Capítulo 9. Regresión lineal simple

9.1 Introducción

Uno de los aspectos más relevantes de la Estadística es el análisis de la relación o dependencia entre variables. Frecuentemente resulta de interés conocer el efecto que una o varias variables pueden causar sobre otra, e incluso predecir en mayor o menor grado valores en una variable a partir de otra. Por ejemplo, supongamos que la altura de los padres influyen significativamente en la de los hijos. Podríamos estar interesados en estimar la altura media de los hijos cuyos padres presentan una determinada estatura.

Los métodos de regresión estudian la construcción de modelos para explicar o representar la dependencia entre una *variable respuesta o dependiente* (Y) y la(s) *variable(s) explicativa(s) o dependiente(s)*, X . En este Tema abordaremos el modelo de regresión lineal, que tiene lugar cuando la dependencia es de tipo lineal, y daremos respuesta a dos cuestiones básicas:

- ¿Es significativo el efecto que una variable X causa sobre otra Y? ¿Es significativa la dependencia lineal entre esas dos variables?.
- De ser así, utilizaremos el modelo de regresión lineal simple para explicar y predecir la variable dependiente (Y) a partir de valores observados en la independiente (X).

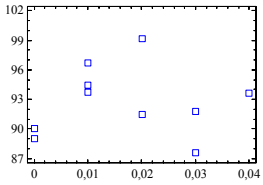
Ejemplo 9.1. El inventor de un nuevo material aislante quiere determinar la magnitud de la compresión (Y) que se producirá en una pieza de 2 pulgadas de espesor cuando se somete a diferentes cantidades de presión (X). Para ello prueba 5 piezas de material bajo diferentes presiones. Los pares de valores observados (x, y) se muestran en la siguiente tabla:

Pieza	Presión (x)	Compresión (y)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

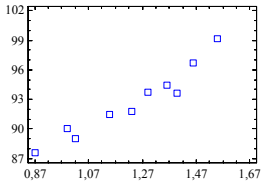
En principio no sabemos si las variables en cuestión están relacionadas o no, o si en caso de haber dependencia es significativa o no. De haber entre ellas una dependencia lineal significativa, podríamos expresar la Compresión (Y) a partir de la Presión (X) mediante una recta, y a partir de ella predecir la compresión que se daría para un determinado nivel de presión.

Una forma de determinar si puede existir o no dependencia entre variables, y en caso de haberla deducir de qué tipo puede ser, es gráficamente representando los pares de valores observados. A dicho gráfico se le llama *nube de puntos o diagrama de dispersión.*

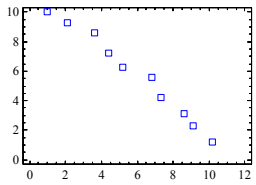
Ejemplos de casos que podrían darse:



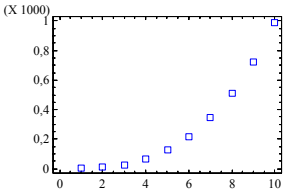
a)



b)



c)



d)

En a) hay ausencia de relación (independencia).

En b) existe asociación lineal positiva (varían en general en el mismo sentido).

En c) existe asociación lineal negativa (varían en sentido contrario).

En d) existe fuerte asociación, pero no lineal.

9.2 El modelo de regresión lineal

La estructura del modelo de regresión lineal es la siguiente:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

En esta expresión estamos admitiendo que todos los factores o causas que influyen en la variable respuesta Y pueden dividirse en dos grupos: el primero contiene a una variable explicativa X y el segundo incluye un conjunto amplio de factores no controlados que englobaremos bajo el nombre de *perturbación o error aleatorio*, ε, que provoca que la dependencia entre las variables dependiente e

independiente no sea perfecta, sino que esté sujeta a incertidumbre. Por ejemplo, en el consumo de gasolina de un vehículo (Y) influyen la velocidad (X) y una serie de factores como el efecto conductor, el tipo de carretera, las condiciones ambientales, etc, que quedarían englobados en el error.

Lo que en primer lugar sería deseable en un modelo de regresión es que estos errores aleatorios sean en media cero para cualquier valor x de X, es decir, $E[\varepsilon/X = x] = E[\varepsilon] = 0$, y por lo tanto:

$$E[Y/X = x] = \beta_0 + \beta_1 x + E[\varepsilon/X = x] = \beta_0 + \beta_1 x$$

En dicha expresión se observa que:

- La media de Y, para un valor fijo x, varía linealmente con x.
- Para un valor x se predice un valor en Y dado por $\hat{y} = E[Y/X = x] = \beta_0 + \beta_1 x$, por lo que el modelo de predicción puede expresarse también como $\hat{Y} = \beta_0 + \beta_1 X$.
- El parámetro β_0 es la ordenada al origen del modelo (punto de corte con el eje Y) y β_1 la pendiente, que puede interpretarse como el incremento de la variable dependiente por cada incremento en una unidad de la variable independiente. Estos parámetros son desconocidos y habrá que estimarlos de cara a realizar predicciones.

Además de la hípotesis establecida sobre los errores de que en media han de ser cero, se establecen las siguientes hipótesis:

ii) La varianza de ε es constante para cualquier valor de x, es decir,

$$Var(\varepsilon/X = x) = \sigma^2$$

iii) La distribución de ε es normal, de media 0 y desviación σ.

iv) Los errores asociados a los valores de Y son independientes unos de otros.

En consecuencia, la distribución de Y para x fijo es normal, con varianza constante σ^2, y media que varía linealmente con x, dada por $\beta_0 + \beta_1 x$. Además los valores de Y son independientes entre sí.

9.3 Estimación de los parámetros del modelo

Partimos de una muestra de valores de X e Y medidos sobre n individuos:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),$$

y queremos estimar valores en Y según el modelo $\hat{Y} = \beta_0 + \beta_1 X$, donde β_0 y β_1 son por el momento desconocidos. Debemos encontrar entonces de entre

todas las rectas la que mejor se ajuste a los datos observados, es decir, buscamos aquellos valores de β_0 y β_1 que hagan mínimos los errores de estimación. Para un valor x_i, el modelo estima un valor en Y igual a $\hat{y_i} = \beta_0 + \beta_1 x_i$ y el valor observado en Y es igual a y_i, con lo cual el error de estimación en ese caso vendría dado por $e_i = y_i - \hat{y_i} = y_i - (\beta_0 + \beta_1 x_i)$. Entonces tomaremos como estimaciones de β_0 y β_1 , que notamos por $\hat{\beta_0}$ y $\hat{\beta_1}$, aquellos valores que hagan mínima la suma de los errores al cuadrado, que viene dada por:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i)^2$$

De ahí que al método de estimación se le llame *método de mínimos cuadrados*. La solución se obtiene por el mecanismo habitual, derivando SSE con respecto a β_0 y β_1 e igualando a 0. Los estimadores resultan:

$$\begin{aligned} \hat{\beta_1} &= \frac{SS_{xy}}{SS_{xx}} \\ \hat{\beta_0} &= \bar{y} - \hat{\beta_1}\bar{x} \end{aligned}$$

siendo:

$$\begin{aligned} SS_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}, \\ SS_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = n\sigma_x^2 \end{aligned}$$

A la recta resultante $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X$ se le llama **recta de regresión lineal de Y sobre X.**

Un último parámetro a estimar en el modelo es la varianza de los errores (σ^2). A su estimador se le denomina *varianza residual* y viene dada por:

$$\hat{s_R}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SS_{yy} - \hat{\beta_1} SS_{xy}}{n-2}$$

Ejemplo 9.2. Para los datos del Ejemplo 9.1. referentes a la cantidad de compresion (Y) de un material aislante a diferentes niveles de presión (X), vamos a determinar la recta de regresión.

$$SS_{xy} = 7, SS_{xx} = 10$$

luego

$$\begin{aligned} \hat{\beta_1} &= \frac{SS_{xy}}{SS_{xx}} = 0.7 \\ \hat{\beta_0} &= \bar{y} - \hat{\beta_1}\bar{x} = -0.1 \end{aligned}$$

La recta de regresión de Y sobre X es por tanto:

$$\hat{Y} = -0.1 + 0.7X$$

9.4 Inferencias sobre el coeficiente de regresión

Observábamos que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ dependen de la muestra seleccionada, por lo tanto son variables aleatorias y presentarán una distribución de probabilidad. Estas distribuciones de probabilidad de los estimadores pueden utilizarse para construir intervalos de confianza o contrastes sobre los parámetros del modelo de regresión.

Al comienzo del capítulo nos planteábamos como uno de los objetivos el decidir si el efecto de la variable independiente es o no significativo para la variable dependiente. Si nos fijamos, esto es equivalente a contrastar si el coeficiente β_1 es o no significativamente distinto de cero. Un $\beta_1 = 0$ implicaría la ausencia de relación lineal entre las variables.

En términos generales planteamos los siguientes contrastes para β_1 :

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Contraste	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 < b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 \neq b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 > b_1$
Estadístico de contraste	$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{s}_R^{\,2}/SS_{xx}}}$, con $\hat{s}_R^{\,2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}$		
Región de rechazo	$t < t_{a,n-2}$	$\lvert t\rvert > t_{1-a/2,n-2}$	$t > t_{1-a,n-2}$

Decíamos que de especial interés es el contraste:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Ejemplo 9.3 Para los datos Ejemplo 9.1 sobre el material aislante, vamos a contrastar si el efecto de la presión sobre la compresión es o no significativo ($\alpha = 0.05$)

$$\begin{aligned}
\hat{\beta}_1 &= 0.7 \\
\hat{s}_R^{\,2} &= \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = 0.367 \\
SS_{xx} &= 10 \\
t &= \frac{\beta_1 - 0}{\sqrt{s^2/SS_{xx}}} = 3.7 \\
t_{0.975,3} &= 3.18
\end{aligned}$$

Como $|t| > t_{0.975,3}$ podemos rechazar H_0 al 5% de significación, por lo tanto el efecto de la presión sobre la compresión es significativo.

9.5 El coeficiente de correlación lineal y el coeficiente de determinación

Nuestro objetivo en adelante será medir la bondad del ajuste de la recta de regresión a los datos observados y cuantificar al mismo tiempo el grado de asociación lineal existente entre las variables en cuestión. A mejor ajuste, mejores serán las predicciones realizadas con el modelo.

La evaluación global de una recta de regresión puede hacerse mediante la varianza residual, que como sabemos es un índice de la precisión del modelo. Sin embargo, esta medida no es útil para comparar rectas de regresión de variables distintas, o comparar el grado de asociación lineal entre distintos pares de variables, ya que depende de las unidades de medida de las variables.

El coeficiente de correlación lineal

Como solución al inconveniente planteado, para medir la asociación lineal entre dos variables X e Y se utiliza una medida adimensional denominada *coeficiente de correlación lineal*, dado por:

$$r = \frac{E[(X - E(X))(Y - E(Y)]}{\sqrt{VAR(X)VAR(Y)}} = \frac{\sqrt{VAR(X)}}{\sqrt{VAR(Y)}}\beta_1$$

y su estimación a partir de datos de una muestra resulta:

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}}\hat{\beta}_1$$

El coeficiente de correlación lineal toma valores entre -1 y 1 y su interpretación es la siguiente:

- Un valor cercano o igual a 0 indica respectivamente poca o ninguna relación lineal entre las variables.
- Cuanto más se acerque en valor absoluto a 1 mayor será el grado de asociación lineal entre las variables. Un coeficiente igual a 1 en valor absoluto indica una dependencia lineal exacta entre las variables.
- Un coeficiente positivo indica asociación lineal positiva, es decir, tienden a variar en el mismo sentido.
- Un coeficiente negativo indica asociación lineal negativa, es decir, tienden a variar en sentido opuesto.

  Nótese que si $\beta_1 = 0$ entonces $r = 0$, en cuyo caso hay ausencia de linealidad. Por lo tanto, contrastar si el coeficiente de correlación lineal es significativamente distinto de 0 sería equivalente a contrastar si β_1 es significativamente distinto de cero, contraste que ya vimos en la sección anterior.

El coeficiente de determinación

Según hemos visto, el coeficiente de correlación lineal puede interpretarse como una medida de la bondad del ajuste del modelo lineal, concretamente, un valor del coeficiente igual a 1 o -1 indica dependencia lineal exacta, en cuyo caso el ajuste es perfecto. No obstante, para cuantificar la bondad del ajuste de un modelo, lineal o no, se utiliza una medida que se denomina *coeficiente de determinación lineal* R^2, que es la proporción de variabilidad de la variable Y que queda explicada por el modelo de entre toda la presente, y cuya expresión es:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y_i} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{SSE}{SS_{yy}},$$

que en modelo de regresión lineal coincide con el cuadrado del coeficiente de correlación lineal:

$$R^2 = r^2$$

El coeficiente de determinación toma valores entre 0 y 1, y cuanto más se aproxime a 1 mejor será el ajuste y por lo tanto mayor la fiabilidad de las predicciones que con él realicemos.

Nótese que si el coeficiente de correlación lineal r es igual a 1 o -1 entonces $R^2 = 1$ y por lo tanto el ajuste lineal es perfecto.

Ejemplo 9.4 En el Ejemplo 9.1 $r = 0.90$ y $R^2 = 0.82$. Esto indica que el grado de asociación lineal entre las variables es alto, y concretamente el 82% de la variación total de los valores de la compresión pueden ser explicados mediante la recta de regresión ajustada.

9.6 Predicción a partir del modelo

Recordamos que en el modelo ajustado de la recta de regresión,

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

y, por otro lado,

$$E[Y/X = x] = \beta_0 + \beta_1 x,$$

luego $\hat{y}$ puede interpretarse de dos formas:

1. Como *predicción* del valor que tomará Y cuando $X = x$.
2. Como *estimación* del valor medio en Y para el valor $X = x$, es decir, $E[Y/X = x]$.

Ambas cantidades están sujetas a incertidumbre, que será tanto mayor cuanto peor sea el ajuste realizado mediante la recta de regresión. Para concluir el tema, establecemos un intervalo de confianza para estas cantidades.

Proposición. Podemos decir que con un $(1-\alpha)x100\%$ de confianza que cuando $X = x$, el valor predicho en Y o el valor medio estimado en $Y, E[Y/X = x]$, se encuentran en el intervalo:

$$\left[\hat{y} \pm t_{1-\alpha/2,n-2}, s_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_{xx}}}\right]$$

Ejemplo 9.4. Para los datos del Ejemplo 9.1,

Pieza	Presíón (x)	Compresión (y)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

a. Predecir el valor en la compresión para un nivel de presión igual a 6.

La recta de regresión ajustada era $\hat{Y} = -0.1 + 0.7X$, con lo cual para un $x = 6$ se predice un valor en Y igual a $\hat{y} = -0.1 + 0.7 * 6 = 4.1$

b. ¿En qué medida son fiables las predicciones realizadas con la recta de regresión ajustada?

Como el coeficiente de determinación es igual a 0.81, las predicciones realizadas con la recta serán fiables en un 81%.

c. Determinar un Intervalo al 95% de confianza para el valor medio de compresión a una presión de 6 .

El intervalo de confianza resulta:

$$\left[\hat{y} \pm t_{1-\alpha/2,n-2}s_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_{xx}}}\right] = \left[4.1 \pm 3.18 * 0.6\sqrt{\frac{1}{5}+\frac{(6-3)^2}{10}}\right] = [2.1, 6.1]$$

9.7 Ejercicios

1. Se supone que el alargamiento de un cable de acero está relacionado linealmente con la intensidad de la fuerza aplicada. Cinco especímenes idénticos de cable dieron los resultados siguientes:

Fuerza (X)	1.0	1.5	2	2.5	3
Alargamiento (Y)	3	3.5	5.4	6.9	8.4

(a) Estudia el grado de asociación lineal entre ambas variables.

(b) Predice el alargamiento para una fuerza de 2.2. ¿En qué medida es fiable tal predicción?.

(c) Contrastar al 5% si la fuerza aplicada influye significativamente sobre el alargamiento.

(d) Obtener un intervalo de confianza al 95% para el valor que se predice en el alargamiento para una fuerza de 2.2

2. Las bodegas modernas utilizan vehículos guiados computarizados y automatizados para el manejo de materiales. En consecuencia, la disposición física de la bodega debe diseñarse con cuidado a modo de evitar el congestionamiento de los vehículos y optimar el tiempo de respuesta. En *The journal of Engineering for Industry (agosto 1993)* se estudió el diseño óptimo de una bodega automatizada. La disposición empleada supone que los vehículos no se bloquean entre sí cuando viajan dentro de la bodega, es decir, no hay congestionamiento. La validez de este supuesto se verificó simulando por ordenador las operaciones de la bodega. En cada simulación se varió el número de vehículos y se registró el tiempo de congestionamiento (tiempo total que un vehículo bloquea a otro). Los datos se muestran en la tabla de abajo. Los investigadores están interesados en conocer la relación entre el tiempo de congestionamiento (Y) y el número de vehículos (X).

X	1	2	3	4	5	6	7	8	9	10
Y	0	0	0.02	0.01	0.01	0.01	0.03	0.03	0.02	0.04

a) Cuantifica la dependencia lineal existente entre ambas variables.

b) ¿Es significativa la dependencia lineal entre las variables?. Tomar $\alpha = 0.05$.

c) Obtén la recta de regresión que expresa el tiempo de congestión en función del número de vehículos.

d) Predice linealmente el tiempo de congestión cuando el número de vehículos es de 12. ¿En qué medida es fiable tal predicción?.

e) Determinar el intervalo en el que se encuentra al 95 de confianza el tiempo medio de congestión para un número de vehículos de 12.

3. Los siguientes datos se refieren al crecimiento de una colonia de bacterias en un medio de cultivo:

X	3	6	9	12	15	18
Y	115000	147000	239000	356000	579000	864000

,

siendo X el número de días desde la inoculación e Y el número de bacterias.

Comprobar gráfica y numéricamente que el tipo de asociación entre ambas variables no es lineal.

4. Se ha realizado un estudio para investigar el efecto de un determinado proceso térmico en la dureza de una determinada pieza. Once piezas se seleccionaron para el estudio. Antes del tratamiento se realizaron pruebas de dureza para determinar la dureza de cada pieza. Después, las piezas fueron sometidas a un proceso térmico de templado con el fin de mejorar su dureza. Al final del proceso, se realizaron nuevamente pruebas de dureza y se obtuvo una segunda lectura. Se recogieron los siguientes datos (Kg. de presión):

Dureza previa	182	232	191	200	148	249	276	213	241	480	262
Dureza post.	198	210	194	220	138	220	219	161	210	313	226

   (a) Calcula la media, mediana, percentiles 25 y 75 de la dureza antes y después del proceso.

   (b) Calcula la desviación típica en ambos casos. ¿En qué caso hay mayor variabilidad?.

   (c) ¿Se puede afirmar que el proceso de templado mejora la dureza de las piezas?.

   (d) Decide si un modelo lineal es adecuado para explicar la dureza posterior en función de la dureza previa. En caso afirmativo obtenlo y predice la dureza tras el proceso de templado de una pieza con un dureza previa de 215.

5. La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial (por $10^3(M)$) desconocida del éster, se han medido las concentraciones del mismo a diferentes tiempos (en minutos) obteniéndose los resultados siguientes:

Tiempo	3	5	10	15	20	30	40	50	60	75	90
Conc.	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.4

   (a) Realiza una nube de puntos de las dos variables. La teoría cinética de este tipo de reacciones nos indica que la evolución de la concentración del éster en función del tiempo se rige por $C_t = C_0 e^{-kt}$, donde C_0 es la concentración inicial. ¿Qué transformación de los datos nos lleva a un modelo lineal?. Realiza esta transformación y obtén la concentración inicial C_0 y la velocidad k de desaparición del éster.

   (b) Suponemos ahora que nos comunican que la concentración inicial del éster es $C_0 = 3.10^{-2}(M)$. ¿Cómo incorporar esta información a nuestro análisis anterior?. Obtén el nuevo valor de k.

6. Para analizar la degradación de la señal emitida por una antena, se tomaron los siguientes datos: la frecuencia de la señal en el momento de ser emitida (X) y la frecuencia de la señal al ser recibida (Y). Los resultados medidos en Megahercios fueron:

X	1.75	1.8	1.78	2.01	2.48	2.58	2.98	2.65	2.01	3.87
Y	1.56	1.45	1.75	0.84	2.02	2.41	2.75	1.44	1.55	2.02

a. Calcular la media, mediana y moda de ambas variables.

b. De las señales emitidas entre 2 y 3 Megahercios ¿Cuál es la proporción de ocasiones en las que la frecuencia recibida fue menor que 2.5 Megahercios?

c. Determinar el intervalo en el que se encuentra el 50% central de la variable Y.

d. ¿Es significativa la relación lineal entre las variables?. ¿Influye significativamente la variable X sobre Y?. Realizar el contraste al 5% de significación.

e. ¿Qué frecuencia se predice en la señal al ser recibida si al ser emitida es de 3.5 Megahercios?. ¿Es fiable la predicción?.

f. Obtener un intervalo de confianza al 95% para la señal recibida si la señal emitida es de 3.5 Megahercios.