Zero-inflated Tweedie Model with Lasso Penalty

Kevin McGregor

January 31^{st}, 2015

Introduction

We attempt to model and predict insurance payouts over a given period of time based on a set of predictor variables. Assume that we have two sets of predictor variables x_i, and z_i, for each of the observations $i \in 1, \ldots, n$. Some or all of the variables may appear in both of these vectors. The sum of each person's payouts over a specified time period is denoted by y_i.

In the past, the Tweedie distribution has been used to model data in this setting, due to its flexibility, its relationships with distributions such as Poisson and Gamma, and its (relatively) straightforward formulation as a generalized linear model (GLM). One remaining problem, however, is that a plurality of subjects in insurance data will have a zero payout. In order to deal with this problem within the Tweedie GLM, one could either model the dispersion parameter, or set up a zero-inflated model. In this project we do the latter; that is, we model both the mean of the Tweedie distribution, as well as a parameter specifying whether the payout amount is zero or non-zero. Additionally, we invoke a Lasso penalty on the parameter modeling the mean in order to do variable selection.

Formulation of Model

We start with the density of the Tweedie distribution, where $\mu > 0$ is the mean, $\phi > 0$ is the dispersion parameter, and $1 < \rho < 2$ is an additional parameter. If we take $a(y, \phi)$ to be a normalizing constant, the density will have the form:

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

Introducing the zero-inflation into the model gives us the following density, conditional on the value of π, which is 1 if the observed value y comes from the Tweedie distribution, and is 0 if y comes from the point-mass at $y = 0$.

$$f(y|\mu, \rho, \phi, \pi) = \left[a(y, \phi) \exp\left\{\frac{1}{\phi}\left(\frac{y\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right)\right\}\right]^{\pi} \mathbb{1}(y = 0)^{1-\pi} \tag{2}$$

In order to find the joint density of y and π, we notice that:

$$f(y, \pi|\mu, \rho, \phi) = f(y|\mu, \rho, \phi, \pi) f(\pi) \tag{3}$$

$$f(y|\mu, \rho, \phi, \pi) = \left[a(y, \phi) \exp\left\{\frac{1}{\phi}\left(\frac{y\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right)\right\}\right]^{\pi} \mathbb{1}(y = 0)^{1-\pi} q^{1-\pi} (1-q)^{\pi}, \tag{4}$$

where q is the probability of having a zero response, a parameter that will be modeled. Therefore, assuming ρ and ϕ are fixed, the log-likelihood is:

$$l(\mu, q|y, \pi) = \sum_{i=1}^{n} \left[\pi_i \log a(y_i, \phi) + \frac{\pi_i}{\phi}\left(\frac{y\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right) + (1-\pi_i) \log \mathbb{1}(y_i = 0) + \log q_i - \pi_i \log q_i + \pi_i \log(1-q_i)\right] \tag{5}$$

Next, we make the following reparameterization, for observed covariates x_i, and z_i, which may or may not share common variables:

$$\mu_i = \exp(\beta^\top x_i) \quad \text{and} \quad \log\left(\frac{1-q_i}{q_i}\right) = \alpha^\top z_i$$

The reparameterized log-likelihood is:

$$l(\beta, \alpha | y, \pi) = \sum_{i=1}^{n} \Bigg[\pi_i \log a(y_i, \phi) + \frac{\pi_i}{\phi} \left(\frac{y_i \exp([1-\rho]\beta^\top x_i)}{1-\rho} - \frac{\exp([2-\rho]\beta^\top x_i)}{2-\rho} \right) + (1-\pi_i) \log \mathbb{1}(y_i = 0) + \log\left(\frac{1}{1+\exp(\alpha^\top z_i)}\right) - \pi_i \alpha^\top z_i \Bigg] \quad (6)$$

Adding the lasso penalty term on the parameter modeling the mean of the Tweedie distribution, we end up with the following optimization problem:

$$\underset{\beta,\alpha}{\operatorname{argmin}} -l(\beta, \alpha | y, \pi) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

The Algorithm

We employ a two-step approach to find the solution to (7). First, we can maximize the objective function with respect to α, for a fixed β. We do this using the Newton-Raphson algorithm. Since the lasso penalty term does not involve α, this term drops out in the differentiation step:

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \left[-z_i \frac{\exp(\alpha^\top z_i)}{1+\exp(\alpha^\top z_i)} - \pi_i z_i \right]. \quad (8)$$

The Hessian matrix is then:

$$\frac{\partial^2 l}{\partial \alpha^2} = \sum_{i=1}^{n} \left[-z_i z_i^\top \frac{\exp(\alpha^\top z_i)}{(1+\exp(\alpha^\top z_i))^2} \right]. \quad (9)$$

So, on the k^{th} iteration, we update α using the following formula:

$$\alpha_{(k+1)} = -\left(-\frac{\partial l}{\partial \alpha}\right)\left(-\frac{\partial^2 l}{\partial \alpha^2}\right)^{-1}\Bigg|_{\alpha=\alpha_{(k)}} \quad (10)$$

$$= \sum_{i=1}^{n} \left[-z_i \frac{\exp(\alpha_{(k)}^\top z_i)}{1+\exp(\alpha_{(k)}^\top z_i)} - \pi_i z_i \right] \left(\sum_{i=1}^{n} \left[-z_i z_i^\top \frac{\exp(\alpha_{(k)}^\top z_i)}{(1+\exp(\alpha_{(k)}^\top z_i))^2} \right] \right)^{-1}. \quad (11)$$

After convergence has been reached, take the final estimate $\hat{\alpha}$ into $-l(\beta, \alpha | y, \pi)$, and then minimize with respect to β using a two-step strategy. First, find the second-order Taylor series approximation $l_Q(\beta_0, \beta)$. Then we minimize:

$$-l_Q(\beta_0, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \quad (12)$$

using blockwise majorization descent. The details of this part of the algorithm are left as future work. The algorithm alternates between minimizing the objective function with respect to α and β until some convergence criterion has been achieved for both of these parameters.