Tema 2

Descripción conjunta de varias variables

2.1. Introducción

Si queremos estudiar k características diferentes de cada individuo (u objeto) de la población, entonces trataremos con una variable aleatoria k-dimensional. [Nota: sólo veremos en este tema el caso $k = 2$, es decir, trabajaremos con 2 variables como mucho, aunque en la práctica 2 con el ordenador se verá un ejemplo con 3 variables].

En este tema, veremos como describir dos variables, de forma análoga a como se estudió en el tema anterior el caso univariante.

Ejemplo 2.1.: estudio de la relación existente entre la variable sexo de la persona y adicción al tabaco.

Ejemplo 2.2.: estudio de la relación existente entre la variable altura y peso de una persona.

Ejemplo 1.3.: (ratón ergonómico para niños), podría estudiarse no sólo la longitud del dedo índice, sino también anchura de la mano, longitud entre dos puntos (marcas) determinados, etc. (bajo el supuesto de un ratón clásico).

Ejemplo 1.4.: (gorras) podría interesarnos no sólo el color preferido sino también la distribución de la talla (tamaño de la cabeza).

2.2. Distribución de frecuencias bivariantes

Ejemplo 2.1.: En este ejemplo, deseábamos estudiar la relación entre las variables sexo (con posibles valores: 1 = Mujer, 2 = Hombre) y adicción al tabaco (1 = Fuma, 2 = No fuma). Como la población en este estudio es grandísima (aunque finita), para realizar el estudio nos basaremos en el análisis de una muestra de 50 personas, seleccionadas aleatoriamente. Por tanto, el tamaño muestral será 50. Igual que hicimos en el tema 1 (sección 1.2.), una primera aproximación para describir las variables, es por medio de una **tabla de contingencia o tabla de frecuencias cruzadas.**

Sexo \ Tabaco	Fuma (1)	No fuma (2)
Mujer (1)	10	18
Hombre (2)	7	15

A partir de esta tabla podemos extraer diversa información, por ejemplo:

- 10 individuos de la muestra fuman y son mujeres (**frecuencia absoluta**). Fijémonos que la suma de todas las casillas es 50, el tamaño muestral.
- $10/50 = 0.2$ (20 %), o sea, el 20 % de los individuos de la muestra fuman y son mujeres (**frecuencia relativa**).
- **Frecuencia marginal (absoluta o relativa)** para cada dimensión, es la frecuencia (absoluta o relativa) para cada variable, sin tener en cuenta los valores de la otra variable.
- Frecuencia marginal (absoluta) de la variable Sexo
  - $10 + 18 = 28$ mujeres en la muestra
  - $7 + 15 = 22$ hombres en la muestra
- Frecuencia marginal (relativa) de la variable Tabaco
  - $(10 + 7)/50 = 17/50 = 0.34$ El 34 % de la muestra fuma
  - $(18 + 15)/50 = 33/50 = 0.66$ El 66 % de la muestra no fuma
- ¿La proporción de fumadores es similar en ambos sexos?, o sea, ¿la adicción al tabaco se distribuye de igual manera para hombres y mujeres? Calcularemos: **frecuencias relativas condicionales** de Tabaco en función de Sexo

	Tabaco:	Sí	No
dado que Sexo = Mujer:		$10/28 = 0.36$	$18/28 = 0.64$
dado que Sexo = Hombre:		$7/22 = 0.32$	$15/22 = 0.68$

En este ejemplo hemos trabajado con 2 variables cualitativas. También se podría construir una tabla de frecuencias cruzadas para variables discretas o continuas (utilizando intervalos para agrupar los valores tal y como se vio en el tema 1 ⏮). En el tema 3, seguiremos trabajando este apartado.

El resto del tema lo dedicaremos al estudio descriptivo de dos variables aleatorias continuas, con el objetivo de explicar la variabilidad de una variable (variable dependiente o explicada) en función de otra (variable explicativa o independiente). Sólo veremos ideas simples sobre regresión (⏭ apartado 2.5), sin adentrarnos en el planteamiento de modelos (bibliografía). En los modelos de regresión, las variables explicativas son generalmente continuas y no controlables por el investigador. En el tema 6 (⏭ apartado 6.4), trataremos el diseño de experimentos donde

las variables explicativas son generalmente cualitativas y controlables: en el ejemplo 1.1. vimos que la variable a estudiar era el contenido en estaño (continua), mientras que el factor a controlar era el tipo de hojalata (cualitativa). Además en el tema 6 (⏭ apartado 6.3) también se verá la relación de dos variables cualitativas (pruebas de independencia y pruebas de homogeneidad). Por ejemplo, **ejemplo 2.3**: se efectúa un estudio sobre los fallos de un componente electrónico. Existen cuatro tipos de fallos posibles y dos posiciones de montaje para el dispositivo. Se toman los siguientes datos y deseamos probar si el tipo de fallo es independiente de la posición de montaje:

	Tipo de fallo			
Posición de montaje	A	B	C	D
1	22	46	18	9
2	4	17	6	12

De manera intuitiva (existe una definición formal), dos variables serán independientes (fíjate que la propia palabra lo expresa) cuando el conocimiento sobre el valor de una de ellas (fijamos el valor de una de ellas), no altera la distribución de valores de la otra, o sea, no nos aportaría información acerca de esta variable.

Lo que no se estudiará serán las series temporales. A veces el factor tiempo contribuye de manera notable en la variabilidad observada en los datos. Esto sólo se tratará en la práctica 3 sobre control de calidad, en las gráficas de control 🔓.

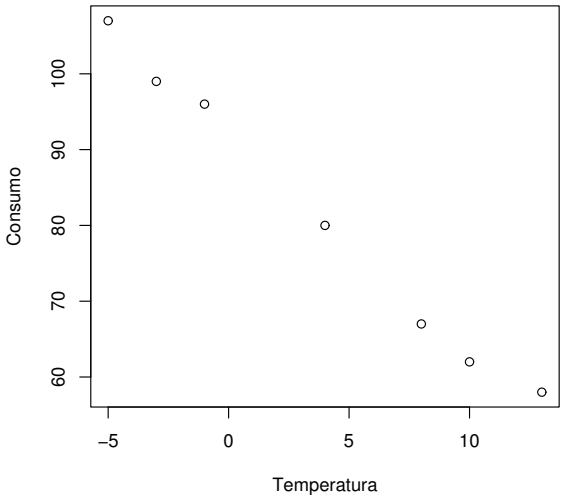
2.3. Representación gráfica

Existen diversas representaciones gráficas que tratan de visualizar el comportamiento conjunto de varias variables (si miráis en el programa de las prácticas, Statgraphics, en el Excel, etc. lo comprobaréis; para variables cualitativas, existe todo tipo de diagramas de barras). Nosotros, nos ceñiremos en este tema al diagrama de dispersión, que es adecuado sobre todo para representar una muestra proveniente de dos variables continuas.

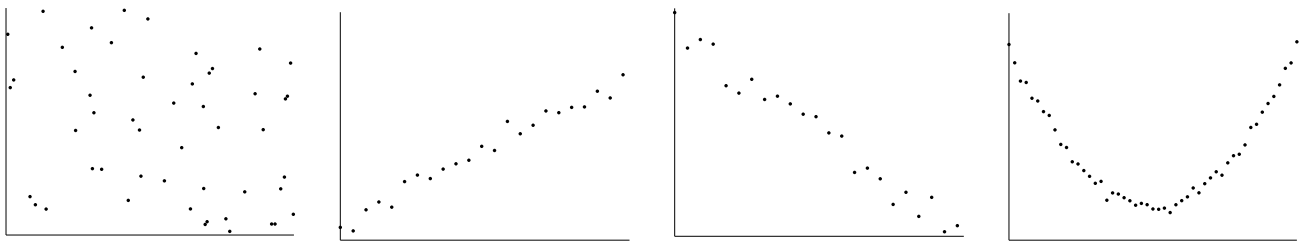
Ejemplo 2.4.: Una compañía local de energía seleccionó una residencia para desarrollar un modelo para el consumo de energía (en Kw por día) como una función de la temperatura promedio diaria durante los meses de invierno para cierto tipo de clientes. Se recogieron para 7 días los datos siguientes:

X = Temperatura (C^o)	13	10	8	4	-5	-1	-3
Y = Consumo	58	62	67	80	107	96	99

Representemos estos datos en un diagrama bivariante:



Según la forma de la nube de puntos, podemos ver la relación entre ambas componentes (si existe).



No relación (independientes) | Relación lineal (pendiente positiva) | Relación lineal (pendiente negativa) | Relación cuadrática

2.4. Medidas de dependencia lineal

Para cada variable por separado, podremos realizar los análisis descriptivos vistos en el tema 1 (medias, varianzas, etc. ⏮ apartado 1.4.). Pero, como estamos tratando con dos variables, además podremos calcular medidas que nos informarán acerca de la dependencia ("relación") lineal de las dos variables. En primer lugar definiremos la covarianza.

Covarianza: consideremos una muestra de N pares de puntos: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. La covarianza se define como:

$$s_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{N-1} = (observación \quad siguiente) = \frac{\sum_{i=1}^{N} x_i \cdot y_i - N \cdot \overline{x} \cdot \overline{y}}{N-1} \qquad (2.1)$$

Observación: *fórmula alternativa de la covarianza:*

$$\sum_{i=1}^{N}(x_i-\overline{x})\cdot(y_i-\overline{y})=\sum_{i=1}^{N}(x_i\cdot y_i-x_i\cdot\overline{y}-\overline{x}\cdot y_i+\overline{x}\cdot\overline{y})=\sum_{i=1}^{N}x_i\cdot y_i-\overline{y}\cdot\sum_{i=1}^{N}x_i-\overline{x}\cdot\sum_{i=1}^{N}y_i+N\cdot\overline{x}\cdot\overline{y}$$

$$=\textstyle\sum_{i=1}^{N}x_i\cdot y_i-\overline{y}\cdot N\cdot\overline{x}-\overline{x}\cdot N\cdot\overline{y}+N\cdot\overline{x}\cdot\overline{y}=\sum_{i=1}^{N}x_i\cdot y_i-N\cdot\overline{x}\cdot\overline{y}$$

[ FÍJATE que en $\sum_{i=1}^{N}x_i\cdot y_i$, se multiplican los pares de datos y luego se suman, NO se multiplican las sumas $\sum_{i=1}^{N}x_i$ y $\sum_{i=1}^{N}y_i$.]

Si $s_{xy}>0$ entonces valores grandes de X corresponden a valores grandes de Y (relación lineal positiva). En cambio si la relación es lineal negativa: valores grandes de X corresponden a valores pequeños de Y y $s_{xy}<0$. Cuando no exista relación lineal entre las variables, s_{xy} será próxima a cero.

 **Ejemplo 2.4.:**

$\overline{x}=$

$\overline{y}=$

$\sum_{i=1}^{7}x_i\cdot y_i=$

$s_{xy}=\frac{\sum_{i=1}^{N}x_i\cdot y_i-7\cdot\overline{x}\cdot\overline{y}}{6}=$

Sin embargo, la covarianza depende de las unidades en que estén expresadas las variables, por ello se define:

 **Coeficiente de correlación lineal:**

$$r_{xy}=\frac{s_{xy}}{s_x\cdot s_y} \tag{2.2}$$

 **Ejemplo 2.4.:**

$\sum_{i=1}^{7}x_i^2=$

$\sum_{i=1}^{7}y_i^2=$

$s_x^2=$

$s_y^2=$

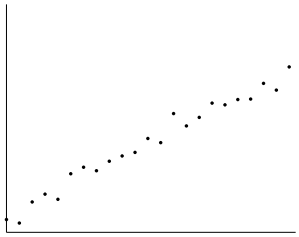
$s_x =$ $s_y =$

$r_{xy} =$

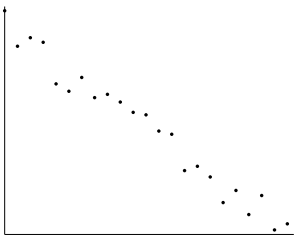
[FÍJATE también que cuando calculábamos la varianza, s_x^2, se sumaba cada dato al cuadrado: $\sum_{i=1}^{N} x_i^2$, NO se hace el cuadrado de la suma de los datos.]

Propiedad: El coeficiente de correlación está SIEMPRE comprendido entre -1 y 1.

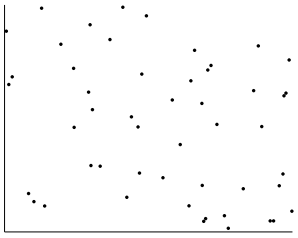
Si r_{xy} es cercano a 1: los puntos se encuentran alineados en una recta con pendiente positiva.



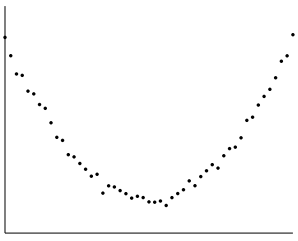
Si r_{xy} es cercano a -1: los puntos se encuentran alineados en una recta con pendiente negativa.



Si r_{xy} es cercano a 0: los puntos no se encuentran alineados, no existe relación lineal entre las variables, lo cual no quiere decir que no haya otro tipo de asociación entre las variables, por ello resulta fundamental examinar los diagramas de dispersión.



$r_{xy} = 0$ (no hay relación)



$r_{xy} = 0$ (pero hay relación cuadrática)

Ejemplo 2.4.: Interpretación de r_{xy}

Si intercambiamos las variables X e Y, el coeficiente de correlación no varía.

El coeficiente de correlación tampoco se ve afectado por transformaciones tales como sumar constantes y multiplicar todos los valores de una variable por una constante, en valor absoluto.

Correlación no implica causalidad, es decir, la observación de una fuerte relación entre las variables no necesariamente supone la existencia de una relación causal entre ellas. **Ejemplo 2.5.:** Número de matrimonios mensuales - Temperatura media mensual

2.5. Recta de regresión

Cuando r_{xy} está cerca de 1 ó -1, se puede ajustar una recta que nos puede servir para predecir otros valores de las variables (como haremos para el ejemplo 2.4). A veces, puede resultar difícil (costoso) medir cierta variable relacionada con otra de la que podemos obtener datos fácilmente, con lo cual la predicción nos puede solucionar las dificultades (problema 6 de este tema).

Supongamos que tenemos las N observaciones $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, y queremos ajustar una recta: $Y = a + b \cdot X$. Podemos hacerlo mediante el método de mínimos cuadrados:

Observación: *consiste en obtener a y b tales que minimicen la expresión siguiente (igualando a cero las derivadas):* $E = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - (a + b \cdot x_i))^2$ *;* $e_i = y_i - (a + b \cdot x_i)$ *se llama residuo o error.*

Recta de regresión de Y sobre X: (algunas calculadoras permiten calcularla)

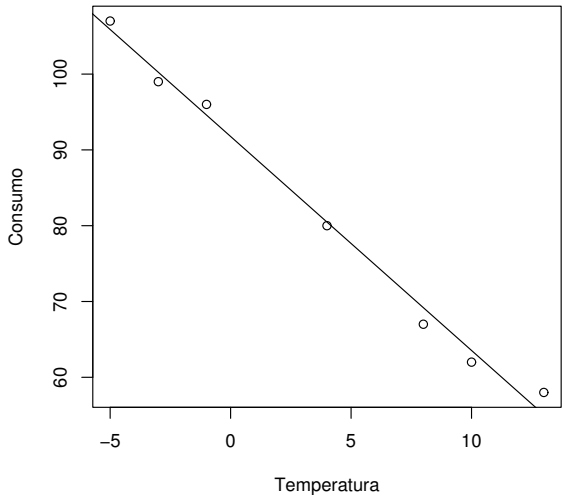
$$Y - \overline{y} = r_{xy} \frac{s_y}{s_x} (X - \overline{x}) = \frac{s_{xy}}{s_x^2} (X - \overline{x}) \tag{2.3}$$

o bien,

$$b = \frac{N \sum_{i=1}^{N} x_i \cdot y_i - (\sum_{i=1}^{N} x_i) \cdot (\sum_{i=1}^{N} y_i)}{N(\sum_{i=1}^{N} x_i^2) - (\sum_{i=1}^{N} x_i)^2}; \quad a = \frac{\sum_{i=1}^{N} y_i - b(\sum_{i=1}^{N} x_i)}{N} \tag{2.4}$$

Ejemplo 2.4.: Recta de regresión

Representémosla gráficamente:



La recta de regresión de X respecto de Y, no se obtiene despejando de la ecuación anterior, sino que es:

$$X - \overline{x} = r_{xy}\frac{s_x}{s_y}(Y - \overline{y}) = \frac{s_{xy}}{s_y^2}(Y - \overline{y}) \tag{2.5}$$

Podemos usar la recta de regresión para predecir valores. ✍ **Ejemplo 2.4.:** Predicción:

Temperatura = 0^0 C $\rightarrow$ Consumo :

Temperatura = 2^0 C $\rightarrow$ Consumo :

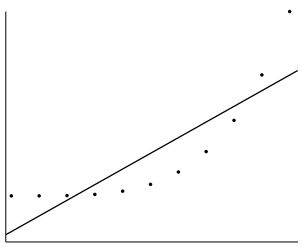
[En los problemas que resuelvas, en general, cuestiónate si el resultado que has obtenido tiene sentido o no, si entra dentro de los posibles resultados. Si la respuesta es negativa, busca las posibles causas y repásalas: redondeo, has usado datos diferentes en partes distintas del problema, te has olvidado alguna operación (suma, cuadrado, producto), has usado mal la calculadora, etc. En caso de no encontrar el error, lo mejor es empezar de cero, pues muy probablemente estés pasando por delante del error sin darte cuenta. Si aún así, no lograras descubrir el error, al menos señala tus dudas].

Sin embargo, sólo podemos fiarnos de la predicción si predecimos valores no demasiado alejados de los que tenemos, ya que nuestros datos únicamente proporcionan información en un cierto rango.

Ejemplo 2.4.: Cuidado con la extrapolación:

Temperatura = 40^0 C $\rightarrow$ Consumo :

También podríamos dar una predicción mala si el ajuste de la recta no es bueno.



Para conocer la calidad del ajuste se realiza un análisis de residuos. [Nota: el hecho de que b sea pequeño en $Y = a + bX$ NO implica un mal ajuste]. Como primera aproximación, se podría calcular la varianza de los residuos (e_i) (cuanto más cerca de cero más clara la relación lineal):

$$s^2_{residuos} = \frac{\sum_{i=1}^N e_i^2 - 0}{N-1} = \frac{\sum_{i=1}^N (y_i - (a + b \cdot x_i))^2}{N-1} = (nuestro \quad caso) = s_y^2(1 - r_{xy}^2)$$

Sin embargo, $s^2_{residuos}$ depende de las unidades de medida, por lo cual utilizaremos el **coeficiente de determinación**:

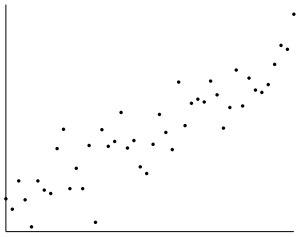
$$R^2 = 1 - \frac{s^2_{residuos}}{s_y^2} = r_{xy}^2$$

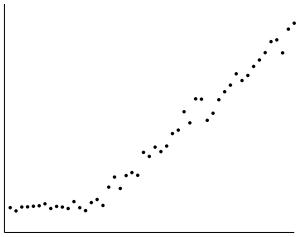
De ahí, es evidente que $0 \leq R^2 \leq 1$. Cuanto más cercano a 1 esté, mejor será el ajuste.

Ejemplo 2.4.: Calidad del ajuste:

$R^2 = r_{xy}^2 =$

Mirar la gráfica en cualquier caso:





A modo meramente orientativo (pues deberían realizarse otros análisis) podría decirse:

R^2	Ajuste
0.8 - 1	bueno
0.5 - 0.8	moderado
0.3 - 0.5	débil
0 - 0.3	malo

Problemas del tema 2

Ejemplo Se dispone de N = 13 datos, que expresan la relación entre la presión de vapor de β -trimetilborazol (Y, en mm de mercurio) y la temperatura (X, en grados centígrados):

X	13.0	19.5	22.5	27.2	31.8	38.4	45.7	56.1	64.4	71.4	80.5	85.7	91.5
Y	2.9	5.1	8.5	10.3	14.6	21.3	30.5	51.4	74.5	10.2	143.7	176.9	216.9

1. Decide, gráficamente, si hay algún outlier.

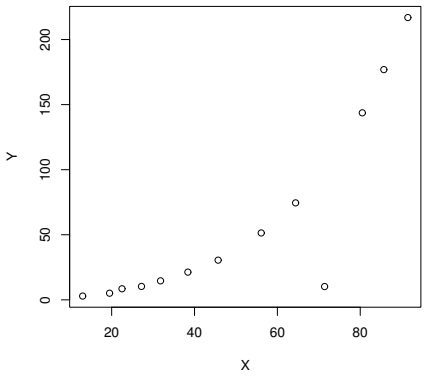
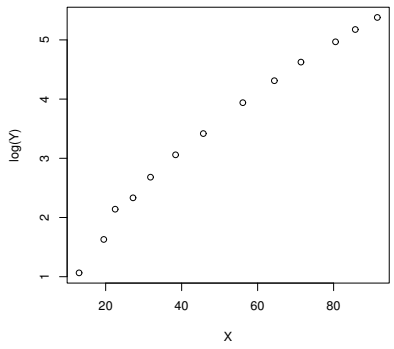


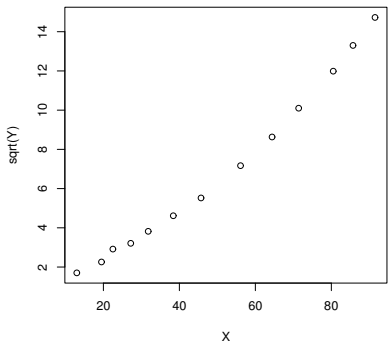
Figura 2.1: Diagrama bivariante

Hay un outlier, el punto (71.4,10.2), parece un error de trascripción, sería (71.4, 102).

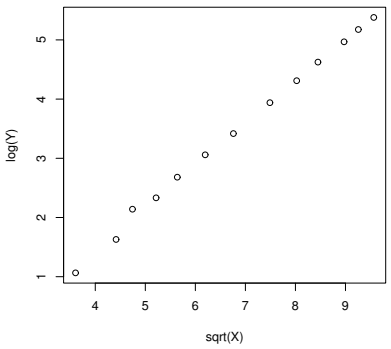
2. Dibuja diagramas bivariantes con los datos corregidos (si había outliers), para diferentes transformaciones de los datos (logaritmos, raíces cuadradas, ...), hasta que la nube se acerque a una recta. En ese caso, calcula la recta de regresión y determina la calidad del ajuste.



X - log(Y)



X - $\sqrt{Y}$



$\sqrt{X}$ - log (Y)

El mejor diagrama es el de la raíz de X (AX) y el logaritmo de Y (LY). Calculemos la recta de regresión:

LY = a + b AX ,

$$b = (\frac{N\sum ax_i ly_i - (\sum ax_i)(\sum ly_i)}{N\sum {ax_i}^2 - (\sum ax_i)^2}) = \frac{13 \cdot 337{,}65 - 88{,}34 \cdot 44{,}72}{13 \cdot 647{,}7 - 88{,}34^2} = 0{,}712$$

$$a = (\frac{\sum ly_i - b\sum ax_i}{N}) = \frac{44{,}72 - 0{,}712 \cdot 88{,}34}{13} = -1{,}39$$

Por tanto, Ly = -1.39 + 0.712 Ax .

$$R^2 = r^2 = \frac{s_{axly}}{s_{ax}s_{ly}} = \frac{\frac{\sum ax_i ly_i - N\overline{ax}\,\overline{ly}}{N-1}}{\sqrt{\frac{\sum ax_i^2 - N\overline{ax}^2}{N-1}}\sqrt{\frac{\sum ly_i^2 - N\overline{ly}^2}{N-1}}} = \frac{\frac{337{,}65 - 13 \cdot 6{,}8 \cdot 3{,}44}{12}}{\sqrt{\frac{647{,}7 - 13 \cdot 46{,}17}{12}}\sqrt{\frac{177{,}93 - 13 \cdot 11{,}83}{12}}} = \frac{2{,}8}{\sqrt{3{,}96 \cdot 2{,}01}} = 0{,}99,$$

es casi 1, es un ajuste muy bueno.

1. La resistencia del papel empleado en la fabricación de unas cajas (Y) se sabe que está relacionada con la concentración de madera dura en la pulpa original (X). Se han extraído las siguientes muestras, a partir de las cuales queremos conocer:

   a) la recta de regresión de la variable Y sobre la X
   b) la calidad del ajuste
   c) la resistencia de una caja fabricada con pulpa con concentración de 2.3

X	1	1.5	1.5	1.5	2	2	2.2	2.4	2.5	2.5	2.8	2.8	3	3	3.2	3.3
Y	101.4	117.4	117.1	106.2	131.9	146.9	146.8	133.9	111	123	125.1	145.2	134.3	144.5	143.7	146.9

(Cálculos: $\overline{x} = 2.325$, $\overline{y} = 129.706$, $\sum {x_i}^2 = 93.66$, $\sum {y_i}^2 = 272841.3$, $\sum x_i y_i = 4937.22$)

(Sol. : Y = 93.34 + 15.64 X, R^2 = 0.479, 129.312)

2. El tiempo que tarda un programa en realizar un determinado cálculo depende de la medida del archivo tratado. En 10 observaciones se han obtenido los siguientes datos:

X (Kb.)	352	387	254	317	428	231	276	324	441	510
Y (segundos)	22	25	20	22	28	17	19	23	25	29

Calcula:

a) la recta de regresión de la variable Y sobre la X

b) la calidad del ajuste

c) cuanto se tardaría en tratar un archivo de 300 Kb.

(Cálculos: $\overline{x} = 352$, $\overline{y} = 23$, $\sum {x_i}^2 = 1310956$,$\sum {y_i}^2 = 5422$, $\sum x_i y_i = 83895$)

(Sol. : Y = 8.63 +0.0408 X, R^2 =0.907, 20.87)

3. Se hace un estudio sobre la cantidad de azúcar refinado que se obtiene al variar la temperatura de un proceso determinado. Los datos se muestran en la tabla:

X (Temp.)	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
Y (Azúcar)	8.1	7.8	8.5	9.8	9.5	8.9	8.6	10.2	9.3	9.2	10.5

Determina:

a) la recta de regresión de la variable Y sobre la X

b) la calidad del ajuste

c) cuanta azúcar se obtendría para una temperatura de 1.25 .

(Cálculos: $\overline{x} = 1.5$, $\overline{y} = 9.127$, $\sum {x_i}^2 = 25.85$, $\sum {y_i}^2 = 923.58$, $\sum x_i y_i = 152.59$)

(Sol. : Y = 6.414 + 1.809 X, R^2 = 0.4999, 8.675)

4. A partir de una muestra de 10 piezas de latón, se quiere estudiar la influencia de la fuerza de tensión (X , en libras por pulgada cuadrada) sobre la dureza (Y en unidades Rockwell) del material. Los datos son:

X	64	65	66	69	73	74	76	79	80	83
Y	45	46	49	49	51	54	57	57	58	59

Determina:

a) la recta de regresión de la variable Y sobre la X

b) la calidad del ajuste

c) qué dureza se predice para X = 70 .

(Cálculos: $\overline{x} = 72.9$, $\overline{y} = 52.5$, $\sum {x_i}^2 = 53549$, $\sum {y_i}^2 = 27803$, $\sum x_i y_i = 38576$)

(Sol. : Y = -2.143 + 0.75 X, R^2 =0.9459, 50.357)

5. Uno de los aspectos de un programa de protección de residuos consiste en medir el contenido de un depósito. La determinación de su volumen se realiza indirectamente midiendo la diferencia de presión entre la parte más alta y más baja del tanque. Por la geometría del tanque, se sabe que la relación entre la presión y el volumen es aproximadamente lineal. Con el objetivo de calibrar la presión respecto al volumen, se colocan en el tanque cantidades conocidas de líquido y se toman lecturas de la presión. Los datos son (P: presión en Pascals, V: volumen en Kilolitros):

P						V					
215	218	633	629	1034	1033	0.189	0.19	0.377	0.379	0.567	0.568
1474	1475	1925	1922	2372	2374	0.757	0.758	0.946	0.947	1.135	1.136
2377	2819	3263	3262	3268	3712	1.137	1.327	1.514	1.515	1.516	1.705

o sea, los datos se leerían (215, 0.189), (218, 0.19), etc. Calcula:

a) la recta de regresión de la variable Volumen sobre la Presión.

b) la calidad del ajuste

c) cuál sería el volumen si la presión vale 2000 Pascals.

(Cálculos: $\overline{x} = 1889.167$, $\overline{y} = 0.9257$, $\sum {x_i}^2 = 85380065$, $\sum {y_i}^2 = 19.366$, $\sum x_i y_i = 40605.07$)

(Sol. : Y = 0.1102 + 0.00043 X, R^2 = 0.9997, 0.97)

6. La calidad de un jabón se determina por el contenido de ácido sebácico, que puede medirse mediante técnicas químicas. Para el uso en control de la calidad en fábricas de jabón, se ha sugerido determinar el porcentaje de ácido sebácico midiendo la conductividad eléctrica del jabón. La conductividad es fácil de medir y puede medirse en el lugar de producción. En la tabla siguiente se muestran una serie de medidas de la conductividad en mS (Milli-Siemens) para un determinado jabón y los correspondientes porcentajes de ácido sebácico.

C						A					
81.3	81.3	81.3	81.3	81.3	81.3	1.20	0.90	1.00	1.08	1.03	0.98
81.3	82.2	82.2	82.2	82.2	82.2	0.88	1.75	1.50	1.70	1.80	1.34
82.2	82.2	82.2	82.3	82.3	82.3	1.44	1.49	1.24	1.52	1.52	1.67
82.3	82.3	82.3	82.3	82.3	83.0	1.67	1.35	1.50	1.30	1.45	2.10
83.0	83.0	83.0	83.0	83.0	83.0	1.95	1.85	1.90	2.35	2.22	2.00

Los datos se leerían (81.3, 1.2), (81.3, 0.9), (81.3, 1), etc. Calcula:

a) la recta de regresión de la variable A (% ácido) sobre la C (conductividad).

b) la calidad del ajuste

c) cuál sería el porcentaje de ácido si la conductividad es 82mS .

(Cálculos: $\overline{x} = 82.2$, $\overline{y} = 1.52$, $\sum {x_i}^2 = 202731.9$, $\sum {y_i}^2 = 74.041$, $\sum x_i y_i = 3761.227$)

(Sol. : Y = -48.128 + 0.604 X, R^2 = 0.832, 1.4)

7. Durante la investigación de la contaminación de un fiordo, se tomaron diversas muestras de agua a diferentes profundidades. Para medir el grado de polución se determina la concentración de una bacteria. La siguiente tabla muestra el logaritmo de la concentración de la bacteria (B) a diferentes profundidades (P).

P						B					
0	0	0	0	0	0	1.95	2.42	2.56	2.08	2.15	2.42
0	0	0	0	4	4	1.90	2.15	1.95	2.42	2.15	2.15
4	4	4	4	4	4	2.15	1.95	2.18	1.95	2.42	1.78
4	4	8	8	8	8	2.15	2.26	2.15	2.26	1.78	1.90
8	8	8	8	8	8	2.15	1.78	2.15	2.56	1.78	1.95

Los datos se leerían: (0, 1.95), (0, 2.42), (0, 2.56). Calcula:

a) la recta de regresión de la variable B sobre la P.

b) la calidad del ajuste

c) Dibuja un diagrama de las variables P y B. ¿Hay relación entre ellas?

(Cálculos: $\overline{x} = 4$, $\overline{y} = 2.12$, $\sum {x_i}^2 = 800$, $\sum {y_i}^2 = 136.3354$, $\sum x_i y_i = 248.24$)

(Sol. : Y = 2.197 - 0.019 X, R^2 = 0.079, No)

8. Se realiza un experimento para determinar la duración de vida de ciertos circuitos electrónicos (Y) en función de cierta variable de fabricación X. Se han obtenido los siguientes resultados:

X	-10	-15	20	-10	5	5
Y	11	8	73	21	46	30

Determina:

a) la recta de regresión de la variable Y sobre la X

b) la calidad del ajuste

c) cuánto durará si $X = 0$.

(Cálculos: $\overline{x}$ = -0.833, $\overline{y} = 31.5$, $\sum {x_i}^2 = 875$, $\sum {y_i}^2 = 8971$, $\sum x_i y_i = 1400$)

(Sol. : Y = 32.99 + 1.788 X, R^2 =0.92, 32.99)

9. En un trabajo sobre seguridad vial se estudió la posible relación entre la velocidad de un determinado vehículo y su distancia de frenado. Uno de los objetivos del análisis es determinar si la relación entre ambas variables es aproximadamente lineal o si la velocidad está relacionada linealmente con la raíz cuadrada de la distancia de frenado, como sugiere una ley física. Los datos siguientes muestran diferentes velocidades del vehículo (en Km/h) con sus correspondientes distancias de frenado (en metros).

V	F			
32	3.74	5.1	4.4	
48	8.2	10.58	9.55	
64	11.67	21.46	22.95	20.95
80	36.81	39.52	37.45	35.45
97	54.48	52.17	46	50.59

Los datos se leerían: (32, 3.74), (32, 5.1), (32, 4.4), (48, 8.2), etc.

a) Calcula la recta de regresión estimada de la distancia de frenado (Y) sobre la velocidad (X).

b) Calcula la recta de regresión estimada de la raíz cuadrada de la distancia de frenado (Z) sobre la velocidad.

c) ¿Cuál parece más adecuada?

(Cálculos: $\overline{x} = 66.889$, $\overline{y} = 26.171$, $\overline{z} = 4.769$, $\sum {x_i}^2 = 89604$, $\sum {y_i}^2 = 17833.2$, $\sum {z_i}^2 = 471.07$, $\sum x_i y_i = 38366.12$, $\sum x_i z_i = 6476.414$)

(Sol. : Y = -24.398 + 0.756 X, R^2 = 0.94 ; Z = -0.6498 +0.081 X, R^2 = 0.96, la segunda)

10. En un estudio para relacionar la longitud de la línea de la vida en la mano izquierda y la vida de una persona, se han obtenido datos de 50 personas con los siguientes resultados (X = longitud de la línea en cm; Y =edad al morir en años):

$\sum y_i = 3333$, $\sum x_i = 459.9$, $\sum {x_i}^2 = 4308.57$, $\sum {y_i}^2 = 231933$, $\sum x_i y_i = 30549$

Calcula la recta de regresión y determina la calidad del ajuste.

(Sol. : Y = 66.66 - 1.38(X - 9.20), R^2 = 0.015)