LECCIÓN 1: ESTADÍSTICA DESCRIPTIVA

1 Introducción

Nuestro objetivo es, dada una población, estudiar cierta característica de esta población.

Ejemplo 1 *Población: personas censadas en Madrid que tienen una edad entre 46 y 55 años; caracteristica que queremos estudiar: peso de la personas censadas en Madrid entre 46 y 55 aõs.*

Ejemplo 2 *Población:Alumnos matriculados en primero del grado de Ingeniería civil en la UPM en el curso 2011-2012; caracteristica que queremos estudiar: asignaturas aprobadas por los alumnos.*

Sería deseable estudiar la característica en todos los elementos de la población, pero esto a veces no es posible, pues requiere demasiado tiempo, dinero, el estudio de un elemento puede ser destructivo,....

¿Qué hacemos?. Tomamos un conjunto de la población, por ejemplo tomemos 25 alumnos de primero de ingeniería civil matriculados en el curso 2011-2012, y anotamos el número de asignaturas, **variable=número de asignaturas**, de estos 25 estudiantes, obteniendo los siguientes datos:

4	6	3	8	0
2	2	7	6	8
6	6	4	7	8
1	3	6	5	7
4	6	3	6	7

Este conjunto de datos representa una **muestra** de la población.

La **estadística descriptiva** tiene por objeto describir y analizar los datos de la muestra, sin pretender sacar conclusiones de tipo más general.

Recordamos algunas definiciones:

Si $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $\cdots$, $x_8 = 8$ son las observaciones de la muestra,

- **Frecuancia absoluta** de la observación x_k es el número de veces n_k que ha salido el dato x_k en la muestra.

1. Frecuencia absoluta de de $x_4 = 4$ es 3.
2. Frecuencia absoluta de $x_5 = 5$ es 1.
3. Frecuencia absoluta de $x_7 = 7$ es 4.

- **Frecuencia relativa** de la observaciń x_k es $f_k = \frac{n_k}{n}$.

1. Frecuencia relativa de de $x_4 = 4$ es $\frac{3}{25}$.
2. Frecuencia relativa de $x_6 = 6$ es $\frac{7}{25}$.
3. Frecuencia relativa de $x_7 = 7$ es $\frac{4}{25}$.

Observese que $0 \leq f_k \leq 1$.

- **Porcentaje o tanto por ciento** de la observación x_k es $100 \cdot f_r\%$

1. Tanto por ciento de $X_4 = 4$ es $100 \cdot \frac{3}{25}\% = 12\%$.
2. Tanto por ciento de $x_6 = 6$ es$100 \cdot \frac{7}{25}\% = 28\%$.
3. Tanto por ciento de $x_7 = 7$ es $100 \cdot \frac{4}{25}\% = 16\%$.

La descripción de la muestra se hace mediante representaciones gráficas y representaciones numéricas.

Si el tamaño de la muestra es pequeño, y cada observación ha aparecido una vez utilizamos

x_k	x_1	x_2	$\cdots$	x_n
n_k	1	1	$\cdots$	1

Este caso no es frecuente.

Si el tamaño de la muestra es grande, pero el número de valores distintos que han aparecido en la muestra es pequeño, es útil utilizar un diagrama de barras, un polígono de frecuencia o diagrama de tallos y hojas.

Ejemplo 3 *En un hospital de Madrid se pretende estudiar el grupo sanguineo del personal que trabaja en el hospital.*

Población: personal que trabaja en el hospital.

Tomamos una muestra de tamaño 100:

- Grupo A: 42 personas.

- Grupo B: 12 personas.
- Grupo AB: 5 personas.
- Grupo O: 41 personas:

Gráfico de barras de frecuencias absollutas

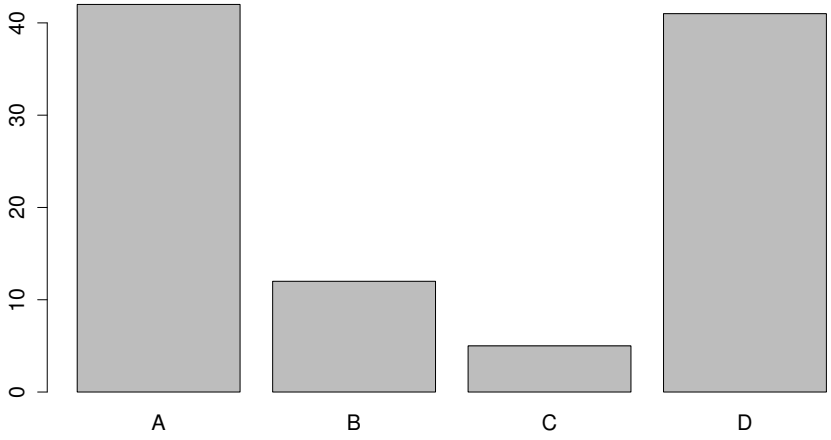
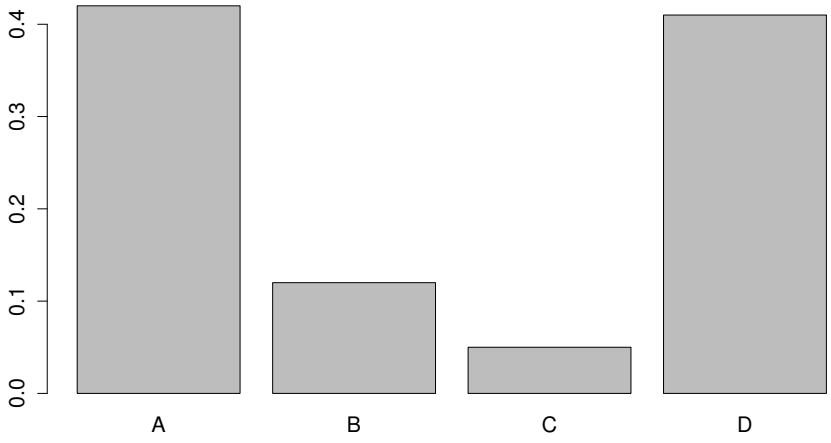


Gráfico de barras de frecuencias relativas



□

Ejemplo 4 *Consideramos como población los terremotos ocurridos en California en el siglo XX. La caracteristica que queremos estudiar es la magnitud del terremoto en la escala de Richter. Tomamos una muestra de 30 terremostos con los siguientes resultados:*

1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	4.1
5.0	2.2	1.2	7.7	1.5

Vamos a hacer un descripción de esta muestra por un **diagrama de tallo y hojas**.

Un diagrama de tallos y hojas consiste en una serie de filas horizontalesde números. El número utilizado para designar una fila es su **tallo**, el resto de los números de la fila son las **hojas** de este tallo. Los primeros dígitos de los números que aparecen en la tabla son 1, 2, 3, 4, 5, 6, 7, 8. Estos dígitos nos van a servir de nombres de los tallos, véase figura (a). A continuación, representamos los datos gráficamente anotando el número que aparece después del punto decimal, como hojas del tallo apropiado. En la fibura (b) se muestran los primeros cuatro datos puntuales (los cuatros datos de la primera columna). La figura (c) se visualiza todo el conjunto de datos. Cada tallo define una clase y se escribe solo una vez. El número de hojas en cada tallo represanta la frecuencia de este tallo o clase.

(a)

1
2
3
4
5
6
7
8

(b)

1	0
2	0
3	3
4	
5	
6	
7	
8	

(c)

1	0	2	4	0	9	2	1	1	4	3	5
2	0	2	7	2	3	4	1	1			
3	3	1	0								
4	1	0	1								
5	0	1									
6	3										
7	7										
8	3										

Ahora tenemos que interpretar este diagrama de tallo y hojas. Giramos el papel hacia un lado y observamos la curva que se ha trazado en la figura (c). Para ello nos vamos a hacer las siguientes preguntas:

1. ¿Tienden a agruparse los datos cerca de un tallo o tallos en particular, o se distrubuyen de forma uniforme por el diagrama?

   Las hojas se agrupan en los tallos 1 y 2, no están distrubuidos de manera uniforme.

2. ¿Tienden a estrecharse los datos hacia un extremo u otro del diagrama?

   Los datos se aproximan al extremo inferior de la escala, y esto lo interpretamos cómo que casi todos los terremotos ocurridos en California en el siglo XX han sido suaves.

3. Si se traza una curva a lo largo de la parte superior del diagrama, ¿dorma más o menos una campana?, ¿es plana?, ¿es simétrica?.

   La curva no es simétrica, hay más bien una cola larga hacia el extremo superior o derecho de la visualización. Decimos que los datos de este tipo están sesgado hacia la derecha.

□

El diagrama de tallo y hojas nos da una idea aproximada de la forma de la distribución de las observaciones, así como de su localización. La técnica funciona bien para las muestras que no tienen una dispersión muy grande. Sin embrago, si el número de observaciones distintas que han aparecido es grande, es difícil escoger tallos adecuados. En este caso es más útil utilizar los **histogramas**.

Ejemplo 5 *En el aereopuerto de Madrid-Barajas se ha recogido una muestra del peso en kilogramos del equipaje de 50 viajeros que realizan el trayecto Madrid-Barcelona:*

20.10	20.25	20.31	20.42	20.58	20.64	20.70	20.90	21.05	21.41
21.57	21.77	21.87	22.01	22.17	22.37	22.49	22.55	22.87	22.95
23.08	23.20	23.37	23.41	23.45	23.47	23.65	23.70	23.98	24.30
24.43	24.43	24.43	24.70	24.81	24.83	24.99	25.13	25.20	25.55
26.27	26.57	27.16	27.60	27.81	28.09	28.19	30.25	32.34	33.46

Vamos a dibujar un histograma de frecuencias absolutas.

Un histograma de frecuencias es un gráfico de barras verticales u horizontales. Ya que el conjunto de observaciones con gran cantidad de valores numéricos distintos no tiene clases naturales obvias, debemos diseñar

un método para definirlas. Queremos definir clases de igual tamaño, de tal forma que cada observación corresponda clara y exactamente a una de ellas.

Reglas para agrupar las observaciones en clases

1. Decidir el número de clases deseado. El número elegido depende de la cantidad de observaciones disponibles.. Está basada en la regla Sturges. Esta regla afirma que

$$\text{número de clases} \simeq 1 + 3.322 \log_{10} n,$$

donde n es el tamaño de la muestra.

En nuesttro caso: número de clases $\simeq 1 + 3.322 \log_{10} 50 = 6.643 \simeq 7$.

2. Localizar la observación mayor y menor y hallar la diferenia entre estas dos observaciones.

$$R = \max x_k - \min x_k = 33.46 - 20.10 = 13.36.$$

A esta diferencia R se le denomina **rango de los valores de la muestra**.

3. Hallar la amplitud, (ancho), mínima de la clase requerida para cubrir este rango, dividiendo el rango por el n'umero de clases deseado.

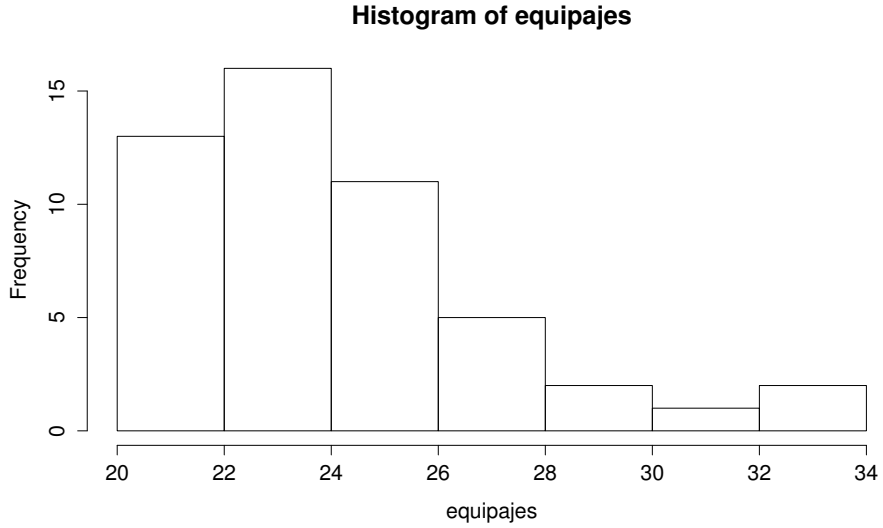
$$\frac{13.36}{7} \simeq \frac{14}{7} = 2.$$

Podemos entonces definir 7 intervalos de amplitud constante desde el valor mínimo $\simeq 20.00$ hasta el valor máximo $\simeq 34.00$.

4. Construimos la tabla:

$L_{k-1} - L_k$	x_k	n_k
20.00-21.99	21.0	13
22.00-23.99	23.0	16
24.00-25.99	25.0	11
26.00-27.99	27.0	5
28.00-29.99	29.0	2
30.00-31.99	31.0	1
32.00-33.99	33.0	2

La representación gráfica de la tabla en un diagrama de barras se denomina histograma:



□

Distinguimos dos tipos de variables o características. Una **variables es continua** cuando puede tomar cualquier valor en algún intervalo de los números reales. Una **variable es discreta** si el conjunto de sus observaciones es finito o numerable.

Ejemplo 6 *Población: estudiantes matriculados en primero del grado de ingeniería civil de la UPM en el curso 2012-13.*

Variable: peso del estudiante.

Esta variable es continua. El peso de un estudiante puede tomar cualquier valor del intervalo $(35, 240)$.

Ejemplo 7 *Población: estudiantes matriculados en primero del grado de ingeniería civil de la UPM en el curso 2012-13.*

Variable: número de cigarrillos fumados por el estudiante..

Esta variable es es discreta. El número de cigarrillos que puede fumar un estudiante es un del conjunto $\{0, 1, 2, \cdots, 150\}$.

2 Medidas de centralización y dispersión

2.1 Medidas de centralización

Supongamos que tomamos una muestra de tamaño n

$$\{x_1, x_2, \cdots, x_n\},$$

de una cierta población. Las medidas de centralización tienen como objetivo dar una idea del valor central alrededor del cual se distrubuyen las observaciones o valores de la muestra.

Definimos la **media** de la muestra $\{x_1, x_2, \cdots, x_n\}$ por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_1}{n}.$$

Ejemplo 8 *Un pastor lleva a sus ovejas a un campo con hierba venenosa. Queremos estudiar el número de horas que tardan en morir la ovejas.*

La población son la ovejas del pastor y la carasteristica o variable que queremos estudiar es el número de horas que tardan en morir las ovejas. Es una variable discreta.

Tomamos unamuestra de tamaño 13

$$44 \quad 27 \quad 24 \quad 24 \quad 36 \quad 36 \quad 44 \quad 44 \quad 120 \quad 29 \quad 36 \quad 36 \quad 36$$

La media de esta muestra es

$$\bar{x} = \frac{44 + 27 + 24 + 24 + 36 + 36 + 44 + 44 + 120 + 29 + 36 + 36 + 36}{13} = 41.23$$

Interpretamos este resultado como que el tiempo medio en que tardan en morir las ovejas es 41.23 horas.

Si ordenamos los valores de la muestra en orden creciente, obtenemos

$$24 \quad 24 \quad 27 \quad 29 \quad 36 \quad 36 \quad 36 \quad 36 \quad 36 \quad 44 \quad 44 \quad 44 \quad 120$$

y observamos que el valor 36 tiene 6 valores a si izquierda $\{24, ;\ 24,\ 27,\ 29,\ 36,\ 36\}$ y 6 valores a su derecha $\{36,\ 36,\ 44,\ 44,\ 44,\ 120\}$. Decimos que el 36 es la **mediana** de la muestra $\{24, ;\ 24,\ 27,\ 29,\ 36,\ 36,\ 36,\ 36,\ 44,\ 44,\ 44,\ 120\}$. □

Ejemplo 9 *El acero es una aleación de hierro y carbono, donde el carbono no supera el* 2.1% *en peso de la composición de la aleación, alcanzando normalmente porcentajes entre el* 0.2% *y el* 0.3%. *Porcentajes mayores que el* 2% *dan lugar a las fundiciones, aleaciones que al ser quebradizas y no poderse forjar, a diferencia de los aceros, se moldean.*

Una fábrica de acero quiere comprobar la calidad del producto que produce.

La población son las piezas de acero que produce la fabrica. La variable o caracteristica que queremos estudiar es porcentaje de carbono en las piezas de acero. Es una variable continua.

Tomamos dos muestras, que las vamos a dar ordenadas en forma creciene:

$$\mathcal{M}_1 \quad 0.2 \quad 0.22 \quad 0.23 \quad 0.23 \quad 0.24 \quad 0.25 \quad 0.25 \quad 0.26 \quad 0.28$$

$$\mathcal{M}_2 \quad 0.12 \quad 0.13 \quad 0.14 \quad 0.14 \quad 0.22 \quad 0.24 \quad 0.26 \quad 0.34 \quad 0.34 \quad 0.35 \quad 0.36$$

Es fácil comprobar que ambas muestras tienen la misma media y mediana. Sin embargo, las dos muestras son muy diferentes. En la primera, en todas las observaciones, el porcentaje de carbono está entre el 0.2% y el 0.3%, mientras en la segunda muestra, solamente tres observaciones de las 11 tienen porcentaje de carbono entre el 0.2% y el 0.3%. Las observaciones están más dispersas en la segunda muestra que en la primera. Esto nos está diciendo que **las medidas de centralización no son suficientes para describir una muestra**.

□

2.2 Medidas de dispersión

Las **medidas de dispersión** nos informan sobre la dispersión de las observaciones con respecto a los valores centrales.

La **varianza** de una muestra $\{x_1, x_2, \cdots, x_n\}$, se define por

$$v_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}.$$

Proposición 10

$$\boxed{v_x = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

Proof.

$$v_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{n} x_i + \frac{n\bar{x}^2}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

□

Vamos a calcular las variaciones de las muestras dadas en ejemplo 9. Para la muestra $\mathcal{M}_1$:

$$v_x = \frac{1}{9}\sum_{i=1}^{9} x_i^2 - \bar{x}^2 = \frac{0.2^2 + 0.22^2 + 2\cdot 0.23^2 + 0.24^2 + 2\cdot 0.25^2 + 0.26^2 + 0.28^2}{9} - 0.24^2$$

$$= \frac{0.5228}{9} - 0.0576 = 0.00048.$$

Para la muestra $\mathcal{M}_2$

$$v_x = \frac{1}{11}\sum_{i=1}^{11} x_i^2 - \bar{x}^2$$

$$= \frac{0.12^2 + 0.13^2 + 2\cdot 0.14^2 + 0.22^2 + 0.24^2 + 0.26^2 + 2\cdot 0,34^2 + 0.35^2 + 0.36^2}{11} - 0.24^2$$

$$= \frac{0.7274}{11} - 0.0576 = 0.00852.$$

Observamos que la varianza de la muestra $\mathcal{M}_2$ es como 20 veces mayor que la varianza de la muestra $\mathcal{M}_1$, como era de esperar.

La varianza nos da la media de los cuadrados de las desviaciones de las observaciones a la media muestral. La **desviación típica** de una muestra es la raiz cuadrada de la varianza de la muestra. La desviación típica vuelve a medir la dispersón pero en unas unidades más naturales que la varianza.

Volvemos al ejemplo 8. Se puede ver que

1. varianza de la muestra es 562.38.

2. desviación típica es 23.71.

La dispersión es grande, y es más natural fijarnos en la desviación típica para medir esta.

Que la desviación típica sea tan grande lo produce el dato atípico 120.

Suprimamos la observación 120 de la muestra. Tenemos la nueva muestra

44 27 24 24 36 36 44 44 29 36 36 36

Calculamos las medidas de centralización y dispersión para esta nueva nuestra:

1. media=34.66
2. En este caso tenemos un número par de observaciones. Tenemos dos valores centrales

   24 24 27 29 36 36 36 36 36 44 44 44

   que son el 36 y el 36. En este caso, la media de estos dos números= $\frac{36+36}{2} = 36$ es la mediana de la muestra.
3. varianza=49.5
4. desviación típica=7.03

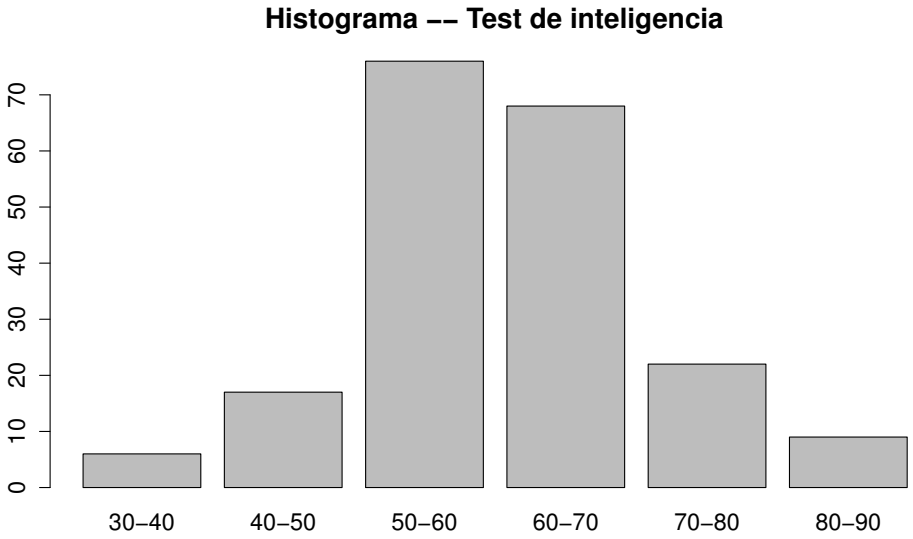
Ahora vemos que ya no hay tanta disperción.

Observamos también que para esta muestra, la media y la mediana están próximas, mientras que para la muestra original, no era así. Esto es otra vez debido a la observación atípica 120. **De aquí sacamos como conclusión que la mediana es más resistente a la influencia de datos atípicos que la media**.

Observación 11 *La media utiliza todos los datos de la muestra y cuando la muestra es homogénea, es la medida de centralización más represntativa. Sin embargo, tiene el inconveniente de que es muy sensible a los datos atípicos, y un error en la recogida de la muestra puede desplazar la media hacia un valor muy distinto del "verdadero". Por otra parte, la mediana se basa más en el orden de los datos que en sus propios valores, por lo que no está afectada por los datos atípicos. Ambas medidas de centralización tienen sus ventajas e inconvenientes y su uso debe ser siempre complementario.*

Ejemplo 12 *Uno de los exámenes para una cierta oposición es un test de inteligencia. La variable o característica que queremos estudiar de la población, (los opositores), es la puntuación en el test de inteligencia.*

De una muestra de tamaño 198 nos dan el siguiente histograma:



1. Intervalo $(30, 40]$ numero 6.
2. Intervalo $(40, 50]$ numero 17.
3. Intervalo $(50, 60]$ numero 76.
4. Intervalo $(60, 70]$ numero 68.
5. Intervalo $(70, 80]$ numero 22.
6. Intervalo $(80, 90]$ numero 9.

Tenemos 6 clases. Como puntuación representativa de cada clase vamos a tomar su valor medio. Por ejemplo, de la tercera clase :$(50, 60]$ = puntuación en el examen tipo test mayor que 50 puntos y menor o igual que sesenta, tomamos $x_2 = \frac{50+60}{2} = 55$.

1. media= $\bar{x} = \frac{6 \cdot 35+17 \cdot 45+76 \cdot 55+68 \cdot 65+22 \cdot 75+9 \cdot 85}{198} = \frac{11990}{198} = 60.55$.
2. Si sumamos los valores que están en las tres promeras clases nos da 99, que corresponde al 50% de la muestra de 198 personas, es decir, que el 50% de la muestra tiene una puntuación menor o igual que 60 puntos,

mientras que el otro 50% tiene una puntuación mayor que 60 puntos. mediana=60.

3. $v_x = \frac{6\cdot 35^2+17\cdot 45^2+76\cdot 55^2+68\cdot 65^2+22\cdot 75^2+9\cdot 85^2}{198} - 60.55^2 = \frac{777750}{198} - 3666.3 = 110.19$

4. desviación típica= $\sqrt{110.19} = 10.49$

□

3 Regresión lineal

Ejemplo 13 *En una empresa, parte de los trabajadores están dedicados a la venta. La empresa quiere conocer la relación entre el tamaño de su equipo de vendedores y sus ingresos anuales, en cientos de miles de dólares. La empresa fué creada en 1975.*

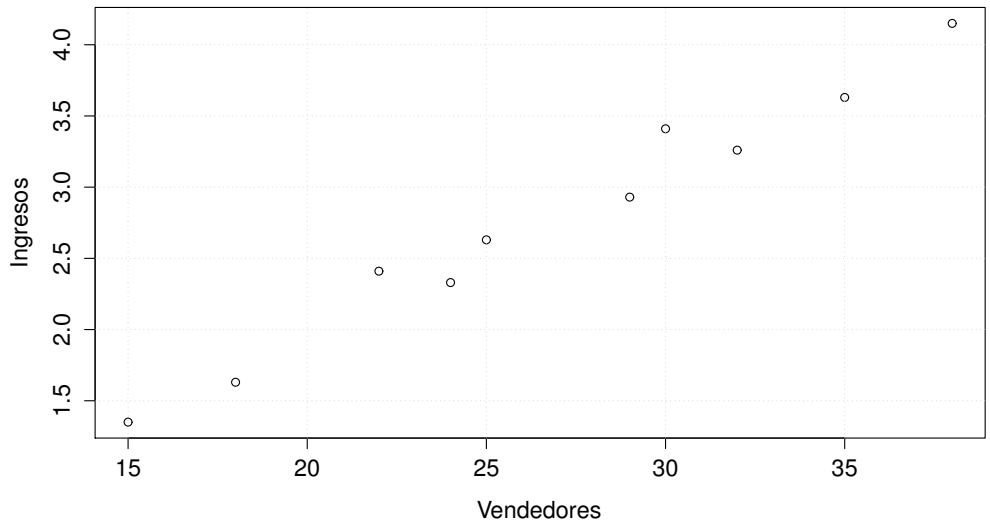
Población= conjunto de años entre 1975 y 2009=$\{1975, 1976, \cdots, 2009\}$.
Consideramos las variables o características
X= número de vendedores.
Y= ingresos anuales en miles de dólares.
Tomamos una muestra de tamaño 10:

Año	Número de vendedores	Ingresos
1991	15	1.35
1993	18	1.63
1995	24	2.33
1997	22	2.41
1999	25	2.63
2001	29	2.93
2003	30	3.41
2005	32	3.26
2007	35	3.63
2009	38	4.15

En este caso, la muestra es un subconjunto de $\mathbb{R}^2$:

$$\{(x_i, y_i)\ i = 1, 2, \cdots, 10\} = \{(15, 1.35), (18, 1.63), (15, 1.35), \cdots, (38, 4.15)\}.$$

Representamos la muestra en $\mathbb{R}^2$



Vamos a definir una recta

$$y = y(x) = a + bx,$$

(x variable independiente, y variable dependiente), tal que si x es el número de vendedores en algún año elegido, $y(x)$ sea un valor próximo a los ingresos en ese año. Por ejemplo, si tomamos el año 2003, entonces $x = 30$ y queremos elegir a y b de tal manera que $y(30) = a + 30b$ sea "próximo" a los ingresos en el año 2003, esto es a 3.41.

Elegimos a y b de tal manera que el punto (a, b) minimice a la función

$$f(u, v) = \frac{1}{10} \sum_{i=1}^{10} (y_i - u - vx_i)^2$$

$$= \frac{1}{10} \left[(1.35 - u - 15v)^2 + (1.63 - u - 18v)^2 + \cdots + (4.15 - u - 38v)^2 \right].$$

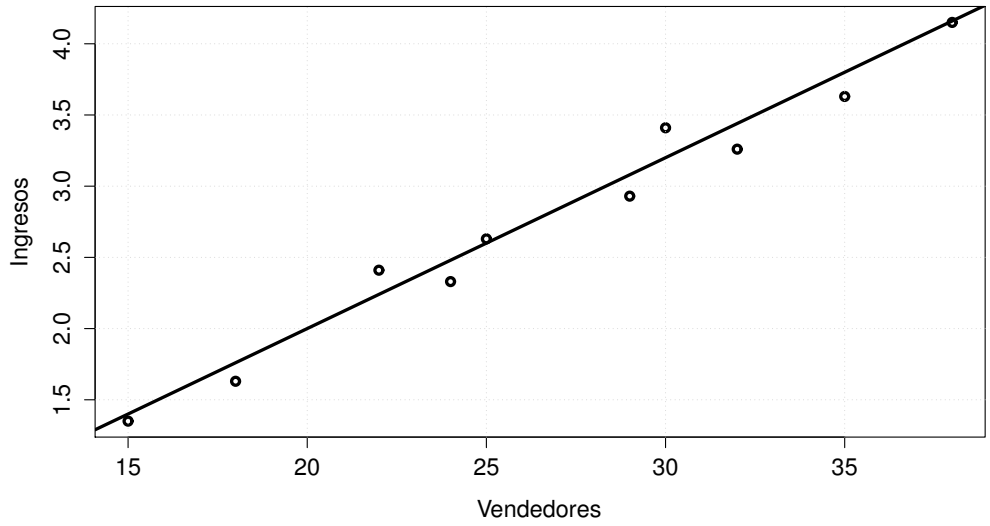
Para minimizar esta función, calculamos los puntos críticos de f, es decir la solución del sistema

$$\nabla f(u, v) = \left(\frac{\partial f}{\partial u}(u, v), \frac{\partial f}{\partial v}(u, v) \right) = (0, 0).$$

Esta ecuación tiene como solución $u = -0.4$ y $v = 0.12$. Puede comprobarse que el punto $(-0.4, 0.12)$ minimiza a la función $f(u, v)$. El alumno no debe

de hacer estos cálculos, pues este problema lo vamos a resolver en su forma general. La recta es

$$y = y(x) = 0.12x - 0.4.$$



Vamos a comparar con los valores de la muestra:

Año	Número de vendedores	Ingresos	$y(x_i) = 0.12x_1 - 0.4$	error= $y(x_1) - y_i$
1991	15	1.35	$y(15) = 1.4$	0.05
1993	18	1.65	$y(18) = 1.7$	0.07
1995	24	2.33	$y(24) = 2.48$	0.15
1997	22	2.41	$y(22) = 2.24$	-0.17
1999	25	2.63	$y(25) = 2.6$	-0.03
2001	29	2.93	$y(29) = 2.93$	0.09
2003	30	3.41	$y(30) = 3.2$	-0.21
2005	32	3.26	$y(32) = 3.44$	0.18
2007	35	3.63	$y(35) = 3.8$	0.17
2009	38	4.15	$y(38) = 4.15$	0.01

Podemos decir que la recta $y = 0.12x - 0.4$ representa bastante bien a la muestra.

¿Podríamos utilizar esta recta para predecir lo que podría ganar la empresa si utiliza 50 vendedores?. La estadiśtica descriptiva no responde a esta

pregunta. Recordamos que la estadística descriptiva solo analiza los valores de la muestra. Sólo si la muestra es "significatíva" podemos utilizar la recta de regresión para predecir. Vamos a suponer que esta muestra es representatíva. Entonces podemos pensar que si la empresa utilizase 50 vendedores, los ingresos que se esperarían serían del orden de

$$y(50) = 0.12 \cdot 50 - 0.4 = 5.6$$

□

Supongamos que de una población queremos estudiar dos variables o caracteristicas X e Y,en el sentido del ejemplo anterior. Tomamos una muestra de tamaño n

$$\{(x_i, y_i)\ i = 1, 2, \cdots, n\},$$

definimos la **covarianza** de la muestra, (de las variables X, Y), por

$$\text{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

La covarianza entre X e Y es una medida de la forma en que las variables X e Y varian conjuntamente.

Proposición 14 *$cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$.*

Proof.

$$\text{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\bar{y}}{n}\sum_{i=1}^{n} x_i - \frac{\bar{x}}{n}\sum_{i=1}^{n} y_i + \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

□

Definición 15 *La* **recta de regresión** *de la variable Y sobre la X asociada a la muestra $\{(x_i, y_i)\ i = 1, 2, \cdots, n\}$, es la recta*

$$y = a + bx,$$

donde a y b son tales que el punto (a,b) minimiza la función

$$f(u,v)=\frac{1}{n}\sum_{i=1}^{n}(y_i-u-vx_i)^2.$$

Vamos a calcular la recta de regresión de la variable Y sobre la X asociada a la muestra $\{(x_i,y_i)\ i=1,2,\cdots,n\}$.

$$f(u,v)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i^2-2y_iu-2y_ivx_i+u^2+2uvx_i+v^2x_i^2\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_i^2-2u\bar{y}-\frac{2v}{n}\sum_{i=1}^{n}x_iy_i+u^2+2uv\bar{x}+\frac{v^2}{n}\sum_{i=1}^{n}x_i^2.$$

Calculamos los puntos críticos de la función $f(u,v)$, es decir la solución del sistema

$$\nabla f(u,v)=\left(\frac{\partial f}{\partial u}(u,v),\frac{\partial f}{\partial v}(u,v)\right)=(0,0).$$

$$\begin{cases}\frac{\partial f}{\partial y}(u,v)=-2\bar{y}+2u+2v\bar{x}=0\\ \frac{\partial f}{\partial v}(u,v)=-\frac{2}{n}\sum_{i=1}^{n}x_iy_i+2u\bar{x}+\frac{2v}{n}\sum_{i=1}^{n}x_i^2=0\end{cases}$$

$$\iff\begin{cases}u+v\bar{x}=\bar{y}\\ u\bar{x}+v\frac{1}{n}\sum_{i=1}^{n}x_i^2=\frac{1}{n}\sum_{i=1}^{n}x_iy_i\end{cases}.$$

Haciendo algunos cálculos, se obtiene que la solución de este sistema es

$$u=\bar{y}-\frac{\mathrm{cov}_{x,y}}{v_x}\bar{x},\qquad v=\frac{\mathrm{cov}_{x,y}}{v_x}.$$

Se puede comprobar que el punto $(a,b)=\left(\bar{y}-\frac{\mathrm{cov}_{x,y}}{v_x}\bar{x},\frac{\mathrm{cov}_{x,y}}{v_x}\right)$, es el que minimiza a la función $f(u,v)$.

Conclusión: recta de regresión de la variable Y sobre la X asociada a la muestra $\{(x_i,y_i)\ i=1,2,\cdots,n\}$ es:

$$\boxed{y=\bar{y}-\frac{\mathrm{cov}_{x,y}}{v_x}\bar{x}+\frac{\mathrm{cov}_{x,y}}{v_x}x}$$

Ejemplo 16 *Una fábrica está dedicada a producir duraluminio, una aleación de alumino con cobre-Cu (3-5%), magnesio-Mg (0.5-2%), manganeso-Mn (0.25-1%) y zinc-Zn (3.5-5%), muy utilizada en casas (puertas, ventanas), transporte, etc. Por alguna razón parece ser que en algunas piezas de duraluminio hay menos proporción de cobre que lo normal. Queremos estudiar las variables o caracteristicas de las piezas de duraluminio:*

X= Proporción de cobre de las piezas,
Y= Proporción de magnesio, manganeso y zing.
Tomamos una muestra de 14 piezas:

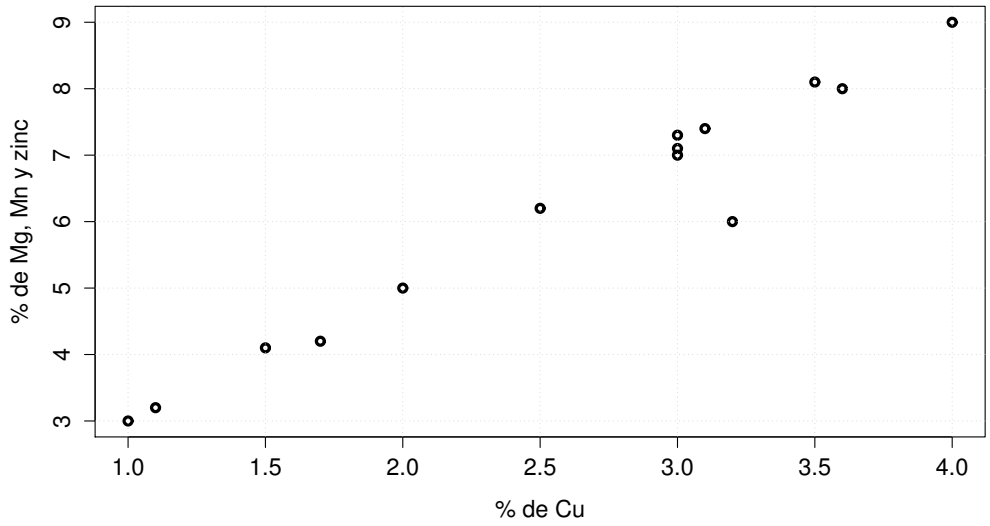
% de Cu	% de Mg, Mn y zinq	% de Cu	% de Mg, Mn y Zn
1.0	3.0	3.0	7.0
1.1	3.2	3.0	7.1
1.5	4.1	3.1	7.4
1.7	4.2	3.2	6.0
2.0	5.0	3.5	8.1
2.5	6.2	3.6	8.0
3.0	7.3	4.0	9.0

La muestra la podemos tambien escribir en la forma

$$M = \{(x_1, y_i),\ i = 1, 2, \cdots, 14\}$$

$$= \{(1.0, 3.0), (1.1, 3.2), (1.5, 4.1), \cdots, (3.6, 8.0), (4.0, 9.0)\}.$$

Empezamos dibujando en $\mathbb{R}^2$ la muestra y luego calculamos la recta de regresión de la variable Y sobre la X asociada a la muestra.



Media de la muestra

$$\{x_1, x_2, x_3, \cdots, x_{13}, x_{14}\} = \{1.0, 1.1, 1.5, \cdots, 3.60, 4.0\}$$

producida por la variable X

$$\bar{x} = \frac{1}{14}\sum_{i=1}^{14} x_i = \frac{1.0 + 1.1 + 1.5 + \cdots + 3.6 + 4.0}{14} = \frac{36.2}{14} = 2.59.$$

Media de la muestra

$$\{y_1, y_2, y_3, \cdots, y_{13}, y_{14}\} = \{3.0, 3.2, 4.1, \cdots, 8.0, 9.0\}$$

producida por la variable Y

$$\bar{y} = \frac{1}{14}\sum_{i=1}^{14} y_i = \frac{3.0 + 3.2 + 4.1 + \cdots + 8.0 + 9.0}{14} = \frac{85.6}{14} = 6.11.$$

Varianza de la muestra

$$\{x_1, x_2, x_3, \cdots, x_{13}, x_{14}\} = \{1.0, 1.1, 1.5, \cdots, 3.60, 4.0\}$$

producida por la variable X

$$v_x = \frac{1}{14}\sum_{i=1}^{14} x_i^2 - \bar{x}^2 = \frac{1.0^2 + 1.1^2 + 1.5^2 + \cdots + 3.6^2 + 4.0^2}{14} - 2.59^2 = \frac{105.66}{14} - 6.7 = 0.85$$

$$\mathrm{cov}_{x,y} = \frac{1}{14}\sum_{i=1}^{14} x_i y_i - \bar{x}\bar{y}$$

$$= \frac{1.1 \cdot 3.0 + 1.1 \cdot 3.2 + 1.5 \cdot 4.1 + \cdots + 3.6 \cdot 8.0 + 4.0 \cdot 9.0}{14} - 2.59 \cdot 6.11$$

$$= \frac{244.8}{14} - 15.8 = 1.68.$$

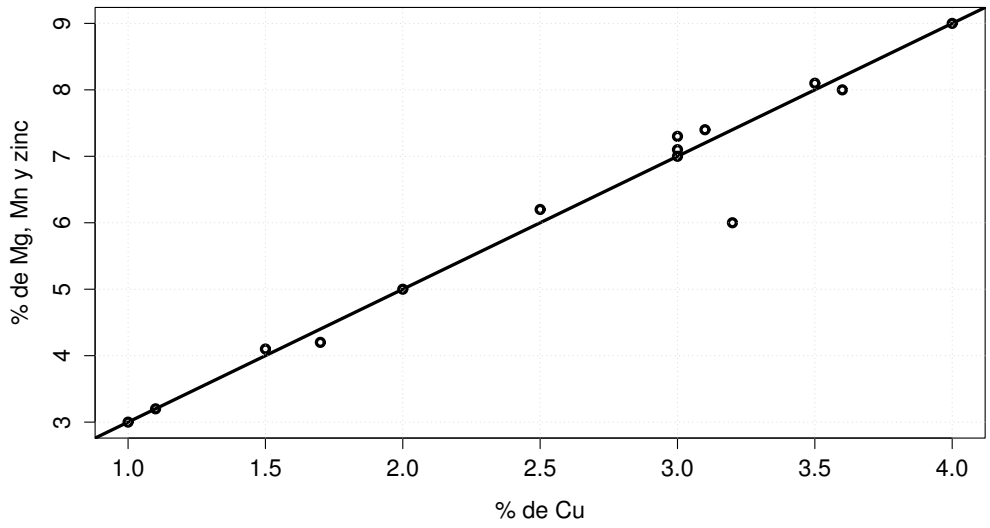
La recta de regresión es $y = a + bx$ donde

$$a = \bar{y} - \frac{\mathrm{cov}_{x,y}}{v_x}\bar{x} = 6.11 - \frac{1.68}{0.85}2.59 = 0.99 \simeq 1,$$

$$b = \frac{\mathrm{cov}_{x,y}}{v_x} = \frac{1.68}{0.85} = 1.97 \simeq 2.$$

Recta de regresión

$$y = y(x) = 1 + 2x.$$



Queremos que la resta de regresión represente a la muestra M, en el sentido de que si tomamos un punto cualquiera de esta muestra, por ejemplo el $(3.0, 7.1)$, se verifique que $y(3.0)$ sea "una buena aproximación" de 7.1. Como $y(3.0) = 1 + 2 \cdot 3.0 = 7$ podemos pensar que si es una "buena aproximación". Pero sólo lo hemos comprobado para el punto $(3.0, 7.1)$?. Quisieramos comprobarlo de una manera conjunta para las 14 observaciones de la muestra. Por el método que hemos seguido para obtener la recta de regresión, sería razonable pensar que si el número

$$E = \frac{1}{14} \sum_{i=1}^{14} (y_i - y(x_i))^2$$

$$= \frac{(3.0 - y(1.0))^2 + (3.2 - y(1.1))^2 + \cdots + (8.0 - y(3.6))^2 + (4.0 - y(4.0))^2}{14}$$

es pequeño, entonces la recta de regresión representa bien a la muestra M en el sentido que hemos indicado.

Vamos a calcular el valor de E.

$y(x_i) = 1 + 2 \cdot x_i =$	$(y_i - y(x_i))^2 =$
$y(1.0) = 1 + 2 \cdot 1.0 = 3.0$	$(3.0 - 3.0)^2 = 0$
$y(1.1) = 1 + 2 \cdot 1.1 = 3.2$	$(3.2 - 3.2)^2 = 0$
$y(1.5) = 1 + 2 \cdot 1.5 = 4.0$	$(4.1 - 4.0)^2 = 0.01$
$y(1.7) = 1 + 2 \cdot 1.7 = 4.4$	$(4.2 - 4.4)^2 = 0.04$
$y(2.0) = 1 + 2 \cdot 2.0 = 5.0$	$(5.0 - 5.0)^2 = 0$
$y(2.5) = 1 + 2 \cdot 2.5 = 6.0$	$(6.2 - 6.0)^2 = 0.04$
$y(3.0) = 1 + 2 \cdot 3.0 = 7.0$	$(7.3 - 7.0)^2 = 0.09$
$y(3.0) = 1 + 2 \cdot 3.0 = 7.0$	$(7.0 - 7.0)^2 = 0$
$y(3.0) = 1 + 2 \cdot 3.0 = 7.0$	$(7.1 - 7.0)^2 = 0.01$
$y(3.1) = 1 + 2 \cdot 3.1 = 7.2$	$(7.4 - 7.2)^2 = 0.04$
$y(3.2) = 1 + 2 \cdot 3.2 = 7.4$	$(6.0 - 7.4)^2 = 1.96$
$y(3.5) = 1 + 2 \cdot 3.5 = 8.0$	$(8.1 - 8.0)^2 = 0.01$
$y(3.6) = 1 + 2 \cdot 3.6 = 8.2$	$(8.0 - 8.2)^2 = 0.04$
$y(4.0) = 1 + 2 \cdot 4.0 = 9.0$	$(9.0 - 9.0)^2 = 0$

$$E = \frac{0 + 0 + 0.01 + 0.04 + 0 + 0.04 + 0.09 + 0 + 0.01 + 0.04 + 1.96 + 0.01 + 0.04 + 0}{14}$$
$$= \frac{2.24}{14} = 0.16.$$

El error es "pequeño" y podemos decir que la recta de regresión de la variable Y sobre X asociada a la muestra M será "una buena aproximación" de la muestra.

□

Supongamos que de una población queremos estudiar dos variables o caracteristicas X e Y. Tomamos una muestra de tamaño n

$$M = \{(x_i, y_i),\ i = 1, 2, \cdots, n\}.$$

Sea

$$y = y(x) = a + bx = \bar{y} - \frac{\text{cov}_{x,y}}{v_x}\bar{x} + = \frac{\text{cov}_{x,y}}{v_x}x$$

la recta de regresión de la variable Y sobre X asociada a la muestra M, se puede demostrar la siguiente expresión para el error E

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - y(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2) = v_y\left(1 - \frac{\text{cov}_{x,y}^2}{v_x v_y}\right).$$

El error E es conocido como la **varianza residual** de la variable Y sobre la X asociada a la muestra M. Se puede demostrar que $0 \leq \frac{\text{cov}_{x,y}^2}{v_x v_y} \leq 1$ y por lo tanto

$$-1 \leq \frac{\text{cov}_{x,y}}{\sqrt{v_x v_y}} \leq 1.$$

Al número

$$r = \frac{\text{cov}_{x,y}}{\sqrt{v_x v_y}}$$

se le denomina **coeficiente de correlación** entre las variables X e Y asociado a la muestra $M = \{(x_i, y_i),\ i = 1, 2, \cdots, n\}$.

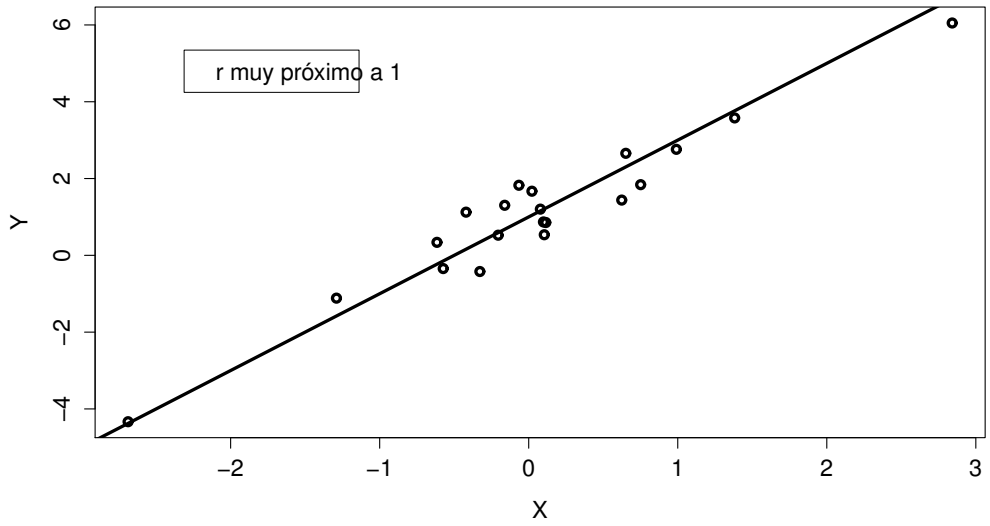
En términos del coeficiente de correlación, el error o varianza residual viene dado por

$$E = v_y(1 - r^2).$$

Si el coeficiente de correlación r es próximo a 1 o -1, el error será pequeño y la recta de regresión de la variable Y sobre X asociada a la muestra M será "una buena aproximación" de la muestra.

Si r es próximo a 1, la pendiente de la recta de regresión sería positiva y estaríamos en una situación como indica la figura 1. En este caso, la recta de regresión de Y sobre X "representa bien a la muestra".

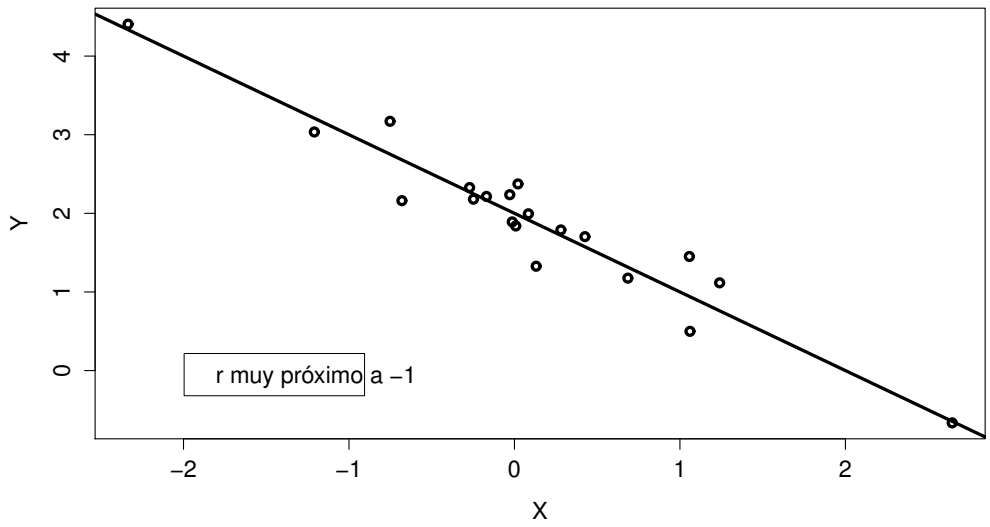
Figura 1



Si r es próximo a -1, la pendiente de la recta de regresión será negativa y estaríamos en una situación como indica la figura 2. En este caso, la recta

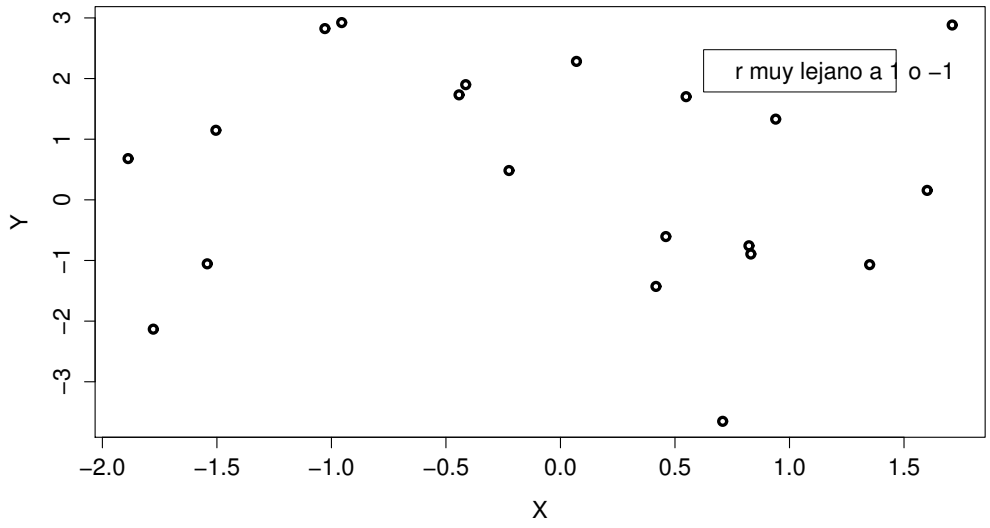
de regresión de Y sobre X "representa bien a la muestra".

Figura 2



Si r es próximo a 0, el error será grande y la recta de regresión de la variable Y sobre X asociada a la muestra M no sería "una buena aproximación" de la muestra, y estaríamos en una situación como indica la figura 3

Figura 3



Ejemplo 17 *Los manatíes son unos animales grandes y dóciles que viven a lo largo de la costa de Florida. Cada año las lanchas motoras hieren o matan muchos de ellos. A continuación se presenta una tabla que contiene, para cada año, el número de licencias para motoras (expresado en miles de licencias) expedidas en Florida y el número de manatíes muertos en los años 1996 y 2009.*

Año	Licencias	Manatíes
1996	$x_1 = 447$	$y_1 = 13$
1997	$x_2 = 460$	$y_2 = 21$
1998	$x_3 = 481$	$y_3 = 24$
1999	$x_4 = 498$	$y_4 = 16$
2000	$x_5 = 513$	$y_5 = 24$
2001	$x_6 = 512$	$y_6 = 20$
2002	$x_7 = 526$	$y_7 = 15$
2003	$x_8 = 559$	$y_8 = 34$
2004	$x_8 = 585$	$y_9 = 33$
2005	$x_{10} = 614$	$y_{10} = 33$
2006	$x_{11} = 645$	$y_{11} = 39$
2007	$x_{12} = 675$	$y_{12} = 43$
2008	$x_{13} = 711$	$y_{13} = 50$
2009	$x_{14} = 719$	$y_{14} = 47$

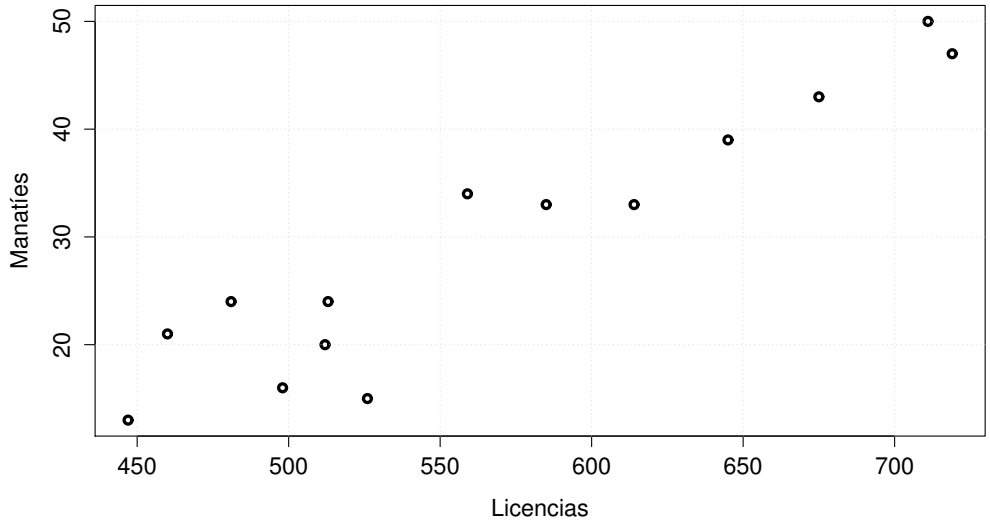
Queremos analizar la relación entre el número de licencias expedidas anualmente en Florida y el número de manatíes muertos. Consideramos las variables o características

$X =$"nḿero de licencias expedidas (en miles)",

$Y =$"número de manatíes muertos".

Vamos a representar gráficamente los puntos

$$\{(x_i, y_i),\ i = 1, 2, \cdots, 14\}.$$



Las nube de puntos nos indica que existe una buena aproximación lineal entre las variables X e Y. Vamos a calcular la recta de regresión.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{14}}{14} = \frac{447 + 460 + \cdots + 719}{14} = 567,5.$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_{14}}{14} = \frac{13 + 21 + \cdots + 47}{14} = 29,43.$$

$$\sum_{i=1}^{14} x_i^2 = 447^2 + 460^2 + \cdots + 719^2 = 4618597.$$

$$\sum_{i=1}^{14} y_i^2 = 13^2 + 21^2 + \cdots + 47^2 = 14056.$$

$$\sum_{i=1}^{14} x_i y_i = 447 \cdot 13 + 460 \cdot 21 + \cdots + 719 \cdot 47 = 247521.$$

$$v_x = \frac{\sum_{i=1}^{14} x_i^2}{14} - \bar{x}^2 = \frac{4618597}{14} - 567,5^2 = 7834,5.$$

$$v_y = \frac{\sum_{i=1}^{14} y_i^2}{14} - \bar{y}^2 = \frac{14056}{14} - 29,43^2. = 137,9.$$

$$\text{cov}_{x,y} = \frac{\sum_{i=1}^{14} x_i y_i}{14} - \bar{x}\bar{y} = \frac{247521}{14} - 567,5 \cdot 29,43 = 978,55$$

La recta de regresión de Y sobre la variable X asociada a la muestra $\{(x_i, y_i),\ i = 1, 2, \cdots, 14\}$ es

$$Y = a + bx,$$

donde

$$a = \bar{y} - \frac{\text{cov}_{x,y}}{v_x}\bar{x} = -41.45,$$

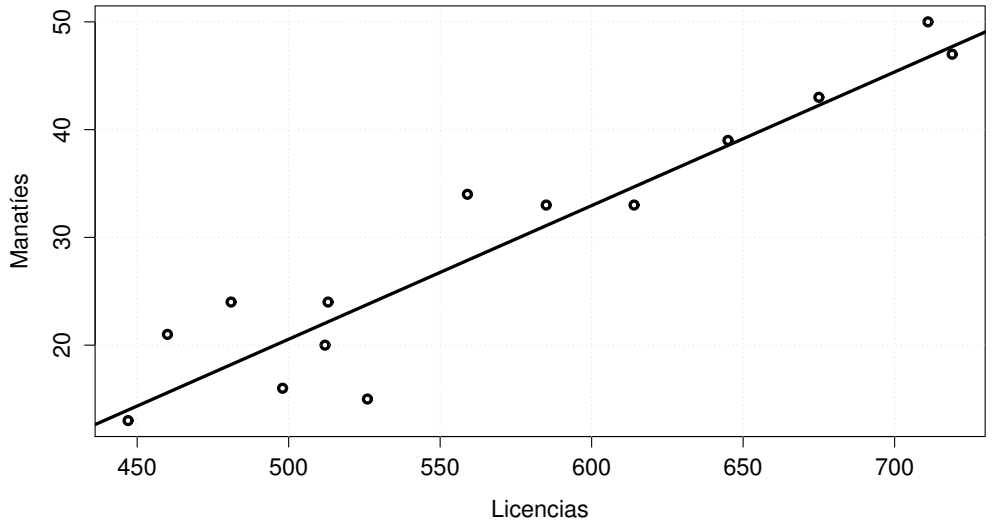
$$b = \frac{\text{cov}_{x,y}}{v_x} = 0.124,$$

$$\boxed{y = 0.124x - 41.45}\,.$$

El coeficiente de correlación lineal

$$r = \frac{\text{cov}_{x,y}}{\sqrt{v_x v_y}} = 0,941.$$

El coeficiente de correlación es muy próximo a 1, por lo tanto la recta de regresión representa bien a la muestra.



□

Ejemplo 18 *Una Variable o característica Y se mide en diez dias sucesivos con los siguientes resultados:*

T	1	2	3	4	5	6	7	8	9	10
Y	0.9	3.6	5.8	6.8	7.1	7.3	7.2	7.4	7.3	7.4

1. *Calcular la recta de regresión de la variable Y sobre la T.*

2. *Calcúlese el coeficiente de correlación mestral entre T e Y. ¿La resta de regresión representa bien a la muestra?.*

3. *Calcúlese la varianza residual de Y sobre T, (error que hemos cometido). Represente la nube de puntos y la recta de regresión.*

4. *Si definimos la nueva variable o característica*

$$X = \log T,$$

calcúlese la recta de regresión de la variable Y sobre la X.

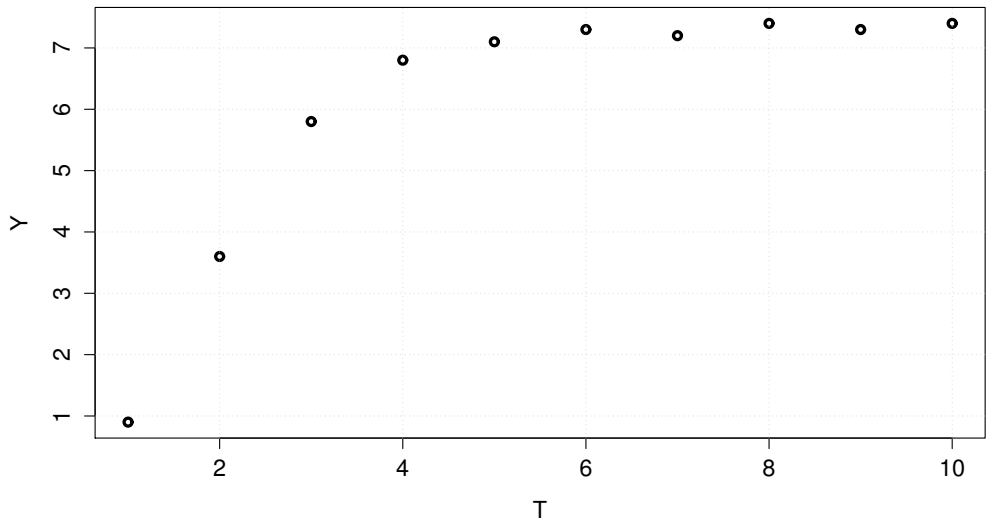
5. *Calcúlese el coeficiente de correlación mestral entre X e Y. ¿La resta de regresión representa bien a la muestra?.*

6. *Calcúlese la varianza residual de Y sobre X, (error que hemos cometido). Represente la nube de puntos y la "cuva de regresión"*

$$y(x) = a + b \log x.$$

Solución

La nube de puntos es:



1. Introducimos la notación:

$$t_1 = 1,\ t_2 = 2,\ t_3 = 3, \cdots, t_{10} = 10,$$

$$y_1 = 0.9,\ y_2 = 3.6,\ y_3 = 5.8, \cdots, y_{10} = 7.4,$$

$$\bar{t} = \frac{\sum_{i=1}^{10} t_i}{10} = 5.5,$$

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 6.08,$$

$$v_t = \frac{\sum_{i=1}^{10} t_i^2}{10} - \bar{t}^2 = \frac{385}{10} - 5.5^2 = 8.25,$$

$$v_y = \frac{\sum_{i=1}^{10} y_i^2}{10} - \bar{y}^2 = \frac{412}{10} - 6.08^2 = 4.2336,$$

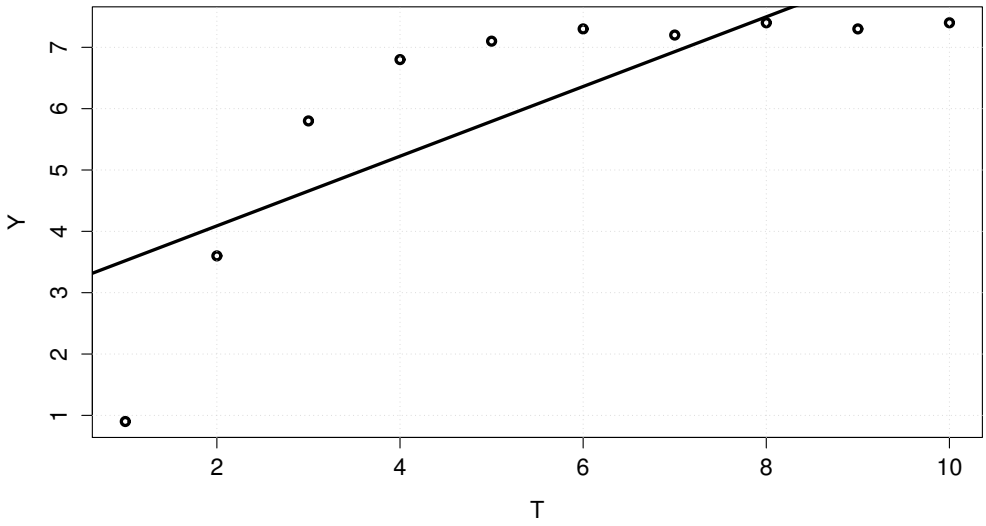
$$\mathrm{cov}_{t,y} = \frac{\sum_{i=1}^{10} t_i y_i}{10} - \bar{t}\bar{y} = \frac{381.3}{10} - \bar{t}\bar{y} = 4.69,$$

$$b = \frac{\mathrm{cov}_{t,y}}{v_t} = \frac{4.69}{8.25} = 0.568,$$

$$a = \bar{y} - \frac{\mathrm{cov}_{t,y}}{v_t}\bar{x} = 2.953$$

La recta de regresión $y = a + bt$ es

$$\boxed{y = 2.953 + 0.568t}$$



2. El coeficiente de correlación muestral entre T e Y es

$$r = \frac{\text{cov}_{t,y}}{\sqrt{v_t v_y}} = 0.7935$$

Esta lejos de 1 y por lo tanto **la recta de regresión de Y sobre T no representa bien a la muestra**.

3. El error que cometemos es

$$\text{varianza resdual} = v_y \left(1 - \frac{\text{cov}_{t,y}^2}{v_t v_y}\right) = v_y(1 - r^2) = 1.568$$

El error es muy grande.

4.

T	X	Y
$t_1 = 1$	$x_1 = \log t_1 = 0$	$y_1 = 0.9$
$t_2 = 2$	$x_2 = \log t_2 = 0.69$	$y_2 = 3.6$
$t_3 = 3$	$x_3 = \log t_3 = 1.09$	$y_3 = 5.8$
$t_4 = 4$	$x_4 = \log t_4 = 1.38$	$y_4 = 6.8$
$t_5 = 5$	$x_5 = \log t_5 = 1.7$	$y_5 = 7.1$
$t_6 = 6$	$x_6 = \log t_6 = 1.79$	$y_6 = 7.3$
$t_7 = 7$	$x_7 = \log t_7 = 1.94$	$y_7 = 7.2$
$t_8 = 8$	$x_8 = \log t_8 = 2.08$	$y_8 = 7.4$
$t_9 = 9$	$x_9 = \log t_9 = 2.2$	$y_9 = 7.3$
$t_{10} = 10$	$x_{10} = \log t_{10} = 2.3$	$y_{10} = 7.4$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{15.1}{10} = 1.51,$$

$$v_x = \frac{\sum_{i=1}^{10} x_i^2}{10} - \bar{x}^2 = \frac{27.65}{10} - 1.51^2 = 0.4849,$$

$$\mathrm{cov}_{x,y} = \frac{\sum_{i=1}^{10} x_i y_i}{10} - \bar{x}\bar{y} = \frac{105.28}{18} - \bar{x}\bar{y} = 1.3472,$$

$$b = \frac{\mathrm{cov}_{t,y}}{v_t} = \frac{1.3472}{0.4849} = 2.78,$$

$$a = \bar{y} - \frac{\mathrm{cov}_{x,y}}{v_t}\bar{x} = 1.885,$$

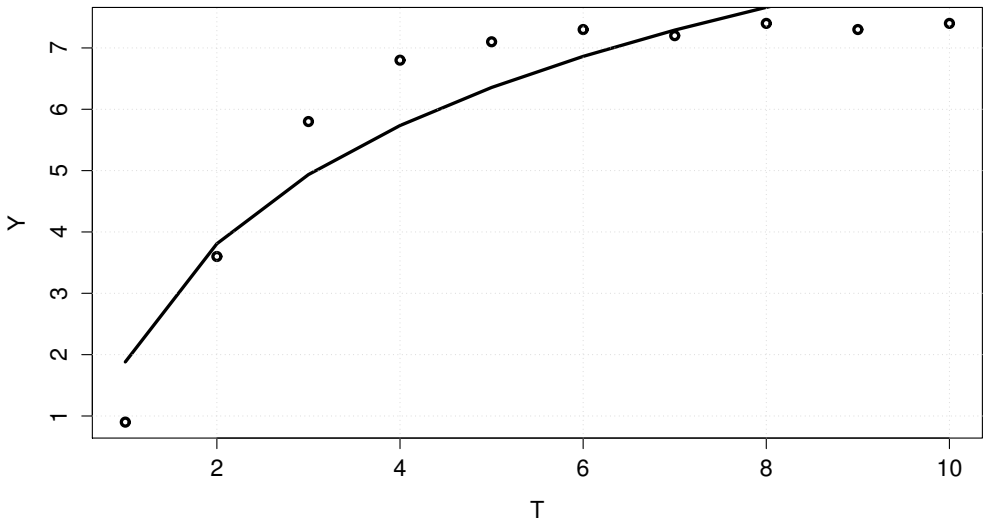
La recta de regresión $y = a + bx$ de Y sobre X es

$$\boxed{y = 1.885 + 2.78x},$$

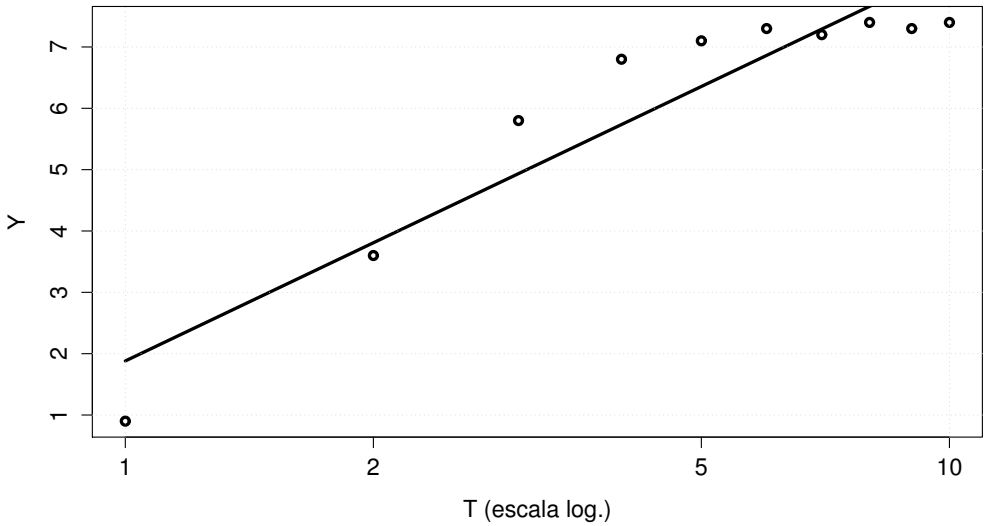
que produce la "curva deregresión" de Y sobre T

$$\boxed{y = 1.885 + 2.78 \log t}.$$

El gráfico en escala lineal muestra una *curva* de regresión:



Pero si tomamos $log(T)$ en el eje horizontal, la curva se transforma en una recta. Esta transformación logarítmica nos permite usar el mtodo de la recta de regresión para ajustar la curva logarítmica.



5. El coeficiente de correlación muestral entre X e Y (T y Y) es

$$r = \frac{\text{cov}_{x,y}}{\sqrt{v_x v_y}} = 0.94.$$

Esta "cerca" de 1 y por lo tanto **la curva de regresión de Y sobre T no representa mal a la muestra a la muestra**.

6. El error que cometemos es

$$\text{varianza resdual} = v_y \left(1 - \frac{\text{cov}^2_{x,y}}{v_x v_y}\right) = v_y(1 - r^2) = 0.49$$

El error es grande.

□

Ejemplo 19 *Una fábrica de cerveza, que se creo en enero del 2008, quiere averiguar si existe una relación lineal entre el dinero, en miles de euros, que gasta cada mes en anuncios de televisión y sus ventas, en miles de euros, de ese més.*

La población serán los meses de los años 2008 y 2009 y los cuatro primeros meses de 2010. Consideramos las variables o caracteristicas

X=dinero que gasta al més en anuncios de televisión

Y= ventas del mes.

Tomamos la siguiente muestra de estas variables de los últimos siete meses

Mes	Ventas	Gastos TV
Octubre 2009	50	0.5
Noviembre 2009	90	0.9
Diciembre 2009	30	0.4
Enero 2010	90	0.7
Febrero 2010	91	1.1
Marzo 2010	95	0.75
Abril 2010	95	0.8

La muestra también la podemos escribir:

$$M = \{(x_i, y_i),\ i = 1, 2, \cdots, 7\}$$

$$= \{(0.5, 50), (0.9, 90), (0.4, 30), (0.7, 90), (1.1, 91), (0.75, 95), (0.8, 95)\}.$$

Lo que tenemos que hacer es hallar la recta de regresión de la variable Y sobre la X asociada a la muestra M.