AJUSTES DE CURVAS
Métodos Lineales y Estimación por Mínimos Cuadrados

Ing. Yamil Armando Cerquera[1]
Esp Sistemas U. Nacional de Colombia
Facultad de Ingeniería
Universidad Surcolombiana

CONTENIDO

[1] Docente de planta. Universidad Surcolombiana. Escalafón Asociado. Programa Ingeniería Electrónica

Preámbulo

A lo largo de la profesión de un ingeniero, un físico, un matemático, frecuentemente se presentan ocasiones en las que deben ajustar curvas a un conjunto de datos representados por puntos. Las técnicas desarrolladas para este fin pueden dividirse en dos categorías generales: interpolación y regresión. Se considerará aquí la primera de estas dos categorías. Más aún, como la teoría de aproximación polinomial es más adecuada para un primer curso de cálculo numérico, será la que se considere principalmente en este trabajo.

Cuando se asocia un error sustancial a los datos, la interpolación polinomial es inapropiada y puede llevar a resultados no satisfactorios cuando se usa para predecir valores intermedios. Los datos experimentales a menudo son de ese tipo. Una estrategia mas apropiada en estos casos es la de obtener una función aproximada que ajuste "adecuadamente" el comportamiento o la tendencia general de los datos, sin coincidir necesariamente con cada punto en particular.

Una línea recta puede usarse en la caracterización de la tendencia de los datos sin pasar sobre ningún punto en particular. Una manera de determinar la línea, es inspeccionar de manera visual los datos graficados y luego trazar la "mejor" línea a través de los puntos. Aunque este enfoque recurre al sentido común y es válido para cálculos a "simple vista" es deficiente ya que es arbitrario. Es decir, a menos que los puntos definan una línea recta perfecta (en cuyo caso la interpolación seria apropiada), cada analista trazará rectas diferentes.

La manera de quitar esta subjetividad es considerar un criterio que cuantifique la suficiencia del ajuste. Una forma de hacerlo es obtener una curva que minimice la diferencia entre los datos y la curva y el método para llevar a cabo este objetivo es al que se le llama *regresión con mínimos cuadrados*.

Introducción

El presente trabajo forma parte de los objetivos y contenidos de aprendizaje de la cátedra MÉTODOS NUMÉRICOS, que pretende desarrollar las habilidades para la utilización de los métodos lineales y estimación de mínimos cuadrados.

En este trabajo básicamente se habla de cómo desarrollar la aplicación de los métodos lineales y estimación por mínimos cuadrados, además de inferencia, predicción y correlación.

Se desarrollaron una serie de ejemplos mediante los cuales se trata de presentar la manera más sencilla de usar estos métodos.

Si se sabe que existe una relación entre una variable denominada dependiente y otras denominadas independientes (como por ejemplo las existentes entre: la experiencia

profesional de los trabajadores y sus respectivos sueldos, las estaturas y pesos de personas, la producción agraria y la cantidad de fertilizantes utilizados, etc.), puede darse el problema de que la dependiente asuma múltiples valores para una combinación de valores de las independientes.

La dependencia a la que hace referencia es relacional matemática y no necesariamente de causalidad. Así, para un mismo número de unidades producidas, pueden existir niveles de costo, que varían empresa a empresa.

Si se da ese tipo de relaciones, se suele recurrir a los estudios de regresión en los cuales se obtiene una nueva relación pero de un tipo especial denominado función, en la cual la variable independiente se asocia con un indicador de tendencia central de la variable dependiente. Cabe recordar que en términos generales, una función es un tipo de relación en la cual para cada valor de la variable independiente le corresponde uno y sólo un valor de la variable dependiente.

Objetivos

Entre los objetivos propuestos en este apartado se puede citar los siguientes:

1. Que sea fácilmente comprensible para los alumnos con un conocimiento mínimo de matemáticas;
2. Capacitar a los alumnos para que practiquen los métodos numéricos en una computadora;
3. Elaborar programas simples que puedan usarse de manera sencilla en aplicaciones científicas;
4. Proporcionar software que resulte fácil de comprender.

La importancia de los métodos numéricos ha aumentado de forma drástica en la enseñanza de la ingeniería y la ciencia, lo cual refleja el uso actual y sin precedentes de las computadoras.

El desarrollo de un programa siempre es importante en el aprendizaje de métodos numéricos. La presentación de resultados calculados con gráficos utilizando algún software, por ejemplo MATLAB, motiva a los alumnos para aprender métodos matemáticos y numéricos que de otra forma podrían resultar tediosos.

Regresión Simple y Correlación

La Regresión y la Correlación son dos técnicas estadísticas que se pueden utilizar para solucionar problemas comunes en los negocios.

Muchos estudios se basan en la creencia de que es posible identificar y cuantificar alguna Relación Funcional entre dos o más variables, donde una variable depende de la otra variable.

Se puede decir que y depende de x, en donde y y x son dos variables cualquiera en un modelo de Regresión Simple.

"y es una función de x" $y = f(x)$

Como y depende de x,

y Es la variable dependiente, y
x Es la variable independiente.

En el Modelo de Regresión es muy importante identificar cuál es la variable dependiente y cuál es la variable independiente.

En el Modelo de Regresión Simple se establece que y es una función de sólo una variable independiente, razón por la cual se le denomina también Regresión Divariada porque sólo hay dos variables, una dependiente y otra independiente y se representa así:

$y = f(x)$ **"Y está regresando por X"**

La variable dependiente es la variable que se desea explicar, predecir. También se le llama REGRESANDO ó VARIABLE DE RESPUESTA.

La variable Independiente x se le denomina VARIABLE EXPLICATIVA ó REGRESOR y se le utiliza para EXPLICAR Y.

En el estudio de la relación funcional entre dos variables poblacionales, una variable x, llamada independiente, explicativa o de predicción y una variable y, llamada dependiente o variable respuesta, presenta la siguiente notación:

$$y = a + bx + e$$

Donde:

a: Es el valor de la ordenada donde la línea de regresión se intercepta con el eje Y.
b: Es el coeficiente de regresión poblacional (pendiente de la línea recta)
e: Es el error que se comete al ajustar los datos.

Suposiciones de la Regresión Lineal

1. Los valores de la variable independiente X son fijos, medidos sin error.
2. La variable Y es aleatoria
3. Para cada valor de X, existe una distribución normal de valores de Y (subpoblaciones Y)
4. Las variancias de las subpoblaciones Y son todas iguales.
5. Todas las medias de las subpoblaciones de Y están sobre la recta.
6. Los valores de Y están normalmente distribuidos y son estadísticamente independientes.

Problemas al Ajustar un Modelo de Regresión Lineal Simple

Al ajustar un modelo de regresión lineal simple se pueden presentar diferentes problemas bien porque no existe una relación lineal entre las variables o porque no se verifican las hipótesis estructurales que se asumen en el ajuste del modelo. Estos problemas son los siguientes:

- ✓ **Falta de Linealidad,** porque la relación entre las dos variables no es lineal o porque variables explicativas relevantes no han sido incluidas en el modelo.
- ✓ **Existencia de valores atípicos e influyentes,** existen datos atípicos que se separan de la nube de datos muestrales e influyen en la estimación del modelo.
- ✓ **Falta de Normalidad,** los residuos del modelo no se ajustan a una distribución normal.
- ✓ **Heterocedasticidad,** La heterocedasticidad es la existencia de una varianza no constante en las perturbaciones aleatorias de un modelo econométrico.
- ✓ **Dependencia (autocorrelación),** existe dependencia entre las observaciones.

En este apartado se estudia como detectar estos problemas, su influencia en el cálculo del modelo de regresión y las posibles soluciones de los mismos.

Un primer paso para el estudio de estos problemas es la realización de un estudio descriptivo, analítico y gráfico, de la muestra. En particular el gráfico de puntos de la muestra bidimensional permite detectar algunos problemas como se deja de manifiesto en las siguientes figuras (1 al 6).

Figura 1. La nube de puntos muestrales bidimensionales parece ajustarse bien a una recta.

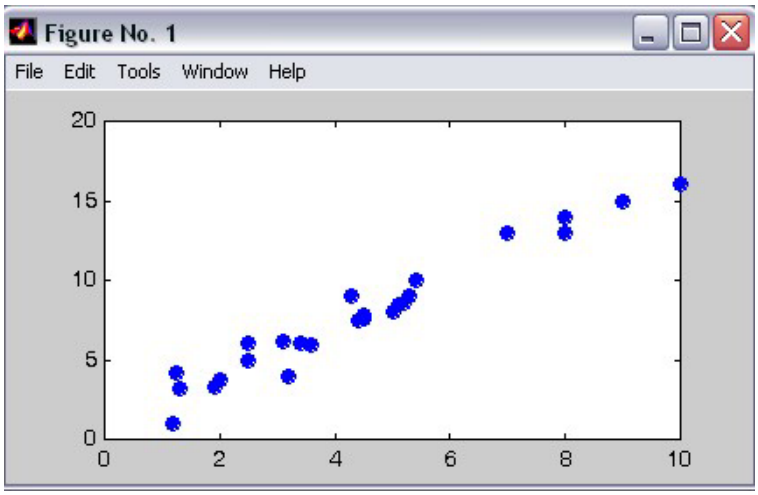


Figura 2. El ajuste lineal no parece adecuado para esta muestra.

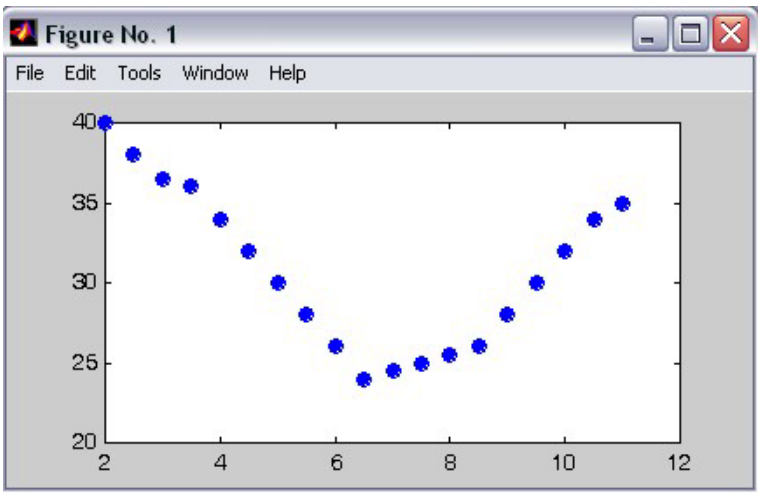


Figura 3. No existe relación lineal entre las dos variables.

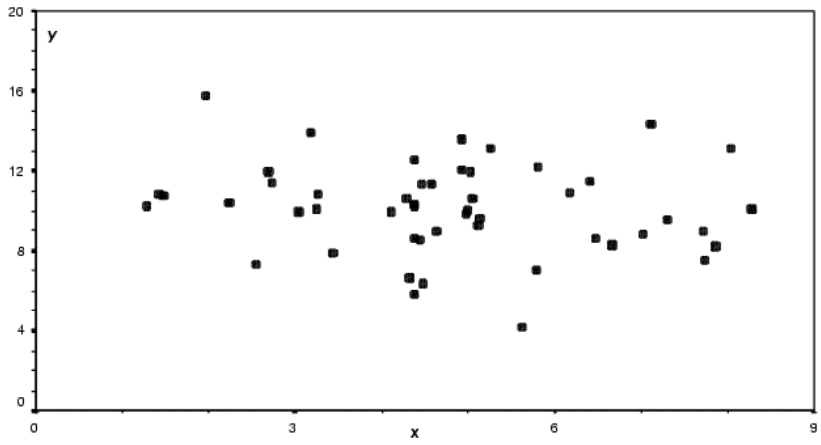


Figura 4. Claros indicios de heterocedasticidad.

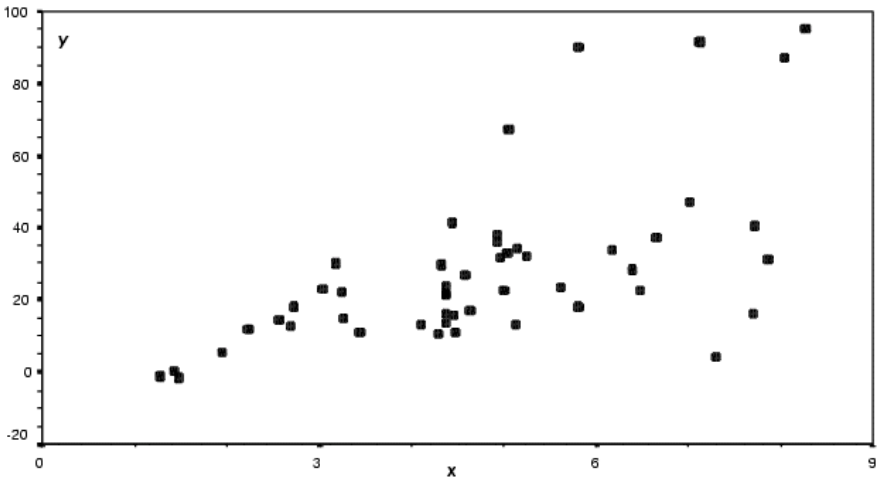


Figura 5. Existen puntos atípicos que probablemente influyan en la estimación de la recta ajustada.

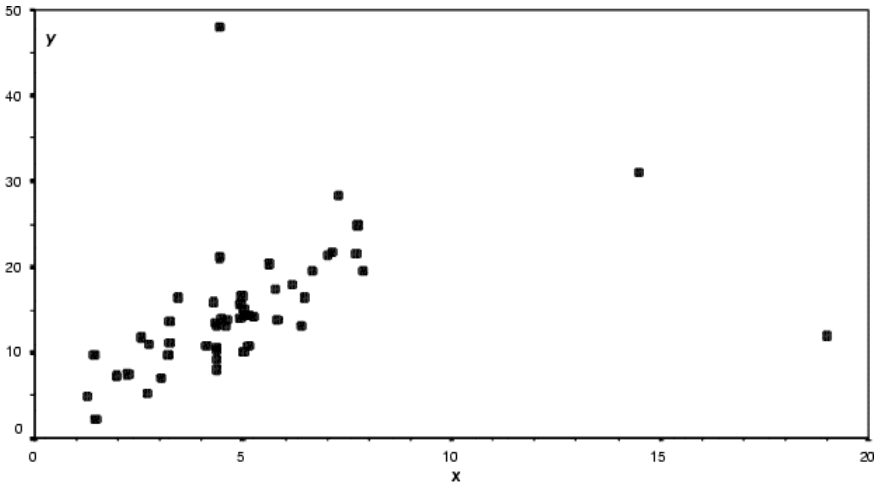
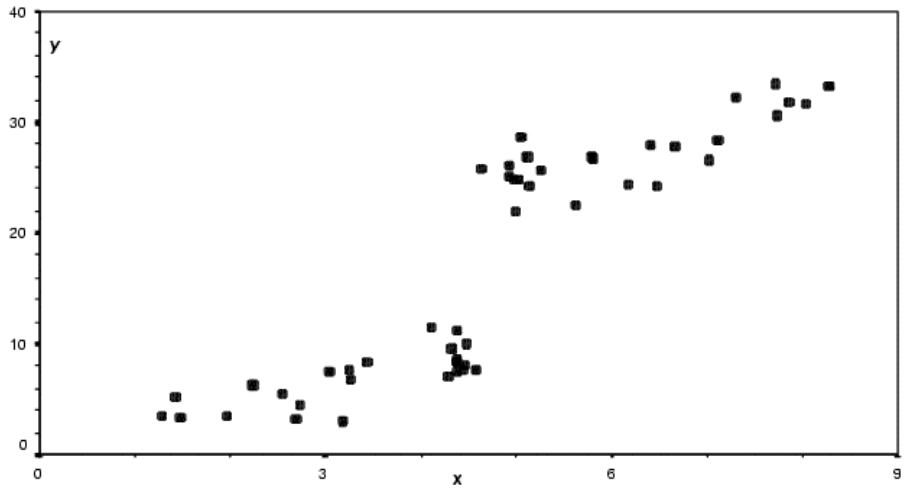


Figura 6. Existe una variable regresora binaria que se debe de incluir en el modelo de regresión.



Método de Mínimos Cuadrados

El procedimiento mas objetivo para ajustar una recta a un conjunto de datos presentados en un diagrama de dispersión se conoce como "el método de los mínimos cuadrados". El ejemplo mas simple de una aproximación por mínimos cuadrados es el ajuste de una línea recta a un conjunto de parejas de datos observadas: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$. La recta resultante $y = a + bx + E$, en donde a y b son coeficientes que representan la intersección con el eje de las abcisas y la pendiente, E es el error o residuo entre las observaciones y el modelo, $E = y - a + bx$, y presenta dos características importantes:

1. Es nula la suma de las desviaciones verticales de los puntos a partir de la recta de ajuste $\sum(\bar{Y} - Y) = 0$.

2. Es mínima la suma de los cuadrados de dichas desviaciones. Ninguna otra recta daría una suma menor de las desviaciones elevadas al cuadrado $\sum(\bar{Y}-Y)^2 \rightarrow 0$ (mínima).

Criterio para un "mejor" ajuste

Una estrategia que obtiene la "mejor" línea a través de los puntos debe minimizar la suma de los errores residuales, como en:

$$\sum_{i=1}^{n} E_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) \qquad \textit{Ec 1}$$

Otro criterio seria minimizar la suma de los valores absolutos de las diferencias, esto es:

$$\sum_{i=1}^{n} |E_i| = \sum_{i=1}^{n} [y_i - a_0 - a_1 x_i] \qquad \textit{Ec 2}$$

Una tercera estrategia en el ajuste de una línea óptima es el criterio de *mínimas*. En este método, la línea se escoge de tal manera que minimice la distancia máxima a la que se encuentra un punto de la línea recta. Esta estrategia esta mal condicionada para regresión ya que influye de manera indebida sobre un punto externo, aislado, cuyo error es muy grande. Se debe notar que el criterio de mínimas, algunas veces esta bien condicionado para ajustar una función simple a una función complicada.

Una estrategia que ignora las restricciones anteriores es la de minimizar la suma de los cuadrados de los residuos, S_r, de la siguiente manera:

$$S_r = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \textit{Ec 3}$$

Este criterio tiene muchas ventajas, incluyendo el que ajusta una línea única a un conjunto dado de datos. Antes de analizar estas propiedades, se muestra un método que determina los valores de a y b que minimizan la ecuación *Ec 3*.

Primera forma de obtener los valores a y b.

La obtención de los valores de a y b que minimizan esta función es un problema que se puede resolver recurriendo a la derivación parcial de la función en términos de a y b: llamemos G a la función que se va a minimizar:

$$G = \sum (y - a - bx)^2 \qquad \textit{Ec 4}$$

Se toma las derivadas parciales de G respecto de a y b que son las incógnitas y se igualan a cero; de esta forma se obtienen dos ecuaciones llamadas ***ecuaciones normales*** del modelo, que pueden ser resueltas por cualquier método ya sea igualación o matrices para obtener los valores de a y b.

Primera Ecuación Normal

La ecuación $G=\sum(y-a-bx)^2$, se deriva parcialmente respecto de a

$\frac{dG}{da}=2\sum(y-a-bx)(-1)=0$ = $\frac{dG}{da}=-2\sum(y-a-bx)=0$, donde

$\frac{dG}{da}=\sum(y-a-bx)=0$, y si se tienen n términos entonces.

$\frac{dG}{da}=\sum y-na-b\sum x=0$, organizando el sistema se tendrá:

$$\sum y=na+b\sum x \quad \textit{Primera ecuación normal Ec 5}$$

Segunda Ecuación Normal

Ahora se deriva parcialmente la ecuación $G=\sum(y-a-bx)^2$ respecto de b

$$\frac{dG}{db}=2\sum(y-a-bx)(-x)=0 \equiv \frac{dG}{db}=-2\sum(y-a-bx)(x)=0$$

$$\frac{dG}{db}=\sum(y-a-bx)(x)=0 \equiv \frac{dG}{db}=\sum(xy-ax-bx^2)=0$$

$\frac{dG}{db}=\sum xy-a\sum x-b\sum x^2=0$, organizando el sistema se tendrá:

$$\sum xy=a\sum x+b\sum x^2 \quad \textit{Segunda ecuación normal Ec 6}$$

Los valores de a y b se obtienen resolviendo el sistema de dos ecuaciones (Primera y segunda ecuación normal) con dos variables (a y b) dados en las Ec 5 y 6. Si se toma lo siguiente, $A=\sum y,\ B=\sum x,\ C=\sum xy,\ D=\sum x^2$, se puede decir que el sistema de ecuaciones quedará así: $\begin{cases} na+Bb=A \\ Ba+Db=C \end{cases}$, resolviendo con el programa MatLab, la expresión:

```
S = solve('n*a+B*b=A','B*a+D*b=C','a,b')
```
dará como resultado

```
S = a: [1x1 sym]
    b: [1x1 sym]
```

Como S es una estructura, toca pedir luego el valor correspondiente al campo b de La siguiente manera:

```
» S.b
ans = -(-n*C+A*B)/(n*D-B^2),
```
que interpretado como función dará:

$b = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$ y $a = \dfrac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}$, o con el valor calculado para b, se puede obtener el valor de a así: $a = y - bx$

EJEMPLO 1

Ajústese una línea recta a los valores x y y de las primeras dos columnas de la siguiente tabla:

x_i	y_i	$x_i y_i$	$(y_i - \bar{Y})^2$	$(y_i - a - bx_i)^2$
1	0.5	0.5	8.5765	0.1687
2	2.5	5.0	0.8622	0.5625
3	2.0	6.0	2.0408	0.3473
4	4.0	16.0	0.3265	0.3265
5	3.5	17.5	0.0051	0.5896
6	6.0	36.0	6.6122	0.7972
7	5.5	38.5	4.2908	0.1993
$\sum \rightarrow 28$	24	119.5	22.7143	2.9911

Se pueden calcular las siguientes cantidades:

$n = 7$ $\quad \sum x_i y_i = 119.5$ $\quad \sum x_i^2 = 140$ $\quad \sum x_i = 28$

$\bar{x} = \dfrac{28}{7} = 4$ $\quad \sum y_i = 24$ $\quad \bar{y} = \dfrac{24}{7} = 3.428571429$

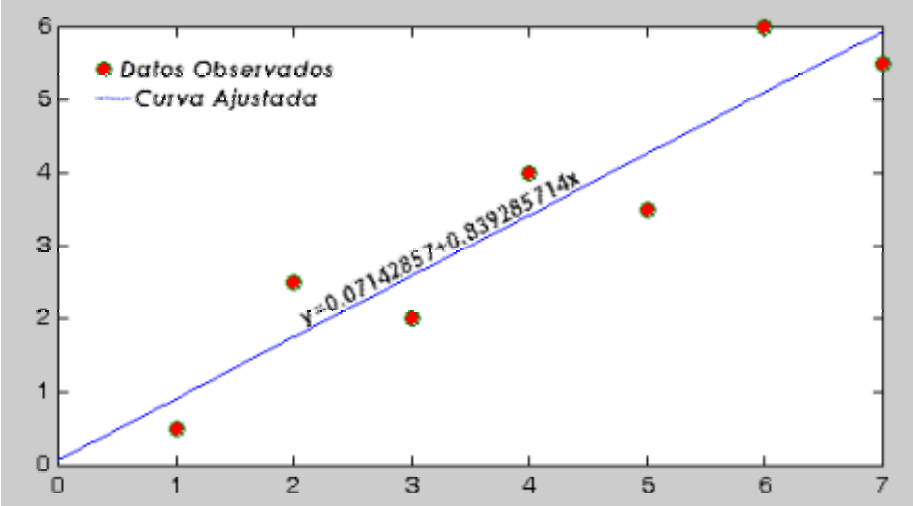
Usando las ecuaciones: $b = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$ y $a = y - bx$, se tiene:

$$b = \frac{7 * 119.5 - 28 * 24}{7 * 140 - 28^2} = 0.839285714$$

$$a = 3.428571429 - 0.829285714 * 4 = 0.07142857$$

Por lo tanto la ecuación lineal con ajuste por mínimos cuadrados es:

$$y = 0.07142857 + 0.839285714x$$



Ejemplo con MatLab:

```
x = [0.1, 0.4, 0.5, 0.7, 0.7, 0.9];
y = [0.61, 0.92, 0.99, 1.52, 1.47, 2.03];
c = polyfit(x,y,1)
c1 = x(1):0.1:x(length(x))
c2 = polyval(c,c1)
plot(c1,c2);hold on
plot(x,y,'x')
axis([0,1,0,2.1])
xlabel('x')
ylabel('y')
```

EJEMPLO 2

Se toma una muestra aleatoria de 8 ciudades de una región geográfica de 13 departamentos y se determina por los datos del censo el porcentaje de graduados en educación superior y la mediana del ingreso de cada ciudad, los resultados son los siguientes:

CIUDAD	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
% de (x) Graduados	7.2	6.7	17	12.5	6.3	23.9	6	10.2
Ingreso (y) Mediana	4.2	4.9	7	6.2	3.8	7.6	4.4	5.4

De las ecuaciones normales:

$$\sum y = na + b\sum x \qquad \text{y} \qquad \sum xy = a\sum x + b\sum x^2$$

Se debe encontrar los términos de las ecuaciones

$\sum y$, $\sum x$, $\sum xy$, $\sum x^2$ Por tanto se procede de la siguiente forma:

n	y	x	xy	x^2
1	4.2	7.2	30.24	51.84
2	4.9	6.7	32.83	44.89
3	7.0	17.0	119.00	289.00
4	6.2	12.5	77.50	156.25
5	3.8	6.3	23.94	39.69
6	7.6	23.9	181.64	571.21
7	4.4	6.0	26.40	36.00
8	5.4	10.2	55.08	104.04
$\sum$	43.5	89.8	546.63	1292.92

Sustituyendo en las ecuaciones los resultados obtenidos se tiene:

$$43.50 = 8a + 89.8b \qquad \textit{Ec. 1}$$
$$546.63 = 89.8a + 1292.92b \qquad \textit{Ec. 2}$$

Para resolver el anterior sistema, se multiplica la primera ecuación por (-89.8) y la segunda por (8) así:

43.50	=	8a	+	89.80b	* (-89.8)	Ec 1	546.63	=	89.8a	+	1292.92b	* (8) Ec 2
-3906.30	=	-718.4a	-	8064.04b			4373.04	=	718.4a	+	10343.36b	

Ec 1 + Ec 2 $\Rightarrow$ 466.74 = -0- 2279.32b

$$b = \frac{466.74}{2279.32} = 0.20477$$

Este valor de *b* se reemplaza en cualquiera de las ecuaciones para obtener el valor de *a*:

Reemplazando b = 0.20477 en la primera ecuación normal. (*Ec. 1*)

$43.50 = 8a + 89.8(0.20477)$, donde $43.50 = 8a + 18.3880$, despejando a se tiene:

$$a = \frac{25.120}{8} = 3.139$$

Se tiene entonces que los coeficientes de regresión son: a = 3.139 y b = 0.20477. Por tanto la ecuación de regresión queda: $\hat{Y} = 3.1390 + 0.20477x$

Significa entonces que por cada incremento en una unidad en X el valor de $\hat{Y}$ se aumenta en 0.20477

Esta ecuación permite estimar el valor de $\hat{Y}$ para cualquier valor de X, por ejemplo: Una ciudad que tiene un porcentaje de graduados a nivel superior del 28% la mediana de ingreso para la ciudad será:

$$\hat{Y} = 3.1390 + 0.20477 * (28)$$

$\hat{Y} = 8.87$ Decenas de miles de $.

Segunda forma de obtener los valores de a y b

Partiendo de las dos ecuaciones normales se tiene:

$$\sum y = na + b\sum x \text{ (Ec 1)}, \qquad \text{y} \qquad \sum xy = a\sum x + b\sum x^2 \text{ (Ec 2)}$$

Si se divide todos los términos de la ecuación normal (Ec 1) entre *n* quedando:

$$\frac{\sum y}{n} = \frac{na}{n} + \frac{b\sum x}{n}$$

Se tiene entonces que el primer término es $\bar{Y}$ el segundo término es la incógnita a y el tercer termino es la incógnita *b* multiplicada por $\bar{X}$, por tanto quedaría de la forma:

$$\bar{Y} = a + b\,\bar{X}, \text{ entonces } \quad a = \bar{Y} - b\,\bar{X}$$

Reemplazando a en la ecuación (Ec 2) se tiene:

$$\sum xy = \left(\bar{Y} - b\bar{X}\right)\sum x + b\sum x^2 \quad \rightarrow \quad b\sum x^2 = \sum xy - (\bar{Y} - b\bar{X})\sum x$$

$$b\sum x^2 = \sum xy - \bar{Y}\sum x + b\bar{X}\sum x$$

$$b\sum x^2 = \sum xy - \frac{n\bar{Y}\sum x}{n} + \frac{nb\bar{X}\sum x}{n}$$

$$b\sum x^2 = \sum xy - n\overline{YX} + nb\bar{X}^2 \quad \rightarrow \quad b\sum x^2 - nb\bar{X}^2 = \sum xy - n\overline{YX}$$

$$b\left(\sum x^2 - n\bar{X}^2\right) = \sum xy - n\overline{YX}$$

$$b = \frac{\sum xy - n\overline{YX}}{\sum x^2 - n\bar{X}^2} \quad \equiv \quad b = \frac{546.63 - 8(5.4375)(11.2250)}{1292.92 - 8(11.2250)^2} = \frac{58.3425}{284.9150} = 0.20477$$

a = 5.4375 - 0.20477 (11.2250) = 5.4375 - 2.2985 = 3.139

Se debe tener presente la diferencia entre el valor de $\hat{Y}$ obtenido con la ecuación de regresión y el valor de Y observado. Mientras $\hat{Y}$ es una estimación y su bondad en la estimación depende de lo estrecha que sea la relación entre las dos variables que se estudian; Y- es el valor efectivo, verdadero obtenido mediante la observación del investigador. En el ejemplo Y^o es el valor mediano del ingreso que obtuvo el investigador

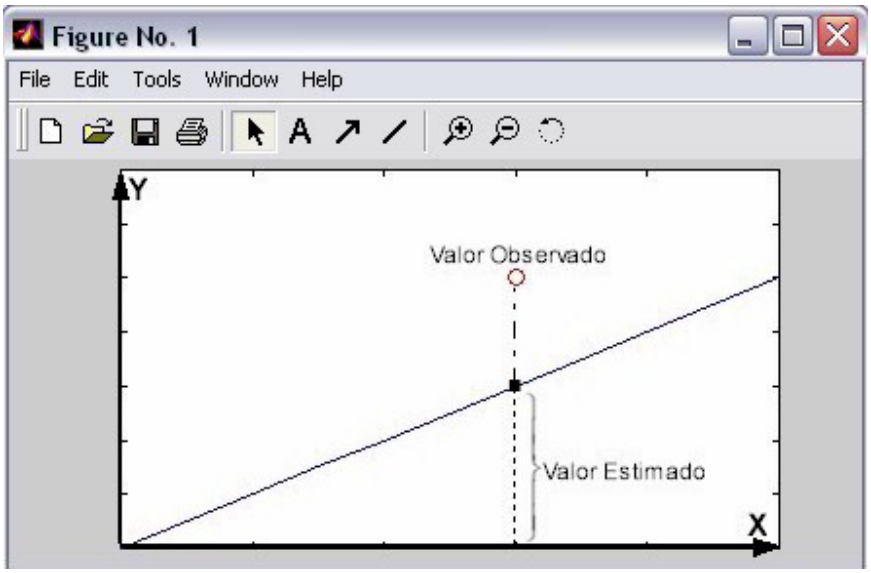
Utilizando todos los ingresos observados en cada ciudad y $\hat{Y}$ es el valor estimado con base en el modelo lineal utilizado para obtener la ecuación de regresión.

Los valores estimados y observados pueden no ser iguales por ejemplo la primera ciudad tiene un ingreso mediano observado de Y^o = 4.2 al reemplazar en la ecuación el porcentaje

De graduados se obtiene un $\hat{Y}$ estimado de

$$\hat{Y} = 3.1390 + 0.20477(1.2) = 4.61$$

Gráficamente lo anterior se puede mostrar así:



Claramente se observa en la gráfica que hay una diferencia entre el valor efectivo de Y^o y el valor estimado; esta diferencia se conoce como error en la estimación, este error se puede medir. A continuación se verá el procedimiento.

Error estándar en la estimación

El error estándar de la estimación designado por S_{yx} mide la disparidad "promedio" entre

Los valores observados y los valores estimados de $\hat{Y}$. Se utiliza la siguiente fórmula.

$$S_{YX} = \sqrt{\frac{\sum\left(Y^o - \hat{Y}\right)^2}{n-2}}$$

Se debe entonces calcular los valores de $\hat{Y}$ para cada ciudad sustituyendo en la ecuación los valores de los porcentajes de graduados de cada ciudad estudiada.

$$Y = 3.139 + 0.20477(x)$$

n	Y^o	X	$\hat{Y}$	$Y^o - \hat{Y}$	$(Y^o - \hat{Y})^2$
1	4.2	7.2	4.6	-0.4	0.16

2	4.9	6.7	4.5	0.4	0.16
3	7.0	17.0	6.6	0.4	0.16
4	6.2	12.5	5.7	0.5	0.25
5	3.8	6.3	4.4	-0.6	0.36
6	7.6	23.9	8.0	-0.4	0.16
7	4.4	6.0	4.4	0.0	0.00
8	5.4	10.2	5.2	0.2	0.04
$\sum$					1.29

$$S_{YX} = \sqrt{\frac{\sum\left(Y^o - \hat{Y}\right)^2}{n-2}} = \sqrt{\frac{1.29}{8-2}} = \sqrt{0.215} = 0.46$$ (Decenas de miles de pesos)

Como esta medida trata de resumir la disparidad entre lo observado y lo estimado, es decir, trata de medir la diferencia promedio entre lo observado y lo estimado ó esperado de acuerdo al modelo, puede considerarse como un indicador del grado de precisión con que la ecuación de regresión, describe la relación entre las dos variables. Este error estándar se ve afectado por las unidades y sus cambios ya que es una medida absoluta, pues, se da en la misma unidad de medida que esta dada la variable Y; en el ejemplo 0.46 serán decenas de miles de pesos, razón por la cual no es posible comparar con las relaciones de variables dadas en distinta unidad de medida. Es necesario entonces calcular una medida que interprete o mida mejor el grado de relación entre las variables.

Coeficiente de determinación

El cambio de la variable Y generalmente depende de muchos factores, en ocasiones, difíciles de identificar; con el modelo lineal simple, sólo tenemos presente uno. Por ejemplo, en nuestro caso la mediana del ingreso depende no sólo del porcentaje de graduados en el nivel superior, que es, el factor que tenemos presente, pueden entrar a jugar factores tales como, la distribución de la edad en la población, la distribución por sexo en la población, la industrialización de la ciudad, el numero de universidades y muchos otros.

El coeficiente de determinación mide o interpreta la cantidad relativa de la variación que ha sido explicada por la recta de regresión, es decir, la proporción de cambio en Y explicado por un cambio en la variable X (X es el factor que se utiliza para calcular la recta de ajuste o ecuación de regresión, en el ejemplo es el porcentaje de graduados en el nivel superior en cada ciudad).

Para el ejemplo el Coeficiente de determinación va a medir la proporción del cambio en el ingreso mediano de cada ciudad, debido o explicado por un cambio en el porcentaje de graduados en el nivel superior.

Vea algunos componentes de la variabilidad en el análisis de regresión:

La diferencia entre cada valor de Y observado y $\overline{Y}$ media se denomina variación de Y.

(Y^o - $\overline{Y}$)=Variación de Y.

La diferencia entre $\hat{Y}$ estimado y $\overline{Y}$ media, es la variación tenida en cuenta por la ecuación de regresión, razón por la cual se denomina variación explicada de Y.

($\hat{Y}$ - $\overline{Y}$) = variación explicada de Y.

La diferencia entre Y^o observado y $\hat{Y}$ estimado, son variaciones consideradas debidas a factores diferentes al tenido presente por la ecuación de regresión por eso se llama: variación no explicada de Y.

($Y^o - \hat{Y}$) = variación no explicada de Y
La sumatoria de las diferencias en cada una de las formas de variación la podemos representar así:

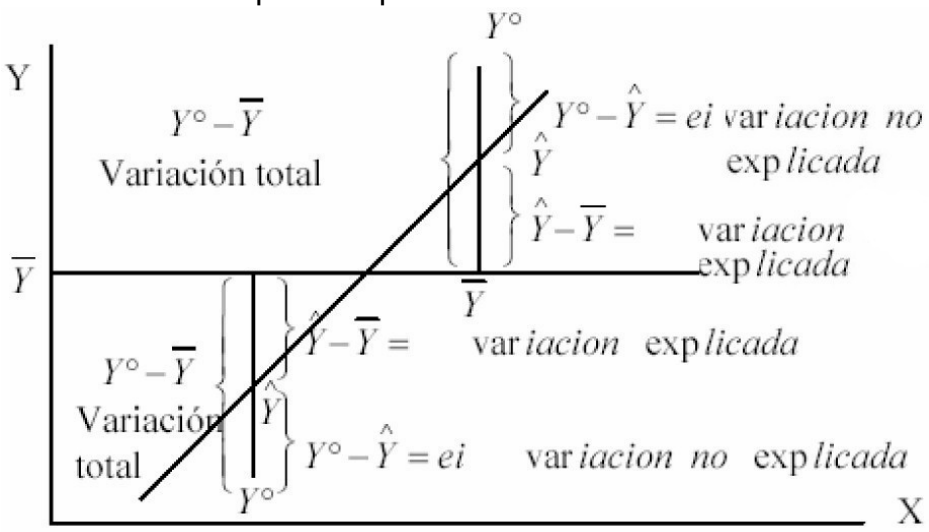
$\sum\left(Y^o - \overline{Y}\right)^2$ = Variación total

$\sum\left(Y^o - \hat{Y}\right)^2$ = Variación no explicada

$\sum\left(\hat{Y} - \overline{Y}\right)^2$ = Variación explicada

$\sum\left(Y^o - \overline{Y}\right)^2 = \sum\left(Y^o - \hat{Y}\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2$

Gráficamente esta relación se puede representar así:



Se mencionó anteriormente, que el coeficiente de determinación es la proporción de cambio explicado en Y, por cambio en X, es decir, la proporción que representa la variación explicada de la variación total. Recuerde una proporción es la relación de una parte con el total, por tanto, el coeficiente de determinación será:

$$r^2 = \frac{\sum\left(\hat{Y} - \overline{Y}\right)^2}{\sum\left(Y^o - \hat{Y}\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2} = \frac{\sum\left(\hat{Y} - \overline{Y}\right)^2}{\sum\left(Y^o - \overline{Y}\right)^2}$$

En otras palabras el coeficiente de determinación es la relación entre la variación explicada y la variación total. Su valor siempre estará $0 \le r^2 \le 1$.

Para su cálculo se procede así:

n	Y^o	$\overline{Y}$	$Y^o - \overline{Y}$	$\left(Y^o - \overline{Y}\right)^2$	$\hat{y}$	$\hat{y} - \overline{Y}$	$(\hat{Y} - \overline{Y})^2$	$Y^o - \hat{y}$	$(Y^o - \hat{Y})^2$
1	4.2	5.44	-1.24	1.5376	4.6	-0.84	0.71	-0.4	0.16
2	4.9	5.44	-1.24	0.29	4.5	-0.84	0.88	0.4	0.16
3	7.0	5.44	1.56	2.43	6.6	1.16	1.35	0.4	0.16
4	6.2	5.44	0.76	0.58	5.7	0.26	0.07	0.5	0.25
5	3.8	5.44	1.64	2.69	4.4	-1.04	1.08	-0.6	0.36
6	7.6	5.44	2.16	4.66	8.0	2.56	6.55	-0.4	0.16
7	4.4	5.44	1.04	1.08	4.4	-1.04	1.08	0.0	0.00
8	5.4	5.44	0.4	0.001	5.2	-0.24	0.06	0.2	0.04
$\sum$	43.5			13.271			11.78		1.29

$$\overline{Y} = 43.5/8 = 5.44$$

$$r^2 = \frac{\sum\left(\hat{y} - \overline{y}\right)^2}{\sum\left(y^o - \overline{y}\right)^2} = \frac{11.78}{13.271} = 0.8876$$

Generalmente esta proporción se expresa como porcentaje, por tanto se puede decir que

$$r^2 = 88.76\%$$

Como conclusión se puede decir que el 88.76% de la variación en el ingreso mediano de las ciudades de la muestra esta relacionada o explicada por la variación en el porcentaje de graduados en Educación Superior en cada ciudad.

Coeficiente de correlación

Este Coeficiente como ya se dijo mide la fuerza de la relación entre las variables. El coeficiente tiene el signo que tiene b y su valor estará $-1 \le r \le 1$ El signo menos en el índice significa una relación negativa y un signo más una correlación positiva. El coeficiente se obtiene sacando la raíz cuadrada al coeficiente de determinación y se simboliza con "r".

$r = \sqrt{\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y^o - \bar{y})^2}}$, por tanto $r = \sqrt{\frac{11.78}{13.2710}} = \sqrt{0.8876} = 0.942125$

En este caso el coeficiente r tiene signo positivo ya que toma el valor de b obtenido con las ecuaciones normales toma valor positivo.

A continuación se da, a modo de orientación, como podrían interpretarse los valores de r (positivo o negativo)

0.0 a 0.2 Correlación muy débil, despreciable
0.2 a 0.4 Correlación débil. Bajo
0.4 a 0.7 Correlación moderada
0.7 a 0.9 Correlación fuerte, alto, importante
0.9 a 1.0 Correlación muy fuerte, muy alto

La correlación entre los valores de dos variables es un hecho. El que lo consideremos satisfactorio o no, depende de la interpretación. Otro problema que representa la correlación es cuando se pregunta si una variable, de algún modo causa o determina a la otra. La correlación no implica causalidad. Si las variables X e Y están correlacionadas, esto puede ser por que X causa a Y, o porque Y causa a X o porque alguna otra variable afecta tanto a X como Y, o por una combinación de todas estas razones; o puede ser que la relación sea una coincidencia.

Modelo de regresión lineal con el uso de matrices y varias variables independientes

Al ajustar un modelo de regresión lineal múltiple, en particular cuando el número de variables pasa de dos, el conocimiento de la teoría matricial puede facilitar las manipulaciones matemáticas de forma considerable. Suponga que el experimentador tiene k variables independientes x_1, x_2,....,x_k, y n observaciones y_1, y_2,...., y_n, cada una de las cuales se pueden expresar por la ecuación: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$

Este modelo en esencia representa n ecuaciones que describen cómo se generan los valores de respuesta en el proceso científico. Con el uso de la notación matricial, podemos escribir la ecuación: $y = X\beta + \varepsilon$, donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \quad x = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ . & . & . & & \\ & & & \dots & x_{k3} \\ . & . & . & & \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}$$

Entonces la solución de mínimos cuadrados para la estimación de ▫ que se ilustra en la sección Estimación de coeficientes, "Regresión lineal múltiple" implica encontrar b para la que: SSE = (y - Xb)'(y - Xb)

Se minimiza. Este proceso de minimización implica resolver para b en la ecuación

$$\frac{\partial}{\partial b}(SSE) = 0$$

No se presentan los detalles relacionados con las soluciones de las ecuaciones anteriores. El resultado se reduce a la solución de b en: $(X'X)b = X'y$

Nótese la naturaleza de la matriz X. Aparte del elemento inicial, el i-ésimo renglón representa los valores x que dan lugar a la respuesta y_i. Al escribir

$$A = X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \quad \text{y} \quad g = X'y = \begin{bmatrix} g_o = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \cdot \\ \cdot \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

Las ecuaciones normales se pueden escribir en la forma matricial AB=g

Si la matriz A es no singular, se puede escribir la solución para el coeficiente de regresión como $b = A^{-1}g = (X'X)^{-1}X'y$

De esta forma se puede obtener la ecuación de predicción o la ecuación de regresión al resolver un conjunto de k + 1 ecuaciones con un número igual de incógnitas. Esto implica la inversión de la matriz $X'X$ de k + 1 por k + 1. Las técnicas para invertir esta matriz se explican en la mayoría de los libros de texto sobre determinantes y matrices elementales. Por supuesto, se dispone de muchos paquetes de computadora de alta velocidad para problemas de regresión múltiple, paquetes que no sólo imprimen estimaciones de los coeficientes de regresión, sino que también proporcionan otra información relevante para hacer inferencias respecto a la ecuación de regresión.

EJEMPLO 3

Se midió el porcentaje de sobre vivencia de cierto tipo de semen animal, después del almacenamiento, en varias combinaciones de concentraciones de tres materiales que se utilizan para aumentar su oportunidad de sobre vivencia. Los datos son los siguientes:

N	y(% sobre vivencia)	x_1(peso %)	x_2(peso %)	x_3(peso %)

1	25,5	1,74	5,30	10,80
2	31,2	6,32	5,42	9,40
3	25,9	6,22	8,41	7,20
4	38,4	10,52	4,63	8,50
5	18,4	1,19	11,60	9,40
6	26,7	1,22	5,85	9,90
7	26,4	4,10	6,62	8
8	25,9	6,32	8,72	9,10
9	32	4,08	4,42	8,70
10	25,2	4,15	7,60	9,20
11	39,7	10,15	4,83	9,40
12	35,7	1,72	3,12	7,60
13	26,5	1,70	5,30	8,20
$\sum$	377.5	59.43	81.82	115.4

Estime el modelo de regresión lineal múltiple para los datos dados.

SOLUCIÓN:

$$X'X = \begin{bmatrix} 13 & 59.43 & 81.82 & 115.4 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.4 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

Por lo tanto, las ecuaciones de estimación de mínimos cuadrados, $[X'X][b]=[X'y]$, son

$$\begin{bmatrix} 13 & 59.43 & 81.82 & 115.4 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.4 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

De los resultados obtenidos con MatLab en una computadora se obtiene los elementos de la matriz inversa

$$(X'X)^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix}$$

Y después, con el uso de la relación $b = [X'X]^{-1} * [X'y]$, los coeficientes estimados de regresión son:

$b = \begin{bmatrix} 39.1574 \\ 1.0161 \\ -1.8616 \\ -0.3433 \end{bmatrix}$, De lo cual se infiere que:

b_0= 39.1574, b_1 = 1.0161, b_2 = -1.8616, b_3 = -0.3433.

En consecuencia la ecuación de regresión estimada es:

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3$$

Modelo de regresión lineal con el uso de matrices y una sola variable independiente

Para el caso de una sola variable independiente, el grado del polinomio de mejor ajuste a menudo se puede determinar al graficar un diagrama de dispersión de los datos que se obtienen de un experimento que da n pares de observaciones de la forma $\{(x_i, y_i); i = 1,2,\ldots,n\}$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \ldots & \sum_{i=1}^{n} x_i^y \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \ldots & \sum_{i=1}^{n} x_i^{y+1} \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \ldots & \sum_{i=1}^{n} x_i^{y+2} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_i^y & \sum_{i=1}^{n} x_i^{y+1} & \sum_{i=1}^{n} x_i^{y+2} \ldots & \sum_{i=1}^{n} x_i^{2y} \end{bmatrix} \begin{matrix} b_0 \\ b_1 \\ b_2 \\ \\ \vdots \\ \\ b_y \end{matrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \vdots \\ \vdots \\ \\ \sum_{i=1}^{n} x_i^y y_i \end{bmatrix}$$

Al resolver estas r + 1 ecuaciones, se obtiene las estimaciones b_0, b_1,....., b_r y por ello se genera la ecuación de predicción de regresión polinomial: $\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_y x^y$

El procedimiento para ajustar un modelo de regresión polinomial se puede generalizar al caso de más de una variable independiente. De hecho, el estudiante de análisis de regresión debe, en esta etapa, tener la facilidad para ajustar cualquier modelo lineal en, k variables independientes. Suponga, por ejemplo, que tiene una respuesta Y con k = 2 variables independientes y se postula un modelo cuadrático del tipo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x^2{}_{1i} + \beta_{22} x^2{}_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

Donde $y_i, i = 1,2,\ldots,n$ es la respuesta para la combinación x_{1i}, x_{2i} de las variables independientes en el experimento. En esta situación n debe ser al menos 6, pues hay seis parámetros a estimar mediante el procedimiento de mínimos cuadrados.

Además, como el modelo contiene términos cuadráticos en ambas variables, se deben usar al menos tres niveles de cada variable. El lector debe verificar con facilidad que las ecuaciones normales de mínimos cuadrados $[X'X][b] = [X'y]$ están dadas por:

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \sum x_{1i}^2 & \sum x_{2i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} & \sum x_{1i}^3 & \sum x_{1i} x_{2i}^2 & \sum x_{1i}^2 x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 & \sum x_{1i}^2 x_{2i} & \sum x_{2i}^3 & \sum x_{1i} x_{2i}^2 \\ \sum x_{1i}^2 & \sum x_{1i}^3 & \sum x_{1i}^2 x_{2i} & \sum x_{1i}^4 & \sum x_{1i}^2 x_{2i}^2 & \sum x_{1i}^3 x_{2i} \\ \sum x_{2i}^2 & \sum x_{1i} x_{2i}^2 & \sum x_{2i}^3 & \sum x_{1i}^2 x_{2i}^2 & \sum x_{2i}^4 & \sum x_{1i} x_{2i}^3 \\ \sum x_{1i} x_{2i} & \sum x_{1i}^2 x_{2i} & \sum x_{1i} x_{2i}^2 & \sum x_{1i}^3 x_{2i} & \sum x_{1i} x_{2i}^3 & \sum x_{1i}^2 x_{2i}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_{11} \\ b_{22} \\ b_{12} \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \\ \sum x_{1i}^2 y_i \\ \sum x_{2i}^2 y_i \\ \sum x_{1i} x_{2i} y_i \end{bmatrix}$$

EJEMPLO 4

Los siguientes datos representan el porcentaje de impurezas que ocurren a varias temperaturas y tiempos de esterilización durante una reacción asociada con la fabricación de cierta bebida.

Tiempo de esterilización, x_2 (minutos)	Temperatura, x_1 (°C)		
	75	100	125
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

Estimar los coeficientes de regresión en el modelo

$$\mu_y \mid x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \ldots + \beta_{12} x_1 x_2$$

SOLUCIÓN:

b_0 = 56,4668	b_{11} =0,00081
b_1 = -0,36235	b_{22} = 0,08171
b_2 = -2,75299	b_{12} = 0,00314

Y la ecuación de regresión estimada es

$$\hat{y} = 56.4648 - 0.36235 x_1 - 2.75299 x_2 + 0.00081 x_1^2 + 0.08171 x_2^2 + 0.00314 x_1 x_2$$

La mayoría de los principios y procedimientos asociados con la estimación de funciones de regresión polinomial caen en la categoría de la metodología de respuesta superficial, un conjunto de técnicas que los científicos e ingenieros han utilizado con bastante éxito en muchos campos. Problemas como la selección de un diseño experimental apropiado, en particular para casos donde hay un número grande de variables en el modelo, y la elección de las condiciones "óptimas" de operación sobre $x_1, x_2, \ldots, x_k$ a menudo se aproximan a través del uso de estos métodos.