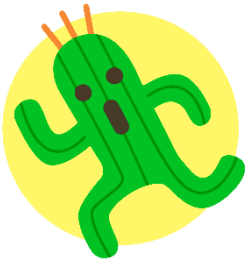
Estadística sin espinas

(7) Regresión lineal



v0.1 24_06_05

Aprende sin espinas
con @carlos_cactus

Sócrates se equivocaba. El conocimiento no es lo único que crece al compartirse: La alegría también.

A la inspiración del bucle_infinito,
al Cibergrupo y al tHash_A, por su amistad,
y sobre todo, a quienes dicen "pero quiero"
cuando sienten "no puedo".

¡Un saludo sin espinas!
@carlos_cactus :D

Y si quieres saber más:
¡Encuéntrame en Telegram como @carlos_cactus o habla con Espinito, el bot Sin Espinas, en @GestionSinEspinasBot.

Únete a la comunidad de Telegram Sin Espinas y no te pierdas nada!

Deja de preocuparte por aprobar y ¡Aprende sin Espinas!

VII. REGRESIÓN LINEAL SIMPLE

Los modelos de regresión fueron introducidos en Astronomía por Laplace y Gauss en el s. XVIII pero toman su nombre del trabajo del biólogo Galton en el s. XIX.

8.1. Concepto de correlación

Las técnicas de correlación permiten evidenciar y calcular relaciones de dependencia entre variables cuantitativas, mediante la construcción de modelos.

Estos modelos permiten predecir el valor que adopta una variable a partir de los valores que adopta otra.

En este apartado se trata el modelo de REGRESIÓN LINEAL SIMPLE.

8.2. Relaciones entre variables

Dos variables pueden estar o no relacionadas. En caso de que SÍ EXISTA RELACIÓN entre ellas, según si existe una expresión matemática que relacione sus valores, se distinguen:

- Relaciones FUNCIONALES o DETERMINISTAS. Sí existe la expresión.
- Relaciones ESTADÍSTICAS o ESTOCÁSTICAS. No existe la expresión.

8.3. Diagrama de dispersión

La representación gráfica más habitual a partir de la cual trazar la correlación es el diagrama de dispersión o de nube de puntos para 2 variables.

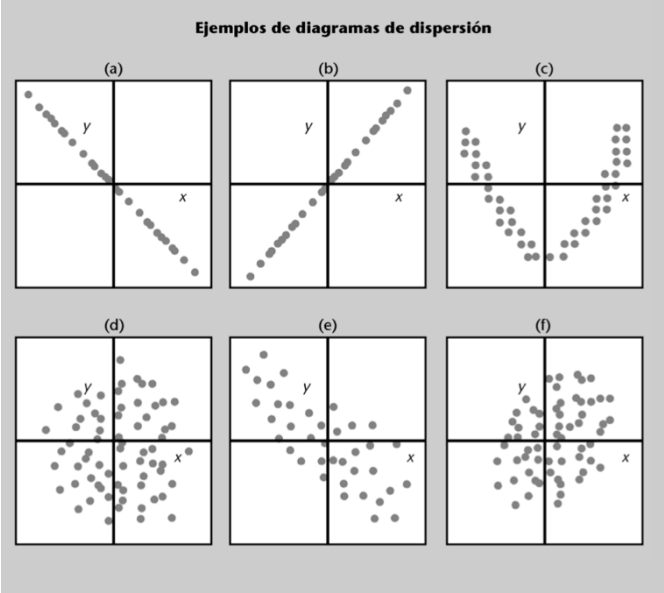
El diagrama de dispersión se obtiene representando en el plano cartesiano los pares (x,y) de forma que:

→ X es la variable EXPLICATIVA (INDEPENDIENTE o EXÓGENA)
→ Y es la variable EXPLICADA o RESPUESTA (DEPENDIENTE o ENDÓGENA).

Para la cuantificación de la dependencia entre ambas variables se dispone de coeficientes de correlación.

En esta materia, solo se usa la correlación LINEAL mediante el coeficiente de correlación LINEAL de Pearson.

Existen multitud de técnicas de correlación lineales y no lineales, que permiten aproximar las nubes de puntos a funciones exponenciales, potenciales, logarítmicas…



Los diagramas a) y b) manifiestan relaciones DETERMINISTAS:

- La relación a) se corresponden con una recta de pendiente positiva.
- La relación b) se corresponden con una recta de pendiente negativa.

El diagrama c) muestra una relación ESTOCÁSTICA, ya que no hay una expresión que refleje EXACTAMENTE la curva sobre la cual se disponen los puntos, pero sí existe una FUERTE correlación entre ambas variables.

El diagrama d) no manifiesta NINGÚN TIPO DE RELACIÓN entre los valores de las variables, NI determinista, NI estocástica. Se dice que no muestra ninguna FORMA TUBULAR definida.

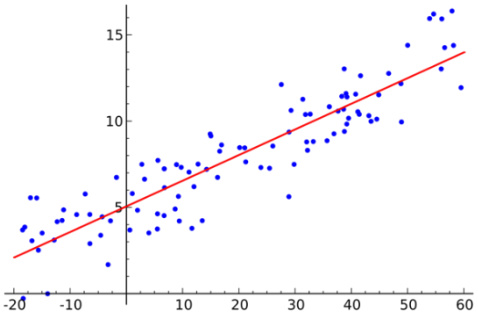
Los diagramas d), e) y f) muestran relaciones ESTOCÁSTICAS, no hay expresión matemática que relacione los valores de ambas variables de forma EXACTA, pero sí se manifiesta una CORRELACIÓN SUAVE, no tan intensa como en a) o b).

8.4. Recta de regresión

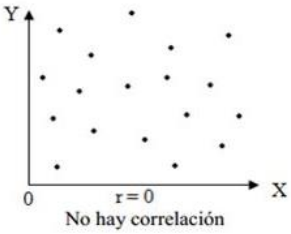
El diagrama de dispersión permite evidenciar una tendencia coordinada entre los valores de ambas variables, ya sea de proporcionalidad directa o inversa, que se puede modelizar mediante una recta, como se muestra a continuación.

Esa recta es la recta de regresión, que aspira a aproximarse lo máximo posible a todos y cada uno de los puntos del diagrama.

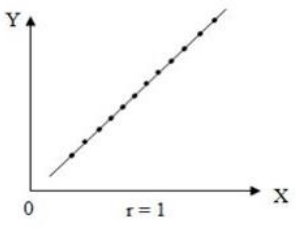
La monotonía de la recta de regresión depende del valor del coeficiente de correlación r

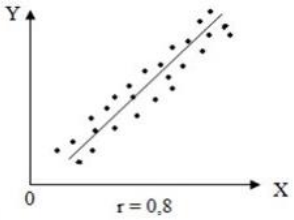


$$\begin{cases} r \to 1 & RECTA\ CRECIENTE \\ r \to 0 & variables\ INCORRELADAS \\ r \to -1 & RECTA\ DECRECIENTE \end{cases}$$

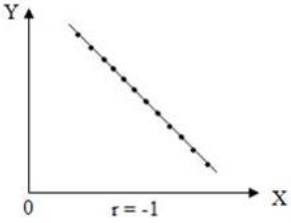


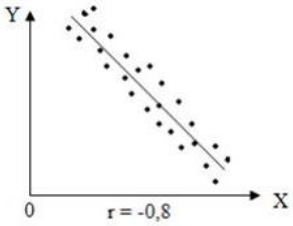
No hay correlación





Correlación Positiva





Correlación Negativa

Se dice que el coeficiente de correlación r es una MEDIDA DE LA BONDAD DEL AJUSTE que realiza sobre los puntos de la nube la recta de regresión.

8.5. Regresión lineal SIMPLE

La regresión lineal permite parametrizar matemáticamente la relación entre una variable X (explicativa) y otra variable Y (explicada) de forma que se puedan PREDECIR propiedades de la POBLACIÓN a partir de conocer propiedades de la MUESTRA.

Habitualmente, se usará para estimar valores de Y a partir de valores de X NO DISPONIBLES EN LA MUESTRA, aunque se puede aplicar el mismo método para explicar X a partir de los valores de Y.

Se dice regresión SIMPLE porque implica solamente 2 variables:

- X independiente o EXPLICATIVA
- Y dependiente o de RESPUESTA

Se dice que es LINEAL porque se basa en la estimación de los valores de una variable en función de la otra utilizando una expresión lineal de la forma:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

Lo cual requiere el cálculo de:

- $\hat{\beta}_1$ Pendiente estimada de la recta.
- $\hat{\beta}_0$ Ordenada en el origen estimada.

De modo que a partir de los valores X de entrada, se obtienen valores APROXIMADOS de Y.

Nótese que esta ESTIMACIÓN APROXIMADA de la variable Y mediante el modelo de regresión se denota como Y estimada $\hat{y}$:

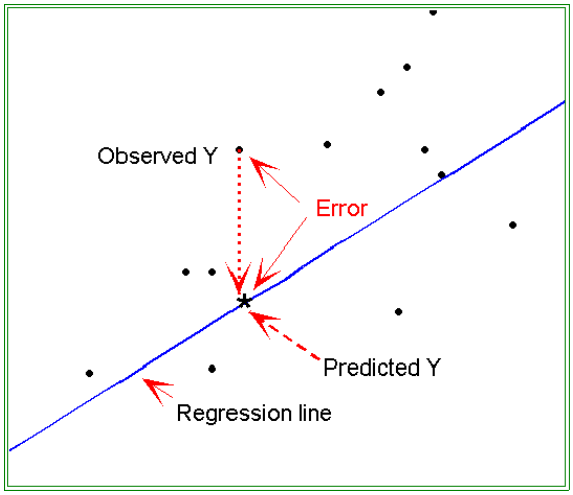
$$valor\ y\ real\ desconocido \approx valor\ \hat{y}\ ESTIMADO\ mediante\ la\ recta$$

8.6. Residuo

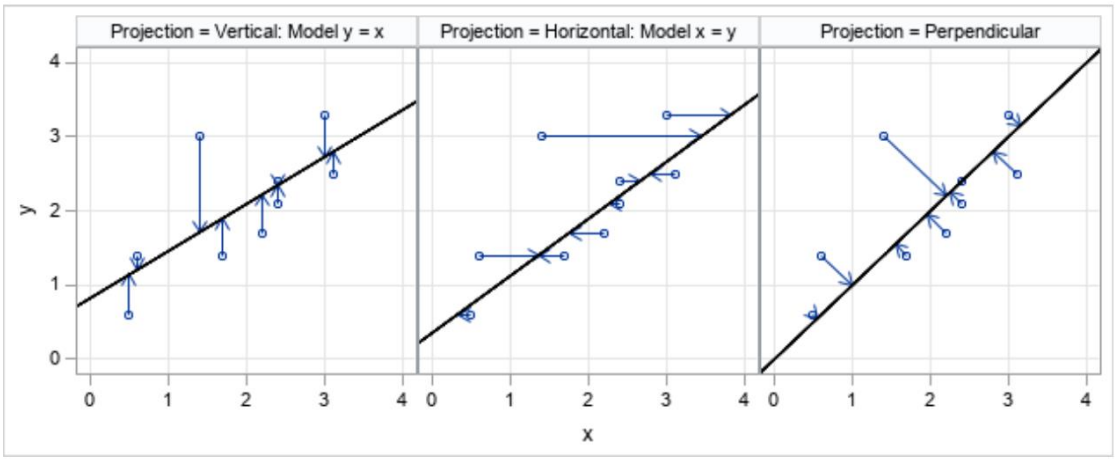
El ERROR que se comete en la regresión se cuantifica mediante el RESIDUO e, que se define como:

$$e = y - \hat{y}$$

Para cada observación (x_i, y_i) en el diagrama de dispersión se define el RESIDUO como la distancia VERTICAL entre el punto (x_i, y_i) y el trazado de la recta, es decir:



Otras aproximaciones consideran la distancia horizontal o bien la perpendicular a la recta:



El residuo e se calcula:

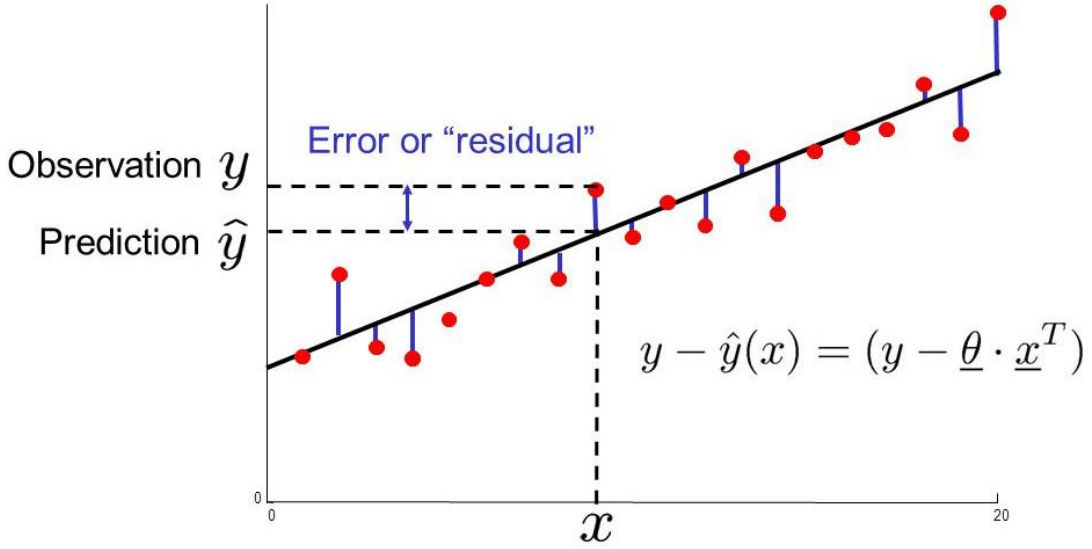
$$e = y - \hat{y} = y_i - (\beta_1 x_i + \beta_0)$$

Donde:

y_i	Es el valor OBSERVADO CIERTO (ordenada en el diagrama de dispersión).
$\hat{y}$	Es el valor ESTIMADO que devuelve la RECTA de regresión.
x_i	Es el valor de ABSCISA del punto del DIAGRAMA asociado a y_i.

Por tanto, el residuo de una observación (x,y) es la diferencia entre el valor y OBSERVADO y el valor $\hat{y}$ ESTIMADO PREDICHO por la recta.

El valor x de la abscisa coincide en la observación y en la recta, ya que se toma solamente la distancia VERTICAL (diferencia $y - \hat{y}$).



Una correlación BIEN ajustada es aquella que permite MINIMIZAR el error, es decir, el RESIDUO.

8.7. Método de los mínimos cuadrados

Es un modelo para la construcción de la recta de regresión que persigue minimizar la suma del cuadrado de los errores, es decir, minimizar el cuadrado de los residuos. Se expresa como:

$$SCR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

8.8. Cálculo de la recta de regresión

Mediante el método de los mínimos cuadrados, se define la expresión de la recta de regresión como:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

Donde:

$\hat{y}$ Valor ESTIMADO de la variable RESPUESTA devuelto por la recta.
$\hat{\beta}_1$ Pendiente de la recta.
$\hat{\beta}_0$ Ordenada en el origen.

→ Hasta ahora, a lo largo de todo este manual, el acento circunflejo ha referido EXCLUSIVAMENTE al carácter CORREGIDO (dividido entre n – 1). En este apartado, además, para $\hat{y}$, $\hat{\beta}$ y $\hat{\beta}_0$ denota el carácter ESTIMADO.

La PENDIENTE $\hat{\beta}_1$, a su vez, se define como:

$$\hat{\beta}_1 = \frac{S_{\text{xy}}}{\hat{s}_x^2}$$

Nótese que se divide entre la cuasivarianza muestral de la variable EXPLICATIVA, con independencia de si se denota por x o bien por y.

Donde:

S_{xy} Es la COVARIANZA MUESTRAL entre ambas variables:

$$S_{\text{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{ij}(x_i \cdot y_j \cdot n_{ij}) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Ambas expresiones son equivalentes.

$\hat{s}_x^2$ Es la CUASIVARIANZA MUESTRAL de la variable EXPLICATIVA.

Y para la ORDENADA $\hat{\beta}_0$ se tiene:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

Donde:

$\bar{y}$ Es la media muestral de la variable RESPUESTA.
$\bar{x}$ Es la media muestral de la variable EXPLICATIVA.
$\hat{\beta}_1$ Es la pendiente de la recta $\hat{\beta}_1 = \frac{S_{\text{xy}}}{\hat{s}_x^2}$

Esta expresión proviene de:

$$\hat{y} - \bar{y} = \frac{S_{xy}}{\hat{s}_x^2} \cdot (x - \bar{x})$$

De lo cual:

$$\hat{y} - \bar{y} = \frac{S_{xy}}{\hat{s}_x^2} x - \frac{S_{xy}}{\hat{s}_x^2} \bar{x}$$

Aislando y:

$$\hat{y} = \underbrace{\frac{S_{xy}}{\hat{s}_x^2}}_{\hat{\beta}_1} x \underbrace{- \frac{S_{xy}}{\hat{s}_x^2} \bar{x} + \bar{y}}_{\hat{\beta}_0}$$

Nótese que tiene sentido escribir la correlación en 2 sentidos:

→ La recta de regresión de Y sobre X, es decir, Y(X) es:

$$\hat{y} - \bar{y} = \frac{\hat{S}_{xy}}{\hat{s}_x^2} \cdot (x - \bar{x})$$

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

En cuyo caso, X EXPLICA el comportamiento de Y.

→ La recta de regresión de X sobre Y, es decir, X(Y) es:

$$x - \bar{x} = \frac{\hat{S}_{xy}}{\hat{s}_y^2} \cdot (y - \bar{y})$$

$$\hat{x} = \hat{\beta}_1 y + \hat{\beta}_0$$

En cuyo caso, Y EXPLICA el comportamiento de X.

O sea, se puede hacer el análisis de la correlación:

- De y sobre x Y(X) en que se pretende explicar la variación de y (explicada) a través de la variación de x (explicativa).
- O de x sobre y X(Y) en que se pretende explicar la variación de x (explicada) a través de la variación de y (explicativa).

¡ATENCIÓN!
Independientemente de la notación (en ocasiones, se deseará explicar una variable denotada como X mediante una variable explicativa denotada como Y y no se debe confundir el carácter EXPLICATIVO o EXPLICADO en el modelo de regresión porque las variables aleatorias tengan nombres como "X" e "Y".

Lo que se debe entender es:

$$\widehat{estimada} = \hat{\beta}_1 \cdot explicativa + \hat{\beta}_0$$

Es decir:

$$\widehat{estimada} = \overbrace{\frac{\hat{S}_{xy}}{\hat{S}^2_{EXPLICATIVA}}}^{\hat{\beta}_1} \cdot explicativa + \overbrace{(\overline{estimada} - \frac{\hat{S}_{xy}}{\hat{S}^2_{EXPLICATIVA}} \cdot \overline{explicativa})}^{\hat{\beta}_0}$$

8.9. Covarianza

La covarianza (ya sea corregida entre n – 1 como CUASIVARIANZA o no) es una medida del grado de variación CONJUNTA de ambas variables respecto sus medias.

Es el parámetro más inmediato para cuantificar si hay relación entre variables.

Se aplica a la determinación de:

- La pendiente de la recta de regresión:

$$\hat{\beta}_1 = \frac{\hat{S}_{xy}}{\hat{s}_x^2}$$

- El coeficiente de correlación lineal de Pearson r:

$$r = \frac{\hat{S}_{xy}}{\hat{s}_x \hat{s}_y}$$

Se calcula como:

$$\hat{S}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

O, alternativamente:

$$\hat{S}_{xy} = \frac{\sum_{ij}(x_i \cdot y_j \cdot n_{ij}) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Se interpreta según:

$$\begin{cases} \hat{S}_{xy} > 0 \rightarrow Relación\ de\ proporcionalidad\ DIRECTA \\ \hat{S}_{xy} = 0 \rightarrow No\ existe\ relación\ LINEAL\ entre\ x\ e\ y \\ \hat{S}_{xy} < 0 \rightarrow Relación\ de\ proporcionalidad\ INVERSA \end{cases}$$

8.10. Inferencia a partir de la pendiente $\hat{\beta}_1$

El parámetro $\hat{\beta}_0$ corresponde al valor que adopta la variable EXPLICADA cuando la EXPLICATIVA es 0. Esto a veces no tiene un sentido realista, de modo que se abandona $\hat{\beta}_0$ como indicar para procesos de inferencia y solo se usa $\hat{\beta}_1$.

El modelo de regresión lineal simple ASUME:

- NORMALIDAD
- INDEPENDENCIA

8.10.1. DISTRIBUCIÓN PROBABILÍSTICA DE LA PENDIENTE

Eso conlleva que la DISTRIBUCIÓN de $\hat{\beta}_1$ sigue una ley NORMAL con:

$$E(\hat{\beta}_1) = \beta_1$$

Es decir, como esperanza, la distribución POBLACIONAL que sigue el ESTIMADOR MUESTRAL $\hat{\beta}_1$ toma el valor REAL β_1 (del cual $\hat{\beta}_1$ es una estimación), que es DESCONOCIDO.
Y con desviación típica correspondiente al ERROR ESTÁNDAR de la pendiente:

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \rightarrow INASEQUIBLE$$

Donde σ es la desviación típica poblacional que se asume DESCONOCIDA.

Ambos parámetros solo pueden conocerse teniendo acceso a TODA LA POBLACIÓN, por tanto, se asumen INALCANZABLES.

En el caso de la esperanza, desconocerlo no es crítico.

En el caso de la desviación típica, se usa un estimador, la desviación típica muestral, $s_{\hat{\beta}_1}$, que es el ERROR ESTÁNDAR de media de la pendiente:

$$s_{\hat{\beta}_1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \rightarrow ASEQUIBLE$$

Donde s_e^2 es la VARIANZA RESIDUAL, es decir, la varianza de los residuos definidos por $e = y - \hat{y}$, que se define como:

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Por tanto, considerando:

- La ESPERANZA de la distribución de la pendiente:

$$E(\hat{\beta}_1) = \beta_1$$

- El ERROR ESTÁNDAR, deducido a partir de la desviación típica muestral $s_{\hat{\beta}_1}$ (a falta de DESVIACIÓN típica poblacional $\sigma_{\hat{\beta}_1}$ y que se calcula a partir de la varianza residual s_e^2):

$$s_{\hat{\beta}_1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Se puede tipificar la variable:

$$\frac{\hat{\beta}_1 - \beta}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

Como una distribución t de Student con n – 2 grados de libertad.

8.10.2. INTERVALO DE CONFIANZA PARA LA PENDIENTE $\hat{\beta}_1$

El intervalo de confianza para la pendiente de la recta de regresión es:

$$(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{\beta}_1})$$

8.10.3. CONTRASTE DE HIPÓTESIS SOBRE LA PENDIENTE $\hat{\beta}_1$

Si se encuentra $\hat{\beta}_1 = 0$ en $\hat{\beta}_1 = \frac{\hat{s}_{xy}}{\hat{s}_x^2}$ se puede afirmar que la variable X NO ES EXPLICATIVA de la variable Y.

Para realizar el contraste de hipótesis:

1. Se establecen la hipótesis nula y la alternativa:

$$\begin{cases} H_0: & \hat{\beta}_1 = 0 & \rightarrow & \text{La variable X NO ES EXPLICATIVA de Y} \\ H_1: & \hat{\beta}_1 \neq 0 & \rightarrow & \text{La variable X SÍ ES EXPLICATIVA de Y} \end{cases}$$

2. Se fija un nivel de significación α
3. Bajo el supuesto de hipótesis nula CIERTA $\hat{\beta}_1 = 0$, se define el estadístico de contraste:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Que corresponde a la observación de una t de Student con n – 2 grados de libertad.

4. Para resolver, se puede recurrir a 2 instrumentos:

a) Valor crítico:

- $Si\ |t| < t_{\frac{\alpha}{2},n-2}$ NO se rechaza H_0
  NO hay relación lineal entre X e Y.
  X NO ES EXPLICATIVA.
- $Si\ |t| > t_{\frac{\alpha}{2},n-2}$ Se rechaza H_0
  Hay relación lineal entre X e Y.

b) p-valor:

Se recurre a la expresión del p-valor para contraste bilateral:

$$p_{valor} = 2 \cdot P(t_{n-2} > |t|)$$

- $Si\ p_{valor} > \alpha$ NO se rechaza H_0
  NO hay relación lineal entre X e Y.
  X NO ES EXPLICATIVA.
- $Si\ p_{valor} \leq \alpha$ Se rechaza H_0
  Hay relación lineal entre X e Y.

8.11. BONDAD DE AJUSTE: Coeficiente de CORRELACIÓN LINEAL de Pearson r

Se define la BONDAD DEL AJUSTE de un modelo de regresión como el grado de solapamiento que hay entre Hay múltiples indicadores del ajuste de la Se destaca el coeficiente de correlación lineal de Pearson r, que se calcula como:

$$r = \frac{\hat{S}_{xy}}{\hat{s}_x \hat{s}_y}$$

Donde:

r Es el coeficiente de correlación lineal de Pearson.
$\hat{S}_{xy}$ Es la COVARIANZA MUESTRAL entre x e y calculada como:

$$S_{xy} = \frac{\sum_{ij}(x_i \cdot y_j \cdot n_{ij}) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

$\hat{s}_x, \hat{s}_y$ Es la CUASIDESVIACIÓN TÍPICA MUESTRAL (¡NO LA CUASIVARIANZA!) de cada variable.

Propiedades de r:

1. Es adimensional (no tiene unidades).
2. $r \in [-1,1]$

Se interpreta según:

$$\begin{cases} r \to 1 & Relación\ de\ proporcionalidad\ DIRECTA \\ r \to 0 & No\ existe\ correlación\ entre\ x\ e\ y\ (variables\ INCORRELADAS) \\ r \to -1 & Relación\ de\ proporcionalidad\ INVERSA \end{cases}$$

La intensidad de la correlación depende por tanto del valor del coeficiente r.

Habitualmente, en el ámbito técnico, se toma:

$$\begin{array}{ll} |r| = 1 & Correlación\ perfecta \\ 0.8 < |r| < 1 & Correlación\ MUY\ ALTA \\ 0.6 < |r| < 0.8 & Correlación\ ALTA \\ 0.4 < |r| < 0.6 & Correlación\ MODERADA \\ 0.2 < |r| < 0.4 & Correlación\ BAJA \\ 0 < |r| < 0.2 & Correlación\ MUY\ BAJA \\ |r| = 0 & Correlación\ NULA \end{array}$$

No hay un criterio ÚNICO a la hora de evaluarlo, sino que depende de la situación y del propósito del modelo. Por ejemplo:

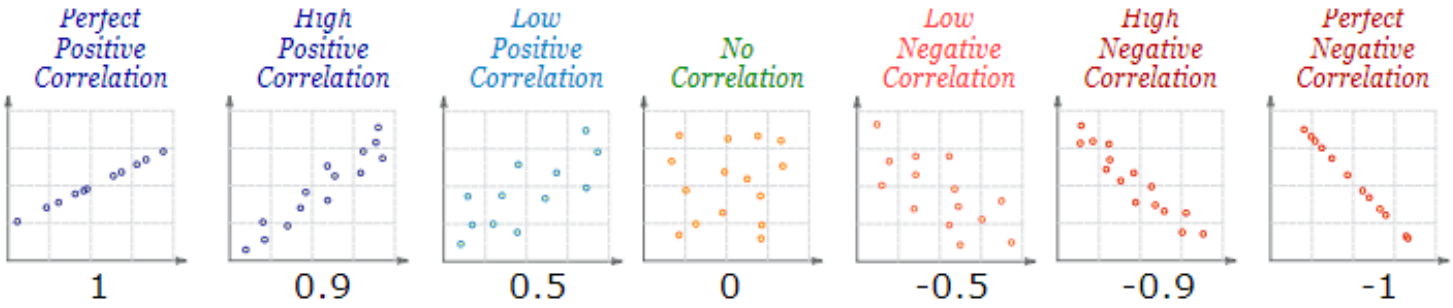
- En Psicología, interesa que los instrumentos de medida del comportamiento humano exhiban respuestas con una baja correlación (r<0.4) para validar su objetividad.

- En Biología, la correlación necesaria para validar resultados de dependencia de poblaciones celulares con marcadores moleculares estudiados suele ser de 0.9.

Si el modelo aspira a la PREDICCIÓN de sucesos se usa 0.8 como referencia.

En esta materia:

$$\begin{cases} |r| < 0.3 & Correlación\ BAJA\ o\ DÉBIL \\ 0.3 < |r| < 0.8 & Correlación\ MODERADA \\ |r| > 0.8 & Correlación\ ALTA\ o\ FUERTE \end{cases}$$



8.12. BONDAD DE AJUSTE: Coeficiente de DETERMINACIÓN R^2

R^2 mide la PROPORCIÓN de la VARIACIÓN de la variable dependiente EXPLICADA por la variable independiente, mientras que el cuadrado de r^2 mide el grado de CORRELACIÓN entre las 2 variables.

En el MODELO DE REGRESIÓN LINEAL SIMPLE, se cumple que el cuadrado del coeficiente de CORRELACIÓN r coincide con el coeficiente de DETERMINACIÓN que se escribe R^2:

$$r^2 = R^2$$

Donde:

r	Coeficiente de correlación $r = \frac{\hat{S}_{xy}}{\hat{s}_x \hat{s}_y}$
R^2	Coeficiente de DETERMINACIÓN

Esto no se cumple en otros modelos.
Dada esta equivalencia, la interpretación del valor absoluto de R^2 es la misma que para r, de modo que:

$$0 < R^2 < 0.5 \rightarrow correlación\ DÉBIL$$
$$0.5 < R^2 < 0.8 \rightarrow correlación\ MODERADA$$
$$0.8 < R^2 < 1 \rightarrow correlación\ FUERTE$$

8.1.3. EJEMPLOS

EJEMPLO 1

Se dispone de las edades en que un grupo de parejas tuvieron hijos por primera vez. Se presentan los datos como una tabla de contingencia:

X = Edad del padre
Y = Edad de la madre

Y(madre) / X(padre)	19	23	27	31	35
20	5	2	-	-	-
24	-	3	9	1	-
28	-	-	4	6	10
32	-	-	-	6	7
36	-	-	-	3	4

Se desea:

a) Estimar mediante una recta de regresión la edad del padre si la madre tiene 25 años.

b) Estimar mediante una recta de regresión la edad de la madre si el padre tiene 25 años.

c) Calcular e interpretar el coeficiente de correlación lineal y calcular el coeficiente de determinación.

1. En primer lugar, se completa la tabla de contingencia con los pares (x,y) para el cálculo de los 2 sumatorios necesarios $\sum_{i=1}^{n} n_i \cdot x_i$ y $\sum_{i=1}^{n} n_i \cdot y_i$

Y_i (madre) / X_i (padre)	19	23	27	31	35	n_i	$n_i \cdot x_i$
20	5	2	-	-	-	7	140
24	-	3	9	1	-	13	312
28	-	-	4	6	10	20	560
32	-	-	-	6	7	13	416
36	-	-	-	3	4	7	252
n_j	5	5	13	16	21	**60**	**1680**
$n_j \cdot y_j$	95	115	351	496	735	**1792**	-

Sumando los productos de cada valor por su frecuencia absoluta, se obtiene:

$$\sum_{i=1}^{n} n_i \cdot x_i = 1680$$

$$\sum_{j=1}^{n} n_j \cdot y_j = 1792$$

2. Se calcula la media de cada variable:

$$\bar{x}_{padre} = \frac{1680}{60} = 28$$

$$\bar{y}_{madre} = \frac{1792}{60} = 29.8\hat{6}$$

3. Se completa la tabla con el recuento de los cuadrados de las desviaciones respecto de la media $\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i$ y $\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot n_i$ para el cálculo de la cuasivarianza muestral $\hat{S}$ de cada variable.

Y(madre) / X(padre)	19	23	27	31	35	n_i	$n_i \cdot x_i$	$(x_i - \bar{x})^2 \cdot n_i$
20	5	2	-	-	-	7	140	448
24	-	3	9	1	-	13	312	208
28	-	-	4	6	10	20	560	0
32	-	-	-	6	7	13	416	208
36	-	-	-	3	4	7	252	448
n_j	5	5	13	16	21	60	1680	1312
$n_j \cdot y_j$	95	115	351	496	735	1792	-	-
$(y_j - \bar{y})^2 \cdot n_j$	590.78	235.98	107.08	20.43	552.65	1506.92	-	-

Se tienen los sumatorios:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i = 1312$$

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 \cdot n_j = 1506.92$$

4. Se calcula la CUASIVARIANZA MUESTRAL de ambas variables:

$$\hat{s}_x^2 = \hat{s}_{padre}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{n-1} = \frac{1312}{59} = 22.237$$

$$\hat{s}_y^2 = \hat{s}_{madre}^2 = \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2 \cdot n_j}{n-1} = \frac{1506.92}{59} = 25.541$$

5. Se calcula la COVARIANZA MUESTRAL:

$$\hat{S}_{xy} = \frac{\sum_{ij}(x_i \cdot y_j \cdot n_{ij}) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

$$\hat{S}_{xy} = \frac{(19 \cdot 20 \cdot 5 + 23 \cdot 20 \cdot 2 + 23 \cdot 24 \cdot 3 + 27 \cdot 24 \cdot 9 + 31 \cdot 24 \cdot 1 + 27 \cdot 28 \cdot 4 + 31 \cdot 28 \cdot 6 + 35 \cdot 28 \cdot 10 + 31 \cdot 32 \cdot 6 + 35 \cdot 32 \cdot 7 + 31 \cdot 36 \cdot 3 + 35 \cdot 36 \cdot 4) - 60 \cdot 28 \cdot 29.7}{59-1}$$

$$S_{xy} = 21.966$$

a) Se escribe la recta de regresión de X (padres) sobre Y (madres), o sea, X(Y), que permite ESTIMAR la edad del PADRE X a partir de la edad CONOCIDA de la MADRE Y.

$y = Edad\ CONOCIDA\ 25\ años\ para\ la\ MADRE$
$\hat{x} = Edad\ ESTIMADA\ del\ PADRE$

El modelo de regresión lineal dicta:

$$\widehat{estimada} = \hat{\beta}_1 \cdot explicativa + \hat{\beta}_0$$

Donde:

$$\hat{\beta}_1 = \frac{\text{Covarianza}}{S^2_{variable\ explicativa}}$$

$$\hat{\beta}_0 = \overline{estimada} - \beta_1 \cdot \overline{explicativa}$$

En este caso:
X = "Edad del padre" = ESTIMADA
Y = "Edad de la madre" = EXPLICATIVA

Entonces:

$$\widehat{padre} = \hat{\beta}_1 \cdot madre + \hat{\beta}_0$$

Donde la PENDIENTE es:

$$\hat{\beta}_1 = \frac{\text{Covarianza}}{S^2_{variable\ explicativa}} = \frac{\text{Covarianza}}{S^2_{madre}} = \frac{21.966}{25.541} = 0.86$$

Y la ORDENADA en el origen:

$$\hat{\beta}_0 = \overline{estimada} - \beta_1 \cdot \overline{explicativa}$$

Es decir:

$$\hat{\beta}_0 = \overline{padre} - 0.86 \cdot \overline{madre} = 28 - 0.86 \cdot 29.86\hat{} = 2.314$$

La recta X (padre) sobre Y (madre) es X(Y):

$$\widehat{padre} = 0.86 \cdot madre + 2.314$$

Es decir:

$$\hat{x} = 0.86 \cdot y + 2.314$$

Para edad de la madre y = 25 años, el padre tendrá:

$$\hat{x} = 0.9878 \cdot (25) + 2.2116 = \ 23.81\ años\ para\ el\ padre$$

Alternativamente, PARA MAYOR CLARIDAD, se recomienda usar SIEMPRE:

$$\hat{x} - \bar{x} = \frac{\hat{S}_{xy}}{\hat{s}_y^2}(y - \bar{y})$$

Es decir:

$$\hat{x} = \frac{\hat{S}_{xy}}{\hat{s}_y^2}(y - \bar{y}) + \bar{x} = 0.86 \cdot y \underbrace{-0.86 \cdot 29.8\hat{6} + 28}_{\hat{\beta}_0}$$

Para y = 25:

$$\hat{x} = 0.86 \cdot (25) - 0.86 \cdot 29.8\hat{6} + 28 = 23.81\ años$$

En este caso, la variable aleatoria denotada por Y es la variable EXPLICATIVA y X es la variable EXPLICADA, ¡que no nos confunda la letra que designa la variable!

ES HABITUAL RESERVAR "X" PARA LA VARIABLE EXPLICATIVA E "Y" PARA LA ESTIMADA.

b) Se escribe la recta de regresión de Y (madres) sobre X (padres), o sea, Y(X) que permite ESTIMAR la edad de la MADRE a partir de la edad CONOCIDA del PADRE.

$x = Edad\ CONOCIDA\ 25\ años\ para\ el\ PADRE$
$\hat{y} = Edad\ ESTIMADA\ de\ la\ MADRE$

En este caso se tiene:

$$\hat{y} - \bar{y} = \frac{\hat{S}_{xy}}{\hat{s}_x^2}(x - \bar{x})$$

Es decir:

$$\hat{y} = \frac{\hat{S}_{xy}}{\hat{s}_x^2}(x - \bar{x}) + \bar{y} = \frac{21.966}{22.237}(25 - 28) + 29.66\hat{7} = 26.703$$

Cuando el padre tenga 25 años, la madre tendrá 26.703 años.

c) Se calcula r y R^2:

$$r = \frac{\hat{S}_{\text{xy}}}{\hat{s}_x \hat{s}_y} = \frac{18.04}{\sqrt{22.237}\sqrt{25.541}} = 0.77 \rightarrow R^2 = r^2 = 0.77^2$$

Sobre el signo de r:
Se observa $r > 0$ lo cual denota una CORRELACIÓN POSITIVA entre ambas variables: cuando una crece, la otra también lo hace, es decir, existe una proporcionalidad DIRECTA entre la edad de ambos progenitores.

Sobre el valor absoluto de r:
Se observa un valor $r = 0.7 \in (0.6,0.8)$ de modo que se puede considerar un ajuste ALTO.

EJEMPLO 2

Se disponen de los siguientes datos sobre los resultados numéricos de las calificaciones de una materia (Y) y del número de horas de estudio semanales dedicadas (X) de 16 estudiantes:

$$\sum_{i=1}^{16} x_i = 96, \sum_{i=1}^{16} y_i = 64, \sum_{i=1}^{16} x_i \cdot y_i = 492, \sum_{i=1}^{16} x_i^2 = 657, \sum_{i=1}^{16} y_i^2 = 526$$

Se desea:

a) Estimar el modelo de regresión lineal que relaciona la calificación obtenida con el número de horas dedicadas semanalmente.

b) Calcular una medida de la bondad de ajuste.

c) Si un estudiante ha estudiado 8 horas, ¿qué nota se espera que obtenga?

d) ¿Cuál es el número mínimo de horas para conseguir un 5?

A partir de los sumatorios dados, se calcula:

1. Las medias muestrales:

$$\bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = \frac{96}{16} = 6$$

$$\bar{y} = \frac{\sum_{i=1}^{16} y_i}{16} = \frac{64}{16} = 4$$

2. Las CUASIVARIANZAS muestrales $\hat{s}_x^2$ y $\hat{s}_y^2$:

$$\hat{s}_x^2 = \frac{(\sum_{i=1}^{16} x_i^2) - n \cdot \bar{x}^2}{n-1} = \frac{657 - 16 \cdot 6^2}{16-1} = 5.4$$

$$\hat{s}_y^2 = \frac{(\sum_{i=1}^{16} y_i^2) - n \cdot \bar{y}^2}{n-1} = \frac{526 - 16 \cdot 4^2}{16-1} = 18$$

3. La COVARIANZA muestral:

$$\hat{S}_{xy} = \frac{\sum_{ij}(x_i \cdot y_j) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Es decir:

$$\hat{S}_{xy} = \frac{492 - 16 \cdot 6 \cdot 4}{16-1} = 7.2$$

a) Se escribe la recta de regresión de Y (calificaciones) sobre X (horas dedicadas), es decir Y(X):

$$\hat{y} = \beta_1 x + \beta_0$$

Para ello:

→ Se calcula la pendiente de la recta de regresión β_1 de Y sobre X:

$$\hat{\beta}_1 = \frac{\hat{S}_{xy}}{\hat{s}_x^2} = \frac{7.2}{5.4} = \frac{4}{3}$$

→ Se calcula la ordenada en el origen $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \cdot \bar{x} = 4 - \frac{4}{3} \cdot 6 = -4$$

→ Se escribe la recta de regresión de Y sobre X

$$\hat{y} = \frac{4}{3}x - 4$$

b) Una medida de la bondad del ajuste es el coeficiente de correlación r:

$$r = \frac{\hat{S}_{xy}}{\hat{s}_x \hat{s}_y} = \frac{7.2}{\sqrt{5.4}\sqrt{18}} = 0.73$$

Por tanto:

- $r > 0 \rightarrow$ Hay correlación POSITIVA entre X e Y

A mayor tiempo de estudio, mayor es la calificación obtenida.

- $r \in (0.6, 0.8) \rightarrow La\ correlación\ es\ ALTA$

Es ajuste es satisfactorio.

c) Para X = 8 horas de estudio, se ESTIMA una calificación $\hat{y}$:

$$\hat{y} = \frac{4}{3}(8) - 4 = 6.\hat{6}$$

d) En este caso, la regresión se realiza de X (horas) sobre Y (calificación), es decir X(Y). La recta es:

→ Se calcula la pendiente de la recta de regresión $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{xy}}{\hat{s}_y^2} = \frac{7.2}{18} = 0.4$$

→ Se calcula la ordenada en el origen $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \cdot \bar{y} = 6 - 0.4 \cdot 4 = 4.4$$

→ Se escribe la recta de regresión de X sobre Y:

$$\hat{x} = 0.4 \cdot y + 4.4$$

Para Y = 5, se ESTIMA un tiempo de estudio $\hat{x}$:

$$\hat{x} = 0.4 \cdot (5) + 4.4 = 6.4\ horas$$

EJEMPLO 3

Se consideran los datos de altura y peso de 10 personas:

individuo i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Altura (x_i)	*161*	*152*	*167*	*153*	*161*	*168*	*167*	*153*	*159*	*173*
Peso (y_i)	*63*	*56*	*77*	*49*	*72*	*62*	*68*	*48*	*57*	*67*

Se desea:

a) Determinar la recta de regresión lineal de esta muestra de Y (peso) sobre X (altura), es decir Y(X).
b) Determinar el coeficiente de correlación lineal de Pearson r.
c) Dar un intervalo de confianza del 95% para la pendiente.
d) Realizar un contraste de hipótesis sobre la media mediante valor crítico y mediante p-valor tomando un nivel de significación $\alpha = 0.05$.

Se construye la tabla necesaria:

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$\hat{y}_i$	$e^2 = (y_i - \hat{y})^2$
1	*161*	*63*	*-0,4*	*1,1*	*0,16*	*1,21*	*-0,44*	*61,50*	*2,22*
2	*152*	*56*	*-9,4*	*-5,9*	*88,36*	*34,81*	*55,46*	*52,69*	*10,90*
3	*167*	*77*	*5,6*	*15,1*	*31,36*	*228,01*	*84,56*	*67,38*	*92,49*
4	*153*	*49*	*-8,4*	*-12,9*	*70,56*	*166,41*	*108,36*	*53,67*	*21,86*
5	*161*	*72*	*-0,4*	*10,1*	*0,16*	*102,01*	*-4,04*	*61,50*	*110,07*
6	*168*	*62*	*6,6*	*0,1*	*43,56*	*0,01*	*0,66*	*68,36*	*40,46*
7	*167*	*68*	*5,6*	*6,1*	*31,36*	*37,21*	*34,16*	*67,38*	*0,38*
8	*153*	*48*	*-8,4*	*-13,9*	*70,56*	*193,21*	*116,76*	*53,67*	*32,22*
9	*159*	*57*	*-2,4*	*-4,9*	*5,76*	*24,01*	*11,76*	*59,55*	*6,50*
10	*173*	*67*	*11,6*	*5,1*	*134,56*	*26,01*	*59,16*	*73,25*	*39,14*

Para elaborar la tabla:

1) Se calculan lo promedios $\bar{x}$ e $\bar{y}$:

$$\bar{x} = 161.4$$
$$\bar{y} = 61.9$$

2) Se calculan las desviaciones respecto la media $x_i - \bar{x}$ e $y_i - \bar{y}$ necesarias para calcular la covarianza.
3) Se calcula su cuadrado para la cuasivarianza muestral:

$$\hat{s}_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} = 52.9\hat{3}$$
$$\hat{s}_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{n-1} = 90.3\hat{2}$$

4) Se calcula el producto de las desviaciones respecto la media $(x_i - \bar{x})(y_i - \bar{y})$ para el cálculo de la covarianza:

$$\hat{S}_{\text{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = 51.8\hat{2}$$

5) Se deduce $\hat{\beta}_1$ a partir de la covarianza S_{xy} y la cuasivarianza muestral de X $\hat{s}_x^2$:

$$\hat{\beta}_1 = \frac{S_{\text{xy}}}{\hat{s}_x^2} = \frac{51.8\hat{2}}{52.9\hat{3}} = 0.979$$

6) Se deduce la ordenada como $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$:

$$\hat{\beta}_0 = 61.9 - 0.979 \cdot 161.4 = -96.112$$

7) Se escribe la recta de regresión:

$$\hat{y} = \hat{\beta}_1 \cdot x + \hat{\beta}_0 = 0.979 \cdot x - 96.112$$

8) Se calculan las estimaciones $\hat{y}$ usando la recta de regresión para cada x.

9) Se calcula el residuo cuadrado de cada estimación según $e^2 = (y_i - \hat{y})^2$ que servirán para determinar la varianza residual s_e^2 que, a su vez, servirá para escribir el error estándar que se usará en el intervalo de confianza y en el contraste de hipótesis.

Para el intervalo de confianza:

1) Se calcula el error estándar de la pendiente:

   i. Se suma las desviaciones cuadradas de x respecto su media:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 476.6$$

   ii. Se usa la columna e^2 para calcular la varianza marginal:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y})^2 = 44.53$$

   iii. El error estándar es:

$$s_{\hat{\beta}_1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \sqrt{\frac{44.53}{476.6}} = 0.305$$

2) Se calcula el valor crítico t para $\alpha = 0.05$ y n = 10 – 2 df:

$$t_{\frac{\alpha}{2},n-2} = t_{\frac{0.05}{2},8} = t_{0.025,8} = \underbrace{2.306}_{\substack{fila\ 8 \\ col\ 0.025}}$$

3) Se escribe el intervalo de confianza:

$$\left(\hat{\beta}_1 \mp t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{\beta}_1}\right) = (0.979 - 2.306 \cdot 0.305, 0.979 + 2.306 \cdot 0.305) = (0.273, 1.684)$$

Para el contraste de hipótesis:

1) Se establecen las hipótesis nula y alternativa:

$$\begin{cases} H_0: & \hat{\beta}_1 = 0 & \rightarrow & \text{La variable X NO ES EXPLICATIVA de Y} \\ H_1: & \hat{\beta}_1 \neq 0 & \rightarrow & \text{La variable X SÍ ES EXPLICATIVA de Y} \end{cases}$$

2) Se fija el nivel de significación $\alpha = 0.05$

3) Bajo el supuesto de hipótesis nula CIERTA $\hat{\beta}_1 = 0$, se define el estadístico de contraste:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.979}{0.305} = 3.201$$

4) Se resuelve:

a. Utilizando el valor crítico t:

Se cumple:

- $Si\ |t| < t_{\frac{\alpha}{2},n-2}$ NO se rechaza H_0
  NO hay relación lineal entre X e Y.
  X NO ES EXPLICATIVA.
- $Si\ |t| > t_{\frac{\alpha}{2},n-2}$ Se rechaza H_0
  Hay relación lineal entre X e Y.

Se tiene:

$$|t| = \left|\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right| = 3.201 > t_{\frac{\alpha}{2},n-2} = t_{0.05,8} = 2.306$$

Se rechaza H_0 y se acepta X como EXPLICATIVA de Y.

b. Usando el p-valor:
Se cumple:

$$p_{valor} = 2 \cdot P(t_{n-2} > |t|)$$

- $Si\ p_{valor} > \alpha$ NO se rechaza H_0
  NO hay relación lineal entre X e Y.
  X NO ES EXPLICATIVA.
- $Si\ p_{valor} \leq \alpha$ Se rechaza H_0
  Hay relación lineal entre X e Y.

Se tiene:

$$p_{valor} = 2 \cdot P\left(t_{n-2} > \left|\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right|\right) = 2 \cdot \underbrace{P(t_8 > 3.201)}_{NO\ EST\acute{A}\ EN\ LA\ TABLA} = 2 \cdot \underbrace{0.0063}_{\substack{calculadora \\ de\ valores\ T}} = 0.0126$$

El valor se ha calculado usando ESTA CALCULADORA, se elige "t score", luego "df" = 8 y "t score" = 3.201.

Se observa: $p_{valor} = 0.0126 < 0.05 \rightarrow SE\ RECHAZA\ H_0\ y\ X\ ES\ EXPLICATIVA$

ENTREGABLE 6 – Cuestionario

PREGUNTA 1

En el modelo de regresión lineal $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ tenemos que

$\hat{y}$	es	el valor esperado de y según la recta de regresión
x	es	la variable independiente o EXPLICATIVA
$y - \hat{y}$	es	el residuo
y	es	la variable dependiente o EXPLICADA
$\hat{\beta}_1$	es	la pendiente de la recta de regresión
$\hat{\beta}_0$	es	la ordenada en el origen

PREGUNTA 2

$\frac{S_{xy}}{s_x s_y}$	Es el coeficiente de correlación
$\frac{S_{xy}}{s_x^2}$	La pendiente de la recta de regresión
$\bar{y} - \hat{\beta}_1 \cdot \bar{x}$	La ordenada del origen
r^2	El coeficiente de determinación

PREGUNTA 3

Se ha tomado una muestra de 7 parejas de datos en la siguiente tabla:

x_i	y_i
2	4
1	1
2	5
2	5
0	3
3	2
1	2

Calculad la recta de regresión de sobre x en la forma $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{x}$

a) S_{xy}
b) $\hat{\beta}_1$
c) $\hat{\beta}_0$
d) r

1. En primer lugar, se calcula la media de cada variable:

$$\bar{x} = \frac{0 \cdot 1 + 1 \cdot 2 + 2 \cdot 2 + 3 \cdot 1}{7} = \frac{11}{7}$$

$$\bar{y} = \frac{1 \cdot 1 + 2 \cdot 2 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 2}{7} = \frac{22}{7}$$

2. En segundo lugar, se construye la tabla para obtener los sumatorios necesarios:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
2	4	$\frac{3}{7}$	$\frac{6}{7}$	$\frac{9}{49}$	$\frac{36}{49}$	$\frac{18}{49}$
1	1	$-\frac{4}{7}$	$-\frac{15}{7}$	$\frac{16}{49}$	$\frac{225}{49}$	$\frac{60}{49}$
2	5	$\frac{3}{7}$	$\frac{13}{7}$	$\frac{9}{49}$	$\frac{169}{49}$	$\frac{39}{49}$
2	5	$\frac{3}{7}$	$\frac{13}{7}$	$\frac{9}{49}$	$\frac{169}{49}$	$\frac{39}{49}$
0	3	$-\frac{11}{7}$	$-\frac{1}{7}$	$\frac{121}{49}$	$\frac{9}{49}$	$\frac{11}{49}$
3	2	$\frac{10}{7}$	$-\frac{8}{7}$	$\frac{100}{49}$	$\frac{9}{49}$	$\frac{-80}{49}$
1	2	$-\frac{4}{7}$	$-\frac{8}{7}$	$\frac{16}{49}$	$\frac{64}{49}$	$\frac{32}{49}$
TOTAL				$\frac{40}{7}$	$\frac{104}{7}$	$\frac{17}{7}$

3. En tercer lugar, se calculan las cuasivarianzas muestrales:

$$\hat{s}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\frac{40}{7}}{6} = \frac{40}{42} = \frac{20}{21}$$

$$\hat{s}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\frac{104}{7}}{6} = \frac{104}{42} = \frac{52}{21}$$

4. Se calcula la covarianza muestral:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\frac{17}{7}}{6} = \frac{17}{42}$$

5. Se calcula la pendiente de la recta de regresión $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{xy}}{\hat{s}_x^2} = \frac{\frac{17}{42}}{\frac{20}{21}} = \frac{17}{40}$$

6. Se calcula la ordenada en el origen $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = \frac{22}{7} - \frac{17}{40} \cdot \frac{11}{7} = \frac{99}{40}$$

7. Se calcula el coeficiente de correlación r:

$$r = \frac{S_{xy}}{\hat{s}_x \cdot \hat{s}_y} = \frac{\frac{17}{42}}{\sqrt{\frac{20}{21}}\sqrt{\frac{52}{21}}} = 0.2635$$

PREGUNTA 4

Se ha tomado una muestra de 70 empresas, de las que se ha obtenido el número de empleados X y el gasto en formación continua Y (en euros). La siguiente tabla muestra las principales estadísticas obtenidas de los datos:

	Variable X	Variable Y
Media	44	244
Varianza	51	1863

Se sabe que el coeficiente de correlación entre X e Y es igual a 0.76. Usando un modelo de regresión lineal de la forma $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ calculad los siguientes valores y la predicción $\bar{y}$ de gasto en formación continua de una empresa con 40 empleados. Se desea calcular:

a) $\hat{\beta}_1$
b) $\hat{\beta}_0$
c) $\hat{y}(40)$

Se tiene:
$\bar{x} = 44$ $\bar{y} = 244$ $\hat{s}_x^2 = 51$ $\hat{s}_y^2 = 1863$ r = 0.76

a) Se calcula la pendiente de la recta de regresión $\hat{\beta}_1$. Para aplicar la definición:

$$\hat{\beta}_1 = \frac{S_{\text{xy}}}{s_x^2}$$

Es necesario conocer la covarianza:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Pero no se disponen de datos para calcularlo. Por tanto, se recurre al valor conocido de r y su expresión para deducir S_{xy}:

$$r = \frac{S_{\text{xy}}}{\hat{s}_x \hat{s}_y} \rightarrow S_{\text{xy}} = r \cdot \hat{s}_x \cdot \hat{s}_y$$

→ Nótese el uso de la desviación muestral S_x y S_y en lugar de la varianza S_x^2 y S_y^2 dadas. En este caso:

$$S_{\text{xy}} = 0.76 \cdot \sqrt{51} \cdot \sqrt{1863} = 234.263$$

Entonces:

$$\hat{\beta}_1 = \frac{S_{\text{xy}}}{\hat{s}_x^2} = \frac{234.263}{51}$$

b) Se calcula la ordenada en el origen $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 244 - \frac{234.263}{51} \cdot 44 = 41.8907$$

c) Se calcula la estimación del gasto para 40 empleados usando la recta de regresión. Se tiene:

$$\hat{y} = \hat{\beta}_1 \cdot x + \hat{\beta}_0 \rightarrow \hat{y} = \frac{234.263}{51} \cdot x + 41.8907$$

Para 40 empleados:

$$\hat{y} = \frac{234.263}{51} \cdot 40 + 41.8907 = 225.6263€ \; de \; gasto \; estimado$$

PREGUNTA 5

Se quiere estudiar la relación entre la variable X y la variable Y y disponemos de los siguientes datos correspondientes a 15 observaciones:

$$\sum_{i=1}^{15} x_i = 431, \sum_{i=1}^{15} y_i = 1292, \sum_{i=1}^{15} x_i^2 = 12789, \sum_{i=1}^{15} x_i y_i = 37810$$

Calculad la recta de regresión de Y sobre X en la forma $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

Se dispone de los sumatorios necesarios para calcular:

a) Las medias muestrales:

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{431}{15}$$

$$\bar{y} = \frac{\sum_{i=1}^{15} y_i}{15} = \frac{1292}{15}$$

b) La cuasivarianza muestral S_x^2 para calcular la pendiente:

$$\hat{s}_x^2 = \frac{\left(\sum_{i=1}^{15} x_i^2\right) - n \cdot \bar{x}^2}{n-1} = \frac{12789 - 15 \cdot \left(\frac{431}{15}\right)^2}{15-1} = 28.924$$

c) La covarianza:

$$S_{\text{xy}} = \frac{\left(\sum_{i=1}^{15} x_i \cdot y_i\right) - n \cdot \bar{x} \cdot \bar{y}}{n-1} = \frac{37810 - 15 \cdot \frac{431}{15} \cdot \frac{1292}{15}}{15-1} = 49.038$$

d) Se calcula la pendiente de la recta de regresión $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{\text{xy}}}{\hat{s}_x^2} = \frac{49.038}{28.924} = 1.695$$

e) Se calcula la ordenada en el origen $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = \frac{1292}{15} - \frac{49.038}{28.924} \cdot \frac{431}{15} = 37.418$$

f) Se escribe la recta de regresión:

$$\hat{y} = 37.418 + \frac{49.038}{28.924} \cdot x$$

PREGUNTA 6

Se ha tomado una muestra de 11 empresas, de las que se ha obtenido el número de trabajadores (X) y el gasto en formación continua Y (en euros). La siguiente tabla muestra las principales estadísticas obtenidas de los datos:

	Variable X	Variable Y
Media	33.846154	576.767692
Varianza	13.307692	42740.169269

Sabiendo además que la covarianza de X e Y es 530.596282, calculad el coeficiente de correlación r y el coeficiente de determinación R^2.

a) Se calcula r:

$$r = \frac{S_{xy}}{\hat{s}_x \hat{s}_y} = \frac{530.596282}{\sqrt{13.307692}\sqrt{42740.169269}} = 0.703549$$

b) Se calcula r^2:

$$r^2 = 0.7035^2 = R^2 = 0.494982$$

Nótese que $r^2 = R^2$ SOLO EN EL CASO DE LA REGRESIÓN LINEAL SIMPLE.

PREGUNTA 7

Hemos obtenido el siguiente modelo de regresión por mínimos cuadrados: $\hat{y} = 0.125 \cdot x - 75$ en que x representa el nivel de ingresos (en euros) y la variable $\hat{y}$ representa el nivel de gastos en cultura (en euros). Si suponemos que el ajuste es bueno y que las siguientes frases son ciertas, calcula el valor de A, B, C, D.

a) Si aumentamos el nivel de ingresos en un euro, el nivel de gastos en cultura aumenta en A unidades.

El valor A equivale al incremento de $\hat{y}$ por cada unidad que x aumenta, es decir, coincide con la pendiente de la recta 0.125 = A.

b) Si el nivel de ingresos fuese 0 €, entonces el nivel de gastos en cultura sería B.

El nivel de gasto estimado $\hat{y}(x)$ para x = 0 es:

$$B = \hat{y}(0) = 0.125 \cdot 0 - 75 = -75€$$

Aunque por contexto no tenga sentido en este caso.

c) Si el nivel de ingresos fuese 1250 euros, entonces el nivel de gastos en cultura sería C euros.

Se tiene:

$$C = \hat{y}(1250) = 0.125 \cdot 1250 - 75 = 81.25€$$

d) El residuo correspondiente al punto observado (2000,250) sería D.

Se tiene (x,y) = (2000,250).
El residuo D es: $y - \hat{y}$
Donde: $\hat{y} = \hat{\beta}_1 \cdot x + \hat{\beta}_0$
Con x = 2000: $\hat{y} = 0.125 \cdot 2000 - 75 = 175$
Entonces: $y - \hat{y} = 250 - 175 = 75€ = D$

ENTREGABLE 6 – Práctica de R

Importación, combinación y selección de los datos

Se recurre a la instrucción:

```
dadesPM10<read.table("C:/Users/Tete/Desktop/AirPollution2000WB_UOC2.csv",header=TRUE,sep=";", na.strings="NA",fileEncoding="UTF-8",quote = "\"", colClasses=c(rep("character",4),rep("numeric",2), rep("character",2)))
```

Para verificar la correcta importación de los datos, se ejecuta head(dadesPM10,3), que devuelve:

	Cod	Country	Citycode	City	Population2000	PM10Concentration1999	Region	IncomeGroup
1	AFG	Afghanistan	40003	Herat	323741	46	South Asia	Low income
2	AFG	Afghanistan	40001	KABUL	2457496	46	South Asia	Low income
3	AFG	Afghanistan	40002	Kandahar (Quandahar)	411752	51	South Asia	Low income

Además, se importa una segunda base de datos mediante la instrucción:

```
dadesAir<-read.table("WHO_AirQuality_Database_2013_UOC.CSV", header=TRUE,
sep=";",dec=".",na.strings="NA",skip=9, fileEncoding = "UTF-8", quote = "\"",fill=TRUE, )
names(dadesAir)
```

Cuya visualización mediante head(dadesAir,3) devuelve:

	iso3	Country	City	pm10	Year	Population	date_compiled	Region	region_abbr
1	ALB	Albania	Tirana	31.61542	2013	4535092	2016	Europe (LMIC)	Eur (LMIC)
2	AUS	Australia	Central Coast	12.82046	2014	297713	2016	Western Pacific (HIC)	Wpr (HIC)
3	AUS	Australia	Devonport	14.91836	2013	29050	2016	Western Pacific (HIC)	Wpr (HIC)

I.

II. Para la comparación de los años se recurre a merge() y se considerarán solamente los registros de ciudades en los cuales constan dos valores de PM10 superiores a 0. En primer lugar:

III.

```
datos<-merge(dadesAir,dadesPM10,by=c("City","Country"))
```

IV. A continuación, se genera un subconjunto de datos para el análisis mediante la instrucción subset():

```
set<-subset(datos,Year==2013)
```

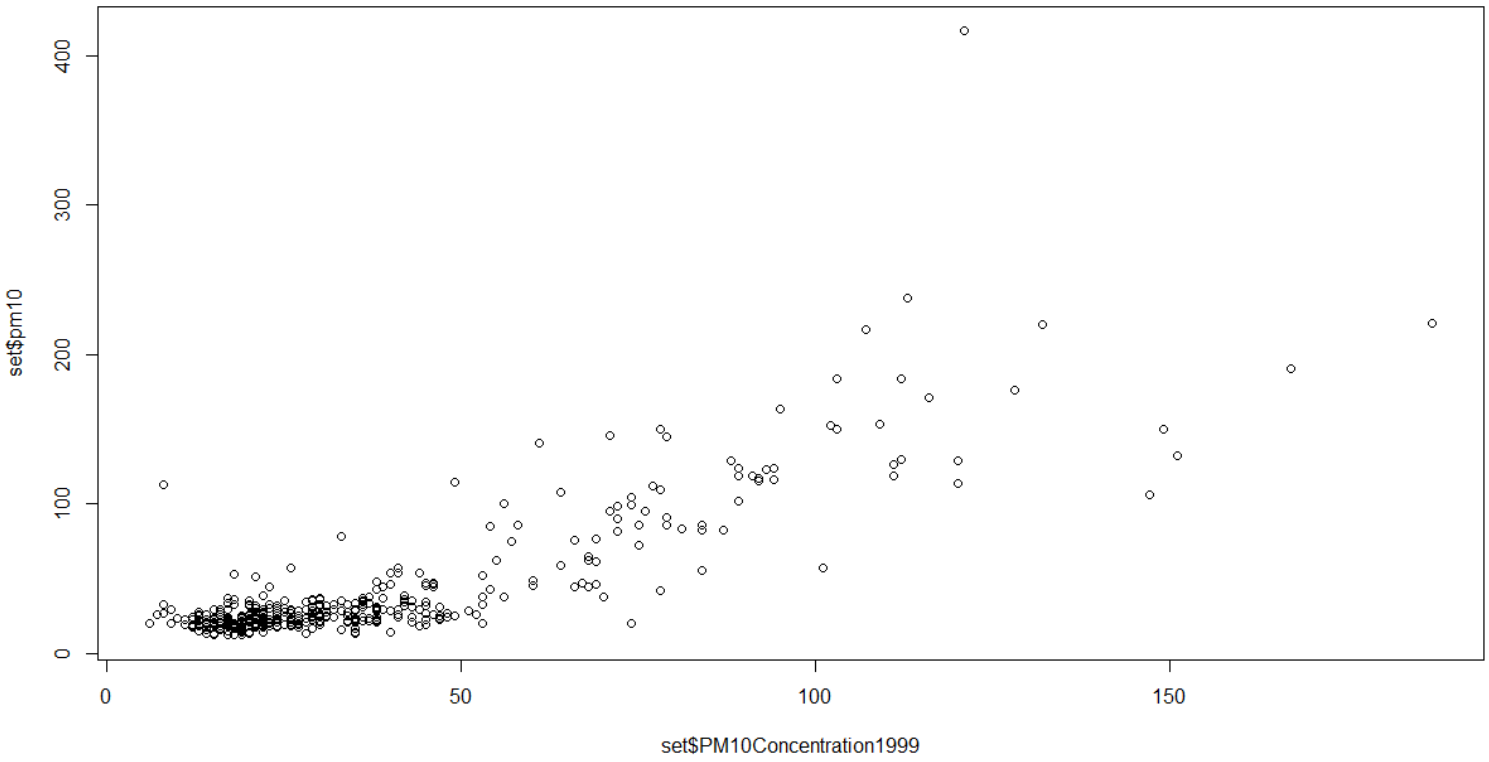
PREGUNTA 1

Haced con R el diagrama de dispersión de la nube de puntos de la variable concentración de PM10 del año 2013 (en el eje de ordenadas) en función de la variable concentración de PM10 del año 1999 (en el eje de abscisas). Comentad la gráfica obtenida.

El diagrama de dispersión se obtiene mediante la instrucción plot() tomando como argumentos ORDENADOS el conjunto de datos para la variable independiente y el conjunto para la variable dependiente:

plot(set$PM10Concentration1999,set$pm10)

Y se obtiene:



Se concluye que existe una relación de proporcionalidad directa entre ambas variables, ya que a medida que el tiempo avanza, se observa un incremento en las partículas contaminantes analizadas.

La dispersión no es especialmente baja y no se observa una masa significativa de puntos separada del conjunto mayor, más allá de los escasos puntos repartidos a lo largo de la línea, que contribuirán a elevar el residuo mucho más significativamente que la inmensa mayoría de puntos.

Por tanto, es plausible la existencia de un modelo de regresión lineal que permita una aproximación relativamente fiel de la nube de puntos observada.

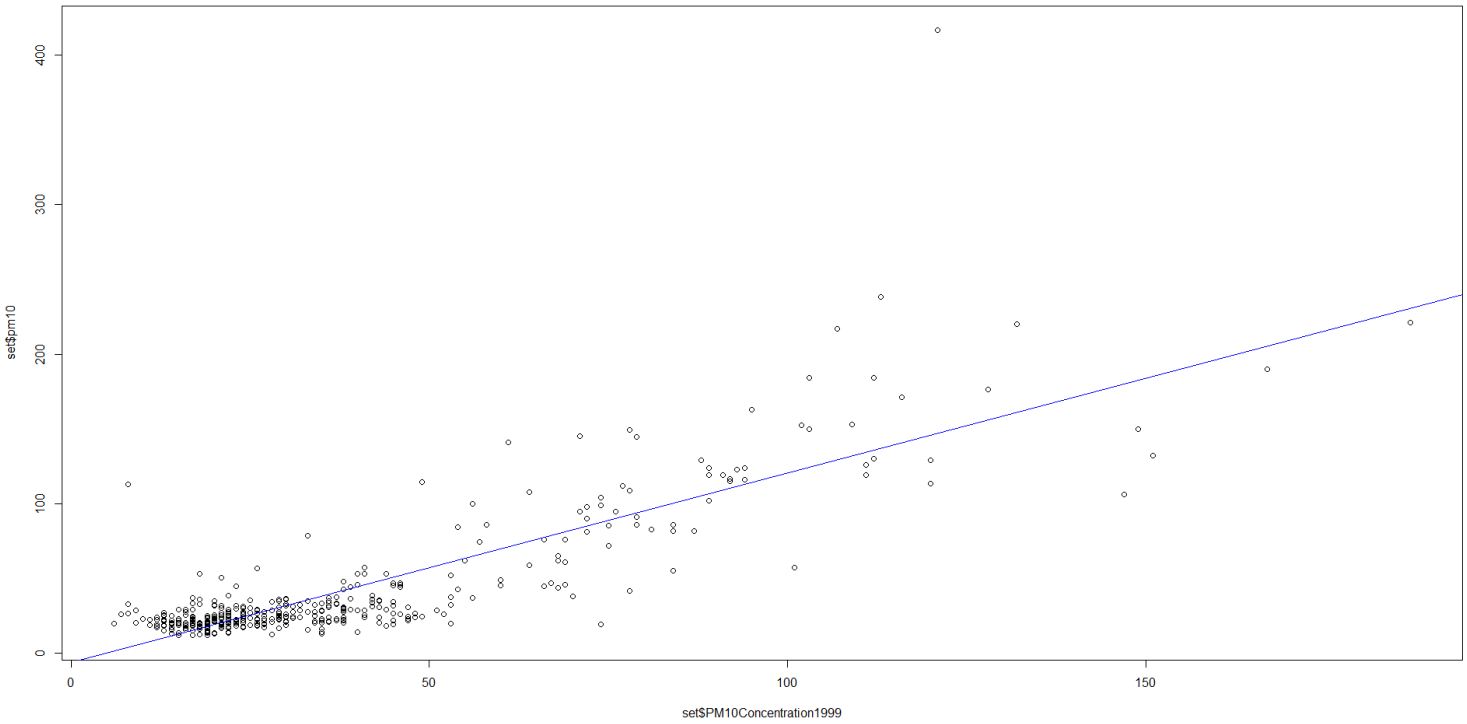
PREGUNTA 2

Calculad con R la recta de regresión de la variable concentración de PM10 del año 2013 en función de la variable concentración de PM10 del año 1999. Dad la pendiente y la ordenada en su origen. Interpretad el valor de la pendiente obtenida.

Para evaluar la correlación en la nube de puntos, se traza una recta de regresión (en color azul en el gráfico) mediante la instrucción abline():

abline(lm(set$pm10 ~ set$PM10Concentration1999),col="blue")

Se obtiene:



Se observa una recta de regresión que ajusta los datos representados al modelo de regresión lineal.

Para evaluar la adecuación del modelo calculado, se debe acceder a la expresión de la recta. Para ello, se recurre a la instrucción summary() tomando como argumentos los mismos que los tomados para la regresión:

summary(lm(set$pm10 ~ set$PM10Concentration1999))

que da lugar a:

```
Call:
lm(formula = set$pm10 ~ set$PM10Concentration1999)

Residuals:
Min              1Q     Median         3Q     Max
-73.859         -9.314         -0.441              7.100  269.299

Coefficients:                          Estimate
(Intercept)                            -5.91938
set$PM10Concentration1999              1.26380
Std. Error
(Intercept)                            1.99078
set$PM10Concentration1999              0.04144
t value
(Intercept)                            -2.973
set$PM10Concentration1999              30.498
Pr(>|t|)
(Intercept)                            0.00312 **
set$PM10Concentration1999              < 2e-16 ***
---
Signif. codes:
0 '***'          0.001 '**'     0.01 '*'              0.05 '.'              0.1 ' '
    1
Residual standard error:           23.95 on 399 degrees of freedom
Multiple R-squared:          0.6998,
Adjusted R-squared:          0.6991
F-statistic:                 930.1 on 1 and 399 DF,
p-value:                     < 2.2e-16
```

La pendiente de la recta es: 1.26380
La ordenada en el origen es: -5.91938

Por tanto, la expresión de la recta de regresión es:

$$\hat{y} = \hat{\beta}_0 + \beta_1 \cdot \hat{x} \rightarrow y = -5.91938 + 1.2638 \cdot x$$

El carácter POSITIVO de la pendiente da lugar a una recta CRECIENTE, lo cual denota la relación de proporcionalidad DIRECTA entre ambas muestras: se reafirma la tendencia creciente de la contaminación en las ciudades de la muestra estudiada.

PREGUNTA 3

¿Cuál es el valor del coeficiente de determinación? ¿Y el valor del coeficiente de correlación? ¿Qué podéis decir sobre la bondad del ajuste?

El coeficiente de determinación obtenido es $r^2 = 0.6998$.
El coeficiente de correlación obtenido es $r = \sqrt{r^2} \rightarrow r = \sqrt{0.6998} = 0{,}83654$.

El modelo presenta un ajuste relativamente adecuado, aunque no se alcanza niveles cercanos a la correlación perfecta (r = 1).

Valores de r bajos no son inherentemente inadecuados, ni viceversa. El coeficiente de correlación aplicado, por ejemplo, en el campo de la psicología, es habitualmente un indicador sólido de la adecuación de los instrumentos de medida del comportamiento humano y se desea que se mantenga en valores bajos, lo cual legitima estadísticamente la ausencia de sesgos en los instrumentos usados.

En cambio, en otras disciplinas como la medicina o la biología celular, niveles de r por debajo de 0.8 permiten desechar habitualmente resultados experimentales.
Así mismo, para evaluar la bondad del ajuste, además del valor de r, sería conveniente consultar, por ejemplo, el gráfico de residuos.

En conclusión, se desea un valor de r cercano a 1 en aquellas situaciones en la cuales el coeficiente de correlación deba servir a la construcción de modelos predictivos.

PREGUNTA 4

Estimad el valor esperado de concentración de PM10 del año 2013 cuando la concentración de PM10 del año 1999 es de 75µg/m^3.

La recta de regresión puede aplicarse a la estimación del valor de contaminación esperado en 2013 a partir de un valor conocido en 1999. Para ello, basta con sustituir en la expresión x = 75. Se obtiene:

$$\hat{y} = -5.91938 + 1.2638 \cdot x = -5.91938 + 1.2638 \cdot 75 = 88.86562$$

PREGUNTA 5

Queremos hacer un contraste de hipótesis con un nivel de significación del 0.05 sobre la pendiente de la recta de regresión obtenida para saber si la variable X es explicativa. Indicad las hipótesis nula y alternativa, el p-valor y la conclusión a la que llegáis.

Para el test de contraste de hipótesis sobre el valor de la pendiente, se toman como hipótesis:

- NULA: Que el valor de la pendiente SEA 0
  (en cuyo caso, la variable X NO es explicativa)
- ALTERNATIVA: Que el valor de la pendiente sea DISTINTO de 0.
  (en cuyo caso, la variable X ES explicativa)

Es decir:

$$\begin{cases} H_0: & B_1 = 0 & \rightarrow & \text{La variable X NO ES EXPLICATIVA} \\ H_1: & B_1 \neq 0 & \rightarrow & \text{La variable X SÍ ES EXPLICATIVA} \end{cases}$$

Para ello, tan solo hace falta comparar el p-valor obtenido en el sumario de la regresión con el nivel de significación con que se desea realizar el contraste de hipótesis. Se tiene:

$$p_{valor} < 2 \cdot 10^{-16} \ll 0.05 = \alpha$$

Por lo tanto, se RECHAZA la hipótesis NULA y se ACEPTA la hipótesis ALTERNATIVA, de modo que se CONFIRMA el carácter EXPLICATIVO de X con una confianza del 99.5%.