>
<th>Model</th>
<th>Group I ($a = .25$)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r=2.0$</td>
<td>$r=1.5$</td>
<td>$r=1.0$</td>
<td>$r=0.5$</td>
<td>$r=2.0$</td>
<td>$r=1.5$</td>
<td>$r=1.0$</td>
<td>$r=0.5$</td>
<td>$r=2.0$</td>
<td>$r=1.5$</td>
<td>$r=1.0$</td>
<td>$r=0.5$</td>
<td>$r=2.0$</td>
<td>$r=1.5$</td>
<td>$r=1.0$</td>
</tr>
<tr>
<td>5. Box-Cox Heteros</td>
<td>6.45</td>
<td>11.39</td>
<td>5.17</td>
<td>10.01</td>
<td>15.95</td>
<td>6.64</td>
<td>19.31</td>
<td>4.79</td>
<td>10.14</td>
<td>24.98</td>
<td>17.73</td>
<td>9.67</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Mean Squared Error

| Model                   | Group I ($a = .25$) |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
|-------------------------|----------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|                         | $r=2.0$              | $r=1.5$ | $r=1.0$ | $r=0.5$ | $r=2.0$ | $r=1.5$ | $r=1.0$ | $r=0.5$ | $r=2.0$ | $r=1.5$ | $r=1.0$ | $r=0.5$ | $r=2.0$ | $r=1.5$ | $r=1.0$ | $r=0.5$ |
| 1. Additive             | 92.2                 | 295.3   | 91.8    | 335.1   | 314.4   | 148.3   | 671.6   | 28.1    | 159.7   | 823.8   | 554.1   | 145.8   |        |         |         |         |
| 2. Loglinear            | 72.9                 | 497.1   | 72.8    | 285.9   | 362.5   | 106.6   | 784.3   | 27.1    | 232.7   | 978.1   | 476.7   | 183.2   |        |         |         |         |
| 3. Mass Hybrid          | 171.1                | 292.5   | 185.0   | 418.6   | 313.3   | 303.9   | 837.6   | 31.1    | 224.6   | 702.7   | 692.4   | 249.1   |        |         |         |         |
| 4. Empirical Bayes      | 45.0                 | 246.9   | 40.9    | 270.9   | 294.4   | 55.4    | 732.1   | 27.8    | 164.8   | 767.1   | 307.6   | 140.3   |        |         |         |         |
| 5. Box-Cox Heteros      | 51.8                 | 204.8   | 107.9   | 231.4   | 498.8   | 68.2    | 829.0   | 28.8    | 172.2   | 978.8   | 490.6   | 186.4   |        |         |         |         |
| 6. Linear w/Interaction | 75.5                 | 281.2   | 101.2   | 365.5   | 355.2   | 132.1   | 827.4   | 28.3    | 163.2   | 873.8   | 511.2   | 139.4   |        |         |         |         |
| 7. Loglinear w/Interaction | 56.8                | 436.2   | 110.5   | 201.9   | 361.1   | 83.6    | 922.3   | 26.2    | 142.7   | 891.9   | 301.9   | 166.5   |        |         |         |         |
| 8. Box-Cox Heteros      | 46.4                 | 244.6   | 46.3    | 291.3   | 548.9   | 57.2    | 785.4   | 28.6    | 126.8   | 854.4   | 293.1   | 133.2   |        |         |         |         |
analysis in this section have been used under the assumption that the disturbance terms are normally distributed. Hence the underlying assumption of normality was tested for each model in each data set. A possible problem with the tests is that the tests using skewness or kurtosis may not be so powerful as in the previous section, because the number of cases for each data set is too small. As White and MacDonald [51] noted in their comparison of the tests for normality, the test by the Shapiro and Wilk's W statistic works well with a small sample size [43]. The approximate test for outliers could also be employed since Lund [34] tabulated the upper bound of maximum absolute standardized residual even for the sample with the very small number of observations.

The measures of W statistic and maximum absolute normalized residuals for each model in each data set are presented in Table 5. It is interesting to see that these measures for the additive or multiplicative models indicate no significant departure of residuals from normality assumption, contrary to the findings in the previous section. This may be attributed to the nature of the data for simulation which are less variable than is real world experience. The measures for the Massachusetts hybrid model (Model 3) and the Box-Cox heteroskedastic pure premium model (Model 5) indicate some suspicion on the normality of the underlying distribution of residuals for these models. However the hypothesis that the residuals from the Box-Cox pure premium model with interaction (Model 8) are normally distributed for any data set could not be rejected.

**Conclusion**

This article shows the power of the flexible functional form estimators or the empirical Bayes estimators based on credibility theory to predict future pure premiums and provide some empirical evidence that the role of interaction terms may be material in improving predictive accuracy, at least for the estimation using flexible functional forms like the Box-Cox heteroskedastic model. However, the findings indicate limitations for such a conclusion.

Given the scope of the study and the nature of the analysis, the results cannot be completely conclusive. In particular, availability of new data sets has been limited, and the simulation is not without drawbacks. Since available information was in the form of aggregate data, the study does not analyze in depth the possibility of heterogeneity within rate classes. Even if more accurate class average rates can be obtained, the within-class heterogeneity conceivably could be increased with increased adverse selection. With respect to interaction terms, other selection methods may be used without the constraint of including all levels of main effects to test the significance of interaction terms. These problems need further discussion from a conceptual perspective as well as empirical analysis.

An important unanswered question is whether it generally would be preferable to conduct separate modeling of claim frequency and severity in
<table>
<thead>
<tr>
<th>Model</th>
<th>Shapiro-Wilk's W</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r=2.0</td>
<td>r=1.5</td>
<td>r=1.0</td>
<td>r=0.5</td>
<td>r=2.0</td>
<td>r=1.5</td>
<td>r=1.0</td>
</tr>
<tr>
<td>1. Additive</td>
<td>0.772</td>
<td>0.804</td>
<td>0.779</td>
<td>0.779</td>
<td>0.777</td>
<td>0.775</td>
<td>0.775</td>
</tr>
<tr>
<td>2. Loglinear</td>
<td>0.787</td>
<td>0.777</td>
<td>0.780</td>
<td>0.753</td>
<td>0.769</td>
<td>0.787</td>
<td>0.787</td>
</tr>
<tr>
<td>3. Mass Hybrid</td>
<td>0.736</td>
<td>0.691*</td>
<td>0.714</td>
<td>0.727</td>
<td>0.735</td>
<td>0.731</td>
<td>0.696*</td>
</tr>
<tr>
<td>4. Empirical Bayes</td>
<td>0.745</td>
<td>0.560*</td>
<td>0.840</td>
<td>0.819</td>
<td>0.821</td>
<td>0.804</td>
<td>0.790</td>
</tr>
<tr>
<td>5. Box-Cox Heteros</td>
<td>0.672*</td>
<td>0.762</td>
<td>0.713</td>
<td>0.728</td>
<td>0.749</td>
<td>0.626*</td>
<td>0.689*</td>
</tr>
<tr>
<td>6. Linear w/Interaction</td>
<td>0.768</td>
<td>0.768</td>
<td>0.768</td>
<td>0.768</td>
<td>0.768</td>
<td>0.769</td>
<td>0.768</td>
</tr>
<tr>
<td>7. Loglinear w/Interaction</td>
<td>0.765</td>
<td>1.124*</td>
<td>0.801</td>
<td>0.762</td>
<td>0.765</td>
<td>0.767</td>
<td>0.767</td>
</tr>
<tr>
<td>8. Box-Cox Heteros</td>
<td>0.789</td>
<td>0.780</td>
<td>0.779</td>
<td>0.786</td>
<td>0.787</td>
<td>0.790</td>
<td>0.787</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Maximum Outlier</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r=2.0</td>
<td>r=1.5</td>
<td>r=1.0</td>
<td>r=0.5</td>
<td>r=2.0</td>
<td>r=1.5</td>
<td>r=1.0</td>
</tr>
<tr>
<td>1. Additive</td>
<td>1.33</td>
<td>1.52</td>
<td>1.36</td>
<td>1.42</td>
<td>1.45</td>
<td>1.31</td>
<td>1.20</td>
</tr>
<tr>
<td>2. Loglinear</td>
<td>1.46</td>
<td>1.43</td>
<td>1.27</td>
<td>1.23</td>
<td>1.31</td>
<td>1.51</td>
<td>1.29</td>
</tr>
<tr>
<td>3. Mass Hybrid</td>
<td>1.75</td>
<td>1.61</td>
<td>1.77</td>
<td>1.76</td>
<td>1.71</td>
<td>1.67</td>
<td>1.91</td>
</tr>
<tr>
<td>4. Empirical Bayes</td>
<td>1.84</td>
<td>2.19*</td>
<td>1.61</td>
<td>1.53</td>
<td>1.49</td>
<td>1.63</td>
<td>1.46</td>
</tr>
<tr>
<td>5. Box-Cox Heteros</td>
<td>2.14*</td>
<td>1.72</td>
<td>2.10*</td>
<td>2.07*</td>
<td>1.96</td>
<td>2.17*</td>
<td>1.97</td>
</tr>
<tr>
<td>6. Linear w/Interaction</td>
<td>1.29</td>
<td>1.30</td>
<td>1.29</td>
<td>1.29</td>
<td>1.29</td>
<td>1.29</td>
<td>1.28</td>
</tr>
<tr>
<td>7. Loglinear w/Interaction</td>
<td>1.33</td>
<td>1.29</td>
<td>1.31</td>
<td>1.29</td>
<td>1.30</td>
<td>1.29</td>
<td>1.29</td>
</tr>
<tr>
<td>8. Box-Cox Heteros</td>
<td>1.19</td>
<td>1.71</td>
<td>1.27</td>
<td>1.79</td>
<td>1.81</td>
<td>1.81</td>
<td>1.81</td>
</tr>
</tbody>
</table>

* Null hypothesis of normality rejected at .01 level
order to estimate relative pure premiums. Moreover, it would be desirable to compare the predictive accuracy of the Box-Cox method or the empirical Bayes estimation for a frequency (or severity) model to that of purely additive or multiplicative models [25]. However, while building more elaborate models may increase accuracy, there also is some advantage to simplicity. Assessment of the methods' performance should take this into account and cost-benefit analysis should be done in selecting the appropriate estimation model. An example of such an effort is Freifelder [21]. The analysis of the modeling approach to rate-making may require theoretical or conceptual analysis in addition to empirical work. Practical issues also must be considered.

References

13. Cummins, J. D. and S. E. Harrington, “Economic Forecasting of Automobile Insurance Paid Claim Costs,” *Strategic Planning and*


A Comparative Analysis of Alternative Pure Premium Models in the Automobile Risk Classification System
Beomha Jee
Stable URL: http://links.jstor.org/sici?sici=0022-4367%28198909%2956%3A3%3C434%3AACAAOAP%3E2.0.CO%3B2-7

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

[Footnotes]

3 Regulation in the Insurance Industry: Determination of Premiums in Automobile Insurance
Yehuda Kahane; Haim Levy
Stable URL: http://links.jstor.org/sici?sici=0022-4367%28197503%2942%3A1%3C117%3ARITIID%3E2.0.CO%3B2-8

5 Empirical Bayes Methods for Two-Way Contingency Tables
N. M. Laird
Stable URL: http://links.jstor.org/sici?sici=0006-3444%28197812%2965%3A3%3C581%3ABMFTC%3E2.0.CO%3B2-E

7 Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches
Herbert I. Weisberg; Thomas J. Tomberlin; Sangit Chatterjee
Stable URL: http://links.jstor.org/sici?sici=0735-0015%28198404%292%3A2%3C170%3APILUCA%3E2.0.CO%3B2-P

NOTE: The reference numbering from the original has been maintained in this citation list.
8 Some Large-Sample Tests for Nonnormality in the Linear Regression Model
Halbert White; Glenn M. MacDonald
Stable URL: http://links.jstor.org/sici?sici=0162-1459%28198003%2975%3A369%3C16%3ASLTFNI%3E2.0.CO%3B2-E

15 Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches
Herbert I. Weisberg; Thomas J. Tomberlin; Sangit Chatterjee
Stable URL: http://links.jstor.org/sici?sici=0735-0015%28198404%292%3A2%3C170%3APILUCA%3E2.0.CO%3B2-P

References

11 Pricing Automobile Insurance under Multivariate Classification of Risks: Additive versus Multiplicative
Lena Chang; William B. Fairley
Stable URL: http://links.jstor.org/sici?sici=0022-4367%28197903%2946%3A1%3C75%3APAIUMC%3E2.0.CO%3B2-W

14 Estimating the Total Claims Distribution Using Multivariate Frequency and Severity Distributions
J. David Cummins; Laurel J. Wiltbank
Stable URL: http://links.jstor.org/sici?sici=0022-4367%28198309%2950%3A3%3C377%3AETTCDU%3E2.0.CO%3B2-Z

15 An Omnibus Test of Normality for Moderate and Large Size Samples
Ralph B. D'Agostino
Stable URL: http://links.jstor.org/sici?sici=0006-3444%28197108%2958%3A2%3C341%3AAOTONF%3E2.0.CO%3B2-Z

NOTE: The reference numbering from the original has been maintained in this citation list.
16 **Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1**
Ralph D’Agostino; E. S. Pearson
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197312%2960%3A3%5C%5C613%5C%5CTFDFNE%5C%5CE2.0.CO%5C%5CB2-Y

18 **The 1982 Massachusetts Automobile Insurance Classification Scheme**
William H. Dumouchel
Stable URL:
http://links.jstor.org/sici?sici=0039-0526%28198303%2F06%2932%3A1%2F2%5C%5C69%5C%5CT1MAIC%5C%5CE2.0.CO%5C%5CB2-Z

19 **Pricing Automobile Insurance under a Cross-Classification of Risks: Evidence from New Jersey**
William B. Fairley; Thomas J. Tomberlin; Hebert I. Weisberg
Stable URL:
http://links.jstor.org/sici?sici=0022-4367%28198109%2948%3A3%5C%5C505%5C%5CPAIUC%5C%5CE2.0.CO%5C%5CB2-6

21 **Estimation of Classification Factor Relativities: A Modelling Approach**
Leonard R. Freifelder
Stable URL:
http://links.jstor.org/sici?sici=0022-4367%28198603%2953%3A1%5C%5C135%5C%5C8E%5C%5CFC%5C%5CE2.0.CO%5C%5CB2-H

24 **Discriminatory Bias in Rates Charged by the Canadian Automobile Insurance Industry**
R. A. Holmes
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197003%2965%3A329%5C%5C108%5C%5CDBR%5C%5CE2.0.CO%5C%5CB2-W

29 **Regulation in the Insurance Industry: Determination of Premiums in Automobile Insurance**
Yehuda Kahane; Haim Levy
Stable URL:
http://links.jstor.org/sici?sici=0022-4367%28197503%2942%3A1%5C%5C117%5C%5CRI%5C%5CE2.0.CO%5C%5CB2-8

**NOTE:** The reference numbering from the original has been maintained in this citation list.
Linking Citations

31 Empirical Bayes Methods for Two-Way Contingency Tables
N. M. Laird
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197812%2965%3A3%3C581%3AEBMFTC%3E2.0.CO%3B2-E

34 Tables for An Approximate Test for Outliers in Linear Models
Richard E. Lund
Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28197511%2917%3A4%3C473%3AATFAAT%3E2.0.CO%3B2-L

36 A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models
W. Oberhofer; J. Kmenta
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28197405%2942%3A3%3C579%3AAGPFOM%3E2.0.CO%3B2-1

37 Tests for Departure from Normality: Comparison of Powers
E. S. Pearson; R. B. D'Agostino; K. O. Bowman
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197708%2964%3A2%3C231%3ATFDNC%3E2.0.CO%3B2-8

40 Power of Some Standard Goodness-of-Fit Tests of Normality Against Asymmetric Stable Alternatives
Erwin M. Saniga; James A. Miles
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197912%2974%3A3%3C861%3AAOSGT%3E2.0.CO%3B2-2

41 Estimating Expected Losses in Auto Insurance
Donald T. Sant
Stable URL:
http://links.jstor.org/sici?sici=0022-4367%28198003%2947%3A1%3C133%3AEELIAI%3E2.0.CO%3B2-4

NOTE: The reference numbering from the original has been maintained in this citation list.
43 An Analysis of Variance Test for Normality (Complete Samples)
S. S. Shapiro; M. B. Wilk
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28196512%2952%3A3%2F4%3C591%3AAAOVTF%3E2.0.CO%3B2-B

44 A Comparative Study of Various Tests for Normality
S. S. Shapiro; M. B. Wilk; H. J. Chen
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196812%2963%3A324%3C1343%3AACSOVT%3E2.0.CO%3B2-2

49 A Statistical Perspective on Actuarial Methods for Estimating Pure Premiums from Cross-Classified Data
Herbert I. Weisberg; Thomas J. Tomberlin
Stable URL:
http://links.jstor.org/sici?sici=0022-4367%28198212%2949%3A4%3C539%3AACSPAM%3E2.0.CO%3B2-O

50 Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches
Herbert I. Weisberg; Thomas J. Tomberlin; Sangit Chatterjee
Stable URL:
http://links.jstor.org/sici?sici=0735-0015%28198404%292%3A2%3C170%3A3PILUC%3E2.0.CO%3B2-P

51 Some Large-Sample Tests for Nonnormality in the Linear Regression Model
Halbert White; Glenn M. MacDonald
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198003%2975%3A3%3C16%3A3LTFN%3E2.0.CO%3B2-E

52 Comment on "Discriminatory Bias in Rates Charged by the Canadian Automobile Insurance Industry"
C. L. Wilcken
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197106%2966%3A334%3C289%3A22BIR%3E2.0.CO%3B2-2

NOTE: The reference numbering from the original has been maintained in this citation list.