Post-selection inference for generalized regression

Jonathan Taylor and Robert Tibshirani

December 31, 2015

Abstract

1 Introduction

- Data $(x_i, y_i), i = 1, 2, \ldots N$ with $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})$. Let $X = \{x_{ij}\}$ be the data matrix.
- Generalized regression model with linear predictor $\eta = \beta_0 + X\beta$ and log-likelihood $\ell(\beta_0, \beta)$. Consider the objective function

$$J(\beta_0, \beta) = -\ell(\beta_0, \beta) + \lambda \cdot \sum_1^p |\beta_j|) \tag{1}$$

- Let $\hat{\beta}_0, \hat{\beta}_1$ be the minimizers of $J(\beta_0, \beta)$. We wish to carry out post-selection inference for any functional $\gamma^T\beta$.
- Leading example: logistic regression. $\pi = E(Y|x)$; $\log \pi/(1-\pi) = \beta_0 + X\beta$. $\ell(\beta_0, \beta) = \sum[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)]$.
- Background: Gaussian case. Selected model M with sign vector s, the KKT conditions state that $\{\hat{M}, \hat{s}\} = (M, s)$ if and only if there exists β and u satisfying

$$\begin{aligned} X_M^{(} X_M^T\beta - y) + \lambda s) + \lambda s &= 0 \\ X_{-M}^T(X_M^T\beta - y) + \lambda s) + \lambda u &= 0 \\ \text{sign}(\beta) &= s \\ ||u||_\infty &< 1 \end{aligned} \tag{2}$$

This allows us to write the set of response y that yield the same M and s in the polyhedral form

$$\left\{ \begin{pmatrix} A_0(M,s) \\ A_1(M,s) \end{pmatrix} y < \begin{pmatrix} b_0(M,s) \\ b_1(M,s) \end{pmatrix} \right\} \tag{3}$$

- A convenient strategy for minimizing (1) to express the usual Newton-Raphson update as an iterative reweighted least squares (IRLS) step, and then replace the weighted least squares step by a constrained weighted least squares procedure.

  We define $u = \partial \ell/\partial \eta$, $W = -\partial^2 l/\partial \eta \eta^T$ and $z = \eta + W^{-1}u$ Then a one-term Taylor series expansion for $\ell(\beta)$ has the form

$$(z-\eta)^T W(z-\eta) \tag{4}$$

  Hence to minimize (1) we use the following procedure:

  1. Fix s and initialize $\hat{\beta} = 0$
  2. Compute η, W and z based on the current value of $\hat{\beta}$
  3. Minimize $(z - \beta_0 - X\beta)^T W(z - \beta_0 - X\beta) + \lambda \cdot \sum |\beta_j|$
  4. Repeat steps (2) and (3) until $\hat{\beta}_0, \hat{\beta}$ don't change.

- KKT

$$-X_M^T W(z - \beta_0 - X_M^T \beta) + \lambda s = 0$$

-

$$\hat{\beta} = (X_M^T W X_M)^{-1}(X_M^T W z - \lambda s)\ (active)$$
$$-X_{-M}^T W(z - X_M\beta) + \lambda u = 0, ||u||_\infty < 1\ (inactive)$$

$$u \quad = \quad X_{-M}^T W P_M W^{-1}(X_M^T)^+ s + \frac{1}{\lambda} X_{-M}^T W(I - P_M)z \tag{5}$$

- For active variables, diag(s)$\beta > 0$ implies $D(X_M^T X W_M)^{-1}(X_m^T W z - \lambda s) > 0$. where $D = \text{diag}(s)$.

  Hence $A_1 = -D(X_M^T W X_M)^{-1} X_M^T W, b_1 = -D(X_M^T W X_M)^{-1}\lambda s$

  For inactive variables, $A_0 = \frac{1}{\lambda}\begin{pmatrix} X_{-M}^T W) \\ -X_{-M}^T W \end{pmatrix}$, $b_0 = \begin{pmatrix} \mathbf{1} + X_{-M}^T W X_M \hat{\beta}/\lambda \\ \mathbf{1} - X_{-M}^T W X_M \hat{\beta}/\lambda \end{pmatrix}$

  Finally, let $A = \begin{pmatrix} A_1 \\ A_0 \end{pmatrix}, b = (b_1, b_0)$

- Idea: take $z \sim N(\mu, W^{-1})$ and apply polyhedral lemma to region $Az \leq b$

- Logistic regression: KKT

$$z = X\beta + \frac{y - \hat{p}}{\hat{p}(1-\hat{p})}$$

$$-X_M^T W(z - X_M^T \beta) + \lambda s = 0$$

- $$\hat{\beta} = (X_M^T W X_M)^{-1}(X_M^T W z - \lambda s) \ (active)$$
$$-X_{-M}^T W(z - X_M\beta) + \lambda u = 0, ||u||_\infty < 1 \ (inactive)$$

$$u \quad = \quad X_{-M}^T W P_M W^{-1}(X_M^T)^+ s + \frac{1}{\lambda} X_{-M}^T W(I - P_M)z \qquad (6)$$

- For active variables, diag(s)$\beta > 0$ implies $D(X_M^T X W_M)^{-1}(X_m^T W z - \lambda s) > 0$. where $D = \text{diag}(s)$.

  Hence $A_1 = -D(X_M^T W X_M)^{-1} X_M^T W, b_1 = -D(X_M^T W X_M)^{-1}\lambda s$

  For inactive variables, $A_0 = \frac{1}{\lambda}\begin{pmatrix} X_{-M}^T W) \\ -X_{-M}^T W \end{pmatrix}$, $b_0 = \begin{pmatrix} \mathbf{1} + X_{-M}^T W X_M \hat{\beta}/\lambda \\ \mathbf{1} - X_{-M}^T W X_M \hat{\beta}/\lambda \end{pmatrix}$

  Finally, let $A = \begin{pmatrix} A_1 \\ A_0 \end{pmatrix}, b = (b_1, b_0)$

- Idea: take $z \sim N(\mu, W^{-1})$ and apply polyhedral lemma to region $Az \leq b$

2 Jon's notes

We are conditioning on the active set and signs. Let $\hat{\beta} = \hat{\beta}_\lambda$ be the LASSO solution. We are going to fix the model M and signs s_M. So, it is a function of $M, X_M^T y, X_M, s_M$. Also, let

$$\hat{\pi} = \pi(X\hat{\beta}_\lambda)$$
$$W = \text{diag}(\hat{\pi}(1 - \hat{\pi}))$$

Let

$$z = X_M\hat{\beta} + \frac{y - \hat{\pi}}{\hat{\pi}(1 - \hat{\pi})}$$

The KKT conditions can then be written as

$$X^T(y - \hat{\pi}) = XW(z - X_M\hat{\beta}) = \lambda u$$

where $u \in \partial(\|\cdot\|_1)(\hat{\beta})$ so

$$u_M = s_M, \quad \|u_{-M}\|_\infty < 1.$$

By construction, we have that

$$\bar{\beta} = (X_M^T W X_M)^{-1}(X_M^T W z) = \hat{\beta} + \lambda(X_M^T W X_M)^{-1} s_M.$$

This is, up to some remainder, the unpenalized logistic regression estimator. The remainder, after rescaling, goes to 0 in probability (p fixed) before selection. So, under suitable assumptions about the selective likelihood ratio, so Lemma 1 of randomized response paper applies, and you can use this for inference about β_M.

Let's look at the inactive block. By construction,

$$\begin{aligned} X_{-M}^T W(z - X_M\hat{\beta}) &= X_{-M}^T(y - \hat{\pi}) \\ &\approx X_{-M}^T(y - \pi) - X_{-M}^T W X_M(\hat{\beta} - \beta_M) \\ &= X_{-M}^T(y - \pi) - X_{-M}^T W X_M(\bar{\beta} - \beta_M) + X_{-M}^T W X_M(X_M^T W X_M)^{-1} s_M \end{aligned}$$

with the remainder also going to 0 in probability after appropriate rescaling.

So, while z is not normally distributed, i.e. the KKT conditions are an affine function of z and the affine functionals are such that, they are asymptotically normally distributed. Further, the variances from Rob's normal approximation work as plugins variance estimators (Section 4.3 of http://arxiv.org/pdf/1507.06739v3.pdf) under the **selected model.**

Since our variance calculations only hold under the selected model, we might be losing some power using polyhedral lemma.

2.1 Selected is the same as full?

An asymptotic variance calculation under pairs model $(y_i, X_i) \overset{IID}{\sim} F$:

$$\mathrm{Cov}_F\left(X_{-M}^T(y-\pi) - E_F((X_{-M}^T W X_M))E((X_M^T W X_M))^{-1} X_M^T(y-\pi)), E_F((X_M^T W X_M))^{-1} X_M^T(y-\pi)\right) = 0$$

yields that the randomness in the inactive block is (asymptotically) independent of $\bar{\beta}$. This assumes that the selected model is correct, or, more precisely that $\hat{\pi}$ is a good estimate of $P_F(y = 1|X)$ so that

$$\frac{1}{n} X^T W X \approx \mathrm{Cov}_F((y - P_F(y = 1|X)) \cdot X)$$

(X on the RHS should be thought of as a random vector). This might not be true if link is misspecified or selected model is poor...

I think then the inactive blocks are not needed.

3 Current favorite version

$$\hat{\beta} = \hat{\beta}_\lambda = \mathrm{argmin}_\beta \ell(\beta) + \lambda\|\beta\|_1$$

$$M = \{j : \hat{\beta} \neq 0\}, s_M = \mathrm{sign}(\hat{\beta}[M])$$

$$\begin{aligned} \bar{\beta}_M &= \hat{\beta}[M] - \left(\nabla^2\ell(\hat{\beta})[M, M]\right)^{-1} \nabla\ell(\hat{\beta})_M \\ &= \hat{\beta}_M + \lambda\left(\nabla^2\ell(\hat{\beta})[M, M]\right)^{-1} s_M \\ &= \hat{\beta}_M + \lambda\ell_M(\hat{\beta}_M)^{-1} s_M \end{aligned}$$

where $\ell_M : \mathbb{R}^M \to \mathbb{R}$ is the objective funtions of the selected model and

$$\nabla \ell^2(\hat{\beta})[M, M] = \frac{\partial^2}{\partial \beta_i \partial \beta_j} \ell(\beta) \Big|_{\hat{\beta}}, \qquad i, j \in M$$

is an $|M| \times |M|$ matrix.

If ℓ is a negative log-likelihood, then under the selected model,

$$\bar{\beta}_M \approx N\left(\beta^*_M, \nabla^2 \ell_M(\hat{\beta}_M)^{-1}\right).$$

subject to affine constraints

$$\left\{\mathrm{diag}(s_M)\left[\bar{\beta}_M - \nabla^2 \ell_M(\hat{\beta}_M)^{-1} s_M\right] \geq 0\right\}.$$

We apply polyhedral lemma to $\bar{\beta}_M$, with M, s_M and $\nabla^2 \ell_M(\hat{\beta}_M)$ fixed.

For logistic regression, these should match your active block KKT conditions exactly where

$$\bar{\beta}_M = (X_M^T W X_M)^{-1} X_M^T W z$$

with

$$z = X_M \hat{\beta}_M + \frac{y - \hat{\pi}}{\hat{\pi}(1 - \hat{\pi})} = X_M \hat{\beta}_M + W^{-1}(y - \hat{\pi}).$$