UNIDAD 3. : REGRESIÓN Y CORRELACIÓN.

Expectativas de logro

- Analizar los modelos de regresión lineal simple.
- Caracterizar el análisis de correlación como una metodología estadística para determinar el grado de relación entre variables cuantitativas.
- Reconocer diferencias entre análisis de regresión y correlación, como así también la complementación de ambos métodos.

Contenidos
Regresión de una o dos variables independientes.
Ecuación y línea de regresión. Análisis e interpretación de un modelo de regresión. Usos.
Correlación Interpretación. Coeficientes de correlación y determinación.

ANÁLISIS DE REGRESIÓN LINEAL

INTRODUCCIÓN

El objetivo de este capítulo es introducir el análisis simultáneo de dos variables y adquirir criterios para el uso de las técnicas de regresión y correlación.

Hasta el capítulo anterior se han introducido métodos estadísticos que se pueden utilizar cuando el interés es analizar el comportamiento de una sola variable, eventualmente, bajo distintas condiciones. Pero frecuentemente se presentan situaciones donde se observan dos o más variables sobre cada unidad experimental y el interés se centra en la forma en que estas variables se relacionan.

Algunos ejemplos de relaciones funcionales que pueden aclarar la aplicabilidad de este tema son: la relación entre el rendimiento de un cultivo y la densidad de siembra, la relación entre la altura y el largo de brazo de un grupo de individuos ,la relación entre la cantidad de horas trabajadas y el grado de concentración en la tarea, etc.

En cada uno de estos casos se pueden plantear los siguientes interrogantes:
¿Existe alguna relación entre las variables?
Si se conoce el comportamiento de una de ellas, ¿se puede predecir el comportamiento de la otra?
La estadística aplicada ofrece dos herramientas que permiten dar respuesta a dichas cuestiones: el Análisis de Regresión y el Análisis de Correlación.

El **Análisis de Regresión estudia la relación funcional que existe entre dos o más variables. Identifica el modelo o función que liga a las variables**, **estima sus parámetros** y, eventualmente, prueba hipótesis acerca de ellos. Una vez estimado el modelo es posible predecir el valor de la variable denominada variable dependiente en función de la o las otras variable/s independiente/s y dar una medida de la precisión con que esa estimación se ha hecho.
Dependiendo del objetivo del estudio, los valores o niveles de la/s variable/s independiente/s pueden ser arbitrariamente modificados por el experimentador, es decir el investigador puede fijar los niveles de la variable independiente para los cuales desea estudiar la respuesta de la variable dependiente. El modelo hallado puede ser usado para predecir el comportamiento de la variable dependiente para otros niveles de la variable independiente, que pertenezcan al dominio del estudio.

El Análisis de Correlación lineal estudia el grado y sentido de la asociación lineal que hay entre un conjunto de variables y, a diferencia del análisis de regresión, no se identifica ni se estima explícitamente un modelo funcional para las variables, este siempre se supone lineal. El interés principal es medir la asociación entre dos variables aleatorias cualesquiera, sin necesidad de distinguir variables dependientes e independientes.

En el análisis de correlación, ninguna de las variables puede ser fijada por el experimentador, ya que éste podría seleccionar niveles de las variables que no son frecuentes y esto podría conducir a una estimación errada del grado de correlación.

ANÁLISIS DE REGRESIÓN LINEAL

El término ,regresión, surgió de estudios de la herencia biológica realizados por Galton durante el siglo pasado. En su conocida experiencia, Galton notó que los padres altos tenían hijos cuya altura era mayor a la altura promedio, pero no eran más altos que sus padres. También, padres bajos tenían hijos con altura menor a la altura promedio pero eran más altos que sus padres. Esta tendencia de las características de los grupos de moverse, en la siguiente generación, hacia el promedio de la población o de regresión hacia la media fue descubierta por Galton. El término no tiene hoy el mismo significado que le dio Galton, pero se usa extensamente para referirse al estudio de relaciones funcionales entre variables cuando hay una componente aleatoria involucrada.

Al estudiar la relación entre dos o más variables surge la idea de encontrar una expresión matemática que la describa. En la práctica es posible adoptar modelos de regresión que se pueden agrupar o clasificar en lineales y no lineales. Los primeros hacen referencia a aquellos modelos en que la función adopta la forma de una suma de términos, cada uno conformado por el producto de un parámetro y una variables independiente. Los modelos no lineales son aquellos donde los parámetros no se encuentran multiplicando a las variables independientes como en el modelo lineal de tal forma que no pueden ser estimados resolviendo un sistema de ecuaciones lineales. Por ejemplo, los parámetros pueden encontrarse como **exponentes** de las variables independientes. La estimación de los parámetros en modelos no lineales se realiza usando herramientas diferentes a las presentadas en este capítulo.

El modelo hallado puede ser usado para predecir el comportamiento de la variable dependiente para otros niveles de la variable independiente, que pertenezcan al **dominio** del estudio.

Los gráficos de dispersión son útiles en la etapa exploratoria, tanto en el análisis de regresión como en el de correlación. La representación gráfica de los datos es frecuentemente el punto de partida de cualquier análisis que involucra más de una variable. En los gráficos de dispersión lo que se ve es una nube de puntos, donde cada punto representa una observación.

La Figura 6.1 muestra los gráficos de dispersión usados en estudios de asociación entre dos variables donde además se ha dibujado sobre la nube de puntos, la posible función de ajuste de esos datos, es decir, se ha identificado el modelo funcional de la relación.

Analicemos las representaciones graficas de varias distribuciones bidimensionales que constituyen nubes de puntos o dispersograma.

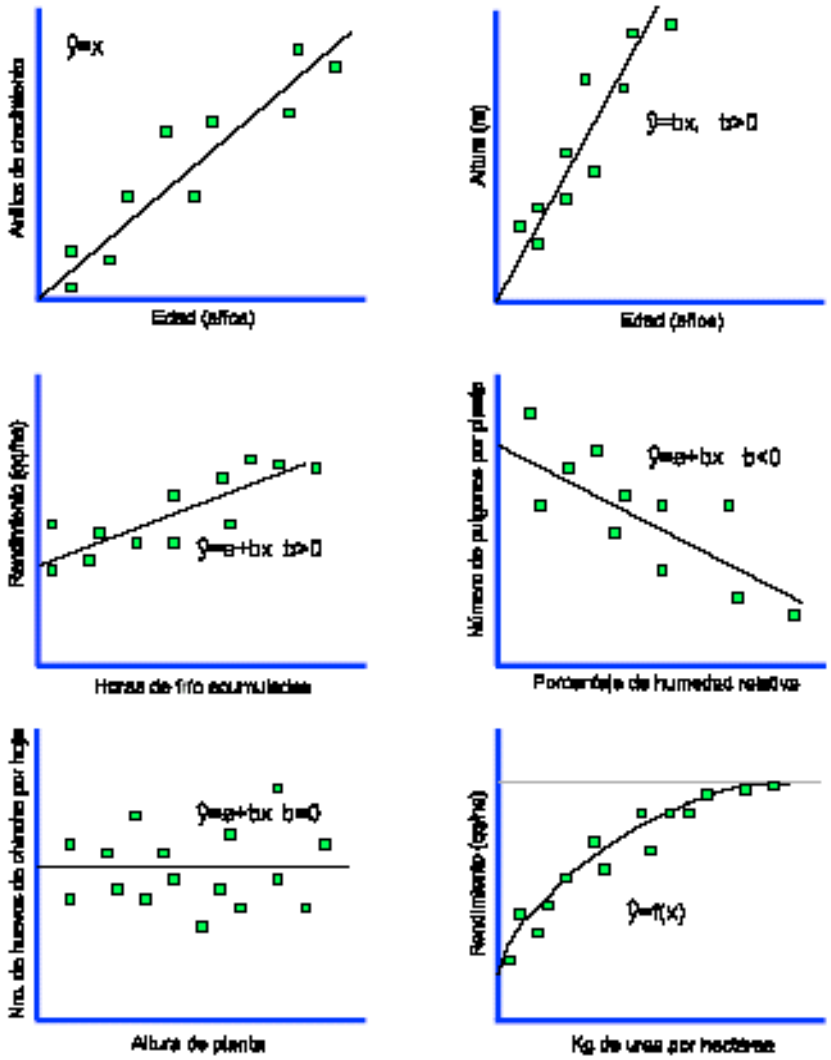


Figura 6.1: Gráficos de dispersión para diferentes modelos de relación entre dos variables.

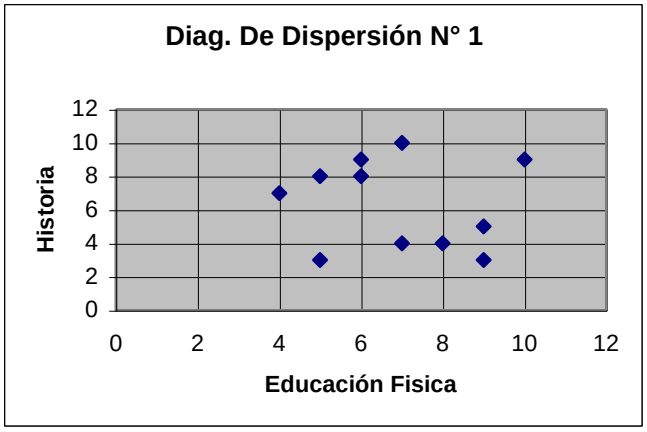
Para cada unidad experimental se graficarán de dos variables en forma conjunta.

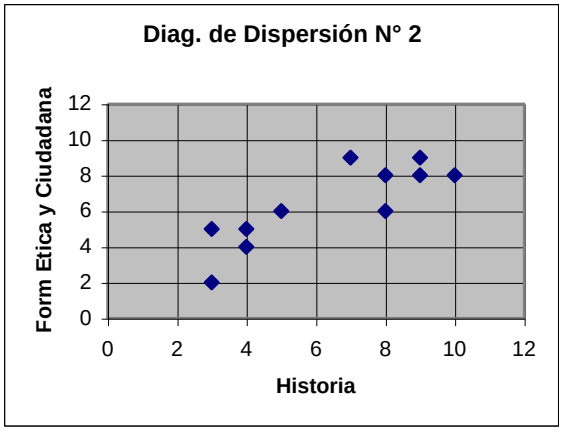
Sea el siguiente ejemplo se presentan calificaciones de un grupo de alumnos en tres materias:

Alumno	A	B	C	D	E	F	G	H	I	J	K
Educ. Física	8	7	9	4	10	9	6	5	7	6	5
Historia	4	10	5	7	9	3	9	8	4	8	3
Form. ética y ciud.	5	8	6	9	9	2	8	8	4	6	5

Un punto de esta grafica representa las dos notas referentes a un mismo alumno. En el gráfico N° 2 se representaron las notas de Historia y Educación Física en el diagrama de dispersión Nº 1 y en el Nº 2 se representan las notas de Historia y las de Formación Ética y Ciudadana.

GRAFICO N° 2





Para el segundo gráfico se observa que hay una cierta relación entre ambas variable (correlación); a mejor nota en Historia se observa una mejor nota en Form. Ética y Ciudadana. Los puntos de esta distribución están próximos a una recta de pendiente positiva, por ello podemos decir que existe una correlación lineal positiva entre ambas variables.
En el primer gráfico no hay una tendencia definida; en este caso se puede afirmar que las variables casi no están correlacionadas.

La simple observación de que dos variables parecen estar relacionadas, no revela gran cosa. Dos importantes preguntas se pueden formular al respecto: ¿Qué tan estrechamente relacionadas se encuentran las variables? o ¿cuál es el grado de asociación que existe entre ambas?

Para responder a la primer pregunta se necesita una medida del grado de asociación entre las dos variables. Esta medida es el **coeficiente de correlación**, que se denota con la letra **r.**
Para medir la correlación hay dos valores extremos: 1 y –1. por consiguiente a cada valor constante comprendido en el intervalo [-1,1], llamado coeficiente de correlación lineal.
En el gráfico 6.1 se observa en los tres primeros el coeficiente de correlación es positivo, mientras que en el siguientes, el coeficiente de correlación es negativo.

Para el caso de dos variables, si se denota como **y** a la variable que se supone dependiente y como **x** a la variable que se postula como independiente, resulta familiar utilizar el concepto de función y decir .**y es función de x**., para indicar que de acuerdo a los valores asignados a **x** se pueden predecir los valores que tomará **y**. Dicho de otra manera, se puede conocer el comportamiento de **y** a través de un modelo que relaciona la variación en **y** con la variación de **x.**
Para determinar el **coeficiente de correlación** se ha establecido la siguiente expresión:

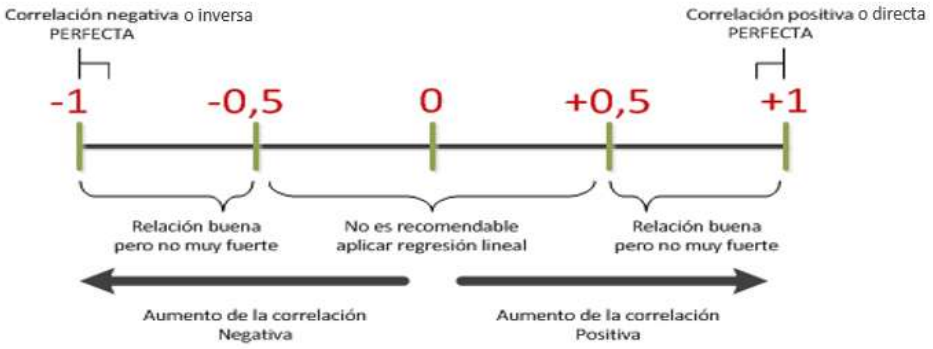
$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y}$$

Si r =0 no hay correlación.

Si $|r|$ =1 la correlación es perfecta

Si r=1 la correlación es directa.

Si r = -1 la correlación es inversa

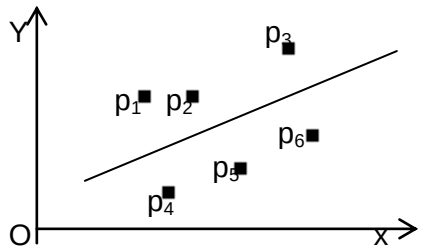


RECTA DE REGRESIÓN

Generalmente, en los casos de correlación lineal, los puntos que representan los pares ordenados de valores de las variables (x,y), no están situados sobre la línea recta (correlación perfecta), sino que se encuentran entorno a ella.

El problema que se plantea es el de determinar la ecuación de la recta mas representativa del conjunto de puntos, es decir, la recta con respecto a la cual la dispersión de los puntos es mínima.

Esta recta se llama recta de regresión.



Para determinar su ecuación se usa el método de los mínimos cuadrados, que consiste en determinar la recta para la cual la suma de los cuadrados de los desvíos de cada punto a la recta sea mínima.

No entraremos en detalle sobre el método para obtener la ecuación, solo nos limitaremos a dar dicha ecuación.

Esta recta pasa por la media de los datos denominados como **x** ,y por la media de los datos de **y**

La ecuación de la **recta de regresión** es: $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

EJEMPLO

Calculemos la correlación para los alumnos de la tabla dada entre Historia y Formac. Ética y Ciudadana.

Alumno	Historia =x_i	Form. E. Y C.=y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\ (y_i - \bar{y})$
A	4	5	-2,3636	-1,3636	3,2230
B	10	8	3,6364	1,6364	5,9504
C	5	6	-1,3636	-0,3636	0,4958
D	7	9	0,6364	2,6364-	1,6778
E	9	9	2,6364	2,6364	6,9500
F	3	2	-3,3636	-4,3636	14,6774
G	9	8	2,6364	1,6364	4,3142
H	8	8	1,6364	1,6364	2,6778
I	4	4	-2,3636	-2,3636	5,5866
J	8	6	1,6364	-0,3636	-0,5949
K	3	5	-3,3636	-1,3636	4,5866
	$\bar{x} = 6{,}36$	$\bar{y} = 6{,}36$			$\sum = 49{,}5429$
	$\sigma_x = 2{,}49$	$\sigma_y = 2{,}14$			

Se determina el **coeficiente de correlación** (Solo trabajaremos en este caso con dos cifras decimales)**:** r = $\frac{1}{11}\frac{49{,}54}{2{,}49\ 2{,}14} \cong 0{,}84$ Este valor nos indica que hay una correlación importante, lo que nos permite continuar con el análisis de regresión

Una vez obtenido el coeficiente de correlación se trazará la recta de regresión, donde esta recta pasa por el punto ($\bar{x}, \bar{y}$), siendo $\bar{x}$ la media de la primer variable, e $\bar{y}$ la media de la segunda variable, y su pendiente está dada por la expresión: $\frac{r\sigma_y}{\sigma_x}$.

Recordando de Matemática que la ecuación de la recta de pendiente conocida (m) por un punto de paso (x_0, y_0) es: ($y-y_0$)= m($x-x_0$)

Por consiguiente la **ecuación de la recta de regresión** es:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Para nuestro caso es:

y-6,36= 0,84 $\frac{2{,}14}{2{,}49}$ (x-6,36) despejando se obtiene y= 0,72 x +1,78

Trace Ud. su gráfica. Utilizando la recta de regresión, estime cual será la nota en Formación Ética y Ciudadana si la nota en Historia es un 6.

Nota: Si te interesa conocer **La desviación o error típico en la recta de regresión** podes encontrar esa información en: Análisis Estadístico multivariable de R Sierra Bravo pag. 15.
En el mismo libro podes encontrar la **Prueba para determinar si la relación entre dos variables es lineal o curvilínea**. Pag 20 **ANEXO N° 6**

A continuación se resumen los conceptos más importantes de:

REGRESIÓN LINEAL

1. El análisis de regresión es una metodología estadística usada para estudiar y modelar la relación entre variables y hacer predicciones.
2. En el caso de regresión lineal simple o bivariada, tendremos dos variables. En el caso de existir más tendremos regresión lineal múltiple.
3. en un diagrama de dispersión se puede observar si existe relación entre las variables de interés. Esta relación podrá ser lineal curvilínea expresable por alguna otra ecuación matemática.
4. En este curso nos concentraremos en las relaciones que se representen por la siguiente ecuación: y = A + B x. Esta es una relación funcional.
5. El análisis de regresión sirve para predecir un valor específico de una variable dado cualquier valor específico de la otra.

Por otro lado cabe recordar que la correlación nos indica:

- la existencia de asociación entre variables
- La dirección de la asociación.
- Grado de asociación

ACTIVIDAD N° 1

Las materias primas que se utilizan en una fábrica en la producción de una fibra sintética son almacenadas en un lugar donde no se tiene control sobre la humedad. Se desea saber si hay una relación entre la humedad del galpón de almacenamiento y la humedad de la materia prima. Grafique los datos en un diagrama de dispersión y estime los parámetros que caracterizan a la recta de regresión. Determine la recta y grafique. Estime cual será la humedad de la muestra si la humedad del lugar es de 40

Humedad del lugar	42	35	50	43	48	62	31	36	44	39	55	48
Humedad de la muestra	12	8	14	9	11	16	7	9	12	10	13	11

ACTIVIDAD N° 2

Los siguientes datos corresponden a los porcentajes de mortalidad obtenidos a dosis crecientes de un insecticida. Se desea estudiar si existe una componente lineal entre la mortalidad y la dosis, expresada como el logaritmo de las concentraciones utilizadas.
El experimento consistió en someter a grupos de 1000 insectos a cada una de las dosis ensayadas. Los resultados fueron los siguientes:

Ln (dosis)	Mortalidad(%)
0	5
1	7
5	10
10	16
15	17
20	25
25	26
30	30

a) Construir un diagrama de dispersión Mortalidad vs. Ln (dosis).
b) De acuerdo al gráfico obtenido, ¿es razonable proponer un ajuste lineal?
c) Escribir el modelo lineal que, se supone, relaciona la mortalidad con la dosis.
d) Calcular los parámetros del modelo de regresión.

ACTIVIDAD N° 3

Determinar si existe alguna correlación, En caso de existir correlación encuentre al menos una recta de regresión.
Consideremos los siguientes datos, donde *x* indica la temperatura media diaria en grados Fahrenheit e *y* el consumo diario correspondiente de gas natural en pies cúbicos.

X	50	45	40	38	32	40	55
Y	2,5	5,0	6,2	7,4	8,3	4,7	1,8