Credibility Prediction using Collateral Information

Edward W. (Jed) Frees, Peng Shi

University of Wisconsin – Madison

March, 2015

Outline

Acknowledgements

- Verisk Analytics provided financial support for the research.
- Our thanks also go to Dave Cummings and John Baldan who actively participated in the discussion during the preparation of the manuscript underpinning this presentation.

Rates from Characteristics and Scores

- The insurer needs to develop rates
  - However, the insurer has limited internal data
  - Either a new line of business, a small market, or a small insurer
- Information available
  - Some (internal) policy characteristics (xs) and claims (ys) upon which rating predictors can be developed
  - An external agency provides one or more scores based on the characteristics in the insurer's portfolio
- The problem
  - decide when to use internal versus external information
  - combine these sets of information efficiently

Using GLM to Incorporate Scores

- Incorporate scores using GLM. Your choices:
  1. use only internal characteristics as covariates
  2. use only the external score
  3. use the external score as an offset, also include internal characteristics as covariates
  4. use internal characteristics and the external score as covariates
  5. combine internal characteristics with external scores, yet allow the actuary to incorporate his/her professional judgement about the relative importance of the external score
- Number (5) introduces our proposal. We want to use the external score as an offset, include internal characteristics as covariates and incorporate a way to allow the actuary to modify results in a disciplined way.

Our Proposal

- The recommended rate formula is

$$Rate = ExtScore \times \exp\left(\mathbf{x}_i'\beta\right) \times CredAdj$$

  - $ExtScore$ is the external score
  - $\exp\left(\mathbf{x}_i'\beta\right)$ represents adjustments from insurer policyholder characteristics $\mathbf{x}_i$, common in log-link GLMs
  - $CredAdj$ is the credibility adjustment, expressed as a weighted average of "1" (our prior) and a weighted average of the data
- Our work provides (minor theoretical) extensions to a long line of papers on Bayesian conjugate priors in exponential families
- This means that our results hold for finite (small) samples and allows the actuary to incorporate his/her judgement into the process through prior beliefs

Our Proposal - Credibility Adjustment

$$Rate = ExtScore \times \exp\left(\mathbf{x}_i'\beta\right) \times CredAdj$$

- $ExtScore$ is the external score
- $\exp\left(\mathbf{x}_i'\beta\right)$ represents insurer adjustments
- $CredAdj$ is the credibility adjustment. That is,

$$CredAdj = \zeta_j + (1-\zeta_j)\overline{(y/\mu)}_{Wj}$$

The term $\overline{(y/\mu)}_{Wj}$ is a weighted average of the data (defined later). The credibility factor is

$$\zeta_j = \frac{\phi}{\phi + \phi_\alpha W_j}$$

- ϕ is the GLM uncertainty parameter,
- ϕ_α represents the actuary's belief in the external score, and
- W_j is a weighted sum of means (all defined in more detail later)

Thinking about Uncertainty of the Score

- Let me go through a chain of reasoning to develop a way to interpret the uncertainty of a score
- Linear Model - simplicity and provide intuition
  - Sample of n observations
- **Base Case - No covariate information**

$$y_i = \tilde{\mu}_i + \varepsilon_i = \mu_{\alpha,i} + \alpha + \varepsilon_i.$$

  - $\mu_{\alpha,i}$ is the external score, varies by policy i
  - α is the uncertainty about the score. This is random, we have some knowledge (prior) about the utility of the score.
  - The external score is an estimate of the $\tilde{\mu}_i$, the true (unobserved) mean

Introducing Covariates

- **Example.**
  - Suppose that y_i representing the claim on the ith personal automobile policy
  - Covariate - whether or not the policyholder also owns a homeowners policy ($x_i = 1$ if yes and $= 0$ otherwise).
  - Incorporate this knowledge using:

$$y_i = \mu_{\alpha,i} + x_i\beta + \alpha + \varepsilon_i,$$

- In the same way, we can think about a host of covariates
  - $\mathbf{x}_i' = (x_{i1}, \ldots, x_{iK})$ represents a set of K explanatory variables
  - β is the corresponding set of parameters
  - Model

$$y_i = \mu_{\alpha,i} + \mathbf{x}_i'\beta + \alpha + \varepsilon_i.$$

Multiple Scores

- Suppose that we have two sets of collateral scores with their associated uncertainties, $\mu_{1,\alpha,i} + \alpha_1$ and $\mu_{2,\alpha,i} + \alpha_2$.
- If we are unsure how to combine them in our claims model, then it would be sensible to use (unknown) scaling factors γ_1 and γ_2.

$$\begin{aligned} y_i &= x_{i1}\beta_1 + \cdots + x_{iK}\beta_K + \gamma_1(\mu_{1,\alpha,i} + \alpha_1) + \gamma_2(\mu_{2,\alpha,i} + \alpha_2) + \varepsilon_i \\ &= x_{i1}\beta_1 + \cdots + x_{iK}\beta_K + \mu_{1,\alpha,i}\gamma_1 + \mu_{2,\alpha,i}\gamma_2 + \gamma_1\alpha_1 + \gamma_2\alpha_2 + \varepsilon_i \\ &= \tilde{\mathbf{x}}_i'\tilde{\beta} + \tilde{\alpha} + \varepsilon_i, \end{aligned}$$

  - where $\tilde{\mathbf{x}}_i' = (x_{i1}, \ldots, x_{iK}, \mu_{1,\alpha,i}, \mu_{2,\alpha,i})$ is a set of known covariates,
  - $\tilde{\beta} = (\beta_1, \ldots, \beta_K, \gamma_1, \gamma_2)'$ is a set of variables to be estimated, and
  - $\tilde{\alpha} = \gamma_1\alpha_1 + \gamma_2\alpha_2$ is a random source of uncertainty.
- incorporating multiple sets of collateral scores follows the same idea

Introducing Multiple Sources of Collateral Information

- **Example.** Personal Automobile
  - One can imagine one set of uncertainties (α_1) for retirees and another set (α_2) for all other drivers.
- In general, we will assume that there are q sets of uncertainties represented as $\alpha = (\alpha_1, \ldots, \alpha_q)'$
  - Could write this as a regression model, with an appropriately defined set of binary variables $\mathbf{z}$, as

$$y_i = \mu_{\alpha,i} + \mathbf{x}_i'\beta + \mathbf{z}_i'\alpha + \varepsilon_i.$$

  - A little cleaner to use the factor model notation and write as

$$y_{ij} = \mu_{\alpha,ij} + \mathbf{x}_{ij}'\beta + \alpha_j + \varepsilon_{ij}.$$

GLM Framework

- Claims y_i follow a generalized linear model (GLM) with conditional mean

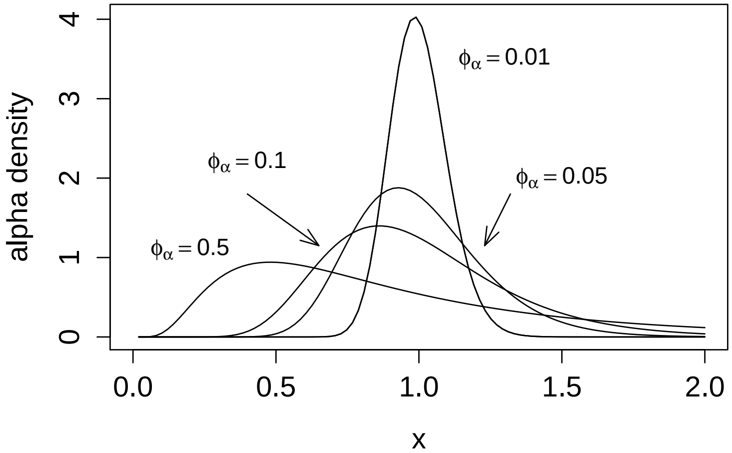
$$\mathrm{E}(y_i|\alpha) = \alpha_j \ \mu_{\alpha,i} \ \exp\left(\mathbf{x}_i'\beta\right)$$

- We continue to assume that the scoring procedure is unbiased and so $\mathrm{E}(\alpha_j) = 1$.
- Thus, the (unconditional) mean is $\mu_i = \mu_{\alpha,i} \exp\left(\mathbf{x}_i'\beta\right)$.
- Although our theory allows μ_i to be a (smooth) nonlinear function of covariates, we focus on a logarithmic link function

Prior Distribution

- For our random α could use just about any distribution and calculate posterior distributions using modern computation methods (e.g., MCMC)
- We follow work of Ohlsson and Johansson (2006) (with some mild extensions) and select a conjugate prior that has mean 1 and dispersion parameter ϕ_α.
- With this, we have closed form expressions for the posterior distribution for several exponential families including the normal, Poisson, gamma, inverse Gaussian, and Tweedie (OJ focussed on the Tweedie).
- In the Tweedie case, we have $\phi_\alpha = \mathrm{Var}(\alpha_j)/\mathrm{E}(\alpha_j)^p$.
- The next overhead gives a few illustrative prior densities.

Prior Distributions for Different Uncertainty Parameters ϕ_α



- Note that the mean is 1 for each density

Posterior Mean

- The posterior mean is

$$\mathrm{E}(\alpha_j|\mathbf{y}) = \zeta_j + (1-\zeta_j)\overline{(y/\mu)}_{Wj},$$

where ζ_j is a credibility factor

$$\zeta_j = \frac{\phi}{\phi + \phi_\alpha W_j},$$

  - that is determined by the sum of weights within the jth factor, $W_j = \sum_{i:z_{ij}=1} b_2(\mu_i)$, and
  - $\overline{(y/\mu)}_{Wj} = \sum_{i:z_{ij}=1}(y_i/\mu_i)b_2(\mu_i)/W_j$, a weighted average.
  - Here, z_{ij} is a binary variable that is one if the ith policyholder is in the jth risk category.
- The parameter ϕ and the function $b_2(\cdot)$ depend on the choice of the outcome (claims) distribution.
  - $b_2(\mu_i) = \mu_i^{2-p}$ for the Tweedie distribution.

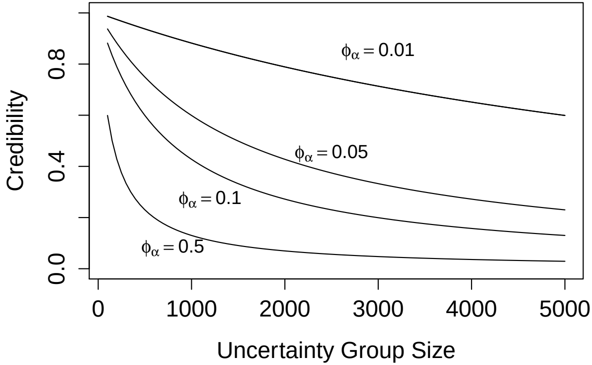
The Usual Credibility Interpretations

The credibility factor is

$$\zeta_j = \frac{\phi}{\phi + \phi_\alpha W_j},$$

- The credibility factor ζ_j tends to one as either $\phi_\alpha \to 0$ or $\phi \to \infty$.
  - In either case, we think of the uncertainty associated with the score being very (increasingly) small relative to the dispersion in the outcome distribution.
- The credibility factor ζ_j tends to zero as either $\phi \to 0$ or $W_j \to \infty$.
  - Intuitively, the credibility (of the score) is small with high precision data or as the number of observations in the jth level of the factor becomes large, indicating substantial information content in the data.

Credibility Factors by Uncertainty Parameters ϕ_α and Group Size



- This example shows how the credibility factors change with group size and the belief parameter ϕ_α

Simulating Policyholder Characteristics

- We simulated portfolios of policies and claims experience based on a sample of Massachusetts automobile experience
- The table provides the policyholder distribution by two rating factors, a driver-based rating group and territory.
- We simulated (i) an external agency's sample, to provide scores, (ii) a sample that an insurer uses to develop rates, and (iii) a sample that an insurer uses to validate its developed rates. We did this 1,000 times.

Table : Proportion of Policies by Rating Group and Territory

Rating Group	Proportion	Territory	Proportion
A – Adult	0.76616	1	0.18410
B – Business	0.01269	2	0.19360
I – Youthful with less than 3 years Experience	0.03453	3	0.11245
		4	0.20300
M – Youthful with 3-6 years Experience	0.04190	5	0.18921
		6	0.11764
S – Senior Citizens	0.14472		

Simulating Claims Distributions

- Simulate claims using the Tweedie distribution with parameters reported below.
- Use a logarithmic link function with scale parameter $p = 1.5$ and dispersion parameter $\phi = 250$.
- To illustrate, for Rating Group "A" and Territory 6,
  - the estimate is $\exp(5.356) = 211.87$
  - corresponding probability of zero claims is $\mu_i^{2-p}/(\phi(2-p)) = 89.0\%$.

Table : Tweedie GLM Coefficients

Rating Group	Estimate	Territory	Estimate
B	0.340	1	-0.743
I	1.283	2	-0.782
M	0.474	3	-0.552
S	-0.033	4	-0.480
		5	-0.269

Intercept is 5.356
Reference levels are "A" for Rating Group and "6" for Territory

External Scores

- External scores are not always great. We generated 9 types of external scores, as follows
- The best (although unattainable in practice) score is the true mean that we label `ScoreTrue`
- For other scores, assume the external agency generates scores using statistical methods. We vary their scores by:
  - sample size, either a relatively large sample size (`LS`) 100,000 or a small sample size (`SS`) 10,000,
  - number of covariates, either including both driver and territory (`Full`) or a reduced set, only driver, (`Red`), and
  - statistical methods, either a GLM using a Tweedie distribution (`GLM`) or a linear model (`LM`).

Rating Predictors

- The analyst working for the insurer has a different data set.
  - This data set may be large ($n = 10,000$) or small ($n = 2,000$)
- In addition to external scores, the analyst has different amounts of information.
  - The analyst has a full set of covariates, including both age and territory (`Full`), or
  - a reduced set, only age, (`Red`)
- The analyst may incorporate the externally provided score as an offset, use company covariates, and may modify the predictors based on the insurer's belief in the scores. For our work, we allowed the belief parameter to vary over $\phi_\alpha = 0.5, 0.1, 0.01, 0$.

Out-of-Sample Summary Measures

- Predictive Modeling Procedure
  1. Generate external agency dataset and derive their scoring mechanism
  2. Generate insurer in-sample. Use this information and external scores to develop rates for subsequent analysis
  3. Generate insurer out-of-sample. Use rates to predict claims. Compare predicted to actual held-out claims
- Out-of-Sample Summary Measures
  - Traditional measures (root mean square error and so forth) do not fare well because the data are (very) non-normal
  - The Gini index is twice the average covariance between the predicted outcome and the rank of the predictor.
  - The Gini correlation is the correlation between the predicted outcome and the rank of the predictor.

Out-of-Sample Statistics - Panel A

- The correlations and the simple Gini statistics suggest that the choice of external score matters. (approximately 5.8-6.0)
  - The external agency is most helpful when they have the right covariates, large data, and appropriate statistical methods
  - We also did an analysis assuming that the insurer had a full set of covariates. Not surprisingly, external scores added little value
- The traditional measures (e.g., root mean square error) give conflicting results.

Table : Out-of-Sample Statistics for Credibility Predictors

$n = 10,000, \quad \phi = 250$

With Company Experience Adjustment, Reduced Covariates

	Mean Absolute Error	Mean Absolute Perc Error	Root Mean Square Error	Correlations Pearson	Correlations Spearman	Correlations Gini	Simple Gini
Panel A. $\phi_\alpha = 0.0$							
LS_Full_GLM	278.920	184.216	721.825	11.031	5.400	7.837	32.680
LS_Red_GLM	279.620	184.127	723.020	9.427	3.368	5.217	21.688
LS_Full_LM	279.301	184.569	722.058	10.680	5.425	7.867	32.802
LS_Red_LM	279.690	184.200	723.084	9.317	3.378	5.224	21.715
SS_Full_GLM	279.427	184.638	722.003	10.865	5.241	7.648	31.890
SS_Red_GLM	279.620	184.127	723.020	9.427	3.368	5.217	21.688
SS_Full_LM	280.002	287.171	722.278	10.485	5.253	7.660	31.937
SS_Red_LM	279.716	216.308	723.105	9.292	3.379	5.226	21.725

- The next overhead present results as the analyst changes his or her belief about the score's precision.
  - As one increases the value of ϕ_α, one places less credibility on the score and more on the data.
- Interestingly, when even acknowledging a small imprecision in the score, the case $\phi_\alpha = 0.01$, the Gini correlation increases from approximately 4.3 to 5.4 even for scores that use only a reduced set of covariates.
- This is because we selected the risk classes to correspond to the set of information, territory, that is "missing" in both the company's covariates and external agency covariates.
  - By averaging over these risk classes in one period, the insurer has a very useful nonparametric predictor of claims in the next period.

Out-of-Sample Gini Correlations for Credibility Predictors - Small Sample

$n = 2,000, \quad \phi = 250$

With Company Experience Adjustment, Reduced Covariates

	Gini Correlation		Gini Correlation
$\phi_\alpha = 0.0$		$\phi_\alpha = 0.01$	
`ScoreTrue`	7.307		
`LS_Full_GLM`	7.224	`LS_Full_GLM`	7.049
`LS_Red_GLM`	4.345	`LS_Red_GLM`	5.423
`LS_Full_LM`	7.276	`LS_Full_LM`	7.095
`LS_Red_LM`	4.395	`LS_Red_LM`	5.467
`SS_Full_GLM`	7.036	`SS_Full_GLM`	6.889
`SS_Red_GLM`	4.345	`SS_Red_GLM`	5.423
`SS_Full_LM`	7.091	`SS_Full_LM`	6.930
`SS_Red_LM`	4.397	`SS_Red_LM`	5.468
$\phi_\alpha = 0.10$		$\phi_\alpha = 0.50$	
`LS_Full_GLM`	6.889	`LS_Full_GLM`	6.759
`LS_Red_GLM`	6.194	`LS_Red_GLM`	6.239
`LS_Full_LM`	6.959	`LS_Full_LM`	6.834
`LS_Red_LM`	6.229	`LS_Red_LM`	6.279
`SS_Full_GLM`	6.807	`SS_Full_GLM`	6.676
`SS_Red_GLM`	6.194	`SS_Red_GLM`	6.239
`SS_Full_LM`	6.843	`SS_Full_LM`	6.743
`SS_Red_LM`	6.230	`SS_Red_LM`	6.276

Legend: `LS` means large sample, `SS` means small sample
`Full` means full set of covariates, `Red` means reduce set of covariates
`GLM` means generalized linear model, `LM` means linear model

Simulation Results

- The paper (available on our web site) gives simulation results for
  - different sample sizes ($n = 2,000$ and $n = 10,000$)
  - different levels of data uncertainty ($\phi = 250$ and $\phi = 100$)
- Results are qualitatively similarly
  - An insurer does better with better quality scores
  - If you are certain of the quality and that it applies to your book of business, use this score
  - If you are uncertain, then acknowledging this uncertainty and blending the score with your own data can improve results. Not necessarily but the simulation study demonstrates situations where this does happen.

Description

- Personal automobile insurance data from the Commonwealth Automobile Reinsurers (CAR) in Massachusetts.
  - The data represent experience from several insurance carriers
  - The dataset contains records about three million policyholders in year 2006
- Information available
  - Limited number of predictors
    - Rating group: policyholder characteristics
    - Territory group: defined by garage town
  - Claims on two mandatory coverage
    - Liability: property damage and bodily injury
    - Peronal injury protection (PIP): no-fault

Predictors

- We take a sample of 100,000 policyholders for this study
  - Use half as training data to develop the model
  - Use the second half as hold-out for validation

Table : Description and Summary Statistics of Basic Rating Information

	Mean	Pure Premium Liability	Pure Premium PIP
Rating Group			
A - Adult	0.776	175.636	17.562
B - Business	0.014	147.043	11.761
I - Youth with $<$3 years experience	0.036	306.594	40.315
M - Youth with 3-6 years experience	0.040	260.558	29.586
S - Senior citizens	0.134	220.561	26.126
Territory Group			
1 - the least risky territory	0.194	115.653	9.201
2	0.201	180.577	13.171
3	0.112	162.928	11.726
4	0.204	162.645	10.935
5	0.184	227.247	21.299
6 - the most risky territory	0.106	354.751	76.009

ISO Risk Analyzer

- ISO Risk Analyzer is a suite of predictive models from Verisk Analytics
- We use two scores from the ISO Risk Analyzer Personal Auto
  - Vehicle Module for Liability (single-limit liability): built on individual vehicle attributes
  - Environmental Module: use data such as environmental, weather-related, traffic and business locations
- See details on ISO Risk Analyzer at `isoriskanalyzer.com`

Incorporating Score

- We consider two ways of using the external score
  - Relativity from the Vehicle Module for Liability is used as an offset in a GLM
  - Relativity from the Environmental Module is used as covariate
- We compare the credibility prediction with and without collateral information
- Consider three scenarios representing the complexity of the predictive models used by the insurer
  1. Rating group only
  2. Rating group + Territory group
  3. Rating group + Territory group + Environmental Module

Summary

- We analyze the Liability and PIP coverage separately
- For each type, we assume
  - The cost of insurance follows a Tweedie GLM
  - There is no uncertainty about the environmental score
  - There are multiple sets of uncertainties about the vehicle liability score according to territory group
- The following two tables present results for Liability and PIP respectively
  - Model parameters are estimated using in-sample data
  - Correlation and Gini statistics are calculated using hold-out observations
  - Allow insurer's belief in the score to vary

Liability

	Pearson	Spearman	Gini	Simple Gini
Panel A : Insurer Information				
Rating	3.389	3.047	1.931	17.361
Rating + Territory	4.833	4.883	3.423	41.328
Rating + Territory + Environmental	4.973	5.535	3.831	46.586
Panel B : $\phi_\alpha = 0$				
Rating	3.916	4.343	2.615	31.791
Rating + Territory	5.276	5.850	3.873	47.101
Rating + Territory + Environmental	5.355	6.318	4.072	49.518
Panel C : $\phi_\alpha = 0.01$				
Rating	4.909	5.702	3.768	45.828
Rating + Territory	5.223	5.839	3.899	47.414
Rating + Territory + Environmental	5.299	6.301	4.089	49.725
Panel D : $\phi_\alpha = 0.1$				
Rating	5.084	5.702	3.858	46.921
Rating + Territory	5.146	5.783	3.880	47.181
Rating + Territory + Environmental	5.224	6.243	4.061	49.390
Panel E : $\phi_\alpha = 0.5$				
Rating	5.091	5.707	3.860	46.940
Rating + Territory	5.132	5.774	3.878	47.162
Rating + Territory + Environmental	5.210	6.230	4.054	49.302

PIP

	Pearson	Spearman	Gini	Simple Gini
Panel A : Inusurer Information				
Rating	1.738	2.637	1.707	2.741
Rating + Territory	4.250	4.439	2.990	6.450
Rating + Territory + Environmental	4.797	4.475	3.075	6.680
Panel B : $\phi_\alpha = 0$				
Rating	1.812	2.872	1.975	4.291
Rating + Territory	4.217	4.524	3.040	6.605
Rating + Territory + Environmental	4.667	4.593	3.137	6.817
Panel C : $\phi_\alpha = 0.01$				
Rating	2.624	3.935	2.730	5.932
Rating + Territory	4.233	4.529	3.045	6.617
Rating + Territory + Environmental	4.677	4.598	3.141	6.824
Panel D : $\phi_\alpha = 0.1$				
Rating	4.071	4.691	3.148	6.840
Rating + Territory	4.280	4.543	3.059	6.646
Rating + Territory + Environmental	4.708	4.615	3.153	6.851
Panel E : $\phi_\alpha = 0.5$				
Rating	4.332	4.722	3.145	6.835
Rating + Territory	4.302	4.547	3.065	6.661
Rating + Territory + Environmental	4.723	4.625	3.160	6.867

Conclusion

Thank you for your kind attention.

You can learn more about our research at:

- `https://sites.google.com/a/wisc.edu/peng-shi/`
- `https://sites.google.com/a/wisc.edu/jed-frees/`