

Capítulo VII

Regresión Lineal Simple (1)

Introducción

Ya que se han tratado problemas inferenciales que incluían el análisis de una variable a la vez. Ahora se discutirá por primera vez un problema inferencial que considera dos variables.

En Biología es frecuente encontrar situaciones en las cuales se desea investigar la relación funcional entre dos variables. Se quiere investigar situaciones en las que se asume que la magnitud de una variable (la variable dependiente **Error! Bookmark not defined.**) será determinada por la magnitud de una segunda variable (función de la variable independiente **Error! Bookmark not defined.**). Por ejemplo, se puede estudiar la suposición de que el número de huevos puesto por una cierta especie de ave se incrementa con la cantidad de comida disponible en su medio ambiente.

Fíjese que en el análisis de **regresión** **Error! Bookmark not defined.** nunca se hace la suposición inversa: que la magnitud de la variable independiente **Error! Bookmark not defined.** depende de la magnitud de la variable dependiente **Error! Bookmark not defined.** . Por ejemplo, aunque el crecimiento de las plantas puede ser función de la temperatura del ambiente, la temperatura no podría ser considerada como función de la tasa de crecimiento.

Es importante tener en cuenta que para muchas clases de datos biológicos la relación entre dos variables no es de dependencia. En tales casos la magnitud de una variable puede cambiar con la magnitud de una segunda, pero no es razonable considerar que una variable es dependiente de la otra. En tales situaciones, se debe emplear un análisis de **correlación** en vez de uno de regresión **Error! Bookmark not defined.**. Un ejemplo de datos de este tipo sería las medidas de la longitud de pico y el peso de huevos de pinzones. Se podría encontrar que aves con picos grandes tienen en general huevos más grandes, pero no hay bases biológicas para decir que el tamaño de los huevos depende del tamaño del pico. En el **Capítulo VIII** se discutirán las técnicas de correlación.

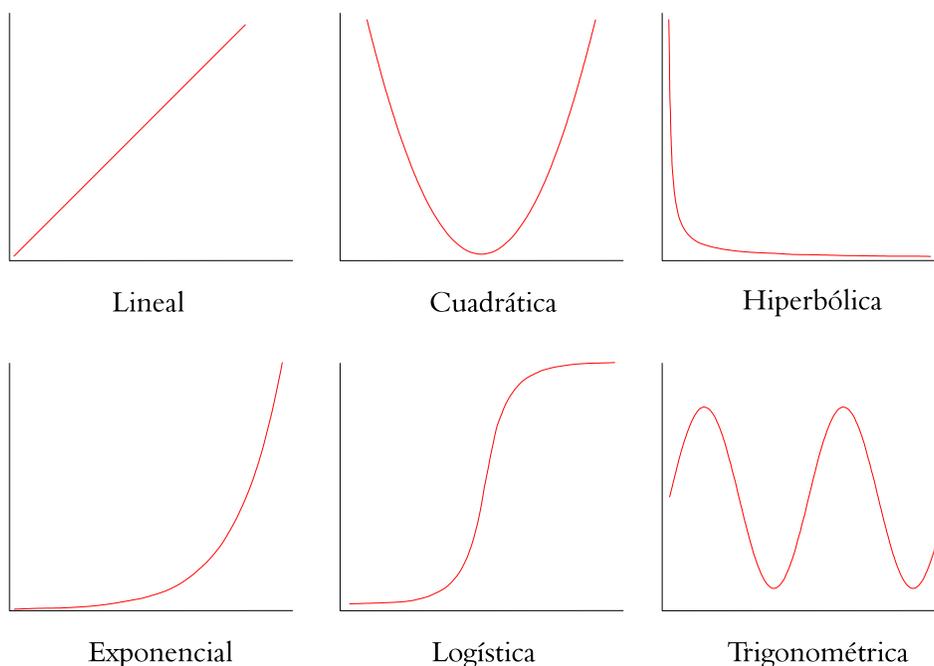
El análisis que se realiza para determinar la naturaleza de la relación entre variables se conoce como **regresión** **Error! Bookmark not defined.**. Cuando se trata con solo dos variables se usa el término **regresión simple**. El adjetivo **lineal** se añade cuando el análisis asume que la relación entre las variables es una línea recta. Se considerará en este libro solo la **regresión lineal simple**, pero muchos de los conceptos que se presentarán son aplicables a la mayoría — sino a todos — de tipos de regresión.

Modelaje Matemático

Las técnicas de regresión **Error! Bookmark not defined.** son parte de un aspecto importante de la estadística conocido como **modelaje matemático** **Error! Bookmark not defined.** Modelado es el proceso de describir datos y relaciones entre variables. La estadística provee un número de herramientas que permiten describir las cosas matemáticamente. Los modelos pueden ser **descriptivos** (así solamente se resumen los datos) o **predictivos** (permiten predecir los valores de variables basados en los valores de una o más variables diferentes). La regresión puede ser descriptiva y predictiva al mismo tiempo, dependiendo de cómo sea usada. La regresión lineal simple es una técnica usada para describir la relación matemática entre una sola variable dependiente **Error! Bookmark not defined.** y una sola variable independiente **Error! Bookmark not defined.**

Aunque solamente se trabajará con modelos de variación lineal, téngase en cuenta que hay un número infinito de posibles modelos no lineales que también pueden ser usados. Algunas de las relaciones más comunes encontradas en Biología se muestran en el Gráfico 7.1.

Gráfico 7.1



Regresión Lineal Simple

La relación funcional más simple, y más común encontrada en Biología es la lineal. La ecuación para la relación lineal en análisis de regresión **Error! Bookmark not defined.** entre dos variables es:

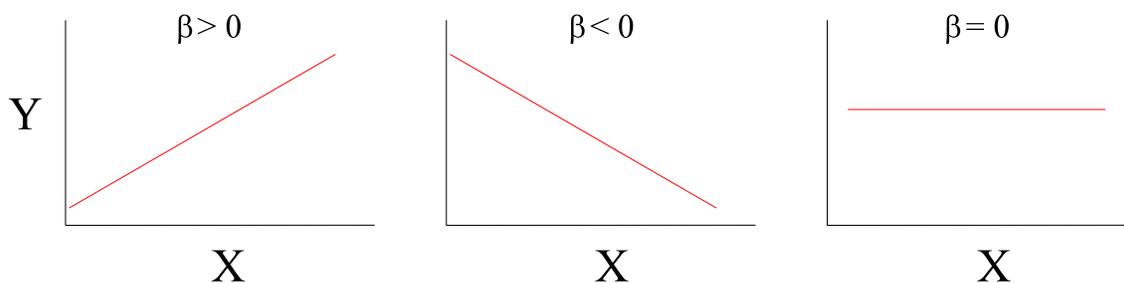
$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (7.1)$$

Esta es la ecuación para una población **Error! Bookmark not defined.** donde: α y β son parámetros de la población (por lo tanto constantes). La ecuación puede ser interpretada como sigue: para cada valor de la variable independiente **Error! Bookmark not defined.** (X_i) el valor correspondiente de la variable dependiente **Error! Bookmark not defined.** (Y_i) es igual a X_i multiplicado por $\beta + \alpha$ + alguna cantidad aleatoria de varianza (ε_i). Más adelante se hablará acerca de ε_i .

Estimación de la Pendiente e Intercepción:

Para una población **Error! Bookmark not defined.** dada (que consiste de pares de valores de X, Y) se quiere conocer los valores únicos de α y β que describen la relación funcional entre las dos variable. El valor de β puede tener valores entre $-\infty$ y $+\infty$. Si β es mayor que 0, entonces hay una relación positiva entre las magnitudes de Y y X . Si β es menor que 0 entonces hay una relación negativa (ej. la magnitud de Y decrece con un incremento en la magnitud de X). Finalmente, si β es igual a 0, entonces no hay relación entre Y y X (ej. la magnitud de Y no depende de la magnitud de X). En el Gráfico 7.2 se puede ver la representación gráfica de estas tres posibilidades.

Gráfico 7.2

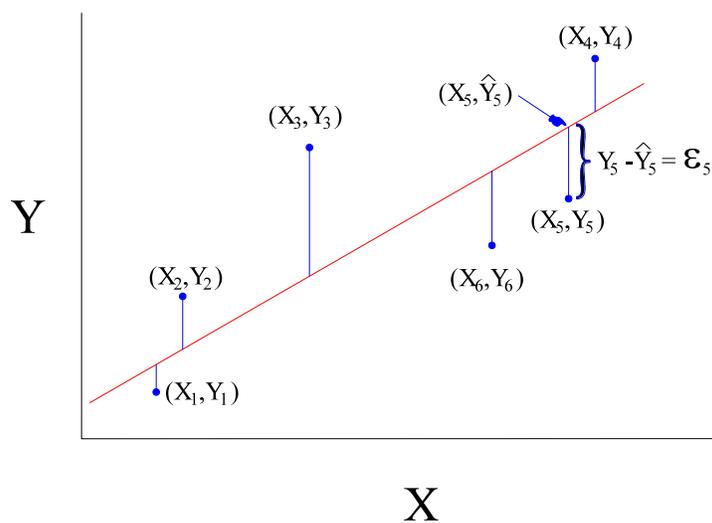


Los datos de regresión **Error! Bookmark not defined.** de dos variables se grafican más convenientemente con la variable independiente **Error! Bookmark not defined.** en el eje horizontal (el eje de las abscisas o eje X) y la variable dependiente **Error! Bookmark not defined.** en el eje vertical (eje de las ordenadas o eje Y). Cada punto en el gráfico corresponde a un par de datos (X_i, Y_i) (véase el Gráfico 7.3).

Para los datos de una muestra de una población **Error! Bookmark not defined.** se quiere estimar los valores de α y β que mejor estimen de la relación lineal entre X y Y . A los estimativos de α y β se les dan los símbolos **a** y **b**.

Para cualquier grupo de datos de una muestra existe sólo una línea que estime mejor la relación lineal entre dos variables, $\hat{Y} = a + bX$ (\hat{Y} se conoce como **Y sombrero**). Cada punto X_i en el gráfico tendrá un valor correspondiente \hat{Y}_i sobre esta línea Y . Por lo tanto, para cada observación (X_i, Y_i) existe un punto (X_i, \hat{Y}_i) en esta línea. La diferencia entre Y_i y \hat{Y}_i es ε_i . Se

refiere a ε_i como el término de error o el residual **Error! Bookmark not defined.** Los residuales son una herramienta importante para chequear que exacta es la aproximación del modelo y que apropiado.

Gráfico 7.3

El método más común (¡pero no el único!) de calcular a y b es el método de **cuadrados mínimos**. Este método calcula los valores de a y b para que el valor de la ecuación 7.2 es un mínimo.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (7.2)$$

Este valor es conocido como **la suma residual de los cuadrados** o el **error de la suma de los cuadrados**. Se hablará de este valor más adelante (si el suspenso es demasiado mire en la página 89).

Para calcular el valor de a y b que minimiza (7.2) primero se calcula:

$$\sum x^2 = \sum X_i^2 - \frac{\sum (X_i)^2}{n} \quad (7.3)$$

$$\begin{aligned} \sum xy &= \sum (X_i - \bar{X}_i)(Y_i - \bar{Y}) \\ &= \sum X_i Y_i - \frac{\sum (X_i) \sum (Y_i)}{n} \end{aligned} \quad (7.4)$$

Ahora se calcula b como:
$$\frac{\sum xy}{\sum x^2} \quad (7.5a)$$

y a se calcula como:
$$\bar{Y} - \beta \bar{X} \quad (7.5b)$$

Como ejemplo se considerará la siguiente situación: una investigadora desea conocer cómo varía la temperatura con respecto a la altura en la isla Santa Cruz. Un buen día ella coloca 9 asistentes a lo largo de una línea que corre de la costa al punto más alto de la isla. A las 12 en punto del medio día exactamente ellos registran la temperatura en cada una de estas localizaciones. Los datos se presentan en la siguiente Tabla:

Tabla 7.1

<u>Altura (m)</u>	<u>Temperatura (°C)</u>
0	25.0
50	24.1
190	23.5
305	21.2
456	20.6
501	20.0
615	18.5
700	17.2
825	17.0

$$\sum X_i = 3642$$

$$\sum Y_i = 187.1$$

$$\sum X_i^2 = 2139412$$

$$\bar{X} = 404.67$$

$$\bar{Y} = 20.78$$

$$\sum X_i Y_i = 68992.1$$

$$\begin{aligned} \sum x^2 &= 2139412 - \frac{3642^2}{9} \\ &= 665616 \end{aligned}$$

$$\begin{aligned} \sum xy &= 68992.1 - \frac{3642 \times 187.1}{9} \\ &= -6721.00 \end{aligned}$$

$$b = \frac{-6721.0}{665616} = -\mathbf{0.0101}$$

$$a = 20.78 - (-0.0101) \times 404.69 = \mathbf{24.88}$$

La línea más apropiada para estos datos es:

$$\text{temp (°C)} = \mathbf{24.88 - 0.0101 \times \text{altura (m)}}$$

en otras palabras: la temperatura al nivel del mar se espera que sea 24.88 °C y que disminuya 0.0101 °C por cada metro de incremento en altitud (1.01 °C cada 100 m).

Utilizando las ecuaciones de regresión**Error! Bookmark not defined.** para predecir valores de Y:

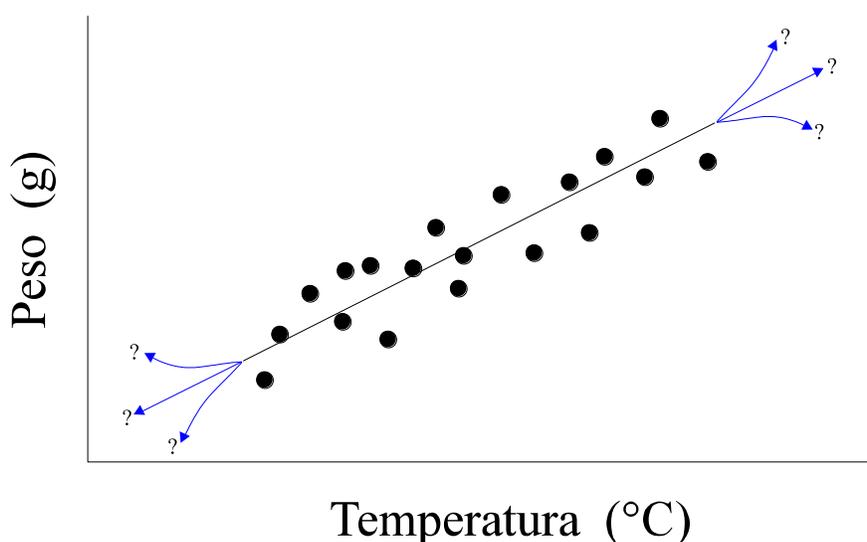
Una vez que se ha calculado **a** y **b** es posible predecir los valores esperados de la variable dependiente**Error! Bookmark not defined.** para un valor dado de X. Note que **nunca** se debe tratar de hacer lo inverso: calcular un valor esperado para X basado en algún valor seleccionado de Y, ya que esto no es válido estadísticamente y puede fácilmente dar resultados erróneos.

Para el ejemplo anterior podemos calcular la temperatura estimada para los 250 m de altitud. Sustituyendo el valor de 250 por X tendremos:

$$\begin{aligned} \text{temp.} &= 24.88 - 0.0101 \times 250 \\ &= \mathbf{22.34 \text{ } ^\circ\text{C}} \end{aligned}$$

CUIDADO: ¡Nunca se debe usar una ecuación de regresión**Error! Bookmark not defined.** para estimar valores de Y que se van a usar para estimar valores de X fuera del rango de los valores de X que se usaron originalmente para estimar la ecuación! Una ecuación de regresión es válida sólo en el rango de valores de X usados para generar la ecuación. El peligro de estimar fuera de este rango es que nunca sabemos con seguridad si la relación que hemos descrito continúa igual fuera del rango de los datos examinados. Siempre es posible, en verdad es hasta común, que para valores intermedios de una relación particular la variable dependiente**Error! Bookmark not defined.** es una función lineal de la variable independiente**Error! Bookmark not defined.**, pero, para valores más extremos esta función se convierte más en una no lineal (véase el Gráfico 7.4).

Gráfico 7.4



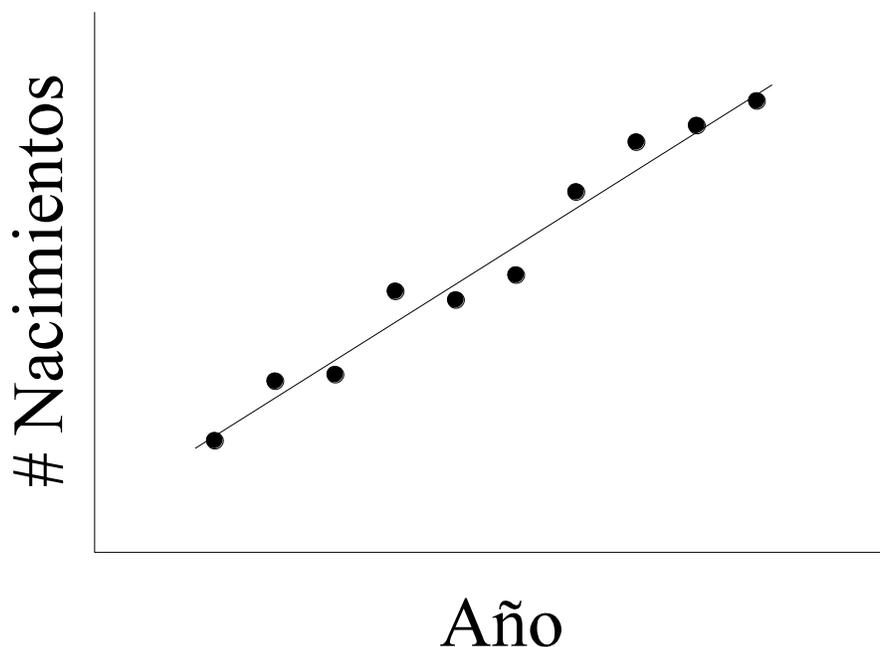
Cuidados con la interpretación de modelos matemáticos:

Un modelo matemático **Error! Bookmark not defined.** no es una prueba de una relación biológica causal. En otras palabras, solo porque se haya desarrollado un análisis de regresión **Error! Bookmark not defined.** y calculado la ecuación de regresión, la cual se adapta a los datos perfectamente, no se ha probado que la magnitud de la variable dependiente **Error! Bookmark not defined.** esté siendo controlada o influida por la magnitud de la variable independiente **Error! Bookmark not defined.**. Considérese el siguiente ejemplo imaginario: en Puerto Ayora se ha notado que tanto la tasa de nacimientos humanos como el número de garzas del ganado han aumentado por varios años los datos se muestran gráficamente en el Gráfico 7.5.

Con base en estos datos exclusivamente, sería muy fácil concluir que el incremento de la tasa de nacimientos en Puerto Ayora al incremento en la población **Error! Bookmark not defined.** de garzas (como no hay cigüeñas en Galápagos...). A menos que se crea en una teoría un tanto no científica de la reproducción humana, sería fácil ver el problema de interpretación de modelos estadísticos sin un entendimiento de la biología relacionada.

Este tema nos lleva también a la discusión de modelos como herramientas descriptivas y/o predictivas. Por ejemplo, podría generarse un modelo de regresión **Error! Bookmark not defined.** que prediga exactamente los valores de una variable (dígase peso total de cuerpo) basado en valores de una segunda (dígase largo del cuerpo). Este entonces sería predictivo en vista de que permite la predicción del peso del cuerpo basado en la longitud. Sin embargo, a menos que la forma del modelo de regresión pudiera ser explicada en términos biológicos no podría ser considerada porque no describe el proceso biológico involucrado. Matemáticamente los modelos son herramientas útiles, pero nunca deben ser usados como sustitutos para un buen entendimiento biológico.

Gráfico 7.5

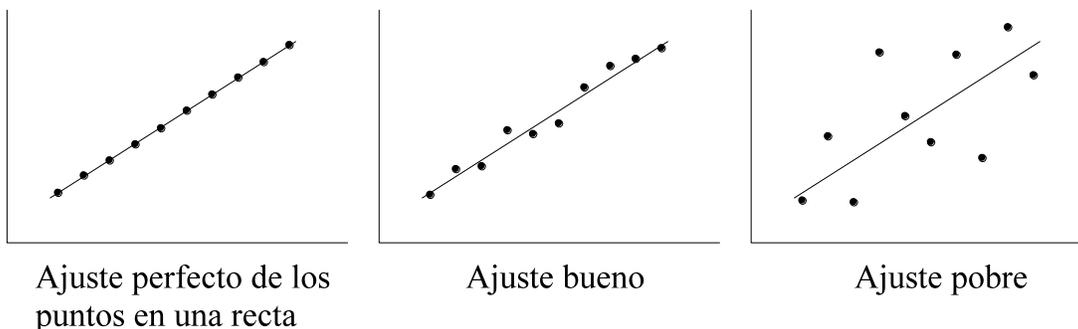


Evaluación de la Significancia de una Regresión

Una vez calculada nuestra ecuación de regresión **Error! Bookmark not defined.** se deben responder dos preguntas: “¿qué tan bien se acomodan los datos a la curva de regresión?” y “¿es la pendiente (β) diferente de 0?” (“¿hay una relación estadística entre las dos variables?”).

Considérese las tres posibilidades del Gráfico 7.6.

Gráfico 7.6



En cada ejemplo la línea de regresión **Error! Bookmark not defined.** es la misma, pero es obvio que hay una gran diferencia en la medida en que cada curva describe los datos utilizados. Lo que se necesita es una forma de describir exactamente “qué tan bien” la línea más apropiada representa los datos.

Para responder la pregunta “¿con qué exactitud la línea representa los datos?” es necesario presentar una nueva variable:

$$\sum y^2 = \sum Y_i^2 - \frac{(\sum Y)^2}{n} = SC_{tot} \quad (7.6)$$

Esta variable se llama la **suma total de los cuadrados** y puede ser interpretada como la cantidad de variación en la variable dependiente. Ahora, usando dos variables ($\sum xy$, $\sum x^2$) calculadas anteriormente:

$$\frac{(\sum xy)^2}{\sum x^2} = SC_{reg} \quad (7.7)$$

Este valor se conoce como **suma de regresión de los cuadrados** y puede ser interpretada como la cantidad de variación en Y explicada por la regresión. SC_{tot} será igual a SC_{reg} solamente cuando todos los puntos caigan sobre la línea de regresión (cuando toda la variación en Y es explicada por la línea). La proporción (%) de variación en Y explicada o representada por la regresión se llama coeficiente de la determinación (r^2) y se calcula como sigue:

$$r^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{\sum xy}{\sum x^2 \sum y^2} \quad (7.8)$$

Los valores de r^2 van desde 0 a 1. Un valor de 0 significa que las líneas de regresión no explican nada de la variación en la variable dependiente. Un valor de 1 significa que la línea de regresión explica 100% de la variación en Y (todos los puntos caen en la línea). Obviamente, mientras más se acerque el valor de r^2 a 1 mejor será la regresión.

La pregunta “¿es β igual a 0?” está relacionada con otra: “¿qué tan bien se ajusta la recta a los datos?” Como se trabaja con datos de muestras, y se está estimando β con b es necesario preguntar “¿es b significativamente diferente a 0?” Esta pregunta es casi idéntica a la de la prueba t para una muestra (donde la pregunta era si $\mu=0$) que se discutió en el Capítulo V. Podría ser de gran ayuda si se regresa a ese Capítulo por un momento y se revisa la prueba t para una muestra. En la misma forma que se prueba si la media de una sola muestra fue significativamente diferente de 0, se puede probar si la estimación de β es significativamente diferente de 0. Se realizará esta prueba usando la prueba t de Student. La misma pregunta se responde con frecuencia usando el análisis de varianza (ANOVA), técnica que está fuera del alcance de este libro.

Como siempre, para cualquier prueba inferencial, se deben definir las hipótesis nula y alternativa:

La Hipótesis Nula (H_0) es: $\beta = 0$

La Hipótesis Alternativa (H_a) es: $\beta \neq 0$

Para calcular el estadístico t se necesita reintroducir una variable:

$$SC_{res} = SC_{tot} - SC_{reg} \quad (7.9)$$

SC_{res} es la **Suma Residual de los Cuadrados** (igual a $\sum (Y_i - \hat{Y}_i)^2$; véase la página 87) y es la cantidad de variación en Y que no es explicada por la regresión. Nótese que cuando todos los puntos caen en la línea $SC_{tot} = SC_{reg}$ y $SC_{res} = 0$.

Al dividir SC_{res} entre $n-2$ (los grados de libertad) se tiene un valor conocido como la **Media Residual de los Cuadrados** (MC_{res}) que es la varianza de las diferencias $Y_i - \hat{Y}_i$ (véase el Gráfico 7.3).

La raíz cuadrada de MC_{res} se da por el símbolo $S_{Y \cdot X}$ y se denomina **error residual de la estimación** (ocasionalmente **error estándar de la regresión**).

$$\frac{SC_{res}}{n-2} = MC_{res} \quad (7.10a)$$

$$\sqrt{MC_{res}} = S_{Y \cdot X} \quad (7.10b)$$

Con $S_{Y \cdot X}$ se puede calcular el **error estándar de la pendiente** (S_b):

$$S_b = \sqrt{\frac{S_{Y \cdot X}^2}{\sum X^2}} \text{ donde } S_{Y \cdot X}^2 = (S_{Y \cdot X})^2 \quad (7.11)$$

El estadístico t se calcula como:

$$t = \frac{b - \beta}{S_b} \quad \text{con: } v = n-2 \quad (7.12a)$$

donde β es el valor de la pendiente de la hipótesis nula.

Cuando $\beta=0$ el mismo estadístico se puede calcular como:

$$t = \frac{b}{S_b} \quad (7.12b)$$

Se mira el valor crítico del valor t para los grados de libertad (v) usados. Si el estadístico t es mayor o igual que el valor crítico, se rechaza la hipótesis nula y se acepta la alternativa que dice que hay una relación significativa entre las dos variables.

Como ejemplo de todo lo que hemos discutido, se considerará el siguiente estudio. Se registró la presión sanguínea de 20 voluntarios. Los datos se muestran en Tabla 7.2.

Tabla 7.2

<u>Clase de edad</u>	<u>Presión Sanguínea</u>	<u>Clase de edad</u>	<u>Presión Sanguínea</u>
30	108	60	148
30	110	60	151
30	106	60	146
40	125	60	147
40	120	60	144
40	118	70	162
40	119	70	156
50	132	70	164
50	137	70	158
50	134	70	159

$$n=20$$

$$\sum X_i = 1050 \quad \sum Y_i = 2744 \quad \sum X_i Y_i = 149240$$

$$\sum X_i^2 = 59100 \quad \sum Y_i^2 = 383346$$

$$\sum x^2 = 3975.0 \quad SC_{tot} = \sum y^2 = 6869.2 \quad \sum xy = 5180.0$$

$$\bar{X} = 52.5 \quad \bar{Y} = 137.2$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{5180.0}{3975.0} = 1.303$$

$$a = \bar{Y} - b\bar{X} = 137.2 - 1.303 \times 52.5 = 68.78$$

$$SC_{reg} = \frac{(\sum xy)^2}{\sum x^2} = \frac{5180.0^2}{3975.0} = 6750.29$$

$$r^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{6750.29}{6869.2} = 0.98$$

$$SC_{res} = SC_{tot} - SC_{reg} = 6869.2 - 6750.29 = 118.91$$

$$S_{y \cdot x} = \sqrt{\frac{SC_{res}}{n-2}} = \sqrt{\frac{118.91}{18}} = 2.57$$

$$S_b = \sqrt{\frac{S_{y \cdot x}^2}{\sum x^2}} = \sqrt{\frac{2.57^2}{3975.0}} = 0.041$$

$$t = \frac{b}{S_b} = \frac{1.303}{0.041} = 31.95 \quad \nu = 20 - 2 = 18$$

Para 18 (20-2) grados de libertad **Error! Bookmark not defined.** el valor crítico de 5% es 2.101; el valor crítico para 1% es 2.878; el valor crítico para 0.1% es 3.922. Dado que 31.95 es mayor que 3.922, se rechaza la hipótesis nula **Error! Bookmark not defined.** que dice que no hay relación entre la edad y la presión sanguínea con un nivel de confianza de $P \leq 0.001$.

Capítulo VIII

Regresión Lineal Simple (2) y Correlación

Introducción

Como se mencionó en el Capítulo VII, en Biología existen muchas ocasiones donde se pueden encontrar relaciones no lineales que se desean modelar. Hay algunas técnicas de regresión no lineal que permiten el análisis directo de estas relaciones no lineales. Desafortunadamente éstas tienden a ser muy complicadas y usualmente se deben llevar a cabo con una computadora a menos que el conjunto de datos que se quiere usar sea muy pequeño. Estas técnicas, entonces, no se discutirán en este libro.

Transformaciones en Regresión

Otra manera de acercarse a las relaciones no lineales es la transformación de los datos y luego el uso de los análisis usuales de regresión lineal simple. Esto se hace frecuentemente. Las relaciones más comunes que se analizan de esta forma son:

$$Y = \alpha e^{\beta X} \quad (8.1)$$

y

$$Y = \alpha X^{\beta} \quad (8.2)$$

donde: e es el logaritmo natural; α y β son los parámetros a ser estimados; Y es la variable dependiente y X es la variable independiente.

Para aplicar las técnicas de regresión lineal simple a ese tipo de datos, necesitamos transformar las funciones para que tengan la siguiente forma:

$$Y = \alpha + \beta X$$

En ambos casos se usa una regresión logarítmica. No importa si la es una transformación logarítmica común o natural, pero usualmente es más conveniente una que trabaje con logaritmos naturales (base e). En el caso de la ecuación 8.1 se toma el logaritmo natural a ambos lados de la ecuación:

$$\ln(Y) = \ln(\alpha e^{\beta X})$$

Si se recuerdan las reglas para manipular logaritmos, esto se convierte en:

$$\ln(Y) = \ln(\alpha) + \ln(e^{\beta X})$$

y con un paso final se obtiene:

$$\ln(Y) = \ln(\alpha) + \beta X \quad (8.3)$$

Esto significa que si se toma cualquier grupo de datos con la misma forma que la ecuación 8.1 y se transforma la variable Y en su logaritmo natural, los datos serán lineales y analizables a través de la regresión **Error! Bookmark not defined.** lineal simple. Esto se conoce como transformación **Error! Bookmark not defined.** log-lineal.

Mediante el mismo proceso, la ecuación 8.2 puede ser transformada para que se parezca a una lineal (Se deja como tarea el hacer la transformación. Pista: ¿De qué otra forma se puede escribir $\ln(X^\beta)$?). La ecuación 8.2 se convierte en:

$$\ln(Y) = \ln(\alpha) + \beta \ln(X) \quad (8.4)$$

La ecuación 8.4 implica que, para datos que tengan la forma de la ecuación 8.2, una transformación logarítmica de ambas variables Y y X producirá una línea recta. Esto se conoce como transformación **Error! Bookmark not defined.** log-log.

Como ejemplo considérese la siguiente situación. El dueño de unas fincas madereras desea predecir el valor de su madera basado en la edad de sus árboles. A lo largo de los últimos años el dueño ha acumulado los siguientes datos de valor de los arboles en el mercado contra la edad de los árboles.

Tabla 8.1

<u>Edad (años)</u>	<u>Valor (\$/ha)</u>	<u>ln(edad+1)</u>	<u>ln(valor)</u>
0	1000	0.000	6.908
1	1105	0.693	7.008
2	1220	1.099	7.107
3	1350	1.386	7.208
5	1650	1.792	7.409
7	2010	2.079	7.606
10	2710	2.398	7.905
12	3320	2.565	8.108
15	4480	2.773	8.407
20	7390	3.045	8.908
25	12180	3.258	9.408
28	16400	3.367	9.705

El primer paso cuando se trata con cualquier problema de regresión **Error! Bookmark not defined.** debe ser siempre mirar los datos gráficamente. El Gráfico 8.1 muestra que los datos **no** son en verdad muy lineales.

Desafortunadamente, a veces los datos que tienen la forma de la ecuación 8.1 son similares en apariencia a la función equivalente de la forma de la ecuación 8.2 y, de inmediato, no resulta claro cuál de las dos debemos usar. A veces hay razones biológicas fuertes para preferir una sobre la otra (ecuación 8.1 sobre 8.2). Estos criterios biológicos deben ser usados cuando sea posible para decidir. En otro caso, el paso siguiente es graficar los datos transformados logarítmicamente (véase el Gráfico 8.2). En el caso que se presenta, los dos gráficos muestran claramente que los datos tienen la misma forma de la ecuación 8.1, dado que solamente la transformación log-lineal resulta en una línea recta.

Gráfico 8.1

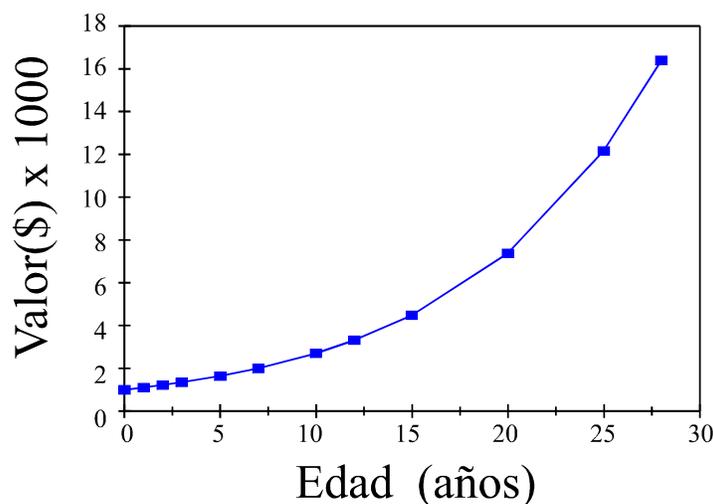
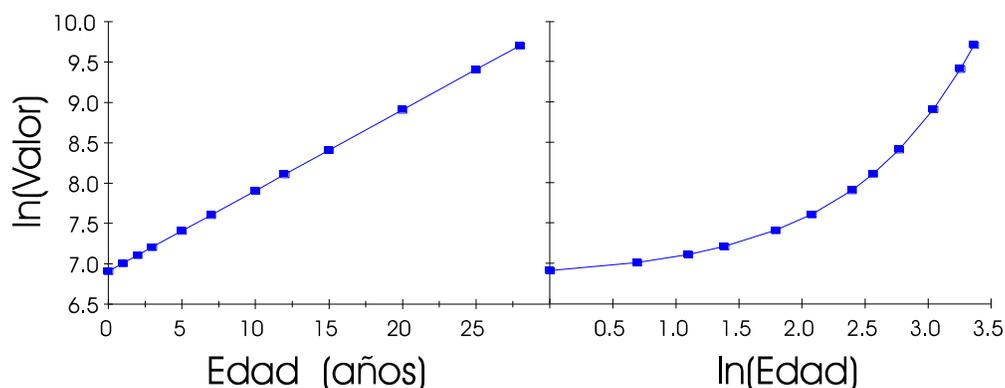


Gráfico 8.2



Lamentablemente, cuando existe mucha dispersión en los datos puede ser difícil decidir cuál de los dos modelos se adapta mejor. En tal caso sería necesario hacer un análisis de regresión para cada uno y escoger el que se adapte mejor a los datos. También puede ser posible que exista un tercer modelo (véase por ejemplo el Gráfico 7.1), el cual se adapta mejor que las ecuaciones 8.1 o 8.2 en cuyo caso se debe aplicar otro tipo de transformación.

Una vez decidido que los datos tienen la forma $Y = \alpha + e^{\beta X}$, se procede con la regresión lineal de los datos transformados para estimar los parámetros α y β . Se encuentra que:

$$r^2 = 1.000$$

para $H_0: \beta = 0$ y $H_a: \beta \neq 0$ $t = 2582.25$, $\nu = 10$

Por lo tanto se rechaza H_0 con un nivel de significancia de $P < 0.001$

La ecuación de regresión es:

$$\ln(\text{valor}) = 6.9074 + 0.1 \times \text{edad}$$

Como se mencionó en el Capítulo II, cuando se transforman datos usando $X+1$ o $Y+1$ se debe corregir por el $+1$. Afortunadamente, la adición de una constante a X o a Y no afecta la pendiente (b). Sin embargo, estas transformaciones sí afectan el valor de 'a' (intersección). Para corregir este efecto debemos seguir los siguientes cálculos:

$$a' = a + bc_x - c_y \quad (8.5)$$

donde: a' es la intersección corregida - a

c_x es la constante que se adiciona a X

c_y es la constante que se adiciona a Y

En el caso que se presenta no usamos una constante ($c_x=0$, $c_y=0$), entonces no tenemos que corregir la ecuación.

Para poner esta ecuación otra vez en la forma de la ecuación 8.1 es necesario hacer una transformación inversa de los datos mediante una potenciación de ambos lados de la ecuación:

$$e^{\ln(\text{valor})} = e^{6.9074 + 0.1 \times \text{edad}}$$

$$\text{Valor} = 1000.0 \times e^{0.1 \times \text{edad}}$$

Una vez más, hay que tener cuidado al interpretar esta ecuación. Primero que todo, este es un modelo puramente predictivo que no pretende describir la biología o la economía detrás de la evaluación de los valores. Un segundo gran cuidado que se debe tener es el de predecir el valor de un lote de madera de más de 28 años de edad — fuera del rango de los datos usados en la regresión **Error! Bookmark not defined.** Por ejemplo, ¿el modelo predice que a una edad de 200 años un lote de madera tendría un valor de \$480,000,000,000! ¡Ni siquiera la inflación crece tan rápido! Es obvio que este modelo es inadecuado para árboles maduros. La única manera de modelar la función de valor para lotes de árboles de edad es tener datos para tales lotes.

Introducción a la Correlación

Simple

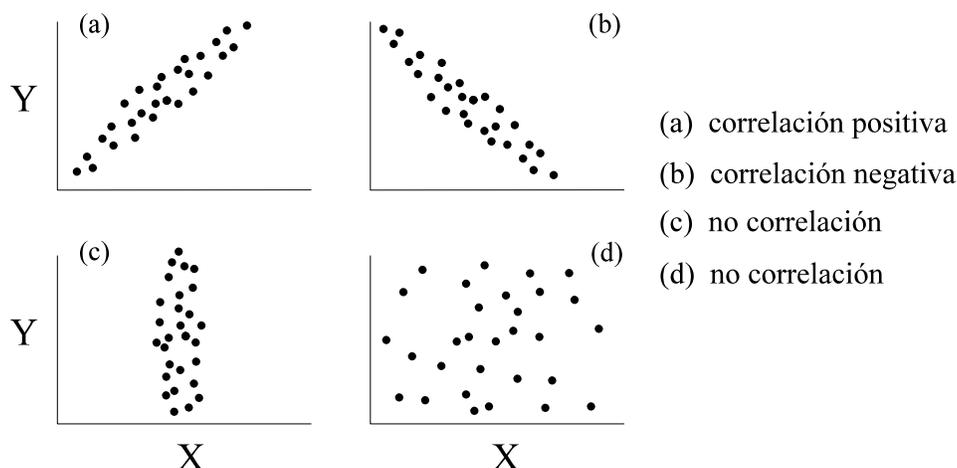
Hasta este punto, se ha considerado el caso donde se asume que una variable Y es la variable dependiente y la otra variable X es independiente. En correlación las dos variables pueden ser intercambiadas sin alterar la interpretación de los resultados. Además no se supone una relación funcional entre las variables.

El **Coefficiente de Correlación** (también conocido como el **Coefficiente de Correlación de Pearson**) se calcula:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad (8.6)$$

Dado que el producto $\sum xy$ puede ser positivo, cero, o negativo, r puede ser positivo, cero, o negativo. Una correlación positiva implica que un incremento en una de las variables implica un incremento en la otra. En una correlación negativa al incrementar una variable, la otra disminuye. Una correlación de cero significa que no hay relación lineal entre las variables (esto es, que un cambio en la magnitud de una variable no implica un cambio en la magnitud de la otra). El Gráfico 8.3 presenta estas posibilidades.

Gráfico 8.3



El rango de r va de un mínimo de -1 a un máximo de 1 y no tiene unidades de medida.

$$-1 \leq r \leq 1$$

Notase que el coeficiente de correlación (r) y el coeficiente de determinación (r^2) son muy similares y que este último puede ser calculado elevando al cuadrado el primero. Aunque están matemáticamente relacionadas, las dos cantidades tienen diferentes interpretaciones. r^2 se puede considerar como una medida de la fuerza de la relación de línea recta entre X y Y. Por otro lado, r no es la medida del cambio de una variable con respecto a otra, sino una medida de la intensidad de asociación entre las dos.

Como ejemplo, se tienen datos de longitudes del ala y cola para una especie particular de ave.

Tabla 8.2

<u>Longitud de Ala (cm)</u> (X)	<u>Longitud de Cola (cm)</u> (Y)
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

$$n = 12$$

$$\sum X_i = 128.2\text{cm} \quad \sum Y_i = 90.8\text{cm} \quad \sum X_i Y_i = 971.4\text{cm}^2$$

$$\sum X_i^2 = 1371.32\text{cm}^2 \quad \sum Y_i^2 = 688.40\text{cm}^2$$

$$\sum x^2 = 1.72\text{cm}^2 \quad \sum y^2 = 1.35\text{cm}^2 \quad \sum xy = 1.32\text{cm}^2$$

$$\text{Coeficiente de correlación} = r = \frac{1.32}{\sqrt{1.72 \times 1.35}} = \mathbf{0.8703}$$

Evaluación de la Significancia de una Correlación **Error! Bookmark not defined.**

El coeficiente de correlación que se calculó a partir de una muestra es una estimación del coeficiente de correlación de la población **Error! Bookmark not defined.** de la que sacamos la muestra. Se ha dado a este parámetro el símbolo ρ (símbolo griego que se pronuncia **rho** **Error! Bookmark not defined.**). Si se desea preguntar si en efecto hay una correlación entre las dos variables en la población se puede probar la hipótesis:

$$H_0: \rho = 0 \text{ contra } H_a: \rho \neq 0$$

Podemos realizar esto utilizando la familiar prueba de Student **Error! Bookmark not defined.**:

$$t = \frac{r - \rho}{S_r} \quad \text{con } \nu = n - 2 \quad (8.7a)$$

donde: ρ es el valor de la hipótesis nula **Error! Bookmark not defined.** que no hay una correlación.

Cuando $\rho=0$, el mismo estadístico se puede calcular como:

$$t = \frac{r}{S_r} \quad (8.7b)$$

donde: S_r es el error estándar **Error! Bookmark not defined.** de r y se calcula como:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} \quad (8.8)$$

Utilizando los datos del ejemplo:

$$\begin{aligned} S_r &= \sqrt{\frac{1 - 0.8703^2}{12 - 2}} = 0.1558 \\ t &= \frac{0.8703}{0.1558} = \mathbf{5.5872} \end{aligned} \quad (8.9)$$

$t_{0.001(2),10} = 4.587 \Rightarrow$ Por lo tanto se rechaza H_0 a un nivel de $P \leq 0.001$.

La misma prueba puede ser realizada directamente usando las Tablas de valores críticos de r (véase la Tabla en página 133). Para encontrar el valor crítico mire el valor para los grados de libertad **Error! Bookmark not defined.** ($\nu = n - 2$) y el nivel de significancia (α). Para $\nu = 10$ y $\alpha = 0.001$ el valor crítico $r_{0.001(2),10} = 0.823$. Como el valor de r que se obtuvo (0.8703) es mayor que este valor crítico, de nuevo se rechaza H_0 a un nivel de $P \leq 0.001$.

