FRESH LOOK By STEPHEN J. MILDENHALL

Bailey Simon Minimum Bias Reexamined, Part 2

Actuarial Review *introduces a new column, Fresh Look, that aims to reassess core areas in actuarial science with more current tools and practices. Part 1 appears in* Actuarial Review *November-December 2020.*

Development

In a GLM, an observation's mean value is a function of a linear combination of covariates, and the observation is sampled from an exponential family distribution. The parameters are determined using maximum likelihood. The function linking the mean domain to the linear domain is called the link function, customarily denoted g.

Exponential family distributions are assumed to be nondegenerate. They are parameterized by a canonical parameter θ that is a function of the mean, and which we will identify in a moment. Most importantly, their density (or probability mass function) factors as

$$f(y;\theta) = c(y)b(\theta)e^{y\theta},$$

with symmetric roles for the observation y and parameter θ. Both c and b are non-negative functions. The factorization reflects the dual meaning of the density: It is the probability of observing y if the true parameter is θ as well as the likelihood of the parameter θ given an observation y.

Since b is non-negative, we can write $b(\theta) = e^{-\kappa(\theta)}$ on the support of f, giving $f(y;\theta) = c(y)e^{Y\theta-\kappa(\theta)}$. It follows that the log likelihood of θ is $l(y;\theta) = \log(c(y)) + y\theta - \kappa(\theta)$. Differentiating with respect to θ and setting equal to zero shows the maximum likelihood estimator (MLE) of θ given y solves the score equation $y - \kappa'(\theta) = 0$. Given a sample of independent observations $y_1, \dots, y_n$, the MLE solves $\bar{y} - \kappa(\theta) = 0$, where $\bar{y}$ is the sample mean. Thus, the mean is a sufficient statistic for θ in an exponential family.

If a random variable Y has an exponential family distribution with density f, then it has a cumulant generating function $K(t) := \log E[e^{tY}]=\kappa(t+\theta)-\kappa(\theta)$. The mean of Y is given by $K'(0)=\kappa'(\theta)=\mu$, which identifies the relationship between μ and θ. $\kappa'(\theta)$ is often denoted $\tau(\theta)$. The variance of Y is given by $K''(0) = \kappa''(\theta) = \tau'(\theta)$. By assumption, exponential family distributions are nondegenerate and therefore have a strictly positive variance. Three important conclusions follow:

1. That K is a convex function, and hence l is concave ensuring a unique maximum likelihood estimate.
2. That τ is increasing and hence invertible.
3. This implies that the variance of Y is a function of its mean.

If we start with a variance function defined on a mean domain we can work backwards, solving two differential equations, to determine a cumulant generating function and hence a unique exponential family distribution with that variance function and domain.

The third conclusion, the mean-variance relationship, is captured by the variance function, $V(\mu)=\kappa''(\tau^{-1}(\mu))=1/(\tau^{-1})'(\mu)$ (chain rule).

If we start with a variance function defined on a mean domain we can work backwards, solving two differential equations, to determine a cumulant generating function and hence a unique exponential family distribution with that variance function and domain. V only determines the distribution uniquely within the exponential family, not within all distributions. For example, kX for any X with $E[X]=1$ and $\mathrm{Var}(X)=1$ has $V(\mu)=\mu^2$, but the only exponential family distribution with variance function $V(\mu)=\mu^2$ is the gamma (with a different parameterization).

It is possible to show that using the exponential family distribution with variance function V is equivalent to making no assumptions other than the mean-variance relationship. Technically, the exponential family has minimal Fisher information. This is a very reassuring fact for the modeler, who must specify some distribution to build a statistical model necessary to evaluate Bailey and Simon's criteria. But making a choice is fraught: what evidence backs it up?

The actuary knows from the physical, economic and contractual operation of insurance that a reasonable V will fall between a linear and a quadratic function. Using an exponential family distribution can test various alternatives in this range while making no additional assumptions. And the story gets better. It turns out that every $1<p<2$ determines an exponential family distribution with $V(\mu)=\mu^p$, called a Tweedie distribution. Tweedie distributions are ideal for modeling insurance losses because they are compound Poisson distributions with a gamma severity (the identification is made by solving the differential equations alluded to above and identifying the resulting cumulant generating function). They take non-negative values and are continuous except for a probability mass at 0. As p approaches 1 from above ($p{\downarrow}1$), the Tweedie approaches a Poisson, and as p approaches 2 from below ($p{\uparrow}2$), a gamma.

GLMs encompass a wide range of model forms. They are much more flexible than normal-error general linear models because they separate the linearizing transformation, the link function, from the error distribution. A linear model uses the same function to linearize and to stabilize the variance.

Now consider the fourth criterion: chance. Let's model Y using an exponential family distribution with the identity link function. Given an observation y in a cell with fitted mean μ, how should we evaluate whether the difference $y-\mu$ "could reasonably be caused by *chance*"? The residual error, $y-\mu$, lacks scale and context. The theory of linear models suggests various standardized residuals, such as the Pearson residual $(y-\mu)/\sqrt{(V(\mu))}$. A frequentist creates a confidence interval such as $y\pm2\sqrt{(V(\mu))}$ for the class mean. If μ falls within the confidence interval, then the experience could reasonably occur by chance. An obvious problem with this approach is the need for it to hold simultaneously for many observations, which will be vanishingly small.

Alternatively, we can use likelihood to evaluate chance. A class rate is likely if its likelihood is close to the maximum likelihood. In the mean parameterization, the log likelihood becomes $l(y;\mu)=\log(c(y))+y\tau^{-1}(\mu)-\kappa(\tau^{-1}(\mu))$. At the maximum of l, the score function

$$\frac{\partial l}{\partial \mu}=\frac{y-\mu}{V(\mu)}=0.$$

Remember, $\kappa'(\tau^{-1}(\mu))=\mu$ by definition. Thus the score is a good measure of chance. For the most likely parameter it is zero. When the score is small the rate μ is reasonably likely, but when it has a large absolute value, l falls off quickly from its maximum value and μ is much less likely. Although dividing by the variance rather than standard deviation seems odd from a classical statistics perspective, it makes sense when considering likelihoods.

Finally, we need an overall assessment of model fit that avoids arbitrary choices. We can create one from the likelihood function. We can compare the model-constrained likelihood with an unconstrained, saturated model likelihood to get a measure called model deviance. Since we already know the maximum likelihood estimate for μ is y, the deviance will be

$$d(y;\mu)=2(l(y;y)-l(y;\mu))\geq0.$$

The factor of 2 is included to ensure agreement with the normal distribution. Since $\partial d/\partial\mu$=-$2\partial l/\partial\mu$ we see

$$d(y;\mu)=2\int_{\mu}^{y}\frac{y-m}{V(m)}dm.$$

The limits of integration are chosen so that d has the correct derivative, forcing μ on the bottom, and $d(y;y)=0$ forcing y on top. Notice that the nuisance $\log(c(y))$ term in l disappears in d.

What is the deviance for a Tweedie, $V(\mu)=\mu^p$? For $p\neq1,2$, simply integrate:

$$\frac{d(y;\mu)}{2}=\int_{\mu}^{y}\frac{y-m}{m^p}dm=\frac{ym^{-p+1}}{1-p}-\frac{m^{-p+2}}{2-p}\Bigg|_{\mu}^{y}$$
$$=-\frac{y^{2-p}}{(2-p)(p-1)}+\frac{y\mu^{1-p}}{p-1}+\frac{\mu^{2-p}}{2-p}.$$

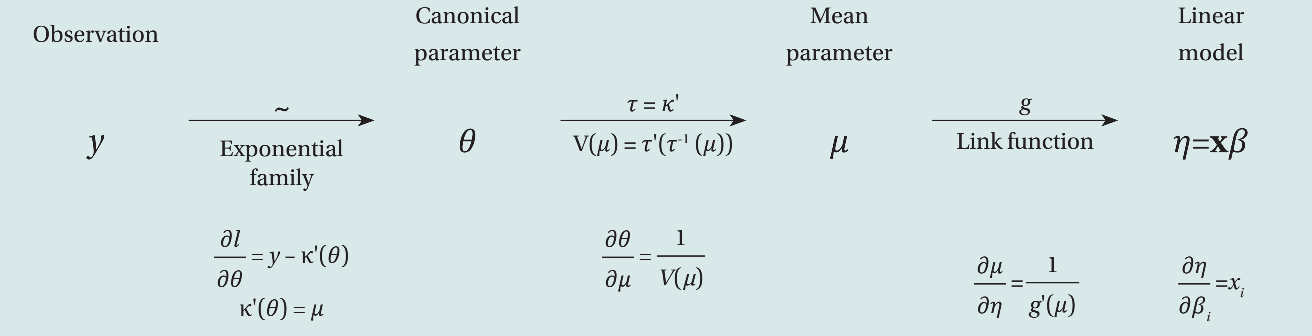
The density of the exponential family can be expressed in terms of the deviance as

$$f(y;\mu)=c_0(y)\exp\left\{-\frac{d(y;\mu)}{2}\right\}$$

where $c_0(y)=c(y)\exp(l(y;y))$. It is an easy exercise to check that when $V(\mu)=1$ the deviance is $(y-\mu)^2$, and so the corresponding exponential family distribution is the normal. (Exercise: Work out which distribution corresponds to $V(\mu)=\mu$.)

To summarize: we can fit a GLM using maximum likelihood or, equivalently, using minimum deviance. The deviance provides a measure of model fit customized to each exponential distribution family and can be used to compare models using that error distribution. Scaled differences in devi-

Figure 1


Observation
Canonical parameter
Mean parameter
Linear model
y
$\sim$
Exponential family
θ
$\tau=\kappa'$
$V(\mu)=\tau'(\tau^{-1}(\mu))$
μ
g
Link function
$\eta=\mathbf{x}\beta$
$\frac{\partial l}{\partial\theta}=y-\kappa'(\theta)$
$\kappa'(\theta)=\mu$
$\frac{\partial\theta}{\partial\mu}=\frac{1}{V(\mu)}$
$\frac{\partial\mu}{\partial\eta}=\frac{1}{g'(\mu)}$
$\frac{\partial\eta}{\partial\beta_i}=x_i$

ance have an asymptotic χ^2 distribution. Other methods are needed to choose between models using different error distributions. Deviance generalizes the fact that maximum likelihood for the normal is the same as minimum square error.

GLMs encompass a wide range of model forms. They are much more flexible than normal-error general linear models because they separate the linearizing transformation — the link function — from the error distribution. A linear model uses the same function to linearize and to stabilize the variance. Linear, logistic and Poisson regressions, and analysis of variance are all special cases of GLMs.

Suppose the linear predictor for a unit (observation) y is specified as $\eta=\mathbf{x}\beta$, where $\mathbf{x}$ is a vector of covariates and β is a parameter vector, and the mean of y is linked to η by $g(\mu)=\eta$. Then the log likelihood function becomes $l(y;\mu)=\log(a(y))+y\tau^{-1}(g^{-1}(\mathbf{x}\beta))-\kappa[\tau^{-1}(g^{-1}(\mathbf{x}\beta))]$. Therefore, using the chain rule, the score for β_i is given by

$$\frac{\partial l}{\partial\beta_i}=\frac{\partial l}{\partial\theta}\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\beta_i}=\left(\frac{y-\mu}{V(\mu)}\right)\frac{1}{g'(\mu)}x_i.$$

The decomposition of the score reflects the components of the GLM. (See Figure 1.)

When the linear model is a two-way classification, the score equations $\partial l/\partial\beta_i=0$ give the famous Bailey minimum bias iterations, only substituting a variance-adjusted $(y-\mu)/V(\mu)$ bias measure in place of the normal model's $y-\mu$. While not recommended for production work, the iterative solution is easy to implement in a spreadsheet, providing an excellent way to test your understanding and confirm results from R `glm` or SAS `proc genmod` or other implementations—see the example below.

Parameters determined by solving a minimum bias iterative scheme generally agree with the maximum likelihood estimates of a GLM with some variance function, even when the scheme is formulated without an explicit statistical model. The situation is analogous to Mack's identification of the stochastic model underlying the chain-ladder method. Before Mack, we happily squared triangles without knowing the underlying assumptions. But knowing the implied statistical model is an essential part of assessing whether the model is appropriate for its intended use.

Examples

Here are two simple examples which capture the essence of the modeling problem. Assume that each cell contains the same number of exposures and model using an exponential family distribution with variance function $V(\mu)=\mu^p$.

The first example is a two-way classification, with each level taking two values. You can think: youthful operator yes/no and prior accidents yes/no. The observations for no/no, no/yes, yes/no, yes/yes are $y_0=1$, $y_1=2$, $y_2=3$, and $y_3=7$. The linear model has means β_0, β_1, β_2 and $\beta_1+\beta_2-\beta_0$ (equivalently, β_0, $\beta_0+\beta_1$, $\beta_0+\beta_2$ and $\beta_0+\beta_1+\beta_2$).

The second is a linear regression, with covariate taking values 0, 1, 2 and outcomes 1, 2 and 4.

In both cases it is clear the model does not fit perfectly. How should the "bias" be apportioned between the classes? The appropriate bias is variance-adjusted, $(y-\mu)/V(\mu)$.

In the first model the bias for each cell has the same absolute value b, and is split b,-b, -b,b, to achieve balance by class and in total. In the linear model it will be b,-2b,b, achieving a covariate-weighted analog of balance ($\partial\eta/\partial\beta_1$=0,1,2 for the three observations). The value of b depends on V, i.e., on p, reflecting the fact there are many balanced models.

To find a specific solution, set up a spreadsheet as shown below and use Solver to minimize the deviance (computed in the Development section) over β_i. The tables show the solution for p=1.6. Solver will readily handle the problem because the deviance is a well-behaved, concave function with a unique maximum. You could also use the minimum bias iterations, or mimic the GLM iteratively re-weighted least squares algorithm. All of these are easy to implement in Excel. It is worth noting that the solutions are maximum likelihood parameter estimates for a density that you can't actually write down in closed form!

Exercise: What happens to the fit as you vary p? Why? (See Figure 2.)

It's always good to double check your work. The R code below reproduces the Excel Solver solution.

```
library(tidyverse)
library(statmod)

# two way classification
df = tibble(a=c(1,0,0,-1),
b=c(0,1,0,1), c=c(0,0,1,1),
y=c(1,2,3,7))
m1 = glm(data=df,
family=tweedie(var.power=1.6,
link.power=1), y~a+b+c-1)
summary(m1)
```

```
# Coefficients:
#  Estimate Std. Error t value
Pr(>|t|)
# a 0.91075  0.50969 1.78687
0.32481
# b 2.42873  1.04994 2.31320
0.25977
# c 3.92350  1.38479 2.83328
0.21600
#
# Residual deviance: 0.3086021
on 1 degrees of freedom

# linear regression
df2 = tibble(x=c(0,1,2),
y=c(1,2,5))
m2 = glm(data=df2,
family=tweedie(var.power=1.6,
link.power=1), y~x)
summary(m2)

# Coefficients:
#       Estimate Std. Error t
value Pr(>|t|)
# (Intercept) 0.939632
0.342186 2.74597 0.22233
# x      1.684947  0.525511
3.20630 0.19247
#
# Residual deviance: 0.1422328
on 1 degrees of freedom
```

Lessons

GLMs allow actuaries to model with an error distribution that incorporates known facts about the loss generating process, but overlays no further arbitrary assumptions. The distribution is specified by the relationship between the mean and variance. It provides a variance-adjusted score, or measure of bias, that satisfies the balance equations and a quantification of model fit. Model parameters can be estimated using an efficient algorithm, implemented in R and Python, or from first principles in a simple spreadsheet. GLMs naturally extend Bailey and Simon's four criteria, giving them more exact meaning. Since GLMs assume the input data is representative, unbiased and credible, the modeler must always exercise good judgment. Nevertheless, GLMs provide an excellent framework that the actuary can use to build fair and transparent rates. Long live statistics and rational, fact-based government!

Stephen J. Mildenhall, FCAS, CERA, PhD, ASA, MAAA, is a consultant with Convex Risk LLC and a member of the CAS Board of Directors.

Figure 2.

x_{i0}	x_{i1}	x_{i2}	y_i	β	μ	$V(\mu)$	Score b	$d(y;\mu)$
1	0	0	1	0.91075	0.91075	0.86107	0.10365	0.00880
0	1	0	2	2.42871	2.42871	4.13620	-0.10365	0.04917
0	0	1	3	3.92352	3.92352	8.91006	-0.10365	0.10996
-1	1	1	7		5.44148	15.03651	0.10365	0.14068
								0.30860

Constant	x_i	y_i	β	μ	$V(\mu)$	Score b	$d(y;\mu)$
1	0	1	0.93963	0.93963	0.90518	0.06669	0.00389
1	1	2	1.68495	2.62458	4.68269	-0.13338	0.09586
1	2	5		4.30952	10.35349	0.06669	0.04248
							0.14223