MAT-468: Sesión 4, Cálculos en regresión II

Felipe Osorio

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Método gradientes conjugados (GC) en regresión lineal

En el contexto de regresión lineal, considere:

$$\phi(\boldsymbol{\beta}) = \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

El objetivo del procedimiento GC[1] es producir la secuencia:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \lambda_k \boldsymbol{p}_k, \qquad k = 0, 1, \dots. \tag{1}$$

El algortimo básico considera:

$$\lambda_k = \frac{\boldsymbol{p}_k^\top \boldsymbol{g}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{p}_k}, \qquad \boldsymbol{g}_k = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}).$$

(En efecto, $\partial\phi(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$) y actualizamos la dirección de búsqueda como:

$$\boldsymbol{p}_{k+1} = \boldsymbol{g}_{k+1} + \delta_k \boldsymbol{p}_k, \qquad \delta_k = -\frac{\boldsymbol{g}_{k+1}^\top \boldsymbol{p}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{p}_k}.$$

[1] McIntosh (1982), Lecture Notes in Statistics 10.

Método gradientes conjugados (GC) en regresión lineal

Se ha sugerido usar:

$$\lambda_k = \frac{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{y}}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k},$$

y actualizar

$$\boldsymbol{p}_{k+1} = \boldsymbol{g}_{k+1} + \delta_{k+1} \boldsymbol{p}_k, \qquad \delta_{k+1} = -\frac{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{g}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k}.$$

Para hacer el proceso más simple es recomendable calcular

$$\boldsymbol{h}_k = \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k.$$

De este modo el requerimiento de almacenamiento del algoritmo es sólo $4p$.

Gradientes conjugados en regresión lineal

Algoritmo 1: Gradientes conjugados para regresión lineal.

Entrada : Datos $\boldsymbol{X}$ y $\boldsymbol{y}$

Parámetros: Tolerancia τ.

begin

  Hacer $\boldsymbol{\beta} = \boldsymbol{0}$, $\boldsymbol{p} = \boldsymbol{g} = -\boldsymbol{X}^\top \boldsymbol{y}$, $\delta = 0$ y $\gamma = \|\boldsymbol{g}\|^2$

  while $\gamma > \tau$ **do**

    Calcular $\boldsymbol{h} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{p}$ y $u = \boldsymbol{p}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{p} = \boldsymbol{p}^\top \boldsymbol{h}$

    if $k \neq 1$ **then**

      $\delta = -\boldsymbol{h}^\top \boldsymbol{g}/u$

      $\boldsymbol{p} = \boldsymbol{g} + \delta\, \boldsymbol{p}$

    $\lambda = -\boldsymbol{p}^\top \boldsymbol{g}/u$

    $\boldsymbol{\beta} = \boldsymbol{\beta} + \lambda\, \boldsymbol{p}$

    $\boldsymbol{g} = \boldsymbol{g} + \lambda\, \boldsymbol{h}$

  end

  return $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

end

Alternativas a mínimos cuadrados

- Soluciones regularizadas: Regresión ridge.
- Estimación vía IRLS:
  - Modelos lineales generalizados.
  - Estimación L_1.
  - Estimación M.
  - Regresión lineal considerando distribuciones con colas pesadas.

Colinealidad en regresión lineal

Considere el modelo de regresión lineal

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

con $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\mathsf{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ y $\mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

Es bien conocido que cuando $\boldsymbol{X}$ es mal-condicionada, el sistema de ecuaciones

$$\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y},$$

puede ser inestable (ver, Stewart, 1987 y Belsley, 1991).

Considere la descomposición valor singular (SVD) de $\boldsymbol{X}$,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

donde $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, tal que $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_r$, $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_r)$ con $d_1 \geq \cdots \geq d_r > 0$, $\boldsymbol{V} \in \mathbb{R}^{r\times r}$ es matriz ortogonal y $r = \operatorname{rg}(\boldsymbol{X})$.

La detección de colinealidad en el modelo lineal puede ser llevada a cabo por medio del número condición

$$\kappa(\boldsymbol{X}) = \|\boldsymbol{X}\|\|\boldsymbol{X}^+\| = \frac{d_1}{d_r},$$

y $\kappa(\boldsymbol{X})$ "grande" es indicador de colinealidad.

Considere la matriz

$$\boldsymbol{A} = \begin{pmatrix} 1.000 & 0.500 \\ 0.667 & 0.333 \end{pmatrix}, \qquad \boldsymbol{A}^{-1} = \begin{pmatrix} -666 & 1000 \\ 1344 & -2000 \end{pmatrix}.$$

El número condición se define como $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|$ para $\|\cdot\|$ alguna norma matricial.

Por ejemplo[2],

$$\kappa_1(\boldsymbol{A}) = \|\boldsymbol{A}\|_1\|\boldsymbol{A}^{-1}\|_1 = (1.667)(3000) = 5001$$

$$\kappa_\infty(\boldsymbol{A}) = \|\boldsymbol{A}\|_\infty\|\boldsymbol{A}^{-1}\|_\infty = (1.500)(3344) = 5016$$

$$\kappa_2(\boldsymbol{A}) = \frac{\max_{x\neq 0}\|\boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|}{\min_{x\neq 0}\|\boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|} = \left|\frac{\lambda_{\max}}{\lambda_{\min}}\right| = \frac{1.333375}{0.000375} = 3555.778$$

[2] $\|\boldsymbol{A}\|_p = \max_{\|x\|_p=1}\|\boldsymbol{A}\boldsymbol{x}\|_p$, $\|\boldsymbol{A}\|_2 = \sqrt{\rho(\boldsymbol{A}^\top\boldsymbol{A})}$

Cemento Portland (Woods, Steinour y Starke, 1932)

Estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de cementos Portland. Woods et al. (1932) consideraron cuatro compuestos para los clinkers desde los que se produce el cemento.

La respuesta (Y) es la emisión de calor después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos: aluminato tricálcico (X_1), silicato tricálcico (X_2), ferrito aluminato tetracálcico (X_3) y silicato dicálcico (X_4).

Cemento Portland (Woods, Steinour y Starke, 1932)

Siguiendo a Woods et al. (1932) consideramos un modelo lineal sin intercepto (modelo homogéneo). El número condición escalado es $\kappa(\boldsymbol{X}) = 9.432$, esto es $\boldsymbol{X}$ es bien condicionada. (variables centradas, $\kappa(\tilde{\boldsymbol{X}}) = 37.106$)

Por otro lado, Hald (1952), Gorman y Toman (1966) y Daniel y Wood (1980) adoptaron un modelo con intercepto (modelo no homogéneo). En cuyo caso, $\kappa(\boldsymbol{X}) = 249.578$, sugiriendo la presencia de colinealidad.

El aumento en el número condición se debe a que existe una relación lineal aproximada, pues

$$x_1 + x_2 + x_3 + x_4 \approx 100.$$

de modo que incluir el intercepto causa una colinealidad severa.

Tratamiento de colinealidad en regresión lineal

El estimador ridge (Hoerl y Kennard, 1970)[3],

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \qquad \lambda > 0.$$

puede ser visto como la solución del problema regularizado:

$$\min_{\beta} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2,$$

o bien como el problema mínimos cuadrados con datos aumentados:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_* \end{pmatrix}.$$

En este contexto λ es un parámetro de regularización (parámetro ridge).

[3] Technometrics **12**, 55-67.

El mejor método para obtener $\widehat{\boldsymbol{\beta}}_\lambda$ es usar la descomposición SVD

$$\widehat{\boldsymbol{\beta}}_\lambda = \boldsymbol{V}\widehat{\boldsymbol{\alpha}}_\lambda,$$

con

$$\widehat{\boldsymbol{\alpha}}_\lambda = (\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}\boldsymbol{z} = \begin{pmatrix} z_1 d_1/(d_1^2 + \lambda) \\ \vdots \\ z_p d_p/(d_p^2 + \lambda) \end{pmatrix},$$

donde $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$. Un procedimiento recomendado para seleccionar el parámetro ridge es validación cruzada generalizada (Golub, Heath y Wahba, 1979), definido como:

$$\mathsf{GCV}(\lambda) = \frac{1}{n}\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2}{\{\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{H}(\lambda))\}^2},$$

con $\boldsymbol{H}(\lambda) = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top$ y definimos edf $= \operatorname{tr} \boldsymbol{H}(\lambda)$.

Es fácil mostrar que:

$$\operatorname{tr} \boldsymbol{H}(\lambda) = \operatorname{tr} \boldsymbol{D}^{\mathbf{2}}(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p)^{-1} = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

Así,

$$\mathsf{GCV}(\lambda) = \frac{\|\boldsymbol{z} - \boldsymbol{D}\widehat{\boldsymbol{\alpha}}_\lambda\|^2/n}{(1 - \mathsf{edf}\,/n)^2}.$$

Esto permite evaluar la función $\mathsf{GCV}(\lambda)$ de forma simple. Para escoger un λ_{opt} se ha sugerido:

- Considerar una grilla de valores para λ.
- Optimizar $\mathsf{GCV}(\lambda)$ usando un procedimiento de minimización unidimensional.

Cemento Portland

Resultados de estimación:

Estimador	β_0	β_1	β_2	β_3	β_4	σ^2
LS (homogéneo)		2.193	1.153	0.759	0.486	4.047
LS (No homog.)	62.405	1.551	0.510	0.102	-0.144	3.682
Ridge, LW	17.189	2.016	0.976	0.578	0.313	3.874
Ridge, HKB	8.587	2.105	1.065	0.668	0.400	3.953
Ridge, GCV	0.085	2.165	1.159	0.738	0.489	4.055

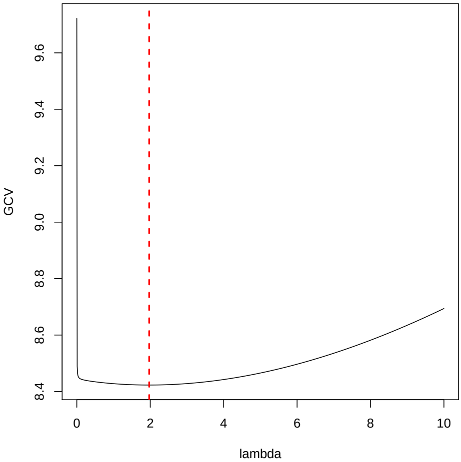
Se utilizó

$$\widehat{\lambda}_{\mathsf{HKB}} = ps^2/\|\widehat{\boldsymbol{\beta}}_{\mathsf{LS}}\|^2 = 0.00767 \qquad \text{(Hoerl, Kennard y Baldwin, 1975).}$$

$$\widehat{\lambda}_{\mathsf{LW}} = ps^2/\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{LS}}\|^2 = 0.00321 \qquad \text{(Lawless y Wang, 1976).}$$

Además, se consideró una grilla de valores para $\lambda = 0.00, 0.01, 0.02, \dots, 10.00$, obteniendo $\widehat{\lambda}_{\mathsf{opt}} = 1.97$.

Cemento Portland



Selección del parámetro ridge usando GCV, $\lambda_{\text{opt}} = 1.97$.

Uno de los primeros procedimientos robustos en regresión[4] corresponde al problema:

$$\min_{\beta} \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}|.$$

Mínimo desvío absoluto (LAD) o regresión L_1 puede ser planteado como un problema de programación lineal, considerando las partes positivas y negativas de los residuos, $\boldsymbol{e}^+$ y $\boldsymbol{e}^-$, respectivamente, y análogamente para $\boldsymbol{\beta}^+$, $\boldsymbol{\beta}^-$.

Así, el problema puede ser expresado como (Charnes, Cooper y Ferguson, 1955):

$$\min_{\beta} \mathbf{1}^\top(\boldsymbol{e}^+ + \boldsymbol{e}^-),$$

$$\text{sujeto a: } \boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + (\boldsymbol{e}^+ - \boldsymbol{e}^-),$$

con $\boldsymbol{\beta}^+$, $\boldsymbol{\beta}^-$, $\boldsymbol{e}^+$, $\boldsymbol{e}^-$ deben ser todos $\geq \mathbf{0}$.

Observación: Barrodale y Roberts (1973, 1974) presentan un algoritmo de propósito especial para resolver este problema modificando el método simplex y la estructura de datos requerida.

[4] Este método es, de hecho, anterior a LS!

Uno de los primeros procedimientos robustos en regresión[4] corresponde al problema:

$$\min_{\beta} \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}|.$$

Mínimo desvío absoluto (LAD) o regresión L_1 puede ser planteado como un problema de programación lineal, considerando las partes positivas y negativas de los residuos, $\boldsymbol{e}^+$ y $\boldsymbol{e}^-$, respectivamente, y análogamente para $\boldsymbol{\beta}^+$, $\boldsymbol{\beta}^-$.

Así, el problema puede ser expresado como (Charnes, Cooper y Ferguson, 1955):

$$\min_{\beta} \mathbf{1}^\top(\boldsymbol{e}^+ + \boldsymbol{e}^-),$$

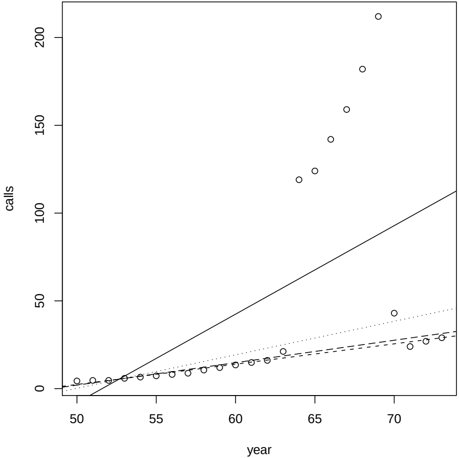
$$\text{sujeto a: } \boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + (\boldsymbol{e}^+ - \boldsymbol{e}^-),$$

con $\boldsymbol{\beta}^+$, $\boldsymbol{\beta}^-$, $\boldsymbol{e}^+$, $\boldsymbol{e}^-$ deben ser todos $\geq \mathbf{0}$.

Observación: Barrodale y Roberts (1973, 1974) presentan un algoritmo de propósito especial para resolver este problema modificando el método simplex y la estructura de datos requerida.

[4] Este método es, de hecho, anterior a LS!

Llamadas telefónicas en Bélgica 1950-73 (Rousseeuw y Leroy, 1987)



Ajustes: LS, normal contaminada ($\epsilon = .15, \gamma = .10$), Student-$t$ ($\nu = 2.5$), L_1.

Regresión L_1

- Schlossmacher (1973) originalmente propuso calcular estimadores LAD usando IRLS.
- Lange y Sinsheimer (1993) y Phillips (2002) identificaron que este procedimiento IRLS corresponde a un algoritmo EM.
- Sin embargo, también ha sido reportado que este algoritmo puede ser incapaz de detectar la observaciones básicas de manera eficiente.

Regresión lineal: función LAD de L1pack

Considere el modelo

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\sqrt{2}\tau_i)^{-1}\epsilon_i, \qquad i = 1, \dots, n$$

donde $\epsilon \overset{\text{ind}}{\sim} \mathsf{N}(0, \phi)$ y τ_i tiene función de densidad

$$g(\tau_i) = \tau_i^{-3} \exp(-\tfrac{1}{2}\tau_i^{-2}).$$

El algoritmo EM procede a llevar a cabo la estimación de $\boldsymbol{\beta}$ y ϕ iterativamente mediante maximizar la función:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^n W_i^{(k)}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W}^{(k)}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

donde $\boldsymbol{W}^{(k)} = \operatorname{diag}(W_1^{(k)}, \dots, W_n^{(k)})$ y los pesos son dados por:

$$W_i^{(k)} = \mathsf{E}(\tau_i^2|Y_i, \boldsymbol{\theta}^{(k)}) = \sigma^{(k)}/\sqrt{2}|Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}^{(k)}|,$$

para $|Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}^{(k)}| > 0$.

Regresión lineal: función LAD de L1pack

Considere el modelo

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\sqrt{2}\tau_i)^{-1}\epsilon_i, \qquad i = 1, \dots, n$$

donde $\epsilon \overset{\text{ind}}{\sim} \mathsf{N}(0, \phi)$ y τ_i tiene función de densidad

$$g(\tau_i) = \tau_i^{-3} \exp(-\tfrac{1}{2}\tau_i^{-2}).$$

El algoritmo EM procede a llevar a cabo la estimación de $\boldsymbol{\beta}$ y ϕ iterativamente mediante maximizar la función:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^{n} W_i^{(k)}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W}^{(k)}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

donde $\boldsymbol{W}^{(k)} = \operatorname{diag}(W_1^{(k)}, \dots, W_n^{(k)})$ y los pesos son dados por:

$$W_i^{(k)} = \mathsf{E}(\tau_i^2|Y_i, \boldsymbol{\theta}^{(k)}) = \sigma^{(k)}/\sqrt{2}|Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}^{(k)}|,$$

para $|Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}^{(k)}| > 0$.

Algoritmo IRLS

Paso de coeficientes:

- Calcular $\boldsymbol{r}^{(k)} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}$ y ${\boldsymbol{W}^{(k)}}^{1/2} = \operatorname{diag}(W_1^{(k)}, \dots, W_n^{(k)})$
- Obtener $\boldsymbol{\delta}^{(k)}$ como solución del problema WLS

$$\min_{\delta} \|{\boldsymbol{W}^{(k)}}^{1/2}(\boldsymbol{r}^{(k)} - \boldsymbol{X}\boldsymbol{\delta})\|^2$$

- Actualizar $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \boldsymbol{\delta}^{(k)}$.

Paso de escala:

$$\phi^{(k+1)} = \frac{1}{n}\|{\boldsymbol{W}^{(k)}}^{1/2}\boldsymbol{r}^{(k+1)}\|^2$$

Criterio de convergencia: basado en el criterio usado en la función glm.fit().

Sea $\boldsymbol{X}_* = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}$, $\boldsymbol{Y}_* = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{Y}$ y calcular la descomposición QR de $\boldsymbol{X}_*$

$$\text{(DGEQRF)} \qquad \boldsymbol{X}_* = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix}, \quad \boldsymbol{Q} \in \mathcal{O}_n \text{ y } \boldsymbol{R} \in \mathbb{R}^{p\times p} \text{ triangular superior,}$$

considere $\boldsymbol{c} = \boldsymbol{Q}^\top \boldsymbol{Y}_*$, entonces

$$\text{(DORMQR)} \qquad \boldsymbol{Q}^\top \boldsymbol{e}_* = \boldsymbol{Q}^\top \widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{Y} - \boldsymbol{X\beta}) = \begin{pmatrix}\boldsymbol{c}_1 - \boldsymbol{R\beta}\\ \boldsymbol{c}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\end{pmatrix}$$

de este modo, $\boldsymbol{\delta}$ es solución del sistema triangular

$$\text{(DTRTRS)} \qquad \boldsymbol{R\delta} = \boldsymbol{r}_1 \qquad \Rightarrow \qquad \boldsymbol{R\beta}^{(k+1)} = \boldsymbol{c}_1,$$

actualizar $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \boldsymbol{\delta}$ (DAXPY) y $\phi^{(k+1)} = \|\boldsymbol{r}_2\|^2/n$ (DNRM2). Finalmente, note que

$$\text{(DORMQR)} \qquad \widehat{\boldsymbol{Y}}_* = \boldsymbol{X}_*\boldsymbol{\beta}^{(k+1)} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{R\beta}^{(k+1)}\\ \boldsymbol{0}\end{pmatrix} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{c}_1\\ \boldsymbol{0}\end{pmatrix}$$

Referencias bibliográficas

Golub, G.H., Heath, M., Wahba, G. (1979).
Generalized cross-validation as a method for choosing a good ridge parameter.
Technometrics **21**, 215-223.

Hoerl, A.E., Kennard, R.W., Baldwin, K.F. (1975).
Ridge regression: some simulations.
Communications in Statistics **4**, 105-123.

Lange, K., Sinsheimer, J.S. (1993).
Normal/independent distributions and their applications in robust regression.
Journal of Computational and Graphical Statistics **2**, 175-198.

Lawless, J.F., Wang, P. (1976).
A simulation study of ridge and other regression estimators.
Communications in Statistics – Theory and Methods **14**, 1589-1604.

Phillips, R.F. (2002).
Least absolute deviations estimation via the EM algorithm.
Statistics and Computing **12**, 281-285.