EJERCICIOS RESUELTOS TEMA 7 Y TEMA 8:
CORRELACIÓN Y REGRESIÓN

Ejercicio 1.

En 8 empresas de una industria textil de la región se han recogido datos sobre la "facturación en miles de euros" (X) y el "número de trabajadores" (Y). Con esa información se han calculado los estadísticos que se indican a continuación:

$$\sum_{i=1}^{8} x_i = 56 \qquad \sum_{i=1}^{8} y_i = 40 \qquad \sum_{i=1}^{8} x_i \times y_i = 1290 \qquad S_X^2 = 120 \qquad S_Y^2 = 146$$

a. Calcula la recta de regresión de mínimos cuadrados que explica el número de trabajadores como una función lineal de la facturación.
b. Analiza la bondad del ajuste obtenido mediante el coeficiente de determinación y la varianza explicada por la regresión
c. Predice según el modelo el número de trabajadores que tendrá una empresa con una facturación de 52000 euros

SOLUCIÓN:

a) Para estimar la recta de regresión, esto es obtener sus coeficientes, simplemente aplicamos las expresiones de éstos y calculamos sus valores a partir de los estadísticos facilitados en el enunciado. Con base al criterio de mínimos cuadrados, la estimación óptima de los parámetros de la recta de regresión viene dada por las fórmulas resaltadas. Éstas se utilizarán ampliamente a lo largo de este documento. Nótese que se denotan con circunflejo a fin de poner de manifiesto que se trata de estimaciones de los coeficientes del modelo de regresión formal: Y=a+bX+e

$$\hat{b} = \frac{S_{X,Y}}{S_X^2} \qquad \hat{a} = \bar{y} - \hat{b} \times \bar{x}$$

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{56}{8} = 7 \qquad \bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{40}{8} = 5 \qquad S_X^2 = 120 \quad S_Y^2 = 146$$

$$S_{X,Y} = \frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x}) \times (y_i - \bar{y}) = \frac{1}{8}\sum_{i=1}^{8} x_i \times y_j - \bar{x} \times \bar{y} = \frac{1240}{8} - 7 \times 5 = 99$$

$$\hat{b} = \frac{S_{X,Y}}{S_X^2} = \frac{99}{120} = 0{,}825$$

$$\hat{a} = \bar{y} - \hat{b} \times \bar{x} = 5 - 0{,}825 \times 7 = -0{,}775$$

Ecuación de mínimos-cuadrados: $Y = -0{,}775 + 0{,}825\,X$

b) En este apartado nos piden evaluar la bondad del ajuste obtenido en el apartado anterior de dos maneras: mediante el valor del coeficiente de determinación (lineal) y mediante la varianza residual. La expresión del coeficiente de determinación lineal que utilizamos que no requiere del cálculo de los residuos es:

$$r^2 = \frac{S_{XY}^2}{S_X^2 \times S_Y^2} = \frac{99^2}{120 \times 146} \cong 0{,}56$$

El valor obtenido es bajo e indica que la bondad del ajuste realizado es moderada o baja. Si se hubiera obtenido previamente el coeficiente de correlación lineal de Pearson r, su valor: $r = \sqrt{r^2} = 0{,}74$ indica que el grado de asociación lineal entre las variables es moderado/bajo; esto es, la relación lineal del número de trabajadores respecto de la facturación no es suficiente para explicar el número de trabajadores en las empresas del sector.

Para el cálculo de la varianza residual aunque no disponemos de los datos observados de las variables X e Y para obtener los residuos, es posible acudir a su relación con el coeficiente de determinación lineal en su expresión general: $r^2 = 1 - \frac{S_{res}^2}{S_Y^2}$

Desde ésta, se puede despejar:

$$S_{res}^2 = (1 - r^2) \times S_Y^2 = (1 - 0{,}56) \times 146 = 64{,}24$$

c) Para un valor de $x = 52$ (recordar que las variables están expresadas en miles de €) sustituyendo en la ecuación de la recta ajustada en el apartado a, se obtiene:

$$\hat{y}_{|x=52} = -0{,}775 + 0{,}825 \times 52 \cong 28$$

Ejercicio 2.

Se sabe que existe una relación lineal entre la *cantidad de lluvia* caída en el mes anterior a la recolección (X) en Dl. por m^2 y la *producción de sandías* (Y) en kg por m^2. En un estudio llevado a cabo en diferentes lugares del país, se han obtenido los siguientes resultados:

Y	40	44	46	48	52	58	60	68	74	80
X	6	10	12	14	16	18	22	24	26	32

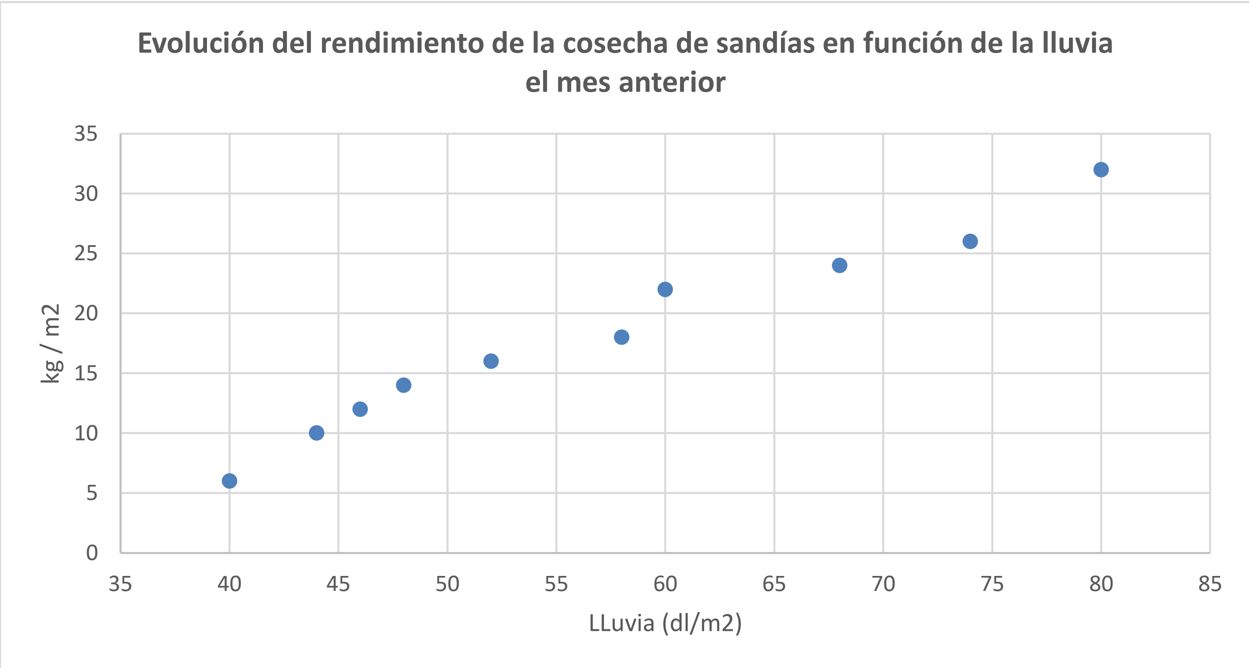
a) Representa gráficamente las variables mediante el diagrama de dispersión.
b) Calcula la covarianza y el coeficiente de correlación lineal ¿Son las variables incorreladas? ¿Existe una relación lineal directa, inversa o no la hay?

c) Obtén la recta de regresión mínimo cuadrática de la producción de sandías sobre la cantidad de lluvia caída en el mes anterior a la recolección.

d) Predice la producción de sandías si la lluvia caída ha sido de 23,5 Dl por m^2 ¿Es una predicción fiable?

SOLUCIÓN:

a) El diagrama de dispersión representa los pares de puntos en unos ejes coordenados como los que se recogen en el gráfico:



b) En este apartado se calcula la covarianza como medida del grado de correlación entre las variables analizadas mediante su fórmula abreviada. Se comienza calculando las medias marginales de ambas variables:

$$\bar{x} = \frac{1}{10}[6 + 10 + 12 + 14 + 16 + 18 + 22 + 24 + 26 + 32] = 18$$

$$\bar{y} = \frac{1}{10}[40 + 44 + 46 + 48 + 52 + 58 + 60 + 68 + 74 + 80] = 57$$

$$S_{X,Y} = \frac{1}{10}\sum_{i=1}^{10} x_i \times y_i - \bar{x} \times \bar{y} =$$

$$= \frac{1}{10}[40 \times 6 + 44 \times 10 + 46 \times 12 + 48 \times 14 + 52 \times 16 + 58 \times 18 + 60 \times 22 + 68 \times 24 + 74 \times 26 + 80 \times 32] - 18 \times 57 = 95{,}6$$

Para el cálculo del coeficiente de correlación lineal de Pearson se necesita calcular los valores de la varianza marginal de ambas variables.

$$S_X^2 = \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = \frac{1}{10}[6^2 + 10^2 + 12^2 + 14^2 + 16^2 + 18^2 + 22^2 + 24^2 + $$
$$+ 26^2 + 32^2] - 18^2 = 57{,}6$$

$$S_Y^2 = \frac{1}{10}\sum_{i=1}^{10} y_i^2 - \bar{y}^2 = \frac{1}{10}[40^2 + 44^2 + 46^2 + 48^2 + 52^2 + 58^2 + 60^2 + 68^2 + $$
$$+ 74^2 + 80^2] - 57^2 = 163{,}4$$

Llevando los valores obtenidos a la fórmula del coeficiente:

$$r = \frac{S_{X,Y}}{S_X \times S_Y} = \frac{95{,}6}{\sqrt{57{,}6} \times \sqrt{163{,}4}} = 0{,}9854$$

Del valor obtenido se concluye que las variables X e Y no son incorreladas: el grado de asociación es fuerte y la dependencia es directa por el signo positivo de la covarianza.

c) En este apartado ya disponemos de los estadísticos necesarios para obtener los coeficientes de la recta de mínimos cuadrados por lo que sustituyendo se obtienen sus expresiones:

$$\hat{b} = \frac{S_{X,Y}}{S_X^2} = \frac{95{,}6}{57{,}6} \cong 1{,}66 \qquad \hat{a} = \bar{y} - b \times \bar{x} = 57 - 1{,}66 \times 18 = 27{,}125$$

La ecuación de la recta que explica la producción de sandía (Kg/m^2) (Y) en función de la lluvia el mes anterior (Dl/m^2) (X) es: Y = 27,125 + 1,66 X

d) Para un mes con lluvia de 23,5 dl/m^2 la predicción en base al modelo de producción de sandía en el mes siguiente será:

$$\hat{y}|_{x=23{,}5} = 27{,}125 + 1{,}66 \times 23{,}5 = 66{,}13\ kg/m^2$$

La predicción obtenida es fiable ya que el valor del coeficiente de determinación indica que el ajuste es bueno y además el valor para el que predecimos está en el rango de los datos utilizados en la estimación. En otras palabras, se está interpolando en la ecuación del modelo.

Ejercicio 3.

A partir de las siguientes observaciones, obtenidas al estudiar la "inversión en I + D" (X) y los "beneficios" (Y), ambos en miles de euros, en una muestra de 8 empresas, se pide:

X	**12**	**22**	**7**	**23**	**19**	**10**	**13**	**17**
Y	115	149	90	151	142	108	127	135

a) Estima la ecuación potencial que explica Y en función de X: (1) $Y = aX^b$

b) Predice en base al modelo el beneficio que tendría una empresa con una inversión de 9000 euros ¿Es fiable? Evalúa los elementos necesarios para argumentar tu respuesta.

SOLUCIÓN:

a) Como cualquier otro modelo de regresión no lineal, el modelo potencial que se propone para la relación entre las variables se debe linealizar para transformarlo a un modelo lineal auxiliar. Tomando logaritmos en una base cualquiera (en particular logaritmo neperiano) en ambas partes del modelo $\ln Y = lna + blnX$

Definiendo las variables auxiliares $V = \ln Y$ y $U = \ln X$ la ecuación (1) se puede escribir como $V = A + BU$ (donde $A = lna$ y $b = B$) obteniendo una relación lineal auxiliar entre las variables V y U. Así la estimación de los parámetros del modelo potencial se realiza a partir de los valores de los coeficientes de la regresión lineal auxiliar, deshaciendo posteriormente el cambio de variable en la obtención de los coeficientes a y b.

Como en cualquier regresión lineal, las expresiones para los estimadores de mínimos cuadrados de los coeficientes A y B de la regresión V=A+BU vienen dadas por: $\hat{B} = \frac{S_{U,V}}{S_U^2}$ y $\hat{A} = \bar{v} - \hat{B} \times \bar{u}$

Se deben obtener, a partir de los datos iniciales, los valores de las variables U=lnX y V=lnY; éstos se recogen en la tabla inferior (se han mantenido 3 cifras decimales).

U	**2,485**	**3,091**	**1,946**	**3,135**	**2,944**	**2,303**	**2,565**	**2,833**
V	4,745	5,004	4,500	5,017	4,956	4,682	4,844	4,905

Se obtienen las medidas necesarias para el cálculo de la estimación de los coeficientes A y B:

$$\bar{u} = \frac{1}{8}(2{,}845 + 3{,}091 + 1{,}946 + 3{,}135 + 2{,}944 + 2{,}303 + 2{,}565 + 2{,}833) = 2{,}663$$
$$\bar{v} = \frac{1}{8}(4{,}745 + 5{,}004 + 4{,}500 + 5{,}017 + 4{,}956 + 4{,}682 + 4{,}844 + 4{,}905) = 4{,}832$$

$S_{U,V} = \frac{1}{8}(2{,}845 \times 4{,}745 + 3{,}091 \times 5{,}004 + 1{,}946 \times 4{,}500 + 3{,}135 \times 5{,}017 + 2{,}944 \times 4{,}956 + 2{,}303 \times 4{,}682 + 2{,}565 \times 4{,}844 + 2{,}833 \times 4{,}905)$-$2{,}663 \times 4{,}832$=$0{,}0644$

$$S_U^2 = \frac{1}{8}(2{,}845^2 + 3{,}091^2 + 1{,}946^2 + 3{,}135^2 + 2{,}944^2 + 2{,}303^2 + 2{,}565^2 + 2{,}833^2) - 2{,}663^2 = 0{,}1500$$

Con todos los estadísticos calculados, se sustituye en las expresiones de los estimadores:

$$\hat{B} = \frac{S_{U,V}}{S_U^2} = \frac{0{,}0644}{0{,}1500} \cong 0{,}429 \qquad \hat{A} = \bar{v} - \hat{B} \times \bar{u} = 4{,}832 - 0{,}429 \times 2{,}663 \cong 3{,}689$$

Identificando la relación entre los coeficientes de la regresión potencial A y B y los de la lineal a y b, $A = lna$ $B = b$ podemos obtener una estimación indirecta de a y b a partir de:

$$\hat{a} = e^{\hat{A}} = e^{3,689} \cong 40$$
$$\hat{b} = \hat{B} = 0,429$$

La ecuación del modelo potencial estimado para la dependencia del beneficio Y con la inversión en I+D (X) es:

$$Y = 40X^{0,43}$$

b) En base al modelo estimado en el apartado anterior, y sin más que sustituir en la ecuación para x= 9, se obtiene: $\hat{y}|_{x=9} = 40 \times 9^{0,43} = 102,89\ (en\ miles\ de\ euros)$

En principio, al tratarse de una interpolación la predicción obtenida sería fiable, siempre y cuando la bondad del ajuste sea elevada. Se debe entonces evaluar la calidad del ajuste mediante el coeficiente de determinación general R^2. Su expresión en un modelo no lineal viene dada por:

$$R^2 = 1 - \frac{ECM}{S_Y^2} \qquad donde\ ECM = \frac{1}{8}\sum_{i=1}^{8} e_i^2 = \frac{1}{8}\sum_{i=1}^{8}(y_i - \hat{y}_i)^2$$

Se obtienen mediante la ecuación del modelo estimado los valores teóricos del modelo recogidos en la tabla inferior (se han mantenido dos cifras decimales) y a partir de ellos los residuos o errores del modelo potencial como diferencia entre esos valores.

$\hat{\mathbf{Y}}$	116,44	151,11	92,35	154,03	141,88	107,66	120,52	135,25
Y	115	149	90	151	142	108	127	135
e	-1,44	-2,11	-2,35	-3,03	0,12	0,34	6,48	-0,25

$$ECM = \frac{1}{8}(-1,44^2 + -2,11^2 + -2,35^2 + -3,03^2 + 0,12^2 + 0,34^2 + 6,48^2 + -0,25^2)$$
$$= 7,93$$

Por otra parte, la variabilidad en los beneficios de las empresas S_Y^2 vale:

$$S_Y^2 = \frac{1}{8}(115^2 + 149^2 + 90^2 + 151^2 + 142^2 + 108^2 + 127^2 + 135^2) - 127,125^2 \cong 402,86$$

Llevando los valores obtenidos a la expresión del coeficiente de determinación general, su valor es:

$$R^2 = 1 - \frac{ECM}{S_Y^2} = 1 - \frac{7,93}{402,86} \cong 0,98$$

El elevado valor obtenido indica que los residuos son proporcionalmente pequeños y el ajuste es muy bueno. Con todo ello, la predicción realizada para una inversión en I+D de nueve mil euros, resulta bastante fiable.

Ejercicio 4.

Los directivos de una empresa multinacional de cosméticos quieren analizar la posible relación entre los Beneficios Anuales (Y) y el Gasto en Publicidad (X) para varios productos comercializados analizando los históricos recogidos en la tabla.

X (millones €)	Y (millones €)
2	32
2,8	35,2
3,4	37,2
4,3	38,9
5,2	39,5
6,1	40,3

Determina la regresión entre los Beneficios Anuales y los Gastos en Publicidad utilizando:

a) Una función lineal

b) Una función logarítmica

c) ¿Cuál de las dos funciones presenta un mejor ajuste a los datos? ¿Por qué?

SOLUCIÓN:

a) Como se ha ido viendo en los ejercicios anteriores, para obtener la ecuación de la recta de regresión se van a calcular las medias y varianzas marginales y la covarianza. Aplicando las correspondientes fórmulas de forma similar a la de los ejercicios anteriores, se obtienen los siguientes valores de los estadísticos:

$$\bar{x} = 4{,}05 \qquad \bar{y} = 36{,}8 \qquad S_{X,Y} = 3.205 \qquad S_X^2 = 1{,}6625 \qquad S_Y^2 = 7{,}087$$

Y a partir de éstos, se obtienen los coeficientes de la recta de regresión:

$$\hat{b} = \frac{S_{X,Y}}{S_X^2} = \frac{3{,}205}{1{,}6625} \cong 1{,}93 \qquad \hat{a} = \bar{y} - \hat{b} \times \bar{x} = 36{,}8 - 1{,}93 \times 4{,}05 \cong 30.07$$

La evaluación de la bondad del ajuste del modelo lineal se realiza mediante el coeficiente de determinación lineal. Sustituyendo los estadísticos en su expresión, se obtiene su valor:

$$r^2 = \frac{S_{XY}^2}{S_X^2 \times S_Y^2} = \frac{3{,}205^2}{1{,}6625 \times 7{,}087} = 0{,}87$$

Hay que señalar que, al tratarse de una medida invariante ante cambios de escala, se puede continuar trabajando con las unidades simplificadas.

b) Ajustaremos y evaluaremos un modelo alternativo al lineal: un modelo logarítmico en este apartado. El modelo logarítmico general es un modelo paramétrico cuya ecuación es: $Y = a + b \times lnX$ y para estimar los parámetros de este modelo se debe transformar a una ecuación lineal. Para ello, realizando el cambio de variable $U = lnX$ el modelo logarítmico se expresa en forma lineal entre las variables Y, U:

$$Y = a + b \times U$$

De este modo, utilizando las expresiones conocidas para la estimación de mínimos cuadrados de un modelo de regresión lineal, las expresiones para los estimadores de los parámetros del modelo logarítmico vendrán dadas por:

$$\hat{b} = \frac{S_{Y,U}}{S_U^2} \qquad \hat{a} = \bar{y} - \hat{b} \times \bar{u}$$

En la tabla se recogen los valores de las variables U, $u_i = \ln(x_i)$, e Y (se han redondeado a dos decimales para simplificar). Se trata ahora de obtener los estadísticos en ambas:

U (millones €)	**Y (millones €)**
0,92	**32**
1,03	**35,2**
1,22	**37,2**
1,46	**38,9**
1,65	**39,5**
1,80	**40,3**

Se recomienda al lector apoyarse en alguna herramienta tipo hoja de cálculo para facilitar la obtención. Los valores requeridos son:

$$\bar{u} = 1{,}35 \qquad \bar{y} = 36{,}8 \qquad S_{U,Y} = 3{,}205 \qquad S_U^2 = 1{,}6625 \qquad S_Y^2 = 7{,}087$$

Y los coeficientes del modelo:

$$\hat{b} = \frac{S_{Y,U}}{S_U^2} = 8{,}46 \qquad \hat{a} = \bar{y} - \hat{b} \times \bar{u} = 36{,}8 - 8{,}46 \times 1{,}35 = 25{,}41$$

La ecuación del modelo logarítmico ajustado para la relación entre los beneficios y los gastos realizados en publicidad queda:

$$Y = 25{,}41 + 8{,}46\, lnX$$

c) A fin de comparar la calidad de ambos ajustes (se denomina selección del modelo), la evaluación de la bondad del ajuste logarítmico requiere calcular los residuos y a partir de ellos el error cuadrático medio global del modelo. Para ellos, lanzando las predicciones o valores teóricos en base al modelo logarítmico, éstos, junto con los residuos y sus cuadrados se recogen en la tabla siguiente:

Y (millones €)	$\widehat{Y}$(millones €)	e_i	e_i^2
32	33,16	-1,06	1,13
35,2	34,12	1,08	1,16
37,2	35,77	0,63	0,40
38,9	37,75	-0,05	0,00
39,5	39,36	0,14	0,02
40,3	40,71	-0,81	0,66

El error cuadrático medio (que, en modelos no lineales no coincide con la varianza de los residuos), toma el valor:

$$ECM = \frac{1}{6}\sum_{i=1}^{6} e_i^2 = 0{,}56$$

Y el valor del coeficiente de determinación general para el ajuste logarítmico es:

$$R^2 = 1 - \frac{ECM}{S_Y^2} = 1 - \frac{0{,}56}{7{,}087} = 0{,}92$$

Resumiendo, el análisis de regresión realizado concluye que el mejor ajuste corresponde al modelo logarítmico debido a que presenta un menor error cuadrático medio, o en otras palabras residuos globalmente menores que los que se obtienen a partir de las predicciones en los datos conocidos que proporciona el modelo lineal.

Ejercicio 5.

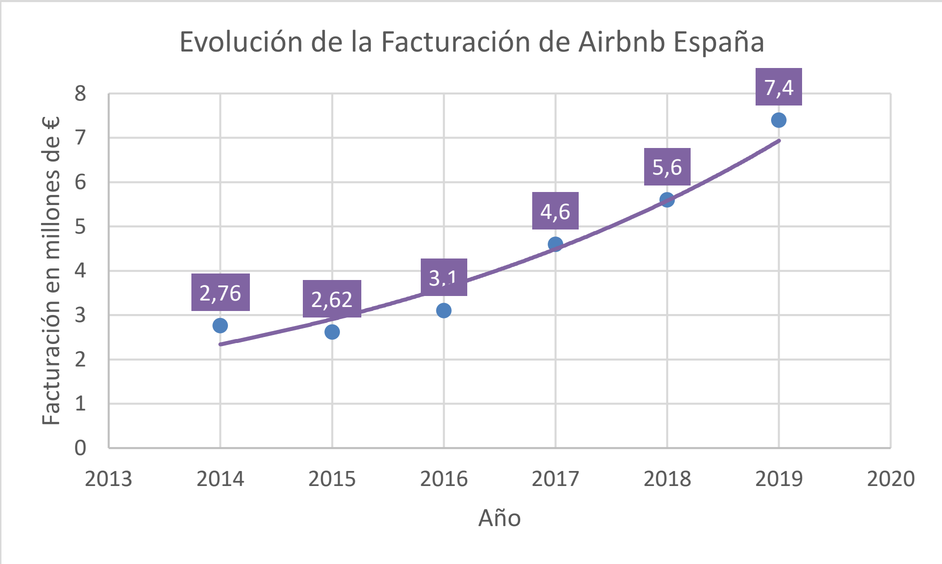
Los datos de facturación (en millones de €) de AirBnb en España (pre-pandemia) según la fuente Statista 2022 vienen recogidos en la tabla inferior. Se pide:

a) ¿Se puede establecer un crecimiento exponencial en la facturación? Representa el diagrama de dispersión y argumenta en base a la disposición de la nube de puntos.
b) En caso afirmativo, determina utilizando ajuste de regresión con respecto del tiempo el ritmo estimado en el crecimiento anual de la facturación y evalúa su adecuación a los datos observados.

Año	Facturación
2014	2,76
2015	2,62
2016	3,1
2017	4,6
2018	5,6
2019	7,4

SOLUCIÓN:

a) Se comienza como se sugiere en el enunciado elaborando el gráfico de dispersión para los datos de la evolución de la facturación de AirBnb en los últimos seis años. Cabe señalar que la variable explicativa en este estudio es el tiempo (los años); se continuará no obstante denotando por X a esta variable.



El gráfico de dispersión muestra un ritmo de crecimiento de la facturación superior al lineal y aunque es posible que se pudiera encontrar alguna función potencial de grado superior a 1 que se adaptase a la tendencia marcada por la nube de puntos, también podría creerse que un modelo exponencial se adaptaría para capturar la evolución permitiendo, además, establecer el ritmo de crecimiento de ésta. No obstante, como toda representación gráfica, la curva sugerida por el diagrama de dispersión es sólo un punto de partida que, en todo caso, deberá ser corroborada por medidas numéricas objetivas que permitan evaluar y concluir sobre la adecuación de un modelo de tipo exponencial para la evolución de la facturación.

b) La ecuación de un modelo exponencial general es $Y = Ae^{BX}$ y para el estudio que nos ocupa Y representa la facturación mientras que X representa el tiempo. Por sencillez de cálculo y porque resulta irrelevante para la operativa, es posible adoptar valores más simples, tales como *t=1,2,.,6* en lugar de los propios años del estudio. La estimación de un modelo exponencial concierne con la estimación de los parámetros o coeficientes del modelo, en particular B que representará la tasa (ritmo) de crecimiento exponencial de la facturación en el transcurso de los años.

Para la linealización de un modelo exponencial, resulta bastante natural, comenzar por adoptar logaritmos neperianos en ambas partes de la ecuación del modelo exponencial. De este modo, el modelo exponencial entre las variables es equivalente al modelo con ecuación:

$$\ln Y = lnA + BX$$

Inmediatamente, y mediante un único y sencillo cambio de variable de tipo $V = lnY$ se identifica la existencia de una relación lineal entre las variables V y X :

$$V = a + bX \qquad dónde \qquad a = lnA \quad B = b$$

Será esta la regresión lineal auxiliar que va a permitir estimar indirectamente los coeficientes A y B del modelo exponencial general.

Las expresiones de los coeficientes de la regresión lineal (adaptada a las variables involucradas) será:

$$\hat{b} = \frac{S_{V,X}}{S_X^2} \qquad \hat{a} = \bar{v} - \hat{b} \times \bar{x}$$

Se deben obtener los datos transformados (variable V) y entonces las medidas descriptivas requeridas entre estas variables:

X=t	1	2	3	4	5	6
Y	2,76	2,62	3,1	4,6	5,6	7,4
V	1,015	0,963	1,131	1,526	1,723	2,001

$$\bar{v} = \frac{1}{8}(1{,}015 + 0{,}963 + 1{,}131 + 1{,}526 + 1{,}723 + 2{,}001) = 1{,}393$$

$$\bar{x} = \frac{1}{8}(1 + 2 + 3 + 4 + 5 + 6) = 3{,}5$$

$$S_{V,X} = \frac{1}{8}(1 \times 1{,}015 + 2 \times 0{,}963 + 3 \times 1{,}131 + 4 \times 1{,}526 + 5 \times 1{,}723 + 6 \times 2{,}001) - 1{,}180 \times 4{,}5 \cong 0{,}634$$

$$S_X^2 = \frac{1}{8}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - 3{,}5^2 \cong 2{,}92$$

Sustituyendo:

$$\hat{b} = \frac{S_{V,X}}{S_X^2} = \frac{0{,}634}{2{,}92} \cong 0{,}217 \qquad \hat{a} = 1{,}393 - 0{,}217 \times 3{,}5 \cong 0{,}633$$

Finalmente, llevados a los coeficientes del modelo exponencial:

$$A = e^a \cong 1{,}883 \quad B = b = 0{,}217$$

El mejor ajuste o modelo exponencial que más se adecúa globalmente a los datos observados, tiene como ecuación:

$$Y = 1{,}833e^{0{,}217X} \ (1)$$

No obstante, resulta conveniente (y así se pide en el enunciado) valorar o evaluar la adecuación del modelo estimado; en términos estadísticos se requiere evaluar la bondad de ajuste mediante el cálculo de alguna medida global de la calidad predictiva del modelo ajustado. La medida de bondad de ajuste en modelos no lineales es el coeficiente de determinación general; se va a calcular a continuación a partir de los valores teóricos (o pronósticos) del modelo, sus errores o residuos y el error cuadrático medio (*ECM*) como medida global del tamaño de los residuos.

Los pronósticos del modelo se obtiene por sustitución en la ecuación (1). Así por ejemplo, para el primer año (2014) t=1 el pronóstico será → $\hat{y} = 1{,}833e^{0{,}217} = 2{,}34$ y su residuo:

$$e = y - \hat{y} = 2{,}76 - 2{,}340 = 0{,}42$$

En la tabla se recogen los pronósticos y residuos así calculados para cada uno de los 6 años:

$\hat{Y}$	2,340	2,907	3,613	4,490	5,579	6,934
e	0,420	-0,287	-0,513	0,110	0,021	0,466

Y el valor del error cuadrático medio es:

$$ECM = \frac{1}{8}(0{,}420^2 + -0{,}287^2 + -0{,}513^2 + 0{,}110^2 + 0{,}021^2 + 0{,}466^2) = 0{,}7525$$

Por último, se valora el tamaño del error cuadrático medio en términos relativos, obteniendo el coeficiente de determinación general (requiere previamente obtener la variabilidad en la facturación, S_Y^2)

$$\bar{y} = \frac{1}{8}(2{,}76 + 2{,}62 + 3{,}1 + 4{,}6 + 5{,}6 + 7{,}4) = 4{,}347$$

$$S_Y^2 = \frac{1}{8}(2{,}76^2 + 2{,}62^2 + 3{,}1^2 + 4{,}6^2 + 5{,}6^2 + 7{,}4^2) - 4{,}347^2 = 3{,}002$$

$$R^2 = 1 - \frac{ECM}{S_Y^2} = 1 - \frac{0{,}7525}{3{,}002} \cong 0{,}75$$

El valor obtenido del coeficiente de determinación lineal (en la frontera adoptada usualmente) refleja que la bondad del ajuste exponencial es moderada/alta, por lo que podría concluirse que la relación exponencial es admisible; No obstante, sería recomendable testar y re-estimar el modelo incorporando los nuevos resultados anuales de facturación con el fin de, por un parte, disponer de un mayor número de datos con los que validar el modelo y por otra, observar si éstos contribuyen a una mejor adecuación de un modelo de tipo exponencial, o si por el contrario es posible determinar algún otro tipo de función de crecimiento más suave (posiblemente potencial). Se propone como ejercicio final para el lector el de ajustar y evaluar la calidad predictiva de un modelo de tipo potencial para la evolución de la facturación de AirBnb.