Test (hard)

Alessandro De Bettin

March 22, 2016

1 Introduction

The objective of this document is to derive the mathematical optimization problem for an elastic-net regularized log-logistic model. All kinds of censoring will be considered. Being $\mathbf{t}$ the vector of the observed values, from now on I'll consider $\mathbf{y}$=log($\mathbf{t}$) as the response values; it is easier to deal with logistic distributed data than with log-logistic data; the estimates are exactly the same, as known from the theory. Section 2 shows the log-likelihood and its first and second derivatives functions for the logistic model. Section 3 shows the optimization algorithm via coordinate-descent.

2 Likelihood functions

Since the likelihood function is proportional to the product of the probability of the single observations, censored and non censored data can be considered apart of each other. Let θ be the vectorial parameter and y the observed values,

$$L(\theta;y)=L_u(\theta;y_u) \cdot L_r(\theta;y_r) \cdot L_l(\theta;y_l) \cdot L_{in}(\theta;y_{in}),$$

where u stands for uncensored, r for right censored, l for left censored and in for interval censored. The likelihood based on all of the observations is the product of the likelihoods based on the 4 types of observation. Therefore, the log-likelihood is

$$l(\theta;y)=l_u(\theta;y_u) + l_r(\theta;y_r) + l_l(\theta;y_l) + l_{in}(\theta;y_{in}).$$

It is clear that the first derivative of the log-likelihood in the parameters is

$$l_*(\theta;y)=l_{*u}(\theta;y_u) + l_{*r}(\theta;y_r) + l_{*l}(\theta;y_l) + l_{*in}(\theta;y_{in}),$$

where $l_{*u}(\theta;y_u)$, $l_{*r}(\theta;y_r)$, $l_{*l}(\theta;y_l)$, $l_{*in}(\theta;y_{in})$ are the first order derivatives for each kind of censoring status. These are vectorial functions of dimension p=dim(θ).
The same reasoning can be applied to second order derivatives, being

$$l_{**}(\theta;y)=l_{**u}(\theta;y_u) + l_{**r}(\theta;y_r) + l_{**l}(\theta;y_l)+ l_{**in}(\theta;y_{in}),$$

where $l_{**u}(\theta;y_u)$, $l_{**r}(\theta;y_r)$, $l_{**l}(\theta;y_l)$, $l_{**in}(\theta;y_{in})$ are the second order derivatives for the 4 censoring states. $l_{**}(\theta;y)$ is a $p \times p$ matrix.

2.1 Parametrization

Before reporting these quantities for the logistic model, let's explain the parametrization. The logistic distribution is a position and scale family, therefore it can be split in a deterministic part (the position parameter, the mean) and a stochastic one (whose only parameter is the scale parameter). In the context of the regression model, we have

$$y=X\beta+\epsilon.$$

X is a $n \times (p-1)$ matrix, containing the information of the explanatory variables; every row is a different observation and every column is a different variable, as usual. β is a $p-1$ length vector; these are the regression parameters. $X\beta$ is the deterministic part of our model. ϵ is a random variable whose distribution is logistic; formally,

$$\epsilon \sim Logistic(0, s),$$

where s is the scale parameter, common to all of the observations. Therefore, the framework is similar to the one of the normal regression model, but the estimation process is entirely based on likelihood rather than least squares. Obviously, θ is a vector of length p, in particular

$$\theta = (\beta^T, s)^T.$$

2.2 Non censored data

Let n_u be the number of uncensored observations. The log-likelihood is

$$l_u(\beta, s; y_u)=-n_u \log(s)+\frac{\sum_{i=1}^{n_u} y_{iu}-\sum_{i=1}^{n_u} x_i^T\beta}{s} - 2\sum_{i=1}^{n_u} \log(1+\exp(\frac{y_{iu}-x_i^T\beta}{s})).$$

The first derivative is

$$l_{*u}(\beta, s;y_u) = \begin{cases} -\frac{\sum_{i=1}^{n_u} x_{ij}}{s} + 2\sum_{i=1}^{n_u} \frac{x_{ij}}{s(1+\exp(-\frac{y_{iu}-x_i^T\beta}{s}))}, & j \text{ in (1,...,p-1)} \\ -\frac{n_u}{s} - \frac{\sum_{i=1}^{n_u} y_{iu}-\sum_{i=1}^{n_u} x_i^T\beta}{s^2} + 2\sum_{i=1}^{n_u} \frac{y_{iu}-x_i^T\beta}{s^2(1+\exp(-\frac{y_{iu}-x_i^T\beta}{s}))}, & j=p \end{cases}.$$

The second order derivative is

$$l_{**u}(\beta, s;y_u) =$$

$$\begin{cases} -2\sum_{i=1}^{n_u} \frac{x_{ik}x_{ij}\exp(\frac{y_{iu}-x_i^T\beta}{s})}{s^2(1+\exp(\frac{y_{iu}-x_i^T\beta}{s}))^2}, & j,k \text{ in (1,...,p-1)} \\ \frac{\sum_{i=1}^{n_u} x_{ij}}{s^2} - 2\sum_{i=1}^{n_u} \frac{x_{ij}\exp(\frac{y_{iu}-x_i^T\beta}{s})[1+\exp(\frac{y_{iu}-x_i^T\beta}{s})+\frac{y_{iu}-x_i^T\beta}{s}]}{s^2(1+\exp(\frac{y_{iu}-x_i^T\beta}{s}))^2}, & j \text{ in (1,...,p-1)}, k=p. \\ \frac{n_u}{s^2} + 2\frac{\sum_{i=1}^{n_u} y_{iu}-\sum_{i=1}^{n_u} x_i^T\beta}{s^3} - 2\sum_{i=1}^{n_u} \frac{\frac{y_{iu}-x_i^T\beta}{s}\exp(\frac{y_{iu}-x_i^T\beta}{s})[2+2\exp(\frac{y_{iu}-x_i^T\beta}{s})+\frac{y_{iu}-x_i^T\beta}{s}]}{s^2(1+\exp(\frac{y_{iu}-x_i^T\beta}{s}))^2}, & j,k=p \end{cases}$$

For the $J = p$ and k in $(1, ..., p-1)$ case, the equation is the same as the second one, just substitute j with k.

2.3 Left censored data

Let n_l be the number of left censored observations. The log-likelihood is

$$l_l(\beta, s; y_u) = -\sum_{i=1}^{n_l} \log(1 + \exp(-\frac{y_{il} - x_i^T \beta}{s})).$$

The score function is

$$l_{*u}(\beta, s; y_u) = \begin{cases} -\sum_{i=1}^{n_l} \frac{x_{ij}}{s(1+\exp(\frac{y_{il} - x_i^T \beta}{s}))}, & j \text{ in (1,...,p-1)} \\ -\sum_{i=1}^{n_l} \frac{y_{il} - x_i^T \beta}{s^2(1+\exp(\frac{y_{il} - x_i^T \beta}{s}))}, & j = p \end{cases}.$$

The second derivative is

$$l_{**l}(\beta, s; y_u) = \begin{cases} -\sum_{i=1}^{n_l} \frac{x_{ik} x_{ij} \exp(\frac{y_{il} - x_i^T \beta}{s})}{s^2(1+\exp(\frac{y_{il} - x_i^T \beta}{s}))^2}, & j, k \text{ in (1,...,p-1)} \\ \sum_{i=1}^{n_l} \frac{x_{ij}[1+\exp(\frac{y_{il} - x_i^T \beta}{s}) - \exp(\frac{y_{il} - x_i^T \beta}{s})\frac{y_{il} - x_i^T \beta}{s}]}{s^2(1+\exp(\frac{y_{il} - x_i^T \beta}{s}))^2}, & j \text{ in (1,...,p-1)}, k = p. \\ \sum_{i=1}^{n_l} \frac{\frac{y_{il} - x_i^T \beta}{s}[2+2\exp(\frac{y_{il} - x_i^T \beta}{s}) - \exp(\frac{y_{il} - x_i^T \beta}{s})\frac{y_{il} - x_i^T \beta}{s}]}{s^2(1+\exp(\frac{y_{il} - x_i^T \beta}{s}))^2}, & j, k = p \end{cases}$$

For the $J = p$ and k in $(1, ..., p-1)$ case, the equation is the same as the second one, just substitute j with k.

2.4 Right censored data

Let n_r be the number of right censored observations. The log-likelihood is

$$l_r(\beta, s; y_u) = -\sum_{i=1}^{n_r} \log(1 + \exp(\frac{y_{il} - x_i^T \beta}{s})).$$

The score function is

$$l_{*r}(\beta, s; y_u) = \begin{cases} \sum_{i=1}^{n_r} \frac{x_{ij}}{s(1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}))}, & j \text{ in (1,...,p-1)} \\ \sum_{i=1}^{n_r} \frac{y_{ir} - x_i^T \beta}{s^2(1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}))}, & j = p \end{cases}.$$

The second derivative is

$$l_{**r}(\beta, s; y_u) = \begin{cases} -\sum_{i=1}^{n_r} \frac{x_{ik} x_{ij} \exp(-\frac{y_{ir} - x_i^T \beta}{s})}{s^2(1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}))^2}, & j, k \text{ in (1,...,p-1)} \\ -\sum_{i=1}^{n_r} \frac{x_{ij}[1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}) + \exp(-\frac{y_{ir} - x_i^T \beta}{s})\frac{y_{ir} - x_i^T \beta}{s}]}{s^2(1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}))^2}, & j \text{ in (1,...,p-1)}, k = p. \\ -\sum_{i=1}^{n_r} \frac{\frac{y_{ir} - x_i^T \beta}{s}[2+2\exp(-\frac{y_{ir} - x_i^T \beta}{s}) + \exp(-\frac{y_{ir} - x_i^T \beta}{s})\frac{y_{ir} - x_i^T \beta}{s}]}{s^2(1+\exp(-\frac{y_{ir} - x_i^T \beta}{s}))^2}, & j, k = p \end{cases}$$

For the $J = p$ and k in $(1, ..., p-1)$ case, the equation is the same as the second one, just substitute j with k.

2.5 Interval censored data

Let n_{in} be the number of interval censored observations. Let y_L be the vector of the lower bounds of the intervals, and y_U be the one of the upper bounds. The log-likelihood is

$$l_{in}(\beta, s; y_u) = \sum_1^{n_{in}} \frac{x_i^T \beta}{s} + \sum_1^{n_{in}} \log(\exp(-\frac{y_{Li}}{s}) - \exp(-\frac{y_{Ui}}{s})) - \sum_{i=1}^{n_r} \log(1 + \exp(\frac{y_{iL} - x_i^T \beta}{s})) -$$
$$\sum_{i=1}^{n_r} \log(1+\exp(\frac{y_{iU} - x_i^T \beta}{s})) = \sum_1^{n_{in}} \frac{x_i^T \beta}{s} + \sum_1^{n_{in}} \log(\exp(-\frac{y_{Li}}{s}) - \exp(-\frac{y_{Ui}}{s})) + l_r(\beta, s; y_U) + l_r(\beta, s; y_L).$$

The score function is

$$l_{*in}(\beta, s; y_u) =$$
$$\begin{cases} \sum_{i=1}^{n_{in}} \frac{x_{ij}}{s} + l_{*r}(\beta, s; y_U)_j + l_{*r}(\beta, s; y_L)_j, & j \text{ in (1,...,p-1)} \\ -\sum_1^{n_{in}} \frac{x_i^T \beta}{s^2} + \sum_1^{n_{in}} \frac{\frac{y_{Li}}{s^2} \exp(-\frac{y_{Li}}{s}) - \frac{y_{Ui}}{s^2} \exp(-\frac{y_{Ui}}{s})}{\exp(-\frac{y_{Li}}{s}) - \exp(-\frac{y_{Ui}}{s})} + l_{*r}(\beta, s; y_U)_p + l_{*r}(\beta, s; y_L)_p, & j = p \end{cases}.$$

The second derivative is

$$l_{**in}(\beta, s; y_u) =$$
$$\begin{cases} l_{**r}(\beta, s; y_U)_{jk} + l_{**r}(\beta, s; y_L)_{jk}, & j, k \text{ in (1,...,p-1)} \\ \sum_{i=1}^{n_{in}} -\frac{x_{ij}}{s} + l_{**r}(\beta, s; y_U)_{jp} + l_{**r}(\beta, s; y_L)_{jp}, & j \text{ in (1,...,p-1)}, k = p \\ 2\sum_1^{n_{in}} \frac{x_i^T \beta}{s^3} + \sum_1^{n_{in}} \frac{-2y_{Li} \exp(-2\frac{y_{Li}}{s}) - 2y_{Ui} \exp(-2\frac{y_{Ui}}{s}) + \exp(-\frac{y_{Ui}+y_{Li}}{s})[-\frac{y_{Ui}^2}{s} + 2y_{Ui} - \frac{y_{Li}^2}{s} + 2y_{Li} + 2\frac{y_{Li}y_{Ui}}{s}]}{s^3(\exp(-\frac{y_{Li}}{s}) - \exp(-\frac{y_{Ui}}{s}))^2} & \\ + l_{**r}(\beta, s; y_U)_{pp} + l_{**r}(\beta, s; y_L)_{pp}, & j, k = p \end{cases}.$$

For the $J = p$ and k in $(1, ..., p-1)$ case, the equation is the same as the second one, just substitute j with k.

3 Optimization

The objective is to maximize the log-likelihood under certain constraints. In particular, being $M(\theta; y, \lambda, \alpha)$ the function to maximize,

$$M(\theta; y, \lambda, \alpha) = l(\theta; y) - \lambda P_\alpha(\beta),$$
$$\lambda P_\alpha(\beta) = \lambda(\alpha \sum_{i=1}^{p-1} |\beta_i| + \tfrac{1}{2}(1-\alpha) \sum_{i=1}^{p-1} \beta_i^2).$$

The penalty is only applied to the β coefficients. Unluckily, the estimates cannot be calculated in closed form, even for $\lambda = 0$. Therefore, we need an efficient strategy to fit the model in reasonable time.

3.1 Approximation of the log-likelihood

The first step is the quadratic approximation of the log-likelihood in a point $\tilde{\theta}$ using Taylor's series. Indeed,

$$l(\theta; y) \approx l(\tilde{\theta}) + (\theta - \tilde{\theta})^T l_*(\tilde{\theta}) + \tfrac{1}{2}(\theta - \tilde{\theta})^T l_{**}(\tilde{\theta})(\theta - \tilde{\theta}).$$

Being $z(\tilde{\theta}) = \tilde{\theta} - l_{**}(\tilde{\theta}; y)^{-1} l_*(\tilde{\theta}; y)$, the quadratic approximation can be written as

$$l(\theta; y) \approx \tfrac{1}{2}(z(\tilde{\theta}) - \theta)^T l_{**}(\tilde{\theta})(z(\tilde{\theta}) - \theta) + c(\tilde{\theta}).$$

Let's now apply this approximation to the objective function.

$$\tilde{M}(\theta; y, \lambda, \alpha) = \tfrac{1}{2}(z(\tilde{\theta}) - \theta)^T l_{**}(\tilde{\theta})(z(\tilde{\theta}) - \theta) - \lambda P_\alpha(\beta).$$

Since the goal is the maximization of this function, it is possible to get rid of the constants. With this approximation, it is possible to find an explicit solution. In practice, we need to maximize

$$\tilde{M}_p(\theta; y, \lambda, \alpha) = \textstyle\sum_{i=1}^{p} \frac{1}{2p} w(\tilde{\theta})_i (z(\tilde{\theta})_i - \theta_i)^2 - \lambda P_\alpha(\beta),$$

where $w(\tilde{\theta})_i$ is the i-th diagonal entry of $l_{**}(\tilde{\theta}; y)$ (this is an approximation). The derivative of this function is

$$\tfrac{\partial \tilde{M}_p(\theta; y, \lambda, \alpha)}{\partial \beta_k} = -\tfrac{1}{p} w(\tilde{\theta})_k (z(\tilde{\theta}_k) - \theta_k) - \lambda(1 - \alpha)\beta_k - \lambda \alpha sgn(\beta_k).$$

The corresponding estimator for β_k is

$$\hat{\beta_k} = \tfrac{S(\frac{1}{p} w(\tilde{\theta})_k z(\tilde{\theta}_k), \lambda\alpha)}{\frac{1}{p} w(\tilde{\theta})_k - \lambda(1-\alpha)},$$

Where the $S(.,.)$ operator is the one defined by Friedman et al. (2007).
Since s is not subject to constraints, its estimate is $z(\tilde{\theta})_p$.

3.2 Algorithm

Let α be fixed. The idea is to estimate the parameters of various models using various values of λ. For a fixed value of λ the algorithm to estimate the parameters is the coordinate descent (ascent in this case). From a starting value of $\tilde{\theta}$, for each θ_i, $i = 1, ..., p$, the $\hat{\theta}_i$ are calculated, updating at each step the quantities $z(\tilde{\theta})$ and $w(\tilde{\theta})$. The procedure is iterated until convergence. Formally,

1. Choose a starting value $\tilde{\theta}$
2. For each parameter
   (a) Calculate $z(\tilde{\theta})$ and $w(\tilde{\theta})$
   (b) Put in the current slot of $\tilde{\theta}$ the estimate of the current parameter.
3. Verify if $\sum_{i=1}^{p} (\tilde{\theta}_i^{current} - \tilde{\theta}_i^{previous})^2 < threshold$
4. If the convergence condition is not reached return to step 2.

Applying the procedure to decreasing values of λ, we can use the estimates found for a value of λ as warm starts for the estimates of the following, slightly smaller, value of λ. Doing so we can see how estimates "evolve" in relation to λ.

This algorithm can be used in conjunction with the quantities reported in Section 2 in order to efficiently estimate the parameters of a log-logistic model.