

CAPITULO 2

REGRESION Y CORRELACION

2.1. REGRESION RECTILINEA EN UNA VARIABLE

2.1.1. El modelo de regresión rectilínea.

Las relaciones entre variables surgen como tema de interés en diversos campos de la investigación. a) ¿Cómo varían los precios con los años? b) ¿Cómo varía el peso de los animales con la edad? c) ¿Cómo se incrementa el rendimiento del trigo con el nivel de nitrógeno en el suelo? d) ¿Cómo evoluciona el peso de los niños menores a un año? e) ¿Cuál es la relación entre la edad de las personas y su presión sanguínea?

Las relaciones entre variables pueden ser graficadas y, en base a ello, se clasifican en rectilíneas y curvilíneas. Relaciones rectilíneas son las que se presentan cuando los incrementos de una variable (Y) son proporcionales a los de otra (X), y se expresan por fórmulas del tipo $Y = \alpha + \beta X$, donde α y β son parámetros que indican el valor de Y cuando $X = 0$, y el incremento de Y con la variación de una unidad de X, respectivamente.

Relaciones curvilíneas son las que se presentan cuando los incrementos en una variable dependiente frente a una variable independiente no son constantes sino que varían según el nivel de esta última. De acuerdo a nuestros objetivos, las relaciones curvilíneas pueden clasificarse en polinomiales y de otro tipo. Las relaciones polinomiales son aquellas que se expresan por polinomios de la forma: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k$ de los cuales los más importantes son los de segundo y tercer grado (sin considerar el de primer grado que proporciona una recta):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

Las relaciones del tipo no polinómico que más nos interesan son las asintóticas y exponenciales, como: $Y = \alpha \cdot \beta^X$; $Y = \alpha \cdot X^\beta$;

o la de Mitscherlich: $Y = A[1 - 10^{-c(X+b)}]$.

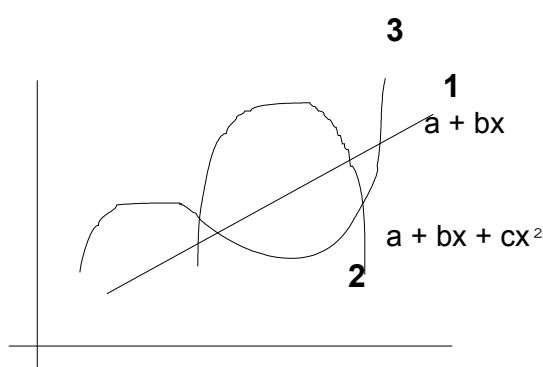


Figura 2.1. Funciones polinomiales

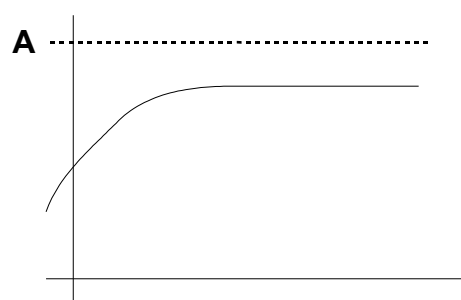


Figura 2.2. Función de Mitscherlich

Si la relación entre variables es perfecta, se puede considerar a una de ellas como variable independiente y, consecuentemente, a la otra como dependiente de la variación de aquella. Si llamamos a la primera X y a la segunda Y, se pueden obtener los valores de Y a partir de los de X, por lo que se dice que Y es función de X: $Y = f(X)$. Este tipo de relación proporciona los llamados modelos determinísticos de explicación de la realidad. Para

este tipo de modelo se pueden calcular los parámetros, en caso de no conocerlos, si se posee un número determinado de puntos.

Ejemplo 2.1. Conociendo que la circunferencia y el radio de los círculos son proporcionales, se puede establecer que: $C = k \cdot R$, donde k es la constante de proporcionalidad desconocida. Si midiendo un círculo de radio 2 tiene 12,566 de circunferencia, es posible determinar experimentalmente la constante como $k = 12,566/2 = 6,283$.

Ejemplo 2.2. La equivalencia entre los grados para medir temperaturas de las escalas Fahrenheit y Centígrada no es generalmente recordada. No obstante, muchas veces se recuerda que el agua hierve a 212 °F y a 100 °C, de modo que postulando la relación $^{\circ}\text{C} = k \cdot ^{\circ}\text{F}$, se puede intentar despejar la constante. Una vez intentado, aparece claro que no es posible ya que los ceros de ambas escalas no coinciden. Esta es una situación particular de la regla general de que el ajuste de una recta exige dos puntos, por los menos. Si, exigiendo a la memoria, se recuerda que el agua congela a 0 °C y 32 °F se puede, ahora sí, despejar la constante del sistema de ecuaciones:

$$\begin{array}{rcl} 0^{\circ}\text{C} & \text{-----} & 32^{\circ}\text{F} \\ 100^{\circ}\text{C} & \text{-----} & 212^{\circ}\text{F} \end{array} \quad \text{obteniendo } ^{\circ}\text{C} = (5/9)^{\circ}\text{F}.$$

Existe otro tipo de situación donde la relación entre variables no es perfecta sino promedial. Estas relaciones, que proporcionan los modelos probabilísticos de explicación de la realidad, se presentan con mucha frecuencia en las ciencias biológicas y en economía, donde el gran número de factores que afectan a las variables de interés impide la visualización de relaciones determinísticas entre ellas.

Ejemplo 2.3. Es natural pensar que el rendimiento de un cultivo es proporcional al nivel de nitrógeno en el suelo. No obstante ello, no siempre que se aumenta el nivel de nitrógeno en el suelo se incrementa el rendimiento de los cultivos, ya que otros factores pueden estar disminuyéndolo simultáneamente. La situación es que, en promedio, el rendimiento de los cultivos se incrementa con el nivel de nitrógeno de los suelos. Para cada nivel de nitrógeno los rendimientos de un cultivo forman una población. La distribución se dice condicional para ése nivel de nitrógeno. Las distribuciones condicionales difieren en la media, la que es mayor para niveles de nitrógeno más altos, como se muestra en la figura 2.3.

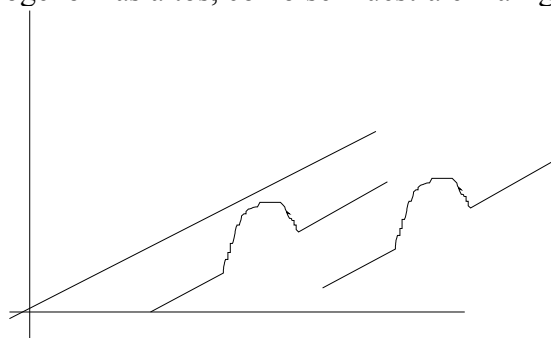
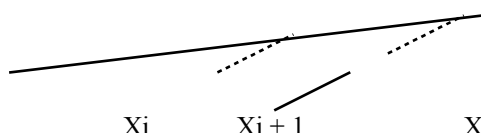


Figura 2.3. Modelo de regresión rectilínea.



2.1.2. Parámetros y estimadores

El modelo de regresión rectilínea en una variable es de la forma $Y_i = \alpha + \beta X_i + \varepsilon_i$, donde Y_i representa los valores observados de la variable de interés, X_i representa valores observados de la variable independiente, los parámetros β y α indican el incremento de rendimiento por unidad de variación de X , y el rendimiento cuando X vale cero, respectivamente; y finalmente ε_i indica un error no explicado y que se comporta como una variable aleatoria. Generalmente, se completa el modelo con el supuesto de que los errores provienen de una población con distribución normal, media cero, varianza finita, e independientes entre sí: $\varepsilon \sim \text{NID}(0, \sigma^2)$.

Los parámetros de los modelos de regresión son generalmente desconocidos y estamos interesados en conocerlos. El método adecuado de solucionar el problema implica estimar los parámetros del modelo mediante una muestra aleatoria. Se buscará obtener estimaciones de los parámetros y, con ellas, construir ecuaciones de predicción de la forma $\hat{Y}_i = a + bX_i$, con **a** como estimador de α y **b** como estimación de β . El valor \hat{Y}_i simboliza el valor predicho para Y cuando $X = X_i$, por la recta determinada por nosotros. Por lo tanto, es un estimador de $E(Y|X_i)$ el valor proporcionado por la verdadera relación. A la diferencia entre el valor observado Y_i y el predicho \hat{Y}_i , se le conoce como desvío o error de predicción $e_i = Y_i - \hat{Y}_i = Y_i - a - bX_i$.

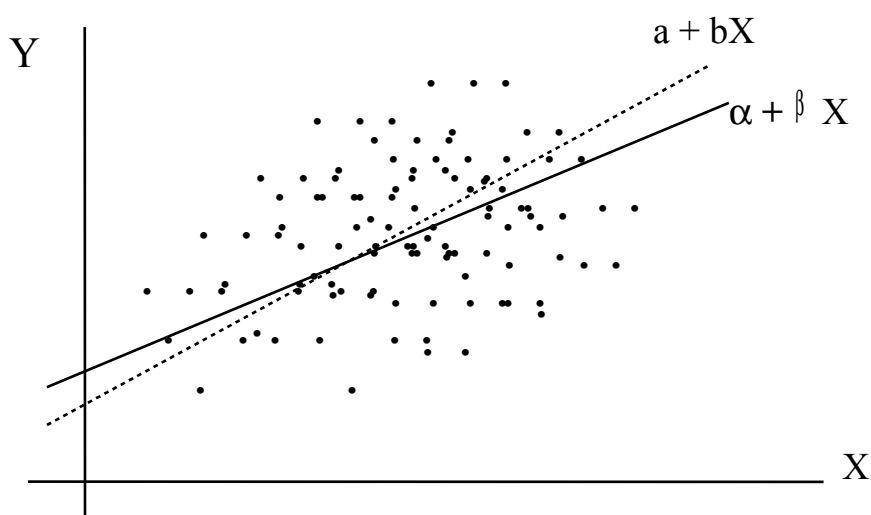


Figura 2.4. Línea de regresión y su estimación. La línea $\alpha + \beta X$ es imaginaria, la línea $a + bX$ es la calculada por nosotros.

2.1.3. Estimación de los parámetros por mínimos cuadrados.

El problema de la estimación de los parámetros del modelo lineal se puede resolver adecuadamente por los métodos de máxima verosimilitud y de mínimos cuadrados. El método de máxima verosimilitud exige conocer la distribución de los errores del modelo: en tanto, el de los mínimos cuadrados no impone esa exigencia sino solamente que los errores sean incorrelacionados, con media cero y varianza común. Para el caso de distribución normal de los errores ambos métodos coinciden en los mismos estimadores, de modo que nos referiremos de preferencia al de los mínimos cuadrados que es intuitivamente más atrayente. Por una discusión acerca de estos enfoques puede consultarse Littel et al. (2005).

El método de estimación de mínimos cuadrados se basa en la técnica de aproximación similar, que postula como la recta de mejor ajuste a una serie de puntos, a aquella que minimice la suma de cuadrados de los desvíos o erroresⁱ:

$$SCE = \sum e^2_i = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2$$

Esta suma de cuadrados es función de los valores de **a** y de **b**, de modo que según sean estos aquella tomará diferentes valores, existiendo un valor de cada uno de ellos que haga mínima la SCE. Entonces, la mejor recta será la que tenga como valores los que hagan simultáneamente mínima esa suma para **a** y **b**. Esos valores son obtenibles por derivación parcial:

$$\partial SCE / \partial a = -2 \sum (Y - a - bX) = 0$$

$$\partial SCE / \partial b = -2 \sum (Y - a - bX) X = 0$$

Los valores de **a** y **b** que cumplan con el sistema de ecuaciones anterior, cumplen con proporcionar la recta que tenga menor SCE, por lo cual, la importancia de este sistema es primordial en regresión; el sistema, conocido como de ecuaciones normales, se escribe generalmente del modo siguiente:

$$\sum Y = na + b \sum X$$

$$\sum XY = a \sum X + b \sum X^2$$

y de él se pueden despejar los estimadores de α y de β : $a = \bar{Y} - b\bar{X}$

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum xy}{\sum x^2} = \frac{\text{Covarianza}(x, y)}{\text{Varianza}(x)}$$

Uso de variables reducidas. Es muy útil tomar $x = X - \bar{X}$ $y = Y - \bar{Y}$. Por lo tanto la fórmula queda $b = \sum xy / \sum x^2$. Al reemplazar **a** por su valor, la ecuación de predicción $\hat{Y} = a + bX$ queda: $\hat{Y} = \bar{Y} - b\bar{X} + bX = \bar{Y} + b(X - \bar{X}) = \bar{Y} + bx$

restando \bar{Y} de ambos lados y llamando $\hat{y} = Y - \bar{Y} = b(X - \bar{X})$; tenemos: $\hat{y} = bx$

Esto corresponde con un modelo $Y = \mu_Y + \beta(X - \bar{X}) + \varepsilon$ donde se toma como parámetro la media de Y (y no α). Este tipo de modelo se usa en algunas situaciones como en genética (ver ejemplo 2.14, pg 69).

Máxima verosimilitud. Si el modelo es Gauss Markov Normal el método de Máxima Verosimilitud proporciona los mismos estimadores que Mínimos Cuadrados.

ⁱ Omitiremos el uso del subíndice por simplicidad.

2.1.4. Fórmulas de cálculo.

Ejemplo 2.4. Peso vs. Edad en niños. Estos datos son artificiales basados en las curvas de crecimiento que publica el Ministerio de Salud Pública de Uruguay.

X	Y	x	Y	x2	y2	xy	Y	e	e2
1	4	-3	-2,4	9	5,92	7,3	4,34	-0,34	0,12
2	5,9	-2	-0,5	4	0,28	1,07	5,04	0,86	0,74
3	4,7	-1	-1,7	1	3	1,73	5,74	-1,04	1,07
4	7,1	0	0,67	0	0,44	0	6,43	0,67	0,44
6	7,7	2	1,27	4	1,6	2,53	7,83	-0,13	0,02
8	9,2	4	2,77	16	7,65	11,1	9,22	-0,02	0,00
24	38,6	0	0	34	18,9	23,7	38,6	0	2,393
4	6,43	0	0	5,67	3,15	3,95	6,433	0	0,3988

Para calcular los estimadores $b = 3,95/5,67=0,70$ y $a = \bar{Y} - b\bar{X} = 6,43-(0,70)(4,0) = 3,65$. Por lo tanto, la función para predecir el peso a partir de la edad sería: $\hat{Y}_i = 3,65 + 0,70 X$

Cálculo abreviado. Conviene en este punto, introducir el concepto de que las cantidades de interés para cálculos de regresión son seis: Σx^2 , Σy^2 , Σxy , n , \bar{X} , \bar{Y} , y salen de igual número de valores más en bruto: ΣX^2 , ΣX , ΣY^2 , ΣY , ΣXY , n . Recordemos que las primeras son sumas de cuadrados y productos corregidas (por media diferente de cero) y que se pueden calcular a partir de las correspondientes cantidades sin corregir, por medio de las llamadas fórmulas abreviadas de cálculo. En el presente ejemplo tenemos:

$n = 6$ $\Sigma X = 24$ $\Sigma Y = 38,6$ $\Sigma Y^2 = 130$ $\Sigma Y^2 = 267,24$ $\Sigma XY = 178,1$	→	$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 130 - 24^2/6=34$ $\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 267,24 - 38,6^2/6=18,9$ $\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 178,1 - (24)(38,6)/6=23,7$	→	$\Sigma x^2 = 34$ $\Sigma y^2 = 18,9$ $\Sigma xy = 23,7$ $\bar{X} = 6$ $\bar{Y} = 4$ $n = 6$
---	---	---	---	---

Si bien las fórmulas de cálculo abreviado fueron desarrolladas principalmente para el trabajo con calculadoras, tienen un cierto valor interpretativo y conviene tenerlas presente aún cuando se usen otros medios de computación para procesar los datos. Las cantidades como ΣX^2 son llamadas sumas de cuadrados sin corregir, mientras que las del tipo Σx^2 son sumas de cuadrados corregidas (por media diferente de cero, como se dice a veces). La cantidad $C = (\Sigma X)^2/n$ se conoce como factor de corrección.

2.1.5. Inferencia sobre los coeficientes.

DISTRIBUCIÓN DE LOS ESTIMADORES. Los valores de los estimadores están contruidos a partir de valores muestrales constituyendo, pues, variables que dependen de la muestra aleatoria elegida. Si se cumple el supuesto de normalidad de los errores, la distribución de los estimadores es normal (ver página 13). Encontraremos la esperanza y varianza de **b**:

$$E[b] = E\left[\frac{\sum xy}{\sum x^2}\right] = \frac{1}{\sum x^2} \sum xE(Y) = \frac{1}{\sum x^2} \sum x(\alpha + \beta X) = \frac{\sum x\alpha}{\sum x^2} + \beta \frac{\sum xX}{\sum x^2} = \beta$$

$$V(b) = V\left[\frac{\sum xy}{\sum x^2}\right] = \frac{1}{(\sum x^2)^2} V(\sum xy) = \frac{V(Y)\sum x^2}{(\sum x^2)^2} = \frac{\sigma^2}{\sum x^2}$$

De modo que **b** es un coeficiente que se distribuye normalmente en el muestreo (ver página 13), insesgadamente y con varianza $\sigma^2/\sum x^2$. Si se estandariza la distribución, tenemos:

$$z = \frac{b - \beta}{\frac{\sigma}{\sqrt{\sum x^2}}}$$

y usando $\hat{\sigma} = \sqrt{(\sum e^2)/(n-2)}$ como estimador de σ_e , obtenemos una distribución t:

$$t = \frac{b - \beta}{\frac{\hat{\sigma}}{\sqrt{\sum x^2}}}$$

VARIANZAS DE LOS ESTIMADORES. Del mismo modo se puede demostrar que el estimador **a** tiene distribución normal, con media en α y varianza: $\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right)$. Otras

varianzas se presentan en la tabla 2.1 al final de esta sección. En el ejemplo 2.4 las varianzas son: $V[b] = 0,598/34 = 0,0176$ y $V[a] = 0,598 (1/6 + 4^2/34) = 0,381$ y los errores estándar (las raíces cuadradas de los anteriores) son: $\hat{\sigma}_b = 0,133$ y $\hat{\sigma}_a = 0,617$

INTERVALOS DE CONFIANZA PARA PARÁMETROS. El conocimiento de la distribución permite elaborar intervalos de confianza para los parámetros. En efecto, si $P[-t_{(n-2;0,975)} < t_{(n-2)} < t_{(n-2;0,975)}] = 0,95$ entonces, tomando como ejemplo el parámetro β ,

$$P\left[-t_{(n-2;0,975)} < \frac{b - \beta}{\hat{\sigma} / \sqrt{\sum x^2}} < t_{(n-2;0,975)}\right] = 0,95$$

y despejando β en la doble desigualdad entre paréntesis

$$P\left[b - t_{(n-2;0,975)} \cdot \hat{\sigma} / \sqrt{\sum x^2} < \beta < b + t_{(n-2;0,975)} \cdot \hat{\sigma} / \sqrt{\sum x^2}\right] = 0,95$$

obtenemos un intervalo de confianza para β al 95%. En general, todos los intervalos de confianza serán de la forma $P[\hat{\theta} - t\hat{\sigma}_{\hat{\theta}} < \theta < \hat{\theta} + t\hat{\sigma}_{\hat{\theta}}] = \gamma$. En el ejemplo 2.4, un intervalo de confianza para β sería: $P[b - t(4) \hat{\sigma}_b < \beta < b + t(4) \hat{\sigma}_b] = \gamma$ o sea que con números sería:

$$P[0,697 - (2,776)(0,133) < \beta < 0,697 + (2,776)(0,133)] = P[0,328 < \beta < 1,066] = 0,95$$

Mientras que para α sería: $P[a-t(4)\hat{\sigma}_a < \alpha < a+t(4)\hat{\sigma}_a] =$

$$P[3,645-(2,776)(0,617) < \alpha < 3,645+(2,776)(0,617)] = P[1,932 < \alpha < 5,358] = 0,95$$

INTERVALOS DE CONFIANZA PARA LA MEDIA CONDICIONAL. La media de valores de Y para un valor determinado de X, llamémosle X_0 es el resultado de reemplazar $X = X_0$ en la ecuación de regresión: $3,64+0,70(7) = 8,54$. La varianza está en el cuadro abajo

como: $\sigma_e^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right]$. Notemos que la precisión depende del cuadrado de la distancia

entre X_0 y la media. Esto da lugar a cinturones de confianza curvilíneos.

INTERVALOS DE PREDICCIÓN. No olvidemos que uno de los principales objetivos de la regresión es predecir Y en función de X. La predicción puntual consiste en substituir el valor de X en la función de regresión, pero si deseamos una predicción por intervalos tenemos que hacer una pequeña corrección a la fórmula de la varianza para la media condicional:

$$\sigma_{\hat{Y}}^2 = \sigma_e^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2} \right]$$

En este caso queda:

$$P[8,54-(2,776)(0,937) < Y^* < 8,54+(2,776)(0,937)] = P[5,97 < Y^* < 11,11] = 0,95$$

Tabla 2.1. Varianzas de los estimadores.

Coficiente		Varianza
B	Pendiente de la línea de regresión	$\sigma_e^2 \left[\frac{1}{\sum x^2} \right]$
A	Ordenada en el origen.	$\sigma_e^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2} \right]$
μ	Media de la variable de interés	$\sigma_e^2 \left[\frac{1}{n} \right]$
$a+bX_0$	Media condicional para $X=X_0$	$\sigma_e^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right]$
Predicción	condicional para $X=X_0$	$\sigma_e^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right]$

PRUEBA DE HIPÓTESIS SOBRE LOS COEFICIENTES. Si queremos probar $H_0: \beta = 0$ vs $H_A: \beta \neq 0$ usamos como variable $t = \frac{b - \beta}{\hat{\sigma}_b} = 0,70 / 0,13 = 5,25^{**}$. Comparando con el valor crítico $t_{(0,975;5)} = 2,776$ vemos que el valor calculado cae en la región crítica, eso se señala con los dos asteriscos. Es decir que se rechaza H_0 , por lo tanto se concluye que β es diferente de 0. Notemos que el intervalo de confianza en la sección anterior no incluye el valor cero.

Si se quiere probar $H_0: \alpha = 0$

$$H_A: \alpha \neq 0$$

la variable es $t = \frac{a - \alpha}{\hat{\sigma}_a} = \frac{3,65}{0,62} = 5,90^{**}$

lo que es significativo, es decir se rechaza H_0 y se decide que $\alpha \neq 0$. También acá vemos que el intervalo de confianza no incluye al valor cero.

Muchas veces en el proceso de una prueba de hipótesis se considera mejor reportar el "p-value" como se hace en la página 72, y que el lector decida el nivel de significación que utilizará. Este aspecto ya fue discutido en la sección 1.2.2 (página 16).

2.1.6. Análisis de varianza para la regresión rectilínea.

La técnica de análisis de varianza se desarrollará, como es habitual, descomponiendo la variación total y logrando estimadores independientes de la varianza poblacional, comparables por la prueba F.

PARTICION DE LA SUMA DE CUADRADOS. Cada observación es igual a la media, más un aporte de la regresión con X, más un error:

$$Y_i = \bar{Y} + bx_i + e$$

elevando al cuadrado y tomando la suma de los cuadrados de todas las observaciones:

$$\sum Y_i^2 = n\bar{Y}^2 + b^2 \sum x_i^2 + \sum e_i^2 + 2 \text{ productos que se anulan.}$$

Esta desigualdad se estudia generalmente restando $n\bar{Y}^2$ de ambos lados y recordando que $\sum Y_i^2 - n\bar{Y}^2 = \sum (Y_i - \bar{Y})^2 = \sum y_i^2$ por lo que tenemos:

$$\sum y_i^2 = b^2 \sum x_i^2 + \sum e_i^2$$

tomando los siguientes nombres $SC = \sum y_i^2$; $SCE = \sum e_i^2$ y $SCR = b^2 \sum x_i^2$,

podemos escribir $SC = SCR + SCE$. Esta última expresión nos dice que la suma de cuadrados de los valores de la variable de interés se pueden descomponer en una parte debida a la regresión (SCR), y otra, no explicada por el modelo, debido al error (SCE).

HOMOGENEIDAD DE LOS ESTIMADORES. La suma de cuadrados de la regresión tiene un grado de libertad. En efecto,

$$\begin{aligned} \varepsilon &= Y - \mu - \beta x && \text{sumando y restando } \hat{Y} = \bar{Y} + b(X - \bar{X}) \text{ tenemos:} \\ &= Y - \mu - \beta x - \hat{Y} + \bar{Y} + b x && \text{y reorganizando esa expresión:} \\ &= (Y - \hat{Y}) + (\bar{Y} - \mu) + (b - \beta) x \end{aligned}$$

Si ahora elevamos al cuadrado de ambos lados y tomamos sumatorias, tenemos:

$$\sum \varepsilon^2 = \sum (Y - \hat{Y})^2 + n(\bar{Y} - \mu)^2 + (b - \beta)^2 \sum x^2 + \text{productos que se anulan.}$$

Dividiendo por σ^2 , tenemos una variable χ^2 :

$$\begin{aligned} \frac{\sum \varepsilon^2}{\sigma^2} &= \frac{\sum e^2}{\sigma^2} + \frac{n(\bar{Y} - \mu)^2}{\sigma^2} + \frac{\sum x^2 (b - \beta)^2}{\sigma^2} \\ \frac{\sum \varepsilon^2}{\sigma^2} &= \frac{\sum e^2}{\sigma^2} + \frac{(\bar{Y} - \mu)^2}{\sigma^2/n} + \frac{(b - \beta)^2}{\sigma^2/\sum x^2} \end{aligned}$$

La expresión de la izquierda es una χ^2 con n grados de libertad, las dos últimas de la derecha también son χ^2 con 1 grado de libertad cada una, y como, $\chi_n^2 = \chi_{(n-2)}^2 + \chi_{(1)}^2 + \chi_{(1)}^2$ la suma de cuadrados del error tiene distribución χ^2 con (n-2) grados de libertad. Por lo tanto, el cociente entre $SCR/1$ y $SCE/(n-2)$ tiene distribución F bajo el supuesto de que $\beta=0$:

$$\frac{SCR/1}{SCE/(n-2)} = \frac{CMR}{CME} \sim F_{n-2}^1$$

Esta distribución posibilita probar la hipótesis de que $\beta=0$. Todo esto generalmente se presenta en una tabla de análisis de la varianza.

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Regr.rectilínea	$SCR = b \sum xy$	1	SCR/GLR	CMR/CME
Error o desvíos	$SCE = \sum e^2$	n - 2	SCE/GLE	
TOTAL	$SC = \sum y^2$			

En el ejemplo 2.4 tenemos:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Regr.rectilínea	16,52	1	16,52	27,61
Error o desvíos	2,39	4	0,60	
TOTAL	18,91	5		

Probar la hipótesis $\beta=0$ por la prueba t y la F es totalmente equivalente al ser: $t^2_{(n-2)} = F_{(1,n-2)}$

$$t^2 = \left(\frac{b - \beta}{\sigma / \sqrt{\sum e^2}} \right)^2 = \frac{(b - \beta)^2 \sum x^2}{\sigma_e^2} = \frac{CMR}{CME} = F_{(1,n-2)}$$

en este caso $5,25^2 = 27,61$.

2.2. REGRESIÓN MÚLTIPLE.

2.2.1. Concepto de regresión múltiple.

Las variables de mayor interés en biología y otras ciencias dependen generalmente de varios factores. El comportamiento de una variables de interés (por ejemplo el rendimiento de un cultivo) frente a variaciones de las variables independientes es el tema de la regresión múltiple.

Ejemplo 2.5. El rendimiento de una pastura es una variable que está influida por una gran cantidad de factores: temperatura, humedad, luz, disponibilidad de los diferentes nutrientes como nitrógeno, fósforo, potasio, etc. intensidad y época de pastoreo, y muchos otros estarán influyendo para que el rendimiento obtenido sea ése y no otro. Si se conoce como reacciona la pradera al agregado de N, P, o K se puede determinar rendimientos para distintos agregados de cada uno y en base a ello decidir cual o cuales conviene utilizar. El conocimiento de la variación del rendimiento frente a cambios de valor de otra variable como tenor de N, P, o K del suelo se expresa generalmente como una función de repuesta. Si $Y = f(N)$ la función de respuesta sería una línea que se puede graficar en el plano, si $Y = f(N,P)$ la función de respuesta sería un plano que debe representarse en el espacio; finalmente para función de más de dos variables como $Y = f(N, P, K)$ la representación gráfica es imposible en el espacio de tres dimensiones y se habla de hiperplanos.

Ejemplo 2.6. Una función de respuesta del tipo mencionado puede ser: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ donde β_1 es el incremento de Y por unidad de X_1 , β_2 es el incremento de Y por unidad de X_2 , y β_0 es el valor de Y cuando X_1 y X_2 son nulas. La gráfica de esa función sería como la mostrada en la figura 2.5.

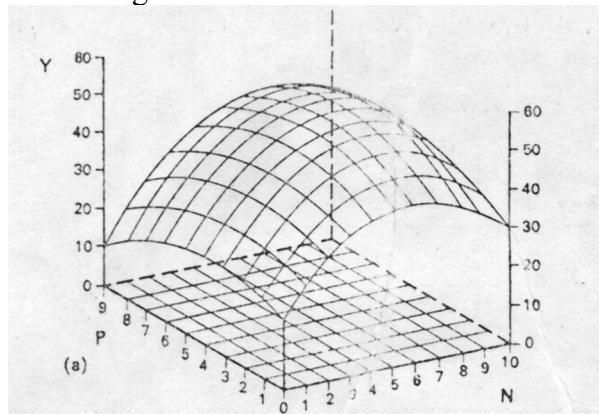


Fig. 2.5. Representación gráfica de una función de dos variables.

Ejemplo 2.7. Una función de respuesta es: $Y = 1.000 + 200 N + 250 P$. La ecuación expresa que por cada kilo de nitrógeno agregado se incrementa el rendimiento de la pastura en 200 kg y por cada kilo de fósforo en 250 kg, siendo el rendimiento de 1.000 kg cuando no se fertiliza.

En este tipo de situaciones no se puede postular la existencia de una función matemática entre el rendimiento y los niveles de fertilidad ya que todos los demás factores que condicionan el rendimiento distorsionan la relación. Por ello es que se postula en cambio la existencia entre variables de una función de regresión, es decir que el promedio de los valores de rendimiento está en función de los niveles de fertilidad.

2.2.2. Modelo de regresión múltiple.

El modelo de regresión múltiple es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

donde Y_i es la observación i -ésima de la variable aleatoria Y

β_j son parámetros: los coeficientes de regresión parcial es decir el incremento de la variable rendimiento por cada unidad de incremento de la variable X_j .

X_j son variables matemáticas no aleatorias

ε representa un error aleatorio que generalmente se supone proviene de una distribución normal con media cero, varianza finita y que es independiente del valor de cada X_j : $\varepsilon \sim \text{NID}(0; \sigma^2)$.

Es usual tomar una variable X que siempre valga 1: $X_0=1$, con lo que se puede abreviar la

escritura del modelo: $Y_i = \sum_{j=0}^k \beta_j X_{ji} + \varepsilon_i$

y el modelo expresa que para un conjunto de valores de las X_j que se consideran los valores de Y que tienen una distribución normal con media en $E[Y|x] = \sum \beta_j X_j$ y con varianza σ^2 . La media así considerada se llama condicional pues es la media para ése conjunto de valores de las variables independientes y cambiará al cambiar estos. El conjunto de las medias condicionales formaría el plano de regresión.

El modelo se puede escribir matricialmente como

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

o sea: $\mathbf{Y} = \mathbf{X} \mathbf{\beta} + \mathbf{\varepsilon}$ #

La principal finalidad en el planteamiento de un modelo es determinar estimaciones de los parámetros de modo que se pueden encontrar fórmulas de predicción de los valores de la variable de interés a partir del conocimiento de los valores de las variables independientes. En general, nos interesa conocer los coeficientes de regresión parcial (los b), probar si esos coeficientes son cero (individualmente), probar la significación del modelo (probar si los coeficientes son todos cero), construir intervalos de confianza para esos coeficientes, y hacer predicciones puntuales y por intervalos.

2.2.3. Estimaciones por mínimos cuadrados y ecuaciones normales.

Llamemos $\hat{Y} = \sum b_j X_j$ a la ecuación de predicción que buscamos. En ella b son las estimaciones de los coeficientes. Cada valor predicho se separaría del valor observado correspondiente en una cantidad que llamaremos error de predicción y simbolizaremos con e :

$$e = Y - \hat{Y}$$

El criterio de los mínimos cuadrados propone como mejores estimadores de los parámetros a los valores de b que cumplan con que los desvíos e tienen una suma de cuadrados mínima. Como:

$$e = Y - \hat{Y} = Y - \sum b_j X_j$$

$$\sum e^2 = \sum (Y - \hat{Y})^2 = \sum (Y - \sum b_j X_j)^2$$

la suma de cuadrados del error es función de los valores de b ; derivando con respecto a b se encuentran los valores de estimadores que minimizan la suma de cuadrados del error. Esos valores son los que cumplen con las ecuaciones normales:

$$\begin{array}{l} nb_0 + b_1 \sum X_1 + b_2 \sum X_2 + \dots + b_k \sum X_k \\ b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + \dots + b_k \sum X_1 X_k \\ b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + \dots + b_k \sum X_2 X_k \\ \vdots \\ b_0 \sum X_k + b_1 \sum X_1 X_k + b_2 \sum X_2 X_k + \dots + b_k \sum X_k^2 \end{array}$$

Matricialmente la ecuación de predicción es: $Y = Xb$, los errores son: $e = Y - \hat{Y} = Y - Xb$

la suma de cuadrados del error es: $\sum e^2 = \sum (Y - \hat{Y})^2 = (Y - Xb)'(Y - Xb)$

y las ecuaciones normales son: $X'Xb = X'Y$ donde

$$X'X = \begin{bmatrix} n & \sum X_1 & \sum X_2 & \dots & \sum X_k \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 & \dots & \sum X_1 X_k \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 & \dots & \sum X_2 X_k \\ \dots & \dots & \dots & \dots & \dots \\ \sum X_k & \sum X_k X_1 & \sum X_k X_2 & \dots & \sum X_k^2 \end{bmatrix}; b = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix}; X'y = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \\ \dots \\ \sum X_k Y \end{bmatrix}$$

Se logra recordar más fácilmente las k ecuaciones normales si se observa que resultan de multiplicar la ecuación de predicción sucesivamente por 1, X_1 , X_2 , ..., X_k y tomar sumatorias. Los valores buscados de las estimaciones se obtienen despejando del sistema de ecuaciones normales.

Las ecuaciones normales tienen una estructura muy simple, la solución es $b = (X'X)^{-1}X'y$. Si el modelo no fuera lineal no pasaría eso, las ecuaciones no-lineales no tienen una solución explícita. El PROC NLIN de SAS ajusta regresiones no-lineales.

2.2.4. Recursos de cálculo. Ecuaciones reducidas.

El manejo de las ecuaciones normales se puede facilitar por el uso de variables reducidas:
 $x_1 = X_1 - \bar{X}_1$; $x_2 = X_2 - \bar{X}_2$; ... ; $x_k = X_k - \bar{X}_k$ e $y = Y - \bar{Y}$ con lo que el sistema se puede escribir:

$$b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \dots + b_k \sum x_1 x_k = \sum x_1 y$$

$$b_1 \sum x_1 x_2 + b_2 \sum x_2^2 + \dots + b_k \sum x_2 x_k = \sum x_2 y$$

$$\vdots \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$b_1 \sum x_1 x_k + b_2 \sum x_2 x_k + \dots + b_k \sum x_k^2 = \sum x_k y$$

obteniéndose una ecuación y una incógnita menos: b_0 , la que se puede obtener de:

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 - \dots - b_k \bar{X}_k$$

Ejemplo 2.8. A los efectos de estudiar la influencia de la fertilización NP en praderas naturales se tomaron observaciones de rendimiento bajo diferentes niveles de fertilización. Los resultados se muestran en la tabla.

Tabla 2.1. Rendimiento de praderas naturales sobre Basalto con fertilización .

N	X ₁	P	X ₂	Y
80	-1	80	-1	1125
80	-1	240	1	1135
240	1	80	-1	1345
240	1	240	1	2185
160	0	160	0	1435
0	-2	160	0	480
320	2	160	0	1895
160	0	0	-2	850
160	0	320	2	2250
0	-2	0	-2	505
0	-2	320	2	615
320	2	0	-2	860
320	2	320	2	3290

Para encontrar las estimaciones de los incrementos basta despejarlos del sistema de ecuaciones normales:

$$\sum Y = nb_0 + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1 Y = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2$$

$$\sum X_2 Y = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$$

En este caso se utiliza una técnica usual de facilitar el trabajo codificando los valores de las variables independientes. Como los intervalos entre los valores son fijos se resta la media y se divide entre esa cantidad fija. $X_1 = (N - 160)/80$ y $X_2 = (P - 160)/80$, de modo que las ecuaciones resultan:

$$17.970 = 13 b_0 + 0 b_1 + 0 b_2$$

$$10.160 = 0 b_0 + 28 b_1 + 0 b_2$$

$$8.730 = 0 b_0 + 0 b_1 + 28 b_2$$

y despejando se obtienen los valores de los coeficientes:

$$b_0 = 17.970/13$$

$$b_1 = 10.160/28$$

$$b_2 = 8.730/28$$

La función de respuesta estimada es $\hat{Y} = 1382,31 + 362,86 X_1 + 311,79 X_2$ expresada en unidades de X_1 y X_2 es decir en unidades de 80 kg. Se pueden decodificar los valores del siguiente modo:

$$\hat{Y} = 1382,31 + 362,86 \left[\frac{N - 160}{80} \right] + 311,79 \left[\frac{P - 160}{80} \right]$$

con lo que la función de respuesta es: $\hat{Y} = 33,01 + 4,536 N + 3,90 P$

En el ejemplo ya está implícito en la codificación el uso de las variables reducidas, ya que los valores de X_1, X_2, \dots, X_k tienen media cero con lo que $x_1 = X_1; x_2 = X_2; \dots, x_k = X_k$.

Las ecuaciones normales usando las variables reducidas serían:

$$b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum x_1 y$$

$$b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum x_2 y$$

reemplazando los valores se ve que la ortogonalidad de X_1 y X_2 que hace que $\sum x_1 x_2 = 0$, facilita los cálculos. $10.160 = 28 b_1 + 0 b_2$

$$8.730 = 0 b_1 + 28 b_2$$

y $b_0 = \bar{Y} = 17.790/13$ lo que proporciona idénticas estimaciones que antes.

2.2.5. Regresión polinomial.

Las funciones de regresión del tipo polinomial son: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \epsilon$. Definiendo a las sucesivas potencias de la variable X como diferentes variables se puede realizar el estudio de los polinomios de regresión con los métodos de regresión multivariante que estamos estudiando: $X_2 = X^2$; $X_3 = X^3$; ... ; $X_k = X^k$.

Los polinomios más importantes son los de primer grado (que da una línea recta y generalmente no se menciona como polinomio) y de segundo grado que proporciona gráficamente una parábola. La importancia de la parábola radica en estudios como los de fertilidad de suelos en que interesa el valor de la variable que produce un máximo rendimiento, para lo cual se asume como conocido que la respuesta a fertilización no continúa por siempre.

Ejemplo 2.9. Ajustar una parábola de la forma $Y = \beta_0 + \beta_1 N + \beta_2 N^2 + \epsilon$ a los datos del ejemplo 2.8.

Como antes codificando la variable $X_1 = (N-160)/80$ las ecuaciones normales son:

$$17.970 = 13 b_0 + 0 b_1 + 28 b_2$$

$$10.160 = 0 b_0 + 28 b_1 + 0 b_2$$

$$36.370 = 28 b_0 + 0 b_1 + 100 b_2$$

y los coeficientes se obtienen despejando:

$$b_0 = 1.508,988$$

$$b_1 = 362.857$$

$$b_2 = 58.817$$

con lo que $\hat{Y} = 1.508,988 + 362,857 X + 58,816 X^2$

Los métodos anteriores se pueden combinar para estudiar modelos como el siguiente:

$$\hat{Y} = \beta_0 + \beta_1 N + \beta_2 N^2 + \beta_3 P + \beta_4 P^2 + \beta_5 NP + \varepsilon$$

que es un modelo de una complejidad superior donde se incluyen términos de segundo grado en cada una de las variables y un sumando con coeficiente β_5 que mide la interacción. El concepto de interacción se discutirá más extensamente en otra parte de estas notas.

Ejemplo 2.10. Ajuste una función del tipo anterior a los datos de la tabla 2.1.
Las ecuaciones normales son:

$$\begin{bmatrix} 13 & 0 & 0 & 28 & 28 & 0 \\ 0 & 28 & 0 & 0 & 0 & 0 \\ 0 & 0 & 28 & 0 & 0 & 0 \\ 28 & 0 & 0 & 100 & 68 & 0 \\ 28 & 0 & 0 & 68 & 110 & 0 \\ 0 & 0 & 0 & 0 & 0 & 68 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 17970 \\ 10160 \\ 8730 \\ 36370 \\ 39270 \\ 10110 \end{bmatrix}$$

cuya inversa es:

$$\begin{bmatrix} 0,256352 & 0 & 0 & -0,04728 & -0,03602 & 0 \\ 0 & 0,035714 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0,035714 & 0 & 0 & 0 \\ -0,04728 & 0 & 0 & 0,025972 & -0,00402 & 0 \\ -0,03602 & 0 & 0 & -0,00402 & 0,020746 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,014705 \end{bmatrix} \text{ con solución: } \begin{bmatrix} 1472,325 \\ 362,857 \\ 311,786 \\ -62,910 \\ 21,116 \\ 148,676 \end{bmatrix}$$

2.2.6. Análisis de varianza.

El análisis de varianza consiste en separar la SC total (corregida) en dos partes: una debida a la regresión (SCR) y otra debida al error (SCE) para compararlas y de se modo probar la hipótesis de que los coeficientes del modelo son todos cero.

La suma de cuadrados del error constituye una medida de la dispersión de los datos no explicada por el modelo y según la ecuación es $SCE = \mathbf{e'e} = (\mathbf{Y-Xb})' (\mathbf{Y-Xb}) = \mathbf{Y'Y} - \mathbf{b'X'Y}$, y como: $SC = \mathbf{Y'Y}$ y $SC = SCR + SCE$, se deduce que $SCR = \mathbf{b'X'Y}$.

La expresión $SC = SCR + SCE$ se conoce como partición de la suma de cuadrados total en dos partes ortogonales o independientes: una atribuible al modelo de regresión y una al error. Este resultado (como lo demuestran por ej Mood y Graybill; 1967; pág 406) posibilita el estudio de la significación de la regresión por medio de una prueba F. El análisis de varianza es generalmente presentado en una tabla como la que sigue.

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Regresión	$\mathbf{b'X'Y} - (\Sigma Y)^2/n$	k - 1		
Error o residuo	por diferencia	n - k		
TOTAL	$\mathbf{Y'Y} - (\Sigma Y)^2/n$	n - 1		

En el ejemplo tenemos una $\Sigma Y^2 = 32:999.700$ de modo que

$$SC = \Sigma y^2 = 32:999.700 - (17.970)^2/13 = 8:159.630,77$$

$$SCR = b_1 \Sigma x_1 y + b_2 \Sigma x_2 y = (362,86)(10.160) + (311,79)(8.730) = 6:408.584,3$$

$$SCE = SC - SCR = 8:159.630,77 - 6:408.584,3 = 1:751.046,47$$

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Regresión	6:408.517,30	2	3:204.258,65	18,30
Error o residuo	1:751.046,47	10	175.104,65	
TOTAL	8:159.630.77	12		

2.2.7. Intervalos de confianza para los parámetros.

Además de la estimación puntual que hemos analizado se pueden construir intervalos de confianza para los parámetros del modelo de regresión utilizando el conocimiento de la distribución de los estimadores que da la teoría. Si se cumplen los supuestos del modelo el vector de estimadores tiene una distribución: $\mathbf{b} \sim N(\mathbf{\beta}; \sigma^2[\mathbf{X}'\mathbf{X}]^{-1})$. De modo que cada estimador se distribuye con media en el parámetro que estima con distribución y varianza conocidas: $b_j \sim N(\beta_j; \sigma^2 c_{jj})$. La varianza del estimador es pues la del error multiplicada por c_{jj} el elemento de la j -ésima columna y fila de la matriz $[\mathbf{X}'\mathbf{X}]^{-1}$. La distribución normal de los estimadores posibilita su estandarización: $\frac{b - \beta}{\sigma / c_{ii}} = z \sim N(0,1)$ y reemplazando el parámetro σ^2 , generalmente desconocido, por una estimación (la del residuo en el análisis de varianza) tenemos: $\frac{b - \beta}{\hat{\sigma} / c_{ii}} = t_{(n-k)}$ que es la distribución generalmente utilizada para construir los intervalos de confianza. Estos intervalos de confianza serán de la forma: $P[b_j - t \cdot \hat{\sigma} \sqrt{c_{jj}} < \beta_j < b_j + t \cdot \hat{\sigma} \sqrt{c_{jj}}] = \gamma$

En el ejemplo 2.8: $[\mathbf{X}'\mathbf{X}]^{-1} = \begin{bmatrix} 1/13 & 0 & 0 \\ 0 & 1/28 & 0 \\ 0 & 0 & 1/28 \end{bmatrix}$

de modo que: $b_0 \sim N(\beta_0; \sigma^2/13)$

$b_1 \sim N(\beta_1; \sigma^2/28)$

$b_2 \sim N(\beta_2; \sigma^2/28)$

por lo tanto un intervalo de confianza para β_1 , por ejemplo, será: $P[b_1 - t\hat{\sigma} < \beta_1 < b_1 + t\hat{\sigma}] = \gamma$ que para el 95% de confianza es: $P[362,86 - (2,23)(79,08) < \beta_1 < 362,86 + (2,23)(79,08)] = 0,95$

o? $P[186,51 < \beta_1 < 539,21] = 0,95$

2.2.8. Funciones lineales y predicciones.

Se puede demostrar que las funciones lineales de los parámetros del tipo: $l = \sum a_j b_j = \mathbf{a}'\mathbf{b}$ tienen la siguiente varianza y esperanza:

$$E[l] = \sum c_j \beta_j \quad V[l] = \sigma^2 \mathbf{a}'[\mathbf{X}'\mathbf{X}]^{-1} \mathbf{a}$$

Este conocimiento se puede utilizar a los efectos de confeccionar intervalos para las predicciones, ya que son casos particulares de los anteriores donde $\mathbf{a}=\mathbf{x}_0$: $E[Y] = \mathbf{x}_0'\mathbf{b}$

$$P\{x_0.b-t.\sqrt{\sigma^2[1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]} < Y < x_0.b+t.\sqrt{\sigma^2[1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]}\}=\gamma$$

En el ejemplo si se desea predecir el rendimiento que se obtendría con 80 kg de P_2O_5 y sin nitrógeno, tenemos:

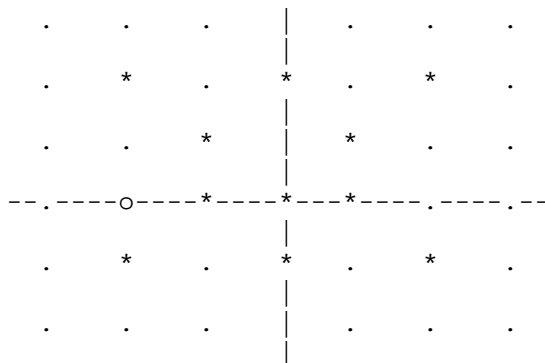


Fig 2.6. Representación de los valores de fertilización con rendimiento observado (*) y con rendimiento a predecir (o).

$$X = (0-160)/80 = -2 \text{ y } X = (80-160)/80 = -1$$

$$\hat{Y}_C = 1.382 + 362,86 (-2) + 311,79 (-1) = 344,8. \text{ La varianza de la predicción será}$$

$$V[\hat{Y}_C] = [\mathbf{x}_C[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_C]_{-}^2 = \begin{bmatrix} 1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1/13 & 0 & 0 \\ 0 & 1/28 & 0 \\ 0 & 0 & 1/28 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} = 93 \sigma^2/364$$

$$V[\hat{Y}_C] = \frac{93}{364} \sigma^2 = \frac{93}{364} \times 175.104,65 = 44.738,274$$

la raíz cuadrada de lo cual proporciona el error estándar: $\sqrt{44.738,274} = 211,51$

el intervalo de confianza para la media condicional será: $P[Y_C - t\sigma_{Y^{\wedge}} < \mu_{Y|X} < Y_C + t\sigma_{Y^{\wedge}}] = P[344,8 - (2,23)(211,51) < \mu_{Y|X} < 344,8 + (2,23)(211,51)] = P[-126,88 < \mu_{Y|X} < 816,47] = 0,95$

Mientras que la predicción tiene varianza: $V[Y_C] = (1 + \mathbf{x}'_C[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_C) \sigma^2 = (1 + \frac{93}{364}) \sigma^2 =$

$$\frac{457}{364} \sigma^2 = \frac{457}{364} \times 175.104,65 = 219.842,92 \text{ y entonces: } \sqrt{219.842,92} = 468,87 \text{ con lo que el}$$

intervalo de predicción queda: $P[Y_C - tY < Y < Y_C + tY] =$

$$P[344,8 - 2,23 \times 468,87 < Y < 344,8 + 2,23 \times 468,87] = P[-700,79 < Y < 1390,39] = 0,95$$

Notemos: 1. La precisión del intervalo de confianza es diferente para los distintos puntos del plano, es decir para las diferentes fertilizaciones.

2. Los puntos a predecir deben ubicarse en la región estudiada, evitándose la extrapolación.

3. Valores de rendimiento negativos aparecen en los intervalos de confianza, lo que no es lógico y muestra la no significación de los coeficientes (es decir que un valor probable es cero)

2.3. CORRELACION.

2.3.1. El coeficiente de correlación y la regresión.

La relación entre el coeficiente de correlación r y el coeficiente de regresión b es muy estrecha.

$$b = \frac{\sum xy}{\sum x^2} \quad y \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad (\text{ver sección 1.1.4, página 9}).$$

Los numeradores de ambas

estadísticas son iguales y los denominadores son siempre positivos. Por lo tanto, si $r=0$ también $b=0$, no existiendo relación rectilínea entre las variables X e Y . Si r es positivo también b es positivo, y la relación entre las variables es de proporcionalidad directa. Finalmente, si r es negativo, b también lo es, de modo que la relación entre las variables es inversa.

Podemos escribir:
$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \cdot \frac{\sqrt{\sum x^2}}{\sqrt{\sum x^2}} = b \cdot \frac{S_x}{S_y}$$

Con los elementos que hemos visto debe sorprendernos que probar la existencia de una regresión significativa por la prueba t implique probar la existencia de una correlación

significativa. En efecto, si $\beta=0$
$$t = \frac{b - \beta}{\hat{\sigma}_b} = \frac{\sum xy / \sum x^2}{\sqrt{\sum y^2}} = \frac{r - \rho}{\hat{\sigma}_r}$$

(*) $\sum y^2 (1-r^2+r^2) = \sum y^2 (1-r^2) + \sum y^2 r^2$
pero como $\sum y^2 r^2 = SCR$ y $SCE = \sum y^2 (1-r^2)$ tenemos $\sum y^2 = SCR + SCE$

si se cumple que $\rho = 0$ y $\sigma_r = \sqrt{\frac{1-r^2}{n-2}}$

LA CORRELACIÓN ES ANTERIOR HISTÓRICAMENTE, PERO TIENE LIMITANTES. El coeficiente de correlación es muy importante en Biometría. Cuando Karl Pearson lo propuso, nació la Biometría. No obstante eso, modernamente se tiende a darle más importancia al enfoque de regresión que a la correlación. Una de las razones puede ser vista en el siguiente ejemplo.

Ejemplo 2.11. Los datos que se muestran en la tabla corresponden a dos variables, llamémosle X y Y .

	X	Y	$x=X-\bar{X}$	$y=Y-\bar{Y}$	Xy
	0	4	-2	2	-4
	1	1	-1	-1	+1
	2	0	0	-2	0
	3	1	1	-1	-1
	4	4	2	2	+4
TOTALES	10	10	0	0,0	0,0

Las respectivas medias son: $\bar{X} = 2$ y $\bar{Y} = 2$. En la tercer y cuarta columnas se presentan los desvíos con respecto a las medias de los valores de X y de Y , se puede verificar que suman cero. Finalmente, en la quinta columna se presentan los productos. La covarianza es el promedio de esos productos de desvíos con respecto a la media: $S_{xy} = \sum xy / n = 0$. El

coeficiente de correlación de Pearson es la covarianza dividida por el producto de las desviaciones estándares: $r=0$ en este ejemplo. Y sin embargo $Y=X^2$. Cuando la relación entre las variables es perfecta decimos que tenemos una situación determinística o matemática en oposición a la situación probabilística o estadística. Dos variables tienen correlación cero cuando no hay relación rectilínea entre ellas, pudiendo haber relaciones curvilíneas. Por lo tanto, que dos variables sean incorrelacionadas no quiere decir que sean independientes.

LA REGRESIÓN ES MEJOR. Entonces actualmente se considera que el estudio de la regresión proporciona mejores herramientas que el estudio de la correlación. El estudio de la regresión es el estudio de la forma de la relación entre variables. No obstante, interesa también medir el grado de la relación así determinada y, para ello, contamos con distintos mecanismos. Uno de ellos es el propio estudio de la regresión: si esta es significativa lo que se determina como significativo es, en realidad, el grado de la relación. Otro coeficiente muy relacionado a

lo anterior es el error estándar de estimación:
$$\hat{\sigma}_e = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}}$$

cuanto mayor sea este, menor es el ajuste de la regresión determinada a los puntos muestrales, por lo que tendremos menor confianza en los valores predichos por la línea y mayor radio para el intervalo de confianza de la predicción.

Un tercer mecanismo para cuantificar la relación entre variables, es el uso de los coeficientes de determinación y de correlación. El coeficiente de determinación (r^2) es la proporción de la variación de la variable de interés, que resulta explicada por la relación con la variable independiente, medidas a través de la suma de cuadrados; es decir, el cociente entre la suma de cuadrados de la regresión y la suma de cuadrados total de la variable de interés: $r^2 = SCR/SC$. Generalmente, el coeficiente de determinación se expresa, indistintamente, en porcentaje o proporción, y puede variar entre 0, si nada de la variación de interés es explicada por la regresión, y 100% si toda la variación está explicada por el modelo.

Como se puede demostrar fácilmente, el coeficiente de correlación es la raíz cuadrada del coeficiente de determinación anterior:

$$r = \sqrt{r^2} = \sqrt{\frac{SCR}{SC}} = \sqrt{\frac{b \sum xy}{\sum x^2}} = \sqrt{\frac{(\sum xy)^2}{\sum x^2 \sum y^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

2.3.2. Modelo I y modelo II.

Muchas veces interesa distinguir entre los casos en los que X es una variable fija de los que es una variable aleatoria. Snedecor (1956) llama respectivamente modelo I y II de Eisenhart a esos casos. El modelo I es cuando la variable X es fijada por el experimentador de tal modo que puede ser manejada. Por el contrario, en el modelo II los valores de X son aleatorios, es decir que no se pueden prever con certeza. Supongamos que queremos estudiar la relación entre el peso de animales y su edad. Si elegimos animales de 1, 2, 3, meses y medimos su peso, estamos en un caso de modelo I. Si elegimos animales al azar y registramos su edad y su peso, estamos en un caso de modelo II. En el primer caso los valores de la edad de los animales son elegidos, en la segunda situación los valores de edad son aleatorios. El uso de los coeficientes de correlación tiene únicamente un sentido pleno en el caso de modelo II. Por una discusión más avanzada vease Graybill (1976). El siguiente es un ejemplo de modelo II.

Ejemplo 2.12. En un estudio se recogieron datos para estudiar la relación entre el contenido en colesterol, la edad y la masa corporal de las personas. Los datos son:

OBS	COLEST	EDAD	MASA
1	5.94	52	20.7
2	4.71	46	21.3
3	5.86	51	25.4
4	6.52	44	22.7
5	6.80	70	23.9
6	5.23	33	24.3
7	4.97	21	22.2
8	8.78	63	26.2
9	5.13	56	23.3
10	6.74	54	29.2
11	5.95	44	22.7
12	5.83	71	21.9
13	5.74	39	22.4
14	4.92	58	20.2
15	6.69	58	24.4
16	6.48	65	26.3
17	8.83	76	22.7
18	5.10	47	24.5
19	5.81	43	20.7
20	4.65	30	18.9
21	6.82	58	23.9
22	6.28	78	24.3
23	5.15	49	23.8
24	2.92	36	19.6
25	9.27	67	24.3
26	5.57	42	22.0
27	4.92	29	22.5
28	6.72	33	24.1
29	5.57	42	22.7
30	6.25	66	27.3

ANOVA SECUENCIAL EN UN MODELO II. Supongamos que corremos los tres modelos:

$$\text{Colesterol} = \beta_0 + \beta_1 \text{EDAD} + \beta_2 \text{MASA}$$

$$\text{Colesterol} = \beta_0 + \beta_1 \text{EDAD}$$

$$\text{Colesterol} = \beta_0 + \beta_2 \text{MASA}$$

Un análisis tipo SAS con el resultado de la regresión de los datos de Dobson (1990) es:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	22.31445	11.15723	10.999	0.0003
Error	27	27.38830	1.01438		
C Total	29	49.70275			
Root MSE	1.00716	R-square	0.4490		
Dep Mean	6.00500	Adj R-sq	0.4081		
C.V.	16.77211				
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	18.06660	18.06660	15.990	0.0004
Error	28	31.63615	1.12986		
C Total	29	49.70275			
Root MSE	1.06295	R-square	0.3635		
Dep Mean	6.00500	Adj R-sq	0.3408	C.V.	17.70108
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	12.73096	12.73096	9.642	0.0043
Error	28	36.97179	1.32042		
C Total	29	49.70275			
Root MSE	1.14910	R-square	0.2561		
Dep Mean	6.00500	Adj R-sq	0.2296	C.V.	19.13565

El análisis de varianza muestra un pvalue de 0,0003, es decir que el modelo es significativo al uno por mil. El R^2 es 0,449 de modo que el 44,9% de la variación en colesterol está asociada con las variaciones de edad y peso corporal.

Modelo: $\text{Colesterol} = -0,435 + 0,042 \text{EDAD} + 0,184 \text{MASA}$ con $R^2=0,45$

El modelo (2) $\text{Colesterol} = 3,296 + 0,053 \text{EDAD}$ tiene un $R^2=0,36$

El modelo (3) $\text{Colesterol} = -0,827 + 0,293 \text{MASA}$ tiene un $R^2=0,26$

$R(E,M) = 22,31$	$R^2=0,45$
$R(E) = 18,07$	$R^2=0,36$
$R(M E) = 4,24$	$R^2=0,09$
$R(M) = 12,73$	$R^2=0,26$
$R(E M) = 9,58$	$R^2=0,19$

Notemos, entre otras cosas, que $18,07+12,73 \neq 22,31$, de modo que el modelo completo explica menos que la suma de los dos modelos simples. Esto se debe a que existe correlación entre las variables edad y peso. Las suma de cuadrados parciales son: $R(M|E) = 22,31-18,07 = 4,24$ de modo que el modelo completo explica 22,31 la edad explica 18,07 y el aporte adicional del peso en un modelo que ya tenía a la edad es: 4,24. Similarmente podemos calcular la suma de cuadrados adicional (marginal) para la edad, $R(E|M) = 22,31-12,73=9,58$. La partición de la suma de cuadrados secuencial se llama en SAS tipo I, y depende del orden en que coloquemos a las variables en el modelo. La suma de cuadrados marginal se llama tipo III y no depende del orden, pero tampoco suman el total.

Cuando observamos los coeficientes individuales vemos que el de la edad es claramente significativo, pero el de la masa corporal es casi-no-significativo al 5%.

2.3.3. Correlación total y parcial.

Cada variable independiente tiene su correlación individual con la variable de interés, la

$$\text{que se mide por el coeficiente de correlación } r: r = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}}$$

Este coeficiente de correlación se individualiza en estudios de correlación múltiple con el nombre de coeficiente de correlación total o simple, ya que se desinteresa de los efectos de los otros factores. El problema de la correlación total es que incluye todo. La correlación total determinada entre la variable de interés y cada una de las variables independientes puede no indicar mucho, ya que la influencia de otros factores puede proporcionar una imagen distorsionada de la relación.

Ejemplo 13. Los datos de colesterol, edad y masa corporal de personas permiten calcular la siguiente matriz de correlaciones totales.

Correlation Analysis			
Pearson Correlation Coefficients / N = 30			
	COLEST	EDAD	MASA
COLEST	1.00000	0.60290	0.50610
EDAD	0.60290	1.00000	0.39371
MASA	0.50610	0.39371	1.00000

La *correlación total* o *correlación simple* vale en este caso 0,603 entre colesterol y edad, 0,506 entre colesterol y masa corporal y 0,394 entre edad y masa corporal. Nos podemos preguntar ¿Cuál es la correlación entre el colesterol y la edad descontado el efecto de la masa? Que es como si nos preguntaáamos: ¿Qué correlación hay entre colesterol y edad si las personas no engordaran? Los coeficientes de correlación parcial se encuentran a partir de

$$\text{los de correlación total del siguiente modo: } r_{YX1.X2} = \frac{r_{YX1} - r_{YX2} * r_{X1X2}}{\sqrt{(1 - r_{YX2}^2)(1 - r_{X1X2}^2)}}$$

donde $r_{YX1.X2}$ es el coeficiente de correlación parcial entre Y y X_1 descontado el efecto de X_2 y r_{X1X2} , r_{YX1} y r_{YX2} son los coeficientes de correlación total. En el presente caso el cálculo de la *correlación parcial*, se muestra en el siguiente proceso.

$$r_{CE.P} = \frac{0,6029 - (0,5061)(0,3937)}{\sqrt{(1 - 0,5061^2)(1 - 0,3937^2)}} = \frac{0,6029 - 0,19925}{\sqrt{(0,7439)(0,8450)}} = \frac{0,40365}{0,7928} = 0,5091$$

Coeficientes de paso.

$$\begin{vmatrix} 1 & 0,39371 \\ 0,39371 & 1 \end{vmatrix} = \begin{vmatrix} P1 \\ P2 \end{vmatrix} = \begin{vmatrix} 0,5061 \\ 0,39371 \end{vmatrix}$$

O sea que la situación es:

Edad 0,60290	Edad 0,477689
0,39371 Colesterol	0,39371 Colesterol
Masa 0,50610	Masa 0,318029

$$\text{Entonces } r_{EC} = 0,60290 = 0,477689 + 0,39371 * 0,318029$$

$$r_{MC} = 0,50610 = 0,318029 + 0,39371 * 0,477689$$

En el ejemplo de Bottaro (1973), la correlación entre rendimientos y nitrógeno es:

$$r_{YN} = \frac{10160}{28,8159670,77} = \frac{10160}{15115,212} = 0,672$$

mientras que entre rendimiento y fósforo es: $r = 8730/28,8159630,77 = 8730/15115,212 = 0,577$

$$\text{Tenemos: } r_{YN,P} = \frac{r_{YN} - r_{YP} * r_{NP}}{\sqrt{(1 - r_{YP}^2)(1 - r_{NP}^2)}} = \frac{0,672 - 0,577 * 0}{\sqrt{(1 - 0,577^2)(1 - 0)}}$$

Finalmente se puede destacar que la relación entre el coeficiente de correlación múltiple y los coeficientes de correlación parcial esta dada por la expresión:

$$1 - R^2 = (1 - r^2_{YX1})(1 - r^2_{YX2.X1}) \dots (1 - r^2_{YXk.X1 \dots X(k-1)})$$

por ejemplo: $1 - 0,886^2 = 0,541672 * 0,2208506 = 0,672 = 0,548 * 0,393159 = 0,215$

Ejemplo 2.12. En el caso que venimos estudiando el 78,5% de la variación en el rendimiento esta asociado a variaciones en los valores de N y P agregados a la pradera ($R^2 = 0,785$), el 45,2% esta explicado por la relación con el nitrógeno ($r^2_{YN} = 0,452$) y el restante 33,3% por el fósforo ($r^2_{YP} = 0,333$). Es decir que la variación explicada por ambas variables es la suma de las variaciones asociadas a cada una de ellas por separado. Este resultado es debido a que el diseño fue construido de modo que las variables actuaran independientemente, es decir ortogonalmente. En efecto la correlación entre ambas variables es nula: $r_{NP} = 0$.

2.3.4. Correlación múltiple y parciales.

A la proporción, o más comúnmente el porcentaje, de la variación explicada por el modelo de la variación total de los datos se le conoce como **coeficiente de determinación** R^2 :

$R^2 = \frac{SCR}{SC}$. El **coeficiente de correlación múltiple** es la raíz cuadrada positiva del coeficiente

de determinación: $R = \sqrt{R^2} = \sqrt{\frac{SCR}{SC}}$. cuadrada de esa cantidad, por lo tanto

$R = \sqrt{0,4490} = 0,67$. La correlación múltiple es siempre positiva, y varía entre 0 y 1. En el

ejemplo de Bottaro (1973) tenemos: $R^2 = \frac{6 : 408.584,3}{8 : 159.630,7} = 0,785$ y $R = 0,886$. El coeficiente de

determinación indica que el 78,5% de la variación de la variable de interés esta “explicado” por la regresión múltiple.

Ejemplo 2.13. El siguiente cuadro muestra la relación entre el valor de granjas, el número de hectáreas irrigadas (X_1), y el número de hectáreas de secano (X_2). Ejemplo ficticio tomado de Ezequiel, 1942).

Granja	Y	X_1	X_2
1	1.900	8	5
2	1.500	4	5
3	1.500	3	10
4	1.860	7	8
5	1.900	7	10
6	2.100	8	15
7	1.840	6	12
8	1.400	1	15
9	1.740	4	17
10	1.640	2	22
11	1.800	4	20
12	1.760	5	13

Si ajustamos el modelo $Y = B_0 + B_1 X_1 + E$ tenemos: $Y = 1365 + 77,14 X_1$ con $R^2 = 0,780$. El 78% de la variación de Y resulta explicada por su relación con X_1 .

Si ajustamos el modelo para la relación con X_2 y realizamos el mismo análisis, tenemos: $Y = 1754 - 0,74 X_2$ con $R^2 = 0,00$. No habría relación entre Y y X_2 .

Finalmente, si ajustamos el modelo $Y = B_0 + B_1 X_1 + B_2 X_2 + E$ tenemos:

$Y = 1000 + 100 X_1 + 20 X_2$ con un $R^2 = 1,00$!

Por tanto el 100% de la variación de Y resulta ahora explicada por el modelo.

Moraleja: la relación entre Y y X_2 es perfecta si consideramos X_1 mientras que anteriormente habíamos pensado que no había relación entre ellas.

¿Cuanto valdrá el coeficiente de correlación parcial de Y con cada una de las variables independientes? Compare con el coeficiente de correlación total.

En el ejemplo, el porcentaje de la variación de la variable Y asociado con X_1 es de 1,65% ($r^2_{YX_1} = 0,01652$) y el asociado con X_2 63,53% ($r^2_{YX_2} = 0,63534$). Si bien la parte explicada por las variables independientes no suman 100% el modelo multivariante tiene un $R^2 = 100\%$.

Es decir, que el aporte de X_2 sola es 63,53% pero el aporte de X_2 en adición a lo que aporta X_1 es de 99,98%! A su vez, X_1 tomada sola explica el 1,65% de la variación de Y , pero tomada en adición a X_2 explica el 36,466%. La relación entre Y y X_1 está totalmente desfigurada por la presencia de X_2 de modo que, a pesar de que la variable independiente tiene en realidad gran influencia en Y , el coeficiente de correlación total es solo 0,128.

2.3.5. Correlaciones y anova secuencial.

Puede existir interés en probar la significación de los coeficientes de correlación parcial o múltiple, del mismo modo que se prueban los de correlación total. El modo más fácil de hacerlo es a través de un análisis de varianza secuencial que veremos ahora.

Las preguntas son: ¿cuánto aporta el modelo con X_1 y X_2 ? ¿Cuánto aporta X_1 sola? ¿Cuánto agrega X_2 por sobre lo que aporta X_1 ? ¿Vale la pena conservar a X_2 o por el contrario el modelo explica casi lo mismo con o sin ella?. Viceversa, ¿cuánto aporta X_1 por sobre lo que aporta X_2 ? puede suceder que todo el valor del modelo este dado por X_2 y no haya ventaja en mantener a X_1 en el modelo. Todas estas respuestas están contestadas por $SC(X_1|X_2)$.

Cuando las X 's son variables aleatorias (y no fijas) el modelo se conoce como modelo II por algunos autores (Snedecor y Cochran, 1963; Steel y Torrie, 1980). Los casos más presentados son cuando las X 's y la variable Y tienen conjuntamente una distribución, como la normal multivariante. En estos casos el problema de interés es encontrar funciones de predicción de la esperanza condicional de la variable Y . En realidad ahora todas las variables están en un plano de mayor igualdad y se puede predecir cualquiera de ellas a partir de las otras. El tema se vincula entonces a un enfoque de como encontrar predictores óptimos (BP-best predictors), predictores que sean óptimos dentro de los lineales (BLP-best linear predictors), o predictores que sean óptimos entre los lineales insesgados (BLUP- best linear unbiased predictors). El tema tiene gran aplicación en genética y otros campos y lo trataremos en otra parte de estas notas. Por un enfoque aplicado se puede ver a Henderson (1984) y Graybill (1976).

2.3.6. Coeficiente de correlación intraclase

Un coeficiente de correlación definido diferente es el coeficiente de correlación intraclase. Este coeficiente es generalmente usado cuando hay valores agrupados (en grupos de más de dos) donde el concepto de correlación entre dos valores no tienen vigencia. Por detalles del concepto de correlación intraclase puede verse Steel & Torrie (1980), Snedecor & Cochran (1967), o en otras de estas notas.

UNA SITUACIÓN DIFERENTE.

Ejemplo 2.14. La tabla siguiente muestra los aumentos de peso vivo, 252 días post-destete por día y en kg., realizados por animales vacunos. Para cada animal se muestra su caravana y el valor del crecimiento de peso y el promedio de ambos padres.

Progenie		Padre		Madre		Promedio
No	Aumento	No	Aumento	No	Aumento	Padres
1	0,782	183	0,793	44	0,562	0,678
2	0,817	188	0,884	140	0,610	0,747
3	0,763	189	0,873	86	0,522	0,698
4	0,815	189	0,873	122	0,575	0,724
5	0,775	189	0,873	58	0,628	0,751
6	0,720	191	0,698	65	0,507	0,603
7	0,782	191	0,698	33	0,594	0,646
8	0,759	191	0,698	39	0,564	0,631
9	0,853	192	0,976	114	0,520	0,748
10	0,807	192	0,976	17	0,632	0,804
11	0,800	194	0,964	28	0,615	0,790
12	0,856	194	0,964	9	0,504	0,734

Algunos análisis que pueden interesar son: el cálculo de la regresión de los valores fenotípicos de la progenie sobre el valor de alguno o el promedio de los padres, encontrar el coeficiente de correlación, etc.

A los efectos de ajustar la regresión de los valores fenotípicos de la progenie sobre el promedio de los padres consideramos Y a la segunda columna y X a la sexta. Tenemos entonces: $\sum Y = 9.529$ $\sum X = 8,5515$ las sumas de cuadrados y productos sin corregir son:

$$\sum Y^2 = 7,583631 \quad \sum X^2 = 6,137417 \quad \sum XY = 6,808671$$

de modo que las medias son: $\bar{Y} = 0,794083$ $\bar{X} = 0,712625$ y las sumas de cuadrados y productos corregidas son: $\sum y^2 = 0,016810$ $\sum x^2 = 0,043404$ $\sum xy = 0,018067$

El coeficiente de regresión es: $b = 0,018067 / 0,043404 = 0,416255$

y el de correlación es: $r = 0,018067 / 0,027012 = 0,68854$

Considerando que $a = 0,497449$, la recta de regresión de los valores fenotípicos de la progenie sobre el promedio de los padres sería: $\hat{Y} = 0,497449 + 0,416255 X$ lo cual no es muy representativo de la situación (¿qué interés tiene el valor de Y cuando X es cero? Es decir ¿Qué valor tendría un animal cuyos padres tengan 0 de crecimiento?). Por lo tanto optaremos por una recta de regresión de la forma:

$$\hat{Y} = \mu_Y + b (X - \mu_X)$$

que ser: $\hat{Y} = 0,794083 + 0,416255 (X - 0,712625) = 0,794083 + 0,416255 x$

Nótese a este como un típico ejemplo de modelo II, ya que los valores de los progenitores son aleatorios.

RESUMEN DE ELEMENTOS TEÓRICOS DE INTERÉS

Por la naturaleza de esta publicación no hemos enfatizado mucho los conceptos teóricos. Puede ser de interés recopilar acá algunos elementos teóricos de los modelos de regresión, no solo por la importancia que tienen en sí mismo, sino porque son la base de modelos que veremos mas adelante.

Propiedades.

1) $\sum e = 0$. La suma de los errores es cero. Toda recta de regresión que pase por el punto \bar{X}, \bar{Y} tiene una suma de errores nula.

1b) $\sum X_1 e = 0$. También la suma de los productos de los errores por los valores de la variable X son nulos.

1c) $\sum X_2 e = 0$ Estas ecuaciones (1) tomadas en conjunto son las ecuaciones normales.

2) $\sum Y = \sum \hat{Y} + \sum e$ La suma de cuadrados total es la suma de cuadrados de la regresión mas la suma de cuadrados del error.

3) $\sum Y e = 0$ Así como la propiedad (1) esta dice que la suma de los productos de los valores de \hat{Y} por los errores es nula.

$$4) R^2_{0(12)} + R^2_{0.12} = 1$$

$$5) SCR = b_1 \sum x_1 y + b_2 \sum x_2 y = b_1^2 \sum x_1^2 + 2b_1 b_2 \sum x_1 x_2 + b_2^2 \sum x_2^2$$

Dos maneras ilustrativas de calcular la suma de cuadrados de la regresión.

6) $R = r_{Y\hat{Y}}$ El coeficiente de correlación múltiple es la correlación entre los valores predichos de Y y los observados.

$$7) \sum Y\hat{Y} = \sum \hat{Y}^2 = \sum Y(\hat{Y} + e) = \sum \hat{Y}^2 + \sum \hat{Y}e = 0$$

2.4. COMPUTACIÓN Y REGRESIÓN

2.4.1. Regresión en el SAS

Como decíamos en la sección 1.4 anterior, SAS dispone de diferentes mecanismos para estudiar regresiones y correlaciones. El PROC CORR calcula las correlaciones de "todos contra todos" de una serie de variables. Específicamente para regresión se dispone de: PROC REG, PROC RSQUARE, PROC RSREG, PROC NLIN, PROC STEPWISE. El PROC REG es el sistema básico para regresión y lo veremos con más detalle a continuación. El PROC RSQUARE, el PROC RSREG y el PROC STEPWISE sirven para modelación. El PROC NLIN sirve para el ajuste de regresiones no lineales.

DESCRIPCION DEL PROC REG

```
PROC REG ajusta estimadores por mínimos cuadrados a modelos de regresión
PROC REG DATA= OUTEST= OUTSSCP= NOPRINT SIMPLE USSCP ALL COVOUT CORR
SINGULAR=n;
  MODEL variable dependiente=regresores /
METHOD=NONE|FORDWARD|BACKWARD|STEPWISE|MAXR|MINR|RSQUARE
  SLENTY= SLSTAY= SELECT= INCLUDE= START= STOP=
  NOPRINT NOINT ALL XPX I SS1 SS2 STB P R
  CLM VIF COVB CORRB COLLIN COLLINOUT TOL
  CLI DW INFLUENCE PARTIAL DETAILS SIGMA= ADJRSQ
  AIC BIC CP GMSEP JP MSE PC RMSE SBC SP SSE B;
  VAR variables;
  ID variables;
  FREQ variables;
  WEIGHT variables;
  ADD variables;
  DELETE variables;
  DELOBS n;
PRINT ALL XPX I SS1 SS2 STB TOL VIF COVB CORRB COLLIN COLLINT
  P R CLM CLI DW INFLUENCE PARTIAL ANOVA;
OUTPUT OUT=SASdataset PREDICTED= RESIDUAL= L95M= U95M=
  L95= U95 STDP= STDR= STDI= STUDENT= COOKD= H= PRESS=
  RSTUDENT= DFFITS= COVRATIO= ;
TEST ecuación1,...,ecuaciónk / PRINT;
MTEST ecuación1,...,ecuaciónk / PRINT CANPRINT DETAILS ;
BY variables;
```

El PROC REG se acompaña siempre de una o más sentencias MODEL para especificar modelos de regresión. Una sentencia OUTPUT debe seguir a cada sentencia MODEL. Varias sentencias TEST y MTEST pueden seguir a cada MODEL. Las sentencias WEIGHT, FREQ, y ID se especifican opcionalmente, una para el PROC en su totalidad.

2.4.2. Ejemplos.

Veremos la aplicación del PROC REG a través de los siguientes ejemplos. El ejemplo 2.4 corrido en el SAS es:

```
data uno; input edad peso; cards;
1 4.0
2 5.9
3 4.7
4 7.1
6 7.7
8 9.2
proc reg; model peso=edad; run;
proc glm; model peso=edad/xpx i; run;
```

y la salida es:

Model: MODEL1					
Dependent Variable: PESO					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	16.52029	16.52029	27.614	0.0063
Error	4	2.39304	0.59826		
C Total	5	18.91333			
Root MSE		0.77347	R-square	0.8735	
Dep Mean		6.43333	Adj R- sq	0.8418	
C.V.		12.02289			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	3.645098	0.61744959	5.903	0.0041
EDAD	1	0.697059	0.13264945	5.255	0.0063

General Linear Models Procedure					
Number of observations in data set = 6					
The X'X Matrix					
	INTERCEPT	EDAD	PESO		
INTERCEPT	6	24	38.6		
EDAD	24	130	178.1		
PESO	38.6	178.1	267.24		
X'X Inverse Matrix					
	INTERCEPT	EDAD	PESO		
INTERCEPT	0.637254902	- 0.117647059	3.6450980392		
EDAD	- 0.117647059	0.0294117647	0.6970588235		
PESO	3.6450980392	0.6970588235	2.3930392157		
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.52029412	16.52029412	27.61	0.0063
Error	4	2.39303922	0.59825980		
Corrected Total	5	18.91333333			
R- Square		C.V.	Root MSE	PESO Mean	
0.873473		12.02289	0.77347256	6.43333333	
T for H0: Pr > T Std Error of					
Parameter	Estimate	Parameter=0		Estimate	
INTERCEPT	3.645098039	5.90	0.0041	0.61744959	
EDAD	0.697058824	5.25	0.0063	0.13264945	

- ❶ Los coeficientes coinciden con lo calculado en la página 57.
- ❷ Los errores estándar se comparan con lo obtenido en la página 58.
- ❸ El valor de t para probar la hipótesis de que el parámetro respectivo es cero, como se mostró en la página 60.
- ❹ SAS reporta el p-value permitiendo que hagamos la inferencia al nivel de significación deseado.

El PROC GLM de SAS es también muy eficiente en el análisis de regresión aunque tiene algunas diferencias con el PROC REG. Analizaremos el ejemplo 2.8 con el PROC GLM. El programa sería:

```
DATA DOS8;
INPUT N X1 P X2 Y;
X3=X1*X1; X4=X2*X2; X5=X1*X2;
CARDS;
80 -1 80 -1 1125
80 -1 240 1 1135
240 1 80 -1 1345
240 1 240 1 2185
160 0 160 0 1435
0 -2 160 0 480
320 2 160 0 1895
160 0 0 -2 850
160 0 320 2 2250
0 -2 0 -2 505
0 -2 320 2 615
320 2 0 -2 860
320 2 320 2 3290
PROC GLM; MODEL Y=X1 X2; RUN;
PROC GLM; MODEL Y=X1 X2 X3 X4 X5; RUN;
```

y la salida es:

General Linear Models Procedure					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6408517.8571	3204258.9286	18.30	0.0005
Error	10	1751112.9121	175111.2912		
Corrected Total	12	8159630.7692			
	R-Square	C.V.	Root MSE	Y Mean	
	0.785393	30.272783	418.46301	1382.3076923	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	3686628.571	3686628.571	21.05	0.0010
X2	1	2721889.286	2721889.286	15.54	0.0028
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	3686628.571	3686628.571	21.05	0.0010
X2	1	2721889.286	2721889.286	15.54	0.0028
Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate	
INTERCEPT	1382.307692	11.91	0.0001	116.06075704	
X1	362.857143	4.59	0.0010	79.08207563	
X2	311.785714	3.94	0.0028	79.08207563	
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8076072.6953	1615214.5391	135.31	0.0001
Error	7	83558.0739	11936.8677		
Corrected Total	12	8159630.7692			
	R-Square	C.V.	Root MSE	Y Mean	
	0.989760	7.9038823	109.25597	1382.3076923	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	3686628.571	3686628.571	308.84	0.0001
X2	1	2721889.286	2721889.286	228.02	0.0001
X3	1	137317.010	137317.010	11.50	0.0116
X4	1	27118.711	27118.711	2.27	0.1755
X5	1	1503119.118	1503119.118	125.92	0.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	3686628.5714	3686628.5714	308.84	0.0001

X2	1	2721889.2857	2721889.2857	228.02	0.0001
X3	1	156382.2769	156382.2769	13.10	0.0085
X4	1	27118.7110	27118.7110	2.27	0.1755
X5	1	1503119.1176	1503119.1176	125.92	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	1462.727273	25.64	0.0001	57.05707119
X1	362.857143	17.57	0.0001	20.64743819
X2	311.785714	15.10	0.0001	20.64743819
X3	-63.981331	-3.62	0.0085	17.67684264
X4	26.643669	1.51	0.1755	17.67684264
X5	148.676471	11.22	0.0001	13.24923289

La matriz $X'X$ y su inversa se proporcionan por haber incluido la opción /XPX I; luego del modelo. Compárense los resultados de esta salida con lo visto en la sección 2.2.6 y 2.3.6, donde se proporciona el coeficiente de determinación.

vemos que, básicamente, el PROC GLM proporciona la misma información que el PROC REG en los casos sencillos de regresión que venimos estudiando.

El ejemplo 2.11. corrido en el SAS. Comencemos por la parte de regresión simple:

```
DATA DOS11; INPUT NO AP NP APA NM AM AMP; CARDS;
1 0.782 183 0.793 44 0.562 0.678
2 0.817 188 0.884 140 0.610 0.747
3 0.763 189 0.873 86 0.522 0.698
4 0.815 189 0.873 122 0.575 0.724
5 0.775 189 0.873 58 0.628 0.751
6 0.720 191 0.698 65 0.507 0.603
7 0.782 191 0.698 33 0.594 0.646
8 0.759 191 0.698 39 0.564 0.631
9 0.853 192 0.976 114 0.520 0.748
10 0.807 192 0.976 17 0.632 0.804
11 0.800 194 0.964 28 0.615 0.790
12 0.856 194 0.964 9 0.504 0.734
PROC REG; MODEL AP=APA; RUN;
PROC REG; MODEL AP=AM; RUN;
PROC REG; MODEL AP=AMP; RUN;
```

y las salidas serían:

Model: MODEL1					
Dep Variable: AP					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.00978	0.00978	13.896	0.0039
Error	10	0.00704	0.00070		
C Total	11	0.01681			
Root MSE		0.02652	R-Square	0.5815	
Dep Mean		0.79408	Adj R-Sq	0.5397	
C.V.		3.34018			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	0.561741	0.06279720	8.945	0.0001
APA	1	0.271481	0.07282805	3.728	0.0039

Model: MODEL1					
Dep Variable: AP					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.00000	0.00000	0.000	0.9848
Error	10	0.01681	0.00168		
C Total	11	0.01681			
Root MSE		0.04100	R-Square	0.0000	
Dep Mean		0.79408	Adj R-Sq	-0.1000	
C.V.		5.16323			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	0.000549	0.00115202	0.477	0.6450
APA	1	0.499716	0.00078298	638.224	0.0001
AM	1	0.499829	0.00181593	275.246	0.0001
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	0.791183	0.14914287	5.305	0.0003
AM	1	0.005093	0.26109613	0.020	0.9848

Model: MODEL1					
Dep Variable: AP					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.00747	0.00747	8.005	0.0179
Error	10	0.00934	0.00093		
C Total	11	0.01681			

	Root MSE	0.03056	R-Square	0.4446	
	Dep Mean	0.79408	Adj R-Sq	0.3891	
	C.V.	3.84797			
Parameter Estimates					
		Parameter	Standard	T for H0:	
Variable	DF	Estimate	Error	Parameter=0	Prob > T
INTERCEP	1	0.498128	0.10497387	4.745	0.0008
AMP	1	0.415182	0.14674204	2.829	0.0179

Comparese con lo visto en la seccion 2.3.3, la raiz cuadrada de 0,4446 es $r=0,6668$.

Ahora, con los mismo datos, consideramos la regresión múltiple. Para ello utilizamos el siguiente programa:

```
PROC REG DATA=DOS11; MODEL AMP= APA AM/XPX I; QUIT;
```

y obtenemos la siguiente salida:

Model Crossproducts X'X X'Y Y'Y					
X'X	INTERCEP		APA	AM	AMP
INTERCEP	12		10.27	6.833	8.554
APA	10.27		8.922048	5.856069	7.391159
AM	6.833		5.856069	3.915483	4.887193
AMP	8.554		7.391159	4.887193	6.140936
INVERSE	INTERCEP		APA	AM	AMP
INTERCEP	16.660154706		-5.136303642	-21.39207055	0.000549017
APA	-5.136303642		7.6958885087	-2.546631242	.49971587824
AM	-21.39207055		-2.546631242	41.39598265	.49982873212
AMP	0.000549017		.49971587824	.49982873212	7.1694186E-7
Dep Variable: AMP Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Prob>F
Model	2	0.04336	0.02168	272149.366	0.0001
Error	9	0.00000	0.00000		
C Total	11	0.04336			
	Root MSE	0.00028	R-Square	1.0000	
	Dep Mean	0.71283	Adj R-Sq	1.0000	
	C.V.	0.03959			
Parameter Estimates					
		Parameter	Standard	T for H0:	
Variable	DF	Estimate	Error	Parameter=0	Prob > T
INTERCEP	1	0.000549	0.00115202	0.477	0.6450
APA	1	0.499716	0.00078298	638.224	0.0001
AM	1	0.499829	0.00181593	275.246	0.0001

Entre los comentarios está el hecho obvio de que a efectos del SAS la distinción entre modelos de regresión simple o múltiple no tiene ninguna importancia, usándose la misma metodología en ambos casos.

Model: MODEL1 Dependent Variable: COLEST					
Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Prob>F
Model	2	22.31445	11.15723	10.999	0.0003
Error	27	27.38830	1.01438		
C Total	29	49.70275			
Root MSE	1.00716	R-square	0.4490		
Dep Mean	6.00500	Adj R-sq	0.4081		
C.V.	16.77211				
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-0.435015	1.94151154	-0.224	0.8244
EDAD	1	0.042341	0.01377528	3.074	0.0048
MASA	1	0.184421	0.09012105	2.046	0.0506

Model: MODEL2 Dependent Variable: COLEST					
Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Prob>F
Model	1	18.06660	18.06660	15.990	0.0004
Error	28	31.63615	1.12986		
C Total	29	49.70275			
Root MSE	1.06295	R-square	0.3635		
Dep Mean	6.00500	Adj R-sq	0.3408		
C.V.	17.70108				
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	3.295612	0.70480188	4.676	0.0001
EDAD	1	0.053440	0.01336405	3.999	0.0004

Model: MODEL3 Dependent Variable: COLEST					
Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Prob>F
Model	1	12.73096	12.73096	9.642	0.0043
Error	28	36.97179	1.32042		
C Total	29	49.70275			
Root MSE	1.14910	R-square	0.2561		
Dep Mean	6.00500	Adj R-sq	0.2296		
C.V.	19.13565				
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-0.827263	2.21032123	-0.374	0.7110
MASA	1	0.293482	0.09451642	3.105	0.0043

CORRELATION ANALYSIS

3 'VAR' Variables: COLEST EDAD MASA						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
COLEST	30	6.0050	1.3092	180.1500	2.9200	9.2700
EDAD	30	50.7000	14.7698	1521	21.0000	78.0000
MASA	30	23.2800	2.2576	698.4000	18.9000	29.2000
Pearson Correlation Coefficients / Prob > R under Ho: Rho=0 / N = 30						
		COLEST	EDAD	MASA		
COLEST		1.00000	0.60290	0.50610		
		0.0	0.0004	0.0043		
EDAD		0.60290	1.00000	0.39371		
		0.0004	0.0	0.0313		
MASA		0.50610	0.39371	1.00000		
		0.0043	0.0313	0.0		

1 'PARTIAL' Variables: MASA 2 'VAR' Variables: COLEST EDAD				
Simple Statistics				
Variable	N	Mean	Std Dev	Sum
MASA	30	23.280000	2.257615	698.400000
COLEST	30	6.005000	1.309155	180.150000
EDAD	30	50.700000	14.769843	1521.000000
Simple Statistics				
Variable	Minimum	Maximum	Partial Variance	Partial Std Dev
MASA	18.900000	29.200000	.	.
COLEST	2.920000	9.270000	1.320421	1.149096
EDAD	21.000000	78.000000	190.916119	13.817240
Pearson Partial Correlation Coefficients/Prob> R under Ho: Partial Rho=0 / N = 30				
		COLEST	EDAD	
COLEST		1.00000	0.50913	
		0.0	0.0048	
EDAD		0.50913	1.00000	
		0.0048	0.0	

1 'PARTIAL' Variables: EDAD 2 'VAR' Variables: COLEST MASA				
Simple Statistics				
Variable	N	Mean	Std Dev	Sum
EDAD	30	50.700000	14.769843	1521.000000
COLEST	30	6.005000	1.309155	180.150000
MASA	30	23.280000	2.257615	698.400000
Simple Statistics				
Variable	Minimum	Maximum	Partial Variance	Partial Std Dev
EDAD	21.000000	78.000000	.	.
COLEST	2.920000	9.270000	1.129863	1.062950
MASA	18.900000	29.200000	4.460574	2.112007
Pearson Partial Correlation Coefficients/Prob> R under Ho: Partial Rho=0 / N = 30				
		COLEST	MASA	
COLEST		1.00000	0.36643	
		0.0	0.0506	
MASA		0.36643	1.00000	
		0.0506	0.0	

2.4.3. Valoración de la regresión.

a. Como lo hemos discutido, la determinación de una relación entre variables no implica que esta exista, sino que la relación puede ser aparente debido al azar.

b. La línea de regresión determinada no tiene valor fuera del rango de variación de los datos que permitieron estimarla. Intentar extrapolar los resultados implica suponer que la relación se mantendrá; lo cual es un supuesto que no tiene ningún método estadístico de análisis para probarlo. Extrapolar implica asumir un riesgo que solo en casos excepcionales podrá justificarse por la naturaleza del problema.

c. La determinación de una línea de regresión y de una relación entre variables no implica, de modo alguno, encontrar una relación de causalidad entre las variables. Pueden determinarse ajustes excelentes entre variables que no tienen ninguna relación en lo que a veces se denomina correlaciones espurias. Este tipo de situación muchas veces se presenta cuando ambas variables se relacionan con una tercera, por ejemplo el tiempo.

d. La existencia de un coeficiente de correlación rectilínea no significativo no implica independencia, excepto en el caso de que ambas variables se distribuyan según una normal bivalente. Puede existir una relación de un tipo que no sea la rectilínea, por ejemplo cuadrática, y el coeficiente ser nulo o cercano a cero.

2.4.4. Diseño de experimentos para estudios de regresión.

El cumplimiento de las propiedades de la recta de regresión comentadas presenta puntos a ser tenidos en cuenta en el diseño de un estudio de regresión. Según lo que hemos visto la

varianza de la predicción del valor de Y cuando $X=x_0$ es proporcional a:
$$CME \left[\frac{1}{n} + \frac{x_0^2}{\sum X^2} \right]$$

de modo que, al aumentar la distancia entre X_0 y la media, aumenta x_0^2 y, por ende, todo el radio del intervalo de confianza. Este tipo de situación implica menor precisión, por lo que al diseñar un experimento debe buscarse que el promedio de los valores de X caiga en las cercanías del valor de X_0 para el cual interesa realizar predicción. Es decir, deben tomarse valores de X a ambos lados del valor que interesa predecir. El hecho de que los valores de X están distantes entre sí contribuye a una mayor precisión, ya que esto hace que $\sum x^2 = \sum (X - \bar{X})^2$ aumente, con lo que disminuye el factor arriba presentado y se estrecha el intervalo de confianza para la predicción. Por el contrario, si se desea estudiar la curvilinearidad de una regresión, puede interesar que los valores de la variable independiente estén a distancias equidistantes, en todo el rango de interés de la variable. En el capítulo 5 nos dedicaremos a estudiar con más detalle este tipo de temas, pero adelantamos algunos conceptos.

Importancia de la ortogonalidad. La ortogonalidad de los factores asegura que la suma de los productos es cero, por lo tanto facilita los cálculos y las interpretaciones. Muchos investigadores no se toman el trabajo de buscar ortogonalidad y luego se quejan de la dificultad de interpretar situaciones con variables no ortogonales. Tampoco es una cosa que no se puede vivir sin ella, solo que hay que sobrellevar las consecuencias. En los "observational studies" la falta de ortogonalidad no se puede eludir y es una de las características de tales estudios.

2.4.5. Funciones linearizables y no lineales.

Dentro de las relaciones curvilíneas existe un conjunto que son susceptibles de ser estudiadas por los métodos descriptos para relaciones rectilíneas, mediante transformaciones de variables. Este tipo de fenómeno se denomina linealidad por anamorfosis por algunos autores (Pimentel Gomes, 1975).

Por ejemplo, en estudios de nutrición animal se considera que el consumo de energía es proporcional a una potencia del peso del animal: $C=A.W^k$. Si aplicamos logaritmos a esa relación: $\log C = \log A + k*\log W$ y tomando $\log C = Y$; $\log A = a$; y $\log W = X$; tenemos: $Y = a + k X$

Otro ejemplo está dado por las relaciones similares a la de Mitscherlich: $Y = A.B^X$. Tomando logaritmos: $\log Y = \log A + (\log B) X$ y se expresa linealmente como $Y' = a + bX$. Las transformaciones de éste tipo se conocen como semilogarítmicas, al tomarse el logaritmo de solamente una de las variables; por oposición, las del ejemplo anterior se conocen como logarítmicas al ser lineal cuando ambas variables son transformadas.

Figura 2.10. Diferentes tipos de curvas y transformaciones sugeridas para linealizarlas.