GLM Invariants

Fred Klinker, FCAS, MAAA

Abstract: Those familiar with classic linear regression, as many actuaries are, are aware that, for any regression including an intercept term, there is an exact balance (equality) between (weighted) fitted and (weighted) observed values in aggregate over the whole dataset. Many are also aware that this balance also holds in aggregate over any level of any classification variable appearing in the regression as a main effect. What many may not be aware of is where these balances come from or the fact that they sometimes, but not always, extend to the GLM setting. This paper will discuss the source of the balance conditions in the so-called GLM "Normal Equations". In those cases where balance does not hold, the Normal Equations imply another invariance. The paper will also discuss some applications of these invariants.

Keywords: Generalized Linear Models (GLMs), balance conditions, invariants.

1. INTRODUCTION

Again and again over the years I have heard actuarial students new to GLM modeling express concern that, after fitting a GLM, their mean fitted values sometimes do not match their mean observed values. For anyone raised on classical linear regression, this match between mean fitted and mean observed is taken as a given, at least for regressions including an intercept term. Furthermore, in classical linear regression the match between mean fitted and mean observed holds for each and every level of any classification variable appearing as a model main effect, and this match also frequently appears to vanish in the GLM setting. What is going on? There are invariants for each GLM, depending on the distribution and the link function, but they aren't always the match between mean fitted and mean observed.

The reason has to do with what are called the "GLM Normal Equations". With each GLM combination of assumed distribution of the dependent variable and link function there is associated a set of "Normal Equations", one equation for each model regressor. These can be rewritten in the form of an equality between two quantities, one involving the observed values of the dependent variable, the other where the observed values have been replaced by their fitted values. So these are of the form of an invariance, where the GLM fit preserves some quantity when observed values are replaced by their fitted values. Frequently, this invariance is of the form of an equality between a weighted sum of fitted values and the same weighted sum of observed values. Actuaries would recognize this as the balance condition they have come to expect from classic regression.

Indeed, classic regression, assuming normally distributed errors and an identity link, produces just such balance conditions. So do Poisson count GLMs, assuming Poisson distributed dependent

variables with a log link, and so do logistic regressions, assuming binomially distributed 0/1 dependent variables and a logit link. But, before we become complacent and start to think that these balance conditions are universal, note that balance need not be preserved for severity model GLMs assuming the dependent variable gamma distributed with a log link, which is a very common actuarial model, not only for severities but also for pure premiums and loss ratios. The normal equations for GLMs with gamma distributed dependent variable and log link produce a different set of invariants other than classic balance. Even when balance is not preserved, however, the normal equations can shed some light on the sign and magnitude of the off-balance, as well as some insight as to the source of the off-balance.

It is also sometimes the case that we have available a number of possible weighting variables for our GLMs, and diagnostic residual plots may fail to give clear guidance as to which of these might be preferred. Sometimes the GLM normal equations and their implied invariants will indicate that one weighting variable will come closer to preserving balance than its competitors, and this admittedly extra-statistical, actuarial consideration may be enough to tip the balance in favor of choosing this weight over its competitors.

1.1 Outline of Remainder of Paper

The remainder of this paper proceeds as follows. Section 2 will discuss GLM normal equations and the invariants they imply. Section 3 will show what the normal equations can reveal in an off-balance situation, at least with respect to the very common GLM with gamma distributed dependent variable and log link. Section 4 will show how the normal equations can provide some guidance with regard to choice of weights, and section 5 summarizes.

2. GLM NORMAL EQUATIONS AND THEIR RESULTING INVARIANTS

One solves for the regression coefficients associated with explanatory variables in a regression or GLM by taking partial derivatives of the loglikelihood with respect to each of the regression coefficients and equating them to zero, verifying that each local extremum is indeed a local max, and hopefully a global max as well. These partial derivatives set to zero are the GLM "Normal Equations". They are a set of simultaneous equations, one equation for each regressor, actually one equation for each column of the model design matrix X, where the classic linear regression is written as the matrix expression $Y \sim X\beta$ (Y varies as X beta), where Y is an n-vector of observations, X is an n by p matrix whose columns are the regressor values, for which each row represents one observation, and β is a p-vector of the regression coefficients. If the model includes an intercept term, then X includes a column of all ones to capture that intercept. Each model regressor is

included as a column of X where the values in that column are the values of the regressor. If the model includes classification variables, then there are columns in X that are indicator variables for membership in each level of the class variable (1 if in that level, 0 otherwise), etc.

GLMs are based on distributions in the exponential family, which produce loglikelihoods and normal equations of a particularly simple form. Letting x_i represent a column of the model design matrix, the normal equation associated with that x_i is:

$$0 = \sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j) g'(\mu_j)} x_{ji} \quad (2.1)$$

The sum is over all observations j. w is weight. y is observed value of the dependent variable. μ is fitted value in the original scale of the observation y and relates to the linear predictor via the link function g, where g(μ) equals the linear predictor, linear in x_i and the other regressors. g'(μ) is the first derivative of the link function evaluated at the fitted value. V(μ) is the so-called variance function associated with the distribution being assumed for this particular GLM. It expresses variance of the individual observations y as a function of their expected values. There is a variance function associated with each distribution in the exponential family. For more on GLMs and their associated loglikelihoods and normal equations, one could probably consult any standard text on Generalized Linear Models, but my personal favorite is chapter 2 of McCullagh and Nelder [1].

One can think of the above normal equation (2.1) as an invariance, because one could express this sum as a difference of two sums, one being a weighted mean y, the other being a weighted mean μ. And the normal equation says that these two sums are equal, in other words, here is a value preserved by the fitting process, the same whether we plug observed values or fitted values into it.

Now assume that our GLM includes an intercept term and consider the normal equation associated with the column of X that is all ones, the column representing the intercept term. Then the sum remains over all observations but x_i appears to drop out, because the x_{ji} are all identically one. Next, assume the model includes a classification (as opposed to continuous regressor) main effect where the classification has L levels. This class variable is encoded into the design matrix X via the inclusion of columns for indicator variables for membership in each of the class levels. (For a technical correction to this last statement and how it impacts the following argument, see the appendix.) Let x_i be one of those columns. x_i has elements equal to 1 if the observation is in that level and 0 otherwise. So the above sum becomes a sum over just the observations in that one level, with x_i again appearing to drop out, because it is identically 1 in that level and 0 elsewhere. In all the cases discussed above in this paragraph, the normal equations associated with these variables become:

$$0 = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{V(\mu_j) g'(\mu_j)} \quad (2.2)$$

The sum is over a subset of the data, either all the data or just the observations in one level of a class variable appearing as a main effect in the model. The variable x_i appears to have disappeared, but not really; it was just an indicator variable that selected out the subset of data in the sum.

Now suppose that one has made a particularly judicious choice of link function relative to the assumed distribution for the dependent variable such that $V(\mu)g'(\mu)=1$. (Although, admittedly, this is not how one chooses a link. Rather one chooses a link based on some combination of a priori reasoning and empirical evidence that under that link $g(\mu)$ becomes at least approximately linear in the explanatory variables.) Then:

$$0 = \sum_{j \in sub} w_j (y_j - \mu_j) \quad (2.3)$$

or, equivalently:

$$\langle y \rangle_{sub,w} = \langle \mu \rangle_{sub,w} \quad (2.4)$$

In other words, mean observed y is equal to mean fitted μ, where the mean is taken over a certain subset of the data weighting on w. This is the balance we sought.

When is it the case that $V(\mu)g'(\mu)=1$? For classical linear regression, the dependent variable is assumed normally distributed ($V(\mu)=1$), and the identity link is assumed, $g(\mu)=\mu$, hence $g'(\mu)=1$. Hence the condition is indeed satisfied, and we get our classical balance. The condition is also satisfied for the following distribution/ link function pairs:

- Poisson count models with Poisson distributed dependent variables ($V(\mu)=\mu$) and log link ($g(\mu)=\ln(\mu)$, where ln is the natural logarithm).
- Logistic regression with binomially distributed dependent variables ($V(\mu)=\mu(1-\mu)$) and logit link ($g(\mu)=\ln(\mu/(1-\mu))$.
- GLMs with gamma distributed dependent variable ($V(\mu)=\mu^2$) and reciprocal link ($g(\mu)=1/\mu$).

For more on distributions, their associated variance functions, and link functions, again see reference [1]. Two important comments at this point: First, from the above, we can expect our Poisson count models and logistic regressions to preserve classical balance. Second, although the above gamma model will also preserve classical balance, it is not usually the case that we will expect reciprocal expectations to be approximately linear in explanatory variables. So, although it would be convenient for purposes of preserving classical balance to adopt a reciprocal link for gamma models, we probably will not usually, because it will probably not preserve linear response. Whether the

chosen link preserves linear response can be tested via various GLM diagnostics, including diagnostic plots.

It is common to build gamma distribution models for severity, or pure premium (loss divided by exposure), or loss ratio (loss divided by premium), but assuming a log link to yield a multiplicative model. What are the normal equations, and what GLM invariants are preserved by such gamma models ($V(\mu)=\mu^2$) with log link ($g(\mu)=\ln(\mu)$)?

$$0 = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{V(\mu_j) g'(\mu_j)} = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{\mu_j} \quad (2.5)$$

or, equivalently:

$$\langle \frac{y}{\mu} \rangle_{sub,w} = 1 \quad (2.6)$$

Recall that $\langle y/\mu \rangle$ need not equal $\langle y \rangle / \langle \mu \rangle$, depending on the distributions of y and μ. So this identity may not be the classical balance we were hoping for, close perhaps, but not exact. What this says is that the w weighted mean of the ratio y/μ over all the data or over any individual level of any class effect appearing as a main effect in the model equals 1; the GLM fitting algorithm forces these constraints. If, in a given subset of data, there are a few significantly peculiar values of either y or μ, the mean ratio would still be constrained to be 1, but the ratio of means $\langle y \rangle / \langle \mu \rangle$ might be significantly distorted from 1 by the fitting algorithm's attempts to satisfy the constraints. It might prove interesting and possibly an important model diagnostic to drill in to see which individual observations were the greatest source of that discrepancy.

For many of the same situations for which it is common to build gamma models, it is not uncommon to also consider the alternative of a Tweedie distribution ($V(\mu)=\mu^p$) model with log link ($g(\mu)=\ln(\mu)$). p is frequently between 1 and 2, quite often 1.67. (I have heard this more than once at Predictive Modeling and RPM Seminars, but I don't have a reference. My apologies.) A Tweedie with p between 1 and 2 is a compound distribution with Poisson count process and gamma severity. As p tends down to 1 from above, the Tweedie tends towards a pure Poisson process. As p tends up to 2 from below, the Tweedie tends towards a pure gamma severity process. The relevant normal equations are:

$$0 = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{V(\mu_j) g'(\mu_j)} = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{\mu_j^{(p-1)}} \quad (2.7)$$

, which imply yet another set of invariants, but close to those we have already considered.

3. OFF-BALANCE

As shown in the previous section, for many distribution/ link function combinations the normal equations imply an exact balance between <y> and <μ>. For other distribution/ link combinations, even when the GLM invariants implied by the normal equations may differ from exact balance, the normal equations may provide some insight into the cause, the sign, and the magnitude of the off-balance.

The case of a gamma model with log link is particularly interesting. Suppose we measure off-balance by the quantity <y>/<μ>-1, where <> denotes means over subsets weighting on w, but I have suppressed the sub,w subscripts of the previous section. Then:

$$\frac{\langle y\rangle}{\langle\mu\rangle}-1=\frac{\langle y\rangle}{\langle\mu\rangle}-\frac{\langle\mu\rangle}{\langle\mu\rangle}=\langle\frac{y}{\langle\mu\rangle}-\frac{\mu}{\langle\mu\rangle}\rangle=\langle\frac{y}{\mu}\frac{\mu}{\langle\mu\rangle}-\frac{\mu}{\langle\mu\rangle}\rangle=\langle\left(\frac{y}{\mu}-1\right)\frac{\mu}{\langle\mu\rangle}\rangle$$
$$=\langle\left(\frac{y}{\mu}-1\right)\left(\frac{\mu}{\langle\mu\rangle}-1\right)+\left(\frac{y}{\mu}-1\right)\rangle=Cov\left[\frac{y}{\mu},\frac{\mu}{\langle\mu\rangle}\right]\quad(3.1)$$

On the second line, the second term in the mean vanishes because <y/μ> equals 1 by equation (2.6), and the first term is recognizable as a covariance because the means of both y/μ and μ/<μ> are 1. Equation (2.6) drives this derivation and is therefore the point of contact in trying to explain off-balance via the normal equations.

This equating of an off-balance to a covariance is interesting. Suppose that for a particular subset of the data (either the whole of the data or a particular level of a particular classification variable in the model) the model is off-balance on the low side, in other words, <μ> less than <y>. Then that covariance is positive. Then, when μ/<μ> is on the high side of its mean 1, so is the ratio y/μ, on average. Likewise, on average y/μ is less than its mean 1 when μ/<μ> is. These observations taken together imply that y grows faster than μ on average in order to yield this behavior of the ratio y/μ, in other words, there is something about the model such that μ is tempered relative to y. On the other hand, when the model (for a particular subset) is off-balance on the high side, in other words, <μ> greater than <y>, then the relevant covariance is negative, and the above argument cuts the opposite way to imply that there is something about the model that causes μ to be somewhat over-responsive and to grow faster than y on average.

It is also possible, however, that the covariance tells us little about the behavior of the ratio y/μ on average but indicates only that there are some very anomalous values of either y or μ that throw off both the covariance and the approximate balance between <μ> and <y>. This alternative possibility was already noted in the commentary following equation (2.6). More research is probably

needed to clarify what the covariance result (3.1) is telling us, but it is certainly interesting and suggestive.

While on the subject of off-balance, there is a particularly useful scatterplot for flagging those levels of those classification variables that are more off-balance than one might expect, those levels most in need of investigation and explanation. Each data point in this scatterplot represents one level of one classification variable, and every level of every classification variable in the model has a representative point somewhere in the plot. On the y-axis we plot the ratio $\langle y \rangle / \langle \mu \rangle$, the means taken over the observations in this level of this classification variable. We draw a horizontal reference line at 1 to draw attention to deviations from 1. On the x-axis we plot class level aggregate exposures (on a log scale because these aggregate exposures might vary over a few orders of magnitude). The reason for this x-axis is that, with decreasing aggregate exposures in the level, we expect increasing scatter of the ratio $\langle y \rangle / \langle \mu \rangle$ about its hoped-for value of 1, just due to random fluctuation. We label each point according to the classification variable and level it represents, identify those variables and levels that appear to stand out from the overall pattern, and drill further into them to see if we can understand the greater than expected degree of off-balance.

It was considering just such a scatterplot that led to the investigations that led to equation (3.1) above. Certain extreme levels of one particular classification variable were flagged as excessively off-balance, and, in hindsight, taking into account known issues with the data and the model, it became clearer how under- or over-responsiveness of μ to y (in those particular levels) was responsible for that excessive off-balance.

4. IMPLICATIONS FOR WEIGHTS

Many models are built on dependent variables that are a ratio of aggregate loss to something else. If the denominator is aggregate exposure, the ratio is pure premium; if premium, loss ratio. These models are frequently built not on data at the level of individual risk but rather on data aggregated into cells defined as crossings on all the classification rating variables. The volume of business in these cells can frequently vary by orders of magnitude from one cell to the next, so some form of weighting will be needed, as the dependent variable ratios will tend to be far more volatile in low volume cells than in high. In actuarial circles it is generally assumed that a large volume cell can be treated as a sum of "independent" smaller cells, leading to variance of the dependent variable proportional to the reciprocal of some measure of business volume, which implies weights varying as some measure of business volume. But which measure? Common actuarial intuition and practice would argue for using the quantity already in the denominator of the ratio as the weight as well. It is

useful to see how the normal equations and GLM invariants of this paper bear out this choice.

I have generally found in the past, when I have fit models using each of several candidate volume measures for weights and then examined various residual plots hoping to see in those plots a signature that one particular choice of weight "outperformed" the rest, that rarely, if ever, did the plots clearly indicate one weighting scheme over the others. Some other extra-statistical, actuarial criterion has had to be imposed in order to select one weight over the rest. The normal equations can provide some guidance.

Consider first the case that our choice of distribution/ link function combination is such that equations (2.3) and (2.4) hold. y is a ratio of loss to some denominator, L/D. μ, the fitted ratio, can similarly be thought of as a ratio of fitted loss to the same denominator, $\hat{L}$/D. In effect, we define the fitted loss $\hat{L}$ as the product μD. If we select the weights w equal to the D, then in equations (2.3) and (2.4) the w and D cancel each other, and these equations say simply that aggregate fitted losses are equal to (in balance with) aggregate observed losses. The equality between aggregate observed and fitted losses is not a statistical necessity (Nothing in the statistical diagnostics argues against a choice of weights other than the D from among a number of reasonable measures of business volume, but only for the w equal to the D do we achieve balance), but it seems a reasonable extra-statistical, actuarial constraint to impose as a means of rationally selecting one weighting scheme over others. Then the aggregate fitted ratio, being the ratio of aggregate fitted losses to aggregate D, equals the aggregate observed ratio, being the ratio of aggregate observed losses to aggregate D. Choice of weights w other than the denominators D in equations (2.3) and (2.4) would result in other "weighted mean fitted ratios" in balance with their corresponding "weighted mean observed ratios", but the interpretation of those "weighted mean ratios" would be far more strained than the interpretation of the more natural weighted mean ratios when w equals D. This is the gist of the usual actuarial intuition regarding weights.

Consider next the Tweedie distribution/ log link normal equations (2.7). If we again select the w equal to the D, w and D again cancel one another, we again have L-$\hat{L}$, but now divided by $\mu^{(p-1)}$:

$$0 = \sum_{j \in sub} \frac{L_j - \hat{L}_j}{\mu_j^{(p-1)}} \quad (4.1)$$

If this denominator in μ were a constant across the sum, we would again have aggregate fitted loss in balance with aggregate observed loss, but it is not constant. How non-constant is it, because, if close to constant, perhaps aggregate fitted loss and aggregate observed loss may not be far out of balance? First, across much of the data, the range of μ may be relatively modest. Second, in those applications of Tweedie I have seen, p is rarely less than 1.5 or greater than 1.67, so the power of μ

is something like 1/2 or 2/3, which further tempers the range of values in the denominator and so brings aggregate fitted and observed losses closer to in balance. In Tweedie/ log link models, I have seen aggregate fitted and observed losses in balance to within a few percent of one another when w is selected equal to D, whereas out of balance by as much as a few tens of percent when another weighting scheme is selected.

Lastly, considering the gamma/ log link model yielding equations (2.5) and (2.6), these look like the Tweedie/ log link case of equation (2.7) but with a p of 2. Hence the exponent on μ in the denominator is 1 rather than the 1/2 to 2/3 of the Tweedie case, there is less tempering of the range of the power of μ, and aggregate fitted losses and aggregate observed losses can be more out of balance than in the Tweedie case, even when we select the w equal to the D.

5. SUMMARY AND CONCLUSIONS

Many of us are familiar with the balance between weighted mean fitted values and weighted mean observed values in standard linear regression settings. These same balance conditions extend to many GLM settings with various combinations of assumed distribution of the dependent variable and link function. The source of these balance conditions are the so-called "GLM Normal Equations". Even for those GLMs with distribution/ link function combinations not preserving the usual balance conditions, there is always another GLM invariance implied by the normal equations. The normal equations can also help us to understand the direction and degree of off-balance when off-balance exists as well as understand the consequences to balance or off-balance when choosing a weighting scheme for our weighted GLMs.

Appendix: A Technical Refinement of the Balance Equations Argument for Model Design Matrices of Full Rank

In section 2 of this paper, the argument leading from the normal equations to the balance equations assumed that, for each level of each classification variable appearing in the model as a main effect, the model design matrix included a column equal to the indicator variable for that level of that classification variable. For technical reasons this may not be quite true, and the argument requires a technical refinement, but the conclusion re the balance equations is still true.

What could be wrong with the original argument? Suppose the design matrix includes a column of ones, representing the intercept, and columns for indicator variables for each and every level of a classification variable. Because the sum of all these indicator variable columns reproduces the column of one's (because every observation is in precisely one level), there is a linear dependence

among the design matrix columns, the design matrix is less than full rank, and its inverse is not uniquely defined, which creates problems solving for the regression coefficients. To resolve this issue, many stat packages arbitrarily select one of the levels of the classification variable to serve as the reference level for that classification variable, remove the indicator variable for that level from the design matrix, and peg the regression coefficient and standard error estimate for that level to zero. The resulting reduced design matrix is now full rank and invertible, at the cost of having arbitrarily selected a reference level and removed that indicator variable from the design matrix.

Because the design matrix still includes the column of all ones, the argument of section 2 of this paper establishes that overall balance still holds. Also because of the section 2 argument, balance still holds in each level of the classification variable for which there is still an indicator variable in the design matrix, but does balance still hold for the reference level, given that there is now no indicator variable for that level in the design matrix? One would think so, given overall balance and balance in every other level, that this would imply balance in the reference level as well, and one would be right, for the following reason. Because equation (2.1) holds for each column of the design matrix, it also holds for all linear combinations of those columns:

$$\sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} \sum_i \alpha_i x_{ji} = \sum_i \alpha_i \sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} x_{ji} = \sum_i \alpha_i * 0 = 0 \quad \text{(A.1)}$$

where j indexes the observations (the rows of the design matrix), i indexes the variables (the columns of the design matrix), the order of summation can be reversed, and the inner sum of the second expression is zero by equation (2.1). The indicator variable for the reference level of the classification variable in question can indeed be expressed as a linear combination of the other columns of the design matrix, which is why it was declared to be a reference level in the first place. So balance holds for this reference level as well. QED. In fact, equation (A.1) establishes that a normal equation holds not just for any column of the design matrix but also for any other variable that can be expressed as a linear combination of the columns of the design matrix.

6. REFERENCES

[1] McCullagh, P., and J. A. Nelder, Generalized Linear Models, Second Edition, Chapman and Hall/CRC, 1989.

Author's Biography

Fred Klinker (FCAS, MAAA) is an actuarial manager with ISO's Modeling Division. He is involved with various efforts to introduce modern predictive modeling techniques, especially GLMs, into standard ISO ratemaking. Much of his career prior to ISO was spent as a research actuary involved in financial and statistical modeling at CIGNA P&C, with brief forays into regulation at the Massachusetts Insurance Department and reinsurance at Zurich Re North America, later Converium.