4

Regresión Lineal Simple y Correlación

4.1. Fundamentos teóricos

4.1.1. Regresión

La *regresión* es la parte de la estadística que trata de determinar la posible relación entre una variable numérica Y, que suele llamarse *variable dependiente*, y otro conjunto de variables numéricas, $X_1, X_2, \dots, X_n$, conocidas como *variables independientes*, de una misma población. Dicha relación se refleja mediante un modelo funcional $y = f(x_1, \dots, x_n)$.
El caso más sencillo se da cuando sólo hay una variable independiente X, y entonces se habla de *regresión simple*. En este caso el modelo que explica la relación entre X e Y es una función de una variable $y = f(x)$.
Dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Familia de curvas	Ecuación genérica
Lineal	$y = b_0 + b_1 x$
Cuadrática	$y = b_0 + b_1 x + b_2 x^2$
Cúbica	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Potencia	$y = b_0 \cdot x^{b_1}$
Exponencial	$y = b_0 \cdot e^{b_1 x}$
Logarítmica	$y = b_0 + b_1 \ln x$
Inversa	$y = b_0 + \frac{b_1}{x}$
Compuesto	$y = b_0 b_1^x$
Crecimiento	$y = e^{b_0 + b_1 x}$
G (Curva-S)	$y = e^{b_0 + \frac{b_1}{x}}$

Para elegir un tipo de modelo u otro, se suele representar el *diagrama de dispersión* , que consiste en dibujar sobre unos ejes cartesianos correspondientes a las variables X e Y, los pares de valores (x_i, y_j) observados en cada individuo de la muestra.

Ejemplo En la figura 4.1 aparece el diagrama de dispersión correspondiente a una muestra de 30 individuos en los que se ha medido la estatura en cm (X) y el peso en kg (Y). En este caso la forma de la nube de puntos refleja una relación lineal entre la estatura y el peso.

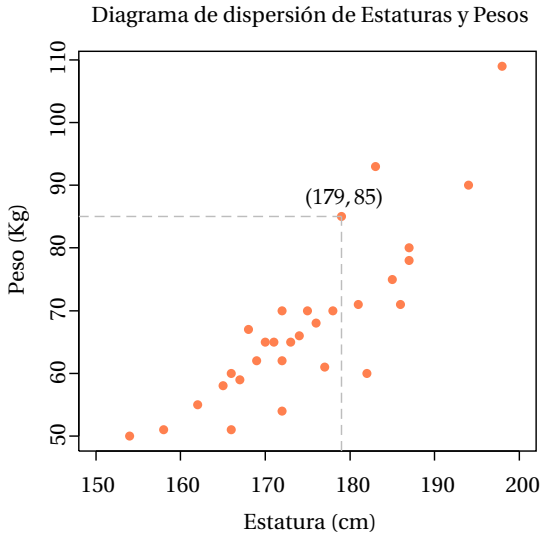
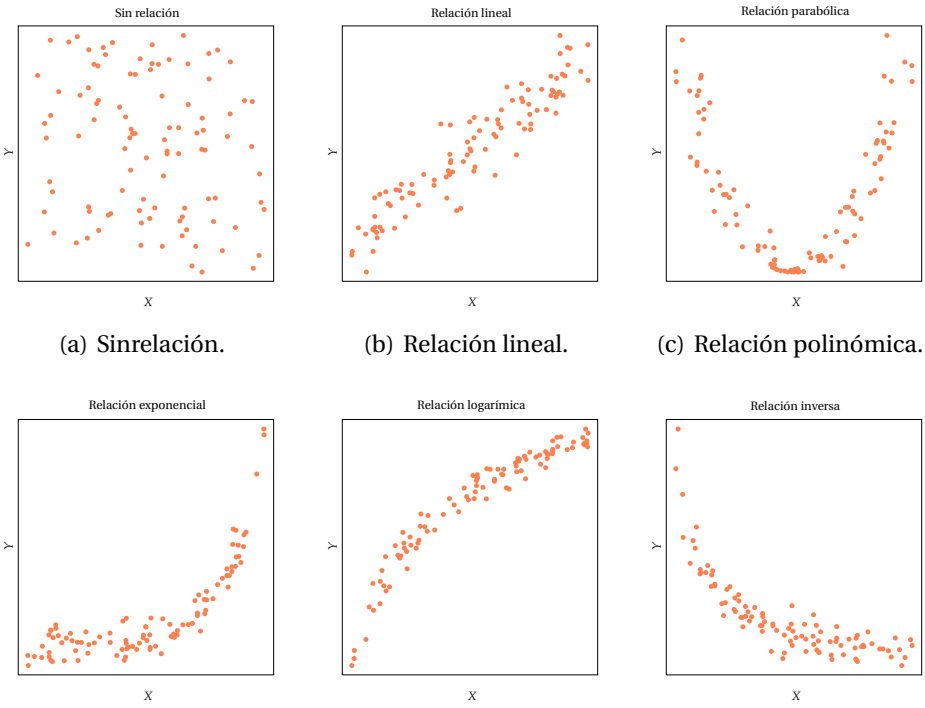


Figura 4.1: Diagrama de dispersión. El punto (179,85) indicado corresponde a un individuo de la muestra que mide 179 cm y pesa 85 Kg.

Según la forma de la nube de puntos del diagrama, se elige el modelo más apropiado (figura 4.2), y se determinan los parámetros de dicho modelo para que la función resultante se ajuste lo mejor posible a la nube de puntos.

El criterio que suele utilizarse para obtener la función óptima, es que la distancia de cada punto a la curva, medida en el eje Y, sea lo menor posible. A estas distancias se les llama *residuos* o *errores* en Y (figura 4.3). La función



(a) Sinrelación. (b) Relación lineal. (c) Relación polinómica.

(d) Relación exponencial. (e) Relación logarítmica. (f) Relación inversa.

Figura 4.2: Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

que mejor se ajusta a la nube de puntos será, por tanto, aquella que hace mínima la suma de los cuadrados de los residuos. [*]

Rectas de regresión

En el caso de que la nube de puntos tenga forma lineal y optemos por explicar la relación entre X e Y mediante una recta $y = a + bx$, los parámetros a determinar son a (punto de corte con el eje de ordenadas) y b (pendiente de la recta). Los valores de estos parámetros que hacen mínima la suma de residuos al cuadrado, determinan la recta óptima. Esta recta se conoce como *recta de regresión de* Y *sobre* X y explica la variable Y en función de la

[*]El cuadrado es para evitar que se compensen los residuos positivos con los negativos.

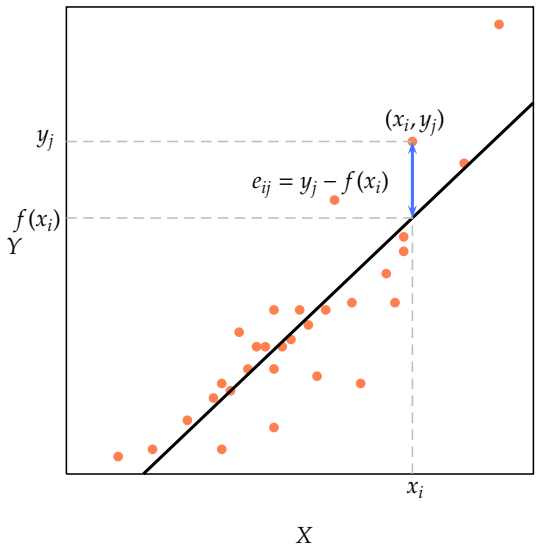


Figura 4.3: Residuos o errores en Y. El residuo correspondiente a un punto (x_i, y_j) es la diferencia entre el valor y_j observado en la muestra, y el valor teórico del modelo $f(x_i)$, es decir, $e_{ij} = y_j - f(x_i)$.

variable X. Su ecuación es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

donde s_{xy} es un estadístico llamado *covarianza* que mide el grado de relación lineal, y cuya fórmula es

$$s_{xy} = \frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}$$

Ejemplo En la figura 4.4 aparecen las rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura del ejemplo anterior.

La pendiente de la recta de regresión de Y sobre X se conoce como *coeficiente de regresión de Y sobre X*, y mide el incremento que sufrirá la variable Y por cada unidad que se incremente la variable X, según la recta.
Cuanto más pequeños sean los residuos, en valor absoluto, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación

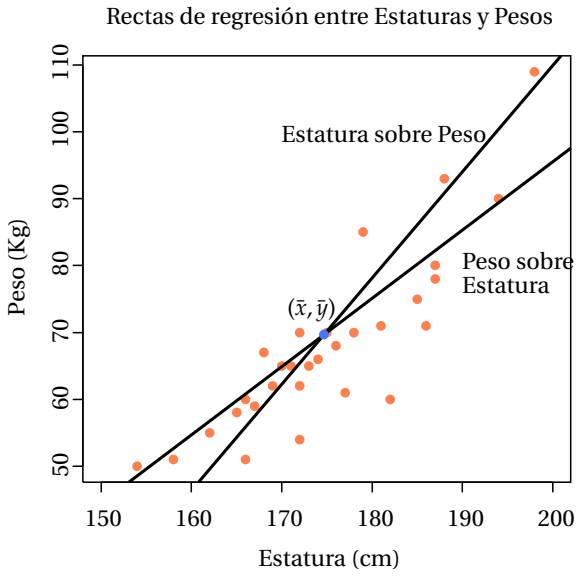
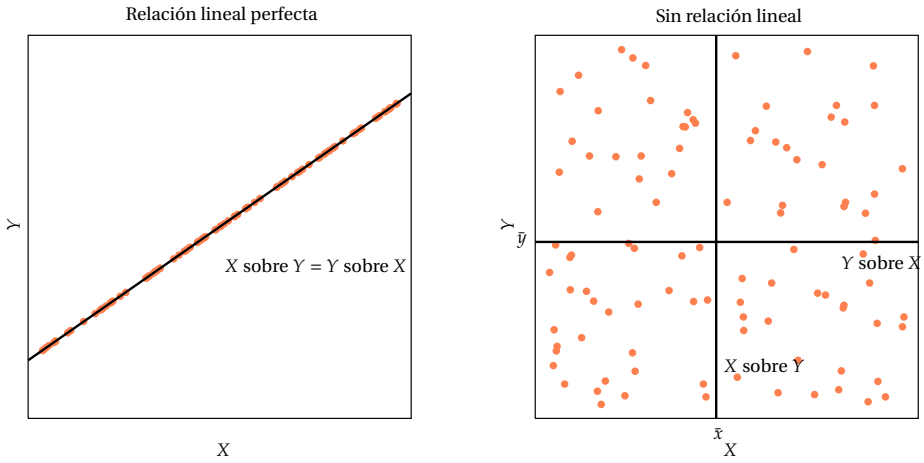


Figura 4.4: Rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura. Las rectas de regresión siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$

entre X e Y. Cuando todos los residuos son nulos, la recta pasa por todos los puntos de la nube, y la relación es perfecta. En este caso ambas rectas, la de Y sobre X y la de X sobre Y coinciden (figura 4.5(a)).
Por contra, cuando no existe relación lineal entre las variables, la recta de regresión de Y sobre X tiene pendiente nula, y por tanto la ecuación es $y = \bar{y}$, en la que, efectivamente no aparece x, o $x = \bar{x}$ en el caso de la recta de regresión X sobre Y, de manera que ambas rectas se cortan perpendicularmente (figura 4.5(b)).

4.1.2. Correlación

El principal objetivo de la regresión simple es construir un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables X (variable independiente) e Y (variable dependiente) medidas en una misma muestra. Generalmente, el modelo construido se utiliza para realizar inferencias predictivas de Y en función de X en el resto de la población. Pero aunque la regresión garantiza que el modelo construido es el mejor posible,



(a) Dependencia funcional lineal.

(b) Independencia lineal.

Figura 4.5: Distintos grados de dependencia. En el primer caso, la relación es perfecta y los residuos son nulos. En el segundo caso no existe relación lineal y la pendiente de la recta es nula.

dentro del tipo de modelo elegido (lineal, polinómico, exponencial, logarítmico, etc.), puede que aún así, no sea un buen modelo para hacer predicciones, precisamente porque no haya relación de ese tipo entre X e Y. Así pues, con el fin de validar un modelo para realizar predicciones fiables, se necesitan medidas que nos hablen del grado de dependencia entre X e Y, con respecto a un modelo de regresión construido. Estas medidas se conocen como medidas de *correlación*.

Dependiendo del tipo de modelo ajustado, habrá distintos tipos de medidas de correlación. Así, si el modelo de regresión construido es una recta, hablaremos de correlación lineal; si es un polinomio, hablaremos de correlación polinómica; si es una función exponencial, hablaremos de correlación exponencial, etc. En cualquier caso, estas medidas nos hablarán de lo bueno que es el modelo construido, y como consecuencia, de si podemos fiarnos de las predicciones realizadas con dicho modelo.

La mayoría de las medidas de correlación surgen del estudio de los residuos o errores en Y, que son las distancias de los puntos del diagrama de dispersión a la curva de regresión construida, medidas en el eje Y, tal y como se

muestra en la figura (4.3). Estas distancias, son en realidad, los errores predictivos del modelo sobre los propios valores de la muestra.
Cuanto más pequeños sean los residuos, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la curva de regresión pasa por todos los puntos de la nube, y entonces se dice que la relación es perfecta, o bien que existe una dependencia funcional entre X e Y (figura 4.5(a)). Por contra, cuando los residuos sean grandes, el modelo no explicará bien la relación entre X e Y, y por tanto, sus predicciones no serán fiables (figura 4.5(b)).

Varianza residual

Una primera medida de correlación, construida a partir de los residuos es la *varianza residual*, que se define como el promedio de los residuos al cuadrado

$$s_{ry}^2 = \frac{\sum_{i,j} e_{ij}^2 n_{ij}}{n} = \frac{\sum_{i,j} (y_j - f(x_i))^2 n_{ij}}{n}$$

Cuando los residuos son nulos, entonces $s_{ry}^2 = 0$ y eso indica que hay dependencia funcional. Por otro lado, cuando las variables son independientes, con respecto al modelo de regresión ajustado, entonces los residuos se convierten en las desviaciones de los valores de Y con respecto a su media, y se cumple que $s_{ry}^2 = s_y^2$. Así pues, se cumple que

$$0 \leq s_{ry}^2 \leq s_y^2$$

Según esto, cuanto menor sea la varianza residual, mayor será la dependencia entre X e Y, de acuerdo al modelo ajustado. No obstante, la varianza tiene como unidades las unidades de Y al cuadrado, y eso dificulta su interpretación.

Coeficiente de determinación

Puesto que el valor máximo que puede tomar la varianza residual es la varianza de Y, se puede definir fácilmente un coeficiente a partir de la comparación de ambas medidas. Surge así el *coeficiente de determinación* que se define como

$$R^2 = 1 - \frac{s_{ry}^2}{s_y^2}$$

Se cumple que

$$0 \leq R^2 \leq 1$$

y además no tiene unidades, por lo que es más fácil de interpretar que la varianza residual:

- $R^2 = 0$ indica que existe independencia según el tipo de relación planteada por el modelo de regresión.
- $R^2 = 1$ indica dependencia funcional.

Por tanto, cuanto mayor sea R^2, mejor será el modelo de regresión.
Si multiplicamos el coeficiente de determinación por 100, se obtiene el porcentaje de variabilidad de Y que explica el modelo de regresión. El porcentaje restante corresponde a la variabilidad que queda por explicar y se corresponde con el error predictivo del modelo. Así, por ejemplo, si tenemos un coeficiente de determinación $R^2 = 0{,}5$, el modelo de regresión explicaría la mitad de la variabilidad de Y, y en consecuencia, si se utiliza dicho modelo para hacer predicciones, estas tendrían la mitad de error que si no se utilizase, y se tomase como valor de la predicción el valor de la media de Y.

Coeficiente de determinación lineal

En el caso de que el modelo de regresión sea lineal, la fórmula del coeficiente de determinación se simplifica y se convierte en

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

que se conoce como *coeficiente de determinación lineal.*

Coeficiente de correlación

Otra medida de dependencia bastante habitual es el *coeficiente de correlación*, que se define como la raíz cuadrada del coeficiente de determinación:

$$R = \pm\sqrt{1 - \frac{s_{ry}^2}{s_y^2}}$$

tomando la raíz del mismo signo que la covarianza.
La única ventaja del coeficiente de correlación con respecto al coeficiente de determinación, es que tiene signo, y por tanto, además del grado de dependencia entre X e Y, también nos habla de si la relación es directa (signo +) o inversa (signo -). Su interpretación es:

- $R = 0$ indica independencia con respecto al tipo de relación planteada por el modelo de regresión.
- $R = -1$ indica dependencia funcional inversa.
- $R = 1$ indica dependencia funcional directa.

Por consiguiente, cuanto más próximo esté a -1 o a 1, mejor será el modelo de regresión.
Coeficiente de correlación lineal Al igual que ocurría con el coeficiente de determinación, cuando el modelo de regresión es lineal, la fórmula del coeficiente de correlación se convierte en

$$r = \frac{s_{xy}}{s_x s_y}$$

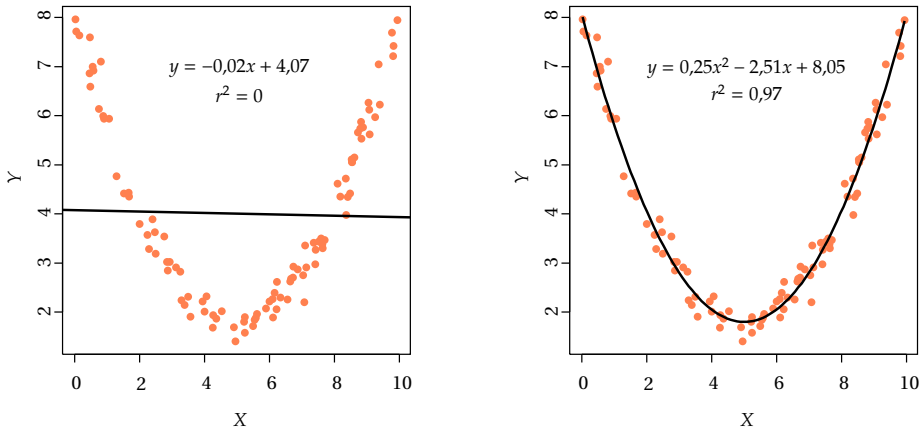
y se llama *coeficiente de correlación lineal.*
Por último, conviene remarcar que un coeficiente de determinación o de correlación nulo, indica que hay independencia según el modelo de regresión construido, pero puede haber dependencia de otro tipo. Esto se ve claramente en el ejemplo de la figura 4.6.

Fiabilidad de las predicciones

Aunque el coeficiente de determinación o de correlación nos hablan de la bondad de un modelo de regresión, no es el único dato que hay que tener en cuenta a la hora de hacer predicciones.
La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuando mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variablidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones del modelo.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Además, hay que tener en cuenta que un modelo de regresión es válido para el rango de valores observados en la muestra, pero fuera de ese rango no tenemos información del tipo de relación entre las variables, por lo que no deberíamos hacer predicciones para valores que estén lejos de los observados en la muestra.



(a) Dependencia lineal débil.

(b) Dependencia parabólica fuerte.

Figura 4.6: En la figura de la izquierda se ha ajustado un modelo lineal y se ha obtenido un $R^2 = 0$, lo que indica que el modelo no explica nada de la relación entre X e Y, pero no podemos afirmar que X e Y son independientes. De hecho, en la figura de la derecha se observa que al ajustar un modelo parabólico, $R^2 = 0{,}97$, lo que indica que casi hay una dependencia funcional parabólica entre X e Y.

4.2. Ejercicios resueltos

1. Se han medido dos variables A y B en 10 individuos obteniendo los siguientes resultados:

A	0	1	2	3	4	5	6	7	8	9
B	2	5	8	11	14	17	20	23	26	29

Se pide:

a) Crear las variables A y B e introducir estos datos.

b) Dibujar el diagrama de dispersión correspondiente.

i

1) Seleccionar el menú `Gráficos` ‣ `Cuadros de diálogo antiguos` ‣ `Dispersión/Puntos...`, elegir la opción Dispersión simple y hacer click sobre el botón Definir.

2) Seleccionar la variable B en el campo Eje Y del cuadro de diálogo.
3) Seleccionar la variable A en el campo Eje X del cuadro de diálogo y hacer click sobre el botón Aceptar.

En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre B y A?

c) Calcular la recta de regresión de B sobre A.

i

1) Seleccionar el menú `Analizar`▸`Regresión`▸`Lineales...`.
2) Seleccionar la variable B en el campo Dependientes del cuadro de diálogo.
3) Seleccionar la variable A en el campo Independientes del cuadro de diálogo y hacer click sobre el botón Aceptar.
4) Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada `Coeficientes`, y en la columna `B` de los `Coeficientes no estandarizados`, encontramos en la primera fila la `constante` de la recta y en la segunda la `pendiente`.

d) Dibujar dicha recta sobre el diagrama de dispersión.

i

1) Editar el gráfico realizado anteriormente haciendo un doble click sobre él.
2) Seleccionar los puntos haciendo click sobre alguno de ellos.
3) Seleccionar el menú `Elementos`▸`Linea de ajuste total` (También se podría usar en lugar del menu, la barra de herramientas)
4) Cerrar la ventana Propiedades.
5) Cerrar el editor de gráficos, cerrando la ventana.

e) Calcular la recta de regresión de A sobre B y dibujarla sobre el correspondiente diagrama de dispersión.

i Repetir los pasos de los apartados anteriores pero escogiendo como variable Dependiente la variable A, y como variable Independiente la variable B.

f) ¿Son grandes los residuos? Comentar los resultados.

2. En una licenciatura se quiere estudiar la relación entre el número medio de horas de estudio diarias y el número de asignaturas suspensas. Para ello se obtuvo la siguiente muestra:

Horas	Suspensos	Horas	Suspensos	Horas	Suspensos
3,5	1	2,2	2	1,3	4
0,6	5	3,3	0	3,1	0
2,8	1	1,7	3	2,3	2
2,5	3	1,1	3	3,2	2
2,6	1	2,0	3	0,9	4
3,9	0	3,5	0	1,7	2
1,5	3	2,1	2	0,2	5
0,7	3	1,8	2	2,9	1
3,6	1	1,1	4	1,0	3
3,7	1	0,7	4	2,3	2

Se pide:

a) Crear las variables horas y suspensos e introducir estos datos.

b) Calcular la recta de regresión de suspensos sobre horas y dibujarla.

i
1) Seleccionar el menú `Analizar`▸`Regresion`▸`Lineales...`.
2) Seleccionar la variable suspensos en el campo Dependientes del cuadro de diálogo.
3) Seleccionar la variable horas en el campo Independientes del cuadro de diálogo y hacer click sobre el botón Aceptar.
4) Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada `Coeficientes`, y en la columna `B` de los `Coeficientes no estandarizados`, encontramos en la primera fila la `constante` de la recta y en la segunda la `pendiente`.

5) Seleccionar el menú `Gráficos` ‣ `Cuadros de diálogo antiguos` ‣ `Dispersión/Puntos...`, elegir la opción Dispersión simple y hacer click sobre el botón Definir.
6) Seleccionar la variable suspensos en el campo Eje Y del cuadro de diálogo.
7) Seleccionar la variable horas en el campo Eje X del cuadro de diálogo y hacer click sobre el botón Aceptar.
8) Editar el gráfico realizado haciendo un doble click sobre él.
9) Seleccionar los puntos haciendo click sobre alguno de ellos.
10) Seleccionar el menú `Elementos` ‣ `Linea de ajuste total` (También se podría usar en lugar del menu, la barra de herramientas)
11) Cerrar la ventana Propiedades.
12) Cerrar el editor de gráficos, cerrando la ventana.

c) Indicar el coeficiente de regresión de suspensos sobre horas. ¿Cómo lo interpretarías?

i El coeficiente de regresión es la pendiente de la recta de regresión, que este caso vale −1,23 e indica que por cada hora de estudio adicional se obtienen 1,23 suspensos menos.

d) La relación lineal entre estas dos variables, ¿es mejor o peor que la del ejercicio anterior? Comentar los resultados a partir las gráficas de las rectas de regresión y sus residuos.

i La relación lineal entre estas dos variables es peor que la del ejercicio anterior, pues en este caso hay residuos.

e) Calcular los coeficientes de correlación y de determinación lineal. ¿Es un buen modelo la recta de regresión? ¿Qué porcentaje de la variabilidad del número de suspensos está explicada por el modelo?

i Observaremos en la ventana de resultados obtenida la tabla denominada `Resumen del modelo`, y en ella encontramos los valores

del coeficiente de correlación lineal `R` y del coeficiente de determinación lineal `R cuadrado`.

f) Utilizar la recta de regresión para predecir el número de suspensos correspondiente a 3 horas de estudio diarias. ¿Es fiable esta predicción?

i

1) Crear una nueva variable valores e introducir los valores de las horas de estudio para los que queremos predecir.
2) Seleccionar el menú `Transformar`▸`Calcular variable...`
3) Introducir el nombre de la nueva variable prediccion en el campo Variable de destino del cuadro de diálogo.
4) Introducir la ecuación de la recta en el campo Expresión numérica, utilizando los coeficientes calculados anteriormente y la variable valores y hacer click sobre el botón Aceptar.

g) Según el modelo lineal, ¿cuántas horas diarias tendrá que estudiar como mínimo un alumno si quiere aprobarlo todo?.

i

Seguir los mismos pasos de los apartados anteriores, pero escogiendo como variable dependiente horas, y como independiente suspensos.

3. Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180	210
Concentración (gramos/litro)	1,6	1,7	1,5	1,1	0,7	0,2	2,1

Se pide:

a) Crear las variables tiempo y alcohol e introducir estos datos.

b) Calcular el coeficiente de correlación lineal e interpretarlo.

i

1) Seleccionar el menú `Analizar`▸`Correlaciones`▸`Bivariadas...`
2) Seleccionar ambas variables en el campo Variables del cuadro de diálogo y hacer click sobre el botón Aceptar.

c) Dibujar el diagrama de dispersión junto con la recta ajustada correspondiente a alcohol sobre tiempo. ¿Existe algún individuo con un residuo demasiado grande? Si es así, eliminar dicho individuo de la muestra y volver a calcular el coeficiente de correlación. ¿Ha mejorado el modelo?

i

1) Seleccionar el menú `Gráficos` ‣ `Cuadros de diálogo antiguos` ‣ `Dispersión/Puntos...`, elegir la opción Dispersión simple y hacer click sobre el botón Definir.
2) Seleccionar la variable alcohol en el campo Eje Y del cuadro de diálogo.
3) Seleccionar la variable tiempo en el campo Eje X del cuadro de diálogo y hacer click sobre el botón Aceptar.
4) Editar el gráfico realizado anteriormente haciendo un doble click sobre él.
5) Seleccionar los puntos haciendo click sobre alguno de ellos.
6) Seleccionar el menú `Elementos` ‣ `Linea de ajuste total` (También se podría usar en lugar del menu, la barra de herramientas)
7) Cerrar la ventana Propiedades.
8) Cerrar el editor de gráficos, cerrando la ventana.
9) Si existe algún individuo con un residuo demasiado grande, ir a la ventana del `Editor de datos`, y eliminarlo.
10) Repetir los pasos del apartado anterior.

d) Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0,5 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley? ¿Es fiable esta predicción?

i

1) Seleccionar el menú `Analizar` ‣ `Regresión` ‣ `Lineales...`.
2) Seleccionar la variable tiempo en el campo Dependientes del cuadro de diálogo.
3) Seleccionar la variable alcohol en el campo Independientes del cuadro de diálogo y hacer click sobre el botón Aceptar.

4) Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada `Coeficientes`, y en la columna `B` de los `Coeficientes no estandarizados`, encontramos en la primera fila la `constante` de la recta y en la segunda la `pendiente`.
5) Crear una nueva variable valores e introducir los valores que queremos estudiar.
6) Seleccionar el menú `Transformar`▸`Calcular variable...`.
7) Introducir el nombre de la nueva variable prediccion en el campo Variable de destino del cuadro de diálogo.
8) Introducir la ecuación de la recta en el campo Expresión numérica, utilizando los coeficientes citados anteriormente y la variable valores y hacer click sobre el botón Aceptar.

4.3. Ejercicios propuestos

1. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

   Se desea calcular:

   a) La relación fundamental (recta de regresión) entre actividad restante y tiempo transcurrido.
   b) ¿En qué porcentaje disminuye la actividad cada año que pasa?
   c) ¿Cuándo tiempo debe pasar para que el fármaco tenga una actividad del 80 %? ¿Cuándo será nula la actividad? ¿Son igualmente fiables estas predicciones?

2. Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días.

Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 5 días. Se pide:

a) Calcular la recta de regresión del tiempo de curación con respecto a la dosis suministrada.

b) Calcular los coeficientes de regresión. Interpretar los resultados.

c) Determinar el tiempo esperado de curación para una dosis de 5 mg diarios. ¿Es fiable esta predicción?

d) ¿Qué dosis debe aplicarse si queremos que el paciente tarde 4 días en curarse? ¿Es fiable la predicción?