Bailey Simon Minimum Bias Reexamined

Stephen J. Mildenhall

2020-10-20

1 Bailey Simon Minimum Bias Reexamined

1.1 Bias

Society is beset with problems of bias, inequity, and unfairness. Insurance, in particular, relies on the perception and reality of fairness. Insureds will only pool their risk with one another when they believe everyone pays a fair and unbiased rate. So insurers must not only treat all insureds equitably, but they must also be able to demonstrate that they do so. However, their complex and granular rating plans make that more challenging. As a result, media and regulators have begun to question and investigate rating models. The NAIC Casualty Actuarial and Statistical Task Force drafted a white paper describing the Regulatory Review of Predictive Models and completed a Price Optimization White Paper in 2015, and the CAS is reconsidering its position. Against this backdrop, now seems a good time to reexamine two very famous Proceedings papers proposing to minimize bias in insurance rates.

Insurance rates should be based on data and not prejudice. Establishing fairness is challenging and encompasses many issues, such as proxy variables and differential impact. The modeler must use a rigorous and transparent framework that avoids arbitrary, unnecessary, or hidden assumptions. This article explains that a generalized linear model (GLM), the natural outgrowth of minimum bias methods, satisfies these requirements, providing an ideal model-building platform. While it is possible to build a flawed GLM model, it is reassuring to know they provide a neutral starting point.

It is important to remember that residual error is a modeling fact of life. An oft-quoted aphorism states, "All models are wrong; some are useful." Models simplify to be useful, but by simplifying, they omit details and are wrong. A statistical model balances fidelity to sample data with out-of-sample predictive power to maximize its usefulness. An actuarial statistical model creates a rating plan to predict expected loss costs and distributions for each insured. Various standards are used to judge if a rating plan is acceptable. US actuaries are familiar with the CAS Ratemaking Principle that rates should be "**Reasonable, not excessive, not inadequate, and not unfairly discriminatory**."

Another set of criteria, almost as well known and pre-dating the CAS principles by nearly 30 years, was written down by Robert Bailey and LeRoy Simon in their 1960 Proceedings paper "Two Studies in Automobile Insurance Ratemaking." A 1963 follow-up by Bailey, "Insurance Rates with Minimum Bias," developed them further. It is instructive to reexamine Bailey and Simon's criteria in the light of what we have learned since then and the issues we currently face as the front-line guardians of fair insurance rates.

1.2 Criteria

Bailey and Simon's concern was personal automobile ratemaking in Canada. At the time, pricing used a two-way classification plan combining a (very coarse) class plan and a merit (experience) rating plan. Their four criteria are as follows, with italics in the original. A set of pricing relativities is acceptable if

1. It reproduces the experience for each class and merit rating class and also the overall experience; i.e., is *balanced* for each class and in total.
2. It reflects the relative *credibility* of the various groups involved.
3. It provides a minimal amount of *departure* from the raw data for the maximum number of people.
4. It produces a rate for each sub-group of risks which is close enough to the experience so that the differences can reasonably be caused by *chance*.

1.3 Assumptions

The Bailey-Simon criteria rely on several assumptions.

Balanced by class means the average rate equals each class's experience rate, summed over remaining classes. This formulation gives particular prominence to the average, or mean, and uses the difference to measure balance (residual error). It also implies that each class is large enough to be fully credible.

The discussion of relative credibility appeals to the general statistical principle of weighting an estimate in inverse proportion to its standard error. Bailey and Simon give each cell's experience a weight proportional to the square root of its expected number of losses because they assume the variance of experience loss grows with its expectation.

Bailey and Simon frame the third criterion in terms of "inequity" or deviation from experience. It is worth quoting their discussion because of its topical relevance.

> *Anyone who has dealt directly with insureds at the time of a rate increase, knows that you can be much more positive when the rate for his class is very close to the indications of experience. The more persons involved in a given sized inequity, the more important it is.*

The ability to explain rates was as necessary in 1960 as it is today! Bailey and Simon quantified the departure criteria using the average absolute deviation.

Chance, the fourth criterion, they assessed using a weighted χ^2-statistic. Based on Canadian experience, they determined that the difference between actual and expected relative loss ratio, scaled by the former's standard deviation, is approximately a standard normal, justifying their selection. They then derived a minimum bias iterative scheme to solve for the minimum χ^2 relativities and show it is balanced.

Bailey's 1963 paper generalized the minimum bias iterative scheme, discussed additive (cents) and multiplicative (percents) models, and the need for a measure of model fit distinct from average model bias (which is zero, by design). He proposed minimum square error and minimum absolute error measures.

Bailey and Simon's principal innovation was to calculate all class relativities at once, reflecting different mixes across each variable. Until their work, rating factors were computed one at a time, in a series of univariate analyses. (This is different from considering interactions between rating factors. Their two-factor rating plan was too simple to allow for interactions.) The minimum bias method was, and remains, very appealing: it is easy to explain and intuitively reasonable (who doesn't want their rating plan to be balanced by class?), and is simple to program. It is no wonder it proved so popular.

1.4 Critique

Certain aspects of Bailey and Simon's work may be tricky for today's statistically trained actuary to follow. The use of the word bias is non-standard. In statistics, an estimator is unbiased if its expected value over samples gives the correct value. Bailey and Simon use bias to mean residual error, the difference between a fitted value and an observation, and as a measure of overall model fit. Balance is also used to describe the residual error in a fitted value.

The focus on the sample mean as a sufficient statistic for the population mean needs no explanation.

The concept of balance by class relies on the form of the linear model underlying the classification. Bailey and Simon use a two-way classification model. The rate for risks in class (i, j) is $x_i + y_j$, in the additive model. The underlying design matrix only has elements 0 and 1. In a more general setting, including continuous covariates, the design matrix would be more complex. Some analog of balance would still apply, but it would be more complicated to explain.

Bailey and Simon place great emphasis on the concept of fully credible rating classes, meaning ones where the model rate should exactly equal the experience rate. A statistical approach quantifies the outcome distribution explicitly and produces tighter and tighter confidence intervals for the model rate, rather than insist on equality. Some sampling error or posterior uncertainty remains for the largest cells, even if very small.

The claim that the variance of experience grows linearly with expected losses in each class is most interesting for the modeler. It reflects a traditional actuarial compound Poisson claims generating process. A severity distribution and an annual frequency characterize each risk cell. The distribution of aggregate losses has a Poisson frequency distribution, with mean proportional to expected losses, and a fixed severity distribution. Its variance is proportional to its mean. These assumptions can fail in at least two ways.

First, there can be common risk drivers between insureds, such as macroeconomic conditions or weather. These result in a correlation between insureds. A negative binomial frequency captures the effect, replacing the Poisson. The resulting aggregate distribution has a variance of the form $\mu(a + b\mu)$ for constants a and b, where μ is the mean. The variance of a large portfolio grows almost quadratically with its mean.

Second, a quadratic mean-variance relationship can arise for catastrophe risks, where portfolio growth corresponds to paying a greater proportion of losses over a fixed set of events. The actuary's understanding of the loss generating process informs the possible relationship between the mean and the variance of losses in a cell. It should fall somewhere between linear and quadratic.

Bailey and Simon test the fourth criterion, that each sub-group's experience should be close enough to its rate that differences could reasonably be caused by chance, using an aggregate χ^2-statistic. There is a clear opportunity to enhance model assessment using a granular, cell-by-cell evaluation of chance deviation, based on the modeled distribution of losses.

Finally, the discussion of both the third and fourth criteria introduce modeler discretion: which measure of overall model bias should be employed? Least squares, minimum absolute deviation, and minimum χ^2 are all mooted. The modeler should avoid unnecessary choices. Is there a better way to select a measure of model fit?

1.5 Homework

In the next sections, we will see how modern statistics has developed the ideas presented so far. As an ex-college professor, I would be remiss if I didn't give you some homework to prepare. Although data science deals with massive data sets and builds very complex models, you can understand its fundamental problems by considering straightforward examples. Here are two that capture our essential conundrum. It would help if you considered how to solve them before reading the sequel.

The first is a two-way classification, with each level taking two values. You can think: youthful operator yes/no and prior accidents yes/no. The data is laid out below. You want to fit an additive linear model.

Level 2 \ Level 1	No	Yes
No	1	2
Yes	3	7

The second is a simple linear regression problem. You want to fit a line through the following data.

Observation i	Covariate x_i	Observation y_i
1	0	1
2	1	2
3	2	4

In both examples, assume the same volume of data underlies each observation, so there is no need for weights. In the first, make the Bailey and Simon assumption that the total experience across each level of each dimension is credible, i.e., the row and column totals are credible.

For partial credit, start by laying out the first question so it looks more like the second one.

The difficulty is clear: there are fewer parameters than data points, and so the requested model will not fit exactly. How should you apportion the model miss? Obviously, with a clever selection of response function you can create many models that do fit exactly—or *over*-fit exactly. Please resist the urge to expound upon these and focus on the stated question.

1.6 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 non-degenerate. 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

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

with symmetric roles for the observation y and parameter θ. Both c and k 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 k 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[1] $K(t) := \log \mathsf{E}[e^{tY}] = \kappa(t+\theta) - \kappa(\theta)$. 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 non-degenerate and therefore have a strictly positive variance. Three important conclusions follow:

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

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 $\mathsf{E}[X] = 1$ and $\mathsf{Var}(X) = 1$ has $V(\mu) = \mu^2$,

[1] Cumulants are combinations of higher moments that are additive for independent variables. The first three cumulants are the mean, variance and third central moment; thereafter they are different from central and non-central moments. The cumulant generating function behaves analogously to the moment generating function.

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 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 $p \downarrow 1$, the Tweedie approaches a Poisson and as $p \uparrow 2$, a gamma.

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 $y \pm 2\sqrt{V(\mu)}$ 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[2]

$$d(y; \mu) = 2(l(y; y) - l(y; \mu)) \geq 0.$$

[2] There is a hidden assumption here; can you see it?

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 $p\neq 1,2$, simply integrate:

$$\begin{aligned}\frac{d(y;\mu)}{2} &= \int_{\mu}^{y}\frac{y-m}{m^p}dm \\ &= \left.\frac{ym^{-p+1}}{1-p}-\frac{m^{-p+2}}{2-p}\right|_{\mu}^{y} \\ &= -\frac{y^{2-p}}{(2-p)(p-1)}+\frac{y\mu^{1-p}}{p-1}+\frac{\mu^{2-p}}{2-p}.\end{aligned}$$

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 $c_0(y)=c(y)e^{l(y;y)}$. 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 deviance have an asymptotic chi^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 regression, 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:

Observation | Canonical parameter | Mean parameter | Linear model

$$\begin{tikzcd}
y \arrow[r, "\sim", "\text{Exponential family}"'] & \theta \arrow[r, "\tau = \kappa'", "V(\mu)=\tau'(\tau^{-1}(\mu))"'] & \mu \arrow[r, "g", "\text{Link function}"'] & \eta = \mathbf{x}\beta
\end{tikzcd}$$

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

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 [1], 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.

1.7 Examples

Here are simple two 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?

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

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

```r
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
```

1.8 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 the actuary can use to build fair and transparent rates. Long live statistics and rational, fact-based government.

References

1. Mildenhall, S.J.: A systematic relationship between minimum bias and generalized linear models. Proceedings of the Casualty Actuarial Society. LXXXVI, 393–487 (1999)