

MAT-468: Sesión 3, Solución de sistemas lineales II

Felipe Osorio

<http://fosorios.mat.utfsm.cl>

Departamento de Matemática, UTFSM



Considere resolver el sistema lineal

$$Ax = b,$$

donde $A \in \mathbb{R}^{n \times n}$ y asumiremos que el sistema admite una **solución exacta** \hat{x} .

A continuación se describe:

- ▶ El método Jacobi.
- ▶ Método Gauss-Seidel.
- ▶ Método de gradientes conjugados.



Para uniformizar la presentación de todos los métodos, sea D , L y U las partes diagonal, triangular inferior y triangular superior de la matriz A , respectivamente¹.

Es decir $D = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$, mientras que

$$L = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ a_{21} & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & 0 & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \dots & 0 & a_{n-1,n} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

¹Evidentemente $A = D + L + U$.

Un **método iterativo** para la solución de sistemas lineales construye una secuencia $\{\mathbf{x}^{(k)}\}$, para $k = 0, 1, \dots$, que bajo ciertas condiciones converge a $\hat{\mathbf{x}}$.

$\mathbf{x}^{(0)}$ representa una **estimación inicial** y la secuencia es construída como una **regla** del tipo:

$$\mathbf{x}^{(k+1)} = \mathbf{B}_k \mathbf{x}^{(k)} + \mathbf{C}_k \mathbf{b}, \quad k = 0, 1, \dots, \quad (1)$$

donde $\mathbf{B}_k, \mathbf{C}_k \in \mathbb{R}^{n \times n}$.

Las diferentes elecciones de \mathbf{B}_k y \mathbf{C}_k definen los distintos algoritmos.



La condición básica sobre B_k y C_k que asegura la convergencia del algoritmo es la siguiente:

$$B_k + C_k A = I.$$

En efecto, se debe tener que

$$\hat{x} = B_k \hat{x} + C_k b,$$

y dado que \hat{x} es solución del sistema de ecuaciones $Ax = b$, tenemos

$$\begin{aligned}\hat{x} &= B_k \hat{x} + C_k A \hat{x} \\ &= (B_k + C_k A) \hat{x},\end{aligned}$$

desde donde sigue el resultado.



Resultado 1

Una secuencia $\{\mathbf{x}^{(k)}\}$ dada por (1) converge a la solución del sistema $\mathbf{Ax} = \mathbf{b}$ para cualquier $\mathbf{x}^{(0)}$, sólo si

$$\lim_{k \rightarrow \infty} \mathbf{B}_k \mathbf{B}_{k-1} \cdots \mathbf{B}_0 = \mathbf{0}$$

En la práctica se utiliza **métodos iterativos estacionarios**, es decir, tal que $\mathbf{B}_k = \mathbf{B}$ y $\mathbf{C}_k = \mathbf{C}$ son constantes. De este modo, tenemos la siguiente regla:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{C}\mathbf{b}, \quad k = 0, 1, \dots, \quad (2)$$

y la condición de convergencia en el resultado anterior adopta una forma más simple

Resultado 2

Una secuencia $\{\mathbf{x}^{(k)}\}$ dada por (2) converge a la solución del sistema $\mathbf{Ax} = \mathbf{b}$ para cualquier $\mathbf{x}^{(0)}$, si y sólo si el **radio espectral** $\rho(\mathbf{B}) < 1$, donde $\rho(\mathbf{B}) = \max\{\lambda_1, \dots, \lambda_n\}$ con $\lambda_1, \dots, \lambda_n$ los valores propios de \mathbf{B}^2 .

²Esta condición es equivalente a $\|\mathbf{B}\| < 1$ para cualquier norma matricial.



Se suele **declarar convergencia** cuando algún criterio preestablecido es satisfecho.

Específicamente, para un nivel de tolerancia τ , el proceso iterativo se detiene si

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \leq \tau, \quad \|\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}\| \leq \tau, \quad \text{o} \quad \|\mathbf{r}^{(k+1)}\| \leq \tau,$$

donde $\mathbf{r}^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}$ es el **vector residual**.

Además se suele incluir un **número máximo aceptable de iteraciones**.



Suponga que A tiene elementos diagonales no cero. Entonces el sistema $Ax = b$ puede ser escrito como

$$Dx + (L + U)x = b,$$

y de este modo, tenemos que

$$x = D^{-1}[-(L + U)x + b]$$

Esto motiva el siguiente algoritmo:

$$x^{(k+1)} = -D^{-1}(L + U)x^{(k)} + D^{-1}b, \quad k = 0, 1, \dots, \quad (3)$$

Observación:

La condición $\rho(D^{-1}(L + U)) < 1$ es satisfecha para una amplia clase de matrices (por ejemplo, diagonal dominantes, simétricas y definidas positivas).



Análogamente al método Jacobi, podemos reescribir el sistema $A\mathbf{x} = \mathbf{b}$ como

$$(\mathbf{L} + \mathbf{D})\mathbf{x} + \mathbf{U}\mathbf{x} = \mathbf{b},$$

es decir

$$\mathbf{x} = (\mathbf{L} + \mathbf{D})^{-1}[-\mathbf{U}\mathbf{x} + \mathbf{b}]$$

Esto lleva al esquema iterativo:

$$\mathbf{x}^{(k+1)} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}, \quad k = 0, 1, \dots, \quad (4)$$

Observación:

La condición de convergencia $\rho((\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}) < 1$ es satisfecha para matrices diagonal dominantes y definidas positivas.



El método Gauss-Seidel puede ser **inaceptablemente lento**. El método de **sobrerrelajación sucesiva (SOR)** está basado en la siguiente identidad

$$(D + \omega L)x = -[\omega U + (\omega - 1)D]x + \omega b,$$

donde $\omega > 0$ es llamado **factor de relajación**. Se puede mostrar que SOR converge solamente para valores de $\omega \in (0, 2)$ (ver Kahan, 1958).

De este modo, obtenemos:

$$\begin{aligned} x^{(k+1)} &= (D + \omega L)^{-1}[(\omega - 1)D - \omega U]x^{(k)} + \omega(D + \omega L)^{-1}b, \\ &= M_\omega x^{(k)} + d_\omega, \quad k = 0, 1, \dots, \end{aligned} \tag{5}$$

Observación:

Una buena elección del parámetro ω^3 puede acelerar la convergencia del método⁴ (Young, 1954; Hadjidimos, 2000).

³ $\omega_{\text{opt}} = 2/(1 + \rho(B))$, con $B = -(L + D)^{-1}U$.

⁴Un radio espectral menor, indica convergencia más rápida



Método de gradientes conjugados

Resolver el sistema lineal $\mathbf{Ax} = \mathbf{b}$ es equivalente a determinar el **mínimo de la función**:

$$\phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top \mathbf{Ax} - \mathbf{x}^\top \mathbf{b}.$$

En efecto, por hacer la derivada de ϕ igual a $\mathbf{0}$, vemos que el punto estacionario de ϕ ocurre en el punto \mathbf{x} donde $\mathbf{Ax} = \mathbf{b}$.

- ▶ Este procedimiento asume que la matriz \mathbf{A} es **simétrica y definida positiva**.
- ▶ En este caso, el (único) mínimo de ϕ ocurre en $\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b}$ y es dado por $-\frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b}$.

El mínimo puede ser encontrado de forma iterativa, usando el siguiente esquema:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \mathbf{p}^{(k)}, \quad r = 0, 1, \dots, \quad (6)$$

donde $\lambda^{(k)}$ es un escalar (que representa un **largo de paso**), mientras que $\mathbf{p}^{(k)}$ es un vector que indica la **dirección de búsqueda**.



Método de gradientes conjugados

Resolver el sistema lineal $\mathbf{Ax} = \mathbf{b}$ es equivalente a determinar el **mínimo de la función**:

$$\phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top \mathbf{Ax} - \mathbf{x}^\top \mathbf{b}.$$

En efecto, por hacer la derivada de ϕ igual a $\mathbf{0}$, vemos que el punto estacionario de ϕ ocurre en el punto \mathbf{x} donde $\mathbf{Ax} = \mathbf{b}$.

- ▶ Este procedimiento asume que la matriz \mathbf{A} es **simétrica y definida positiva**.
- ▶ En este caso, el (único) mínimo de ϕ ocurre en $\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b}$ y es dado por $-\frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b}$.

El mínimo puede ser encontrado de forma iterativa, usando el siguiente esquema:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \mathbf{p}^{(k)}, \quad r = 0, 1, \dots, \quad (6)$$

donde $\lambda^{(k)}$ es un escalar (que representa un **largo de paso**), mientras que $\mathbf{p}^{(k)}$ es un vector que indica la **dirección de búsqueda**.



Método gradientes conjugados

Note que

$$d_x \phi(\mathbf{x}) = \frac{1}{2} (d\mathbf{x})^\top \mathbf{A} \mathbf{x} - (d\mathbf{x})^\top \mathbf{b} = (d\mathbf{x})^\top (\mathbf{A} \mathbf{x} - \mathbf{b}),$$

es decir,

$$\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A} \mathbf{x} - \mathbf{b}, \quad -\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} := \mathbf{r}$$

es un residuo.

Definición 1

Un conjunto de vectores $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_k\}$ se dicen **conjugados**⁵ con respecto a la matriz simétrica y definida positiva \mathbf{A} si

$$\mathbf{z}_i^\top \mathbf{A} \mathbf{z}_j = 0, \quad \forall i \neq j.$$

Un conjunto de vectores satisfaciendo esta propiedad también son **linealmente independientes**.

⁵también son dichos \mathbf{A} -conjugados.



Método gradientes conjugados

Ahora, note que

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + \lambda^{(1)}\mathbf{p}^{(1)} + \dots + \lambda^{(k-1)}\mathbf{p}^{(k-1)}.$$

Escogiendo $\lambda^{(k)}$ como

$$\lambda^{(k)} = \frac{\mathbf{r}^{(k)\top} \mathbf{p}^{(k)}}{\mathbf{p}^{(k)\top} \mathbf{A} \mathbf{p}^{(k)}},$$

obtenemos que la dirección de búsqueda $\mathbf{p}^{(k)}$ es \mathbf{A} -conjugada con $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(k-1)}$.

- ▶ La secuencia $\{\mathbf{x}^{(k)}\}$ definida por el [algoritmo gradiente conjugado](#) (6) converge a la solución $\hat{\mathbf{x}}$ en a lo más n pasos.
- ▶ El algoritmo puede ser implementado muy eficientemente y por tanto es una opción atractiva para [problemas de gran tamaño](#).



Algoritmo 1: Gradientes conjugados.

Entrada : Estimación inicial $\mathbf{x}^{(0)}$

Parámetros: Tolerancia τ .

```
1 begin
2    $k \leftarrow 0$ 
3   Calcular  $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$ ,  $\mathbf{p}^{(k)} = \mathbf{r}^{(k)}$  y  $\gamma^{(k)} = \|\mathbf{r}^{(k)}\|^2$ 
4   while  $\gamma^{(k)} > \tau$  do
5      $\lambda^{(k)} = \frac{\mathbf{r}^{(k)\top} \mathbf{p}^{(k)}}{\mathbf{p}^{(k)\top} \mathbf{A} \mathbf{p}^{(k)}}$ 
6      $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \mathbf{p}^{(k)}$ 
7      $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)}$ 
8      $\beta^{(k)} = \frac{\mathbf{r}^{(k+1)\top} \mathbf{p}^{(k)}}{\mathbf{p}^{(k)\top} \mathbf{A} \mathbf{p}^{(k)}}$ 
9      $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta^{(k)} \mathbf{p}^{(k)}$  y  $\gamma^{(k)} = \|\mathbf{r}^{(k)}\|^2$ 
10    Hacer  $k \leftarrow k + 1$  y  $\mathbf{r}^{(k)} \leftarrow \mathbf{r}^{(k+1)}$ ,  $\mathbf{p}^{(k)} \leftarrow \mathbf{p}^{(k+1)}$ 
11  end
12  return  $\hat{\mathbf{x}} = \mathbf{x}^{(k)}$ 
13 end
```



MAT-468: Sesión 3, Cálculos en regresión

Felipe Osorio

<http://fosorios.mat.utfsm.cl>

Departamento de Matemática, UTFSM



Antes de describir una variante para obtener la [descomposición ortogonal-triangular \(QR\)](#) de una matriz es conveniente revisar algunas propiedades fundamentales de las matrices ortogonales:

- ▶ $QQ^T = Q^T Q = I$.
- ▶ $\langle Qx, Qy \rangle = x^T Q^T Q y = x^T y = \langle x, y \rangle$.
- ▶ $\|Qx\| = \|x\|$.
- ▶ Si $B = Q^T A Q$, entonces A y B tienen los mismos valores propios para Q matriz ortogonal.

Existen diversas variantes del algoritmo para implementar la [descomposición QR](#). A continuación veremos una basada en [transformaciones Householder](#).



Problema 1

Para $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x} \neq \mathbf{0}$, hallar una **matriz ortogonal** $M \in \mathbb{R}^{p \times p}$ tal que

$$M^T \mathbf{x} = \|\mathbf{x}\| \mathbf{e}_1,$$

donde $\mathbf{e}_1 = (1, 0, \dots, 0)^T$ denota el primer **vector unidad**.

Definición 1 (Reflexión)

Sea \mathbf{u} y \mathbf{v} **vectores ortonormales** y \mathbf{x} vector generado por \mathbf{u} y \mathbf{v} . Entonces

$$\mathbf{x} = c_1 \mathbf{u} + c_2 \mathbf{v},$$

para escalares c_1, c_2 . El vector

$$\tilde{\mathbf{x}} = -c_1 \mathbf{u} + c_2 \mathbf{v},$$

el llamado una **reflexión** de \mathbf{x} a través de la línea definida por el vector \mathbf{v} (o \mathbf{u}^\perp).



Definición 2 (Transformación Householder)

Sea $\mathbf{x} = c_1\mathbf{u} + c_2\mathbf{v}$, con \mathbf{u} y \mathbf{v} vectores generadores de \mathbf{x} y considere la matriz

$$\mathbf{H} = \mathbf{I} - \lambda\mathbf{u}\mathbf{u}^\top, \quad \lambda = 2/\mathbf{u}^\top\mathbf{u}.$$

Note que $\mathbf{H}\mathbf{x} = \tilde{\mathbf{x}}$, es decir \mathbf{H} es un reflector.

La **transformación Householder** satisface las siguientes propiedades:

- ▶ $\mathbf{H}\mathbf{u} = -\mathbf{u}$.
- ▶ $\mathbf{H}\mathbf{v} = \mathbf{v}$ para cualquier \mathbf{v} ortogonal a \mathbf{u} .
- ▶ $\mathbf{H}^\top = \mathbf{H}$.
- ▶ $\mathbf{H}^{-1} = \mathbf{H}^\top$.



Considere $\mathbf{u} = \mathbf{x} + \delta \mathbf{e}_1$, donde $\delta^2 = \mathbf{x}^\top \mathbf{x}$. Es fácil notar que,

$$\begin{aligned}\mathbf{u}^\top \mathbf{u} &= (\mathbf{x} + \delta \mathbf{e}_1)^\top (\mathbf{x} + \delta \mathbf{e}_1) = \mathbf{x}^\top \mathbf{x} + 2\delta \mathbf{x}^\top \mathbf{e}_1 + \delta^2 \mathbf{e}_1^\top \mathbf{e}_1 \\ &= \delta^2 + 2\delta x_1 + \delta^2 = 2(\delta^2 + \delta x_1),\end{aligned}$$

así

$$\lambda = \frac{2}{\|\mathbf{u}\|^2} = \frac{1}{\delta^2 + \delta x_1}.$$

Además,

$$\mathbf{u}^\top \mathbf{x} = (\mathbf{x} + \delta \mathbf{e}_1)^\top \mathbf{x} = \mathbf{x}^\top \mathbf{x} + \delta \mathbf{e}_1^\top \mathbf{x} = \delta^2 + \delta x_1.$$

De este modo, \mathbf{H} satisface:¹

$$\begin{aligned}\mathbf{H}\mathbf{x} &= (\mathbf{I} - \lambda \mathbf{u}\mathbf{u}^\top)\mathbf{x} = \mathbf{x} - \lambda(\mathbf{u}^\top \mathbf{x})\mathbf{u} \\ &= \mathbf{x} - \lambda(\delta^2 + \delta x_1)(\mathbf{x} + \delta \mathbf{e}_1) = \mathbf{x} - \mathbf{x} - \delta \mathbf{e}_1 \\ &= -\delta \mathbf{e}_1.\end{aligned}$$

¹Note que $\mathbf{H}\mathbf{x}$ puede ser obtenido usando un `axpy`.



Descomposición QR

La **descomposición QR** de una matriz $A \in \mathbb{R}^{n \times p}$ ($n > p$), puede ser construída a través de una **secuencia de matrices** Q_1, \dots, Q_p tales que

$$Q_p \cdots Q_1 A = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

donde **todas** Q_1, \dots, Q_p son ortogonales. De este modo,

$$A = Q_1^\top \cdots Q_p^\top \begin{pmatrix} R \\ 0 \end{pmatrix} = Q \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

A continuación se describe el algoritmo para obtener la **descomposición QR** usando **transformaciones Householder**.² Sea $M(x)$ la matriz ortogonal desde el **Problema 1** basada en un vector x .

²Otro método popular para obtener la descomposición QR es usando rotaciones Givens.



Algoritmo 1: Descomposición QR

Entrada: Matriz $A \in \mathbb{R}^{n \times p}$.

Salida : Factores Q y R , matrices ortogonal y triangular superior, respectivamente.

```
1 begin
2   Hacer  $Q = I_n$  y  $R = A$ 
3   for  $i = 1$  to  $p$  do
4      $x = (R_{1i}, \dots, R_{pi})^\top$ 
5      $Q_i = \begin{pmatrix} I_{i-1} & \mathbf{0} \\ \mathbf{0} & M(x) \end{pmatrix}$ 
6     /*  $M(x)$  obtenido usando reflexiones Householder */
7      $Q = Q_i Q$ 
8      $R = Q_i R$ 
9   end
10   $Q = Q^\top$ 
11   $R = (R_{ij})$  para  $i, j = 1, \dots, p$ .
end
```



Aplicación en regresión lineal

Sea el **modelo de regresión lineal**:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\mathbf{X} \in \mathbb{R}^{n \times p}$ con $\text{rg}(\mathbf{X}) = p$ y $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$. El **estimador mínimos cuadrados (LS)** de $\boldsymbol{\beta}$ es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \quad \text{con} \quad \text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Además,

$$s^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2.$$

Adicionalmente, si $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$, entonces³

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}), \\ \frac{(n-p)s^2}{\sigma^2} &\sim \chi^2(n-p). \end{aligned}$$

³En cuyo caso, $\hat{\boldsymbol{\beta}}$ corresponde al estimador ML de $\boldsymbol{\beta}$.



Aplicación en regresión lineal

Las ecuaciones de estimación para obtener $\hat{\beta}$ son $\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$, o bien

$$\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{Y}. \quad (1)$$

Así, podemos resolver (1) usando la **descomposición Cholesky**, de

$$\mathbf{X}^\top \mathbf{X} = \mathbf{U}^\top \mathbf{U},$$

con \mathbf{U} matrix triangular superior. De este modo:

$$\mathbf{U}^\top \mathbf{z} = \mathbf{X}^\top \mathbf{Y}, \quad \text{y} \quad \mathbf{U} \hat{\beta} = \mathbf{z},$$

para obtener s^2 considere

$$RSS = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \mathbf{Y}^\top \mathbf{Y} - \mathbf{z}^\top \mathbf{z}.$$

Invirtiendo \mathbf{U} (in-place)⁴, podemos hacer

$$\mathbf{U}^{-1} \mathbf{U}^{-\top} = (\mathbf{X}^\top \mathbf{X})^{-1}.$$

⁴Haciendo $\mathbf{U} \leftarrow \mathbf{U}^{-1}$, tenemos $(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{U} \mathbf{U}^\top$



Aplicación en regresión lineal

Considere

$$\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n \times (p+1)},$$

luego

$$\mathbf{Z}^\top \mathbf{Z} = \begin{pmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Y} \\ \mathbf{Y}^\top \mathbf{X} & \mathbf{Y}^\top \mathbf{Y} \end{pmatrix} \in \mathbb{R}^{(p+1) \times (p+1)}.$$

Aplicando el **operador Sweep** sobre las **primeros p** elementos diagonales de $\mathbf{Z}^\top \mathbf{Z}$, obtenemos:

$$\begin{aligned} \mathbf{B} &= \prod_{i=1}^p \text{Sweep}(i) \mathbf{Z}^\top \mathbf{Z} \\ &= \begin{pmatrix} (\mathbf{X}^\top \mathbf{X})^{-1} & (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ -\mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} & \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{X}^\top \mathbf{X})^{-1} & \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}}^\top & RSS \end{pmatrix}. \end{aligned}$$



Considere la **descomposición QR** (Ortogonal-Triangular) de X , como:

$$X = QR, \quad R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

con $R_1 \in \mathbb{R}^{p \times p}$ matriz triangular superior ($n > p$). Si $\text{rg}(X) = p$, entonces R_1 es no singular. Además, considere la transformación:

$$Q^T Y = c, \quad c = (c_1^T, c_2^T)^T.$$

El **estimador LS** minimiza la función objetivo:

$$\begin{aligned} \|Y - X\beta\|^2 &= \|Q^T(Y - X\beta)\|^2 = \|Q^T Y - Q^T QR\beta\|^2 \\ &= \|c - R\beta\|^2, \end{aligned}$$

Es fácil notar que:

$$\|c - R\beta\|^2 = \|c_1 - R_1\beta\|^2 + \|c_2\|^2.$$



Finalmente, el **estimador de mínimos cuadrados** $\hat{\beta}$ está dado por la solución del sistema triangular:

$$\mathbf{R}_1 \hat{\beta} = \mathbf{c}_1.$$

El mínimo de la función objetivo está dado por $\|\mathbf{c}_2\|^2$. Note además que

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{c}_2\|^2,$$

corresponde al estimador máximo verosímil para σ^2 .⁵ Por otro lado,

$$\mathbf{R}_1^{-1} \mathbf{R}_1^{-\top} = (\mathbf{X}^\top \mathbf{X})^{-1},$$

permite obtener $\text{Cov}(\hat{\beta})$.

⁵bajo el supuesto de normalidad.



Considere la **descomposición valor singular (SVD)** de X ,

$$X = UDV^T,$$

donde $U \in \mathbb{R}^{n \times p}$ tal que $U^T U = I_p$, $D = \text{diag}(d_1, \dots, d_p)$ y V es matriz ortogonal $p \times p$. De este modo, podemos escribir el modelo:

$$Y = X\beta + \epsilon = UD\alpha + \epsilon,$$

con $\alpha = V^T \beta$. Haciendo $Z = U^T Y$, tenemos el modelo en **forma canónica**:

$$Z = D\alpha + \delta, \quad \delta = U^T \epsilon,$$

donde $E(\delta) = \mathbf{0}$ y $\text{Cov}(\delta) = \sigma^2 U^T U = \sigma^2 I_p$.



El **estimador LS** de α en el modelo canónico es:

$$\hat{\alpha} = D^{-1}Z, \quad \Rightarrow \quad \hat{\beta} = V\hat{\alpha}.$$

Además,

$$\|Y - X\hat{\beta}\|^2 = \|Y - UDV^T\hat{\beta}\|^2 = \|Z - D\hat{\alpha}\|^2.$$

Finalmente,

$$(X^T X)^{-1} = (VD^2V^T)^{-1} = VD^{-2}V^T.$$

