A Supplemental Material for "Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models"

Yi Yang[*], Wei Qian[†] and Hui Zou[‡]

April 20, 2016

Part A: A property of Tweedie distributions

For completeness, we give here a known result of the Tweedie distribution and its detailed proof.

Proposition 1. *Let* $Z_i = \sum_{d_i=1}^{N_i} \tilde{Z}_{d_i}$ *is the total claim amount. Let* $Y_i = Z_i/w_i$*, where* ω_i *is the duration. Assume* N_i *is Poisson distributed* $\mathrm{Pois}(\lambda_i w_i)$*. Conditional on* N_i*, assume* Z_{d_i}*'s* $(d_i = 1, \ldots, N_i)$ *are i.i.d.* $\mathrm{Gamma}(\alpha, \gamma_i)$*. Assume that under unit duration (i.e.,* $w_i = 1$*), the mean-variance relation satisfies* $Var(Y_i^*) = \phi[E(Y_i^*)]^\rho$*, where* Y_i^* *is the pure premium under unit duration,* ϕ *is a constant, and* $\rho = (\alpha+2)/(\alpha+1)$*. Then for the pure premium* Y_i *under duration* ω_i

$$Y_i \sim \mathrm{Tw}(\mu_i, \phi/w_i, \rho).$$

Proof. Note that under unit duration $w_i = 1$,

$$\mu_i^* := E(Y_i^*) = E(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i^*) = E(Var(Y_i^*|N_i)) + Var(E(Y_i^*|N_i)) = \lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2.$$

[*] McGill University

[†] Rochester Institute of Technology

[‡] Corresponding author, zoux019@umn.edu, University of Minnesota

Similarly, under any duration w_i,

$$\mu_i := E(Y_i) = \frac{1}{w_i} E(Z_i) = \lambda_i \alpha \gamma_i,$$
$$Var(Y_i) = \frac{1}{w_i^2} Var(Z_i) = (\lambda_i \alpha \gamma_i^2 + \lambda_i \alpha^2 \gamma_i^2)/w_i.$$

As a result, we can obtain the mean-variance relation for the pure premium Y_i that

$$Var(Y_i) = \frac{1}{w_i} Var(Y_i^*) = \frac{\phi}{w_i} (\mu_i^*)^\rho = \frac{\phi}{w_i} \mu_i^\rho, \tag{1}$$

where the second equation follows by

$$Var(Y_i^*) = \phi [E(Y_i^*)]^\rho. \tag{2}$$

By the scale-invariance property of Tweedie distribution, the proof is complete. □

Part B: Computational issues for profile likelihood

There are some computational issues, which must be taken care of when evaluating the log-likelihood functions in (20) and (21) of Section 4.2: since in general there are no closed forms for Tweedie densities, in likelihood evaluation one must deal with an infinite summation in the normalizing function $a(y, \phi, \rho) = \frac{1}{y} \sum_{t=1}^{\infty} W_t$. For numerical evaluation of Tweedie densities, Dunn and Smyth (2005) proposed a series expansions approach, which sums an infinite series arising from a Taylor expansion of the characteristic function. Alternatively, Dunn and Smyth (2008) developed a Fourier inversion approach, which consists of an inversion of the characteristic function based on numerical integration methods for oscillating functions. These two numerical methods turn out to be complementary since each has advantages under a certain situation: when only considering the case $1 < \rho < 2$, the series approach performs very well for small y but gradually loses computational efficiency as y increases, whereas the inversion approach performs very well for large y but gradually fails to provide accurate results as y decreases. Hence the inversion approach is preferred for large y and the series approach for small y. Dunn and Smyth (2008) provided a simple guideline to choose between the two methods. In this paper we use their R package "tweedie" (available at `http://cran.r-project.org/web/packages/tweedie/index.html`) for evaluating

Tweedie densities in our profile likelihood computation. For further details regarding their algorithms, the reader may refer to Dunn and Smyth (2005, 2008).

Part C: Bias-adjusted variable importance measure

Following Sandri and Zuccolotto (2008) and Sandri and Zuccolotto (2010), we compute the biased-adjusted VI measure for each explanatory variable in the following way:

(1) For $s = 1, \dots, S$, repeat steps (2)–(4).

(2) Generate a matrix $\mathbf{z}^s$ by randomly permutating (without replacement) the n rows of the design matrix $\mathbf{x}$, while keeping the order of columns unchanged.

(3) Create an $n \times 2p$ matrix $\tilde{\mathbf{x}}^s = [\mathbf{x}, \mathbf{z}^s]$ by binding $\mathbf{z}^s$ with $\mathbf{x}$ matrix by column.

(4) Use the data $\{y, \tilde{\mathbf{x}}^s\}$ to fit the model, and compute VI measures $\mathcal{I}^s_{X_j}$ for X_j and $\mathcal{I}^s_{Z^s_j}$ for Z^s_j, where Z^s_j (jth column of Z^s) is the pseudo-predictor corresponding to X_j.

(5) Compute the VI measure $\overline{\mathcal{I}}_{X_j}$ as the average of $\mathcal{I}^s_{X_j}$ and the baseline $\overline{\mathcal{I}}_{Z_j}$ as the average of $\mathcal{I}^s_{Z^s_j}$

$$\overline{\mathcal{I}}_{X_j} = \frac{1}{S}\sum_{s=1}^{S} \mathcal{I}^s_{X_j} \qquad \overline{\mathcal{I}}_{Z_j} = \frac{1}{S}\sum_{s=1}^{S} \mathcal{I}^s_{Z^s_j}. \tag{3}$$

(6) Report the adjusted VI measure as $\mathcal{I}^{\text{adj}}_{X_j} = \overline{\mathcal{I}}_{X_j} - \overline{\mathcal{I}}_{Z_j}$ for the variable X_j.

The basic idea of the above algorithm is the following: the permutation breaks the association between the response variable Y and each pseudo-predictor Z^s_j, but still preserves the association between Z^s_j and Z^s_k ($k \neq j$); since Z^s_j is re-shuffled from X_j, Z^s_j has the same number of possible splits as the corresponding predictor X_j and has approximately the same probability of being selected in split nodes. Therefore, $\overline{\mathcal{I}}_{Z_j}$ can be viewed as a bias approximation of the importance of X_j.

Part D: Descriptive statistics for real data

The descriptive statistics of Yip and Yau (2005) data used in Section 6 are provided in Table A1, A2 and A3.

Total Claim Amount	% obs.	% of total sum	Mean	Median
0	61.1	0	0	0
$(0, 10000]$	29.6	36.0	4902	4842
$(10000, 50000]$	9.1	61.5	27144	27679
> 50000	0.2	2.5	52157	51986

Table A1: Description of the individual total claim amount in the last five years.

	AGE	HOMEKIDS	BLUEBOOK	KIDSDRIV
Min.	16.00	0.0000	1500	0.0000
1st Qu.	39.00	0.0000	9200	0.0000
Median	45.00	0.0000	14405	0.0000
Mean	44.84	0.7199	15666	0.1694
3rd Qu.	51.00	1.0000	20900	0.0000
Max.	81.00	5.0000	69740	4.0000

	NPOLICY	RETAINED	TRAVTIME	MVR_PTS
Min.	1.000	1.000	5.00	0.000
1st Qu.	1.000	1.000	22.00	0.000
Median	1.000	4.000	33.00	1.000
Mean	1.695	5.328	33.42	1.709
3rd Qu.	2.000	7.000	44.00	3.000
Max.	9.000	25.000	142.00	13.000

Table A2: Descriptive statistics for the continuous variables in the claim history data set in Section 6.

AREA	MARRIED	REVOKED	GENDER
Rural: 20.2% Urban: 79.8%	No: 39.9% Yes: 60.1%	No: 87.8% Yes: 12.2%	F: 53.8% M: 46.2%

CAR_USE	MAX_EDUC	CAR_TYPE	JOBCLASS
Private: 63.2% Commercial: 36.8%	<High School: 14.6% Bachelors: 27.3% High School: 28.7% Masters: 20.2% PhD: 9.2%	Panel Truck: 8.3% Pickup: 17.3% Sedan: 26.2% Sports Car: 11.4% SUV: 27.9% Van: 8.9%	Blue Collar: 22.2% Clerical: 15.5% Professional: 13.6% Manager: 12.2% Lawyer: 10.0% Student: 8.7% (Other): 17.8%

Table A3: Descriptive statistics for the categorical variables in the claim history data set in Section 6.

Part E: Identifying important interactions

In this section, we demonstrate that the nonparametric approach described in this paper can serve as an important complement to the traditional GLM model in insurance rating. Even under strict circumstances that the final model must have a GLM structure, our approach can still be quite helpful due to its ability to automatically identify additional information such as important interactions. It is often challenging for a GLM approach alone to capture such information, especially if many explanatory variables are discrete (which is quite common for insurance data sets). For example, if there are eight discrete explanatory variables each with eight different values, there are $\binom{8}{2} \times 7 \times 7 = 1372$ possible two-way interaction terms. Even for data sets with millions of observations, it is in general not practical to fit simultaneously all interaction terms in a GLM model.

We continue using the real data example in Section 6. Suppose one builds a TGLM model with all main effects and applies the stepwise selection for variable selection (the p-values for entering and removal of an variable are set to be 0.05 and 0.10, respectively). The resulting model TGLM1 is showed in Table A4.

We next show that TDboost can provide insights into the structure of interaction terms, which can be subsequently integrated into TGLM1. Elith et al. (2008) proposed a relative importance measure to quantify magnitudes of fitted interaction effects. By adopting this method for TDboost, we can calculate the relative importance of two-way interactions for all possible pairs of predictors in TGLM1. Table A5 provides a summary list of 10 two-way interactions with the highest relative importance. To improve TGLM1, we then add to TGLM1 the two strongest interactions MVR_PTS:AREA and REVOKED:AREA, which account for approximately 88.33% of the total relative importance. We denote the adjusted model with interactions as TGLM2. Table A4 suggests that both interactions in TGLM2 are significant at 0.05 significance level.

To compare TGLM1 and TGLM2, we use the Gini index as the criterion. As shown in Table A6, we find that the maximal Gini index is 9.751 when using TGLM1 as the base premium, and -2.172 when using TGLM2. Therefore, TGLM2 is more favorable than TGLM1. We also compare the TGLM2 model against the TGLM1 model using the likelihood ratio test and get the same conclusion ($\chi^2 = 371.79$, $df = 2$, $p \approx 0$). Therefore, with the help of TDboost, the overall model performance is improved under a GLM model structure.

Variable	TGLM1		TGLM2	
	Estimate	Std.Error	Estimate	Std.Error
Intercept	-2.93**	0.20	-4.61**	0.54
KIDSDRIV	0.10**	0.04	0.10**	0.05
REVOKED	1.54**	0.06	2.47**	0.44
MVR_PTS	0.20**	0.01	0.58**	0.07
MARRIED	-0.17**	0.04	-0.17**	0.05
AREA	1.22**	0.07	2.12**	0.27
CAR_TYPE_2	-0.08	0.10	-0.07	0.11
CAR_TYPE_3	-0.07	0.10	-0.07	0.11
CAR_TYPE_4	0.22*	0.11	0.23*	0.12
CAR_TYPE_5	0.09	0.10	0.10	0.11
CAR_TYPE_6	0.13	0.11	0.13	0.12
JOBCLASS_2	-0.17	0.11	-0.17	0.12
JOBCLASS_3	-0.07	0.12	-0.07	0.13
JOBCLASS_4	-0.47**	0.17	-0.49**	0.19
JOBCLASS_5	-0.07	0.13	-0.06	0.15
JOBCLASS_6	-0.42**	0.13	-0.43**	0.14
JOBCLASS_7	-0.26**	0.12	-0.27**	0.13
JOBCLASS_8	-0.21*	0.11	-0.21*	0.12
JOBCLASS_9	0.05	0.13	0.02	0.14
MVR_PTS:AREA			-0.20**	0.03
REVOKED:AREA			-0.49**	0.23

Note. * $p < 0.10$; ** $p < 0.05$.

Table A4: TGLM1 and TGLM2 model coefficient estimates.

	Variable 1	Variable 2	Importance
1	AREA	MVR_PTS	73.66
2	AREA	REVOKED	14.67
3	AREA	CAR_TYPE	7.34
4	MVR_PTS	REVOKED	6.80
5	CAR_TYPE	REVOKED	2.49

Table A5: Summary of the top 10 most important two-way interactions in the TDboost model in the automobile claims data example.

	Competing Premium	
Base Premium	TGLM1	TGLM2
TGLM1	0	9.751 (0.213)
TGLM2	-2.172 (0.260)	0

Table A6: The averaged Gini indices and standard errors for TGLM1 and TGLM2 in the auto insurance claim data example based on 20 random splits.

References

Dunn, P. K. and Smyth, G. K. (2005), "Series evaluation of Tweedie exponential dispersion model densities," *Statistics and Computing*, 15, 267–280.

— (2008), "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion," *Statistics and Computing*, 18, 73–86.

Sandri, M. and Zuccolotto, P. (2008), "A bias correction algorithm for the Gini variable importance measure in classification trees," *Journal of Computational and Graphical Statistics*, 17.

— (2010), "Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms," *Statistics and Computing*, 20, 393–407.

Yip, K. C. and Yau, K. K. (2005), "On modeling claim frequency data in general insurance with extra zeros," *Insurance: Mathematics and Economics*, 36, 153–163.