## 4.4 Aplicación de la factorización QR en Regresion

Consideremos el modelo de regresión lineal multiple

$$\mathbf{Y=XB+e}$$

Donde **Y** es un vector n dimensional de observaciones de la variable de respuesta,
**X** es una matriz n x (p+1) de observaciones de p variables predictoras, incluyendo una primera columna de unos (si se desea un modelo con intercepto),
**B** es un vector de dimensión p+1 de parámetros a estimar, y
**e** es un vector n dimensional de errores aleatorios. Para estimar **B** usando la técnica de cuadrados mínimos, hay que minimizar la suma de cuadrados de los errores. Es decir,

$$\|\mathbf{XB}-\mathbf{y}\|_2^2 \qquad (1)$$

Sea Q una matriz ortogonal cuadrada de orden n y **x** un vector n dimensional entonces

$$\|Q'\mathbf{x}\|_2^2=(Q'\mathbf{x})'(Q'\mathbf{x})=\mathbf{x}'QQ'\mathbf{x}=x'x=\|\mathbf{x}\|_2^2$$

Luego (1) puede ser escrita como

$$\|\mathbf{XB}-\mathbf{y}\|_2^2= \quad \|Q'\mathbf{XB}-Q'\mathbf{y}\|_2^2 \qquad (2)$$

Por otro lado, X puede ser factorizada como **X=QR**, donde Q es una matriz orthogonal n por n y R es de orden n por (p+1) y de la forma

$$R=\begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

donde $\mathbf{R_1}$ es una matriz triangular superior de orden p+1 por p+1 y **0** es una matriz de ceros n-p-1 por p+1. Haciendo

$$Q'y=\begin{pmatrix} c \\ d \end{pmatrix}$$

La ecuación (2) puede ser escrita como

$$\left\|\begin{pmatrix} R_1B \\ 0 \end{pmatrix}-\begin{pmatrix} c \\ d \end{pmatrix}\right\|_2^2=\left\|\begin{pmatrix} R_1B-c \\ -d \end{pmatrix}\right\|_2^2=\left((R_1B-c)' \quad -d'\right)\begin{pmatrix} R_1B-c \\ -d \end{pmatrix}=\|R_1B-c\|_2^2+\|d\|_2^2$$

En consecuencia, (1) sera mínimo con respecto a B cuando B es tal que

$$R_1B=c$$

Y la suma de cuadrados residual estará dado por

$$SSE=\|d\|_2^2$$

La matriz $(X'X)^{-1}$ puede ser calculada como $(R'R)^{-1}=R^{-1}(R')^{-1}$

**Algoritmo para calculos de regresión usando QR-Householder**

1. Aplicar Householder a la matriz X para obtener la matriz triangular superior $R_1$ y los vectores $u_1, u_2, \ldots, u_n$ que definen las matrices Householder $H_1, H_2, \ldots H_n$ respectivamente.
2. Fornar $H_n H_{n-1} \ldots H_2 H_1 y = \begin{pmatrix} c \\ d \end{pmatrix}$
3. Resolver $R_1 B = c$

La siguiente funcion lshouse en Matlab calcula el estimador minimo cuadrático de B

```
function [b,c,d] = lshouse(X,y);
%LSHOUSE Resuelve el problema de regresion usando
Householder QR
%b = lshouse(X,y) calcula el estimado del vector de
%parametros B
% del modelo de regresion Y=XB +e
%c y d son las componentes de Q'y
%Este  programa llama a los programs BACKSUB, HOUSECERO and
PMULHOUSE.
%input  : Matrix X and vector y
%output : vector b

        [m,n] = size(X);
        b= zeros(n,1);
        s= min(n,m-1);
        for k = 1 : s
          [u,sigma] = housecero(X(k:m,k));
          X(k:m,k:n) = pmulhouse(X(k:m,k:n),u);
          y(k:m) = pmulhouse(y(k:m),u);
        end;
        r1 =  X;
        ran = rank(r1);
        c=y(1:ran)
        d=y(ran+1:m)
        R = r1(1:ran,:);
        b = backsub(R,c);
```

**La función pmulhouse es como sigue:**

```
function A = pmulhouse(A,u)
%PMULHOUSE Pre Multiplicacion por una matriz Householder
%A = pmulhouse(A,u) calcula la pre-multiplicacion
%de una matriz A por la matriz Householder generada
%por el vector u. La matriz  output  A contiene el producto
```

```
%HA.
%input   : Matriz A y vector u
%output  : Matriz A

   [m,n] = size(A);
        beta = 2/(u'*u);
        for j = 1 : n
          alpha = 0;
          alpha =  alpha + u(1:m)'*A(1:m,j);
          alpha =  beta * alpha;
          A(1:m,j) = A(1:m,j) - alpha * u(1:m);
         end;
```

Ejemplo: Hallar El modelo de regression lineal para el siguiente conjunto de datos

| Y1 | X1 | X2 |
|---|---|---|
| 4 | 3 | 4 |
| 7 | 6 | 9 |
| 8 | 9 | 12 |
| 12 | 8 | 18 |
| 15 | 9 | 15 |
| 21 | 12 | 23 |

Solucion:

```
» X=[1 3 4; 1 6 9;1 9 12; 1 8 18; 1 9 15; 1 12 23]

X =

     1     3     4
     1     6     9
     1     9    12
     1     8    18
     1     9    15
     1    12    23

» y=[4 7 8 12 15 21]'

y =

     4
     7
     8
    12
    15
```

```
    21

» [b,c,d]=lshouse(X,y)


c =

  -27.3526
   12.2988
    4.1807


d =

   -4.1730
    2.0945
    0.5426


b =

   -1.2772
    0.3675
    0.7085

» Calculo de la suma de cuadrados total

» y'*y-6*mean(y)^2

ans =

  190.8333

» calculo de la suma de cuadrados residual

» norm(d)^2

ans =

   22.0948
```

**Ejemplo 2**: Hallar los coeficientes del modelo de regresión y la suma de cuadrados del conjunto de datos **homedat**

**Solucion**

```
» addpath c:\matlab\acuna
» load c:\matlab\acuna\homedat.txt

» y=homedat(:, 1)

y =

        2050
        2150
        2150
        1999
        1900
        1800
        1560
        1449
        1375
        1270
        1250
        1235
        1170
        1155
        1110
        1139
         995
         900
         960
        1695
        1553
        1020
        1020
         850
         720
         749
        2150
        1350
        1299
        1250
        1239
        1125
        1080
        1050
        1049
         934
         875
         805
```

```
         759
         729
         710
         975
         939
        2100
         580
        1844
         699
        1160
        1109
        1129
        1050
        1045
        1050
        1020
        1000
        1030
         975
         940
         920
         945
         874
         872
         870
         869
         766
         739

» size(homedat)

ans =

    66     8

» homedat(:,1)=ones(66,1)


» [b,c,d]=lshouse(homedat,y)

c =

  1.0e+003 *

   -9.4938
   -2.8818
```

```
   -0.4350
    0.0984
    0.0292
    0.5566
   -0.3164
    0.5001


d =

   46.0017
   40.4566
  -85.2902
 -302.7359
 -162.7525
   52.9118
  103.9650
  161.6211
   21.5400
   55.7500
  -66.6844
   55.6660
  176.3852
  212.3979
  108.0963
  104.9959
   78.0572
 -270.9350
   46.8564
 -236.4600
 -553.5367
 -106.2789
  152.2594
 -187.0408
 -316.9902
  -57.2189
  -37.8992
  -81.5145
  157.4806
  -45.1424
  107.4597
   45.8426
   17.1765
 -112.3433
   70.5644
  476.6995
  -29.5058
```

```
  292.9683
 -206.5248
   60.3407
   31.3890
  103.6587
   21.6429
   35.8144
 -162.3750
   -8.6055
  129.1937
  104.6487
 -125.4896
  113.5470
  -57.0026
  -62.9529
   38.6287
   42.1136
   52.6013
   70.5756
   50.6687
  155.6671


b =

   92.7448
    0.3522
   -0.5651
    4.3896
  -17.3853
  174.9411
  -73.5823
    0.4989
```

**Calculo de la suma de cuadrados residual**

```
» sse=norm(d)^2

sse =

  1.4641e+006
```

**Calculo de la suma de cuadrados  total**

```
» sst=norm(y-mean(y))^2
```

```
sst =

  1.0629e+007

calculo de la suma de cuadrados de regression

» ssreg=sst-sse

ssreg =

  9.1647e+006

» Calculo del coeficiente de determinacion R^2

» R2=(ssreg/sst)*100

R2=

  86.2251
```