Estimates of the Error in GLM Coefficients, Understanding The Sources of the Errors, and Some Ideas for Troubleshooting the Errors

Presentation to 2021 CASE Seminar

by

Joseph Boor, FCAS, Ph.D., CERA

Retired, consulting occasionally

joeboor@comcast.net

850-766-6322

Antitrust Notice from CAS

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding expressed or implied that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Reason for Analysis

When I reviewed GLMs as a regulator I sometimes saw

- Inconsistent relationships among coefficients for related rating criteria
  - e.g factor for 2 accidents is lower than factor for a single accident
- Negative lift: Model doesn't improve accuracy—reduces it
- Poor performance of some rating values on sequential F test

Regulatory Issues

Previous slide has what are really business issues, pure regulatory concerns could be

- Coefficients create rating factors. Regulators and other constituents need to know they are not just random.
- Most insurance laws say rates should not be excessive, inadequate, or unfairly discriminatory—Problem coefficients touch all three
- Generally, "not arbitrary' is preferred
- Social issues are outside the scope of this presentation

Contrast — Present View of Many GLM Practitioners

- Most GLM practitioners happy as long their model predicts the dependent variable.
  - No focus on the coefficients other than as step along the way to the model.

Goals for Discussion

- Straightforward estimates of error in each coefficient
- Detailed formula for error of whole set of coefficients (root expected sum of squares)
  - Splits error drivers between statistical limits of data vs. structure of rating variables.
- Suggestions for identifying real problem and what do about it.

As Promised- Easy Computation of Variances (SD's) of Coefficients

- Split the data randomly into 5 equal parts
  - "Random" is important
- Create separate GLMs for each of the 5 datasets
- Final coefficients are average of values from 5 GLMs. Error Variance is sample variance of 5 estimates $\div$ 5

Setup of the Core Linear Model Within a GLM

- Will use vector $\boldsymbol{X}$ (Using bold for vectors and matrices) of predictor variables X_1, X_2, ... X_p (using uppercase for individual variables that could be random) to predict the "dependent" random variable Y (loss ratio, frequency, etc.) with linear formula using X's. I.e., want β coefficients so that est $(Y) = \beta_1 \times X_1 + \beta_2 \times X_2 + ...\beta_p \times X_p$
  - Will plug in different values of $\boldsymbol{X}$ for different risks with different characteristics–to predict each one's y.
- Have a "training dataset" consisting of "n" joint simultaneous observations of the predictor variables $\boldsymbol{X}$ (the set of X's) and Y that we use to estimate the β's

Setup of the Linear Model Continued

In a world of complete knowledge and infinite computer precision, the vector of coefficients $\boldsymbol{\beta}$ is per the matrix equation.

$$\boldsymbol{\beta} = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] & \cdots & Cov[X_1, X_p] \\ Cov[X_2, X_1] & Var[X_2] & \cdots & Cov[X_2, X_p] \\ \ldots & \ldots & \ldots & \ldots \\ Cov[X_p, X_1] & Cov[X_p, X_2] & \cdots & Var[X_p] \end{bmatrix}^{-1} \times \begin{bmatrix} Cov[X_1, Y] \\ Cov[X_2, Y] \\ \ldots \\ Cov[X_p, Y] \end{bmatrix}. \tag{1}$$

Prediction is $\boldsymbol{\beta}^T$ times the $\boldsymbol{X}$ vector for an insured. Conceptually $Var[X_1]$, say, is the variance of X_1 within the general target population of this type of insured, but it is estimated using the "n" records in the training dataset. Similarly for the other values.

Restatement of the Linear Model

In a world of complete knowledge and infinite computer precision, the vector of coefficients $\boldsymbol{\beta}$ is determined by solving a matrix equation, or symbolically,

$$\boldsymbol{\beta} = \boldsymbol{V}^{-1} \times \boldsymbol{C} \ . \tag{2}$$

Sources of Error: What Happens When the Coefficients are Computed

- Computer arithmetic is imperfect.

- The data has statistical limitations (possible limited credibility)

  - Especially when high CV/high volatility data such as loss ratios or severity is to be predicted.

Errors and the Linear Model

- The world that the pure model came from is not the world we live in. In our world the actual β's result from
-

$$\boldsymbol{\beta} + \boldsymbol{\tau} \;= [\boldsymbol{V}+d\boldsymbol{\Delta}]^{-1} \times [\boldsymbol{C}+d\boldsymbol{\epsilon}], \tag{3}$$

- Note that in our world the error in computing $\boldsymbol{V}$, $d\boldsymbol{\Delta}$, and the error in computing the covariance vector with Y, $d\boldsymbol{\epsilon}$ cause error in the final estimate of $\boldsymbol{\beta}$. That error is $\boldsymbol{\tau}$.
- d is used because $\boldsymbol{\Delta}$ and $\boldsymbol{\epsilon}$ are fixed across sample sizes, and the error in approximating $\boldsymbol{V}$ and $\boldsymbol{C}$ have the same relationship to the sample size n. d represents this impact

The Linear Model that Actually Happens

$$\boldsymbol{\beta} = \begin{bmatrix} Var[x_1] & Cov[x_1, x_2] & \cdots & Cov[x_1, x_p] \\ Cov[x_2, x_1] & Var[x_2] & \cdots & Cov[x_2, x_p] \\ \ldots & \ldots & \ldots & \ldots \\ Cov[x_p, x_1] & Cov[x_p, x_2] & \cdots & Var[x_p] \end{bmatrix}^{-1} \times \begin{bmatrix} Cov[x_1, y] \\ Cov[x_2, y] \\ \ldots \\ Cov[x_p, y] \end{bmatrix}. \tag{4}$$

where each variance or covariance is the sample variance or covariance across the n samples/observations in the training dataset.

Considerations About the Error in $d\boldsymbol{\Delta}$and $d\boldsymbol{\epsilon}$

- Three considerations
  - Computer arithmetic
  - Standard deviation of each estimator
    - $\sum_{k=1}^{n}\{(x_i - \mu_i)(x_j - \mu_j) - Cov(X_i, X_j\})$ (similarly for Y) used in estimating $\boldsymbol{C}$.
  - Error reduction through sampling all the "n" observations in training data.

How Bad Can Computer Arithmetic Be?

- Standard double precision arithmetic relative error a little more than $1{\times}10^{-16}$.
- Multiplying typically creates minor errors, but adding smaller number to a sum generally creates more meaningful relative error. More additions=more error
- Typically truncation when adding smaller number to a sum is about $n/2$ (midway in sum) times $1{\times}10^{-16}$.
- Overall error has approximate size n, n additions gives relative error of $n/2 \times 10^{-16}$, about 6-7 good digits when adding a billion observations.

The Impact of the Standard Deviation of Error in the Estimates of $\boldsymbol{V}$ and $\boldsymbol{C}$

Covariance estimate is the average of a number of “sample calculations” $(x_j - \mu_{X_j})(y - \mu_Y)$ of the covariance (with overall means, not those of the individual records)

$$Var[\text{estimate of } Cov[X_j, Y]] = \frac{Var[(X_j - \mu_{X_j})(Y - \mu_Y)]}{n}, \quad (5)$$

- The fact that the means are also estimated might mean $n-1, n-2,$ or $n-3$ should be in the denominator, but the numbers are usually large so the difference from n is not material.

Standard Deviation of Error in Estimation of $\boldsymbol{C}$ Due to Randomness—Part 2

- Can estimate the error in the entries in $\boldsymbol{V}$ and $\boldsymbol{V}$ with the sample variance of the "sample calculations" $(x_j - \mu_{X_j})(y - \mu_Y)$ for each entry, across all the records in the training dataset, divided by n.

- Relative error analogue = CV. With a very low underlying CV of .10, a billion samples would have 4-5 good digits. Ignoring computer error henceforth to focus on sample size-induced error.

Estimating the Variances of the ϵ_j's and $\Delta_{i,j}$'s

- Appears to work when means are determined from data.

- Compute the quantity on previous slide (the covariance error of each ϵ_j and $\Delta_{i,j}$) in each sample.

- Sum and divide by, maybe by n, take square root for standard deviation.

- All error terms in this case have a mean of zero.

Total Relative Error in Estimating $\boldsymbol{V}$ and $\boldsymbol{C}$

- Quick answer from numerical analysis is $\frac{||\boldsymbol{\Delta}||}{||\boldsymbol{V}||}$ and $\frac{||\boldsymbol{\epsilon}||}{||\boldsymbol{C}||}$.
  - "||...||" is the 2-norm, square root of the sum of squares.

- Problem: We don't know what values $\boldsymbol{\Delta}$ and $\boldsymbol{\epsilon}$ take. It's random. But this formula applies to all $\boldsymbol{\epsilon}$'s

- Solution: Use RSES: square Root of the Sum of <u>Expected</u> Squares for the "norm" ||...||. Now can mix random and constant components.

Total Relative Error in Estimating $\boldsymbol{V}$ and $\boldsymbol{V}$: The Formula

- Again, use $t(X_j, Y) = [(X_j - \mu_{X_j}) \times (Y - \mu_Y) - Cov(X_j, Y)]^2$ representing the squared error one data point makes in approximating the covariance. This is based on the "sample calculation" earlier.

- Then, up to whether "n" is the exact correct value

$$||\boldsymbol{\Delta}|| = RSES(\boldsymbol{\Delta}) = \frac{\sqrt{\sum_{i=1,j=1}^{p} Var[t(X_i, X_j)]}}{\sqrt{n}}, \tag{6}$$

$$||\boldsymbol{\epsilon}|| = RSES(\boldsymbol{\epsilon}) = \frac{\sqrt{\sum_{j=1}^{p} Var[t(X_j, Y)]}}{\sqrt{n}}. \tag{7}$$

- Relative to standard math, I took some liberties defining the norm of a matrix.

Conclusion on $\frac{1}{\sqrt{n-2}}||\boldsymbol{\Delta V}^{-2}||$

Since t is a random variable, we can estimate the variance of the average across n observations and get a standard deviation for the total relative error

$$E[||\boldsymbol{\tau}||] \leq \frac{1}{\sqrt{n}} cond(\boldsymbol{V}) \sqrt{\frac{\sum_{j=1}^{p} E[t(X_j, Y)]}{||\boldsymbol{C}||^2} + \frac{\sum_{i,j=1}^{p} E[t(X_j, X_j)]}{||\boldsymbol{V}||^2}}.$$

What's This $cond$ Stuff

- You may remember "eigenvalues" or "characteristic values" from linear algebra. They are λ's that have corresponding "eigenvectors" "$\boldsymbol{X}$'s where multiplication by $\boldsymbol{V}$ magnifies $\boldsymbol{X}$ but otherwise leaves it unchanged, e.g. $\boldsymbol{V} \times \boldsymbol{X} = \lambda \boldsymbol{X}$.

- Underlying the inequality on the last slide is an analysis by James Wilkinson that showed that (norm of) the error propagated through solving a matrix equation was capped at the absolute value of the "condition number" ($cond(\boldsymbol{V})$, or the ratio of the highest to lowest eigenvalue) times the norm of the error entering the process.

Error in Estimating $\boldsymbol{V}$ vs. that in Estimating $\boldsymbol{C}$

- $\boldsymbol{V}$ would be based on products of census or coding-type variables.
- $\boldsymbol{C}$ is based on products of those variables with pure premium, loss ratios, frequency or severity.
- Except for frequency and very large accounts, all those are highly volatile and highly skewed. One would expect standard deviations of items in $\boldsymbol{\epsilon}$ to be larger than those of items in $\boldsymbol{\Delta}$.

What Can I Do With This?

- If there are questions about the coefficients:
  - Is it my data structure or do I have too few records?
  - Need to look at it in light of the error equation and size of condition number (say 10,000 for 10-15 variables?).
  - Probably actually easier to get condition number, given some software. Almost all software lets you see $\boldsymbol{V}$, freeware computes eigenvalues or condition number.
  - If it is not obviously the condition number, suggest computing the error in estimating $\boldsymbol{V}$ and $\boldsymbol{C}$

Fixing Overlapping Rating Variable Structure/Condition Number

- May be able to use somewhat different variables that cover the same ground.
- E.g. replace number of traffic tickets in territory + number of accidents with number of traffic tickets + % of cars in city with high hp/mass ratio
- Consider pruning variables
  - LASSO
  - Rating variables that most closely match eigenvector that goes with highest eigenvalue. Seems to better avoid throwing out a variable that would useful after throwing out the next two or so.

Fixing Number of Records Problems

- Is there another dataset I an use?.
- Can I modify the data to reduce the variance? Maybe cap the claim sizes or use frequency only.
- Note that reducing the condition number will give you more "room".

Last Step-Effect of Inverse Link

- Often rating variables in context are not linear-e.g., multiplicative rating factors.
- E.g. for multiplicative rating equation, take log (the link function)of rating equation, solve the linear problem, the apply inverse function of link function (inverse link) to linear model.
- Final relative error under log link is additive relative error $\times$ derivative of exponential of factor $\times$ value in additive of additive factor $\div$ final log link factor.

???