Apuntes de Métodos estadísticos de la Ingeniería, Segundo de Ingeniería Industrial

Mathieu Kessler
Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena
mathieu.kessler@upct.es

Ésta es una versión preliminar, comentarios bienvenidos, 2005
Todos los gráficos de estos apuntes han sido realizados con el programa estadístico freeware R, (http://cran.r-project.org)

TEMA I

Exploración de datos

I.1. Introducción

La estadística utiliza datos para conseguir comprensión sobre un fenómeno. Básicamente, esta comprensión es una consecuencia de la combinación entre conocimientos previos sobre el fenómeno y nuestra capacidad para utilizar gráficos y cálculos para extraer información de los datos.

En contextos industriales se recogen a menudo grandes conjuntos de datos correspondientes a un gran número de variables. Un efecto contradictorio aparece: por una parte, cuanto más datos, más información podemos extraer sobre las variables de interés, pero a la vez es más difícil su extracción.

En este contexto aparece una primera etapa fundamental frente a un conjunto de datos: la *exploración*, que se realiza a través de representaciones gráficas y del cálculo de unas cuantas medidas numéricas bien escogidas.

Para tener las ideas claras, unos cuantos gráficos pueden proporcionarnos información más valiosa que procedimientos sofisticados que no dominamos. En esta asignatura, veremos en temas posteriores métodos más sofisticados de análisis pero dedicamos ahora un capítulo a recordar las técnicas elementales con el objetivo de fomentar reacciones sanas frente a un conjunto de datos.

Aun cuando el conjunto de datos presenta varias variables, se debe empezar por el estudio individual de cada una.

I.2. Unos cuantos términos

- Un conjunto de datos describe **individuos**, que pueden ser personas pero también objetos. Por ejemplo, asociados a esta clase, podemos considerar que los individuos son los alumnos.

- Consideramos variables asociadas a este conjunto de datos, distinguiremos entre **variable cuantitativa** , que asocia un número a cada individuo, o **va-**

riable cualitativa , que coloca cada individu en una categoría. Ejemplos de variables cuantitativas asociadas a la clase: peso, altura o edad. El sexo o el grupo sanguíneo son en cambio variables cualitativas.

- Un concepto fundamental que utilizaremos con frecuencia corresponde a la **distribución** de una variable X asociada a un conjunto de datos. Describir la distribución de X corresponde a establecer la lista de los valores que toma X junto con la frecuencia con la que toma cada valor. Hablaremos de **frecuencia absoluta** de un valor para denotar el número de veces que aparece este valor en el conjunto de datos, mientras que la **frecuencia relativa** corresponde a la proporción (o el porcentaje) de veces que aparece este valor.

  En particular, una de las características interesantes de un conjunto de datos consiste en determinar si presenta mucha o poca **variabilidad**.

Ejemplo I.2.1 *Consideremos por ejemplo la distribución del grupo sanguíneo en una clase presentada en la tabla siguiente:*

Grupo	*Frec. absoluta*	*Frec. relativa*
A	*51*	*51/145=0.35*
B	*19*	*0.13*
O	*5*	*0.03*
AB	*70*	*0.49*

¿Qué representa la suma de la segunda columna (Frec. absoluta)? ¿Cuanto vale la suma de la tercera columna?

I.3. Tabulación y representaciones gráficas

Las representaciones gráficas son una herramienta fundamental para extraer información de forma visual de un conjunto de datos. Pueden ser mucho más útiles que procedimientos sofisticados que uno no domina...

I.3.1. Gráficas para variable cualitativa

Para un conjunto de datos descritos por una variable cualitativa, podemos realizar dos tipos de gráficas:

I.3.1.1. Diagrama de barras

Para cada valor que toma la variable en el conjunto y que indicamos en el eje horizontal, representamos en el eje vertical su frecuencia absoluta o relativa, en forma de una barra. En el caso del ejemplo I.2.1, obtenemos el diagrama de barra de la figura I.1. Cabe destacar que se suelen ordenar los valores de la variable por orden decreciente de frecuencias.

I.3.1.2. Diagrama de sectores

Si el conjunto no presenta demasiados valores distintos, también podemos utilizar el diagrama de sectores, donde cada valor ocupa un sector circular cuya área es proporcional a su frecuencia.

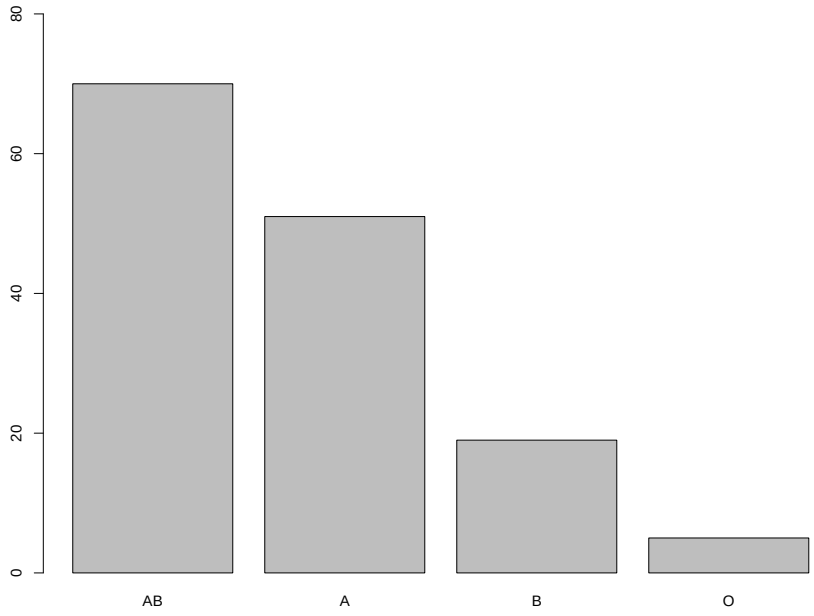

Figura I.1: Diagrama de barras, frecuencias absolutas, para el ejemplo I.2.1 del grupo sanguíneo,

Para el ejemplo I.2.1, calculemos el ángulo que ocupará el sector para cada uno de los valores AB, A, B, O. Por una regla de tres, deducimos que si el círculo entero (360 grados) representará el número total de datos en el conjunto, es decir 145 individuos, el valor AB con una frecuencia de 70 individuos deberá ocupar un sector de $70/145 \times 360 = 174^o$. Asimismo, el valor A ocupará 126^o, el valor B 48^o, mientras que el valor O ocupará solamente 12^o. El diagrama de sectores correspondiente se representa en la figura I.2.

I.3.2. Gráficas para una variable cuantitativa

Nos centramos ahora en variables cuantitativas. Los conjuntos que examinaremos se presentarán o bien en forma bruta: un fichero con una columna para cada variable, donde cada fila representa un individuo, o bien en forma ya tabulada, es decir donde los datos están agrupados.

Para datos agrupados, consideremos mediciones del contenido en nitrato de una muestra de agua:

Valor	Frecuencia	Valor	Frecuencia
0.45	1	0.49	8
0.46	2	0.50	10
0.47	4	0.51	5
0.48	8	0.51	8

También se puede representar gráficamente mediante un diagrama de barras esta distribución de frecuencias, indicando en el eje Ox los valores que puede tomar la

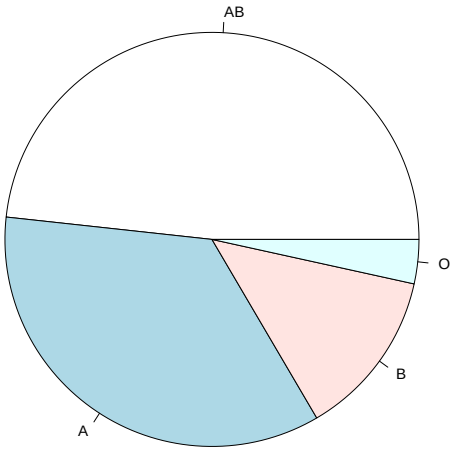

Figura I.2: Diagrama de sectores para el ejemplo I.2.1 del grupo sanguíneo,

variable y en el eje Oy sus frecuencias. Obtenemos así un diagrama de barras en el ejemplo de las mediciones de la concentración en nitrato, ver figura I.3.

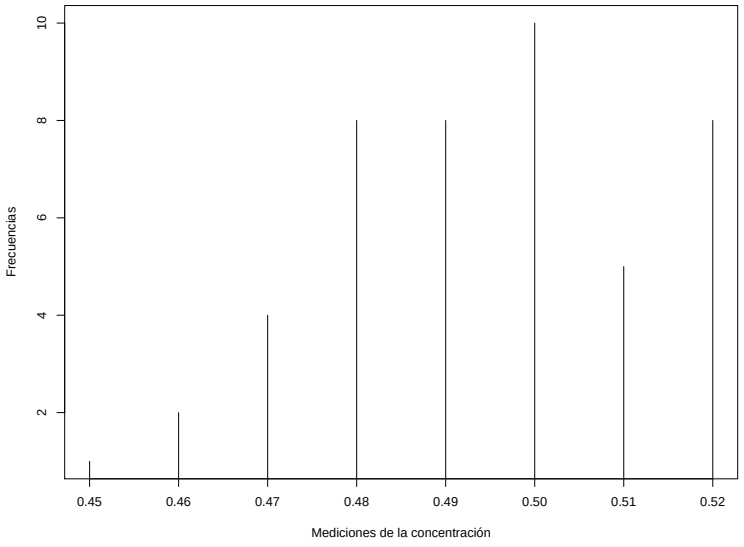

Figura I.3: Diagrama de barras para las concentraciones de nitrato

En el caso en que el conjunto presente muchas valores próximos pero distintos,

agrupamos los datos por clases, tal como lo veremos en los apartados siguientes.

I.3.2.1. Ejemplo: mediciones de la velocidad de la luz

Consideramos para ilustrar los conceptos que introduciremos en el resto del temael conjunto de datos de Newcomb (http://www.dmae.upct.es/∼mathieu). Newcomb fue el primero en conseguir ¡en 1882! una estimación bastante precisa de la velocidad de la luz. Las mediciones recogidas a continuación corresponden a los tiempos codificados que tardó un rayo de luz en recorrer el camino de ida y vuelta desde el laboratorio de Simon Newcomb situado en el Río Potomac hasta un espejo situado en la base del "Washington Monument", en total una distancia de 7400m. Para obtener los tiempos en nano segundos ($10^{-9}s$) no codificados, hay que añadir 24800 a cada dato.[1]

Tiempos codificados: 28, 26, 33, 24, 34, -44, 27, 16, 40, -2, 29, 22, 24, 21, 25, 30, 23, 29, 31, 19, 24, 20, 36, 32, 36, 28, 25, 21, 28, 29, 37, 25, 28, 26, 30, 32, 36, 26, 30, 22, 36, 23, 27, 27, 28, 27, 31, 27, 26, 33, 26, 32, 32, 24, 39, 28, 24, 25, 32, 25, 29, 27, 28, 29, 16, 23

Al observar estos datos, podemos realizar dos comentarios:

1. ¿Por qué Newcomb repitió tantas veces las mediciones, y no se limitó a realizar el experimento una vez? Porque los datos resultados del experimento presentan una cierta variabilidad: por mucho que haya intentado controlar las condiciones experimentales para mantenerlas constantes, el resultado es imprevisible. La medición está siempre perturbada por un "ruido" incontrolable...

2. ¿Qué hacer con estos datos? A la vista de estos datos, ¿cuál es el valor que podríamos tomar como la velocidad de la luz? Debemos encontrar un valor que sea representativo de las 66 mediciones realizadas. Se suele escoger la media, pero para asegurarnos de que ésta es representativa del conjunto, es útil establecer la tabla de frecuencias y visualizar el conjunto a través de un histograma, tal como lo vemos en la sección siguiente...

I.3.2.2. Tabla de frecuencias y histograma

En el caso en que el conjunto presente muchas valores próximos pero distintos, empezamos por agrupar los datos por clases: ordenamos los datos por orden creciente, dividimos el rango de los valores en clases de igual amplitud, y colocamos cada dato en la clase que le toca. A continuación podemos realizar el recuento de las frecuencias de cada clase.

¿Cuántas clases escoger? *La elección del número de clases es una problema que no admite una solución perfecta que sirva para todos los conjuntos de datos. Una regla aproximada llamada regla de Sturges consiste en escoger* $1+\log_2(n)$ *clases para un conjunto con n datos.*

Para le ejemplo de las mediciones de Newcomb, los datos ordenados se presentan como:

[1] Fuente: Moore, David S. and McCabe, George P. (1989). Introduction to the Practice of Statistics, W. H. Freeman and Company: New York, NY, pp 3-16.

Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dato	-44	-2	16	16	19	20	21	21	22	22	23	23	23	24	24
Pos	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Dato	24	24	24	25	25	25	25	25	26	26	26	26	26	27	27
Pos	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Dato	27	27	27	27	28	28	28	28	28	28	28	29	29	29	29
Pos	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Dato	29	30	30	30	31	31	32	32	32	32	32	33	33	34	36
Pos	61	62	63	64	65	66									
Dato	36	36	36	37	39	40									

Utilizamos por ejemplo clases de amplitud 5 empezando en -45 y acabando en 40, y realizamos el recuento de las frecuencias de cada clase:

Clase	Frecuencia	Clase	Frecuencia	Clase	Frecuencia
$]-45,-40]$	1	$]-15,-10]$	0	$]15,20]$	4
$]-40,-35]$	0	$]-10,-5]$	0	$]20,25]$	17
$]-35,-30]$	0	$]-5,0]$	1	$]25,30]$	26
$]-30,-25]$	0	$]0,5]$	0	$]30,35]$	10
$]-25,-20]$	0	$]5,10]$	0	$]35,40]$	7
$]-20,-15]$	0	$]10,15]$	0		

Cuando establecemos la tabla de frecuencias de una variable cuantitativa, indicamos también las **frecuencias acumuladas** de cada clase: la frecuencia absoluta (relativa) acumulada de una clase es el número (proporción) de datos que pertenecen a esta clase o a alguna clase anterior.

La tabla completa de frecuencias tal como nos la suele presentar un programa de estadística incluye las frecuencias absolutas y relativas así como las frecuencias acumuladas absolutas y relativas. Para el ejemplo de las mediciones de Newcomb, la tabla completa se puede ver en la Tabla I.1 más abajo.

Por otra parte, los datos tabulados se examinan con más comodidad a través de representaciones gráficas. En el eje Ox aparecen las clases y en el eje Oy las frecuencias, el diagrama resultante se llama histograma. En la figura I.4, aparece el histograma para las mediciones de Newcomb. Se pueden representar histogramas de frecuencias absolutas, relativas, absolutas acumuladas o relativas acumuladas.

I.3.2.3. Cómo interpretar un histograma

Las representaciones gráficas describen la distribución de la variable en el conjunto. Al examinarlos hay que que intentar contestar a las siguientes preguntas, para resumir las características de la distributión.

1. ¿ Es el histograma simétrico? Es decir, ¿aparece un punto central, respecto al cual, los valores se van repartiendo de manera aproximadamente simétrica? Esta es la situación clásica para un conjunto de mediciones: el valor central sería lo más representativo de lo que intentamos medir, y las mediciones van sobrevalorando e infravalorando de manera simétrica este valor. Si no consideramos los valores -44 y -2 en el conjunto de Newcomb, por ser muy diferentes

Clase	Frecuencias		Frec. Acumuladas	
	Absolutas	Relativas(%)	Absolutas	Relativas(%)
$]-45,-40]$	1	1.5	1	1.5
$]-40,-35]$	0	0.0	1	1.5
$]-35,-30]$	0	0.0	1	1.5
$]-30,-25]$	0	0.0	1	1.5
$]-25,-20]$	0	0.0	1	1.5
$]-20,-15]$	0	0.0	1	1.5
$]-15,-10]$	0	0.0	1	1.5
$]-10,-5]$	0	0.0	1	1.5
$]-5,0]$	1	1.5	2	3.0
$]0,5]$	0	0.0	2	3.0
$]5,10]$	0	0.0	2	3.0
$]10,15]$	0	0.0	2	3.0
$]15,20]$	4	6	6	9
$]20,25]$	17	25.7	23	34.7
$]25,30]$	26	39.3	49	74
$]30,35]$	10	15.3	59	89.3
$]35,40]$	7	10.7	66	100
TOTAL	66	100.0		

Tabla I.1: Tabla de frecuencias, mediciones de Newcomb.

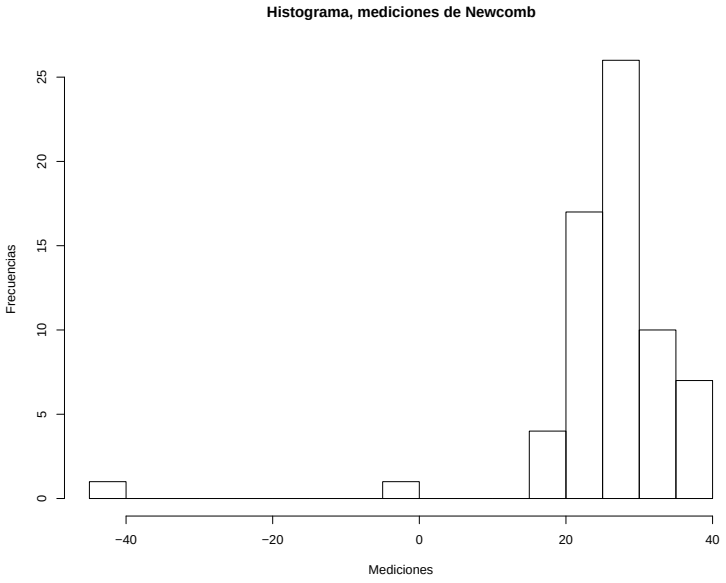

Figura I.4: Histograma para las mediciones de Newcomb

del resto del conjunto, podemos decir que la distribución de las mediciones es aproximadamente simétrica.

2. ¿Posee la distribución colas largas?

3. ¿Posee el histograma un máximo claro único? En este caso hablamos de histograma unimodal.

4. ¿Aparecen datos atípicos?, es decir datos que se alejan del patrón global de los datos. Para el conjunto de Newcomb, dos datos aperecen claramente atípicos: -44 y -2, mientras que las 64 mediciones restantes están entre 15 y 40. Al detectar datos atípicos, debemos comprobar que no se deban a errores tipográficos, y buscar si están asociados a unas circunstancias experimentales especiales. Podremos entonces decidir corregirlos u omitirlos del estudio.

5. ¿Donde localizamos el centro aproximado de los datos?

6. ¿Presentan los datos mucha dispersión?, lo que se traduce en la forma puntiaguda o chata del histograma. En el caso de mediciones, el hecho de que los datos estén concentrados revela que se consiguió una buena regularidad en el proceso de medición...

En la figura I.5, presentamos varios patrones de histogramas.

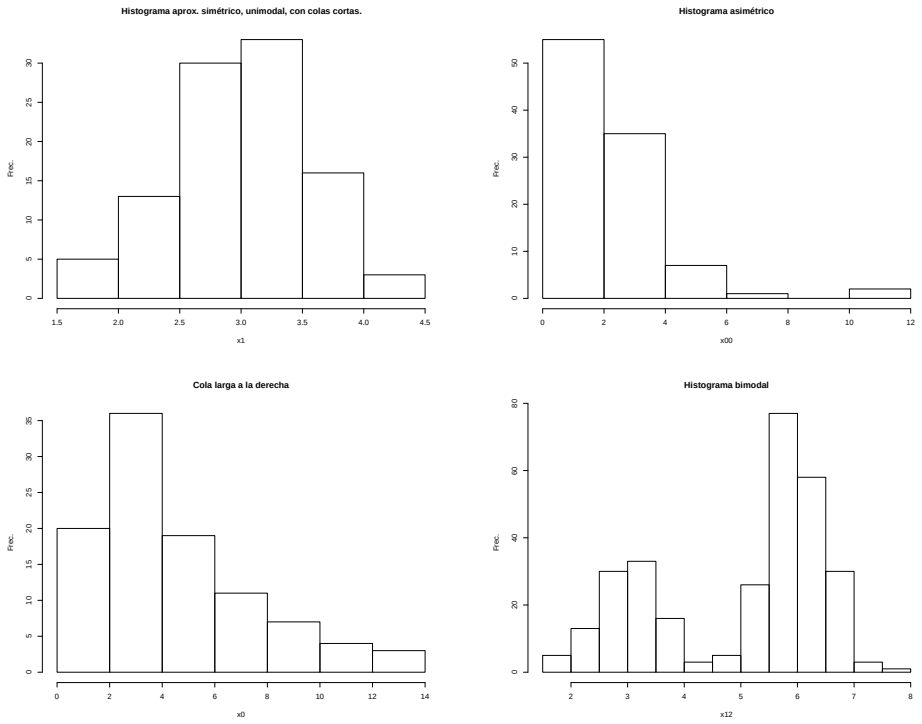

Figura I.5: Distintos patrones de histogramas.

I.4. Medidas numéricas

Para variables cuantitativas, se suele acompañar las representaciones gráficas de las distribuciones con medidas numéricas que proporcionen un resumen de sus características principales. Existen medidas numéricas para contestar a cada pregunta

(y alguna más...) planteadas en el apartado anterior a la hora de examinar el histograma. Nos limitaremos a las medidas de centro y de dispersión, es decir las que proporcionen una respuesta a las preguntas 5 y 6.

I.4.1. Medidas de centro

Buscamos ahora medidas numéricas que sean representativas del centro del conjunto de dato.

I.4.1.1. La media:

Si $x_1, \ldots, x_n$ son los datos, sabemos todos que la media es

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

En el caso en que los datos ya están tabulados y tenemos los valores distintos $x_1, \ldots, x_m$ junto con sus frecuencias $n_1, \ldots, n_m$, deberemos tener en cuenta estas frecuencias para el cálculo de la media:

$$\bar{x} = \frac{n_1 x_1 + \cdots + n_m x_m}{(n_1 + \ldots + n_m)}.$$

En este caso, ¿cuántos individuos tenemos en el conjunto?

Nota: Representa el centro de gravedad de los datos, es decir que si a cada dato le damos un peso unidad, la media representa el punto en el que el conjunto está en equilibrio.
En particular, deducimos que la media es muy sensible a datos atípicos en el conjunto de datos: si añado un dato (peso) alejado del centro de gravedad, el punto de equilibrio debe desplazarse mucho hacia éste para que se mantenga el equilibrio.

Para paliar estos inconvenientes, se considera también la mediana:

I.4.1.2. La mediana:

La mediana es el punto que deja el 50 % de los datos a su izquierda y el otro 50 % a su derecha. Es una medida de centralización más adecuada que la media en el caso en que la distribución de los datos es asimétrica (lo que se ve en el histograma) o si hay datos atípicos. Si la distribución es simétrica, la media y la mediana coinciden.

Para calcular la mediana de un conjunto de n datos, $x_1, x_2, \ldots, x_n$, empiezo por ordenar los datos por orden creciente. La mediana es el dato ordenado n^o $(n+1)/2$.

Ejemplo: 125, 129, 134, 185, 200. La mediana es el dato ordenado n^o 3, y es igual a 134.

11, 15, 20, 23: la mediana es el dato ordenado n^o 2.5, que tomamos por convención igual al punto medio entre el dato n^o 2 y el dato n^{o}3. En este caso, la mediana es igual a 17.5.

La mediana no es sensible a datos atípicos, para convencerse de ello, se puede considar el ejemplo anterior donde se sustituye el valor 23 por 1000... La mediana no cambia... Por lo tanto, la mediana es más representativa del centro del conjunto si hay algun dato atípico o si la distribución es algo asimétrica...

I.4.2. Medidas de dispersión

I.4.2.1. La desviación típica

Mide lo lejos que están situados los datos respecto de su centro de gravedad, la media. Empezamos por definir **la varianza**:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}, \tag{I.1}$$

que representa aproximadamente el promedio de las distancias al cuadrado entre los datos y su media. La desviación típica s es la raíz cuadrada de s^2.

Para calcularla en la práctica se suele preferir la fórmula siguiente

$$s^2 = \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2), \tag{I.2}$$

donde $\overline{x^2}$ representa la media de los datos que hemos previamente elevado al cuadrado, mientras que $(\bar{x})^2$ representa el cuadrado del valor de la media. Como ejemplo, supongamos que quiero calcular la varianza de los datos siguientes 4, 5,5, 6,5, 8.

Necesito por una parte $\bar{x}$, que calculo como $\bar{x} = (4 + 5{,}5 + 6{,}5 + 8)/4 = 6$, y por otra parte $\overline{x^2}$ que calculo como $\overline{x^2} = (4^2 + 5{,}5^2 + 6{,}5^2 + 8^2)/4 = 38{,}125$. Por lo tanto, deduzco

$$s^2 = \frac{3}{4}[38{,}125 - (6)^2] = 2{,}8333.$$

Naturalmente, la desviación típica es representativa de la dispersión del conjunto de datos solo si la media es representativa de su centro.

Es bueno ser consciente de que la desviación típica, al igual que la media, se expresa en las mismas unidades que los datos, mientras que la varianza en $(unidades)^2$.

Una medida alternativa de dispersión que puede ser más representativa en el caso en que la distribución es asimétrica o en presencia de datos atípicos, es el **rango intercuartílico**.

I.4.2.2. El rango intercuartílico (RIC)

Hemos definido la mediana como el punto que separa el conjunto en dos partes de mismo tamaño. Definimos de la misma manera los cuartiles como los puntos que separan el conjunto en cuatro partes de mismo tamaño. El primer cuartil Q_1 deja el 25 % de los datos ordenados a su izquierda, y el otro 75 % a su derecha, mientras que el tercer cuartil Q_3 deja el 75 % de los datos ordenados a su izquierda, y el otro 25 % a su derecha. Por lo tanto el par (Q_1, Q_3) nos proporciona información sobre la dispersión presente en los datos: cuanto más alejados estén los cuartiles, más dispersos están los datos. Por ello, calculamos el rango intercuartílico RIC como la diferencia entre Q_3 y Q_1.

Para calcular los cuartiles, empezamos por calcular la mediana Me de los datos. El primer cuartil es la mediana del grupo de datos que queda a la izquierda de Me (Me excluida), mientras que el tercer cuartil se calcula como la mediana del grupo que queda a su derecha (Me excluida).

El RIC también se utiliza para detectar datos atípicos:
Regla: Se consideran como atípicos los datos que son menores de $Q_1 - 1{,}5 \times RIC$, o mayores de $Q_3 + 1{,}5 \times RIC$.

I.4.3. Un resumen gráfico: el diagrama de caja-bigotes

El diagrama de caja-bigotes es un resumen gráfico que permite visualizar, para un conjunto de datos, la tendencia central, la dispersión y la presencia posible de datos atípicos. Para realizarlo se necesita calcular la mediana, el primer cuartil, y el tercer cuartil de los datos:

El diagrama de caja-bigotes presenta de manera gráfica estas informaciones, tal como está recogida en la figura I.6.

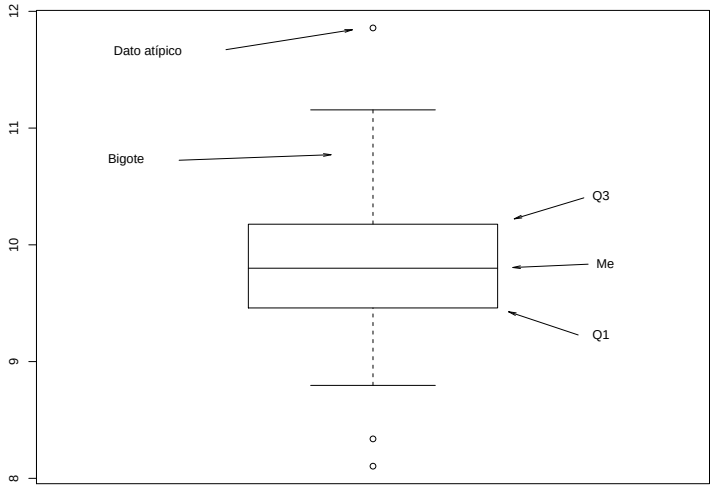

Figura I.6: Componentes del diagrama caja-bigotes

Los segmentos 1.5 RIC (llamados bigotes) se recortan hasta : el dato del conjunto inmediatamente superior a $Q_1 - 1{,}5 \times RIC$ para el bigote inferior, y el dato inmediatamente inferior a $Q_3 + 1{,}5 \times RIC$, para el bigote superior.

La mayor utilidad de los diagramas caja-bigotes es para comparar dos o más conjuntos de datos.

Ejemplo

La puntuación de los equipos de la liga española al final de las temporadas 01/02 y 02/03 en primera división se pueden comparar con un diagrama caja-bigotes, como aparece en la figura I.7

Comentarios: No hay datos atípicos, es decir que no hay equipo que se haya destacado por arriba o por abajo del resto de los equipos. Hay más diferencia de puntos entre el primer y el último clasificado para la liga 02/03 que en la liga anterior. Los equipos del tercer cuarto de la clasificación están muy apelotonados en la liga 02/03.

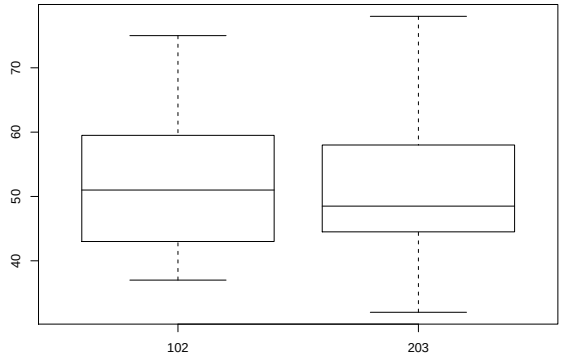

Figura I.7: Comparación puntuación final, temporadas 01/02 y 02/03

I.5. Ajuste por mínimos cuadrados

I.5.1. Planteamiento

Es muy normal considerar más de una variable asociada a un experimento. En este caso, más que la distribución de cada variable por separado, nos puede interesar en particular las relaciones que existan entre ellas. Nos centraremos aquí en el caso en que distinguimos una variable llamada "respuesta", cuya amplitud depende de los valores de otras variables llamadas "explicativas", y aprenderemos cómo deducir un modelo para la evolución de la primera en función de estas últimas.

Hay dos utilidades principales al disponer de un modelo: podemos primero explicar la manera en la que cambios en los valores de una variable explicativa induce cambios en el valor de la variable respuesta. Por ejemplo, si pienso que la temperatura media Y en agosto en San Javier evoluciona en función del año según el modelo:

$$Temperatura = -582{,}5 + 0{,}31 año,$$

deduciré que en promedio, la temperatura media en agosto aumenta de 0.3 grados cada año.

Por otra parte, si dispongo de un modelo para la evolución de la variable respuesta, me permite también realizar predicciones del valor que tomará para valores de las explicativas que no hemos observado.

Acabamos esta sección de presentación con cuatro ejemplos con datos reales tomados de campos diferentes. Las nubes de puntos correspondientes están presentadas en la figura I.8

- Estudio de la resistencia del cemento en función del tiempo de fraguado en días. Fuente: Hald, A. (1952) *Statistical theory for engineering applications*, Wiley & Sons New-York, pág 541. ¿Cómo evoluciona la resistencia de piezas de cemento en función del tiempo de fraguado? ¿Cuánto tiempo hay que esperar para conseguir el 90 % de la resistencia máxima? Este es el tipo de preguntas a las que podemos contestar con el estudio de este conjunto de datos.

- Todos los años Venecia se inunda durante las "acqua alta". Sin embargo, parece que el nivel máximo al que llega el mar está cada año más alto, haciendo temer por la conservación de la ciudad y de sus monumentos. Es por lo tanto de interés estudiar la evolución del nivel máximo del mar en función del año. Fuente: Smith, R.L (1986) "Extreme value theory based on the r largest annual events", *Journal of Hydrology*, **86**.

- Evolución de la producción mundial de petróleo desde 1880. Fuente: Data and Stories Library http://lib.stat.cmu.edu/DASL/.

- En 1929, Edwin Hubble investigó la relación entre la distancia de una galaxia a la tierra y la velocidad con la que está alejándose. En efecto se piensa que las galaxias se alejan como consecuencia del "Big Bang". Hubble pensó que disponiendo de un modelo que relacionara la velocidad de recesión con la distancia a la tierra proporcionaría información sobre la formación del universo y sobre lo que podría pasar en el futuro. Los datos recogidos incluyen distancias en megaparsecs (1 megaparsec= 3.26 años luz) y velocidad de recesión en km/s. Fuente: Data and Stories Library, http://lib.stat.cmu.edu/DASL.

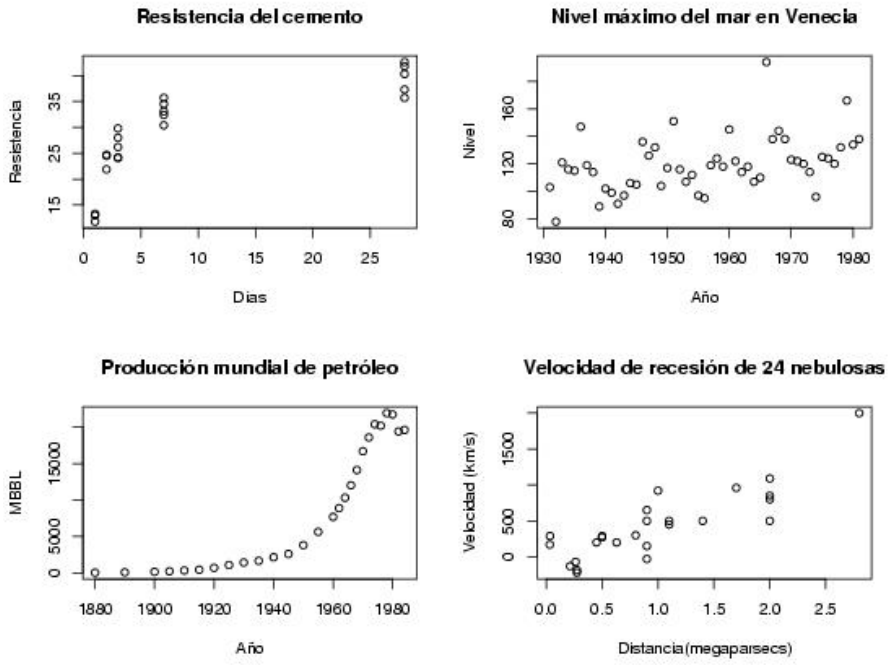

Figura I.8: Cuatro ejemplos de conjuntos de datos

I.5.2. Criterio de mínimos cuadrados

Para ilustrar las nociones nos limitamos primero al caso de una variable respuesta que llamaremos Y y una variable explicativa que llamaremos X.

Los datos se presenta en forma de pares:

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

es decir que, para varios valores X observamos los valores correspondientes de Y. Para visualizar el conjunto recurrimos a la nube de puntos, también llamada diagrama de dispersión, en el que representamos los pares $(x_i, y_i),\ i = 1, \cdots, n$, en unos ejes $Ox,\ Oy$, ver figura I.9

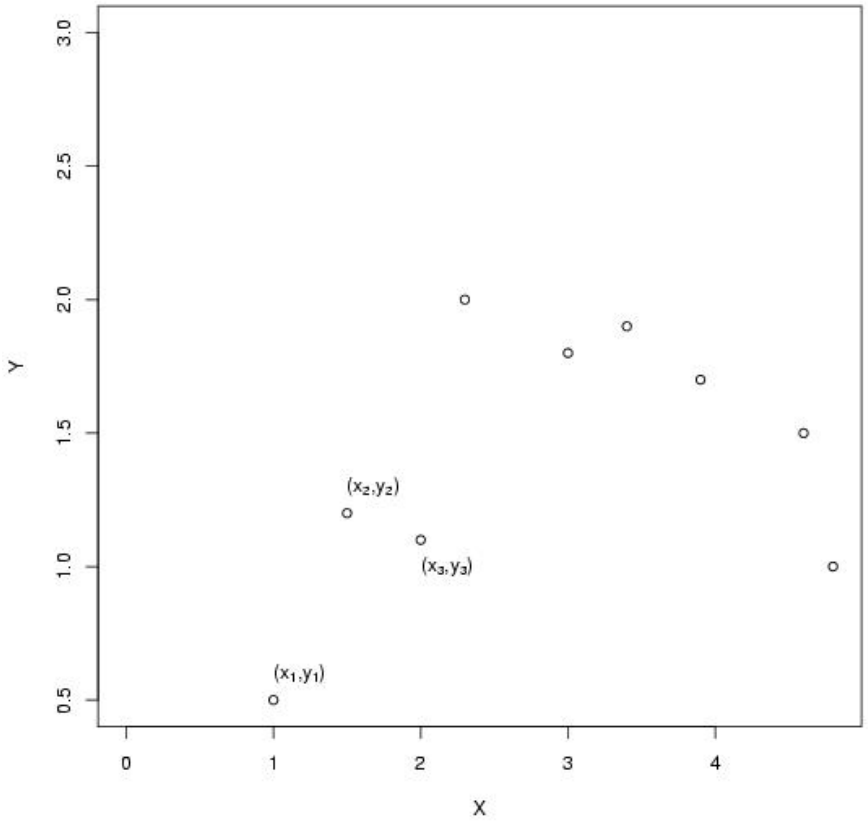

Figura I.9: Ejemplo de nube de puntos

Por conocimientos previos sobre el fenómeno que estudiamos o por la propia nube de puntos, decidimos ajustar a ésta una curva de una determinada forma funcional: podría ser por ejemplo una recta, de ecuación $Y = aX + b$, o una parábola $Y = a_0 + a_1X + a_2X^2$. La forma de la curva está fijada pero intervienen en la ecuación constantes, también llamadas parámetros, cuyo valor tenemos que ajustar para obtener el "mejor" ajuste posible: en el caso de la recta, debemos encontrar los valores de la pendiente b y de la ordenada en el origen a.

En una formulación general, escogemos una familia paramétrica de funciones

$$x \mapsto f(\theta, x) \quad \theta = (\theta_1, \ldots, \theta_k), \tag{I.3}$$

donde θ es el vector de parámetros. Buscar la función de la familia que mejor se ajusta

a la nube de puntos es equivalente a encontrar el valor $\hat{\theta}$ de θ, que corresponde a esta función.

Debemos ahora dar sentido a la noción de "mejor"; debemos fijarnos un criterio que nos permita decidir que una función de la familia se ajusta mejor a la nube de puntos que otra. El criterio que seguimos en este tema es el de **mínimos cuadrados**.

Definimos la suma de cuadrados asociada a una función de la familia como la suma de los cuadrados de las distancias verticales entre la curva correspondiente y los datos observados de la nube de puntos. Tal como viene reflejado en la figura I.10, la distancia vertical entre por ejemplo el punto (x_3, y_3) y la curva es $y_3 - f(\theta, x_3)$, por lo tanto la suma de cuadrados se escribe

$$SC(\theta) = (y_1 - f(\theta, x_1))^2 + (y_2 - f(\theta, x_2))^2 + \cdots + (y_n - f(\theta, x_n))^2. \tag{I.4}$$

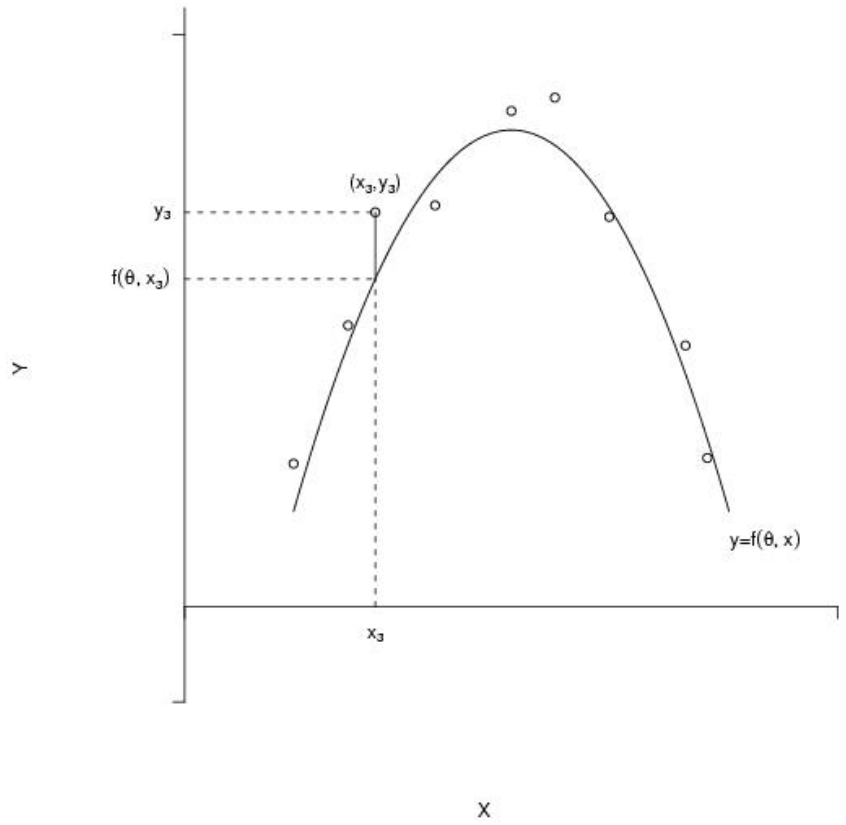

Figura I.10: Ajuste de una curva a la nube de puntos.

Buscamos el valor $\hat{\theta}$ de θ que minimiza la cantidad $\theta \mapsto \theta$, en muchos casos, es imposible encontrar este mínimo explícitamente y tenemos que recurrir a algoritmos numéricos. Nos centraremos en este tema en elcaso en que la forma paramétrica de f es particularmente simple y permite el cálculo explícito de $\hat{\theta}$.

Supongamos que hemos ajustado la curva, es decir que hemos encontrado el valor $\hat{\theta}$ de θ que minimiza la suma de cuadrados, introduzcamos unos cuantos términos:

- La curva de ecuación $y = f(\hat{\theta}, x)$ se llama la **curva ajustada**.
- Los ordenadas de la curva ajustada correspondientes a los datos observados, es decir los valores $\hat{y}_1 = f(\hat{\theta}, x_1), \ldots, y_n = f(\hat{\theta}, x_n)$ se llaman los **valores ajustados**.
- Las distancias verticales entre los puntos observados y la curva ajustada se llaman los **residuos** $e_1, \ldots, e_n$. Tenemos

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n.$$

- La suma de cuadrados

$$SC(\hat{\theta}) = \sum_{i=1}^{n} e_i^2$$

se llama **suma de cuadrados residuales**.

- Calcularemos en algunas ocasiones la varianza de los residuos, también llamada varianza residual

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i - \bar{e})^2.$$

I.5.3. Casos concretos

Describimos ahora con más detalle unos pocos casos concretos en los que es posible obtener de manera explícita la expresión de $\hat{\theta}$, que minimiza la suma de cuadrados residuales. Estos casos corresponden todos a la llamada regresión lineal: son casos para los cuales los parametros $(\theta_1, \ldots, \theta_k)$ intervienen de manera lineal en la ecuación (I.3)

I.5.3.1. Recta $y = ax + b$

El caso más utilizado de ajuste por mínimo por mínimos cuadrados al ajuste por una recta, es decir cuando consideramos una variable explicativa X y buscamos ajustar un modelo de la forma

$$Y = aX + b.$$

Corresponde al caso en que θ consta de dos parámetros a y b, y la función f descrita en la sección I.5.2 es $f(\theta, x) = ax + b$. En este caso, decimos que el ajuste corresponde a la regresión lineal simple.

En el caso en que la pendiente a es positiva, hablamos de asociación positiva entre X e Y: cuando crece X, crece Y, cuando decrece X, decrece Y, y viceversa. En cambio, si la pendiente a es negativa, hablamos de asociación negativa entre X e Y (cuando crece una variable, decrece la otra).

a). Obtención de la recta ajustada La suma de cuadrados se escribe

$$SC(\theta) = SC(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2,$$

Los candidatos a alcanzar el mínimo de esta función satisfacen

$$\frac{\partial}{\partial a}SC(a,b) = 0$$
$$\frac{\partial}{\partial b}SC(a,b) = 0.$$

Deducimos de unas cuantas manipulaciones algebraicas que las soluciones a este sistema de ecuaciones son

$$\begin{aligned}\hat{a} &= \tfrac{\overline{xy}-\bar{x}\bar{y}}{\overline{x^2}-(\bar{x})^2}\\ \hat{b} &= \bar{y} - \hat{a}\bar{x}.\end{aligned}$$

Introducimos la cantidad

$$s_{xy} = \frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y}), \tag{I.5}$$

que llamamos la **covarianza** de X e Y. El coeficiente $\hat{a}$ se puede por lo tanto escribir como

$$\hat{a} = \frac{s_{xy}}{s_x^2},$$

donde s_x^2 es la varianza de X que introdujimos en la sección I.4.2.1. Con estas notaciones, se puede escribir la ecuación de la recta ajustada en una forma compacta:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Nota La covarianza es una cantidad que puede ser positiva o negativa. De hecho tiene el mismo signo que la pendiente de la recta ajustada. Por lo tanto, si la covarianza es positiva, Y y X presentan una asociación positiva mientras que, si la covarianza es negativa Y y X presentan una asociación negativa.

b). Bondad del ajuste Para la regresión lineal simple, los residuos son

$$\begin{aligned}e_1 &= y_1 - f(\hat{\theta}, x_1) = y_1 - \hat{a}x_1 - \hat{b}\\ \vdots &= \vdots\\ e_n &= y_n - f(\hat{\theta}, x_n) = y_n - \hat{a}x_n - \hat{b},\end{aligned}$$

y tienen las siguientes propiedades

Propiedades de los residuos

- La media de los residuos es nula.
  Demostración:

$$\begin{aligned} \bar{e} = \frac{e_1 + \cdots + e_n}{n} &= \frac{1}{n}[(y_1 + \ldots + y_n) - \hat{a}(x_1 + \ldots + x_n) - n\hat{b}] \\ &= \bar{y} - \hat{a}\bar{x} - \hat{b} = 0 \end{aligned}$$

- Se puede demostrar sin dificultad que la varianza residual se escribe como

$$s_e^2 = s_y^2 \left(1 - \frac{(s_{xy})^2}{s_x^2 s_y^2}\right). \tag{I.6}$$

  De esta ecuación deducimos que la cantidad $\frac{(s_{xy})^2}{s_x^2 s_y^2}$ puede medir la calidad del ajuste. De hecho le damos un nombre especial:

  Definición I.5.1 *La cantidad* $r = \frac{s_{xy}}{s_x s_y}$ *se llama coeficiente de correlación (de Pearson) de* X *e* Y*.*
  La cantidad $R^2 = \frac{(s_{xy})^2}{s_x^2 s_y^2}$ *se llama coeficiente de determinación.*

Propiedades de r y R^2 De la fórmula $s_e^2 = s_y^2(1 - R^2)$, ver (I.6), deducimos

- R^2 está siempre comprendido entre 0 y 1, y cuanto más cercano esté de 1, mejor es el ajuste, puesto que corresponderá a una varianza residual menor. En particular, deducimos que si $R^2 = 1$, la varianza residual s_e^2 es nula, lo que quiere decir que la dispersión de los residuos es nula: todos los residuos son iguales, y por lo tanto iguales a su media, que vale 0, todos los puntos de la nube están situados en la recta, el ajuste es perfecto. Se suela considerar un valor de R^2 mayor que 0.8 como correspondiente a un ajuste bueno, mientras que un valor mayor que 0.9 corresponde a un ajuste muy bueno.

- Puesto que $R^2 = r^2$ y $0 \leq R^2 \leq 1$, deducimos que el coeficiente de correlación r está siempre comprendido entre -1 y 1. Si $r = \pm 1$, el ajuste de los puntos observados por una recta es perfecto. El coeficiente de correlación se interpreta en general como una cantidad que cuantifica la asociación lineal que existe entre dos variables: cuanto más cerca de ± 1, más se aproxima la nube de puntos a una recta.

  Además por la definición de r, sabemos que r es del mismo signo de la covarianza. Por lo tanto, si r es positivo y cercano a 1, los datos apoyan la existencia de una asociación lineal positiva entre las dos variables, mientras que si es negativo y cercano a -1, presentan una asociación lineal negativa.

  Sin embargo, es necesario tener precaución a la hora de interpretar valores del coeficiente de correlación: sólo es un resumen, fiable en el caso en que está próximo a ± 1 para indicar que existe una fuerte asociación lineal entre las variables pero mucho menos fiable si toma un valor alejado de ± 1. Anscombe (1973), "Graphs in statistical analysis", *American Statistician*, **27**, pp 17-21,

construyó cuatro conjuntos de datos artificiales que dan lugar al mismo coeficiente de correlación y a las mismas rectas de regresión, pero cuyos aspectos son completamente diferentes. Los datos se presentan en el apéndice, y se deja su estudio en ejercicio.

c). Un ejemplo Para ilustrar el procedimiento que se sigue para calcular los valores de $\hat{a}$ y $\hat{b}$, consideremos el ejemplo muy sencillo a continuación:

Queremos estudiar la relación entre el peso y la altura en un grupo de individuos. Los datos son

Peso(kg)	54	70	65	78	68	85	Y
Altura(cm)	160	170	172	185	160	175	X

Se deja en ejercicio al lector la representación de este conjunto a través de una nube de puntos... Buscamos ajustar una recta a la nube y pasamos a calcular la ecuación de la recta de regresión que en su forma compacta se escribe

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Para calcular s_{xy} y s_x^2 utilizaremos las fórmulas (I.2) y (I.5), necesitamos por lo tanto $\bar{x}$, $\overline{x^2}$, $\bar{y}$, $\overline{y^2}$ y $\overline{xy}$. Tenemos

$$\bar{x} = \tfrac{160+170+...+175}{6} = 170{,}33, \qquad \bar{y} = \tfrac{54+70+...+85}{6} = 70,$$

$$\overline{x^2} = \tfrac{160^2+170^2+...+175^2}{6} = 29089, \quad \overline{y^2} = \tfrac{54^2+70^2+...+85^2}{6} = 4995{,}7,$$

$$\overline{x\,y} = \tfrac{160\times 54+170\times 70+...+175\times 85}{6}$$

Deducimos que

$$\begin{aligned} s_x^2 &= \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2) = \frac{6}{5}[29089 - (170{,}33)^2] \simeq 90{,}7, \\ s_y^2 &= \frac{n}{n-1}(\overline{y^2} - (\bar{y})^2) = \frac{6}{5}[4995{,}7 - (70)^2] \simeq 144{,}8, \\ s_{xy} &= \frac{n}{n-1}(\overline{xy} - (\bar{x})(\bar{y})) = \frac{6}{5}[11984{,}2 - 170{,}33 \times 70] \simeq 73. \end{aligned}$$

La ecuación de la recta es por lo tanto $y - 70 = \frac{73}{90{,}7}(x - 17033)$, es decir

$$y = 0{,}80x - 67{,}1.$$

El modelo teórico propuesto para relacionar el Peso y la Altura es $Peso \simeq 0{,}8 Altura - 67{,}1$.

En cuanto a la bondad del ajuste, tenemos que

$$R = \frac{s_{xy}}{s_x\, s_y} = \frac{7}{3}\sqrt{90{,}7}\sqrt{114{,}8} \simeq 0{,}715,$$

lo que implica que $R^2 \simeq 0{,}51$, un ajuste malo.

d). Predicción Tal como lo mencionamos en la introducción del tema, si disponemos del modelo ajustado podemos utilizarlo para predecir el valor de la respuesta para valores no observados de X:

Si x_0 es un valor no observado, nuestra predicción del valor de Y será

$$y_{x_0} = \hat{a}\, x_0 + \hat{b}.$$

Si consideramos el ejemplo de la relación entre peso y altura del apartado anterior, podemos contestar a la pregunta ¿a qué peso correspondería una altura de 180cm? Sustituimos x por 180 en la ecuación de la recta ajustada, y encontramos que el peso asociado sería $0{,}80 \times 180 - 67{,}1 \simeq 76{,}9kg$.

Sin embargo, debemos tener mucho cuidado al extrapolar nuestro modelo fuera del rango de valores de X que hemos observado, al no disponer de valores fuera de este rango, tampoco sabemos si el modelo deducido seguirá valido. Para el ejemplo de los pesos, si queremos utilizar el modelo ajustado para saber a qué peso correspondería la altura de un niño de 80cm por ejemplo, obtenemos $0{,}80 \times 80 - 67{,}1 \simeq -3{,}1kg$, ¡lo que no tiene sentido!

Nota. El motivo por el cual, muy a menudo una recta suele ajustarse bastante bien a una nube de puntos, corresponde a que la fórmula de Taylor nos dice que localmente, cualquier función derivable se puede aproximar por una recta: aunque la relación entre Y y X no sea lineal sino de la forma $Y = f(\theta, X)$, f general, si f es derivable y observamos valores de X no muy dispersos alrededor, f se comporta aproximadamente como la tangente en un X central.

I.5.3.2. Recta forzada por el origen

Hay situaciones en las que pensamos recurrir a un ajuste lineal, pero sabemos por motivos físicos que un valor de X nulo corresponde necesariamente a un valor de Y nulo también. En este caso, no tenemos por que considerar todas las rectas, sino podemos restringirnos a las rectas que pasan por el origen $(0, 0)$. La ecuación de una recta forzada por el origen es

$$y = ax.$$

Dos ejemplos de situaciones en las que un valor nulo de X implica un valor nulo de Y:

- Medimos la evolución en función del tiempo (X) de la concentración (Y) de un producto que se va creando en una reacción química. Cuando empezamos la reacción $X = 0$, todavía no puede haber producto, por lo tanto $Y = 0$.
- Queremos medir el tiempo t que tarda un objeto que soltamos desde una altura h, en alcanzar el suelo. La relación física proporcionada por la teoría es $h = gt^2$, donde g es la constante de la gravedad. Si queremos comprobar que los datos empíricos confirman esta relación, buscaremos si es cierto que

  $$t = \frac{1}{\sqrt{g}}\sqrt{h}.$$

  Consideraremos $X = \sqrt{h}$, $Y = t$, y buscaremos ajustar una recta $y = ax$.

Las fórmulas que vimos para el caso de una recta general ya no son válidas. Calculemos la ecuación de la recta forzada por el origen: disponemos de n pares de datos $(x_1, y_1), \ldots, (x_n, y_n)$, puesto que la función que buscamos ajustar es $f(\theta, x) = ax$, $\theta = a$ y la suma de cuadrados de la fórmula (I.4) se escribe

$$SC(\theta) = SC(a) = \sum_{i=1}^{n} (y_i - ax_i)^2.$$

El candidato a minimizar $SC(a)$ satisface la ecuación $\frac{dSC(a)}{da} = 0$. Calculamos

$$\frac{dSC(a)}{da} = \sum_{i=1} -x_i \, 2(y_i - ax_i) = 2[-\sum_{i=1} x_i y_i + a \sum_{i=1}^{n} x_i^2].$$

Por lo tanto, la solución a la ecuación $\frac{dSC(a)}{da} = 0$ es

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\overline{xy}}{\overline{x^2}}.$$

Puesto que la derivada segunda de $SC(a)$ es positiva, se trata efectivamente de un mínimo.

I.5.3.3. Algunas transformaciones útiles

Sólo hemos hemos descrito cómo calcular la curva ajustada para dos familias específicas de funciones $y = ax$ e $y = ax + b$. Para una especificación más general de la función f que queremos ajustar, se recurre a algoritmos numéricos para encontrar el valor de los parámetros que minimicen la suma de cuadrados $SC(\theta)$.

Sin embargo, hay algunos tipos de modelos lineales que se pueden abordar con los resultados del caso lineal después de realizar unas transformaciones convenientes.

a). Modelo exponencial Supongamos que queremos ajustar un modelo exponencial a una nube de puntos. La ecuación de las funciones que consideramos son $y = be^{ax}$, con $b > 0$. En el caso en que a es positivo, modelizamos un crecimiento exponencial, mientras que, si a es negativa, modelizamos un decrecimiento exponencial.

La relación entre Y y X es altamente no lineal, sin embargo una simple transformación puede llevarlo a un modelo lineal:

Modelo teórico original		Modelo transformado
$y = be^{ax}$	$\xrightarrow{\text{cojo } \ln}$	$\ln(y) = \ln(b) + ax$
		$y' = b' + a'x'$

Si introducimos las variables tranformadas $Y' = \ln(Y)$, y $X' = X$, éstas satisfacen una relación lineal: $Y' = a'X' + b'$.

Nuestro procedimiento para ajustar un modelo exponencial consistirá por lo tanto en

1. Calculamos los datos transformados, es decir pasar de

X	x_1	x_2	$\dots$	x_n
Y	y_1	y_2	$\dots$	y_n

$y = be^{ax}$

a

X'	x_1	x_2	$\dots$	x_n
Y'	$\ln(y_1)$	$\ln(y_2)$	$\dots$	$\ln(y_n)$

$y' = a'x' + b'$

2. Ajustamos una recta a las variables transformadas, encontramos $y' = \hat{a}'x' + \hat{b}'$.
3. Volvemos al modelo original, haciendo la transformación inversa (en este caso exponencial)

$$y' = \hat{a}'x' + \hat{b}' \xrightarrow{\text{cojo exp}} y = e^{\hat{a}'x'+\hat{b}'} = e^{\hat{b}'}e^{\hat{a}'x'}.$$

Ejemplo. Queremos ajustar un modelo exponencial a los siguientes datos

X	2.3	5	7.1	8
Y	2.92	3.69	6.19	6.36

Transformamos los datos:

X'	2.3	5	7.1	8
$Y' = \ln(Y)$	1.07	1.31	1.82	1.85

Ajustamos una recta a los datos transformados, calculando $\bar{x'}$, $\overline{x'^2}$, $\bar{y'}$, $\overline{y'^2}$ y $\overline{x'y'}$, para obtener $\hat{a'}$ y $\hat{b'}$: $y' = 0{,}148x' + 0{,}682$, es decir que $\ln(y) = 0{,}148x + 0{,}682$, lo que implica que

$$y = e^{0{,}148x}e^{0{,}682} = 1{,}18e^{0{,}148x}.$$

b). Modelo potencial El modelo potencial es de la forma $y = bX^a$. La forma de la nube de puntos correspondiente depende del valor de a. La transformación que utilizamos es la misma que para el modelo exponencial: aplicamos los logaritmos.

Modelo teórico original		Modelo transformado
$y = bx^a$	$\xrightarrow{\text{cojo ln}}$	$\ln(y) = \ln(b) + a\ln(x)$
		$y' = b' + a'x'$

Introducimos las variables tranformadas $Y' = \ln(Y)$, y $X' = \ln(X)$, éstas satisfacen una relación lineal: $Y' = a'X' + b'$. Seguimos los mismos pasos que en el apartado anterior con los datos transformados.

Ejemplo. Queremos ajustar un modelo potencial a los siguientes datos

X	3	7.34	20.1	54.6
Y	10.3	13.5	18.2	24.5

Transformamos los datos:

$X' = \ln(X)$	1.1	2	3	4
$Y' = \ln(Y)$	2.3	2.6	2.9	3.2

Ajustamos una recta a los datos transformados, calculando $\bar{x'}$, $\overline{x'^2}$, $\bar{y'}$, $\overline{y'^2}$ y $\overline{x'y'}$, para obtener $\hat{a'}$ y $\hat{b'}$: $y' = 0{,}298x' + 2{,}006$, es decir que $\ln(y) = 0{,}298\ln(x) + 2{,}006$, lo que implica que

$$y = e^{0{,}298\ln(x)}e^{2{,}006} = 7{,}433x^{0{,}298}.$$

Apéndice

A continuación se presentan los datos de Anscombe (1973), "Graphs in statistical analysis", *American Statistician*, **27**, pp 17-21, se recomienda calcular las medias de X_1, X_2, X_3, y X_4 asi como de Y_1, Y_2, Y_3 y Y_4 y a continuación calcular las rectas de regresión de Y_i sobre X_i para i=1, 2, 3, 4. Finalmente, realizar las cuatro gráficas de Y_i en función de X_i para i=1, 2, 3, 4.

X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6