Daena: International Journal of Good Conscience. 7(3) 67-81. Noviembre 2012. ISSN 1870-557X

Análisis de Regresión Lineal Simple para Predicción

(Simple Linear Regression Analysis for Prediction)

Badii, M.H.; A. Guillen; E. [1]Cerna; J. [1]Valenzuela & J. [1]Landeros

UANL, San Nicolás, N.L. & UAAAN, Buenvista, Coah., México

Resumen. Se analizan las nociones de regresión y correlación lineal simple, presentando ejemplos para clarificar el papel de estos modelos estadísticos en predicción de los procesos o fenómenos. Se explica la forma de verificar la significancia estadística de los parámetros de dichos modelos y se abunda sobre la noción de intervalo de confianza para cada uno de los parámetros de regresión.

Palabras claves. Correlación, intervalo de confianza, parámetros, regresión

Abstract. Simple linear regression and correlation are analyzed emphasizing the role these models play in predicting processes or phenomenon. Methods for validation of statistical significance of regression parameters are explained and the notion and calculation of confidence intervals for these parameters are discussed.

Keywords. Confidence intervals, correlation, parameters, regression.

Introducción

Estadística, ciencia que estudia las probabilidades, en base a la recolección, análisis e interpretación de datos, ya sea para ayudar en la resolución de la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Según Sokal & Rholf (2006), la bioestadística se puede definir como el estudio científico de datos numéricos basados en fenómenos naturales.

La regresión y la correlación son dos técnicas estrechamente relacionadas y comprenden una forma de estimación. En forma más especifica el análisis de correlación y regresión comprende el estudio de los datos muestreales para saber qué es y cómo se relacionan entre sí dos o más variables en una población. El análisis de correlación produce un número que resume el grado de la correlación entre dos variables; y el análisis de regresión da lugar a una ecuación matemática que explica y predice dicha relación.

El análisis de correlación generalmente resulta útil para un trabajo de exploración cuando un investigador trata de determinar que variables son potenciales importantes, el interés radica básicamente en el grado de la relación y la regresión da lugar a una ecuación que describe, explica y predice dicha relación en términos matemáticos

Según Badii et al. (2007), en la práctica se ha notado que cuando en un individuo, un carácter (variable) se presenta en mayor intensidad, otro se afecta en alguna proporción. Ejemplos de estos podrán ser la altura y el peso en ganado, edad y vigor de rebrote en árboles, etc.; en el primer ejemplo se puede ver que esta relación es positiva, es decir, al aumentar una variable tiende a aumentar el otro, en cambio, en el segundo ejemplo esta relación es negativa, pues al aumentar la edad, el vigor tiende a disminuir. Por lo tanto, los datos necesarios para análisis de regresión y correlación provienen de observaciones de variables relacionadas.

Objetivos generales

1. Calcular el coeficiente de la correlación entre dos variables.
2. Graficar un diagrama de dispersión.
3. Representar la recta que define la relación lineal entre dos variables.
4. Estimar la recta de regresión por el método de mínimos cuadrados.
5. Usar el método de Ji^2 para estimar el ajuste entre datos observados y datos estimados
6. Realizar una prueba de significancia (hipótesis) para determinar si el coeficiente de correlación (*r*), coeficiente de regresión (*b*) y si la intersección con la ordenada (*a*) difieren de cero, es decir la significancia de *r*, *b* y *a*.

Desarrollo

Cuando solamente dos variables están involucradas en el análisis de Regresión y Correlación, se dice que la técnica es una Regresión o Correlación Simple.

Cuando están implicadas tres o más variables, se tratará de una Regresión o Correlación Múltiple.

Coeficiente de Regresión: la técnica de Regresión se refiere al procedimiento de obtener una ecuación con fines de estimación o predicción.
Variable Dependiente: o variable respuesta es la variable a estimar o predecir.
Variable Independiente: o variable predictora aquella variable que proporciona la base para la estimación.
Regresión Simple: existe solamente una variable independiente y una variable dependiente.
Regresión Múltiple: implica dos o más variables independientes y una variable dependiente.

Coeficiente de correlación

Se ocupa de la medición de la cercanía de la relación entre las dos variables involucradas. Es la asociación de estas, más no de la dependencia, por lo tanto y no olvidar que los cambios de valores de una variable no es la causa del cambio de la otra. Denominado *r* (ecuación siguiente), medida usual del grado de correlación basándose en una muestra de *n* pares de observaciones.

$$r=\frac{\sum(xy)-\frac{(\sum x*\sum y)}{n}}{\sqrt{\left[\sum(x^2)\sum\frac{(x)^2}{n}\right]\left[\sum(y^2)\sum\frac{(y)^2}{n}\right]}}$$

Relación Funcional. Es una relación que permite la predicción de la variable dependiente por la(s) variable(s) independiente(s).
Coeficiente de Regresión o la pendiente (***b*).** El coeficiente de regresión mide la relación causa–efecto entre las variables. En otras palabras, la *b* mide el grado de dependencia de *y* sobre la *x*, es decir, el grado de cambio en *y* en función del cambio Unitario en *x*.

Estadísticamente, se puede medir este grado de relación o dependencia, mediante el índice conocido como coeficiente de regresión, denotado por β (parámetro poblacional), y como ***b*** para el estimador muestreal. En realidad, gráficamente, el valor del coeficiente de regresión es la pendiente promedio, o la pendiente de la línea de la tendencia del comportamiento de ambas características estudiadas (Badii et al., 2007).

El estudio de estos temas puede hacerse desde el caso más simple (regresión lineal simple) hasta formas más complicadas, en donde intervengan en forma lineal o aditiva más de dos factores, e inclusive para formas no lineales, plinomiales, armónicas, modelos lineales estructurados (Badii et al, 2007). En el presente estudio solo nos ocuparemos del caso más simple, es decir, aquél en el que sólo intervengan dos caracteres o variables. Generalmente, a una de las variables se le denomina como independiente o predictora (denotada por *X*) y a la otra como dependiente o de respuesta (denotada como *Y*) (Badii et al., 2009).

Correlación Simple. El grado de precisión en la predicción depende de la cercanía de la relación entre ***X*** y ***Y*,** lo cual también se conoce como **Grado de Correlación o asociación** entre las dos variables. Es un modelo matemático que explora la dependencia entre dos variables cuantitativas (supone que en el modelo una es la variable dependiente y otra la independiente), tratando de verificar si la citada relación es lineal y aportando unos coeficientes (***a*** y ***b***) que sirven para construir la ecuación de la recta de predicción. Ambas técnicas, basadas en la media y en la varianza de las variables evaluadas, tienen importantes condiciones de aplicación, entre las que destacan la independencia de las observaciones sobre la variable independiente, la normalidad e independencia de los residuales y la homogeneidad de las varianzas, disponiéndose de alternativas no paramétricas para la correlación cuando estas no se cumplen.

La correlación simple puede presentarse de dos formas: **Correlación Positiva:** cuando ***r*** está entre 0 y +1. Los valores de *X* y *Y* tienden a moverse en la misma dirección. **Correlación Negativa**: cuando ***r*** está entre -1 y 0. Entonces los valores de *X* y *Y* tienden a moverse en dirección opuesta; cuando uno aumenta el otro tiende a disminuir y viceversa.

Regresión lineal. Se refiere a una relación que puede representarse gráficamente mediante una línea recta que describe la dependencia entre dos variables, la que puede ser positiva o negativa (Figuras 1a & 1b).

Objetivo de la regresión lineal: al evaluar la relación entre dos variables es realizar predicciones cuantitativas. La regresión puede utilizadas en diversas situaciones. Se emplean en situaciones en la que las dos variables miden aproximadamente lo mismo, pero en las que una variable es relativamente costosa, o, por el contrario, es poco interesante trabajar con ella, mientras que con la otra variable no ocurre lo mismo (Sokal & Rohlf, 2006). El análisis de regresión únicamente indica qué relación matemática podría haber, de existir una. Ni con regresión ni con la correlación se pude establecer si una variable tiene "causa" es decir provoca cambio en los valores de otra variable, por lo tanto este modelo solo debe utilizar cuando a priori ya se estableció la relación causa-efecto o la dependencia entre las variables bajo el estudio.

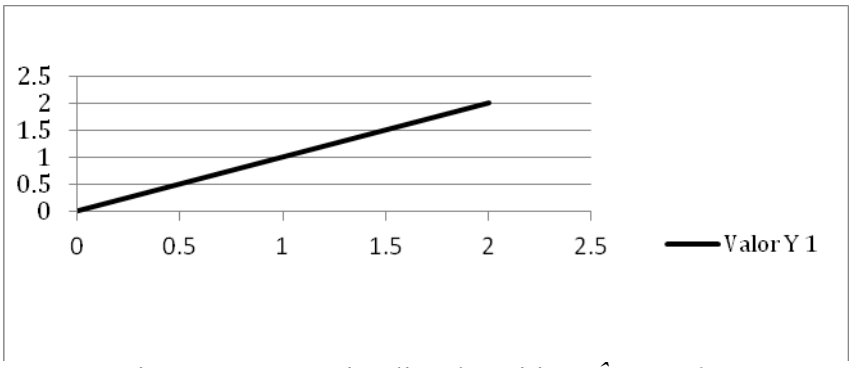


Figura 1a. Regresión lineal positiva: $\hat{Y} = a + bX$

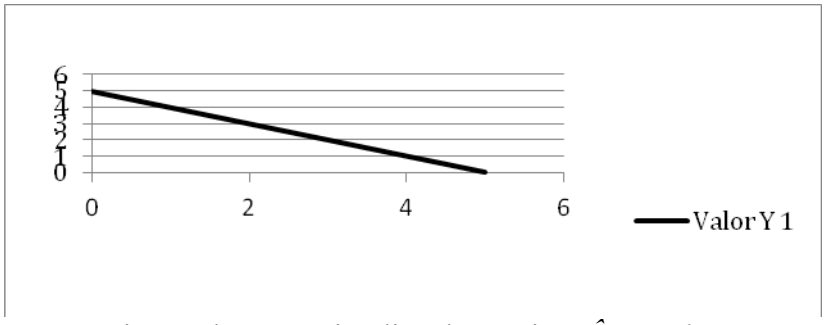


Figura 1b. Regresión lineal negativa: $\hat{Y} = a - bX$

Ecuación Lineal

Dos características importantes de una ecuación lineal son: (1) la pendiente de la recta y (2) la localización de la recta en algún punto. Una ecuación lineal tiene la forma.

$$\hat{Y} = \boldsymbol{a} + \boldsymbol{b}\mathbf{X}$$

En la que *a* y *b* son estimaciones que se determinan a partir de los datos de la muestra. Donde, *a*: indica la altura de la recta cuando $X = 0$. *b*: señala su pendiente de la línea. La variable $\hat{Y}$ es la que se habrá de predecir, y *X* es la variable predictora.

Determinación de la ecuación matemática. En la regresión, los valores de *Y* son predichos a partir de valores de *X* dados o conocidos.

Métodos de mínimos cuadrados. El procedimiento más utilizado por adaptar una recta a un conjunto de punto se le que conoce como método de mínimos cuadrados. La recta resultante presenta 2 característica importantes.
- es nula la suma desviaciones verticales en los puntos a partir de la recta.
- es mínima la suma de los cuadrados de dicha desviaciones.

$$(Ye - Yi)^2$$

Donde,
$Ye = \hat{Y}$ = valor esperado de Y
Yi = valor calculado de Y utilizando la ecuación de mínimos cuadrados con el valor correspondientes X para Yi.
Los valores de a y b para la recta $\hat{Y} = a + bX$ se calculan de tal forma que minimiza la suma de los cuadrados de la desviaciones "ecuaciones normales."

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum(xy) - \frac{\sum x * \sum y}{n}}{SCx}$$

Ejemplo 1. Se eligieron 10 personas desde el día de su nacimiento.
Preguntas. (1) Cuál es el grado de asociación entre la variable altura (X) y peso (Y) y el grado de dependencia entre éstas dos variables. **(2)** Calcular la línea de regresión para pronosticar o predecir $\hat{Y}$ y el error estimado. Los datos y la gráfica se indican en la Tabla 1 y la Figura 2.

Tabla 1. Relación de la altura (cm) y el peso (kg).

Individuos	X Altura en Cm	Y Peso en Kg
1	50	3
2	55	3.9
3	60	5.8
4	65	8.0
5	70	11.0
6	75	11.3
7	80	12.4
8	100	16.7
9	121	32.0
10	145	46.2

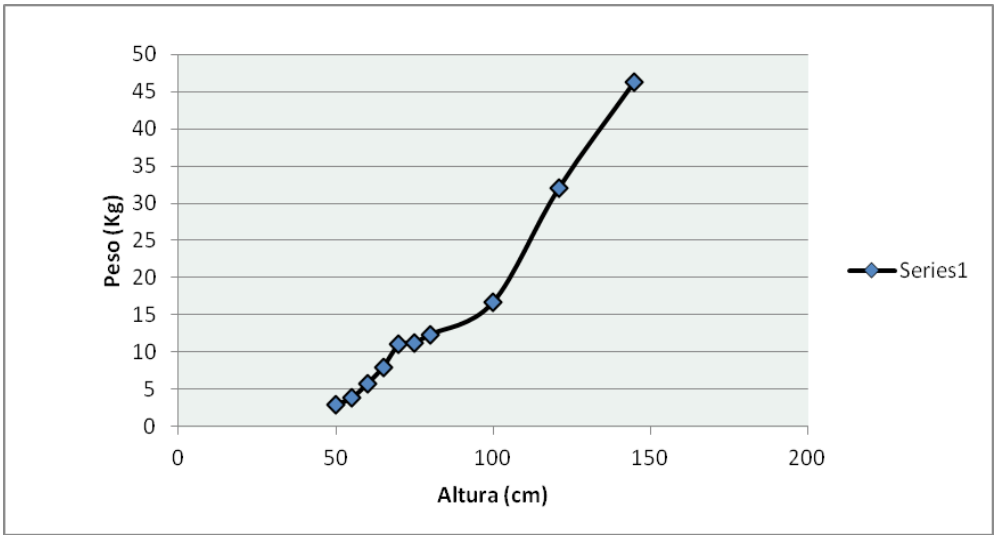


Figura 2. Relación de la altura (cm) y el peso (kg).

$r = [\sum XY - (\sum X * \sum Y)] / \sqrt{SC_X * SC_Y}$
Donde,

SC_X = Suma de cuadrados de X lo cual es igual a: $\sum X^2 - (\sum X)^2/n$
SC_Y = Suma de cuadrados de Y lo cual es igual a: $Y\,Y^2 - (\sum Y)^2/n$

Datos necesarios:

$$r = \frac{16083 - \frac{821 \times 150.3}{10}}{\sqrt{(856.9)(1702.63)}}$$

$$r = \frac{3483.37}{3512.50} = 0.98$$

Altura (X)	**Peso (Y)**
$\overline{X}$ = 82.1	$\overline{Y}$ = 15.03
$\sum X$ = 821	$\sum Y$= 150.3
$\sum X^2$= 75.941	$\sum Y^2$ = 3961.63
$\sum XY$ = 16083	

Existe un 98 % de asociación o correlación positiva entre las variables peso y altura, por lo tanto, a medida que aumenta la altura también aumenta el peso a una tasa de 98% o 98 en 100 veces.

Grado de dependencia o coeficiente de regresión (*b*).

$$b = \frac{\sum(xy) - \frac{\sum x \times \sum y}{n}}{SC_X}$$

$$b = \frac{16 - \frac{82.1 \times 150.3}{10}}{856.9}$$

$$SC_X = \sum X^2 - \frac{\sum(X)^2}{n}$$

b = 0.438 (43.8%). Es decir, la b mide el grado del cambio o aumento de *Y* en función del cambio "unitario" de *X*. Por lo tanto, a medida que una persona crece 1 cm, esperamos que el peso aumente 43.8% de 1 Kg lo que sería 438 gr por cada cm de altura, con un 98% de asociación o correlación entre las dos variables.

Línea de regresión para pronosticar. Se quiere estimar los valores de *Y* esperada o estimado ($\hat{Y}$) para poder calcular el error estimado de nuestro modelo y así saber cuan erróneo es (Spiegel & Stephens, 2001). Para el cálculo del error estimado primero se necesita tener la ecuación siguiente de $\hat{Y}$ (*Y* estimado). La ecuación funcional valores de $\hat{Y}$ y los errores de mínimos cuadrados se indican en la Tabla 2. La relaci

$$\hat{Y} = a + bX$$

$a = \bar{y} - b\bar{x}$

a = 15.03 – (0.438 x 82.1) = **-20.92**

Ecuación funcional de predicción: $\hat{Y}$ **= -20.92 + (0.438** ***X*****)**

Tabla 2. Valores de $\hat{Y}$ y los errores de mínimos cuadrados de datos de Tabla 1.

individuos	*X* Altura (Cm)	*Y* Peso (Kg)	$\hat{Y}$	$(\hat{Y} - Y)$	
1	50	3	.98	-2.02	
2	55	3.9	3.17	-0.73	
3	60	5.8	5.36	-0.44	
4	65	8.0	7.55	-0.45	
5	70	11.0	9.74	-1.26	
6	75	11.3	11.93	0.63	
7	80	12.4	14.12	1.72	
8	100	16.7	22.88	6.18	
9	121	32.0	32.07	0.07	
10	145	46.2	42.59	-3.61	

$\sum(\hat{Y} - Y) = 0.09 \approx 0$ por el redondeo. Como se ve los errores son mínimos lo cual es una característica de la regresión lineal (Figura 3).

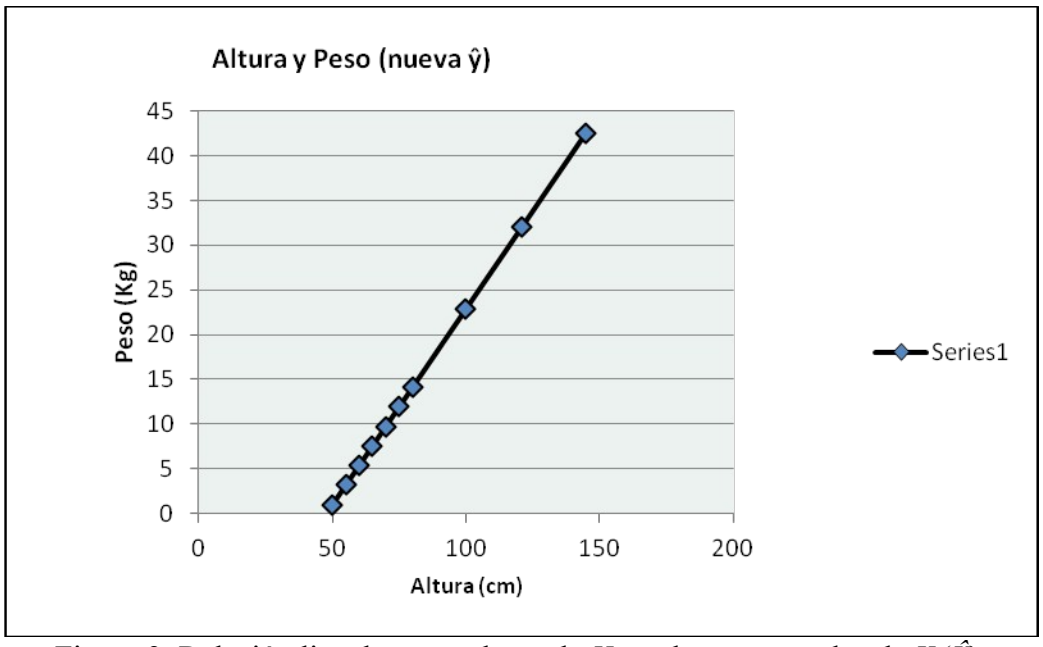


Figura 3. Relación lineal entre valores de X y valores esperados de Y ($\hat{Y}$).

Se puede observar que la nueva gráfica (Figura 3) con los nuevos valores de ŷ es más representativo que el primero (Figura 2) de acuerdo a una regresión lineal.

Pruebas de significancia. Una vez que hemos calculado la recta de regresión, el siguiente paso consiste en analizar si la regresión en efecto es significativa y la podemos utilizar para predicción de los valores de $\hat{Y}$ en función del cambio unitario en los valores de X. Para ello debemos contrastar si la coeficiente de regresión entre ambas variables es distinta de cero o si el modelo de regresión es significativo en el sentido de contrastar si el análisis de nuestra variable endógena (Y) es válido a través de la influencia de la variable explicativa (X).

Aceptación o rechazo de las hipótesis del modelo en estudio, ya sea para coeficiente de regresión (Ho: $b = 0$ vs. Ha: $b \neq 0$), coeficiente de correlación (Ho: $r = 0$ vs. Ha: $r \neq 0$) y la intersección "a" es decir, altura de la línea a partir de la intersección con la ordenada (Ho: $a = 0$ vs. Ha: $a \neq 0$) determina la significancia de dichos parámetros.

Supongamos por un lado que el coeficiente de correlación lineal r, y el coeficiente de regresión (b) tienen valores muy altos, y por tanto solo por sus magnitudes parece indicar la existencia de una correlación y dependencia alta entre los valores de la muestra. Pero la magnitud alta de estos coeficientes muestreales entre ambas variables no necesariamente refleja la misma situación en la población. Para poder contrastar esta suposición, una vez que hemos estimado la recta de regresión y hemos obtenido las estimaciones de los parámetros del modelo, debemos comprobar si estas estimaciones del modelo son significativas de tal forma que la variable (X) es relevante para explicar la variable de respuesta (Y). Entonces debemos contrastar si las "r" y "b" de la recta de regresión poblacional son significativamente distintos de cero, de ahí tendríamos que, en efecto, existe una relación (r) y una dependencia (b) significativa entre ambas variables poblacionales.

Las hipótesis que se ponen a prueba indican que no existen diferencias en las medias de las poblaciones en los diferentes niveles del factor (Ho), es decir que la variable de

respuesta no difiere entre los grupos y que, por lo tanto, la variable independiente no tiene un efecto sobre la variable de respuesta. A continuación se demuestran las pares de hipótesis para cada parámetro.

Hipótesis de coeficiente de correlación (*r*)

r

Ho: $r = 0$ Variables no asociadas. (sin significancia estadística).

Ha: $r \neq 0$ Asociación de las variables. (Con significancia estadística).

Hipótesis para coeficiente de regresión (*b*):

b

Ho: b = 0 No dependencia de las variables. (sin significancia estadística).

Ha: b ≠ 0 Dependencia de la variables. (Con significancia estadística).

Hipótesis de la la intersección con la ordenada (*a*)

a

Ho: $a = 0$ Línea de regresión sale del origen ($X = 0$ & $Y = 0$).

Ha: $a \neq 0$ Línea de regresión no sale del origen.

Para poder comprobar las hipótesis planteadas se utilizan pruebas de comparación de estimadores, como la prueba de t-student (parámetro de comparación de las medias de las variables en estudio). En este caso, lo que se desea investigar es si los promedios de las muestras sometidas a diferentes métodos o tratamientos (distintos niveles de algún factor de variación), manifiestan diferencias significativas, es decir, si los intervalos de confianza de los valores paramétricos estimados no se traslapan.

Cuando sólo se tienen dos niveles, lo común es realizar una prueba de t-student, si se tienen más de dos tratamientos, lo común es realizar un análisis de varianza (ANOVA), lo que se puede clasificar de manera siguiente:

1. Prueba de t para una muestra.
2. Prueba de t para comparación de dos muestras relacionadas.
3. Prueba de t para comparar dos muestras independientes.
4. Análisis de varianza para comparar más de dos medias muestreales (Badii et al., 2009).

Las formulas estadísticas de t-student para comprobar la significancia de las hipótesis planteadas para los coeficientes de regresión y correlación son los siguientes:

1. t-student para el coeficiente de regresión (*b*):

$$t_b = \frac{b_{obs} - 0}{\varepsilon_b}$$

$$\varepsilon_b = \sqrt{V_E / S_X}$$

ANOVA para la regresión

Fuente de variación	Gl	Suma de cuadrados (SC)	Cuadrados medios o varianzas
Regresión	1	$b\sum SC_X$	$V_{Reg} = SC_{Reg.}/gl$
Error		$SC_{Total} - SC_{Reg.}$	$V_{Error} = SC_{Error}/gl$
Total	n – 1	$\sum Y^2 - \frac{(\sum y)^2}{n}$	—

2. t-student para el coeficiente de correlación (*r*):

$$t_r = \frac{r_{obs} - 0}{\varepsilon_r}$$

$$\varepsilon_r = \sqrt{1 - r^2 / n - 2}$$

3. t-student para la pendiente la intersección con la ordenada (*a*):

$$t_a = \frac{a_{obs} - 0}{\varepsilon_a}$$

$$\varepsilon_a = \sqrt{V_E \left[\frac{1}{n} + \frac{m_X^2}{S_X} \right]}$$

Ejemplo 2. Supongamos una población de mariposas elegidas en forma aleatoria. Analizar la relación entre *X* y *Y* (aumento del largo del ala (cm) a medida que van creciendo (días) y si existe asociación significativa entre estas dos variables, para saber que tan real son los datos de la colonia.

Tabla 2. Datos de la población de mariposas.

Edad en días (*X*)	Longitud de ala en cm (*Y*)
3	1.4
4	1.5
5	2.2
6	2.4
8	3.1
9	3.2
10	3.2
11	3.9
12	4.1
14	4.7
15	4.5
16	5.2
17	5.0
$\sum X = 130$	$\sum Y = 44.4$
$\sum X^2 = 1562$	$\sum Y^2 = 171.3$

$\sum XY = 514.8$

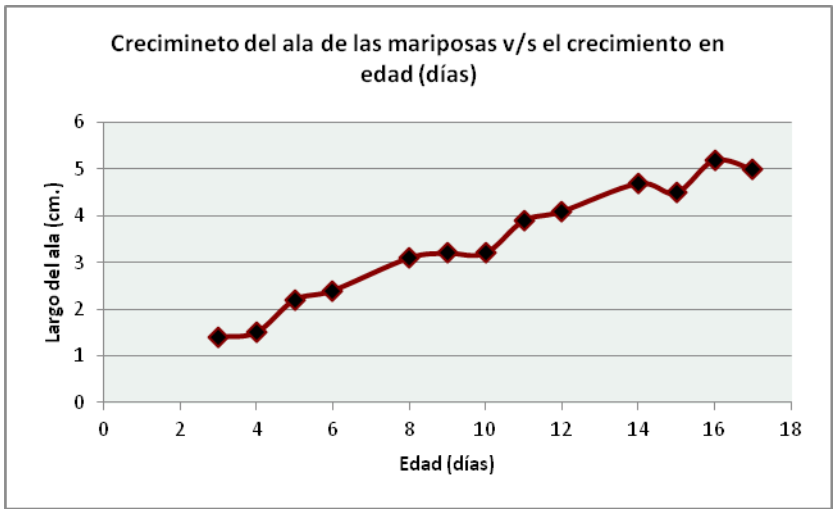


Figura 3. Población de mariposas con sus respectivos largos de alas.

1.- Coeficiente de correlación:
r = 0.9866 ≈ 99 %
Existe una asociación de 99% entre la edad de las mariposas y el crecimiento del ala.

2.- Grado de dependencia de las variables (regresión):

SCx = 262
b = 0.27 ≈ 27%

$$b = \frac{\sum(xy) - \frac{\sum x * \sum y}{n}}{SCx}$$

Cada día que crecen las mariposas, las alas les aumentan un 27%, lo que correspondería a 2.7 mm por día. Con estos resultados podemos predecir cuánto tiempo se puede demorar en crecer una mariposa y en cuanto tiempo va a llegar a ser adulta en término de largo de alas.

Para poder determinar una línea de predicción de crecimiento de las alas, desde el primer día de nacimiento (si las alas comienzan a aparecer una vez que nace la mariposa o si ya nace con una determinada longitud de alas) mediante la ecuación de regresión**:**

$a = \bar{y} - b\bar{x}$ a = 0.7 $\hat{Y} = 0.715 + 0.27X$ Ecuación de línea para predicción

Pruebas de significancia o comparación de las Hipótesis para Regresión, Correlación y la pendiente dentro de la población de mariposas que se están evaluando en el ejemplo:

1.- Significancia en la correlación o asociación entre el crecimiento de las mariposas y el aumento del ala.

$$t_b = \frac{b_c - 0}{\varepsilon_b} \qquad \varepsilon_b = \sqrt{VE / SCX}$$

ANOVA DE REGRESIÓN

Fuente de variación	gl	Suma de cuadrados (SC)	Cuadrado medios (CM = Varianza)
Regresión	1	19.099	19.099
Error	11	0.558	0.0558
Total	12	19.66	

n total = 13
SCx = 262
$SC_Y = SC_{Total}$ = 19.66
b = 0.27

$$t_b = \frac{0.27 - 0}{\sqrt{0.0558 / 262}}$$

tb **t_T**

18.54 **>** **2.201**

α = 0.05

gl = 11 (n – 2)

Debido a que el valor calculado de t-student es menor que el valor tabulado (**t_T**) con 11 gl, se rechaza Ho. Lo que significa que el crecimiento de las alas depende de forma significativa y positiva en la edad de la mariposa.

2.- Significancia en la asociación entre el crecimiento de las mariposas y el crecimiento de ala.

$$\varepsilon_r = \sqrt{1 - r^2 / n - 2} \qquad t_r = \frac{r_c - 0}{\varepsilon_r}$$

$$\varepsilon_r = \sqrt{1 - 0.98^2 / 13 - 2} \qquad \varepsilon_r = 0.0486$$

$$t_r = \frac{0.98 - 0}{0.0486} \qquad t_r = 19.9272$$

tc **t_T**

α = 0.05

19.92 > 2.201

gl = 11 (n - 2)

El valor calculado es menor que tabulado y por tanto, se rechaza Ho, es decir existe una asociación significativa entre la edad y el largo de las alas.

3.- Verificación si la línea de regresión sale del origen.

$$t_c = \frac{a - 0}{s_a} \qquad s_a = \sqrt{VE\left[\frac{1}{n} + \frac{\bar{x}^2}{SCX}\right]}$$

$$s_a = \sqrt{0.0253\left[\frac{1}{13} + \frac{100}{264}\right]} \qquad s_a = 0.159$$

$$t_c = \frac{0.71 - 0}{0.159} \qquad t_c = 4.465$$

tc **t_T**

4.465 > 2.201 **α = 0.05**, **gl = 11 (n – 2)**

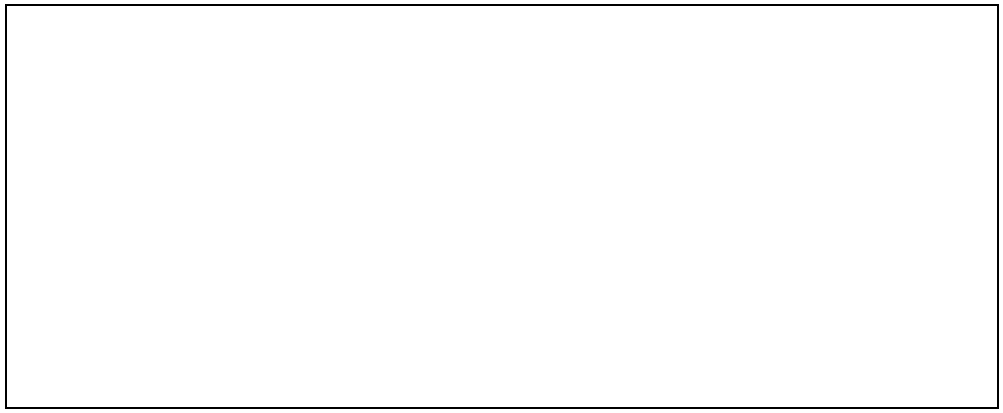
De nuevo el valor pequeño de t calculada en comparación con t tabulada indica que la intersección difiere de forma significativa de cero y por tanto, se rechaza Ho. La línea de regresión no sale del origen, lo que significa que las mariposas no nacen con 0 cm de alas, si no que ya nacen con un largo determinado el que probablemente tenga un rango de longitud.

IC o intervalo de confianza.

Se llama intervalo de confianza en estadística a un par de números entre los cuales se estima que estará cierto valor desconocido (parámetro poblacional) con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un parámetro poblacional. La probabilidad de éxito en la estimación se representa por 1 - α y se denomina nivel de confianza. En estas circunstancias, α es el llamado error nivel uno o nivel de significación, esto es, una medida de las probabilidad de fallar en la estimación mediante tal intervalo, es decir, la probabilidad de rechazar erróneamente una hipótesis cierta (Ostle, 1977).

El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más probabilidad de acierto (mayor nivel de confianza), y viceversa.

Límites de confianza: Son los límites del intervalo de confianza inferior (LIC) y superior (LSC), se determinan sumando y restando a la media de la muestra $\overline{X}$ un cierto número Z de la tabla normal (dependiendo del nivel o coeficiente de confianza) de errores estándar de la media $\sigma_{\overline{X}}$ (Sokal & Rohlf, 2006).



Interpretación del intervalo de confianza: Tener un 95% de confianza en que la media poblacional real y desconocida se encuentra entre los valores LIC y LSC.

El 95% de Nivel de Confianza significa que sólo tenemos un 5% de probabilidad de obtener un punto fuera de ese intervalo. Esto es el 5% total (prueba unilateral), o 2.5% mayor o menor para la prueba bilateral. Si vamos a la tabla Z o normal, veremos que un área de 0.025, corresponde a una Z de 1.960.

Nivel de significancia = 1 - intervalo de confianza = error tipo 1 = alfa

¿Cómo obtenemos un intervalo de confianza?
Estimación puntual $\pm$ error de estimación
¿De dónde viene el error de estimación?

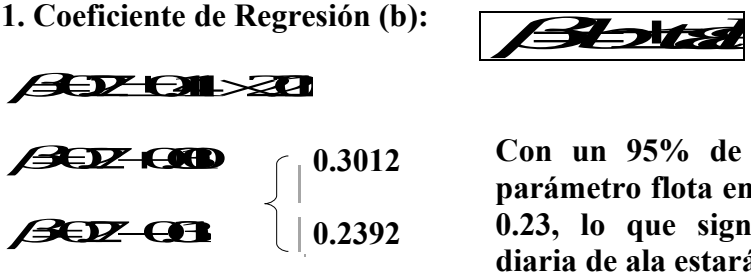
Desviación estándar por multiplicador de nivel de confianza deseado $Z\alpha/2$ (Montgomery et al., 2006).

Objetivos del intervalo de confianza:

1. Si existe diferencia estadística significativa.
2. Si tal diferencia es relevante.

Para el ejemplo del aumento del largo de ala por día de la mariposa, se va a determinar el intervalo para la coeficiente de regresión, correlación y la intersección.

1. Coeficiente de Regresión (b):

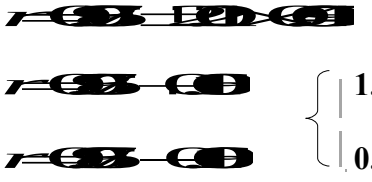


0.3012

0.2392

Con un 95% de confianza (α = 0.05) el parámetro flota entre los extremos de 0.30 y 0.23, lo que significa que el crecimiento diaria de ala estará dentro de éste rango.

2. Coeficiente de asociación (r): 

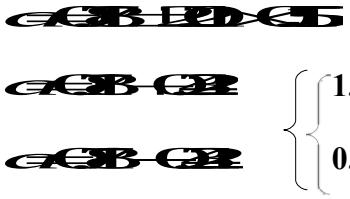


1.0935 (+1)

0.8797 (-1)

Este intervalo por encima de 1 se debe al redondeo ya que la asociación como ya vimos va a ir desde -1 a 1.

3. Intersección con la ordenada(a):



1.0502

0.3658

Al momento de nacer la mariposa tiene un largo de ala de 7.18 mm con *X*= 0 (cero día de edad) y cada día que pasa va aumentando 2.7 mm.

A nivel poblacional el intervalo de crecimiento diario de ala será entre 2.39 a 3 mm con lo que se puede estimar un valor o tiempo aproximado en el que las mariposas alcanzarán su máximo crecimiento de ala o llegaran a ser adultas**.** Por lo tanto, usando el modelo de predicción: $\hat{Y} = a + bX$, y la ecuación funcional con los valores de los parámetros *a* y *b*: $\hat{Y} = .718 + .27X$, cuando la mariposa tiene 3 días de edad se espera que la población tenga 1.52 cm (valor mínimo) de largo del ala y a los 17 días un valor igual a 5.3 cm (valor máximo). Por lo que se espera que el promedio poblacional de ala (44.4/13 = 3.46 cm) esté dentro de estos valores extremos de 1.52 y 5.3.

Referencias

Badii, M.H., J.Castillo, J. Rositas & G.Alarcón. 2007. Uso de un método de pronóstico en investigación . Pp. 137-155. In: M.H. Baddi & J. Castillo (eds). Técnicas Cuantitativas en la investigación. UANL, Monterrey.

Baddi, M.H., J. Castillo, J. Landeros & K. Cortez. 2009. Papel de la estadística en la investigación científica. Pp. 1-43. In: M.H. Badii & J.Castillo (eds). Desarrollo Sustentable: Métodos, Aplicaciones y Perspectivas. UANL. Monterrey.

Montgomery, D.C., E. A. Peck & G.G. Vining. 2006. Introduction to Linear Regression Analysis. Wiley Interscience, USA.

Ostle, B.1977. Estadística Aplicada Técnicas de la Estadística Moderna, Cuando y Donde Aplicarlas. Limusa, México, D.F.
Sokal R.R. & F.J. Rohlf. 2006. Introducción a la bioestadística. Reverté, S.A.
Spiegel, M.R. & L.J. Stephens. 2001. Estadística. Mc.Graw-Hill, México, D.F.