Insurance Programs
and Analytic Services

Using Collateral Information to Blend Predictions

John Baldan, ISO
March 10, 2015

Examples of Blending

Example #1

A company wants to blend its own experience with information provided by a third-party vendor.

Example: a company has personal auto bodily injury loss costs by territory, and wishes to blend them with third-party modeled loss costs by zip code.

Examples of Blending

Example #2

A company wants to blend its own statewide and countrywide reviews.

Example: a company has developed classification relativities on a single state's data, and wants to blend in relativities derived in a previously performed countrywide analysis.

Future Research

We intend to use the methodology discussed at this session to perform the analysis described above in Example #1.

Will share results either at a future CAS meeting or via publication, or both.

Setting up the GLM

Using the company's policy-level data, use a GLM to predict pure premium.

- Regressor: company manual loss cost, including both geographic and non-geographic rating factors.
- Also include third-party zip loss cost as either an offset or a regressor.
- Model form: log link, Tweedie distribution
- Weight: earned exposure
- SAS PROC HPTWEEDIE also estimates optimal Tweedie "power" parameter p. (Single-company estimates I've seen range from 1.41 to 1.67.)

Calculations

Perform the calculations as outlined in the Frees and Shi presentation, slides 14-15.

The key idea is to determine the mean of the posterior Bayesian distribution for each zip code relativity α_j. Use the paper's assumption of a Tweedie-distributed claims distribution and corresponding conjugate prior distribution.

Bayesian "belief" parameter - ϕ_α

The one free parameter in the model specifies the judgmental belief in the third-party's zip loss costs.

- $\Phi_\alpha = 0 \rightarrow$ rely completely on agency prediction
- $\Phi_\alpha = \infty \rightarrow$ rely completely on company's own data

The blending happens here:

$$E(\alpha_j \mid \text{data}) = \frac{\Phi}{\Phi+\Phi_\alpha * f(\mu_i, p)} + \left(1 - \frac{\Phi}{\Phi+\Phi_\alpha * f(\mu_i, p)}\right) * g(y_i, \mu_i, p)$$

How to select ϕ_α?

How do you know how much to believe the third-party data?

Take the “Empirical Bayesian” viewpoint: split the company data into training and holdout, score the holdout data, then select Φ_α that optimizes lift on the holdout data.

How to select ϕ_α?

- Score the holdout data:

  $E(\alpha_j|\text{training data}) * \text{ZIP loss cost} *$

  company non-geographic rating factor

Compare Gini indices measured on the holdout data:

- Prediction using ZIP loss cost offset, optimal $\phi\alpha$
- Prediction using ZIP loss cost regressor, optimal ϕ_α
- Company manual loss cost