Modeling for Bank Loan Data

School of Statistics
University of Minnesota

May 19, 2014

It is know that the Tweedie's Compound Poisson model has the probability form (Dunn and Smyth, 2005) that

$$p(y|\mu,\phi) = a(y,\phi)\exp\Big(\frac{1}{\phi}\Big(\frac{y\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\Big)\Big), \tag{1}$$

where

$$a(y,\phi) = \begin{cases} 1, & \text{if } y = 0, \\ \frac{1}{y}\sum_{t=1}^{\infty} W_t, & \text{if } y > 0, \end{cases}$$

with $\alpha = (2-\rho)/(\rho-1)$ and

$$W_t = \frac{y^{t\alpha}}{(\rho-1)^{t\alpha}(2-\rho)^t\gamma(t\alpha)\phi^{t(1+\alpha)}t!}.$$

Let $(y_i, \mathbf{x}_i)_{i=1}^n$ be i.i.d. observations with sample size n. The link function is $\log\mu_i = \boldsymbol{\beta}^T\mathbf{x}_i$. Given a modeling method, we use the first 50,000 bank loan data observations as the training data, and the rest of the data as the testing data. To measure the model performance, we use the testing data to compute the mean absolute error (MAE), mean squared error (MSE) and gini index (GI). First, we fit the lasso Tweedie model, and choose tuning parameter λ by five-fold cross-validation:

```
fit <- cv.HDtweedie(x,y,p=1.5,pmax=300)
```

The MAE, MSE and GI are 0.0143, 0.0426, and 0.110, respectively.

As another way to quickly summarize the data, we can estimate the marginal distribution for Y. Specifically, given an interval $I_k \subset \mathbb{R}$, using the testing data alone, we can get the empirical estimate for $P(Y \in I_k)$ by

$$\hat{P}_e(Y_i \in I_k) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \in I_k).$$

Table 1: Comparing regular lasso Tweedie model with its zero-inflated version.

Model	MAE	MSE	GI
lasso Tweedie	0.0143	0.0426	0.110
ZIF ($\hat{\phi} = 1$)	0.0143	0.0426	0.129

On the other hand, using the Tweedie model, we can have model based estimate of $P(Y \in I_k)$ by

$$\hat{P}_m(Y \in I_k) = \frac{1}{n}\sum_{i=1}^{n} \hat{P}_m(Y_i \in I_k|\mathbf{x}_i),$$

where conditional distribution $\hat{P}_m(Y_i \in I_k|\mathbf{x}_i)$ is computed by the fitted model. Unlike MAE, MSE and GI that depends only on the estimated μ, the computation of $\hat{P}_m(Y_i \in I_k|\mathbf{x}_i)$ also relies on the dispersion parameter ϕ (e.g., note that $P(Y = 0) = \exp(-\frac{\mu^{2-\rho}}{\phi(2-\rho)})$). Here, we try three methods for estimating ϕ: 1) a constant choice $\hat{\phi} = 1$; 2) Pearson method $\hat{\phi} = \frac{1}{n-df}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i^{\rho}} = 2.60$; 3) MLE $\hat{\phi} = 1.90$. Table 2 summarize the results. Not surprisingly, the value of $\hat{P}_m(Y \in I_k)$ can be quite different with different choice of $\hat{\phi}$, and in this particular example, MLE appears to be closest to $\hat{P}_e(Y_i \in I_k)$ in most intervals.

Table 2: Comparing $\hat{P}_e(Y_i \in I_k)$ and $\hat{P}_m(Y \in I_k)$.

I_k	$\hat{P}_e(Y \in I_k)$		$\hat{P}_m(Y \in I_k)$		
		$\hat{\phi} = 1$	Pearson	MLE	ZIF($\hat{\phi} = 1$)
{0}	0.9035	0.8396	0.9349	0.9121	0.9127
(0,0.01]	0.0124	0.0302	0.0053	0.0095	0.0095
(0.01,0.02]	0.0129	0.0245	0.0048	0.0084	0.0084
(0.02,0.03]	0.0107	0.0198	0.0044	0.0075	0.0075
(0.03,0.04]	0.0098	0.0161	0.0041	0.0067	0.0067
(0.04,0.05]	0.0065	0.0130	0.0037	0.0060	0.0059
(0.05,0.1]	0.0225	0.0364	0.0146	0.0215	0.0213
(0.1,0.15]	0.0086	0.0128	0.0095	0.0121	0.0119
(0.15,0.2]	0.0047	0.0045	0.0062	0.0068	0.0067
(0.2,1]	0.0077	0.0025	0.0120	0.0090	0.0090

1. Zero-Inflated Sparse Tweedie Model

Alternatively, instead of trying to estimate ϕ, we can use a zero-inflated model. Assume that the response is sampled from the following mixture model (sampled from a point mass 0 with probability q_i, and from the Tweedie model (as before, with log link) with probability $1-q_i$):

$$Y_i \sim \begin{cases} 0, & \text{with probability } q_i, \\ \text{Tweedie}(\mu_i, \phi), & \text{with probability } 1-q_i. \end{cases} \tag{2}$$

In the following, we consider the case that $\phi = 1$, and $q_i = q$ for every i. Let $l(\boldsymbol{\beta}, q; \mathbf{y})$ denote the log-likelihood for (2), and we intend to find

$$(\hat{\boldsymbol{\beta}}, \hat{q}) = \arg\min_{(\boldsymbol{\beta}, q)} l(\boldsymbol{\beta}, q; \mathbf{y}) + \lambda P_n(\boldsymbol{\beta}), \tag{3}$$

where $P_n(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_i|$. For the choice of λ, we simply use the tuning parameter obtained previously from five-fold cross validation with lasso Tweedie model, that is, $\lambda = 0.0003566$.

To solve (3), we resort to an EM algorithm. Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$, where π_i is the class variable, i.e., $\pi_i = 1$ if Y_i is sampled from Tweedie(μ_i, ϕ), and $\pi_i = 0$ if Y_i is sampled from the zero point mass. Then the log-likelihood for the joint distribution of $(\mathbf{y}, \boldsymbol{\pi})$ is

$$\begin{aligned} & l(\boldsymbol{\beta}, q; \mathbf{y}, \boldsymbol{\pi}) \\ = & \sum_{i=1}^{n} \Big(\pi_i \log a(y_i, \phi) + \frac{\pi_i}{\phi} \Big(\frac{y_i \mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho} \Big) + (1-\pi_i) \log(I(y_i = 0)) + \pi_i \log(1-q) + (1-\pi_i) \log q \Big). \end{aligned}$$

Taking posterior expectation w.r.t. $\boldsymbol{\pi}$, we have

$$\begin{aligned} & E_{\boldsymbol{\pi} | \mathbf{y}, \hat{\boldsymbol{\beta}}, \hat{q}} l(\boldsymbol{\beta}, q; \mathbf{y}, \boldsymbol{\pi}) \\ = & \sum_{i=1}^{n} \Big(\Delta_{1i} \log a(y_i, \phi) + \frac{\Delta_{1i}}{\phi} \Big(\frac{y_i \mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho} \Big) + \Delta_{0i} \log(I(y_i = 0)) + \Delta_{1i} \log(1-q) + \Delta_{0i} \log q \Big), \end{aligned}$$

where $\hat{\mu}_i = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i)$,

$$\Delta_{1i} = P(\pi_i = 1 | y_i, \hat{\boldsymbol{\beta}}, \hat{q}) \begin{cases} 1, & \text{if } y_i > 0, \\ \frac{(1-\hat{q}) \exp\left(\frac{1}{\phi} \left(\frac{y \hat{\mu}_i^{1-\rho}}{1-\rho} - \frac{\hat{\mu}_i^{2-\rho}}{2-\rho} \right) \right)}{(1-\hat{q}) \exp\left(\frac{1}{\phi} \left(\frac{y \hat{\mu}_i^{1-\rho}}{1-\rho} - \frac{\hat{\mu}_i^{2-\rho}}{2-\rho} \right) \right) + \hat{q}}, & \text{when } y_i > 0, \end{cases} \tag{4}$$

and $\Delta_{0i} = 1 - \Delta_{1i}$.

Thus, we perform the maximization step by updating $(\hat{\boldsymbol{\beta}}, \hat{q})$ with the following scheme.

- Update $\hat{\boldsymbol{\beta}}$ by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{\Delta_{1i}}{\phi}\Big(\frac{y_i \mu_i^{1-\rho}}{\rho - 1} + \frac{\mu_i^{2-\rho}}{2-\rho}\Big) + \lambda P_n(\boldsymbol{\beta});$$

- Update $\hat{q}$ by

$$\hat{q} = \frac{1}{n}\sum_{i=1}^{n} \Delta_{0i}.$$

To perform the expectation step, we update Δ_{1i} and Δ_{0i} by (4). We iterate the E-M steps until convergence $(\max\{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(\text{old}))^2, (\hat{q} - \hat{q}(\text{old}))^2\} < 10^{-6})$.

By the aforementioned steps, the zero-inflated Tweedie model (ZIF) with lasso gives $\hat{q} = 0.661$. The corresponding MAE, MSE and GI are shown in Table 1. We can see that the Gini index of the zero-inflated model increases by almost 2% compare to the regular lasso Tweedie model. We also compute $\hat{P}_m(Y \in I_k)$, which is shown in Table 2. Without the efforts to estimate ϕ (recall that we set $\hat{\phi} = 1$), the zero-inflated model performs competitively compared with the best of the regular lasso Tweedie models.

REFERENCES

Dunn, P. K. and Smyth, G. K. (2005), "Series Evaluation of Tweedie Exponential Dispersion Models Densities," *Statistics and Computing*, 15, 267-280.