
Modelos Lineales

Table of Contents

Regresión lineal simple (modelo univariante)	1
Crecimiento en naranjos.	1
Ejercicio Propuesto. Modelos lineales con polytool	10
Regresión lineal simple con regress	10
Regresión Robusta.	12
Ejercicio Propuesto: Calor en el proceso de creación de Cemento	12
Transformación de un modelo en un modelo lineal. Regresión Logística	14
Modelo lineal en los parámetros (Regresión no lineal)	15
Modelo de regresión "Teórico"	17
Ejercicio	20
Solución	21
Determinación de la adecuación del modelo de regresión	24

En Matlab podemos realizar el cálculo de los coeficientes con la instrucción `polyfit` (usaremos `polyval` para estimación/predicción). Esta instrucción puede realizar ajustes polinómicos, un caso particular es el ajuste lineal, que nos interesa. En ocasiones, problemas no lineales pueden transformarse en un problema lineal utilizando, en lugar de los valores originales, los valores obtenidos al aplicar una determinada función sobre los mismo. Posteriormente veremos la regresión lineal múltiple, implementada en Matlab con la instrucción `regress`.

Regresión lineal simple (modelo univariante)

Crecimiento en naranjos.

El conjunto de datos a usar se encuentra en el fichero `orange.txt` y contiene 35 filas (observaciones) con 3 columnas que se corresponden con las siguientes variables: a) un indicador de árbol; b) edad, días desde el 31/12/1968 y, por último, c) circunferencia del tronco del árbol. Carga el fichero y determina:

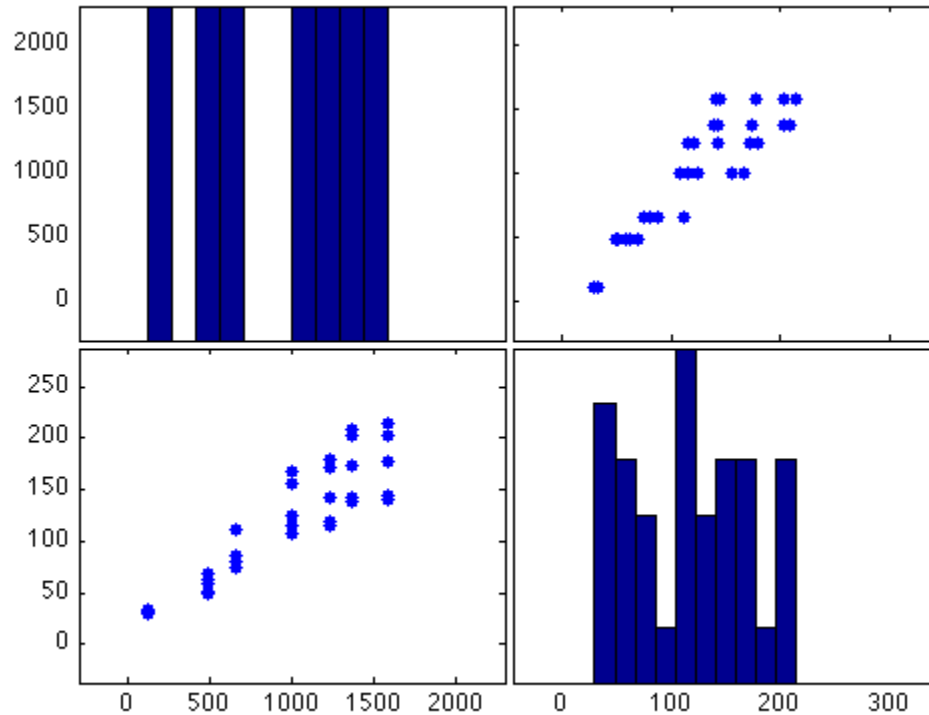
- La posible relación lineal, para cada árbol entre edad y circunferencia (utiliza representaciones gráficas como el diagrama de dispersión y valores numéricos como el coeficiente de correlación para apoyar tu hipótesis).
- Desarrolla un modelo polinómico para la circunferencia del árbol en función de la edad; plantea diferentes órdenes en dicho modelo.
- Analiza los residuos (errores cometidos) de los modelos anteriores (no hace falta que todos!!!) y comprueba, o no, su normalidad (puedes usar histogramas o los gráficos de probabilidad).

```
clear
clc
close all

load orange.txt
x=orange(:,2);
y=orange(:,3);
```

Matriz de dispersión

```
plotmatrix(orange(:,2:3))
```



Matriz de correlación

```
corrcoef(orange(:,2:3))
```

```
ans =
```

```
1.0000    0.9135
0.9135    1.0000
```

```
for N=1:5
```

```
    %N=input('Orden del modelo ? ');
```

```
    [p,s]=polyfit(x,y,N)
```

```
    p
```

```
    yest(:,N) = polyval(p,x,s);
```

```
    residuo(:,N)=y-yest(:,N);
```

```
    %Representamos el ajuste
```

p =

0.1068 17.3997

s =

R: [2x2 double]
 df: 33
 normr: 136.3625

p =

0.1068 17.3997

p =

-0.0000 0.1314 10.2874

s =

R: [3x3 double]
 df: 32
 normr: 135.2841

p =

-0.0000 0.1314 10.2874

p =

-0.0000 0.0001 0.0334 24.2496

s =

R: [4x4 double]
 df: 31
 normr: 132.2515

p =

-0.0000 0.0001 0.0334 24.2496

Warning: Polynomial is badly conditioned. Add points with distinct X values, reduce the degree of the polynomial, or try centering and scaling as described in HELP POLYFIT.

p =

0.0000 -0.0000 0.0004 -0.0934 35.8468

s =

```
R: [5x5 double]
df: 30
normr: 131.7576
```

p =

```
0.0000    -0.0000    0.0004    -0.0934    35.8468
```

Warning: Polynomial is badly conditioned. Add points with distinct X values, reduce the degree of the polynomial, or try centering and scaling as described in HELP POLYFIT.

p =

```
-0.0000    0.0000   -0.0000    0.0041   -1.1788   121.1019
```

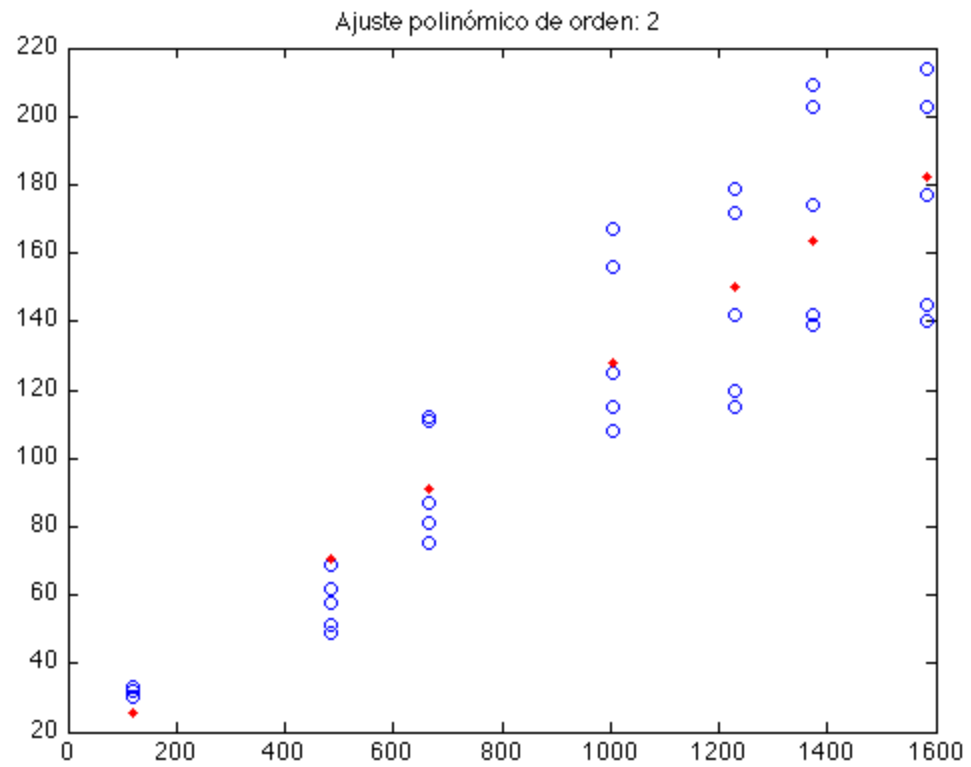
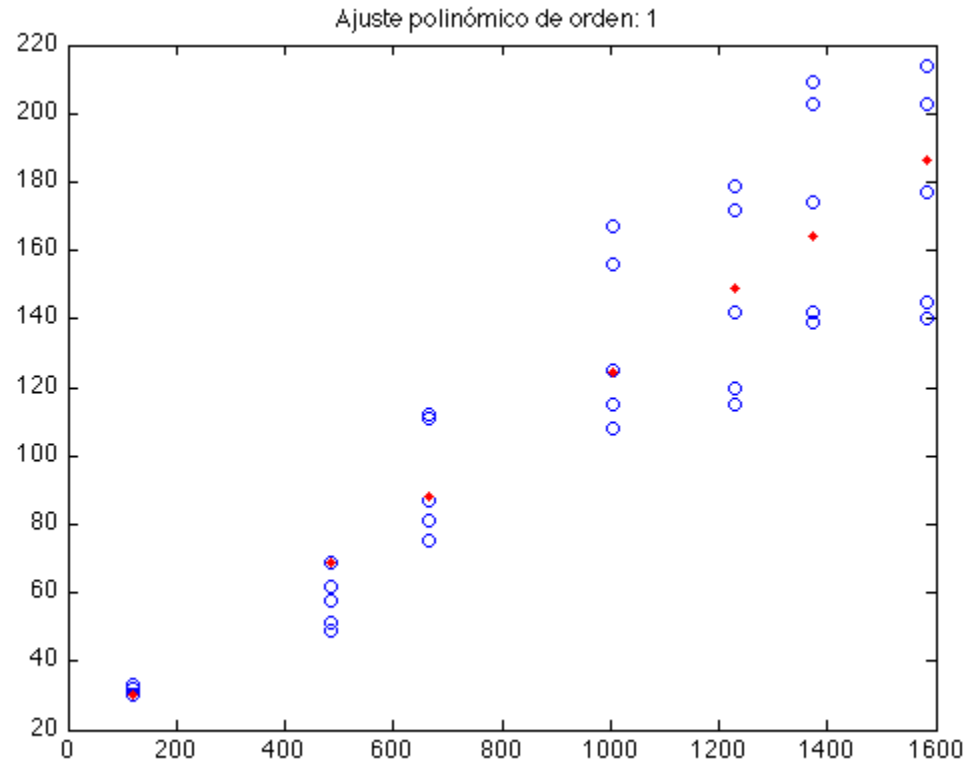
s =

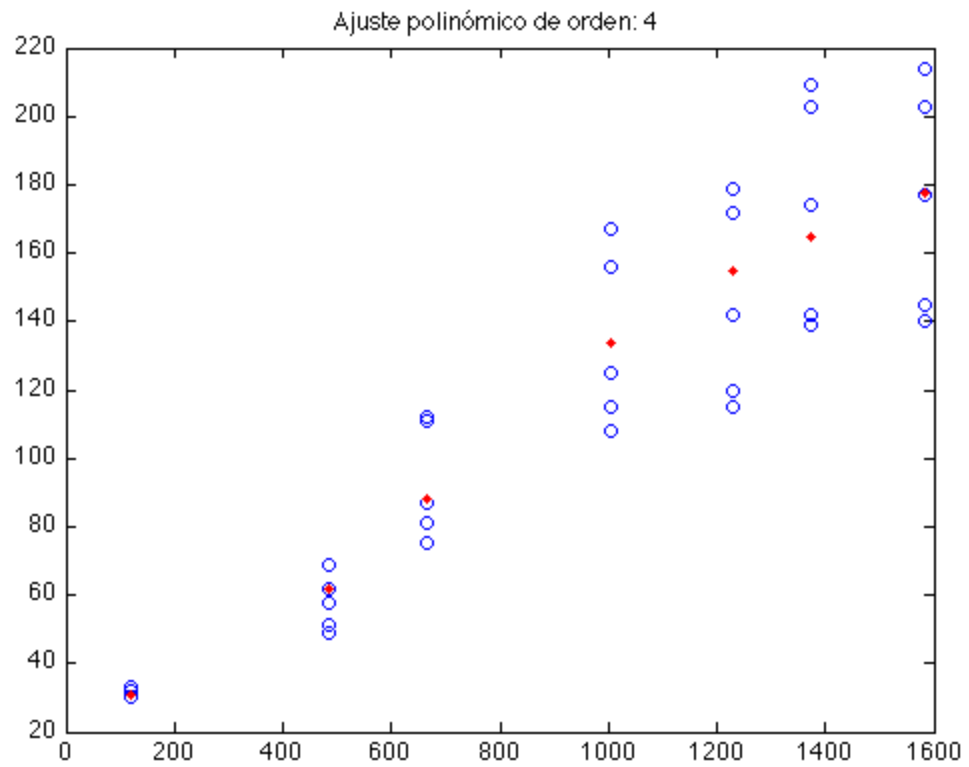
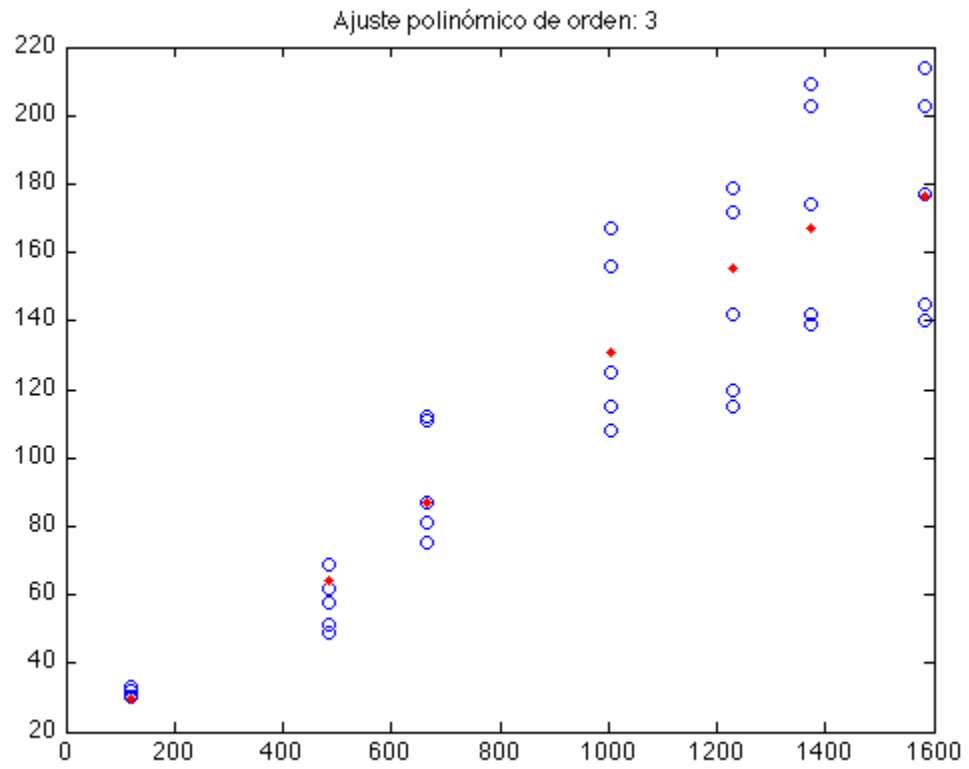
```
R: [6x6 double]
df: 29
normr: 129.1049
```

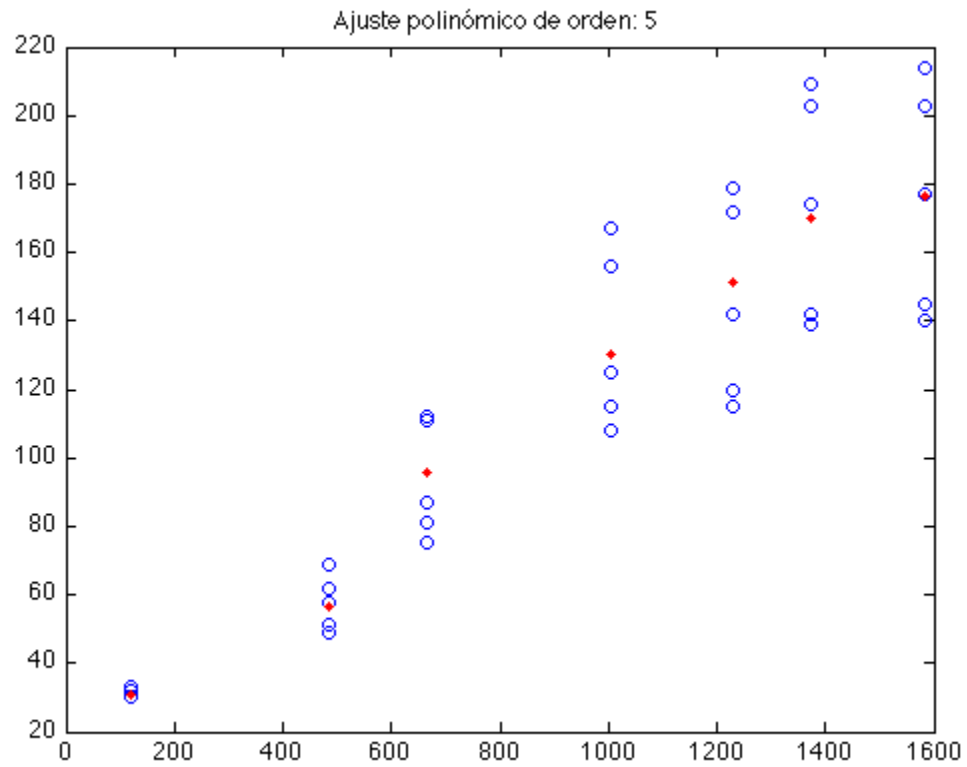
p =

```
-0.0000    0.0000   -0.0000    0.0041   -1.1788   121.1019
```

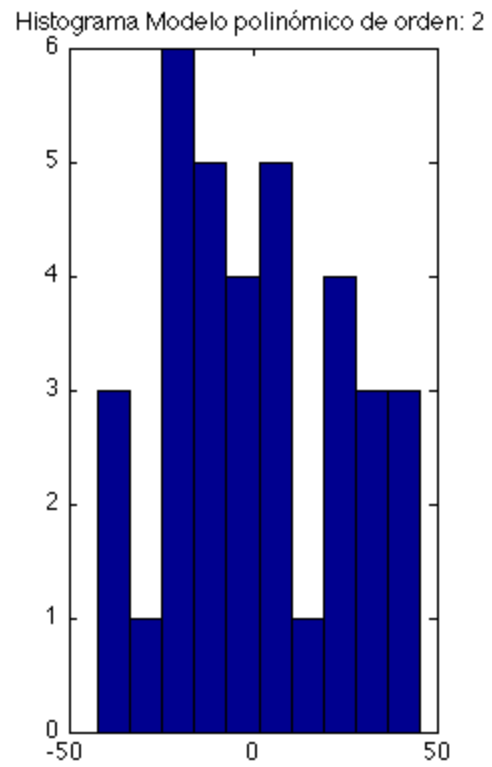
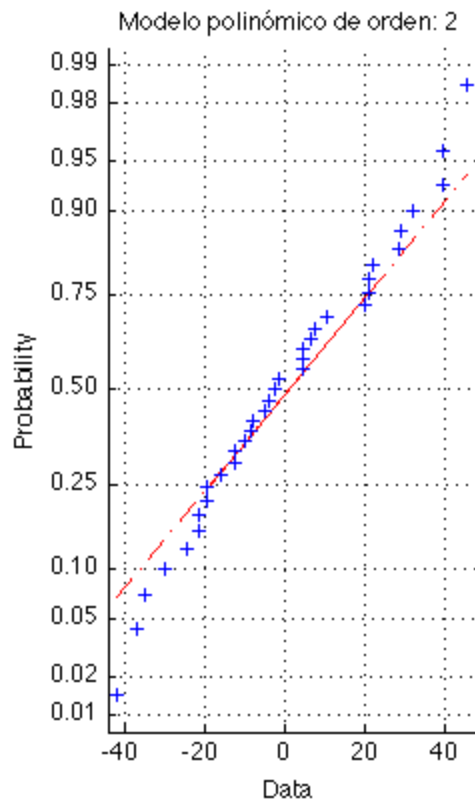
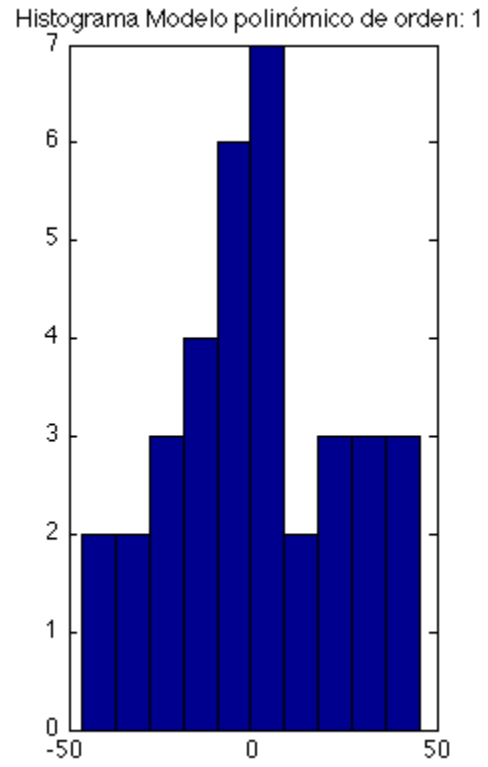
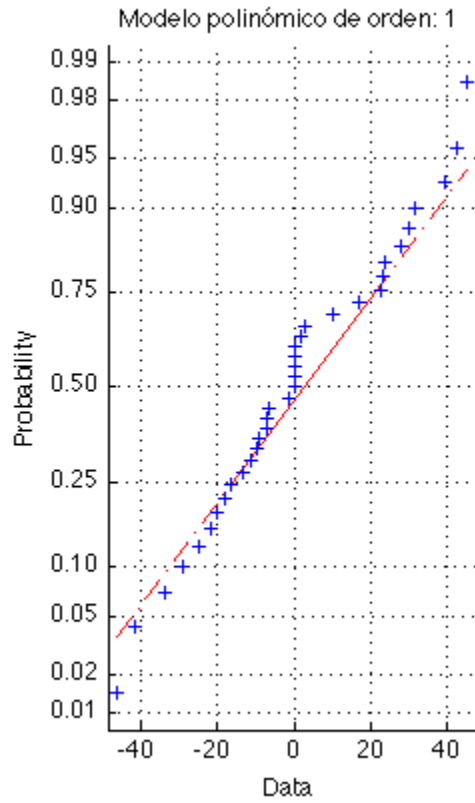
```
figure
plot(x,y,'bo',x,yest(:,N),'r.')
title(['Ajuste polinómico de orden: ' num2str(N)]);
% Analizamos si los residuos siguen una distribución normal
```

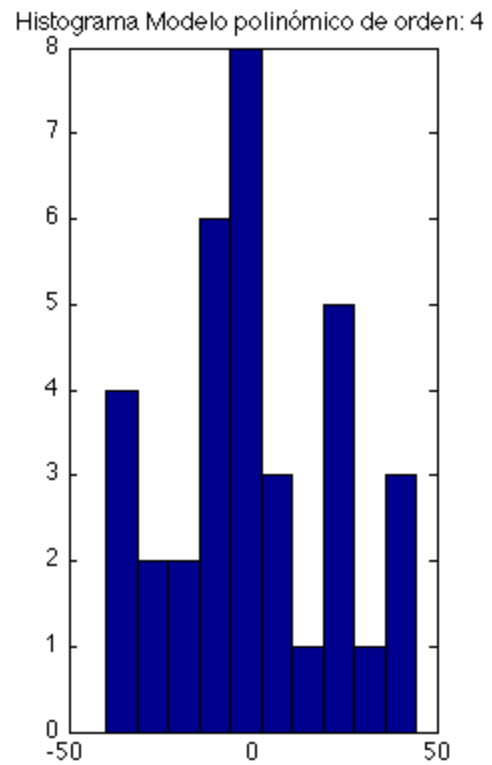
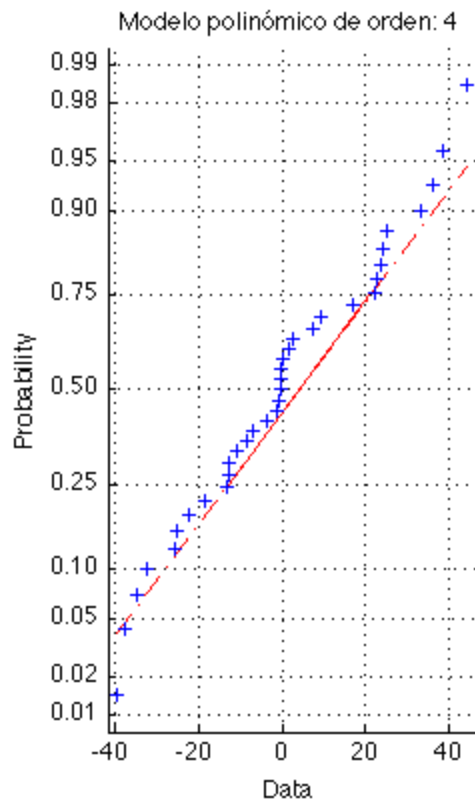
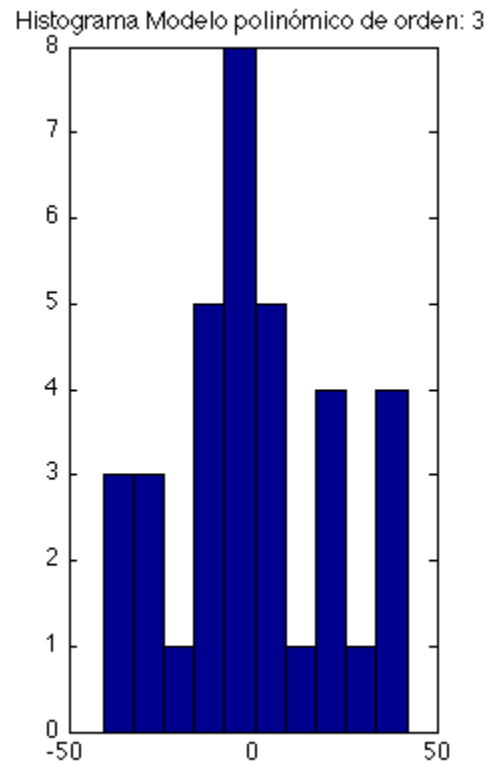
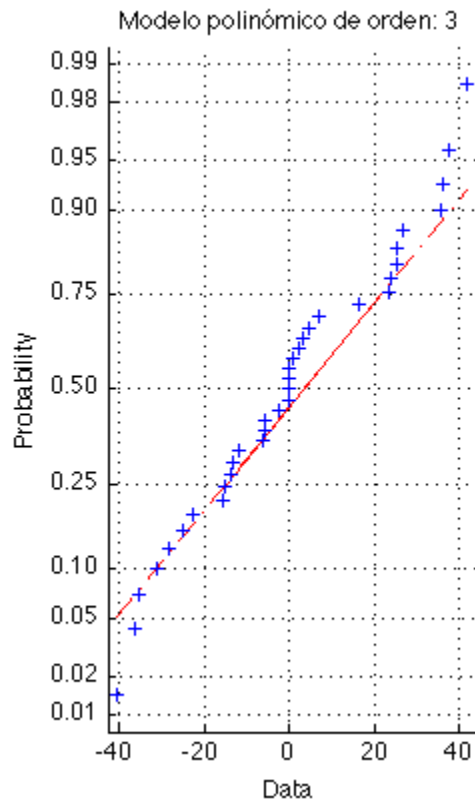


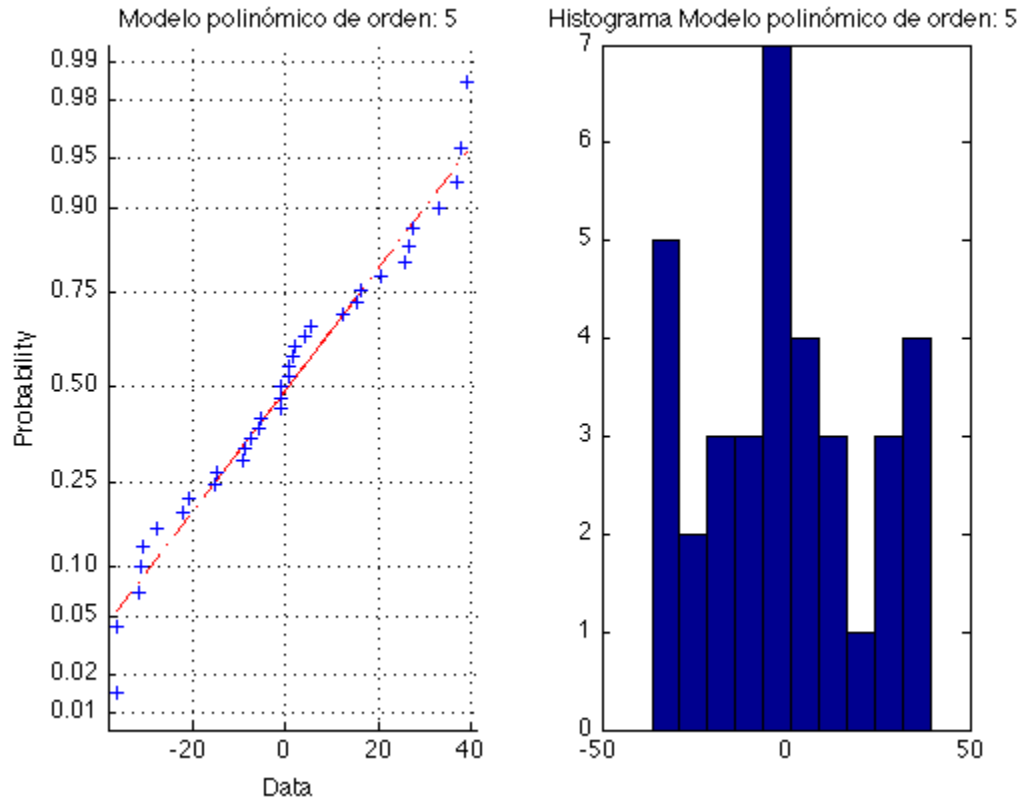




```
figure
subplot(121)
normplot(residuo(:,N))
title(['Modelo polinómico de orden: ' num2str(N)])
subplot(122)
hist(residuo(:,N))
title(['Histograma Modelo polinómico de orden: ' num2str(N)])
%pause
```







end

Ejercicio Propuesto. Modelos lineales con polytool

Utiliza polytool y repite el ejercicio anterior. Prueba el efecto que tiene realizar una regresión robusta en lugar de la habitual

Regresión lineal simple con regress

Dado que regress puede realizar una regresión lineal múltiple, hemos de escribir los valores en forma matricial, siendo la primera columna un vector de unos

```
X=[ones(size(x)), x];
[B,BINT,R,RINT,STATS] = regress(y,X)
```

B =

```
17.3997
0.1068
```

BINT =

```
-0.1433  34.9426
 0.0899  0.1236
```

R =

```
  0.0015
-11.0765
 -1.2951
 -9.5971
-28.8339
-21.8885
-41.3103
  3.0015
 -0.0765
 22.7049
 31.4029
 23.1661
 39.1115
 16.6897
  0.0015
-18.0765
-13.2951
-16.5971
-33.8339
-24.8885
-46.3103
  2.0015
 -7.0765
 23.7049
 42.4029
 30.1661
 45.1115
 27.6897
  0.0015
-20.0765
 -7.2951
  0.4029
 -6.8339
 10.1115
 -9.3103
```

RINT =

```
-46.3392  46.3421
-58.6636  36.5106
-49.4287  46.8384
-57.7911  38.5970
-75.7498  18.0820
-68.9565  25.1794
-85.9026   3.2820
-43.3266  49.3295
-47.8301  47.6771
-24.7332  70.1429
-15.5760  78.3819
-24.1544  70.4865
 -6.4900  84.7129
-29.9278  63.3072
-46.3392  46.3421
-65.3855  29.2325
-61.1928  34.6025
-64.5444  31.3503
-80.3159  12.6480
```

-71.7632	21.9862
-90.2626	-2.3580
-44.3336	48.3365
-54.7622	40.6092
-23.6699	71.0796
-3.4447	88.2506
-16.6414	76.9735
0.2324	89.9905
-18.2455	73.6249
-46.3392	46.3421
-67.2810	27.1280
-55.3594	40.7691
-47.9143	48.7202
-54.8194	41.1515
-37.4715	57.6944
-56.1932	37.5726

STATS =

0.8345 166.4159 0.0000 563.4771

Regresión Robusta.

Ejercicio Propuesto: Calor en el proceso de creación de Cemento

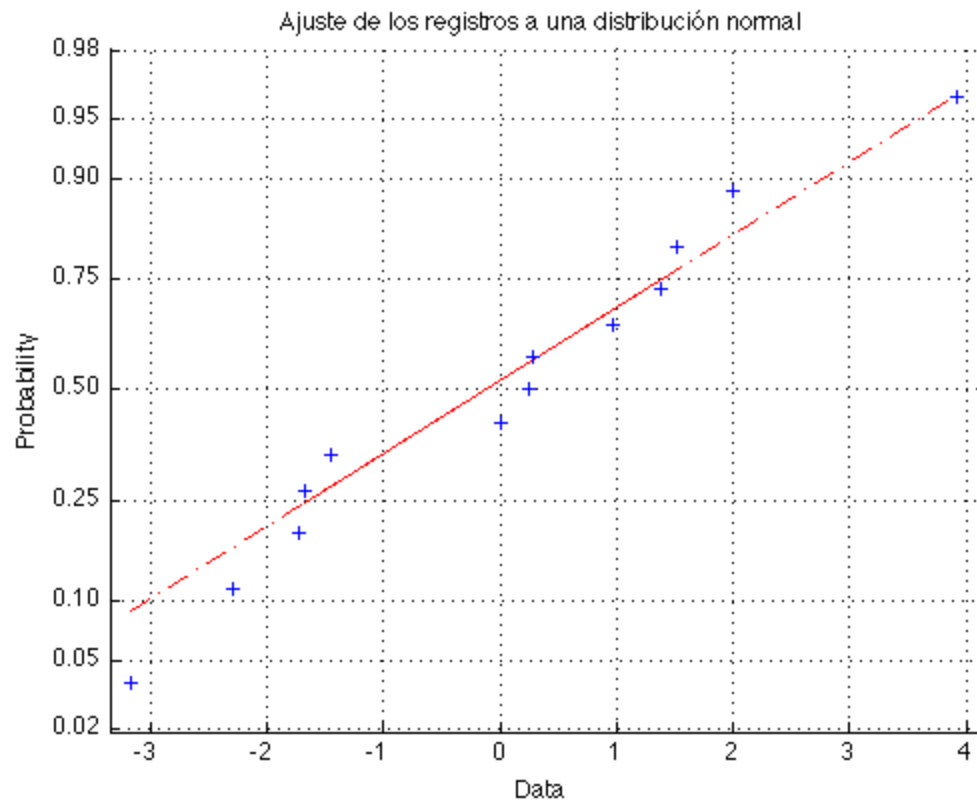
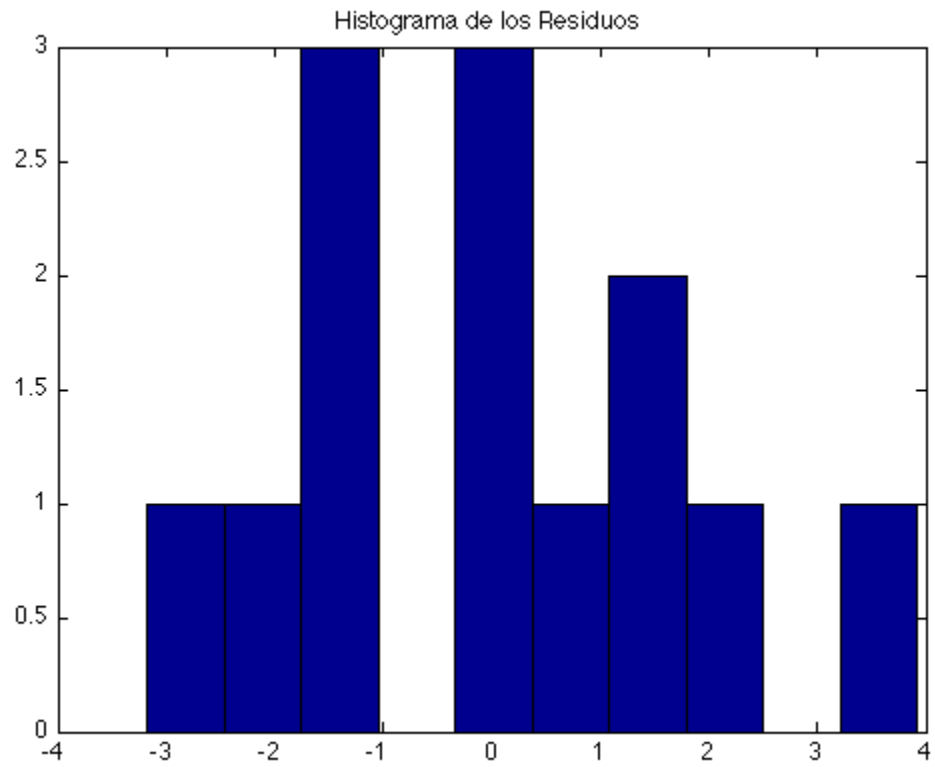
En este problema se plantea el análisis de la cantidad de calor que interviene en el proceso de creación de cemento en función de la proporción de cuatro de sus componentes. Se dispone de un fichero con la composición de 13 cementos diferentes y con el calor necesario en el proceso de creación, fichero cemento.txt. Carga este fichero en MATLAB (las primera cuatro columnas son las proporciones de los componentes y la última se corresponde con el calor) e implementa un modelo lineal multivariante que relacione las proporciones con el calor. Comprueba la bondad y la calidad del modelo obtenido (ayúdate de representaciones gráficas y de los estadísticos de los residuos obtenidos).

```
clear
clc
close all

load cemento.txt

entradas=ones(13,5);
entradas(:,2:5)=cemento(:,1:4); %La primera columna con 1
y=cemento(:,5);
[B,BINT,R,RINT,STATS] = regress(y,entradas);

hist(R);
title('Histograma de los Residuos')
figure
normplot(R)
title('Ajuste de los registros a una distribución normal')
```



Transformación de un modelo en un modelo lineal. Regresión Logística

En este problema se plantea el uso de una transformación para obtener un modelo lineal a partir de uno no lineal. Se conoce que la relación entre un determinado par de variables físicas sigue la relación definida en la siguiente ecuación:

$$y_k = \frac{1}{1 + e^{-K_1 + K_2 \cdot x_k}}$$

Los datos a modelizar se encuentran en el fichero `sigmoide.mat` determina los parametros K_1 y K_2 transformando el modelo de la ecuación anterior en uno lineal. Una vez que tengas dicha transformación determina los parámetros K_1 y K_2 mediante un ajuste de mínimos cuadrados.

```
clear
clc
close all

load sigmoide.mat;

% Realizamos la transformación adecuada para transformarlo en lineal. Esta
% transformación ya la conocemos ya que es la que se utiliza en la Regresión
% Logística.

YT=log(yy./(1-yy));

%%este caso lo podemos resolver con polyfit %%%%%%%%%%%%%%
[p,s]=polyfit(XX,YT,1)
[B,BINT,R,RINT,STATS] = regress(YT',[ones(size(XX')) ,XX'])

p =

    -2.0251    -3.0232

s =

    R: [2x2 double]
    df: 18
    normr: 0.7526

B =

    -3.0232
    -2.0251

BINT =

    -3.1066    -2.9399
    -2.0938    -1.9565

R =

    0.0595
    0.3114
```

```
-0.0415
-0.0687
 0.2234
-0.2388
-0.2331
 0.0175
-0.0502
-0.0144
 0.0632
-0.1140
 0.1541
-0.3949
 0.0743
 0.0295
-0.1557
 0.0511
 0.0873
 0.2400
```

RINT =

```
-0.2852    0.4042
-0.0026    0.6253
-0.3976    0.3146
-0.4281    0.2906
-0.1231    0.5699
-0.5858    0.1082
-0.5837    0.1174
-0.3543    0.3893
-0.4224    0.3221
-0.3880    0.3593
-0.3091    0.4355
-0.4825    0.2546
-0.2094    0.5176
-0.7055   -0.0842
-0.2915    0.4401
-0.3349    0.3939
-0.5080    0.1965
-0.3047    0.4068
-0.2616    0.4362
-0.0838    0.5637
```

STATS =

```
1.0e+03 *
0.0010    3.8415    0.0000    0.0000
```

Modelo lineal en los parámetros (Regresión no lineal)

En este ejemplo vamos a ver como podemos aplicar un modelo de regresión lineal en un problemas en el que hay una dependencia lineal con los parámetros, pero no entre las variables. Se sospecha que en un determinado problema la relación entre dos variables x e y viene dada por la expresión:

$$y_i = \beta + \alpha_1 \cdot x_i + \alpha_2 \cdot \cos(x_i) + \alpha_3 \cdot x_i^2$$

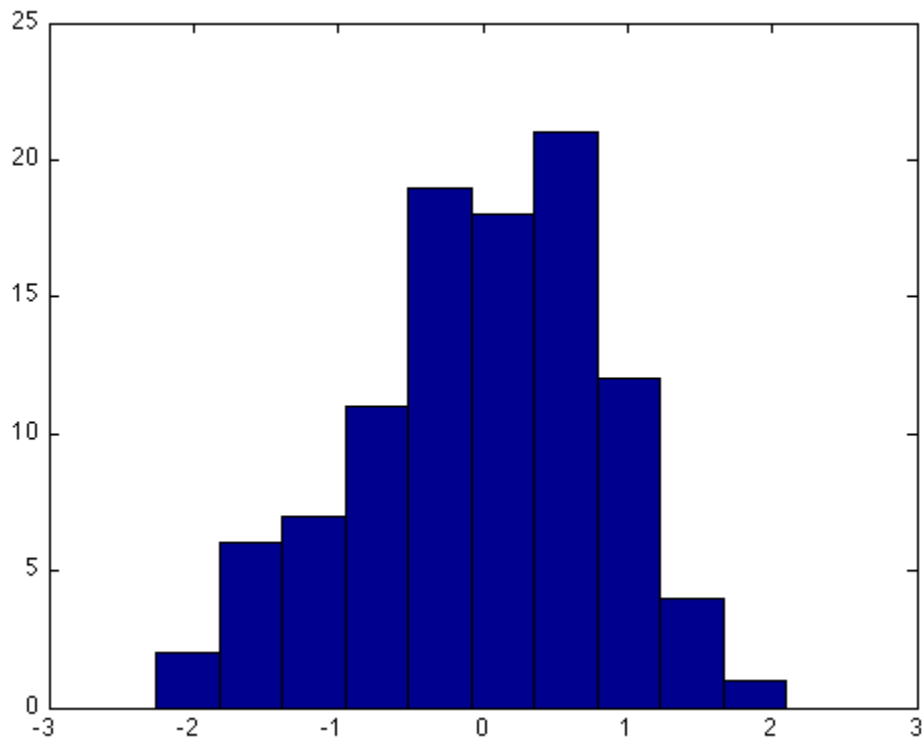
Los pares de datos (x,y) se encuentran en el fichero linealparametros.mat Determina los parámetros del modelo mediante un ajuste de mínimos cuadrados. Para los residuos obtenidos determina:

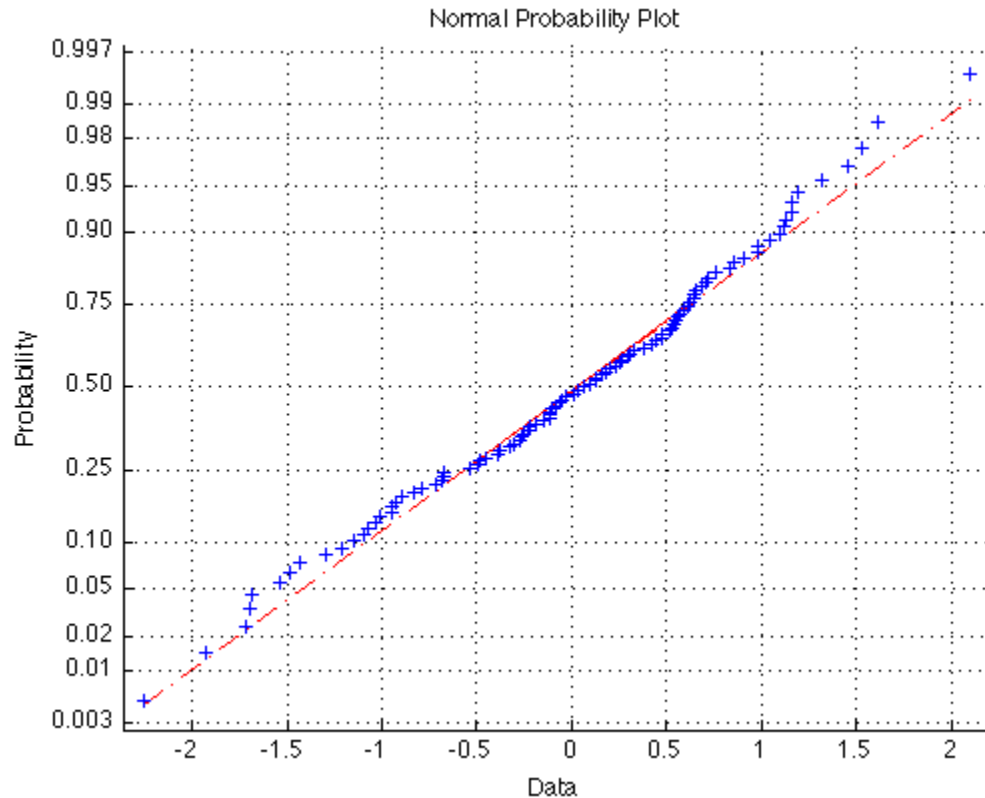
- Un análisis exploratorio de los residuos; ¿qué tipo de distribución quieren ?
- Parámetros estadísticos (v. medio, varianza, etc)
- Dirías que se cumple el modelo planteado.

```
clear
clc
close all

load lineal_parametros;
entradas=ones(101,4);
% La primera columna de la matriz X al utilizar regress debe ser de unos,
% por eso inicializamos de esta forma.
entradas(:,2)=x;
entradas(:,3)=cos(x);
entradas(:,4)=x.^2;
[B,BINT,R,RINT,STATS] = regress(y',entradas);

% Histograma de los residuos
hist(R)
figure
% Comprobamos si los residuos siguen una distribución normal (La línea
% diagonal se corresponde con una distribución normal
normplot(R)
```





Modelo de regresión "Teórico"

En este ejercicio vamos a crear un conjunto de datos artificial, con una relación lineal entre las variables, que cumple las hipótesis necesarias para que se pueda aplicar un modelo de regresión y vamos a ver qué herramientas nos permiten determinar la bondad del ajuste realizado.

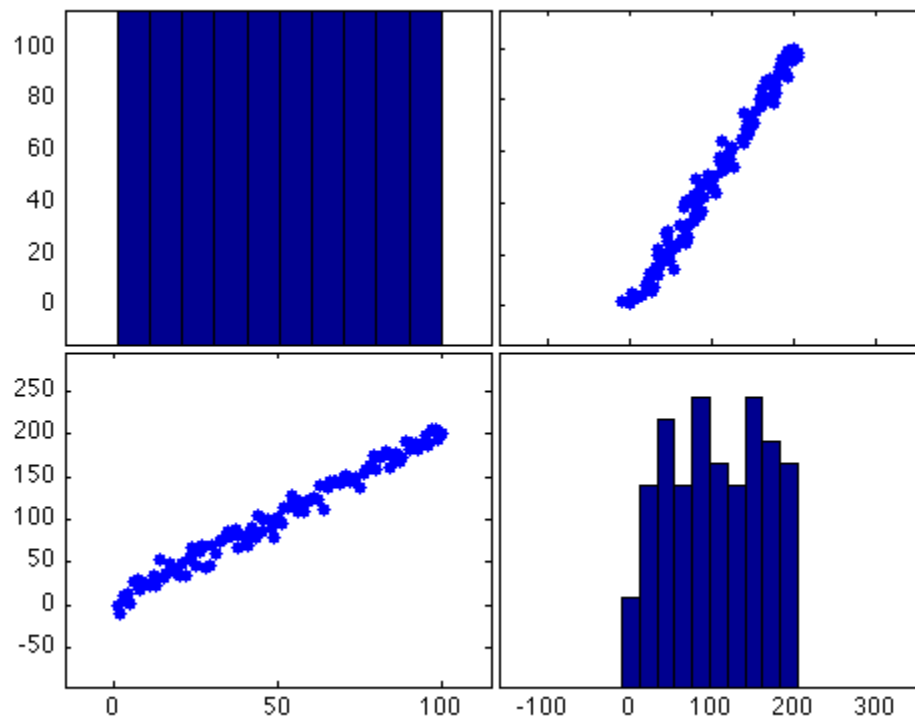
%Creamos unos datos que siguen un modelo lineal, con unos errores que
%siguen una distribución normal de media cero y varianza constante igual a 10.

```
clear  
close all
```

```
randn('state',0); %Así siempre generamos la misma secuencia aleatoria, para obtener  
N=100;  
x=(1:N)'; y=3+2*x+randn(N,1)*10; %Generamos los datos
```

```
% Dibujamos la matriz de dispersión  
plotmatrix([x y])
```

```
% Descomentando las siguientes líneas podemos ver el efecto de diferentes  
% modelos sobre el conjunto de datos  
%polytool(x,y)  
%robustfit(x,y)  
%[B,BINT,R,RINT,STATS] = regress(y,[ones(size(x)),x])
```



```
% El significado de los parámetros es el siguiente:
% B coeficientes del modelo
% BINT intervalo de confianza para B
% R residuo
% RINT
%STATS: cuatro valores en este orden [the R-square statistic, the F statistic and
%model, and an estimate of the error variance]
```

```
%Realizamos el ajuste al modelo lineal
[p,s]=polyfit(x,y,1)
p
N=length(x);
%Calculamos la estimación del modelo.
yest = polyval(p,x,s);
%Calculamos los residuos
residuo=y-yest;
```

$p =$

1.9868 4.1474

$s =$

R: [2x2 double]
df: 98
normr: 86.3305

$p =$

1.9868 4.1474

Análisis de varianza (ANOVA) para la regresión

Este método se puede utilizar para verificar la significación de la regresión. El procedimiento divide la variabilidad total de la variable respuesta (y) en 2 partes significativas. Se obtienen sumando y restando y a la diferencia entre la variable respuesta y su valor medio (que se correspondería con el modelo bobo, que siempre da como salida el valor medio)

$$\sum_{i=1}^n (y - \bar{y})^2 = \sum_{i=1}^n (\hat{y} - \bar{y})^2 + \sum_{i=1}^n (y - \hat{y})^2$$

Estas componentes son

$$SS_T = SS_R + SS_E$$

SSR es la variabilidad de la salida explicada por el modelo de regresión SSE es la variabilidad residual no explicada por el modelo.

Se plantea un contraste de hipótesis para verificar si la hipótesis nula $H_0: b_1(\text{pendiente del modelo lineal})=0$, y la hipótesis alternativa $H_1: b_1$

```
%es distinto de 0 para ello se plantea el estadístico:
%
% $$F_0=\frac{SS_R/1}{SS_E/(n-2)}=\frac{MS_R}{MS_E}$$
%
% Que sigue una distribución
```

$F_{1, n-1}$

distribución F con 1, n-1 grados de libertad La hipótesis nula NO se acepta si

$$f_0 > f_{\alpha, 1, n-1}$$

```
%Calculamos SST, SSR, SSE y R^2 y el error de la varianza
SST=sum((y-mean(y)).^2)
SSR=sum((yest-mean(y)).^2)
SSE=sum(residuo.^2)
R_square=SSR/SST
error_variance=SSE/(N-2)

MSR=SSR/1
MSE=SSE/(N-2)
%Obtenemos el estadístico
Fo=MSR/MSE
```

$SST =$

3.3636e+05

```
SSR =  
    3.2891e+05  
  
SSE =  
    7.4530e+03  
  
R_square =  
    0.9778  
  
error_variance =  
    76.0506  
  
MSR =  
    3.2891e+05  
  
MSE =  
    76.0506  
  
Fo =  
    4.3248e+03
```

A partir del valor de F_o , se calcula la significación estadística p (almacenada en `STATS(3)`) y obtenemos 0.

P-value es el valor mínimo del nivel de significación (α) que nos llevaría a DESECHAR la hipótesis nula. (si $P < \alpha$ se desecha la hipótesis nula y se da por válida la hipótesis alternativa) Ejemplo: Si en un test obtenemos $P=0.038$ significa que la hipótesis nula debe DESECHARSE si hemos fijado un nivel de significación $\alpha > 0.038$ pero no se debe desechar si se ha fijado un nivel de significación inferior p.ej. $\alpha=0.01$; En este caso, dado que habitualmente se fija $\alpha=0.05$, y hemos obtenido $f_o=STATS(3)=0$ $f_o < \alpha$ luego RECHAZAMOS la hipótesis nula que es $b_1=0$ (pendiente es 0) y aceptamos el modelo lineal es adecuado para medir esta relación

Ejercicio

```
%En un estudio Médico se ha determinado una variable (y) en función la  
%edad (x) obteniéndose la siguiente tabla de valores  
% BMI(y) Age(x)  
% 19.92 45.5  
% 20.59 34.6  
% 29.02 40.6  
% 20.78 32.9  
% 25.97 28.2  
% 20.39 30.1  
% 23.29 52.1  
% 17.27 33.3  
% 35.24 47
```

- Verifica la significación de la regresión para $\alpha=0.05$. Determina el valor de P para este test. Puedes concluir si un modelo lineal es adecuado para medir la relación entre ambas variables-
- Estima la varianza del modelo y la desviación típica del parámetro estimado b_1 .
- ¿Cuál es la desviación típica de la ordenada en el origen (b_0) para este modelo ?

Solución

```
clear
close all

Data=[19.92 45.5
20.59 34.6
29.02 40.6
20.78 32.9
25.97 28.2
20.39 30.1
23.29 52.1
17.27 33.3
35.24 47];
x=Data(:,2);
y=Data(:,1);

[B,BINT,R,RINT,STATS] = regress(y,[ones(size(x)),x])
% alpha=0.05, or 95% de intervalo de confianza es el valor por defecto de
% regress
% Obtenemos que P=STATS(3)=0.3105 >0.05 luego se acepta la hipotesis nula y
% NO podemos concluir que haya una relación lineal entre las variables

B =

    13.8201
     0.2558

BINT =

    -7.7957    35.4360
    -0.2975     0.8092

R =

    -5.5413
    -2.0825
     4.8124
    -1.4576
     4.9349
    -1.1312
    -3.8599
    -5.0699
     9.3950

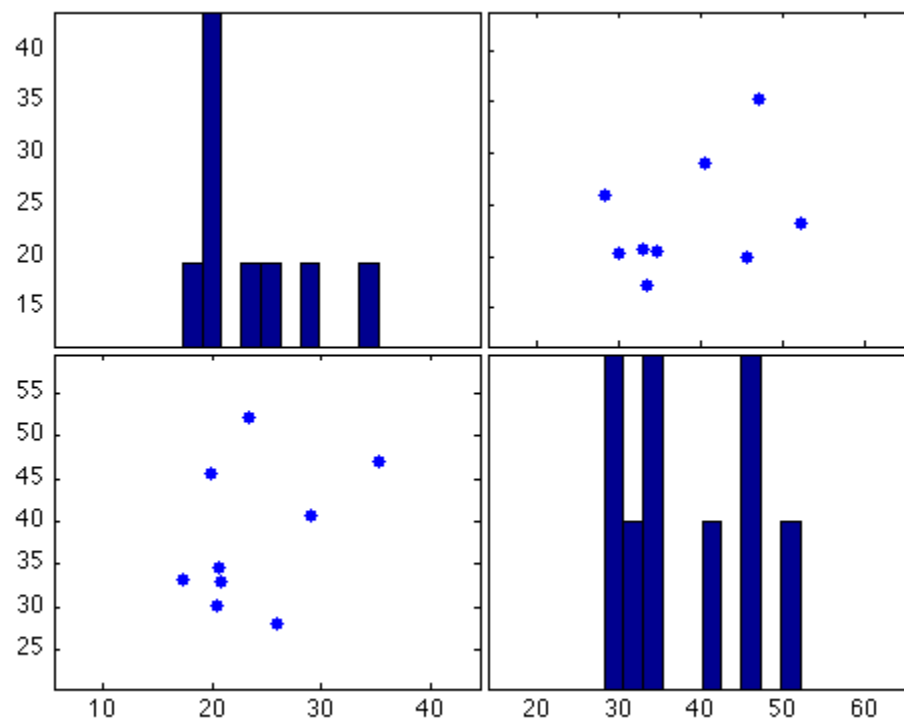
RINT =
```

```
-16.9688    5.8863
-15.0878   10.9228
-7.6137    17.2385
-14.3311   11.4160
-5.9799    15.8498
-13.5004   11.2381
-13.6369    5.9172
-17.1209    6.9811
 1.1257    17.6642
```

STATS =

```
0.1458    1.1952    0.3105   30.6896
```

plotmatrix(Data)



corrcoef(Data)

ans =

```
1.0000    0.3819
0.3819    1.0000
```

La varianza del modelo viene dada por STATS(4)=30.68 Podemos obtener muchos más parámetros

estadísticos de la regresión con la función `regstats`

```
regstats(y,x)
```

```
% Además de los parámetros que hemos indicado también se pueden obtener  
% intervalos de confianza para los parámetros del modelo (ver curvas  
% polytool). Estos modelo son los que se almacenan en la variable BINT  
% devueltas por regress
```

Regstats Export to Workspace

<input type="checkbox"/> Full QR Decomposition	qr
<input type="checkbox"/> Coefficients	beta
<input type="checkbox"/> Coefficient Covariance	covb
<input type="checkbox"/> Fitted Values	yhat
<input type="checkbox"/> Residuals	r
<input type="checkbox"/> Mean Square Error	mse
<input type="checkbox"/> R-square Statistic	rsquare
<input type="checkbox"/> Adjusted R-square Statistic	adirsquare
<input type="checkbox"/> Leverage	leverage
<input type="checkbox"/> Hat Matrix	hatmat
<input type="checkbox"/> Delete-1 Variance	s2_i
<input type="checkbox"/> Delete-1 Coefficients	beta_i
<input type="checkbox"/> Standardized Residuals	standres
<input type="checkbox"/> Studentized Residuals	studres
<input type="checkbox"/> Change in Beta	dfbetas
<input type="checkbox"/> Change in Fitted Value	dffit
<input type="checkbox"/> Scaled Change in Fit	dffits
<input type="checkbox"/> Change in Covariance	covratio
<input type="checkbox"/> Cook's Distance	cookd
<input type="checkbox"/> t Statistics	tstat
<input type="checkbox"/> F Statistic	fstat
<input type="checkbox"/> DW Statistic	dwstat

OK Cancel Help

Determinación de la adecuación del modelo de regresión

Un modelo de regresión presupone varias hipótesis:

- Errores no correlacionados con valor medio cero y varianza constante
- Para realizar el test de hipótesis los errores deben seguir una distribución normal.
- Estamos presuponiendo que un modelo de orden uno es el más adecuado.

Una vez obtenido el modelo debemos comprobar en qué medida se cumplen estas hipótesis. Usamos "normal probability plot" de los residuos (normplot). Podemos estandarizar los residuos $(\text{res} - \text{mean}(\text{res})) / \text{std}(\text{res})$, si están normalmente distribuidos, aproximadamente un 95% de residuos estará entre el -2 y el 2. Residuos fuera de este intervalo son Outliers. También son habituales representaciones de: Serie temporal de los residuos. Residuos frente a yestimada. Residuos frente a x.

Otra medida de bondad es el Coeficiente de determinación $R^2 = \text{SSR} / \text{SST}$,

`%%Ejercicio.`

`% Determina en qué medida se cumplen las hipótesis necesarias para aplicar un modelo`

Published with MATLAB® 7.10