A SYSTEMATIC RELATIONSHIP BETWEEN MINIMUM BIAS AND GENERALIZED LINEAR MODELS

Stephen Mildenhall

Abstract

The minimum bias method is a natural tool to use in parameterizing classification ratemaking plans. Such plans build rates for a large, heterogeneous group of insureds using arithmetic operations to combine a small set of parameters in many different ways. Since the arithmetic structure of a class plan is usually not wholly appropriate, rates for some individual classification cells may be biased. Classification ratemaking therefore requires measures of bias, and minimum bias is a natural objective to use when determining rates.

This paper introduces a family of linear bias measures and shows how classification rates with minimum (zero) linear bias for each class are the same as those obtained by solving a related generalized linear model using maximum likelihood. The examples considered include the standard additive and multiplicative models used by Insurance Services Office (ISO) for private passenger auto ratemaking and general liability ratemaking, see ISO [10] and Graves and Castillo [7], respectively.

Knowing how to associate a generalized linear model to a linear bias function is useful for several reasons. It makes the underlying statistical assumptions explicit so the user can judge their appropriateness for a given application. It provides an alternative method to solve for the model parameters, which is computationally more efficient than using the minimum bias iterative method. In fact not all linear bias functions allow an iterative solution; in these cases, solving a generalized linear model using maximum likelihood provides an effective way to determine model parameters. Finally, it opens up the possibility of using statistical techniques for parameter estimates, analysis of residuals and model fit, significance of effects, and comparison of different models.

Acknowledgments

I would like to thank Bill Emmons, John Huddleston, and particularly David Bassi and Thom McDaniel for their help and support while I was writing this paper. I am very grateful to David Fennell and Sue Groshong for patiently explaining numerous statistical concepts to me. Debra McClenahan gave me the final impetus to finish the paper. Finally, the comments from Stuart Klugman and the other anonymous reviewers were invaluable and they led to significant changes from the first draft.

1. Introduction

1.1 History and Background

Bailey and Simon [2], [3], first considered bias in classification ratemaking and introduced minimum bias models. Since classification plans use fewer variables than underwriting cells and impose an arithmetic structure on the data, fitted rates in some cells may be biased, that is, not equal to the expected rate. Bias is a feature of the structure of the classification plan and not a result of a small overall sample size: bias could still exist even if there were sufficient data for all the cells to be individually credible. Of course, in such a situation an actuary would not use a classification plan.

In [3] Bailey and Simon proposed their famous list of four criteria for an acceptable set of relativities:

BaS1. It should reproduce experience for each class and overall (balanced for each class and overall).

BaS2. It should reflect the relative credibility of the various groups.

BaS3. It should provide the minimum amount of departure from the raw data for the maximum number of people.

BaS4. It should produce a rate for each sub-group of risks which is close enough to the experience so that the differences could reasonably be caused by *chance*.

Condition BaS1 means that classification rates for each class should be balanced, that is, have zero bias. Obviously, zero bias by class implies zero bias overall.

Bailey points out that since more than one set of rates can be unbiased in the aggregate it is necessary to have a method for comparing between them. The average bias has already been set to zero, by criteria BaS1, and so it cannot be used. Bailey suggests the average absolute deviation and the chi-square statistic, particularly if cells are large enough to assume normality. He mentions that neither of these statistics has a known theoretical distribution and stresses that they should be used for comparison between models and not for tests of significance. This paper shows there is a natural correspondence between linear bias functions and generalized linear models. The theory of generalized linear models can then be used to define and analyze various measures of fit statistically, improving upon Bailey's more ad hoc methods.

In 1988 Brown [5] revisited minimum bias. His approach was to replace the bias function with an expression from the likelihood function and then solve for parameters to maximize its value. By assuming a distribution for the underlying quantity being modeled he converts the problem to "an exercise in statistical modeling." This paper takes the opposite approach and goes *from* a particular class of bias functions *to* a statistical distribution. Brown also comments that "[t]o this point we have not been able to use GLIM [generalized linear models] to reproduce results obtained by Bailey's additive model"; see Section 4.4 below for such a reconciliation.

Venter's review [25] of Brown considers four alternatives to Bailey's methods:

V1. Alternatives to the balance principle.

V2. More general arithmetic functions to determine classification rates.

V3. Allow individual cells to vary from an arithmetically defined base.

V4. Do not use an arithmetic function to determine classification rates.

Venter comments that Brown's paper is mainly concerned with V1. This paper is largely concerned with V1 and V2, but also has comments on V3 and V4. Link functions, introduced below, allow more general arithmetic functions. The Box-Cox transformation, which Venter mentions, is an example of a link function. Section 10 mentions a method related to mixed models which is exactly what Venter proposed in V3 to determine unbiased rates.

Venter also comments that "the connection with general linear models does not seem to be the primary emphasis of [Brown's] paper." This paper builds on Brown's initial work by focusing on the connection between the minimum bias methods and generalized linear models and by providing a more in-depth explanation of generalized linear models based on ideas already familiar to actuaries. Showing how they provide a unified treatment of minimum bias models will give actuaries another reason to learn more about generalized linear models. Other actuarial applications of generalized linear models have been proposed in McCullagh and Nelder [16], Renshaw [22], Haberman and Renshaw [8], and Wright [26].

1.2 Contents

Section 2 recalls some familiar material about linear models and sets up the progression from general linear models to generalized linear models by analyzing the three components of a general linear model.

Section 3 explains the non-uniqueness of solutions to a classification plan and how to get around the problem.

Section 4 explains the elementary, but unfamiliar, relationship between the cross classification ratemaking notation used in minimum bias models and the standard statistical, matrix notation used in linear models. It derives a matrix version of Bailey's minimum bias equations, and shows how Bailey's additive model is a simple linear model. The section ends with a general matrix formulation of balance and introduces a numerical example.

Section 5 introduces a family of linear bias functions and an associated measure of model fit called deviance both related to a variance function. By construction, minimum linear bias corresponds to the minimum deviance best-fit model. It also shows how, in some cases, the minimum bias solution can be obtained using iterative equations.

Section 6 defines the exponential family of distributions and gives several examples. It explains the relationship between variance functions and distributions which is then used to convert the minimum bias models of Section 5 into fully defined statistical models.

Section 7 introduces generalized linear models and their connection with minimum linear bias. This correspondence holds regardless of whether an iterative method can be used to solve the minimum bias problem, so generalized linear models extend the existing family of models. A detailed set of examples, comparing different linear bias assumptions, is also given.

Section 8 discusses measures of model fit associated with generalized linear models. Fit is discussed at several different levels, ranging from selection of covariates to selection of link functions and variance functions.

Section 9 is concerned with numerical computations. It explains how and when the iterative equations obtained using Bailey's minimum bias equations converge. It also discusses how to solve generalized linear models using iteratively re-weighted least squares. Appendix B gives SAS computer code illustrating a hands-on example of this approach.

Section 10 gives some suggestions for future work. It touches on some recent work of Lee and Nelder [18] on mixed models and hierarchical generalized linear models, which can be regarded as an extension of the work in this paper and which provides unbiased predictors for all cells.

The theory is illustrated throughout with simple examples the reader can reproduce.

In the first 7 sections of the paper most concepts are developed from first principles and very little background in statistics is assumed. Sections 8 and 9 make greater demands on the reader, assuming more statistical and mathematical background, respectively.

1.3 Notation

Random variables will be denoted by capitals and realized values in lower case. Vectors will be denoted by bold lower case letters. Matrices will be denoted by bold upper case letters, typically $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ and $\mathbf{W}$. The (ij)th element of a matrix $\mathbf{X}$ will be denoted x_{ij}, $x_{i,j}$ or $\mathbf{X}(i,j)$. Some matrices will be given in block form. If $\mathbf{W}$ is a block matrix then $\mathbf{W}_{ij}$ will denote the block in the (i,j)th place. Superscript t denotes transpose. Random observations are denoted r, r_i, r_{ij}; Greek letters typically refer to model parameters or fitted values. Matrix dimensions are denoted $m \times n$. The end of an example or proof is marked off using a box. ▮

2. Linear Models

A statistical model is defined by specifying a probability distribution for the quantity being modeled. Fitted values, predicted by the model, can then be determined from the relevant probability distribution, usually as the mean. The goal of using a model is to replace the data, which may have many thousands of observations, with a far smaller set of parameters without losing too much information. A good model helps the actuary better understand the data and make reasonable predictions from it. Models can be designed to facilitate the construction of classification ratemaking tables.

In a basic **linear** model the fitted values are linear combinations of the model parameters. Examples of linear models include ANOVA's, linear regression, and general linear regression.

In order find model parameter values it is necessary to select an objective function. The objective function can measure the deviance between the underlying data and the fitted values for different parameter choices, or it can be based on other criteria such as minimum variance amongst unbiased estimators. Least squares and maximum likelihood are two common examples of the former type of objective function. A single statistical model can give rise to different parameter solutions depending upon the objective function used. Therefore it is necessary to include the objective function in an effective description of the model.

The input data for all models considered here can be given as a two dimensional array. The rows correspond to the different observations or **units**. The first column corresponds to the **response** variate which can be continuous, such as pure premium, frequency or severity, or discrete, such as claim count. The remaining columns correspond to the explanatory variates, or **covariates**, whose values are supposed to explain the values of the response. Covariates can be qualitative or quantitative. A qualitative covariate, called a **factor**, takes on non-numerical values called **levels**, such as vehicle use, vehicle type or sex. Quantitative covariates have numeric values. Examples include age, time, weight of vehicle, or price of vehicle. Age *group* is a qualitative covariate. If the covariates are all factors then the rows of the input can be labeled by the levels of the factors (as in Example 2.1 below). Classification ratemaking naturally uses these coordinates. However, they are generally not used if some of the some covariates are continuous, as in Example 2.2.

2.1 Example

A two-way analysis of variance with no interactions assumes each observation r_{ij} is a realization of an independent, normally distributed random variable with mean $a_i + b_j$ and variance σ^2. Parameters are selected using either maximum likelihood, minimum square error, or minimum variance amongst unbiased estimators; the three are equivalent for this model. The a_i and b_j are the **effects** corresponding to the different levels of the two factors (classification variables). In texts on linear models this example is often presented in the equivalent form $r_{ij} = a_i + b_j + e_{ij}$ where the errors e_{ij} are independent $N(0, \sigma^2)$ random variables. For example, r_{ij} could be the observed pure premium in cell i, j of an auto classification plan, with a_i the factor for age of operator group i and b_j the factor for vehicle use group j. If r_{ij} is the average of w_{ij} exposures, then it is a realization of a variable with variance σ^2/w_{ij} and w_{ij} is called the **weight** of the i, jth cell. ▮

2.2 Example

A linear regression model assumes each observation r_i is a realization of an independent, normally distributed random variable with mean $a + bx_i$ and variance σ^2. There is a single continuous covariate whose values are given by x_i. The same three objectives can be used to solve for a and b. The model can also be written $r_i = a + bx_i + e_i$ with e_i independent $N(0, \sigma^2)$ random variables. Actuaries use linear regression to compute trends, in which case r_i is the observed pure premium, or log of pure premium, at time i and $x_i = i$. ▮

The input data for a linear model can be compactly described using vectors and matrices. Suppose there are n observations. The responses can be put into an $n \times 1$ column vector $\mathbf{r} = (r_1, \dots, r_n)^t$. The covariates can be arranged into a **design** matrix $\mathbf{X}$ which has one row for each observation and one column for each parameter of the model. Let p be the number of parameters and let $\mathbf{x}_i$ be the ith row of $\mathbf{X}$, so $\mathbf{x}_i$ is a $1 \times p$ row vector. If all the covariates are factors then the design matrix has one column for each level of each factor and consists of 0's and 1's. In Example 2.1, if there are three age groups and three vehicle use classes then the design matrix would have six columns. In Example 2.2, $\mathbf{X}$ has two columns, corresponding to a and b. The first column is all 1's, corresponding to the constant term; the second is given by $(x_1, \dots, x_n)^t$.

The parameters of a linear model can be arranged into a $p \times 1$ column vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$. Finally, let $\mu_i = \mathrm{E}(R_i)$ be the fitted value of the ith response and let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^t$. A general linear model, which includes both analysis of variance and linear regression as special cases, assumes

$$\mathbf{r} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{2.1}$$

where the error term $\mathbf{e} = (e_1, \dots, e_n)^t$, has e_i independent, normally distributed with mean 0 and variance σ^2. Thus R_i is assumed to be independent, normally distributed with mean $\mu_i = \mathbf{x}_i\boldsymbol{\beta}$ and variance σ^2.

Three important assumptions underlie a general linear model.

(1) Constant variance: the σ term does not vary between different responses. When the ith response is an average of w_i individual responses each with variance σ^2 then the variance is σ^2/w_i, and again σ does not vary between observations. The w_i are prior weights.

(2) Normality of errors: the errors e_i are independent, identically distributed normal random variables.

(3) Linear: the fitted value $\mu_i = \mathbf{x}_i\boldsymbol{\beta} = \sum_j \mathbf{x}_{ij} b_j$ is a linear combination of the parameters, so the systematic effects are additive.

In actuarial work it is common that the responses are averages from populations with different sizes. In Example 2.1, there are typically more exposures in the mature operator classes than in youthful and senior operator classes. General linear models allow for such differences in variance by using prior weights which vary by observation—as in (1).

The second assumption, normal errors, is frequently a problem in actuarial applications. Losses, severities, pure premiums and frequencies, are all positive and generally positively skewed; they are therefore not normally distributed. The log transformation is often applied to the data prior to using a linear model in order to improve normality. The log transformation is also applied in order to convert multiplicative effects into additive ones.

2.3 Example

Example 2.1 modeled R_{ij} as normally distributed with mean $a_i + b_j$ and variance σ^2/w_{ij}, where w_{ij} is the number of exposures in the i, jth cell. Applying the log transformation to the response we can consider the same model for $\log(R_{ij})$. On the untransformed scale, the model for R_{ij} is lognormal with parameters $a_i + b_j$ and σ^2/w_{ij}, see the Appendix to Hogg and Klugman, [9], or Appendix A of Klugman, Panjer and Willmot [15]. The systematic effects are now multiplicative. Also $\mathrm{E}(R_{ij}) = \exp(a_i + b_j + \sigma^2/2w_{ij})$ and the variance depends on the fitted mean because $\mathrm{Var}(R_{ij}) = (\mathrm{E}(R_{ij}))^2(\exp(\sigma^2/2w_{ij}) - 1)$. ▮

In order to set up *generalized* linear models, consider a general linear model as split into three components.

GLM1. A random component: observations r_i are assumed to come from an independent normal distribution R_i with $\mathrm{E}(R_i) = \mu_i$.

GLM2. A systematic component: the covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$ give a linear predictor

$$\eta_i = \sum_j x_{ij}\beta_j.$$

GLM3. A link between the random and systematic components:

$$\eta = \mu.$$

The parameters are selected using the maximum likelihood objective.

A generalized linear model allows extensions to GLM1 and GLM3. GLM2 is retained since the model is still assumed to be *linear.*

Assumption GLM1 is generalized to allow the R_i to have a distribution from the exponential family, defined in Section 6. The exponential family includes the normal, Poisson, binomial, gamma and inverse Gaussian distributions. The lognormal distribution is not a member of the exponential family. The recent book by Jørgensen [12] is a good reference on exponential distributions.

In GLM3, the identity link $\eta_i = \mu_i$ between the random and systematic components is generalized to allow $\eta_i = g(\mu_i)$ for any strictly monotonic, differentiable function g. Three common choices are $g(x) = x$, $g(x) = \log(x)$ and $g(x) = 1/x$. The log-link has been discussed above. The reciprocal-link can be understood as representing rates: premium is the dollar rate per year; the reciprocal premium is therefore years of coverage per dollar premium. While not something that has been tried to date in actuarial applications, there is no reason why the systematic effects should not be additive on the reciprocal scale. McCullagh and Nelder, [16] Section 8.4 gives an insurance example.

In a general linear model, scale transformations may be applied to the responses prior to fitting in order to increase the validity GLM1-3. However, the three assumptions may be mutually incompatible and so the question of an appropriate scale can be very problematic—see [16] Section 2.1 for an example. For a generalized linear model, normality and constant variance are no longer required. The choice of link-function (scale) is therefore driven solely by the need to ensure additivity of effects. Since transformations in generalized linear models are used to achieve one end, rather than three in a general linear model, there is more flexibility in the modeling process.

The next three examples illustrate how generalized linear models include, extend, and differ from general linear models.

2.4 Examples

(a) A generalized linear model with identity link function and normal errors is a general linear model.

(b) A generalized linear model version of Example 2.1 with gamma error distribution and a reciprocal link would model R_{ij} as an independent gamma random variable with $\mathrm{E}(R_{ij}) = \mu_{ij} = 1/(a_i + b_j)$ and $\mathrm{Var}(R_{ij}) = \mu_{ij}^2\phi/w_{ij}$. The constant ϕ acts like σ in Example 2.1.

(c) A generalized linear model with log link and normal errors is *not* the same as applying a general linear model to the log responses. The generalized linear model assumes R_{ij} is *normally* distributed with mean $\exp(a_i + b_j)$ and variance σ^2/w_{ij}. The general linear model applied to the log transformed data (Example 2.3) assumes that R_{ij} is *lognormally* distributed and that $\log(R_{ij})$ has mean $a_i + b_j$ and variance σ^2/w_{ij}. In the generalized linear model the log-link is only trying to achieve additivity of effects; the error distribution is specified separately. Exhibit 8, described fully in Section 7.7, shows the differences between these models applied to an example dataset. ∎

3. Uniqueness of Parameters

Going back to Example 2.1, it is clear that the parameters of a linear model need not be unique. If $\mu_{ij} = a_i + b_j$ then

$$\mu_{ij} = (a_i + \alpha) + (b_j - \alpha) \tag{3.1}$$

for all constants α. Similarly, if $\mu_{ij} = a_i b_j$ then $\mu_{ij} = (\alpha a_i)(b_j/\alpha)$ for all constants $\alpha \neq 0$. If the model is $\mu_{ijk} = a_i + b_j + c_k$ then the situation is even worse: there are two degrees of freedom because $\mu_{ijk} = (a_i + \alpha_1 + \alpha_2) + (b_j - \alpha_1) + (c_k - \alpha_2)$ for all α_1 and α_2. In general, it is easy to see there are $q - 1$ degrees of freedom when there are q classification variables. Therefore it is necessary to select $q - 1$ base classes in order to have unique parameters. This is familiar from setting up rate classification plans. For example the personal auto plan has one base rate for the married, aged 25-50, pleasure use, single standard vehicle, zero points class and deviations for all other classes.

There is no canonical method for selecting the base classes needed to ensure unique parameters. Here is one possible approach. First select one classification *cell* as a base. Then, select one classification *variable* which will not have a base. Finally, set the parameters corresponding to the base class in all the other classification variables to zero (additive models) or one (multiplicative models). This specifies the values of $q-1$ parameters and so removes all degrees of freedom. Now the parameters for all the non-base classification variables are deviations from the base class for that variable. Picking different base classes leads to different parameters, but the fitted values remain the same.[1]

In Example 2.1, we could select mature drivers and pleasure use as the base cell, and age as the base classification. This forces pleasure use to be the base class in the vehicle use classification, and so the parameter for pleasure use is set to zero. Since b_1 corresponds to pleasure use, this choice is the same as selecting $\alpha = b_1$ in Equation (3.1).

In conclusion, a linear model or minimum bias method which uses all the available parameters will generally not have a unique solution. However, the non-uniqueness is of a trivial nature and the fitted values will be unique. After making an arbitrary selection of base classes the remaining parameters will be unique. This is what Bailey and Simon [3] mean when they say "[the parameters] can only be regarded in relationship to the coordinate system in which they find themselves."

[1] Selecting base classes corresponds to deleting columns from the design matrix. Selecting $q-1$ base classes ensures that the resulting design matrix $\hat{X}$ has maximal rank. This in turn implies $\hat{X}^t\hat{X}$ is invertible and so the normal equations can be solved uniquely for the remaining parameters. In general linear models, non-uniqueness is handled by computing the generalized inverse of X^tX. The generalized inverses can be regarded as a method for picking base classes. See Rao, [21], Chapter 1b.5 for more details.

4. Matrix Formulation

As Venter noted in his discussion [25] it is not clear to those unfamiliar with linear models how they are related to minimum bias methods. Moreover, the translation from statistical linear models to minimum bias methods is hampered by different uses of the same notation. We will follow Brown's notation as much as possible, since actuaries are probably most familiar with his approach. This section explains the relationship between linear models and minimum bias methods and provides a dictionary to translate between the two. In order to keep difficulties of notation in the background we only consider a simple additive model with two variables. Extensions to more general models are easy to work out—indeed the point of this section is to convince the reader they will work out just as expected. The auto classification plan will be used to provide examples.

4.1 Minimum Bias Method Language

The generic minimum bias method attempts to explain a collection of observed values r_{ij} with two sets of parameters x_i and y_j, $i = 1, \dots, n_1$, $j = 1, \dots, n_2$. For example, r_{ij} could be the pure premium in the (i, j)th cell, x_i may correspond to the ith age classification and the y_j to the jth vehicle use classification such as pleasure, drive to work, or business. Let w_{ij} denote the number of exposures in the (i, j)th cell. Minimum bias methods then give iterative equations to solve for the x_i's and y_j's.

For example, Bailey's additive method models r_{ij} as x_i+y_j (hence the appellation "additive") in such a way that, for all i,

$$\sum_j w_{ij}(r_{ij} - (x_i + y_j)) = 0, \tag{4.1}$$

and similarly for j. Equation (4.1) means that the model is balanced, i.e. has zero weighted bias, for each class i and in total (summing over i), and so minimizes bias. Rearranging Equation (4.1) gives the familiar form of Bailey's additive method:

$$x_i = \sum_j w_{ij}(r_{ij} - y_j) \Bigg/ \sum_j w_{ij}, \tag{4.2}$$

and similarly

$$y_j = \sum_i w_{ij}(r_{ij} - x_i) \Bigg/ \sum_i w_{ij}. \tag{4.3}$$

This notation is shorthand for an iterative procedure, where the transition from the lth to $l+1$st iteration is

$$x_i^{(l+1)} = \sum_j w_{ij}(r_{ij} - y_j^{(l)}) \Bigg/ \sum_j w_{ij},$$

and similarly for $y_j^{(l+1)}$ in terms of $x_i^{(l+1)}$. The final result of the iterative procedure is given by $x_i = \lim_{l\to\infty} x_i^{(l)}$, and similarly for y.

4.2 Translation

The key to translating from minimum bias notation to linear model notation is how the observations are indexed. In linear models they are indexed by one parameter, whereas in the minimum bias method they are indexed by two parameters (or more generally, by the number of classification variables). The translation is described by the following correspondences. In all cases the left hand side gives the minimum bias notation and the right hand side the linear model notation. Also, in this section commas are inserted between subscript indices for clarity. The difference in how observations are indexed is illustrated by the following two correspondences between $n_1 n_2 \times 1$ column vectors:

$$\begin{pmatrix} r_{1,1} \\ r_{1,2} \\ \vdots \\ r_{1,n_2} \\ r_{2,1} \\ \vdots \\ r_{n_1,1} \\ \vdots \\ r_{n_1,n_2} \end{pmatrix} \leftrightarrow \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_{n_2} \\ r_{n_2+1} \\ \vdots \\ r_{(n_1-1)n_2+1} \\ \vdots \\ r_{n_1 n_2} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} w_{1,1} \\ w_{1,2} \\ \vdots \\ w_{1,n_2} \\ w_{2,1} \\ \vdots \\ w_{n_1,1} \\ \vdots \\ w_{n_1,n_2} \end{pmatrix} \leftrightarrow \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n_2} \\ w_{n_2+1} \\ \vdots \\ w_{(n_1-1)n_2+1} \\ \vdots \\ w_{n_1 n_2} \end{pmatrix}.$$

The different levels of the two classifications (or effects) correspond as

$$\begin{pmatrix} x_1 \\ \vdots \\ x_{n_1} \\ y_1 \\ \vdots \\ y_{n_2} \end{pmatrix} \leftrightarrow \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{n_1} \\ \beta_{n_1+1} \\ \vdots \\ \beta_{n_1+n_2} \end{pmatrix}. \tag{4.4}$$

Let $n = n_1 n_2$ be the number of observations and $p = n_1 + n_2$ be the number of model parameters. Our translation assumes there are observations for each of the $n = n_1 n_2$ possible combinations of x_i and y_j. If necessary, the model can be brought into this form by using zero weights in any empty cells.

4.3 Linear Model Language

A statistical linear model attempts to explain a collection of observed values r_i using linear combinations of a smaller number of parameters. In our setting, the model explains pure premiums r_i, $i = 1, \dots, n$ using linear combinations of parameters $\beta_1, \dots, \beta_p$ given by

$$r_i = \sum_j x_{ij}\beta_j + e_i,$$

where e_i is a random error term. In matrix language this can be written

$$\mathbf{r} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{X} = (x_{ij})$ is the $n \times p$ design matrix of covariates, and $\mathbf{r} = (r_1, \dots, r_n)^t$, $\mathbf{e} = (e_1, \dots, e_n)^t$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$ are column vectors.

The design matrix corresponding to the two variable additive linear model is the $n \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{I} \\ \vdots & \vdots \\ \mathbf{A}_{n_1} & \mathbf{I} \end{pmatrix} \tag{4.5}$$

where

$$\mathbf{A}_i = \begin{pmatrix} 0 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & \dots & 0 \end{pmatrix} \qquad \text{dimension } n_2 \times n_1 \tag{4.6}$$

has zero entries except for 1's in the ith column, and $\mathbf{I}$ is the $n_2 \times n_2$ identity matrix. (In Equation (4.6) we have put the dimensions of the matrix to the right. Knowing the dimensions is very useful for keeping track of what is going on.) Using the block matrix form of $\mathbf{X}$, and the translation Equation (4.4) it is easy to see that

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X} \begin{pmatrix} x_1 \\ \vdots \\ x_{n_1} \\ y_1 \\ \vdots \\ y_{n_2} \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_1 + y_{n_2} \\ x_2 + y_1 \\ \vdots \\ x_2 + y_{n_2} \\ \vdots \\ x_{n_1} + y_1 \\ \vdots \\ x_{n_1} + y_{n_2} \end{pmatrix}$$

(dimensions $(n \times p)(p \times 1) = n \times 1$) demonstrating the translation between minimum bias notation and linear model notation.

4.4 Solution of linear models

It is well known that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies the following normal equations under the assumption of independent and identically distributed normal errors (see Rao [21], Section 4a.2)

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{r}. \tag{4.7}$$

If observation i has weight w_i, the solution satisfies

$$\mathbf{X}^t\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{W}\mathbf{r}, \tag{4.8}$$

where $\mathbf{W} = \operatorname{diag}(w_1, \dots, w_n)$ is the diagonal matrix of weights.

Next we compute Equation (4.8), for the two variable additive model presented in Section 4.1 using the translations introduced in Sections 4.2 and 4.3. In minimum bias notation, the matrix of weights can be written as a block matrix

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{W}_{n_1} \end{pmatrix} \qquad \text{dimension } n \times n \tag{4.9}$$

where

$$\mathbf{W}_i = \begin{pmatrix} w_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_{in_2} \end{pmatrix} \qquad \text{dimension } n_2 \times n_2.$$

Using the block matrix form of $\mathbf{X}$ and $\mathbf{W}$ it is a simple computation to show

$$\mathbf{X}^t\mathbf{W}\mathbf{X} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}^t & \mathbf{D} \end{pmatrix} \qquad \text{dimension } p \times p$$

where $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are given by

$$\mathbf{B} = \begin{pmatrix} \sum_j w_{1j} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_j w_{n_1 j} \end{pmatrix} \qquad \text{dimension } n_1 \times n_1,$$

$$\mathbf{C} = \begin{pmatrix} w_{11} & \dots & w_{1n_2} \\ \vdots & & \vdots \\ w_{n_1 1} & \dots & w_{n_1 n_2} \end{pmatrix} \qquad \text{dimension } n_1 \times n_2,$$

and

$$\mathbf{D} = \begin{pmatrix} \sum_i w_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_i w_{in_2} \end{pmatrix} \qquad \text{dimensions } n_2 \times n_2.$$

Therefore

$$\begin{aligned} \mathbf{X}^t\mathbf{W}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^t\mathbf{W}\mathbf{X}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}^t & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}\mathbf{x} + \mathbf{C}\mathbf{y} \\ \mathbf{C}^t\mathbf{x} + \mathbf{D}\mathbf{y} \end{pmatrix}, \end{aligned}$$

giving the $p \times 1$ vector equality

$$\mathbf{X}^t\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} x_1 \sum_j w_{1j} + \sum_j w_{1j} y_j \\ \vdots \\ x_{n_1} \sum_j w_{n_1 j} + \sum_j w_{n_1 j} y_j \\ \sum_i w_{i1} x_i + y_1 \sum_i w_{i1} \\ \vdots \\ \sum_i w_{in_2} x_i + y_{n_2} \sum_i w_{in_2} \end{pmatrix}. \tag{4.10}$$

On the other hand

$$\mathbf{X}^t\mathbf{W}\mathbf{r} = \begin{pmatrix} \sum_j w_{1j} r_{1j} \\ \vdots \\ \sum_j w_{n_1 j} r_{n_1 j} \\ \sum_i w_{i1} r_{i1} \\ \vdots \\ \sum_i w_{in_2} r_{in_2} \end{pmatrix} \qquad p \times 1. \tag{4.11}$$

Equating corresponding rows of Equation (4.10) and Equation (4.11)—the normal equations—gives exactly Equation (4.2) and Equation (4.3), demonstrating that the results of a two effect additive general linear model are the same as the Bailey additive method.

This is a significant result for several reasons. Firstly, it shows the minimum bias parameters are the same as the maximum likelihood parameters assuming normal errors, which the user may or may not regard as a reasonable assumption for his or her application. Secondly, it is much more efficient to solve the normal equations than perform the minimum bias iteration, which typically converges quite slowly (see Section 9). Thirdly, knowing that the result is the same as a linear model allows the statistics developed to analyze linear models to be applied. For example, information about residuals and influence of outliers can be used.

4.5 General theory and a matrix formulation of balance

It is easy to generalize the preceding discussion to the case of a general linear model with q classification variables. Let the ith classification variable have n_i

levels, $i = 1, \ldots, q$. Thus there are $p = n_1 + \cdots + n_q$ different parameters and, assuming no empty cells, $n = n_1 \cdots n_q$ observations.

The minimum bias notation associates an $n \times n_i$ design matrix $\mathbf{A}_i$ and an $n_i \times 1$ parameter vector $\mathbf{a}_i$ with the ith classification variable. The $n \times 1$ vector of modeled rates $\boldsymbol{\mu} = (\mu_{1,\ldots,1}, \ldots, \mu_{n_1,\ldots,n_q})^t$ is

$$\boldsymbol{\mu} = (\mathbf{A}_1 \ldots \mathbf{A}_q) \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_q \end{pmatrix} = \mathbf{A}_1\mathbf{a}_1 + \cdots + \mathbf{A}_q\mathbf{a}_q. \tag{4.12}$$

In linear model language, the design matrix $\mathbf{X}$ has dimension $n \times p$ and equals the horizontal concatenation $(\mathbf{A}_1 \cdots \mathbf{A}_q)$. The parameter vector $\boldsymbol{\beta}$ has dimension $p \times 1$ and equals $(\mathbf{a}_1, \ldots, \mathbf{a}_q)^t$ and Equation (4.12) becomes $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

Note that the linear model notation makes it possible to use two-dimensional matrix notation to describe models with any number of classification variables.

Using this notation and the same approach used to derive Equation (4.10) and Equation (4.11) shows that the normal equation condition

$$\mathbf{X}^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) = \mathbf{0} \tag{4.13}$$

is exactly a matrix formulation of condition BaS1, that relativities be balanced by class. This interpretation of Equation (4.13) is important and will be used repeatedly below.

To see why Equation (4.13) is the balance condition, first use the translation of 4.2 to write it as

$$\mathbf{X}^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{A}_1^t \\ \vdots \\ \mathbf{A}_q^t \end{pmatrix} \mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{A}_1^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) \\ \vdots \\ \mathbf{A}_q^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) \end{pmatrix} = \mathbf{0}. \tag{4.14}$$

Consider balance over the first level of the first classification variable. By permuting columns of $\mathbf{X}$ this can be done without loss of generality. Similarly, by permuting the observations, assume that $\mathbf{A}_1$ has the form

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

the vertical concatenation of n_1 different matrices each with $n_2 \cdots n_q$ rows and n_1 columns and one column of ones. Then the first row of Equation (4.14) is given by the sum product of the first column $\mathbf{A}_1$ (i.e. the first *row* of $\mathbf{A}_1^t$) of with $\mathbf{W}(\mathbf{r} - \boldsymbol{\mu})$, which gives

$$\sum_{j_2,\dots j_q} w_{1,j_2,\dots j_q}(r_{1,j_2,\dots j_q} - \mu_{1,j_2,\dots j_q}) = 0,$$

exactly the sum over all other classes required by the balance condition.

4.6 Numerical Example

We now introduce a numerical example which will be used throughout the paper to illustrate the theory. The data, shown in Exhibit 1, gives average claim severity for private passenger auto collision[2]. The severities have been adjusted for severity trend. There are $n = 32$ observations and two classification variables: age group and vehicle use. Age group has eight levels and vehicle use four. The response variable r is average claim severity. The weights w are given by the number of claims underlying the average severity. Exhibit 2 gives the one way weighted average severities.

[2] The data is derived from McCullagh and Nelder's example [16] Section 8.4.

Exhibit 3 gives the design matrix **A** corresponding to the age group classification. **A** has the block form shown in Equation (4.6). Exhibit 4 gives the design matrix **B** corresponding to the vehicle use classification. Pleasure use has been selected as the base (as in Section 3) and the corresponding column of the design matrix has been deleted; this accounts for the rows of zeros. The design matrix for the whole model is $\mathbf{X} = (\mathbf{A}\ \mathbf{B})$. Except for the deleted column in **B**, **X** has the form given in Equation (4.5).

Exhibit 5 uses the iterative method, Equation (4.2) to fit an additive minimum bias model to the data. There are 50 iterations shown (column 1). Column 2 shows the length of the change in the parameter vector from one iteration to the next. Columns 3-13 show how the parameters change with each iteration. Columns 14-17 will be explained in Section 9. Exhibit 6 shows the solution to the normal equations Equation (4.8). The resulting parameters are all within 2 cents of the values in the last row of Exhibit 5 as expected. Had more iterations been performed the results would have been closer.

This example will be continued in Sections 7, 8 and 9.

5. Bias Functions and Deviance Functions

Bailey's first criterion for a set of classification relativities, that rates be balanced (unbiased) for each class and in total, makes it necessary for the actuary to be able to measure the bias in a set of rates. Bailey's third and fourth conditions, which require a minimum departure from the raw data and a departure that could be caused by chance, make it necessary to measure the deviance between the fitted rates and the data and to quantify its likelihood.

In the papers on minimum bias discussed in the Introduction, none of the authors differentiated between a measure of bias and a measure of deviance. A measure of bias should be proportional to the predicted value minus the observed value and can be positive or negative. A measure of deviance, or model goodness of fit, should be like a distance: always positive with a minimum of zero for an exact fit (zero bias). Deviance need not be symmetric; we may care more about negatively biased estimates than positively biased ones or, vice versa.

This section will introduce three concepts: variance functions, linear bias functions and deviance functions, and then show how they are related. All three concepts have to do with specifying distributions—a key part of a statistical model. However, they are independent of the choice of covariates.

In this section r denotes the response, with individual units being r_i, or r_{ij} in the example. The fitted means are μ or μ_i.

Ordinary bias is the difference $r - \mu$ between an observation r and a fitted value μ. When adding the biases of many observations and fitted values, there are two reasons why it may be desirable to give more or less weight to different observations. Firstly, if the observations come from cells with different numbers of exposures then their variances will be different. As explained in Section 2, this possibility is handled using prior weights for each observation.

The second reason to weight the biases of individual observations differently is if the variance of the underlying distribution is a function of its mean (the fitted value). This is a very important departure from normal distribution models where the prior weights do not depend on the fitted values. In Example 2.1, r_{ij} is a sample from R_{ij} which is normally distributed with mean $\mu_{ij} = a_i + b_j$ and variance σ^2. The variance is independent of the mean. In Example 2.4(b), r_{ij} is a sample from R_{ij} which has a gamma distribution with mean μ_{ij} and variance $\phi\mu_{ij}^2$ (assume all weights are 1). Now the variance of an individual observation is a function of the fitted cell mean μ_{ij}. Clearly, large biases from a cell with a large mean are more likely, and should be weighted less, than those from a cell with a small mean. In this situation we will use variance functions to give appropriate weights to each cell when adding biases. Once again, it is important to realize variance functions are not a feature of normal distribution models and that they represent a substantial generalization.

A **variance function**, typically denoted V, is any strictly positive function of a single variable. Three examples of variance functions are $V(\mu) \equiv 1$ for $\mu \in (-\infty, \infty)$, $V(\mu) = \mu$ for $\mu \in (0, \infty)$, and $V(\mu) = \mu^2$ also for $\mu \in (0, \infty)$. It should not be a surprise that the first can arise from the normal distribution and the last can arise from the gamma distribution.

Combining variance functions and prior weights—the two reasons to weight biases from individual cells differently—define a **linear bias function** to be a function of the form

$$\mathfrak{b}(r; \mu) = \frac{w(r-\mu)}{V(\mu)}$$

where V is a variance function and w is a prior weight. The weight may vary between observations, but is not a function of the observation or of the fitted value.

In applications there would be many observations r_i, each with a fitted value μ_i and possibly different weights w_i. The total bias would then be

$$\sum_i \mathfrak{b}(r_i; \mu_i) = \sum_i \frac{w_i(r_i - \mu_i)}{V(\mu_i)}.$$

The functions $r - \mu$, $(r-\mu)/\mu$, and $(r-\mu)/\mu^2$ are three examples of linear bias functions, each with $w = 1$, corresponding to the variance functions given above.

A **deviance function** is some measure of the distance between an observation r and a fitted value μ. The deviance $d(r; \mu)$ should satisfy the following two conditions common to a distance.

Dev1. $d(r; r) = 0$ for all r, and

Dev2. $d(r; \mu) > 0$ for all $r \neq \mu$.

The weighted squared difference $d(r;\mu) = w(r-\mu)^2$, $w > 0$, is an example of a deviance function.

An important difference between bias and deviance is that deviance, which corresponds to distance, is always positive while bias can be positive or negative. Deviance can be regarded as a value judgment: "how concerned am I that r is this far from μ?" Deviance functions need not be symmetric about $r = \mu$.

It is possible to associate a deviance function to a linear bias function by defining

$$d(r;\mu) := 2w \int_{\mu}^{r} \frac{(r-t)}{V(t)} dt. \tag{5.1}$$

Clearly this definition satisfies Dev1 and Dev2. Note that by the Fundamental Theorem of Calculus

$$\frac{\partial d}{\partial \mu} = -2w \frac{(r-\mu)}{V(\mu)}.$$

5.1 Examples of deviance functions

(a) If $\mathfrak{b}(r;\mu) = r - \mu$ is ordinary bias then

$$d(r;\mu) = 2\int_{\mu}^{r} (r-t)dt = (r-\mu)^2$$

is the squared distance deviance, with weight $w = 1$.

(b) If $\mathfrak{b}(r;\mu) = (r-\mu)/\mu^2$ corresponds to $V(\mu) = \mu^2$ for $\mu \in (0,\infty)$ then

$$\begin{aligned} d(r;\mu) &= 2\int_{\mu}^{r} \frac{(r-t)}{t^2} dt \\ &= 2\left(\frac{r-\mu}{\mu} - \log\left(\frac{r}{\mu}\right)\right), \end{aligned}$$

again with weight $w = 1$. In this case the deviance is not symmetric about $r = \mu$. Figure 1 shows plots of the gamma density and corresponding deviance function for three different means μ.

(c) The deviance $d(r;\mu) = w|r-\mu|$, $w > 0$, is an example which does not correspond to a linear bias function. ∎

Returning to the case of multiple observations r_i with fitted values μ_i, the total deviance is

$$D = \sum_i d_i = \sum_i d(r_i; \mu_i).$$

Suppose $\mu_i = h(\mathbf{x}_i\boldsymbol{\beta})$ is a function of a linear combination of covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ and parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ as it would be in the generalized linear model setting.[3] We find the minimum deviance over the parameter vector $\boldsymbol{\beta}$ by solving the system of p equations

$$\frac{\partial D}{\partial \beta_j} = 0 \tag{5.2}$$

$j = 1, \dots, p$. Using the chain rule and assuming the deviance function is related to a linear bias function as in Equation (5.1) gives

$$\begin{aligned} \frac{\partial D}{\partial \beta_j} &= \sum_i \frac{\partial d_i}{\partial \beta_j} \\ &= \sum_i \frac{\partial d_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} \\ &= -2\sum_i \frac{w_i(r_i - \mu_i)}{V(\mu_i)} h'(\mathbf{x}_i\boldsymbol{\beta})x_{ij}. \end{aligned} \tag{5.3}$$

Let $\mathbf{X}$ be the design matrix with rows $\mathbf{x}_i$, $\mathbf{W}$ be the diagonal matrix of weights with i,ith element $w_i h'(\mathbf{x}_i\boldsymbol{\beta})/V(\mu_i)$, and $\boldsymbol{\mu}$ equal $(h(\mathbf{x}_1\boldsymbol{\beta}), \dots, h(\mathbf{x}_n\boldsymbol{\beta}))^t$. Then Equation (5.3) can be written as

$$\mathbf{X}^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu}) = \mathbf{0} \tag{5.4}$$

which by Equation (4.13) is the zero bias equation. This shows that Bailey and Simon's balance criteria, BaS1, is equivalent to a minimum deviance criteria when

[3] The function h is the inverse of the link function which will be introduced in Section 6. The link function g relates the linear predictor to the mean: $\mathbf{x}_i\boldsymbol{\beta} = g(\mu_i)$.

bias is measured using a linear bias function and weights are adjusted for the link function and form of the model using $h'(\mathbf{x}_i\boldsymbol{\beta})$.

The adjustment in Equation (5.3), given by $h'(\mathbf{x}_i\boldsymbol{\beta})x_{ij}$, depends upon the form of the underlying statistical model. This shows clearly how the bias function (which is related to the underlying distribution) and the form of the linear model (link and covariates) both impact the minimum bias parameters. The separation mirrors that between the error distribution and the link function exhibited in GLM1 and GLM3.

5.2 Examples of minimum bias models

(a) $V \equiv 1$ and $h(x) = x$ reproduces the familiar additive minimum bias model which has already been considered in Section 4.

(b) Let $V(\mu) = \mu$ and $h(x) = e^x$. Using the minimum bias notation from Section 4, the minimum deviance condition Equation (5.3), which sets the bias for the ith level of the first classification variable to zero, is

$$\sum_{j=1}^{n_2} \frac{w_{ij}(r_{ij} - e^{a_i+b_j})}{e^{a_i+b_j}} e^{a_i+b_j} = \sum_{j=1}^{n_2} w_{ij}(r_{ij} - e^{a_i+b_j}) = 0,$$

including the link-related adjustment. Therefore

$$e^{a_i} = \sum_j w_{ij} r_{ij} \Big/ \sum_j w_{ij} e^{b_j}$$

and similarly

$$e^{b_j} = \sum_i w_{ij} r_{ij} \Big/ \sum_i w_{ij} e^{a_i}$$

giving Bailey's multiplicative model. ▮

See Section 7.4 for many more examples.

5.3 Summary

The definitions of linear bias function and deviance function have set up a natural correspondence

$$\text{Deviance} \xrightarrow{\frac{\partial}{\partial\mu}} \text{Linear Bias Function}$$
$$d(y;\mu) \to \frac{\partial d}{\partial\mu}$$
$$\int_\mu^r \mathfrak{b}(r;t)dt \to -\mathfrak{b}(r;\mu)$$
$$\text{Minimum Deviance} \to \text{Zero bias by class.}$$

It follows from these definitions that the balance criterion sets the average bias to zero. However, except in trivial cases the total minimum deviance is non-zero and is available as a model-fit statistic which can be used to select between models. This is an important step, especially since deviance has a reasonably well understood distribution. It is developed in Section 8.

Many minimum bias equations can now be derived using different link functions and linear bias functions, several of which lead to iterative equations. Everything in this section has been developed with no explicitly defined statistical model—since no probability distributions have been mentioned. Leaving out the statistical model makes the presentation more elementary and focuses on the intuitively reasonable roles of bias and deviance. In order to put minimum bias methods onto a firm statistical footing, a goal of the paper, we turn next to the theory of generalized linear models and exponential distributions and its relation to linear bias functions and deviance.

6. Exponential Distributions

The following diagram gives a schematic of Section 5 for the normal distribution.

$$\begin{tikzcd}
\text{Balanced by class} & \text{Linear Bias Function } = r - \mu \arrow[l] \\
\text{Least Squares} \arrow[u, "\text{Differentiation}"] \arrow[d, equal] & \text{Deviance Function } = (r-\mu)^2 \arrow[l] \arrow[u, "\text{Differentiation}"'] \arrow[d] \\
\text{Max Likelihood Params} & \text{Normal Distribution.} \arrow[l]
\end{tikzcd}$$

To generalize to arbitrary linear bias functions we need a family of distributions extending the normal which fills out the lower right hand corner of the diagram. It should have a likelihood function related to the given deviance function in the same way as the normal likelihood is related to the square distance deviance. Solving maximum likelihood for μ should correspond to minimum deviance and will give balanced (according to the appropriate notion of bias) classification factors. The required family of distributions is called the exponential family. This section will define it and derive some of its important properties.

The **exponential family** of distributions[4] is the two-parameter family whose density functions can be written in the form

$$f(r;\mu,\phi) = c(r,\phi)\exp\left(-\frac{1}{2\phi}d(r;\mu)\right), \tag{6.1}$$

where d is a deviance function derived from a linear bias function using Equation (5.1). Using the squared distance deviance, unit weights $w = 1$, and $\phi \equiv \sigma^2$ shows that the normal distribution is in the exponential family, and that it corresponds to $V(\mu) = 1$. The gamma, binomial, Poisson and inverse Gaussian[5] distributions are also members of the exponential family. The exponential distribution, being a special case of the gamma, is also in the exponential family. It is important in Equation (6.1) that the function c depends only on r and ϕ; the same constant has to hold for all values of μ. This is a hard condition to satisfy. For example, it can

[4] This definition is slightly different from that in McCullagh and Nelder [16] and other sources on generalized linear models. See Appendix A for a reconciliation with the usual definition. The approach here is derived from Jørgensen [12] and [16] Chapter 9.

[5] For more information on the inverse Gaussian see Johnson, Kotz and Balakrishnan [11] and Panjer and Willmot [20]. It is similar to the lognormal distribution and can be used to model severity distributions.

be shown there is no such c when the deviance is derived from the variance function $V(\mu) = \mu^\zeta$ with $0 < \zeta < 1$.

Equation (6.1) and the definition of linear bias functions in terms of variance functions imply that an exponential family distribution is determined by the variance function.

If a random variable R has an exponential family distribution given by Equation (6.1) then

$$E(R) = \mu \tag{6.2}$$

and

$$\operatorname{Var}(R) = \frac{\phi}{w} V(\mu), \tag{6.3}$$

which helps to explain the choice of μ as the first parameter and also why V is called the variance function. Because of its role in Equation (6.3), ϕ is called the **dispersion parameter**. Equations (6.2) and (6.3) follow immediately from two well known results about the loglikelihood function $l = l(\mu, \phi; r) = \log f(r; \mu, \phi)$. The first is

$$E\left(\frac{\partial l}{\partial \mu}\right) = 0, \tag{6.4}$$

$(E(\partial l/\partial\mu) = E(f'/f) = \int f' = \partial/\partial\mu \int f = \partial/\partial\mu(1) = 0)$. Equation (6.4) implies Equation (6.2). The second is

$$E\left(\frac{\partial^2 l}{\partial \mu^2}\right) + E\left[\left(\frac{\partial l}{\partial \mu}\right)^2\right] = 0 \tag{6.5}$$

which is derived similarly and which implies Equation (6.3).

The next two subsections derive the deviance functions associated with the gamma distribution and the inverse Gaussian distribution. The gamma example starts with the density and derives the variance function. The inverse Gaussian the example goes in the opposite direction and starts with a variance function. In both cases the reader may (correctly) suspect the calculations are easier if one knows what the answer is going to be! Similar calculations can be performed for the Poisson and binomial distributions.

6.1 Gamma distribution in the exponential family

The usual parameterization of the gamma density is

$$f(r;\alpha,\beta) := \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$$

which has mean α/β and variance α/β^2. Since the parameter of interest is the mean, it makes sense to reparameterize to $\mu = \alpha/\beta$ and $\nu = \alpha$. The variance becomes μ^2/ν and the density becomes

$$f(r;\mu,\nu) := \left(\frac{\nu}{\mu}\right)^\nu \frac{1}{\Gamma(\nu)} r^{\nu-1} e^{-\nu r/\mu}.$$

Assuming weight w, Equation (6.3) gives $\phi\mu^2/w = \mu^2/\nu$, so $\nu = w/\phi$. Rearranging the density gives

$$\begin{aligned} f(r;\mu,\nu) &= \frac{\nu^\nu r^{-1}}{\Gamma(\nu)} \exp(\nu \log(r/\mu) - \nu r/\mu) \\ &= \frac{\nu^\nu r^{-1} e^{-\nu}}{\Gamma(\nu)} \exp\left(-\frac{\nu}{2}\, 2\left(\left(\frac{r-\mu}{\mu}\right) - \log\left(\frac{r}{\mu}\right)\right)\right). \end{aligned} \tag{6.6}$$

Since the deviance $d = 2((r-\mu)/\mu - \log(r/\mu))$ corresponds to the variance function $V(\mu^2)$—see Example 5.1(b)—the gamma distribution is in the exponential family.

▮

6.2 Exponential density corresponding to the variance function $V(\mu) = \mu^3$

The deviance function corresponding to $V(\mu) = \mu^3$ is given by

$$\begin{aligned} d(r;\mu) :&= 2\int_\mu^r \frac{r-t}{t^3} dt \\ &= \frac{1}{r} + \frac{r}{\mu^2} - \frac{2}{\mu} \\ &= \frac{(r-\mu)^2}{\mu^2 r}. \end{aligned}$$

The corresponding exponential family distribution when $w = 1$ is

$$f(r; \mu, \phi) := c(r, \phi) \exp\left(-\frac{1}{2\phi} \frac{(r-\mu)^2}{\mu^2 r} \right),$$

which is exactly the inverse Gaussian distribution. The term $c(r, \phi)$ is given by

$$\sqrt{\frac{1}{2\pi\phi r^3}}.$$

The usual parameters for the inverse Gaussian are $1/\phi$ and $1/\mu$. ▮

The variance function corresponding to the Poisson distribution is $V(\mu) = \mu$; for the binomial distribution it is $V(\mu) = \mu(1-\mu)$.

The modeling interpretation of V is clear from its role in linear bias functions. Now that we know how some variance functions and distributions match up we can make some further observations. The normal distribution model assumes constant variance, which is why the second important adjustment on page 23 is not present in normal theory models. The Poisson model assumes the variance is proportional to the mean. The gamma model assumes the variance is proportional to the square of the mean, that is, that the coefficient of variation is constant. The inverse Gaussian assumes that the variance is proportional to the cube of the mean. The form of the variance function is very important in modeling, since the modeler will generally attempt to give smaller weights to observations with larger variances. Allowing the variance to be a function of the fitted mean gives generalized linear models a significant advantage over normal, constant variance, models.

Section 8 and Jørgensen [12] discuss other members of the exponential family. In particular see [12] Chapter 4 and Table 4.1.

7. Generalized Linear Models and their Connection with Minimum Linear Bias

This section will explain how to solve generalized linear models using a maximum likelihood objective function and show the connection between such solutions and solutions of minimum deviance models using linear bias functions. A thorough understanding of generalized linear models requires a more detailed treatment than can be given in this paper. The book by McCullagh and Nelder [16] is an excellent source for those desiring more information.

Section 2 divided general linear models into three components. The components were a random part, a systematic part and a link between the two—see GLM1-3. The random component can be any member of the exponential family, rather than just the normal distribution. The link function can be any monotonic function. Common choices include $\eta = \mu$, $\eta = \log(\mu)$, $\eta = 1/\mu$, $\eta = 1/\mu^2$ and the logit function $\eta = \log(\mu/(1-\mu))$. The link in a generalized linear model is a function of the predicted mean: $\eta = g(\mu)$, as opposed to the inverse link functions h used in Section 5 which are functions of the linear predictor $\mu = h(\eta)$.

7.1 Specification of a generalized linear model

The full specification of a generalized linear model consists of

- input data,
- model and distribution assumptions, and
- an objective function.

The input data comprises n observations $\mathbf{r} = (r_1, \dots, r_n)^t$, n prior weights $\mathbf{w} = (w_1, \dots, w_n)^t$ and p covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ for each observation $i = 1, \dots, n$. The covariates are the rows of the design matrix $\mathbf{X}$.

The model and distribution assumptions mirror the description GLM1-3. Observations r_i are assumed to be sampled from an exponential family distribution with mean μ_i and second parameter ϕ/w_i. The mean is related to the linear predictor using a link function

$$\mu_i = h(\eta_i), \quad \eta_i = g(\mu_i),$$

and the linear predictor is related to the covariates by

$$\eta_i := \sum_j x_{ij}\beta_j = \mathbf{x}_i\boldsymbol{\beta}$$

for parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$. Finally, the parameters are selected using the maximum likelihood objective.

The differences between a generalized and general linear model are the link function and the exponential family error distribution.

7.2 Maximum likelihood equations for a generalized linear model

Let d be the deviance function associated with the exponential distribution used to define the model. From the definition of the exponential family Equation (6.1), the loglikelihood is given by

$$l = l(\boldsymbol{\beta}; \mathbf{r}) := \sum_{i=1}^{n} -\frac{1}{2\phi} d(r_i; \mu_i) + \log(c(r_i, \phi)). \tag{7.1}$$

To help the reader work through some explicit examples, Table 1 gives a summary

TABLE 1
PARAMETERS FOR EXPONENTIAL FAMILY DISTRIBUTIONS

Quantity	Normal	Gamma	Inverse Gaussian
$V(\mu)$	1	μ^2	μ^3
Deviance, $d(r;\mu)$	$(r-\mu)^2$	$2\left(\frac{r-\mu}{\mu} - \log\left(\frac{r}{\mu}\right)\right)$	$\frac{(r-\mu)^2}{\mu^2 r}$
Dispersion, ϕ	ϕ	$\phi = 1/\nu$	ϕ
c	$(2\pi\phi)^{-1/2}$	$\nu^\nu r^{-1} e^{-\nu}/\Gamma(\nu)$	$(2\pi\phi r^3)^{-1/2}$

of the functions introduced so far for the normal, gamma and inverse Gaussian distributions. If the weights $w \neq 1$ then replace ϕ with ϕ/w.

We find the maximum likelihood parameters $\hat{\boldsymbol{\beta}}$ by solving the system of p equations

$$\frac{\partial l}{\partial \beta_j} = 0$$

for $j = 1, \dots, p$. Calculating from Equation (7.1) gives

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= -\sum_i \frac{1}{2\phi} \frac{\partial d(r_i; \mu_i)}{\partial \beta_j} \\
&= -\sum_i \frac{w_i}{2\phi} \frac{\partial}{\partial \mu_i} \left(2 \int_{\mu_i}^{r_i} \frac{r_i - t}{V(t)} dt \right) \frac{\partial \mu_i}{\partial \beta_j} \\
&= \sum_i \frac{w_i}{\phi} \frac{r_i - \mu_i}{V(\mu_i)} \frac{\partial h(\mathbf{x}_i \boldsymbol{\beta})}{\partial \beta_j} \\
&= \sum_i \frac{w_i}{\phi} \frac{r_i - \mu_i}{V(\mu_i)} h'(\mathbf{x}_i \boldsymbol{\beta}) x_{ij}
\end{aligned}$$

since $\mathbf{x}_i \boldsymbol{\beta} = \sum_j x_{ij} \beta_j$. Equating to zero, the ϕ cancels out (just as σ cancels out of normal error linear models) giving the maximum likelihood equations for β_j:

$$\sum_{i=1}^{n} \hat{w}_i (r_i - \mu_i) x_{ij} = 0, \tag{7.2}$$

where the adjusted weight is defined as

$$\hat{w}_i = \frac{w_i h'(\mathbf{x}_i \boldsymbol{\beta})}{V(\mu_i)}. \tag{7.3}$$

Let $\mathbf{W}$ be the $n \times n$ diagonal matrix of adjusted weights $\hat{w}_i$. Then writing Equation (7.2) in matrix notation gives

$$\mathbf{X}^t \mathbf{W} (\mathbf{r} - \boldsymbol{\mu}) = \mathbf{0}. \tag{7.4}$$

As expected from the definition of exponential densities, Equation (7.4) is the same as the minimum deviance equations Equation (5.4). We have shown that the solution to the generalized linear model specified in Section 7.1 is the same as the solution to the minimum bias model with the same covariates, link function, and associated variance function.

Special cases of the correspondence between generalized linear models and minimum linear bias models include:

$$\begin{aligned} \text{Normal} &\leftrightarrow V(\mu) = 1, \\ \text{Binomial} &\leftrightarrow V(\mu) = \mu(1-\mu), \\ \text{Poisson} &\leftrightarrow V(\mu) = \mu, \\ \text{Gamma} &\leftrightarrow V(\mu) = \mu^2, \text{and} \\ \text{Inverse Gaussian} &\leftrightarrow V(\mu) = \mu^3. \end{aligned}$$

The correspondence holds for all link functions. It also holds regardless of whether the minimum linear bias problem can be converted into a set of iterative equations. If the iterative equations exist, they can be used to solve for the parameters. In all cases, the theory of generalized linear models can be used to find the model parameters.

7.3 Canonical Link

If $\hat{w}_i = w_i$ in Equation (7.3) then h is called the canonical link corresponding to the variance function V. Clearly the canonical link satisfies the differential equation $V(h(\eta)) = h'(\eta)$. For example, if $V(\mu) = \mu$ then $h(\eta) = e^{\eta}$ is the canonical link. It is easier to find the maximum likelihood parameters using the canonical link because the weight matrix $\mathbf{W}$ is independent of the fitted values. If the canonical link is used, then adjusted balance is the same as balance in Bailey's definition. Despite its name, there is no need to use the canonical link associated with a particular variance function.

7.4 Explicit Examples

This subsection presents some explicit forms of the correspondence laid out above, including six out of the eight different minimum bias models given by Brown [5].

Assume there are two classification variables and use the minimum bias notation from Section 4. Thus i and j are used to label both the observations and the parameters. Equation (7.4) translates into

$$\mathbf{0} = \mathbf{X}^t\mathbf{W}(\mathbf{r}-\boldsymbol{\mu}) = \begin{pmatrix} \sum_j \hat{w}_{1j}(r_{1j}-\mu_{1j}) \\ \vdots \\ \sum_j \hat{w}_{n_1j}(r_{n_1j}-\mu_{n_1j}) \\ \sum_i \hat{w}_{i1}(r_{i1}-\mu_{i1}) \\ \vdots \\ \sum_i \hat{w}_{in_2}(r_{in_2}-\mu_{in_2}) \end{pmatrix}, \qquad p\times 1,$$

(compare with Equation (4.9) and Equation (4.10)). An equation from the first block gives

$$\sum_{j=1}^{n_2} \hat{w}_{ij}(r_{ij}-\mu_{ij}) = 0, \qquad i = 1,\ldots,n_1, \tag{7.5}$$

while one from the second block gives

$$\sum_{i=1}^{n_1} \hat{w}_{ij}(r_{ij}-\mu_{ij}) = 0, \qquad j = 1,\ldots,n_2. \tag{7.6}$$

The basic symmetry of the minimum bias method is already clear in the above equations.

(a) *Identity Link Function*

For the identity link function, $\eta_{ij} = \mu_{ij}$ and $d\eta/d\mu = 1$, so

$$\hat{w}_{ij} = \frac{w_{ij}}{V(\mu_{ij})}.$$

Moreover, using an additive model $\eta_{ij} = x_i + y_j$, and so $\mu_{ij} = x_i + y_j$. Substituting into the maximum likelihood equation Equation (7.5) gives

$$\begin{aligned} 0 &= \sum_{j=1}^{n_2} \hat{w}_{ij}(r_{ij}-\mu_{ij}) \\ &= \sum_{j=1}^{n_2} \frac{w_{ij}}{V(\mu_{ij})}(r_{ij}-(x_i+y_j)) \\ &= \sum_{j=1}^{n_2} \frac{w_{ij}}{V(\mu_{ij})}(r_{ij}-y_j) - x_i\sum_{j=1}^{n_2}\frac{w_{ij}}{V(\mu_{ij})}, \end{aligned}$$

for $i = 1, \dots, n_1$. Hence

$$x_i = \sum_{j=1}^{n_2} w_{ij}(r_{ij} - y_j)/V(\mu_{ij}) \Bigg/ \sum_{j=1}^{n_2} w_{ij}/V(\mu_{ij}), \tag{7.7}$$

and similarly

$$y_j = \sum_{i=1}^{n_1} w_{ij}(r_{ij} - x_i)/V(\mu_{ij}) \Bigg/ \sum_{i=1}^{n_1} w_{ij}/V(\mu_{ij}), \tag{7.8}$$

for $j = 1, \dots, n_2$.

For the normal distribution $V(\mu) = 1$. Substituting into Equation (7.7) gives

$$x_i = \sum_j w_{ij}(r_{ij} - y_j) \Bigg/ \sum_j w_{ij} \tag{7.9}$$

which is Bailey's additive model discussed in Section 4.

For the Poisson distribution $V(\mu) = \mu$, and so Equation (7.7) gives

$$x_i = \sum_j w_{ij}(r_{ij} - y_j)/\mu_{ij} \Bigg/ \sum_j w_{ij}/\mu_{ij}. \tag{7.10}$$

which is a new minimum bias method. For the gamma distribution $V(\mu) = \mu^2$, and so Equation (7.7) gives

$$x_i = \sum_j w_{ij}(r_{ij} - y_j)/\mu_{ij}^2 \Bigg/ \sum_j w_{ij}/\mu_{ij}^2. \tag{7.11}$$

which is another new method. Finally, for the inverse Gaussian distribution $V(\mu) = \mu^3$, and so Equation (7.7) gives

$$x_i = \sum_j w_{ij}(r_{ij} - y_j)/\mu_{ij}^3 \Bigg/ \sum_j w_{ij}/\mu_{ij}^3. \tag{7.12}$$

which is a third new method. The binomial distribution, with $V(\mu) = \mu(1-\mu)$ also gives a new method.

The models Equation (7.10) to Equation (7.12) give progressively less and less weight to observations with higher predicted means μ_{ij}.

(b) *Log Link Function*

For the log link function, $\eta = \log(\mu)$, so $d\eta/d\mu = 1/\mu$, which gives

$$\hat{w}_{ij} = \frac{w_{ij}\mu_{ij}}{V(\mu_{ij})}.$$

In this case $\mu_{ij} = \exp(\eta_{ij}) = \exp(x_i + y_j) =: a_i b_j$. As expected the log link converts an additive model into a multiplicative one. Substituting into Equation (7.5) gives,

$$\begin{aligned} 0 &= \sum_{j=1}^{n_2} \hat{w}_{ij}(r_{ij} - \mu_{ij}) \\ &= \sum_{j=1}^{n_2} \frac{w_{ij} a_i b_j}{V(\mu_{ij})}(r_{ij} - a_i b_j) \\ &= \sum_{j=1}^{n_2} \frac{w_{ij} r_{ij} b_j}{V(\mu_{ij})} - a_i \sum_{j=1}^{n_2} \frac{w_{ij} b_j^2}{V(\mu_{ij})}, \end{aligned}$$

for $i = 1, \ldots, n_1$. Hence

$$a_i = \sum_{j=1}^{n_2} w_{ij} r_{ij} b_j / V(\mu_{ij}) \Bigg/ \sum_{j=1}^{n_2} w_{ij} b_j^2 / V(\mu_{ij}), \tag{7.13}$$

and similarly for b_j.

Now substituting V's for the normal, Poisson, gamma and inverse Gaussian distributions gives the following four minimum bias methods:

$$a_i = \sum_j w_{ij} r_{ij} b_j \Bigg/ \sum_j w_{ij} b_j^2, \tag{7.14}$$

$$a_i = \sum_j w_{ij} r_{ij} \Bigg/ \sum_j w_{ij} b_j, \tag{7.15}$$

$$a_i = \sum_j w_{ij} r_{ij} / b_j \Bigg/ \sum_j w_{ij}, \tag{7.16}$$

$$a_i = \sum_j w_{ij} r_{ij} / b_j^2 \Bigg/ \sum_j w_{ij} / b_j. \tag{7.17}$$

Equation (7.15) is Bailey's multiplicative model and Equation (7.16) is Brown's exponential model—which Venter comments also works for the gamma. Equation (7.14) and Equation (7.17) appear to be new. Again, going from Equation (7.14) to Equation (7.17) the models give less and less weight to observations with high predicted means.

(c) *Ad Hoc Methods*

Other functions besides the identity, log, and logit can be used as links. Two common choices are $\eta = 1/\mu$ and $\eta = 1/\mu^2$. For the inverse function, $d\eta/d\mu = -1/\mu^2$ so $\hat{w} = w\mu^2/V(\mu)$, and $\mu_{ij} = 1/\eta_{ij} = 1/(x_i + y_j)$. Substituting into Equation (7.5) it is easy to see that only the inverse Gaussian produces a tractable minimum bias method. For the inverse Gaussian $V(\mu) = \mu^3$, so Equation (7.5) gives

$$\begin{aligned}
0 &= \sum_{j=1}^{n_2} \hat{w}_{ij}(r_{ij} - \mu_{ij}) \\
&= \sum_j \frac{w_{ij}}{\mu_{ij}}(r_{ij} - \mu_{ij}) \\
&= \sum_j \frac{w_{ij} r_{ij}}{\mu_{ij}} - w_{ij} \\
&= \sum_j w_{ij} r_{ij}(x_i + y_j) - w_{ij} \\
&= x_i \sum_j w_{ij} r_{ij} + \sum_j w_{ij} r_{ij} y_j - \sum_j w_{ij}
\end{aligned}$$

and hence we get another new iterative method:

$$x_i = \sum_j w_{ij}(1 - r_{ij} y_j) \Bigg/ \sum_j w_{ij} r_{ij}. \tag{7.18}$$

In this method one set of parameters will be negative and the other positive.

7.5 Other Variance Assumptions

Brown proposes two models where the variance of r_{ij} is proportional to $1/w_{ij}^2$ rather than $1/w_{ij}$. Although, as Venter points out, the latter is a more natural choice, the former assumption can be handled within our framework. Simply use weights w_{ij}^2 rather than w_{ij}. For example, Equation (7.9) becomes

$$x_i = \sum_j w_{ij}^2 (r_{ij} - y_j) \Bigg/ \sum_j w_{ij}^2 \tag{7.19}$$

which is Brown's model 7.

7.6 Correspondence with Brown's models

For the reader's convenience this subsection identifies our models with the nine models in Brown's paper [Br].

B1: Poisson, multiplicative, Equation (7.15).

B2: Normal, additive, Equation (7.9).

B3: Bailey-Simon, multiplicative—see [3], Equation (7) for derivation. This method comes from minimizing a χ^2-statistic, rather than maximizing a likelihood function. Since generalized linear models rely on maximum likelihood, we would not expect to be able to reproduce it. Unlike B4, it does not use the Newton method.

B4: Bailey-Simon, additive—see [3] page 12 for derivation. This method (which certainly puzzled the author as a Part 9 exam candidate!) also a minimizes a χ^2-statistic. Its derivation uses the Newton method.

B5: Gamma, multiplicative, Equation (7.16); note the Exponential is a special case of the gamma.

B6: Normal, multiplicative with variance proportional to $1/w^2$, Equation (7.14) upon replacing w with w^2.

B7: Normal, additive, with variance proportional to $1/w^2$, Equation (7.19).

B8: The same as B1.

B9: Normal, multiplicative, Equation (7.14). Brown derives B9 using least squares and Venter uses maximum likelihood. The two approaches agree because the likelihood of a normally distributed observation is proportional to its squared distance from the mean.

7.7 Numerical Example, continued

We now present the results of fitting ten generalized linear models to the data presented in Section 4.6. The models are described in Table 2.

Table 2
Description of Models

Model Number	Error Distribution	Link Function	Variance Function
1	Normal	Identity	$V(\mu) = 1$
2	Normal	Log	$V(\mu) = 1$
3	Normal	Inverse	$V(\mu) = 1$
4	Gamma	Identity	$V(\mu) = \mu^2$
5	Gamma	Log	$V(\mu) = \mu^2$
6	Gamma	Inverse	$V(\mu) = \mu^2$
7	Inverse Gaussian	Identity	$V(\mu) = \mu^3$
8	Inverse Gaussian	Log	$V(\mu) = \mu^3$
9	Inverse Gaussian	Inverse	$V(\mu) = \mu^3$
10	Inverse Gaussian	Inverse Square	$V(\mu) = \mu^3$

So far we have not been concerned with the value of the parameter ϕ. It is well known that in general linear models parameter estimates and predicted values are independent of the variance of the error term (usually labeled σ^2 rather than ϕ). Since ϕ does not appear in Equation (7.4) the same is true of generalized linear models. However, just as for general linear models, it is necessary to estimate ϕ in order to determine statistics, such as standard errors of predicted values. In general linear models

$$\hat{\sigma}^2 = \sum_i w_i(r_i - \mu_i)^2/(n-p)$$

is used as an estimator of σ^2, where n is the number of observations and p is the number of parameters. In generalized linear models ϕ is estimated using the moment estimator

$$\hat{\phi} = \frac{1}{n-p} \sum_i w_i \frac{(r_i - \mu_i)^2}{V(\mu_i)}. \tag{7.20}$$

It can also be estimated using

$$\hat{\phi} = \frac{D}{n-p} = \frac{1}{n-p} \sum_i d(r_i, \mu_i), \tag{7.21}$$

where D is the total deviance, see McCullagh and Nelder [16]. (Note that the weights are included in the deviance d in Equation (7.21).) Another way to estimate ϕ is to use the maximum likelihood estimate. Equation (7.1) ensures that the maximum likelihood parameters are unchanged whether or not ϕ is estimated. SAS's proc genmod uses maximum likelihood by default, see [23], and the statistics reported below are based on it unless otherwise noted.

Exhibit 7 gives the parameters corresponding to the ten models in Table 2. Each panel of Exhibit 7 shows the parameter estimates, the standard error of the estimate, the χ^2-statistic to test if the parameter is significantly different from zero and the corresponding p-value from the χ^2-distribution. (See Section 8.1 for more discussion of the χ^2-statistic.) When the link function is not the identity, Exhibit 7 also shows the parameter estimates transformed by the inverse link. For example, in the first row of Exhibit 7-2, $265.22 = e^{5.5806}$. The final row gives the scale function, which is equal to $\sqrt{\phi}$ for the normal and inverse Gaussian distributions and $1/\phi$ for the gamma distribution. Again, maximum likelihood is used to estimate ϕ.

Examining Exhibit 7 shows that all parameters except drive to work less than 10 miles are significantly different from zero for all models. All models indicate there is not a statistically significant difference between drive to work less than 10 miles and pleasure use. The other two use classifications are significantly different from one another. The estimates and standard errors within the age classifications show there is not a statistically significant difference between all levels. For example the 35-39 and 40-49 classes are not significantly different for most models, although exact results depend on the choice of ϕ. In the gamma model with identity link using maximum likelihood gives the estimate $\hat{\phi} = 0.9741$ and the contrast between these two classes has a χ^2-statistic of 4.07, which is significant at the 5% level ($p = 4.4\%$). However, using Equation (7.20) results in an estimate $\hat{\phi} = 1.4879$ with a χ^2-statistic of 2.839 which is not significant at the 5% level ($p = 9.2\%$). In the first case the standard error of the 35-39 class is 8.13 (Exhibit 7-4); in the second it is 10.04.

Exhibit 8 compares the fitted values from three models: the standard linear model (column 5), a general linear model applied to log(severity) (column 6), and a generalized linear model with normal errors and log link (column 7). As pointed out in Example 2.4(c), the three are distinct and give different answers.

Exhibit 9 summarizes the predicted severities by class, by model. The choice of link function and error distribution has a considerable impact on the predicted means in some cells. Using a gamma or inverse Gaussian error term generally results in a greater range of estimates, as does the log or reciprocal link function. Since this is only illustrative data we will not comment on the specific results. See Renshaw, [22], for a more detailed analysis of similar data, together with other suggestions for modeling and assessing model fit.

Exhibit 10 gives the average bias

$$\sum_i w_{ij}(r_{ij} - \mu_{ij}) \Big/ \sum_i w_{ij} \tag{7.22}$$

for each j and

$$\sum_j w_{ij}(r_{ij} - \mu_{ij}) \Big/ \sum_j w_{ij}$$

for each i, for each model. For the normal/identity model, the average bias is zero, since this model is Bailey's additive model. The gamma/inverse model and inverse Gaussian/inverse square models are also balanced because the respective link functions are the canonical links (Section 7.3) and so the adjustment to the weights in Equation (7.3) equals 1, reducing Equation (7.4) to Equation (7.22). In the other cases, the parameters are zero bias according to the relevant adjusted bias function, but not according to that given by Equation (7.22). This provides an interesting example of Venter's V1, alternatives to bias functions.

Exhibit 11 gives the average absolute bias suggested by Bailey [2]:

$$\sum_i w_{ij}|r_{ij} - \mu_{ij}| \Bigg/ \sum_i w_{ij} \tag{7.23}$$

for each i, and similarly for j. The gamma/identity model has the lowest average absolute bias. Finally, the value of the likelihood is available as a fit statistic, since these models were fit using maximum likelihood over all parameters (including ϕ). The results are shown in Table 3. Other statistics that can be used to select between models are discussed in Section 8.

Table 3
Model Loglikelihoods

Model	Distribution	Link	Loglikelihood
1	Normal	Identity	-144.303
2	Normal	Log	-144.435
3	Normal	Inverse	-145.792
4	Gamma	Identity	-140.753
5	Gamma	Log	-141.055
6	Gamma	Inverse	-143.267
7	Inverse Gaussian	Identity	-141.078
8	Inverse Gaussian	Log	-141.347
9	Inverse Gaussian	Inverse	-143.343
10	Inverse Gaussian	Sqr Inverse	-147.224

These examples hint at the power of the statistical viewpoint. Using a minimum bias approach not within the statistical framework it would be impossible to discuss the standard error of predicted values and parameters, or to ask whether two parameters are statistically significantly different. Having the tools to answer such questions can provide useful information to help in designing and parameterizing classification plans. The statistical model also gives information on model fit, discussed in the next section, which helps select covariates, as well as link and variance functions within parameterized families. Again, these tools are not available with the minimum bias approach. Fundamentally it is the connection between variance functions and exponential family distributions which makes the statistical viewpoint possible.

8. Model Fit Statistics

Generalized linear model and minimum bias methods allow the actuary to consider a large number of models: different choices of covariates, different link functions and different variance functions. It is obviously important to be able to determine if one model fits the data better than the others. The specification of a generalized linear model in Section 7.1 shows there are at least four distinct model fit questions:

(1) Comparing different sets of covariates for a given link function and variance function (error distribution).

(2) Comparing different link functions and covariates for a given variance function.

(3) Comparing different variance functions for a given set of covariates and link function.

(4) Simultaneously comparing different link and variance functions and covariates.

In this section we will discuss some of the available statistical tests of model fit. These methods extend the earlier work of Bailey and Simon.

8.1 Comparing sets of covariates

The simplest test of model fit looks for information about the best set of covariates assuming given link and variance functions. In the numerical example, is anything really gained from adding a vehicle use classification? Analysis of variance is used in normal-error model theory to assess the significance of effects and answer such questions. For generalized linear models we look at an analysis of deviance table, obtained from a nested sequence of models. Unfortunately, unlike the normal-error theory where the χ^2- and F-distributions give exact results, only approximations and asymptotic results are available for generalized linear models. McCullagh and Nelder recommend analysis of deviance as a screening device for models and regard this as an area where more work is required.

Consider the gamma distribution model with identity link. With two explanatory variables available we can consider a nested sequence of four models: intercept only, age only, age and vehicle type with no interaction, and age and vehicle type with interaction. The last model is complete: it has as many parameters as there are observations and so fits perfectly. Table 4 shows the resulting analysis of deviance. For each model, it shows the deviance, the reduction in deviance from adding covariates, the number of incremental degrees of freedom, and the mean incremental deviance per degree of freedom. The degrees of freedom are computed as the incremental number of parameters from one model to the next. The model with an intercept only has 1 parameter. Including age variables adds 7 more parameters, and so on. The complete model has one parameter for each of the 32 observations.

Table 4
Analysis of Deviance

Model	Deviance	Δ Deviance	Degrees of Freedom	Mean Deviance
Intercept	347.0331			
Age	264.8553	82.1778	7	11.74
Age+Vehicle	31.2453	233.6100	3	77.87
Complete	0	31.2453	21	1.49

The mean deviance has an approximate χ^2-distribution. Adding the age variable and then the vehicle type variable both significantly improve the model fit. When more explanatory variables are available, an analysis of deviance is helpful in deciding which to use in a model, and in particular, in assessing which interaction effects are significant and should be included.

8.2 Comparing Link Functions

The models discussed in Section 7 used the identity, inverse and log links, all of which belong to the power-link family[6]

$$\eta = \begin{cases} \mu^\lambda & \text{for } \lambda \neq 0, \\ \log(\mu) & \text{for } \lambda = 0. \end{cases}$$

According to Nelder and Lee [17] Section 2.3 we can use the deviance to compare different link functions as well as different covariates. Table 5 shows the deviance for various values of λ, again using the gamma distribution.

[6] Considering $(\mu^\lambda - 1)/\lambda$ instead of μ^λ makes the family appear more natural because $(\mu^\lambda - 1)/\lambda \to \log(\mu)$ as $\lambda \to 0$. This form of the power-link function is called the Box-Cox transformation. It is mentioned in Venter's review [25].

Table 5
Deviance vs. Link Power λ

λ	Deviance
-1.800	43.828
-1.300	38.966
-0.800	35.190
-0.300	32.724
0.200	31.464
0.700	31.129
1.200	31.418
1.450	31.717

The deviance is relatively flat across the range $0.325 \leq \lambda \leq 1.075$ which includes the identity link. The deviance for the inverse link $\lambda = -1$ is substantially greater than for λ in this range.

8.3 More on Variance Functions

Before discussing tests over sets of variance functions, we must mention a few facts about them. Jørgensen, [12] and [13], discusses the exponential families corresponding to variance functions beyond the simple examples we have considered so far. His results include the following which are of interest to actuaries.

(1) $V(\mu) = \mu^{\zeta}$ for $1 < \zeta < 2$ corresponds to the Tweedie distribution, which is a compound distribution with Poisson frequency component and gamma severity component. It is a mixed distribution with a non-zero probability of taking the value zero, which makes it useful in modeling aggregate distributions. Jørgensen and deSousa [14] fit the Tweedie model to Brazilian auto data.

(2) $V(\mu) = \mu^{\zeta}$ for $\zeta < 0$ corresponds to an extreme stable distribution. Non-normal stable distributions are thick tailed distributions which may be useful in fitting loss data.

(3) $V(\mu) = \mu^{\zeta}$ for $2 < \zeta < \infty$, $\zeta \neq 3$ corresponds to a positive stable distribution.

(4) $V(\mu) = \mu^\zeta$ for $0 < \zeta < 1$ does not give an exponential family distribution.

(5) $V(\mu) = \mu(1 + \mu/\nu)$ corresponds to the negative binomial distribution.

(6) $V(\mu) = \mu(1 + \tau\mu^2)$ corresponds to the Poisson-inverse Gamma distribution. Renshaw [22] gives the deviance functions for both of the last two distributions.

The power variance function family leads naturally to the question of determining the best estimate for ζ, to which we now turn.

8.4 Comparing variance functions

The deviance cannot be used to select an optimal ζ because the deviance of an individual observation $(r - \mu)/\mu^\zeta \to 0$ as $\zeta \to \infty$ for $\mu > 1$. This means a deviance based objective would generally claim ζ should be very large and that the model fit was excellent. Clearly it is necessary to include some measure of the likelihood of ζ in the objective function to counter-balance the effect of the variance function on the deviance. In general, according to Nelder and Lee [17], deviance cannot be used to compare different variance functions on the same data.

One way to include the likelihood of ζ would be to use the full likelihood function for the corresponding density. This method was used in the examples shown in Section 7 for the normal, gamma and inverse Gaussian distributions—where the densities are known. Unfortunately, for most exponential family distributions, including the Tweedie and stable distributions, there is no simple closed form expression for the density or distribution function. It is therefore not possible to write down the likelihood function.

The way out of this impasse is to use a tractable approximation to the density function, such as the saddlepoint approximation. Details of the derivation are beyond the scope of this paper, but the result is to replace the deviance function

$$d(r_i; \mu) = 2w_i \int_\mu^{r_i} \frac{r_i - t}{V(t)} dt \tag{8.1}$$

with an extended deviance function (extended quasi-likelihood in the literature)

$$d(r_i;\mu) = 2\frac{w_i}{\phi}\int_{\mu}^{r_i}\frac{r_i - t}{V(t)}dt + \log(\phi V(r_i)). \tag{8.2}$$

The added term grows with V thus providing the desired counter-balance to the first term, which shrinks. Note that V is evaluated at the responses r_i rather than the fitted means μ_i. Including the scale parameter ϕ allows Equation (8.2) to be used both for inference over parameterized families of variance functions and for different values of ϕ. Jørgensen [12] Example 3.1 on page 104 explains the saddlepoint approximation for a gamma distribution, which is just Stirling's formula for the gamma function. See McCullagh and Nelder [16] Chapter 9, Nelder and Lee [17], and Renshaw [22] for more about extended deviance functions. [17] also defines and compares other extensions of deviance.

Table 6 shows the extended deviances for various values of ζ modeled with the identity link function.

Table 6
Extended Deviance vs. Variance Functions $V(\mu) = \mu^{\zeta}$

ζ	Deviance
1.20	372.740
1.45	372.020
1.70	371.422
1.95	370.946
2.20	370.597
2.45	370.374
2.70	370.282
2.95	370.321
3.20	370.494
3.45	370.800

The table shows a reasonable range $1.95 \le \zeta \le 3.45$ which includes both the gamma distribution $\zeta = 2$ and inverse Gaussian distribution $\zeta = 3$. Combining the results of Tables 5 and 6 shows the gamma or inverse Gaussian distribution with identity link is still a reasonable choice even if we are free to select from the power link family and power variance function family. These conclusions are in line with the full likelihood results in Table 3 and the average absolute deviations in Exhibit 11 where the gamma/identity and inverse Gaussian/identity models show the best results.

8.5 Deviance Profiles and Comparing Link and Variance Functions

The last step we will consider combines the power link and variance functions and looks for the overall minimum extended deviance estimators. Figure 2 shows a contour plot of extended deviance over ζ and λ. The results are as expected from the one dimensional calculations. The dotted rectangle shows a range of λ from log link to the identity and ζ from gamma distribution to inverse Gaussian.

9. Computations

Section 9 is in two parts. The first discusses the iterative method for solving minimum bias models. For the additive model with identity link it gives a sufficient condition for the iterative method to converge (no matter the initial conditions), explains precisely how it converges in terms of the eigenvectors of a particular matrix, and gives a telescoping argument that jumps to the solution of the iterative process once the first iteration has been computed.

The second section discusses how to find the maximum likelihood parameters in a generalized linear model. Even though commercial software exists to solve generalized linear models it is instructive to perform the calculations by hand, and we explain how to do this. Examples of SAS code to solve generalized linear models using both the SAS/Stat procedure genmod and a "bare hands" approach using matrix algebra are given in Appendix B.

At several points this section discusses a notion of computational efficiency. Two algorithms are of similar computational efficiency if they will run in about the same time for *all sizes of input.* (Technically, if n is the problem size, and $f(n)$ and

$g(n)$ are the number of elementary operations required to solve the problem using two methods, then they are of the same computational efficiency if $f = O(g)$ and $g = O(f)$, Borwein and Borwein [4] Chapter 6. Recall $f = O(g)$ means there is a constant K so that $f(n) \leq Kg(n)$ for all n.)

9.1 Iterative Methods

Bailey's original paper [2] introduces the additive and multiplicative models and suggests the iterative method for finding parameters:

> Using a predetermined set of estimators for each territory, construction, and protection, we can solve the [minimum bias] formula for the estimator for each occupancy. We can then use these calculated estimators for each occupancy to calculate a revised set of estimators for each territory using a similar formula, and continue this process until the estimators stabilize.

Since Bailey's paper, it has become common for actuaries to use this iterative method. For example, ISO [10] explicitly describes the three-way minimum bias model for the personal auto classification plan as iterative.

Just because the minimum bias model *suggests* using an iterative method to solve for the parameters, it does not follow that such a method is the best method to use. Section 4 showed that the usual additive model is simply a general linear model; and so it is far more computationally efficient to solve the normal equations (no iterations, few matrix multiplications and one inverse) than it is to use the iterative method. Any actuaries still using iterative methods should investigate whether the generalized linear model approach outlined in this paper would speed up their calculations—as well as providing them with more useful diagnostic information.

This section considers the iterative method for the additive model with identity link which is used by ISO for the personal auto class plan. The iterative method is considered in detail despite its shortcomings, because many actuaries may have tried the method (perhaps as part 9 students) and may have wondered what initial conditions are required for convergence and may also have noted the strange way the models converge. We explain the convergence in detail and also show it is not necessary to perform many iterations, even if the iterative paradigm is followed. However, the final message of this section is *do not use the iterative method for Bailey's additive model.* Solve the normal equations instead!

We will use the notation of Section 4 and consider two classification variables—extensions are immediate. Assume that base classes have been selected so that the sum-of-squares and products matrix $\mathbf{X}^t\mathbf{W}\mathbf{X}$ is invertible, $\mathbf{a}$ has dimension $n_1 \times 1$ and $\mathbf{b}$ has dimension $n_2 \times 1$. Finally assume $n_2 \leq n_1$; if this is not the case then swap $\mathbf{a}$ and $\mathbf{b}$. For this example the adjusted weights $\hat{w} = w$, see Equation (7.3).

From Equation (4.14) the minimum bias equations can be written as

$$(\mathbf{A}\ \mathbf{B})^t\mathbf{W}\left(\mathbf{r} - (\mathbf{A}\ \mathbf{B})\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right) = \begin{pmatrix}\mathbf{A}^t\mathbf{W}(\mathbf{r} - \mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b})\\ \mathbf{A}^t\mathbf{W}(\mathbf{r} - \mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b})\end{pmatrix} = \mathbf{0}_{(n_1+n_2)\times 1}. \tag{9.1}$$

Re-arranging Equation (9.1) gives

$$\mathbf{a} = (\mathbf{A}^t\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{W}(\mathbf{r} - \mathbf{B}\mathbf{b}) \tag{9.2}$$

$$\mathbf{b} = (\mathbf{B}^t\mathbf{W}\mathbf{B})^{-1}\mathbf{B}^t\mathbf{W}(\mathbf{r} - \mathbf{A}\mathbf{a}) \tag{9.3}$$

The iterative solution starts with some initial choice $\mathbf{b}^{(0)}$ and uses Equation (9.2) to solve for $\mathbf{a}^{(1)}$. Substituting $\mathbf{a}^{(1)}$ into the Equation (9.3) gives an expression for $\mathbf{b}^{(1)}$. Iterating gives $\mathbf{a}^{(2)}$, $\mathbf{b}^{(2)}$ and so forth. The procedure stops when the difference between successive iterations is sufficiently small. Set $\mathbf{v}^{(m)} = \mathbf{b}^{(m)} - \mathbf{b}^{(m-1)}$ equal to the difference in the m and $(m-1)$th iterations for $\mathbf{b}$. Note there is an asymmetry between the $\mathbf{a}$-iterations and the $\mathbf{b}$-iterations based on where we choose to start.

Set

$$\mathbf{M} = (\mathbf{B}^t\mathbf{W}\mathbf{B})^{-1}\mathbf{B}^t\mathbf{W}\mathbf{A}(\mathbf{A}^t\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{W}\mathbf{B}, \tag{9.4}$$

an $n_2 \times n_2$ matrix. A straightforward telescoping argument shows that

$$\mathbf{b} = (\mathbf{I} - \mathbf{M})^{-1}\mathbf{v}^{(1)} + \mathbf{b}^{(0)} \tag{9.5}$$

provided $\mathbf{M}^m \to 0$. We can guarantee that $\mathbf{M}^m \to 0$ as $m \to \infty$ if all the eigenvalues of $\mathbf{M}$ have absolute value less than 1. This gives a necessary condition for the iterative method to converge, and, moreover, Equation (9.5) shows how to "jump" straight to the final solution after computing only one iteration, $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(1)}$. This method of solving the minimum bias problem will run much faster than the iterative method, but will still be slower than solving the normal equations (computing $\mathbf{M}$ alone involves eleven matrix multiplications and two inverses).

It is also possible to show that $\mathbf{v}^{(m)}$ tends to a scalar multiple of the eigenvector associated with the largest eigenvalue of $\mathbf{M}$ and the iterative method converges along the direction of that eigenvector. Moreover, the distance between subsequent iterations of $\mathbf{b}^{(m)}$ decreases by approximately the absolute value of the largest eigenvector for m large enough.

Numerical Example, continued

Exhibit 5 illustrates the above theory. Column 2 gives the length of $\mathbf{v}$, column 14 gives the ratio of successive iterations of $\mathbf{v}$, and columns 15-17 give the three components of $\mathbf{v}^{(m)}$. The ratio of lengths of $\mathbf{v}$ should converge to the largest eigenvalue of the matrix $\mathbf{M}$ defined by Equation (9.4). For the data underlying Exhibit 5

$$\mathbf{M} = \begin{pmatrix} 0.457572 & 0.300972 & 0.118435 \\ 0.431800 & 0.306347 & 0.122798 \\ 0.428349 & 0.309565 & 0.126608 \end{pmatrix} \tag{9.6}$$

which has eigenvalues 0.000541, 0.010541 and 0.859445. This explains the 0.85944's that appear in Exhibit 5; their appearance is quick since the largest eigenvalue is so much greater than the other two. The overall convergence of the model is quite slow, since 0.859 is close to 1.0.

Exhibits 12 and 13 show how the iterative method converges for two other models: gamma/identity in 12 and gamma/inverse in 13. Convergence is particularly slow for the latter; after 25 iterations the parameters are nowhere near their final values. The methods of this section do not apply to non-canonical link functions because the weight matrix **W** must be re-evaluated between each iteration and so the telescoping argument will not hold.

9.2 Solving Generalized Linear Models

One conclusion of this paper is that many useful minimum linear bias models correspond in a natural way with generalized linear models. However, not all minimum linear bias models have a tractable iterative solution. It is therefore useful to know how to solve generalized linear models. Since there are pre-programmed routines for generalized linear models[8] we give only a brief overview here. This section follows McCullagh and Nelder [16]. The notation is the same as the first part of Section 7.

From Equation (7.4) the maximum likelihood equations for the generalized linear model are given by

$$\mathbf{X}^t\mathbf{W}(\mathbf{r}-\boldsymbol{\mu})=\mathbf{0}$$

where $\mathbf{W}$ is the diagonal matrix with entries $\hat{w}_i = w_i h'(\mathbf{x}_i\boldsymbol{\beta})/V(\mu_i)$.

To find the maximum likelihood it is necessary to solve $\partial l/\partial\beta_j = 0$, for $j = 1,\ldots,p$. This can be done using a method related to the Newton-Raphson method. In one dimension the Newton-Raphson method solves an equation $f(x) = 0$ by iterating $x_{n+1} = x_n - f(x_n)/f'(x_n)$. We are trying to solve the vector equation $\mathbf{u}(\boldsymbol{\beta}) = \mathbf{0}$ where

$$\mathbf{u}(\boldsymbol{\beta}) = \mathbf{u} := \partial l/\partial\boldsymbol{\beta} = (\partial l/\partial\beta_1,\ldots,\partial l/\partial\beta_p)^t.$$

[8] As well as GLIM, mentioned by Brown, SAS now includes a procedure, `proc genmod` to solve generalized linear models in its SAS/Stat package. Genmod has the same syntax as `proc glm`.

Looking at Newton-Raphson suggests trying $\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n - (\partial \mathbf{u}/\partial \boldsymbol{\beta})^{-1}\mathbf{u}$. The term $\partial \mathbf{u}/\partial \boldsymbol{\beta}$ is called the Hessian. The negative Hessian is called the observed information matrix (see Hogg and Klugman, [9] page 121). It is generally a random quantity. Fisher's scoring method simplifies the Newton-Raphson method by using the expected value of the Hessian rather than the Hessian itself. It often results in more staightforward calculations.

To apply Fisher's scoring method, let

$$\mathbf{H} = -E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) = -E\left(\frac{\partial \mathbf{u}}{\partial \beta}\right)$$

be the negative expected value of the Hessian matrix. Given an estimate $\boldsymbol{\beta}_n$ of $\boldsymbol{\beta}$ we find the next adjustment $\mathbf{a}$ by solving $\mathbf{Ha} = \mathbf{u}$. (The adjustment term in the Newton-Raphson method, $a := f(x_n)/f'(x_n)$, satisfies $f'(x_n)a = f(x_n)$. Here, $f \leftrightarrow \mathbf{u}$ and $f' \leftrightarrow H$.) From Equation (7.4), $\mathbf{u} = \mathbf{X}^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu})$, and so

$$\begin{aligned}
\mathbf{H} &= -E\left(\frac{\partial \mathbf{u}}{\partial \boldsymbol{\beta}}\right) \\
&= -E\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{X}^t\mathbf{W}(\mathbf{r} - \boldsymbol{\mu})\right) \\
&= -E\left(\frac{\partial(\mathbf{X}^t\mathbf{W})}{\partial \boldsymbol{\beta}}(\mathbf{r} - \boldsymbol{\mu}) + (\mathbf{X}^t\mathbf{W})\frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{r} - \boldsymbol{\mu})\right) && (9.7) \\
&= E\left(\mathbf{X}^t\mathbf{W}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right) && (9.8) \\
&= E\left(\mathbf{X}^t\mathbf{W}\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right) && (9.9) \\
&= \mathbf{X}^t\tilde{\mathbf{W}}\mathbf{X} && (9.10)
\end{aligned}$$

which is the weighted sums of squares and products matrix for the model with weights

$$\tilde{\mathbf{W}} = \mathrm{diag}(\frac{w_i(h'(\mathbf{x}_i\boldsymbol{\beta})^2}{V(\mu_i)}).$$

Equation (9.7) uses the chain rule; Equation (9.8) uses the fact that $\mathrm{E}(\mathbf{r}) = \boldsymbol{\mu}$ and $\partial \mathbf{r}/\partial \boldsymbol{\beta} = \mathbf{0}$ ($\mathbf{r}$ is a vector of numbers); Equation (9.9) uses the chain rule and the fact that $\boldsymbol{\eta} = X\boldsymbol{\beta}$ and finally Equation (9.10) uses the fact that $\mathbf{X}$ is constant. Since $\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + \mathbf{a}$, $\mathbf{H}\boldsymbol{\beta}_{n+1} = \mathbf{H}\boldsymbol{\beta}_n + \mathbf{Ha} = \mathbf{H}\boldsymbol{\beta}_n + \mathbf{u}$, and hence

$$\begin{aligned} \mathbf{X}^t\tilde{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}_{n+1} &= \mathbf{X}^t\tilde{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}_n + \mathbf{u} \\ &= \mathbf{X}^t\tilde{\mathbf{W}}\boldsymbol{\eta}_n + \mathbf{X}^t\tilde{\mathbf{W}}\frac{d\boldsymbol{\eta}_n}{d\boldsymbol{\mu}}(\mathbf{r}-\boldsymbol{\mu}) \\ &= \mathbf{X}^t\tilde{\mathbf{W}}(\boldsymbol{\eta}_n + \frac{d\boldsymbol{\eta}_n}{d\boldsymbol{\mu}}(\mathbf{r}-\boldsymbol{\mu})). \end{aligned} \tag{9.11}$$

Equation (9.11) is the normal equation for a linear weighted least-squares model of the data $\boldsymbol{\eta}_n + (d\boldsymbol{\eta}_n/d\boldsymbol{\mu})(\mathbf{r}-\boldsymbol{\mu})$ using design matrix $\mathbf{X}$ and weights $\tilde{\mathbf{W}}$.

Note that

$$g(\mu) + (r-\mu)g'(\mu) = \eta + (r-\mu)\frac{d\eta}{d\mu}$$

is a linear approximation to $g(r) = h^{-1}(r)$ and that

$$\mathrm{Var}(g(\mu) + (r-\mu)g'(\mu)) = V(\mu)\left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)^2 = \tilde{\mathbf{W}}^{-1}$$

up to a factor involving ϕ.

In order to implement this iteratively re-weighted least squares method we can start by taking $\boldsymbol{\mu} = \mathbf{r}$. Certain observations may need to be adjusted, for example zero values when the log or inverse power links are used. The method is easy to implement in a matrix programming language such as MATLAB, APL or SAS IML. Annotated SAS IML code is given in Appendix B.

10. Future Research

Bailey, [2], points out that in statistics the best estimator is a minimum variance unbiased estimator, but that in classification ratemaking there are typically no unbiased estimators.

Venter's third suggestion, of allowing individual cells to vary from an arithmetically defined base, gives a way to produce unbiased estimators. Credibility weighting the model pure premium with the experience would give asymptotically unbiased rates, because in a large enough sample each cell would be fully credible. Venter notes such an approach was used in the 1981 Massachusetts auto rate hearings. The credibility factor used was Bühlmann credibility

$$Z = \frac{n}{n+K}, \qquad K = \frac{\text{Expected process variance}}{\text{Variance of hypothetical means}}, \tag{10.1}$$

where n is the number of exposures in the cell.

A credibility approach was also hinted at by Bailey who discusses the problem of combining information about youthful drivers and business classes into youthful business drivers. "[The data] may be insufficient to be fully reliable but it will always provide *some information*."

The statistical theory of mixed models provides a method of credibility weighting fitted values and raw data. The details of mixed models are beyond the scope of this paper; the interested reader should consult Searle et al. [24]. In fact, Equation (42) on page 57 of Searle uses mixed models to give an unbiased predictor for a cell pure premium as

$$(1 - Z) \times \text{model fit} + Z \times \text{ cell average},$$

where credibility Z is given by Equation (10.1). A very nice recent paper by Nelder and Verrall [19] extends the same result to a certain family of generalized linear mixed models and discusses some possible actuarial applications. Lee and Nelder [18] gives a more detailed description of the theory, together with some (non-actuarial) examples. Aside from their application to credibility theory, mixed models could also be used in territorial ratemaking, just as they are currently used in geophysical statistics—see Cressie [6].

11. Conclusion

We have introduced generalized linear models by making a connection between them and minimum bias models, with which actuaries are already familiar. The connection is made possible by using variance functions to define linear bias functions and then relating them to the exponential family of distributions. The definitions imply that minimum bias corresponds to the maximum likelihood solution of the associated generalized linear model. By starting with the known and familiar we have provided an introduction to generalized linear models which is easier to understand than ones which start from abstract definitions. We have also explained how generalized linear models extend the well known ANOVA and regression analyses. Two by-products of the exposition were to clarify uniqueness of parameters for class plans and to explain the different notations used in linear models and minimum bias methods. Finally, the iterative paradigm for solving minimum bias models is shown not to be useful given the more efficient algorithms available for solving generalized linear models. Actuaries should not implement the iterative method. Whenever possible, they should use explicit statistical models.

Linear bias functions are an alternative to the usual measure of bias and so extend Venter's first alternative to Bailey's methods. Link functions, introduced as part of the definition of generalized linear models, allow for more general arithmetic functions to determine classification rates. However, since the models are still linear they do not allow functions such as $r_{ijk} = x_i y_j + z_k$ suggested by Venter.

In jumping from actuaries of the second kind, who use risk theory and probabilistic models, to actuaries of the third kind[9], who use stochastic models and financial tools, I believe the profession may have overlooked an important intermediate step: the statistical actuary—perhaps actuary of the 5/2nds kind? A statistical approach

[9] Stephen D'Arcy, "On Becoming An Actuary of the Third Kind", PCAS LXXVI p. 45 1989.

is perfect for data intensive lines, such as personal auto and homeowners. I hope this and other statistical papers which have appeared recently will encourage actuaries working in data intensive lines to take statistics beyond that which is required for an Associateship in either North American actuarial society and to start taking advantage of its power in their work.

References

[1] Abraham, Bovas and Johannes Ledolter, *Statistical Methods for Forecasting*, John Wiley and Sons, 1983.

[2] Bailey, Robert A., *Insurance Rates with Minimum Bias*, PCAS **L**, 4–13.

[3] Bailey, Robert A. and LeRoy J. Simon, *Two Studies in Automobile Insurance Ratemaking*, PCAS **XLVII**, 1–19.

[4] Borwein, Jonathan M. and Peter B. Borwein, *Pi and the AGM*, John Wiley and Sons, 1987.

[5] Brown, Robert L., *Minimum Bias with Generalized Linear Models*, PCAS **LXXV**, 187–217.

[6] Cressie, Noel A. C., *Statistics for Spatial Data*, Revised Edition, John Wiley and Sons, 1993.

[7] Graves, Nancy C. and Richard Castillo, *Commercial General Liability Ratemaking for Premises and Operations*, 1990 CAS Discussion Paper Program **II**, 631–696.

[8] Haberman, Steven and A. R. Renshaw, *Generalized linear models and actuarial science*, The Statistician **45(4)** (1996), 407–436.

[9] Hogg, Robert V. and Stuart A. Klugman, *Loss Distributions*, John Wiley and Sons, 1983.

[10] Minutes of Personal Lines Advisory Panel, *Personal Auto Classification Plan Review*, Insurance Services Office **PLAP-96-18** (1996).

[11] Johnson, Norman L., Samuel Kotz, and N. Balakrishnan, *Continuous Univaritate Distributions, Volume 1*, Second Edition, John Wiley and Sons, 1994.

[12] Jørgensen, Bent, *The Theory of Dispersion Models*, Chapman and Hall, 1997.

[13] Jørgensen, Bent, *Exponential Dispersion Models*, J. R. Statist. Soc. B **49(2)** (1987), 127–162.

[14] Jørgensen, Bent and Marta C. Paes de Souza, *Fitting Tweedie's Compound Poisson Model to Insurance Claims Data*, Scand. Actuarial J. **1** (1994), 69–93.

[15] Klugman, Stuart A., Harry H. Panjer and Gordon E. Willmot, *Loss Models, From Data to Decisions*, John Wiley and Sons, 1998.

[16] McCullagh, P. and J. A. Nelder, *Generalized Linear Models*, Second Edition, Chapman and Hall, 1989.

[17] Lee Y. and J. A. Nelder, *Likelihood, Quasi-likelihood and Pseudolikelihood: Some Comparisons*, J. R. Statist. Soc. B **54** (1992), 273–284.

[18] Lee Y. and J. A. Nelder, *Hierarchical Generalized Linear Models*, J. R. Statist. Soc. B **58** (1996), 619–678.

[19] Nelder, J. A. and R. J. Verrall, *Credibility Theory and Generalized Linear Models*, ASTIN Bulletin **27(1)**, 71–82.

[20] Panjer, H. H., and G. E. Willmot, *Insurance Risk Models*, Society of Actuaries, 1992.

[21] Rao, C. Radhakrishna, *Linear Statistical Inference and Its Applications*, Second Edition, John Wiley and Sons, 1973.

[22] Renshaw A. E., *Modeling the Claims Process in the Presence of Covariates*, ASTIN Bulletin **24(2)**, 265–286.

[23] *SAS/STAT Software, Changes and Enhancements through Release 6.11*, SAS Institute Inc., 1996.

[24] Searle S. R., George Casella and C. E. McCulloch, *Variance Components*, John Wiley and Sons, 1992.

[25] Venter, Gary G., *Discussion of Minimum Bias with Generalized Linear Models*, PCAS **LXXVII**, 337–349.

[26] Wright, Thomas S., *Stochastic Claims Reserving When Past Claim Numbers Are Known*, PCAS **LXXIX**, 255–361.

Figure 1

Figure 2

Appendix A

Reconciliation of Notation with Literature

McCullagh and Nelder [16] define the exponential as a two-parameter family of distributions whose density functions can be written in the form

$$f(r;\theta,\phi) = \exp\left((r\theta - b(\theta))/a(\phi) + c(r,\phi)\right). \tag{12.1}$$

Generally $a(\phi) = \phi/w$ where w is a known prior weight. We will assume a has this form. Thus to reconcile Equation (12.1) with Equation (6.1) it is enough to explain what is meant by the identity

$$r\theta - b(\theta) = -\frac{1}{2}d(r;\mu) = \int_r^\mu \frac{(r-t)}{V(t)}dt. \tag{12.2}$$

We must define the function b. Differentiating Equation (12.2) with respect to θ gives

$$r - b'(\theta) = \frac{r-\mu}{V(\mu)}\frac{d\mu}{d\theta}$$

because r is a constant. Taking expected values over r shows $\mu = b'(\theta)$ since $E(r) = \mu$ by Equation (6.2) and so the right hand side vanishes. Substituting for μ and canceling $r-\mu$ shows $V(\mu) = b''(\theta)$. Thus the function b satisfies the differential equation

$$V(b'(\theta)) = b''(\theta), \tag{12.3}$$

which is enough to determine b; θ is simply an argument.

12.1 Example 6.1 revisited

Example 6.1 showed that the gamma distribution belongs to the exponential family by deriving the deviance function from the density function. We now assume the form of the variance function and derive the density using the function b. $V(\mu) = \mu^2$ corresponds to the gamma distribution, so Equation (12.3) gives

$$(b'(\theta))^2 = b''(\theta)$$

whence

$$\mu = b'(\theta) = -\frac{1}{\theta}$$

and

$$b(\theta) = -\log(-\theta).$$

Plugging into Equation (12.1) gives exactly Equation (6.6) with $\phi = 1/\nu$.

12.2 Connection with generalized linear models

To solve for the parameters of a generalized linear model using maximum likelihood directly from Equation (12.1) it is necessary to differentiate the log likelihood of an observation r_i

$$l(\theta, \phi; r_i) = l = w_i \left(r_i\theta - b(\theta) \right) / \phi + c(r_i, \phi)$$

with respect to β_j. Using the chain rule and substituting $\mu = b'(\theta)$, $d\mu/d\theta = b''(\theta) = V(\mu)$ gives

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \frac{\partial l}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j} \\ &= \frac{w_i(r_i - b'(\theta))}{\phi}\frac{1}{b''(\theta)}\frac{d\mu}{d\eta}x_{ij} \\ &= \frac{w_i(r_i - \mu)}{\phi}\frac{1}{V(\mu)}\frac{d\mu}{d\eta}x_{ij} \end{aligned}$$

which is Equation (5.3) for one observation $r_i = r$, up to a factor of ϕ which cancels out.

Appendix B

Computer solution of generalized linear models

This section contains annotated SAS IML code to compute the parameters for a generalized linear model with log link and gamma errors.

The dataset CARDATA contains the following variables

(1) AGE the age group classification

(2) VUSE the vehicle use classification

(3) LOSS the average severity

(4) NUMBER the number of claim counts

as shown in Exhibit 1.

Comments in SAS are enclosed between * and ;. In IML the statement * denotes matrix multiplication, # denotes componentwise multiplication, and ## denotes componentwise exponentiation.

The author will e-mail a copy of this code on request.

```
DATA CARDATA;
INPUT AGE VUSE LOSS NUMBER;
CARDS;
data lines
;

PROC IML;

* READ ALL VARIABLES INTO IML VARIABLES AGE, VUSE, R AND W ;
USE CARDATA;
READ ALL VAR AGE INTO AGE;
READ ALL VAR VUSE INTO VUSE;
READ ALL VAR LOSS INTO R;
READ ALL VAR NUMBER INTO W;

* COMPUTE DESIGN MATRICES ;
A = DESIGN(AGE);
B = DESIGN(VUSE);
* SELECT A BASE CLASS BY DELETING A COLUMN OF B ;
* [,1:3] MEANS SELECT COLUMNS 1 THRU 3 ;
B = B[,1:3];
* MODEL DESIGN MATRIX = HORIZONTAL CONCATENATION OF A AND B ;
X = A||B;

* DEFINE A FUNCTION TO COMPUTE THE VARIANCE FUNCTION FOR A GAMMA DISTRIBUTION ;
START VARFUN(MUIN);
RETURN(MUIN# MUIN); * COMPONENTWISE MULTIPILCATION ;
FINISH;
```

```
* WEIGHTS FOR THE LOG LINK, PER Equation (7.3);
START W(MUIN);
ANS = MUIN# # 2 / VARFUN(MUIN);
RETURN(ANS);
FINISH;

* INITIALIZE WITH DATA ;
MU = R;
ETA = LOG(MU);

* SET UP HOLDERS FOR CURRENT AND PREVIOUS PARAMETERS ;
* J(NCOL(X),1,10) RETURNS A NCOL(X) x 1 MATRIX WITH VALUE 10, ETC ;
LASTBETA = J(NCOL(X),1,10);
BETA = J(NCOL(X),1,0);

* WHILE SQUARED DISTANCE BETWEEN BETA AND LAST BETA IS LARGE DO ;
DO WHILE((BETA-LASTBETA)‘ * (BETA-LASTBETA) > 1E-9);
  * COMPUTE AUXILLARY VARIABLE ;
  Z = ETA + (R - MU) # DETADMU(MU);

  * SAVE LAST BETA VECTOR ;
  LASTBETA = BETA;

  * DO WEIGHTED LEAST SQUARES;
  * NOTE: GINV = INVERSE ;
  WEIGHT = W(MU) # W;
  BETA = GINV(X‘ * ( WEIGHT # X)) * X‘ * (WEIGHT # Z);

  * COMPUTE PREDICTED VALUES ;
  ETA = X * BETA;
  MU = EXP(ETA);
END;

* PRINT OUT PARAMETERS ;
PRINT I BETA[F=8.4];

* NOW COMPUTE THE VARIOUS STATS, DEVIANCE AND SO FORTH ;
* MU AND ETA ALREADY HOLD THE LAST ESTIMATES OF PRED VALUES ETC;

* COMPUTE VAR;
VAR = VARFUN(MU);

* COMPUTE GAMMA DEVIANCE ;
DEV = 2 # W # (-LOG(R / MU) + ((R-MU) / MU));

* PEARSON RESIDUAL AND DEVIANCE RESIDUALS ;
PEARES = (R - MU ) / SQRT(VAR);
DEVRES = SIGN(R - MU) # SQRT(DEV);

NOBS = NROW(X); * NUMBER OF OBSERVATIONS ;
NPARAM = NCOL(X); * NUMBER OF PARAMETERS ;
DF = NOBS - NPARAM; * NUMBER OF DEGREES OF FREEDOM ;

PEARSON = (PEARES# PEARES)[+]; * [+] = SUM OVER COMPONENETS ;
DEVIANCE = DEV[+];
PRINT PEARSON, (PEARSON / DF)[LABEL="DISPERSION = PEARSON/DF"],
  DEVIANCE, (DEVIANCE / DF)[LABEL="DEVIANCE/DF"];

* LOGLIKELIHOOD FOR GAMMA DISTRIBUTION ;
PHI = DEVIANCE / DF; * ESTIMATE FOR PHI ;
LLH = (W/PHI) # LOG(W # R / (PHI # MU))
- W # R / (PHI # MU) - LOG(R) - LGAMMA(W/PHI);
* LGAMMA = LOG(GAMMA FUNCTION) ;

PRINT (LLH[+]);


** ABOVE CODE WILL GIVE THE SAME RESULT AS THE FOLLOWING CODE ;
** USING THE BUILT-IN SAS GENERALIZED LINEAR MODEL ROUTINE, PROC GENMOD ;

PROC GENMOD DATA=CARDATA;
```

```
CLASS AGE VUSE;
SCWGT NUMBER;
MODEL R = AGE VUSE / NOINT DIST=GAMMA LINK=LOG DSCALE;
RUN;
```