| | | |
|---|---|---|
| | **INTRODUCCIÓN A LA ESTADÍSTICA**<br>*Problemas Tema 3: Estadística Descriptiva Bidemensional* | |

## Pregunta 1ª

La probabilidad de que las personas desarrollen algún tipo de cáncer está relacionada con el nivel de emisiones radiactivas. Se ha estimado esta relación tanto para la central de Chernóbil como para la de Fukushima, en función del nivel de radiación.

Número de enfermos de cáncer por millón de habitantes= 0,87+1,84*Nivel de radiación de Chernóbil

Número de enfermos de cáncer por millón de habitantes = 0,64+2,05*Nivel de radiación de Fukushima

Se pide:

a) ¿A partir de qué nivel de emisión, el número de enfermos de cáncer por millón de habitantes será igual o superior en Fukushima que a los de Chernóbil?

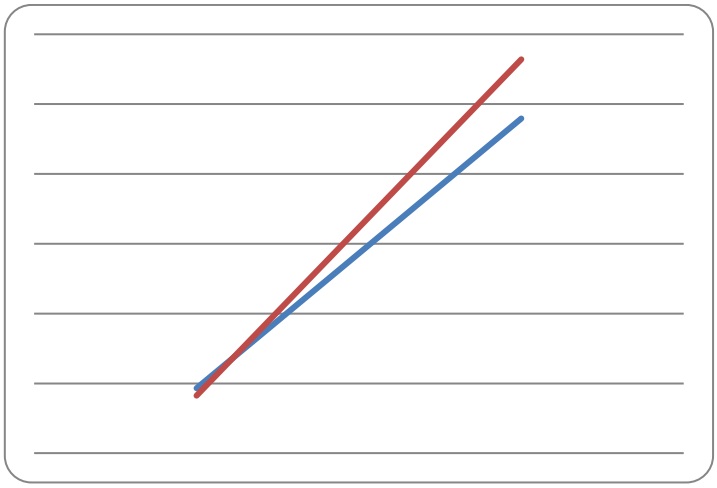
**Se trata de igual ambas rectas**

**0,87+1,84 Nivel de radicación = 0,64+2,05 Nivel de radicación**

**0,21 Nivel de Radiación = 0,23**

**Nivel de Radiación = 1,095. Luego a partir de este nivel de radiación el número de enfermos de cáncer por millón de habitantes será igual o superior en Fukushima que los de Chernóbil.**

**Se puede observar así mismo en la siguiente gráfica:**



b) ¿Qué tasa de cáncer se pronostica para un nivel de radiación de 410 en Fukushima?

**Número de enfermos de cáncer por millón de habitantes en Fukushima= 0,64+2,05 (410)= 841,14**

c) Sabiendo que la varianza residual tiene un valor de 20, ¿entre que límites se espera que oscile el número de enfermos de cáncer para una radiación de 410 en Fukushima? (límites calculados al 95%)

**841,14 ± 2 $S_{resid}$**

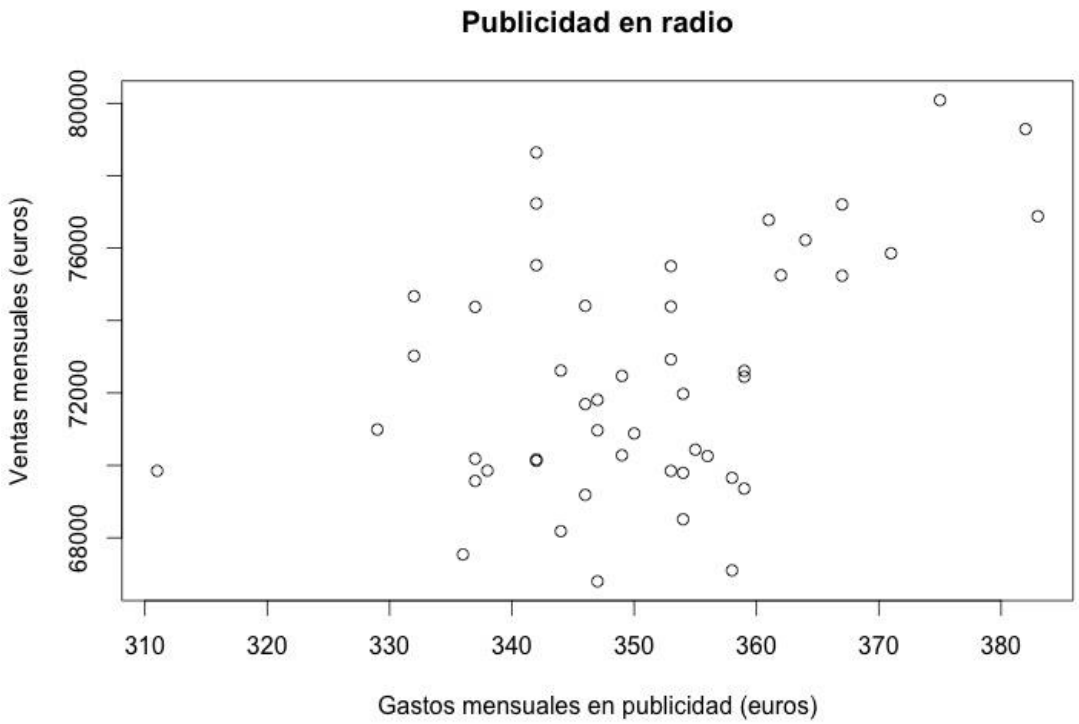
**$S_{resid}$ = $\sqrt{S_{resid}}$=$\sqrt{20}$=4,47**

**Luego sustituyendo [841,14 ± (2*4,47)], es decir, entre 832,2 y 850,08.**

## Pregunta 2ª

Una pequeña empresa ha decidido insertar cuñas de publicidad en dos emisoras de radio locales (A y B), para ver el efecto que tienen estos anuncios sobre sus ventas.

**a)** El siguiente gráfico de dispersión se ha obtenido utilizando los gastos en publicidad y las ventas de los últimos meses.




¿Qué información se desprende del gráfico anterior? Redacta un breve informe en el contexto del problema, sabiendo, además, que el coeficiente de correlación lineal entre ambas variables es r = 0.45.

*Observando el diagrama de dispersión, se aprecia una relación lineal positiva pero débil entre los gastos mensuales en publicidad y las ventas mensuales de esta empresa. En general, valores altos de una variable van asociados con valores altos de la otra, mientras que valores bajos de una variable van asociados con valores bajos de la otra.*

*En el contexto del problema, esto significa que aquellos meses en los que el gasto en publicidad es elevado se obtienen resultados de ventas mejores (más elevados) que en aquellos meses en los que el gasto en publicidad en radio es menor.*

*Esta relación lineal positiva es débil. Así lo corrobora el coeficiente de correlación lineal r=0.45 (valor positivo pero alejado de 1).*

*Si se obtuviese un modelo de regresión lineal para predecir las ventas mensuales en función del gasto en publicidad, el coeficiente de determinación, que mide la bondad del ajuste, sería* $R^2 = 0.45^2 = 0.2025$*, es decir, aproximadamente el 20% de la variabilidad total de las ventas mensuales estaría explicada por el modelo (asociada a los costes mensuales de la publicidad en radio).*

**b)** Observando el gráfico anterior, un estudiante decidió realizar un ajuste de regresión por mínimos cuadrados, obteniendo estas estimaciones para los parámetros del modelo:

Constante = 33524.79 Pendiente = 111.12

Explica qué significado tienen estas estimaciones en el contexto del problema.

*Constante: 33524.79 €. Es la cifra media de ventas de aquellos meses en los que el gasto de publicidad en radio sea de 0€ (aquellos meses en los que la empresa no inserta cuñas de publicidad en la radio). La variable explicativa (gastos mensuales en publicidad) no contempla valores próximos al 0 (ya que el rango de valores observados para esta variable se encuentra entre 310 y 380 € mensuales aproximadamente) y, por lo tanto, este valor debe interpretarse simplemente como un valor de ajuste.*

*Pendiente: 111.12 (€ en ventas/€ en publicidad). Es el incremento de la cifra media mensual de ventas por cada euro adicional que se gasta mensualmente en publicidad en radio.*

**c)** Explica qué son los residuos de este modelo y para qué se utilizan.

*Los residuos de un modelo de regresión son las distancias verticales de los puntos a la recta de regresión y recogen el efecto conjunto de todos los factores que afectan a la variable respuesta y que el modelo de regresión no contempla.*

*En el contexto de este problema, los residuos recogerán el efecto conjunto de todas aquellas variables que influyan en las ventas mensuales de la empresa, además del gasto mensual de publicidad en radio. (Algunas de estas variables podrían ser: el número de días festivos del mes, el número de días del mes en los que se hacen ofertas especiales, el mes en concreto, etc.)*

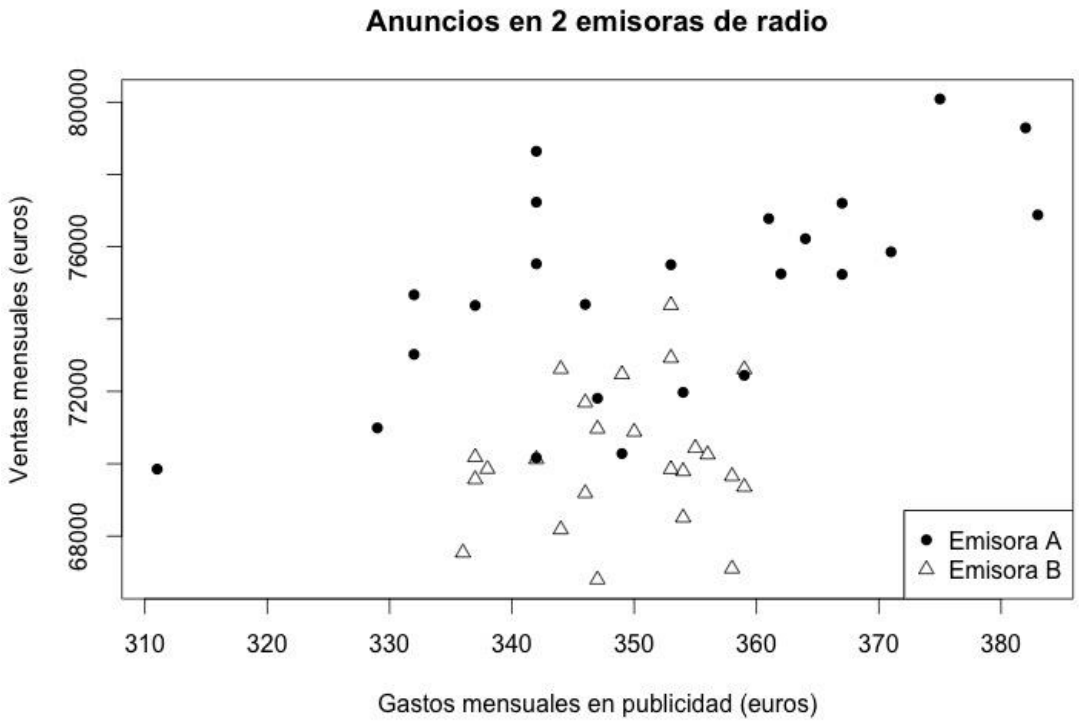
*Los residuos se utilizan para validar el modelo de regresión. Al hacer el ajuste se asume que los residuos son independientes y que siguen un modelo de distribución normal, con media nula y varianza constante. La comprobación de estos supuestos puede y debe hacerse analizando los residuos.*

**d)** Explica por qué el método de ajuste utilizado recibe el nombre de *ajuste por mínimos cuadrados*.

*De todas las posibles rectas que pueden trazarse para aproximar la nube de puntos del diagrama de dispersión, el método de mínimos cuadrados proporciona aquella recta que **minimiza** la suma del **cuadrado de las distancias verticales** de los puntos a la recta de regresión. De ahí el nombre*

*de ajuste por mínimos cuadrados. (Nota: es necesario elevar estas distancias al cuadrado para evitar que, al realizar la suma, se compensen las distancias positivas con las negativas).*

e) Finalmente, se obtuvo el siguiente gráfico, para diferenciar los datos según la emisora de radio en la que se habían insertado las cuñas de publicidad:




¿Cambian las conclusiones del estudio tras observar este gráfico? Justifica en qué emisora de radio debería anunciarse esta empresa

*Al diferenciar los puntos según la emisora en la que se insertan las cuñas de publicidad se observa un aspecto interesante: la relación entre las ventas mensuales y los gastos mensuales en publicidad, depende de la emisora en que se insertan las cuñas.*

*Si consideramos únicamente los datos para la emisora A, se aprecia una relación lineal positiva (más fuerte que al considerar todos los puntos), mientras que si nos fijamos en los datos para la emisora B, la relación entre las ventas mensuales y los gastos mensuales en publicidad es muy débil o prácticamente inexistente.*

*Comparando los dos grupos de datos, vemos que, en general, se obtienen mejores resultados de ventas cuando se insertan las cuñas en la emisora A. Además, valores altos del gasto en publicidad en esta emisora, suelen ir asociados a valores altos de ventas. Por lo tanto este gráfico*

*sugiere que deberíamos elegir la emisora A para insertar las cuñas, si nuestro objetivo es maximizar la cifra mensual de ventas.*

## **Pregunta 3ª**

En las pruebas de acceso del último año se seleccionaron al azar 120 alumnos de tres tipos de colegio y se tomaron en cuenta las notas obtenidas por cada uno de ellos. Con estos datos se definió la variable aleatoria bidimensional (Tipo de colegio, Calificaciones obtenidas) como muestra la tabla siguiente:

| | SUSPENSOS | APROBADOS | NOTABLES | SOBRESALIENTES | Total fila |
|---|---|---|---|---|---|
| PUBLICO | 3 | 15 | 22 | 6 | 46 |
| PRIVADO | 3 | 24 | 8 | 5 | 40 |
| CONCERTADO | 4 | 8 | 17 | 5 | 34 |
| Total columna | 10 | 47 | 47 | 16 | 120 |

a) Completar la tabla anterior calculando las probabilidades de la distribución bidimensional conjunta de la variable (Tipo de colegio, Calificaciones) más interesantes.

| | SUSPENSOS | APROBADOS | NOTABLES | SOBRESALIENTES | Total fila |
|---|---|---|---|---|---|
| PUBLICO | 3/46 | 15/46 / 22/46 | | 6/46 | 46 |
| PRIVADO | 3/40 | 24/40 | 8/40 | 5/40 | 40 |
| CONCERTADO | 4/34 | 8/34 | 17/34 | 5/34 | 34 |
| Total columna | 10 | 47 | 47 | 16 | 120 |

b) Completar la tabla siguiente con las distribuciones unidimensionales marginales de las variables Tipo de colegio y Calificaciones y con loa condicional de las notas en el colegio privado. **(0,5 puntos)**.

| | SUSPENSOS | APROBADOS | NOTABLES | SOBRESALIENTES | Total fila |
|---|---|---|---|---|---|
| PUBLICO | 3 | 15 | 22 | 6 | 46 y 46/120 |
| PRIVADO | 3 | 24 | 8 | 5 | 40 |
| CONCERTADO | 4 | 8 | 17 | 5 | 34 |
| Total columna | 10 10/120 | 47 | 47 | 16 | 120 |

c) ¿A qué tipo de probabilidades corresponderían las frecuencias relativas calculada en el apartado a?

*A las probabilidades condiciónales de Calificación condicionado a centro.*

## **Pregunta 4ª**

Una empresa de telefonía móvil ha llevado a cabo un estudio de consumo telefónico mensual (en euros) para varios de sus clientes, de los cuales ha apuntado también la edad en el momento del estudio. Los datos se muestran a continuación:

| Dato | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Edad | 20 | 45 | 45 | 40 | 56 | 52 | 57 | 37 | 68 | 17 | 43 | 23 | 29 | 58 | 39 | 49 | 53 | 26 | 26 | 27 |
| Consumo | 14 | 9,5 | 3,5 | 8 | 6,2 | 7,4 | 2 | 10,9 | 5,6 | 18,9 | 8,1 | 17,1 | 13,3 | 2,6 | 6,3 | 7,3 | 1,1 | 15,2 | 11,2 | 9,9 |

Se proporciona además la siguiente información:

Edad:

Media: 40,5

Desviación típica: 14,5295

Consumo:

Media: 8,905

Desviación típica: 4,9813

Covarianza entre Edad y Consumo: -62,2711

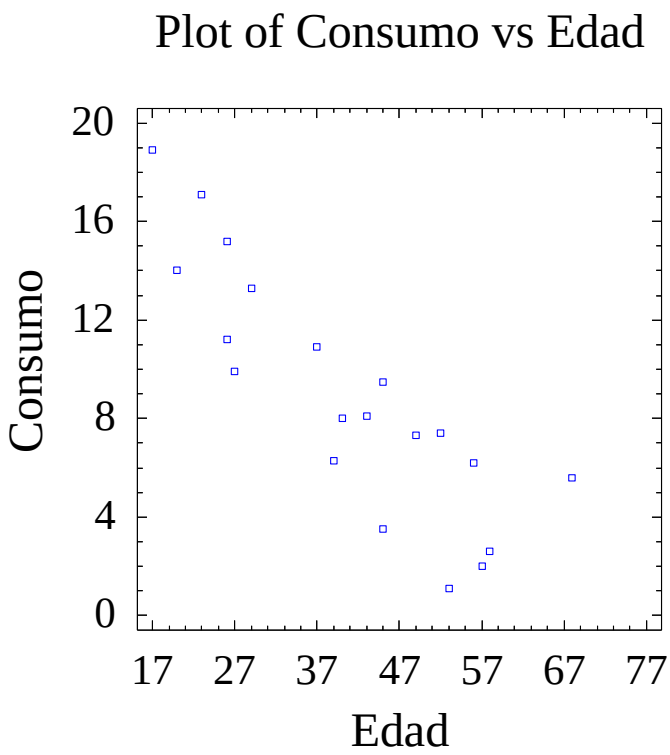
Se pide:

**a)** Define claramente las variables objeto del estudio. Especifica qué variable es dependiente y cuál independiente.

*La variable edad es la variable independiente X y es una variable cuantitativa discreta.*

*La variable Consumo es la variable dependiente Y y es una variable cuantitativa que puede considerarse continua.*

**b)** Dibuja el diagrama de dispersión entre las variables Consumo y Edad. ¿Qué puedes concluir a partir del diagrama?




*Se puede observar como hay una cierta relación lineal inversa. A mayor edad, menor consumo telefónico.*

**c)** Calcula todos los parámetros de la recta de regresión para poder calcular el nivel de Consumo en función de la Edad. Dibuja la recta de regresión en el diagrama de dispersión del apartado anterior. ¿Cuál es la tarifa fija que la compañía cobra a sus clientes con independencia de la edad?

*Queremos calcular la recta:* $y = a + b \cdot x$

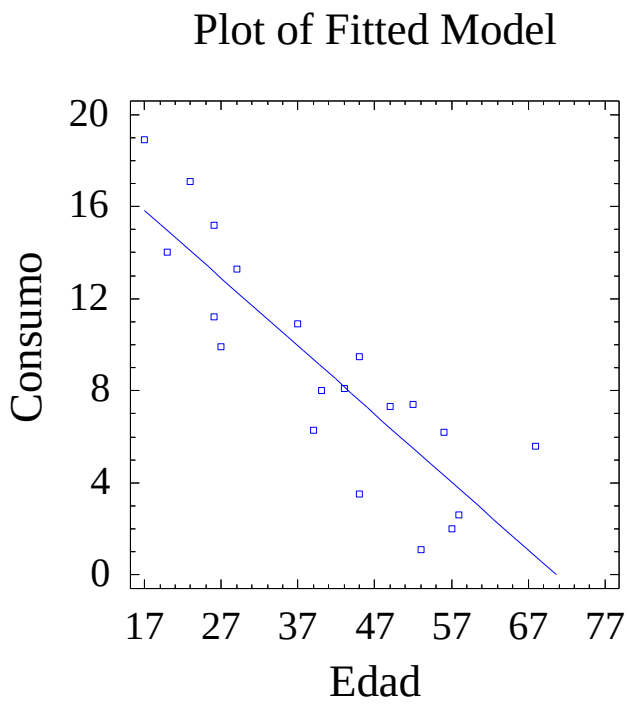
*Para ello aplicaremos las fórmulas:* $b = r_{x,y} \frac{s_y}{s_x} = \frac{Cov_{x,y}}{s_x^2}$ $\qquad a = \bar{y} - b\bar{x}$

*Aprovechando los datos que nos han dado en el problema tenemos que:*

*b=(-62,2711)/(14,5295*14,5295)=-0,295*

*a=8,905-((-0,295)* 40,5)=20,8515*

*Luego ya tenemos la recta que podemos dibujar:*




*La tarifa fija que la compañía cobra a sus clientes con independencia de la edad se podría considerar como el término independiente a de la recta, es decir, 20,8515 €.*

**d)** Comenta la validez del ajuste proporcionado por la recta de regresión, calculando el estadístico correspondiente.

*La validez del ajuste se tiene que obtener a través del coeficiente de determinación:*

que se calcula con la siguiente expresión:

$$R^2 = r(x,y)^2 * 100 = (-0{,}86^2) * 100$$

*Ya que:*

*rx,y=(-62,2711)/( 14,5295*4,9813)=-0,8604.*

*Como ya sabíamos, es un coeficiente de correlación negativo, es bastante cercano a -1 pero no es un ajuste perfecto*.

**e)** De cada 100 clientes analizados de 50 años, ¿cuál es el límite máximo de consumo que se espera exceder en máximo 5 de ellos?.

*Para poder contestar a esta cuestión, necesitamos conocer la varianza residual y el intervalo para las predicciones al 95%:* $s^2_{res} = s^2_y\left(1-r^2_{x,y}\right)$

*Por tanto, Sres2=(4,9813*4,9813)*(1-(-0,8604*-0,8604))=6,4446.*

*A partir de Sres2, sabemos que el 95% de los clientes de 50 años estarán a +- 2 desviaciones típicas:*

*(a+b*50)+-2*sres2= (20,8515 + -0,295*50)+-2*6,4446 = 6,1027+-12,8893. Dicho de otra forma, el consumo esperado para el 95% de los clientes de 50 años es de 6,1027+-12,8893. Por tanto, en el 5% de los casos podemos esperar un consumo superior a 6,1027+12,8893= 18,9920 €*

## **Pregunta 5ª**

Un ingeniero que trabaja en una gran empresa dedicada a la fabricación y montaje de vehículos necesita implementar un modelo para predecir el *tiempo de entrega* de los vehículos nuevos, considerando los complementos extras que incluyen. Se entiende como *tiempo de entrega* el número de días discurridos entre el pedido de un coche y la entrega real del mismo. El ingeniero quiere analizar si entre el número de extras que incluyen los pedidos respecto de la configuración básica del coche solicitado, y el tiempo de entrega del mismo, puede haber una relación lineal. Para formular este modelo de relación selecciona aleatoriamente una muestra de 20 pedidos de los que recoge los siguientes datos relativos al *número de extras* y el *tiempo de entrega* de cada vehículo. A continuación se muestran los datos obtenidos:

Sx=1.8033

Sy=0.5242

cov_xy = 0.92062

$m_x$= 12.5495

$m_y$= 13.7658

**a)** Plantear el modelo de regresión del *tiempo de entrega* en función del *número de extras*, indicando cuáles son sus parámetros y la ecuación del modelo. (0.5 puntos)

La ecuación del modelo será: y = a + bx, siendo:

$b = s_{xy} / s_x^2 = 0.92062 / 3.2519 = 0.2831$

$a = m_y – b.m_x = 13.7658 - 0.2831 \times 12.5495 = 10.213$

La ecuación del modelo es por tanto:

$t_{entrega}$ (días) = 10.213 + 0.2831 × num_extras (número)

**b)** Explicar brevemente la interpretación práctica que tienen los parámetros del modelo.

a: representa el tiempo medio de entrega para un pedido de un vehículo con la configuración básica, sin extras. Se mide en días (y su fracción).

b: representa el número de días (fracción en este caso) en los que aumenta el tiempo de entrega del pedido con cada extra adicional sobre la configuración básica del vehículo. Se mide en días/num_extras.

**c)** ¿Qué porcentaje de la variabilidad del *tiempo de entrega* viene explicada por el *número de extras*? Indica el parámetro que cuantifica dicho porcentaje.

Este porcentaje de variabilidad viene expresado por el valor del Coeficiente de Determinación, que se calcula como:

$R = r_{xy}^2 \times 100 = (s_{xy} / (s_x s_y))^2 \times 100 = (0{,}92062/1{,}8033 \times 0{,}5242)^2 \times 100 = 94{,}85\ \%$

**d)** Si se recibe un pedido de un vehículo con 16 extras, ¿cuántos días, en promedio, predice el modelo para la entrega?

El tiempo de entrega medio para 16 extras se puede calcular directamente a partir de la ecuación del modelo:

$t_{entrega}$ (16 extras) = 10.213 + 0.2831 × 16 = 14.74 días

**e)** ¿Entre que valores aproximadamente estará el *tiempo de entrega*, en promedio, en el 95% de los pedidos en los que se solicitan 16 extras?

La variabilidad residual se puede calcular como:

$s_{res}^2 = s_y^2 (1 - r_{xy}^2) = 0.01415$

$s_{res} = 0.119$

El intervalo de valores es por tanto, asumiendo una distribución normal para la variabilidad residual:

$[t_{entrega}$ (16 extras) $- 2\ s_{res}; t_{entrega}$ (16 extras) $+ 2\ s_{res}] =$

$[14.74 - 2 \times 0.119; 14.74 + 2 \times 0.119] = [14.50 ; 14.98]$

Es decir prácticamente entre 14.5 días como mínimo y 15 días como máximo.

**Problema 6**

Una industria de reciclaje de residuos dispone de un sistema automático de clasificación de los mismos. Con el fin de evaluar su rendimiento y grado de acierto se ha revisado una muestra completa de todos los residuos que ha clasificado durante un turno de trabajo. La clasificación automática de los residuos se realiza según 3 categorías diferentes: Orgánico (O), Envases (E) y Papel (P). A continuación se muestra la denominada matriz de confusión de dicho sistema clasificador:

| | | **Sistema Clasificador** | | |
|---|---|---|---|---|
| | | **Oc** | **Ec** | **Pc** |
| **Realidad** | **Or** | 350 | 10 | 11 |
| | **Er** | 15 | 200 | 12 |
| | **Pr** | 3 | 5 | 150 |

Siendo respectivamente:

Oc = {El sistema clasificador dice que el objeto es orgánico}

Ec = {El sistema clasificador dice que el objeto es envase}

Pc = {El sistema clasificador dice que el objeto es papel}

Or = {El objeto es en realidad orgánico}

Er = {El objeto es en realidad envase}

Pr = {El objeto es en realidad papel}

A la vista de esta información, responde a las siguientes preguntas:

**a)** ¿Qué porcentaje de los objetos son realmente orgánicos?

$P(Or) = (350+10+11)/756 = 0.4907 \Longrightarrow 49.07\%$

**b)** ¿Qué porcentaje de objetos han sido clasificados como envase, aunque realmente eran papel?

$P(Ec \cap Pr) = 5/756 = 0.0066 \Longrightarrow 0.66\%$

**c)** ¿Qué porcentaje de objetos de entre los clasificados como orgánico, eran realmente papel?

$P(Pr \mid Oc) = 3/368 = 0.0082 \Longrightarrow 0.82\%$

**d)** ¿Qué porcentaje de objetos han sido clasificados como envase?

P(Ec) = 215 / 756 = 0.2844 ==> 28.44%

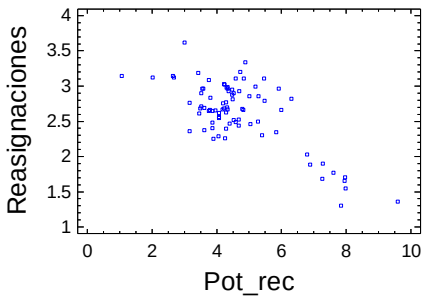
**e)** ¿Cuál es el porcentaje de acierto del sistema clasificador?

P(acierto) = P(Ec ∩ Er) + P(Pc ∩ Pr) + P(Oc ∩ Or) = 350/756 + 200/756 + 150/756 = 700/756 = 0.9259 ==> 92.59%

## Problema 7

Una compañía de telefonía móvil ha venido detectando problemas de funcionamiento en una estación de su red. Para tratar de resolverlos ha estudiado para esta estación un indicador de calidad de la llamada, que pretende estudiar el número de veces que se reasigna el canal de transmisión una vez la llamada está en curso. Para ello, e analiza si se trata de un problema de cobertura y para abordar este estudio, recopila los niveles de potencia recibidos (en mW) por los terminales móviles servidos por esta estación cada vez que inician una llamada. Con el objeto de analizar la posible relación entre ambas variables, en el informe correspondiente se recopilan sus datos medios diarios para todas las llamadas registradas en dicha estación durante 82 días consecutivos. De ese informe se ha extraído la siguiente información:

Gráfico de Reasignaciones frente a Pot_rec




Resumen Estadístico

| | Pot_rec | Reasignaciones |
|---|---|---|
| Frecuencia | 82 | 82 |
| Media | 4.65893 | 2.63361 |
| Mediana | 4.332 | 2.67762 |
| Desv. típica | 1.43172 | 0.450234 |
| Mínimo | 1.072 | 1.30222 |
| Máximo | 9.592 | 3.62142 |
| Rango | 8.52 | 2.3192 |
| Primer cuartil | 3.864 | 2.46312 |
| Tercer cuartil | 5.032 | 2.96102 |
| Rango interc. | 1.168 | 0.4979 |

Coef. correl. lineal rxy = -0.7229

A partir de estos datos, se pide:

**a)** Describir la naturaleza de la relación entre las variables aleatorias estudiadas.

**b)** Calcular la matriz de varianzas y covarianzas entre ambas variables.

**c)** Plantea la ecuación del modelo de regresión correspondiente entre ambas variables.

a) La relación entre ambas variables es moderadamente lineal y negativa ($r_{xy}$ = -0.7229), es decir conforme aumenta la potencia disminuye el número de reasignaciones. Es simplemente moderada puesto que la 'nube de puntos' tiene forma de recta, pero los puntos no están muy agrupados en torno a ella yademásel coeficiente de correlación lineal no es demasiado cercano a -1.

b) La matriz de varianzas y covarianzas puede calcularse a partir de la expresión de la covarianza del formulario:

$$s_{xy} = r_{xy} * s_x * s_y$$

De los datos de la salida de Statgraphics:

Covarianzas

| | Pot_rec | Reasignaciones |
|---|---|---|
| Pot_rec | $s_x^2$= 2.04982 | $s_{xy}$ ($Cov_{xy}$) = -0.466009 |
| Reasignaciones | $s_{xy}$ ($Cov_{xy}$) = -0.466009 | $s_y^2$ = 0.202711 |

c) La ecuación del modelo de regresión sería la dada por la siguiente expresión:

y = a + b.x ==> Reasignaciones = 3.693 – 0.227*Pot_recibida

con b = $s_{xy}/s_x^2$ = -0.7229*1.431*0.45 / 1.431^2 = -0.227

y a = $m_y$ – b.$m_x$ = 2.634 – 0.227*4.659 = 3.693

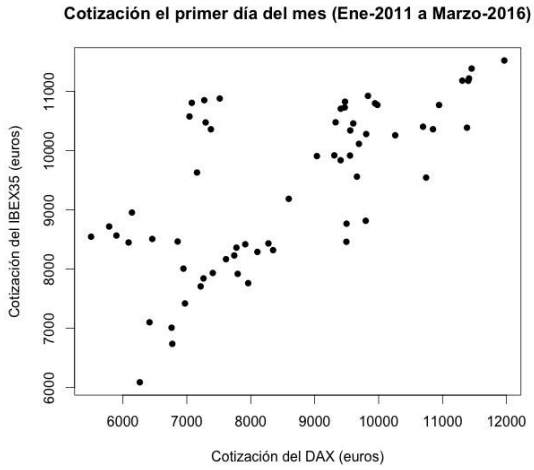
Análisis de Regresión - Modelo Lineal Y = a + b*X

Variable dependiente: Reasignaciones
Variable independiente: Pot_rec

| Parámetro | Estimación | Error estándar | Estadístico T | P-Valor |
|---|---|---|---|---|
| Ordenada | 3.69278 | 0.118335 | 31.206 | 0.0000 |
| Pendiente | -0.227341 | 0.0242919 | -9.35872 | 0.0000 |

**Problema 8**

Un estudiante obtuvo datos sobre dos índices bursátiles: el IBEX35 y el DAX alemán. Se observó la cotización (en euros) de estos índices el primer día de cada mes (desde Enero de 2011, hasta Marzo de 2016), obteniendo un total de 63 casos. Utilizando estos datos se obtuvo el siguiente gráfico:




a) Explica qué ves en el gráfico. Evita utilizar únicamente un lenguaje estadístico y expresa tus conclusiones en el contexto del problema.

*Existe una relación lineal positiva entre la cotización mensual del DAX y del IBEX35: meses en los que la cotización del DAX es elevada están asociados a meses en los que la cotización del IBEX35 es también elevada, y viceversa.*

*Esta relación es débil, ya que la dispersión de los puntos respecto a una hipotética recta es elevada.*

*Se observan dos grupos de puntos. Para cotizaciones del DAX entre los 6000 y 8000 euros, existen dos grupos de cotizaciones para el IBEX35: unos meses en los que este índice ha cotizado entre 6000 y 8000 euros y otro grupo de meses en los que el IBEX35 tuvo cotizaciones mucho más elevadas (entre 8000 y 11000 euros). Este último grupo de meses se aparta del patrón que define el resto de los puntos.*

b) El coeficiente de correlación entre estos índices es *r=0,68*. Explica qué es el coeficiente de correlación y explica el resultado obtenido en el contexto del problema.

*El coeficiente de correlación mide la dirección y la fuerza de la relación lineal que presentan dos variables numéricas medidas en las mismas unidades de observación.*

*Toma siempre valores entre -1 y +1. Valores de r cercanos a 0 indican una relación lineal muy débil (o inexistente). La fuerza de la relación lineal aumenta a medida que r se aleja del 0.*

*En este análisis el coeficiente de correlación vale 0,68, por lo tanto, la relación lineal entre la cotización mensual de los dos índices estudiados (IBEX35 y DAX) es positiva pero débil.*

c) Si obtenemos un modelo de regresión lineal con estos datos para predecir la cotización mensual del DAX en función de la cotización mensual del IBEX35, obtenemos los siguientes resultados:

Constante = 4808,31 Pendiente = 0,536 $R^2$= ????

Calcula el valor de $R^2$ y explica, en el contexto del problema, el significado de estos tres estadísticos (no olvides indicar las unidades en que se expresan los mismos).

*Constante = 4808,31 (euros). Valor que predice el modelo para la cotización (a principios de mes) del IBEX35, para un mes en el que la cotización del DAX es de 0 euros. Como un valor de 0 euros para la cotización del DAX no tiene sentido, esta interpretación carece de interés. El valor obtenido es necesario para realizar el ajuste, pero carece de significado práctico.*

*Pendiente = 0,536 (euros/euros). Por cada euro que se incrementa la cotización del DAX, la cotización prevista del IBEX35 aumenta en 0,536 euros.*

*$R^2$ es el coeficiente de determinación: $R^2 = r^2 = 0{,}68^2 = 0.46 = 46\%$.*

*Este coeficiente mide la bondad del ajuste del modelo de regresión a los datos. En nuestro estudio, un 46% de la variación total de la cotización mensual del IBEX35 está explicada por el modelo de regresión propuesto.*

d) A principios de marzo de 2016, la cotización del IBEX35 fue de 8766,9 euros, y la del DAX de 9498,15 euros. Explica qué son los residuos de un modelo de regresión y calcula el residuo asociado al mes de marzo de 2016.

*Los residuos de un modelo de regresión son las distancias verticales de los valores observados para la variable respuesta a la recta de regresión.*

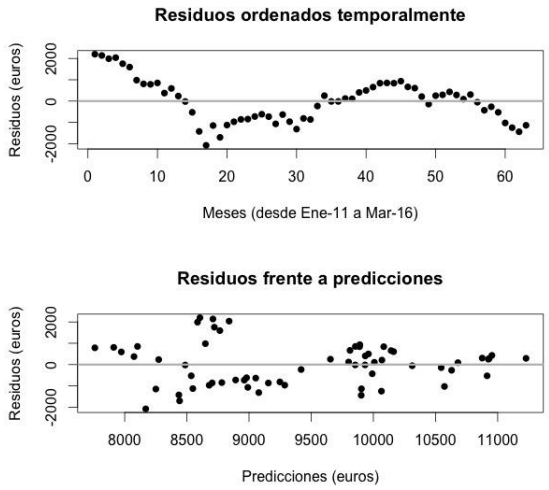
*Para el mes de marzo la cotización que predice el modelo para el IBEX35 es:*

*IBEX35 = 4808,31 + 0,536 x 9498,15 = 9899,32 euros*

*Y el residuo de este mes es:*

*$e_i$ = 8766,9 – 9899,32 = -1132.42 euros*

e) Explica qué información puede extraerse de los siguientes gráficos.




*Los gráficos de residuos se utilizan para validar el modelo de regresión, viendo si se cumplen los supuestos asumidos al hacer el ajuste.*

*El primer gráfico es un diagrama de dispersión de los residuos ordenados en el tiempo. Se aprecia claramente que los residuos están correlacionados. Vemos que presentan una forma de ola: positivos, negativos, positivos y negativos.*

*Si los residuos están correlacionados se viola uno de los supuestos asumidos al hacer el ajuste, por lo que el modelo de regresión obtenido no es válido.*

*El segundo gráfico es un diagrama de dispersión de los residuos en función de los valores que predice el modelo.*

*Se aprecia que los residuos se posicionan formando grupos, y que la dispersión de los residuos no es constante: existe mayor dispersión para predicciones con valores bajos que para valores elevados. El modelo propuesto presenta, por tanto, un problema de heterocedasticidad.*

*Conclusión: el modelo de regresión propuesto no es válido ya que no se cumplen varios de los supuestos asumidos al hacer el ajuste.*

**Problema 9**

Los estudiantes matriculados en un curso de estadística respondieron estas preguntas:

1. ¿Te gusta leer novelas?
    - NO, no me gustan las novelas
    - Sí, pero no tengo tiempo
    - Sí, y soy un lector muy activo

2. Respecto a la religión, te consideras una persona:
    - Agnóstica (no podemos demostrar que dios exista)
    - Atea (no crees que dios exista)
    - Religiosa (crees en dios)

Utilizando sus respuestas, obtuvimos la siguiente tabla de contingencia:

| | No me gustan | No tengo tiempo | Sí y leo muchas |
|---|---|---|---|
| Agnósticos | 10 | 60 | 20 |
| Ateos | 30 | 30 | 40 |
| Religiosos | 40 | 80 | 30 |

Obtén la tabla de frecuencias condicionales que consideres relevante para determinar si estas variables están relacionadas o, por el contrario, son independientes. Escribe un breve párrafo con tus conclusiones, aportando los valores numéricos que creas necesarios.

*Tabla de frecuencias condicionales de Lectura/Religión:*

| | *No me gustan* | *No tengo tiempo* | *Sí y leo muchas* |
|---|---|---|---|
| *Agnósticos* | *0,11* | *0,67* | *0,22* |
| *Ateos* | *0,30* | *0,30* | *0,40* |
| *Religiosos* | *0,27* | *0,53* | *0,20* |

*Observando esta tabla vemos, de forma muy clara, que estas dos variables están relacionadas: un 40% de los estudiantes que se consideran ateos dicen que les gusta leer novelas y que, de hecho, leen bastantes, frente a tan sólo un 20% de los estudiantes agnósticos o a un 22% de los estudiantes religiosos.*

*Conocer el sentimiento religioso de un estudiante aporta información útil para saber si a éste le gustará, o no, leer novelas. Las variables no son independientes: la distribución condicional de una variable depende de los valores considerados para la otra.*

**Problema 10**

Algunos inversores creen que el comportamiento de la bolsa durante el mes de enero es un buen indicador del funcionamiento que tendrá la misma durante todo el año: los cambios que experimenta un índice bursátil en enero están asociados a los cambios de este índice durante todo el año.

Queremos realizar un análisis para comprobar si esta creencia es cierta, utilizando el IBEX35 como referencia. Recogeremos información sobre la variación de este índice desde el año 1970 hasta el año 2015.

El principal objetivo del análisis es obtener un modelo de regresión.

**a)** Identifica y define cuál es la variable respuesta y cuál es la variable explicativa de este análisis. ¿Cuántos casos hay en este estudio? ¿Qué son exactamente los casos de este estudio?

- *Variable respuesta: Y = Variación que ha experimentado el IBEX35 durante todo el año (en %)*
- *Variable explicativa: X = Variación que ha experimentado el IBEX35 durante el mes de enero (en %)*
- *Los casos de este estudio son los años considerados: desde 1970 hasta 2015. Existen un total de 46 casos (años).*

**b)** Explica los pasos que seguirías para obtener un modelo de regresión: qué gráficos y qué estadísticos obtendrías, y en qué orden.

- *El primer paso es obtener un **diagrama de dispersión** de los datos. Este gráfico muestra el tipo de relación que presentan la variable respuesta y la explicativa: ¿es una relación lineal?, ¿existen valores anómalos o valores influyentes?*
- *Si las variables presentan una relación lineal, el **coeficiente de correlación** nos ayuda a medir la fuerza de esa relación.*
- *Por último, si la relación es lineal, sería interesante tratar de obtener un **modelo de regresión**:*

$$y = a + b\,x$$

**c)** Imagina que una vez obtenido el modelo de regresión, obtenemos un valor para el coeficiente de determinación $R^2$ = 89%. ¿Garantiza este resultado que el modelo de regresión obtenido es un buen modelo para predecir la variación anual del IBEX35 en función de su variación durante el mes de enero? Justifica tu respuesta.

- *No. El coeficiente de determinación indica el grado de bondad de ajuste de la recta de regresión a los datos, pero primero debemos comprobar si el modelo obtenido es válido, es decir, necesitamos comprobar si se cumplen las condiciones asumidas por el modelo al realizar el ajuste.*
- *Para comprobar si se cumplen estas condiciones (residuos no autocorrelacionados, distribución Normal de los residuos, homoscedasticidad, etc.) utilizaremos gráficos de residuos.*

**d)** Explica por qué el método de ajuste utilizado para obtener la recta de regresión, se denomina *método de ajuste por mínimos cuadrados*.

- *El método de ajuste se denomina así, porque la recta obtenida es aquella que hace mínima la suma del cuadrado de las distancias verticales de los puntos a la recta de regresión.*
- *Es, por tanto, la recta que minimiza la suma de los residuos al cuadrado.*

## Problema 11

En la siguiente tabla se muestra la salida de la regresión lineal simple realizada entre el precio de los carritos de bebé (en euros) y su altura plegado (en cm)

Variable dependiente: PRECIO

Variable independiente: ALTO PLEGADO

Lineal: Y = a + b*X

**Coeficientes**

| | *Mínimos Cuadrados* | *Estándar* | *Estadístico* | |
|---|---|---|---|---|
| *Parámetro* | *Estimado* | *Error* | *T* | *Valor-P* |
| Intercepto | 969,666 | 241,99 | 4,00705 | 0,0002 |
| Pendiente | -3,54721 | 2,88418 | -1,22988 | 0,2244 |

**Análisis de Varianza**

| *Fuente* | *Suma de Cuadrados* | *Gl* | *Cuadrado Medio* | *Razón-F* | *Valor-P* |
|---|---|---|---|---|---|
| Modelo | 308172, | 1 | 308172, | 1,51 | 0,2244 |
| Residuo | 1,03905E7 | 51 | 203734, | | |
| Total (Corr.) | 1,06986E7 | 52 | | | |

Coeficiente de Correlación = -0,16972
R-cuadrada = ¿? porciento
R-cuadrado (ajustado para g.l.) = 0,976172 porciento
Error estándar del est. = 451,37

Teniendo en cuenta las herramientas estadísticas anteriores, se pide:

a) Identifica la variable dependiente e independiente y analiza si existe algún tipo de relación entre ellas. Utiliza las herramientas estadísticas apropiadas para justificar tu respuesta. ***(0.5 punto)***

b) ¿Es el efecto de la variable independiente o explicativa estadísticamente significativa? Estima el valor del $R^2$. Justifica tu respuesta. ***(1 punto)***

R-cuadrada = 2,88048 porciento

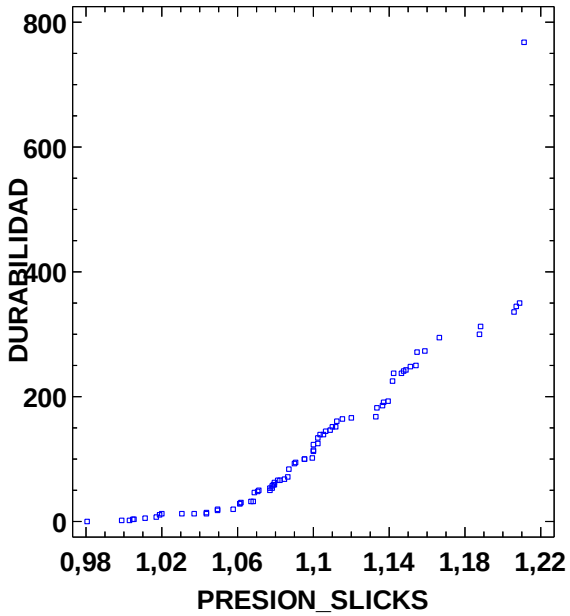
c) Propón una estimación puntual de los parámetros del modelo y calcula, si es posible, ¿qué el precio esperado de un carrito con una altura de plegado de 50? Justifica razonadamente tu respuesta. ***(0.5 punto)***

**Valores Predichos**

| | | *95,00%* | | *95,00%* | |
|---|---|---|---|---|---|
| | *Predicciones* | *Límite* | *Predicción* | *Límite* | *Confianza* |
| *X* | *Y* | *Inferior* | *Superior* | *Inferior* | *Superior* |
| 50,0 | 792,306 | -139,927 | 1724,54 | 573,389 | 1011,22 |

## Problema 12

A un ingeniero le piden obtener un modelo para calcular la relación entre la durabilidad de los neumáticos (Km) y la presión de los mismos (bares) de los coches de carreras. Para ello realiza un estudio de regresión lineal simple de la durabilidad en función de la presión de los mismos en una muestra de 76 vehículos y obtuvo el siguiente diagrama de dispersión:



a) Explica qué ves en el gráfico. Evita utilizar únicamente un lenguaje estadístico y expresa tus conclusiones en el contexto del problema. **(0,5 puntos)**

b) El coeficiente de correlación entre estos índices es *r=0,89*. Explica qué es el coeficiente de correlación y explica el resultado obtenido en el contexto del problema. **(0,5 puntos)**

c) El modelo de regresión obtenido se muestra a continuación:

**Coefficients**

| | *Least Squares* | *Standard* | *T* | |
|---|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Statistic* | *P-Value* |
| Intercept | -2138,25 | 134,153 | -15,9389 | 0,0000 |
| Slope | 2067,08 | 122,344 | 16,8955 | 0,0000 |

**Analysis of Variance**

| *Source* | *Sum of Squares* | *Df* | *Mean Square* | *F-Ratio* | *P-Value* |
|---|---|---|---|---|---|
| Model | 903021, | 1 | 903021, | 285,46 | 0,0000 |
| Residual | 234091, | 74 | 3163,4 | | |
| Total (Corr.) | 1,13711E6 | 75 | | | |

Correlation Coefficient = 0,891143
R-squared = ¿???
Standard Error of Est. = 56,2441

Calcula los parámetros a y b de la recta y el valor de $R^2$ y explica, en el contexto del problema, el significado de estos tres estadísticos (no olvides indicar las unidades en que se expresan los mismos). Efectúa una estimación de la durabilidad media de los neumáticos para una presión de 2 bares. **(0,75 puntos)**

b) Explica por qué el método utilizado para ajustar un modelo de regresión a un conjunto de datos cuantitativos es conocido como método de ajuste de mínimos cuadrados. **(0,25 puntos)**

c) ¿Qué son los residuos de un modelo de regresión? ¿Cuál sería el valor de la varianza residual? Explica cómo podemos utilizarlos para validar un modelo de regresión. **(0,5 puntos)**

*Solución:*

*a) Existe una relación lineal creciente entre la durabilidad de los neumáticos y la presión de los mismos. Esta relación es bastante alta, debido a la compactos que están los puntos respecto a una hipotética recta imaginaria. Se aprecia también un dato aislado o anómalo*

*b) El coeficiente de correlación mide la dirección y la fuerza de la relación lineal que presentan dos variables numéricas medidas en las mismas unidades de observación. Toma siempre valores entre -1 y +1. Valores de r cercanos a 0 indican una relación lineal muy débil (o inexistente). La fuerza de la relación lineal aumenta a medida que r se aleja del 0.*

*En este análisis el coeficiente de correlación vale 0,89, por lo tanto, la relación lineal es positiva y relativamente fuerte.*

*c)* Constante = -2138. *Valor que predice el modelo para los kms que durará el coche cuando la presión es de cero bares. Como un valor negativo para los Kms no tiene sentido, esta interpretación carece de interés. El valor obtenido es necesario para realizar el ajuste, pero carece de significado práctico.*

*Pendiente = 2067,08 (euros/Megabytes). Es la pendiente de la recta, es decir, por cada bar que se incrementa la presión, los neumáticos incrementan su duración en 2067,08 Kms.*

*$R^2$ es el coeficiente de determinación: $R^2 = r^2 = ,89^2 = 79,14\%$.*
*Este coeficiente mide la bondad del ajuste del modelo de regresión a los datos. En nuestro estudio, un 79,14% de la variación total de los kms que duran los neumáticos está explicada por la presión de los mismos.*

DURABILIDAD = -2138,25 + 2067,08* (2) = 1995,91

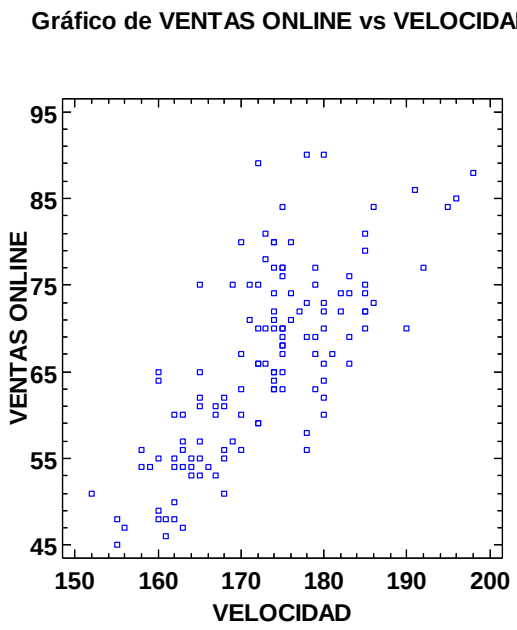
*d)* *De todas las posibles rectas que pueden trazarse para aproximar la nube de puntos del diagrama de dispersión, el método de mínimos cuadrados proporciona aquella recta que **minimiza** la suma del **cuadrado de las distancias verticales** de los puntos a la recta de regresión. De ahí el nombre de ajuste por mínimos cuadrados. (Nota: es necesario elevar estas distancias al cuadrado para evitar que, al realizar la suma, se compensen las distancias positivas con las negativas).*

*e)* *Los residuos de un modelo de regresión son las distancias verticales de los puntos a la recta de regresión y recogen el efecto conjunto de todos los factores que afectan a la variable respuesta y que el modelo de regresión no contempla. Sirven para validar el modelo de regresión ya para ello disponemos de los gráficos de los residuos, que nos permiten verificar las hipótesis de independencia, homocedasticidad y normalidad, requisito para que el ajuste de regresión sea válido.*

$S^2resid = (56,2441)^2 = 3163,398785$

**Problema 13**

Una empresa de zapatillas de deporte que realiza parte de sus ventas a través de su propia página web, desea efectuar un estudio para saber si en función de la velocidad de navegación (MegaBytes) por su catálogo de productos online, se incrementan el número de ventas (euros). Para ello estudió si existía relación entre ambas variables mediante un diagrama d dispersión:




d) Explica qué ves en el gráfico. Evita utilizar únicamente un lenguaje estadístico y expresa tus conclusiones en el contexto del problema. **(0,5 puntos)**

e) El coeficiente de correlación entre estos índices es *r=0,74*. Explica qué es el coeficiente de correlación y explica el resultado obtenido en el contexto del problema. **(0,5 puntos)**

f) Se obtuvo un modelo de regresión lineal con estos datos para predecir el volumen de ventas on-line en función de la velocidad de navegación por el catálogo de productos y se obtuvieron los siguientes resultados:

Constante = -84,07 Pendiente = 0,869 $R^2$= ????

Calcula el valor de $R^2$ y explica, en el contexto del problema, el significado de estos tres estadísticos (no olvides indicar las unidades en que se expresan los mismos). **(0,5 puntos)**

g) Explica por qué el método utilizado para ajustar un modelo de regresión a un conjunto de datos cuantitativos es conocido como método de ajuste de mínimos cuadrados. **(0,5 puntos)**

h) ¿Qué son los residuos de un modelo de regresión? Explica cómo podemos utilizarlos para validar un modelo de regresión. **(0,5 puntos)**

*Solución:*

*f) Existe una relación lineal creciente entre el volumen de ventas y la velocidad de navegación. Esta relación es débil, debido a la dispersión de los puntos respecto a una hipotética recta es elevada.*

*g) El coeficiente de correlación mide la dirección y la fuerza de la relación lineal que presentan dos variables numéricas medidas en las mismas unidades de observación.*

*Toma siempre valores entre -1 y +1. Valores de r cercanos a 0 indican una relación lineal muy débil (o inexistente). La fuerza de la relación lineal aumenta a medida que r se aleja del 0.*

*En este análisis el coeficiente de correlación vale 0,74, por lo tanto, la relación lineal entre es positiva pero no muy fuerte.*

*h)* Constante = -84,07. *Valor que predice el modelo para el volumen de ventas cuando la velocidad de navegación es de 0 megabytes. Como un valor negativo para la las ventas no tiene sentido, esta interpretación carece de interés. El valor obtenido es necesario para realizar el ajuste, pero carece de significado práctico.*

*Pendiente = 0,869 (euros/Megabytes). Es la pendiente de la recta, es decir, por cada megabyte que se incrementa la velocidad, las ventas on-line se incrementan 0,869 euros.*

*$R^2$ es el coeficiente de determinación: $R^2 = r^2 = ,74^2 = 0.548 = 54,8\%$.*

*Este coeficiente mide la bondad del ajuste del modelo de regresión a los datos. En nuestro estudio, un 54,8% de la variación total las ventas on-line está explicada por la velocidad de navegación por el catálogo.*

*i) De todas las posibles rectas que pueden trazarse para aproximar la nube de puntos del diagrama de dispersión, el método de mínimos cuadrados proporciona aquella recta que **minimiza** la suma del **cuadrado de las distancias verticales** de los puntos a la recta de regresión. De ahí el nombre de ajuste por mínimos cuadrados. (Nota: es necesario elevar estas distancias al cuadrado para evitar que, al realizar la suma, se compensen las distancias positivas con las negativas).*

*j) Los residuos de un modelo de regresión son las distancias verticales de los puntos a la recta de regresión y recogen el efecto conjunto de todos los factores que afectan a la variable respuesta y que el modelo de regresión no contempla. Sirven para validar el modelo de regresión ya para ello disponemos de los gráficos de los residuos, que nos permiten verificar las hipótesis de independencia, homocedasticidad y normalidad, requisito para que el ajuste de regresión sea válido.*

**Problema 13**

Un estudiante de administración de empresas recibe una beca para el departamento de administración de una de las mayores productoras de Hollywood. Después de varios fracasos en taquilla seguidos, la dirección decide recortar el presupuesto de las próximas películas. Para demostrar que la dirección no ha tomado la decisión correcta el estudiante plantea un de regresión con el objetivo de predecir la taquilla de una película (en millones de dólares) a partir del presupuesto destinado a la misma (en millones de dólares), para ello ha seleccionado **20 éxitos de la productora**, y ha obtenido los siguientes resultados del análisis:

Sx=159,29

Sy=308.54

cov_xy = 43.249,50

$m_x$= 436,70

$m_y$= 806,54

**a)** Calcula la ecuación del modelo de regresión obtenido **(0.75 puntos).**

*La ecuación del modelo será: y = a + bx, siendo:*

*$b = s_{xy} / s_x^2 = 43.249,50/ 25.372,20 = 1,705$*

*$a = m_y – b.m_x = 806,54- 1,705×436,70= 61,97$*

*La ecuación del modelo es por tanto:*

*Taquilla (Millones de dólares) = 61,97+1,705× Presupuesto (Millones de dólares)*

*En función del modelo planteado, a mayor presupuesto la recaudación en taquilla es superior.*

**b)** ¿Qué porcentaje de la variabilidad de *la taquilla obtenida* viene explicada por la cuantía *presupuesto*? Indica el parámetro que cuantifica dicho porcentaje. **(0.5 puntos).**

*Este porcentaje de variabilidad viene expresado por el valor del Coeficiente de Determinación, que se calcula como:*

$R = r_{xy}^2 \times 100 = (s_{xy} / (s_x s_y))^2 \times 100 = (43.249{,}50/159{,}29 \times 308.54)^2 \times 100 = 77{,}44\%$

**c)** Si una nueva película recibe un presupuesto de 550 Millones de dólares, ¿cuánta taquilla, en promedio, predice el modelo del estudiante que obtendrá? **(0.5 puntos).**

*La taquilla media para para una película de 550 Millones de dólares se puede calcular directamente a partir de la ecuación del modelo:*

*Taquilla (Millones de dólares) = 61,97+1,705× 550 (Millones de dólares) = 999,72 Millones de dólares*

**d)** ¿Entre que valores aproximadamente estará *la taquilla*, en promedio, en el 95% de las películas que tengan 550 Millones de dólares? (**0.75 puntos).**

*La variabilidad residual se puede calcular como:*

$s_{res}^2 = s_y^2 (1 - r_{xy}^2) = 21.476{,}43$

$s_{res} = 146{,}55$

*El intervalo de valores es por tanto, asumiendo una distribución normal para la variabilidad residual:*

*[taquilla (550 Millones de dólares de presupuesto) - 2 $s_{res}$; $t_{entrega}$ (550 Millones de dólares de presupuesto) + 2 $s_{res}$] =*

$[999{,}72 - 2 \times 146{,}55; 999{,}72 + 2 \times 146{,}72] = [706{,}64; 1.292{,}8]$

*Es decir prácticamente entre 706 y 1.292 días como máximo.*

**e)** ¿Qué análisis debería plantear el estudiante para validar el modelo? Cita al menos dos hipótesis que todo modelo de regresión debe cumplir para ser valido (**0.75 puntos).**

*El estudiante tendría que realizar un análisis de los residuos para validar el modelo. Con los gráficos de los residuos debemos comprobar que los residuos son variables aleatorias*

*de media nula, que tienen varianza contante, que no están correlacionados entre sí, que son independientes y que no dependen de otras variables explicativas.*

**f)** Después de los cálculos realizados ¿Consideras que el modelo planteado por el estudiante es correcto y qué la dirección se equivoca al reducir los presupuestos? (**0.75 puntos).**

*El análisis estadístico realizado por el estudiante viene mal planteado desde el principio, puesto que la muestra únicamente incluye los éxitos de la productora, la muestra está sesgada. El ajuste no sería tan alto si hubiese incluido los fracasos de la productora, que es el principal motivo por lo que la dirección de la productora quiere reducir los presupuestos.*

*No se puede considerar que la dirección se equivoque, ni lo contrario, lo que es seguro es que gran* parte del éxito de una película no está explicada por el presupuesto.