Exercise 1

Solutions

David Phelz

9/12/2016

Exercise 1

David Puelz

Linear Regression

- Consider $y = X\beta + \varepsilon$

$$y \in \mathbb{R}^N, \quad X \in \mathbb{R}^{N\times P}, \quad \beta \in \mathbb{R}^P$$

- Principal of weighted least squares:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{w_i}{2}(y_i - x_i^T\beta)^2$$

A) (i) $\sum_{i=1}^{N} \frac{w_i}{2}(y_i - x_i^T\beta)^2$

$$= \frac{1}{2}\sum_{i=1}^{N} (y_i - x_i^T\beta)' w_i (y_i - x_i^T\beta)$$

$$= \frac{1}{2}\sum_{i=1}^{N} (y_i w_i y_i - 2 y_i w_i x_i^T\beta + x_i^T\beta w_i x_i^T\beta)$$

$$= \frac{1}{2}(y^T W y - 2 y^T W X\beta + (X\beta)^T W X\beta)$$

$$= \frac{1}{2}(y - X\beta)^T W (y - X\beta)$$

$\blacksquare$

(ii) Need to find minimum of

$$f(\beta) = y^T W y - 2\beta^T X^T W y + \beta^T X^T W X \beta$$

$$\Rightarrow \frac{\partial f(\beta)}{\partial \beta} = -2X^T W y + 2X^T W X \beta \quad \text{(matrix derivative)}$$

FOC is: $\frac{\partial f(\beta)}{\partial \beta} = 0 = -2X^T W y + 2X^T W X \hat{\beta}$

$\Rightarrow$ $\hat{\beta}$ must satisfy:

$$0 = -2X^T W y + 2X^T W X \hat{\beta}$$

$$\Longleftrightarrow$$

$$X^T W y = (X^T W X)\hat{\beta} \quad \blacksquare$$

(iii) Let's look at the QR factorization.

Theorem: Given problem

$\min_\beta \| y - X\beta \|_2$, its solution set is

equivalent to the solution set of $R\beta = Q^T y$ where Q is

orthonormal, R is upper triangular, and both come from the QR factorization of X.

one approach

A second approach would be to solve the set of normal equations: (FOC from Least-squares)

$$(X^T W X)\beta = X^T W y$$ where the unknown is y.

Write: $C = X^T W X$

$d = X^T W y$

$LL^T = C$ a Cholesky factorization (L upper triangular).

Algorithm: Solve for z $\quad Lz = d$, forward substitution since L is triangular.

Solve for β $\quad L^T \beta = z$, backward substitution

Class notes:

3 standard factorizations

1) Cholesky (LL') $\longrightarrow$ fast, unstable (than the book)

2) QR $\longrightarrow$ fast, stable.

3) SVD $\longrightarrow$ matrix that is close to rank deficient.

↓

Focus on QR:

$$\underset{N\times N}{W^{1/2}}\underset{N\times p}{X} = \underset{N\times p}{Q}\,R \qquad (\text{where } W^{1/2}W^{1/2} = W)$$

Q: orthonormal columns; R: square upper Δ $p\times p$ — reduced QR factorization.

Looking back at normal equations:

$$\Leftrightarrow X^T W Y = X^T W X \beta$$

$$\Leftrightarrow X^T W^{1/2} W^{1/2} Y = X^T W^{1/2} W^{1/2} X \beta$$

$$\Leftrightarrow (QR)^T W^{1/2} Y = (QR)^T QR \beta$$

$$\Leftrightarrow \cancel{R^T} Q^T W^{1/2} Y = \cancel{R^T} Q^T Q R \beta$$

$$Q^T W^{1/2} Y = R\beta$$

$\longleftarrow$ now we have an upper triangular system!!!

Generalized Linear Models

Let $y_i \sim \text{Bin}(m_i, w_i)$

with $w_i(\beta) = \dfrac{1}{1 + \exp\{-x_i^T \beta\}}$.

A) likelihood is:

$$\ell(\beta) = -\log \left\{ \prod_{i=1}^{N} p(y_i \mid \beta) \right\}$$

where $p(y_i \mid \beta) = \binom{m_i}{y_i} w_i^{y_i} (1-w_i)^{m_i - y_i}$

$$\Rightarrow \ell(\beta) = -\log \left\{ \prod_{i=1}^{N} \binom{m_i}{y_i} w_i^{y_i} (1-w_i)^{m_i-y_i} \right\}$$

$$\propto -\sum_{i=1}^{N} \left[y_i \log w_i + (m_i - y_i) \log(1-w_i) \right]$$

let's define

$$l_i(\beta) = y_i \log w_i + (m_i - y_i) \log(1-w_i)$$

so that

$$\ell(\beta) = -\sum_{i=1}^{N} l_i(\beta)$$

↓

→ for a general function, $f(x)$

$$\frac{d}{dx} \log f(x) = \frac{f'(x)}{f(x)} .$$

→ Similarly,

$$\nabla_\beta \log w_i(\beta) = \frac{\nabla_\beta w_i(\beta)}{w_i(\beta)} .$$

Solving for gradient of $w_i(\beta)$:

$$\nabla_\beta w_i(\beta) = \frac{-\exp\{\bar{x}_i^T\beta\}}{(1+\exp\{-x_i^T\beta\})^2} x_i = w_i(1-w_i)x_i$$

$$\Longrightarrow \frac{\nabla_\beta w_i(\beta)}{w_i(\beta)} = \frac{-\exp\{-x_i^T\beta\}}{1+\exp\{-x_i^T\beta\}} x_i$$

$$= -(1-w_i)x_i$$

Now, looking back at the likelihood, we have

$$\nabla l_i(\beta) = \nabla(y_i \log w_i) + (m_i - y_i)\log 1-w_i)$$

$$= -y_i(1-w_i)x_i + (m_i-y_i)w_i x_i$$

$$= -y_i x_i + \cancel{y_i w_i x_i} + m_i w_i x_i - \cancel{y_i w_i x_i}$$

$$= -(y_i - m_i w_i)x_i \quad]$$

where $\nabla_\beta \log(1-w_i(\beta)) = w_i x_i$ by similar argument to above.

Therefore,

$$\nabla l(\beta) = -\sum_{i=1}^{N} \nabla l_i(\beta)$$

$$= -\sum_{i=1}^{N} (y_i - m_i w_i) x_i$$

$$= - X^T S \quad , \quad X \text{ matrix} \quad , \quad S \text{ is vector with } N \text{ components } \{ i^{\text{th}} \text{ component } \{y_i - m_i w_i\}$$

c) Let $\beta_0 \in \mathbb{R}^p$, recall that we write

$\ell(\beta)$ as

$$\ell(\beta) = -\sum_{i=1}^{N} \ell_i(\beta)$$

with $\ell_i(\beta) = y_i \log w_i + (m_i - y_i) \log(1 - w_i)$

Taylor's theorem states (approximating $\ell(\beta)$ near β_0)

$$\ell(\beta) \approx \ell(\beta_0) + \underbrace{\nabla \ell(\beta_0)^T (\beta - \beta_0)}_{\substack{1^{st}\text{ order} \\ \text{term}}} + \underbrace{\frac{1}{2}(\beta - \beta_0)^T \nabla^2 \ell(\beta_0)}_{\substack{2^{nd}\text{ order} \\ \text{term}}} (\beta - \beta_0)$$

Let's solve for the Hessian, $\nabla^2 \ell(\beta_0)$.

Recall:

$$\nabla \ell(\beta) = -X^T S$$
, S vector with $i^{\underline{th}}$ component $y_i - m_i w_i$

$$= -X^T (y - Mw)$$
, M is matrix (diagonal) with $M_{ii} = m_i$.

$$= -X^T y + X^T M w$$

Now, we aim to calculate

$$\nabla^2 \ell(\beta) = \nabla(-X^T y + X^T M w)$$

$$= \nabla(X^T M w)$$

$$= X^T M \nabla w$$

where the gradient of a vector field is a second order tensor (i.e. matrix).

$$\nabla w = \frac{\partial w_i}{\partial \beta_j} e_i \otimes e_j$$ where $\{e_i\}_{i=1}^N$ is standard basis for $\mathbb{R}^N$.

$$= \begin{bmatrix} \frac{\partial w_1}{\partial \beta_1} & \cdots & \frac{\partial w_1}{\partial \beta_N} \\ \vdots & & \vdots \\ \frac{\partial w_N}{\partial \beta_1} & & \frac{\partial w_N}{\partial \beta_N} \end{bmatrix}$$

$$= \left[\nabla_\beta w_1(\beta) \cdots\cdots \nabla_\beta w_N(\beta) \right]^T$$

We've shown!

$$\nabla_\beta w_i(\beta) = w_i(1-w_i)\, x_i$$

$$\Longrightarrow$$

$$\nabla w = \left[w_1(1-w_1)x_1 \cdots\cdots w_N(1-w_N)\, x_N \right]^T$$

$$= \tilde{W} X$$, W diagonal matrix with $W_{ii} = w_i(1-w_i)$

Therefore

So simple.

$$\boxed{\nabla^2 \ell(\beta) = \nabla(X^T M w) = X^T M \nabla w = X^T M \tilde{W} X}$$

So, by Taylor's theorem

$$\ell(\beta) \approx$$

$$\ell(\beta_0) - (X^T S)^T(\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^T X^T A X(\beta-\beta_0)$$

$$\left\{ \begin{array}{l} \underline{\text{with}} : A = M\tilde{W} \\ \tilde{W}_{ii} = w_i(1-w_i) \\ M_{ii} = m_i \end{array} \right.$$

$$=$$

$$\ell(\beta_0) - S^T(X\beta - X\beta_0) + \frac{1}{2}(X\beta - X\beta_0)^T A (X\beta - X\beta_0)$$

$$=$$

$$\ell(\beta_0) - S^T(X\beta - \tilde{z}) + \frac{1}{2}(X\beta - \tilde{z})^T A (X\beta - \tilde{z})$$

$$\left\{ \tilde{z} = X\beta_0 . \right.$$

completing the square! wrt to $\underline{X\beta - \tilde{z}}$

$$\propto \quad \frac{1}{2}(X\beta - \tilde{z} - A^{-1}S)^T A (X\beta - \tilde{z} - A^{-1}S) + c$$

Flipping around signs

$$= \frac{1}{2}(\{\tilde{z} + A^{-1}S\} - X\beta)^T A (\{\tilde{z} + A^{-1}S\} - X\beta) + c$$

$$= \frac{1}{2}(z - X\beta)^T A (z - X\beta) + c$$

$$z = X\beta_0 + (M\hat{W})^{-1}S \quad , \quad A = M\tilde{W}$$