UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

DIVISIÓN DE AGRONOMIA

Temas de Covarianza y Regresión en Agronomía

POR:

ROBERTO BETANCOURT CORVERA

MONOGRAFIA

Presentada como Requisito Parcial para
Obtener el Titulo de:

Ingeniero Agrónomo Fitotecnista

Buenavista, Saltillo, Coahuila, México.

Agosto 2002

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO
DIVISION DE AGRONOMIA

Temas de Covarianza y Regresión en Agronomía

POR:

ROBERTO BETANCOURT CORVERA

MONOGRAFIA

Que somete a la consideración del H. Jurado examinador como requisito parcial para obtener el titulo de:

Ingeniero Agrónomo Fitotecnista

APROBADA

Asesor principal

_____

MC. Jaime M Rodríguez Del Ángel

Asesor	**Asesor**
MC Humberto Macias Hernández	**ING. Jesús Macias Hernández**

El Coordinador de la División de Agronomía

_____

MC Reynaldo Alonso Velasco

Buenavista, Saltillo, Coahuila Agosto 2002

Índice de contenido

Introducción

La calidad del liderazgo científico es ciertamente un factor vital en el éxito de cualquier campaña de producción. Es deplorable pero cierto que numerosos científicos agrícolas de muchos países desarrollados renuncien a su lealtad a la agricultura por razones de conveniencia y de presunto prestigio. Algunas instituciones, a su vez, les hacen una cortina detrás de la cual se pueden ocultar. El caso va más allá todavía, pues según se observa, no pocas organizaciones de educación e investigación restringen la magnitud de la investigación fundamental que puede hacerse bajo la égida de sus departamentos agrícolas, independientemente de cuán básicos sean esos estudios para incrementar la producción de alimentos. Dejemos a los individuos vivir con sus propias motivaciones, dejemos que sirvan a la ciencia y a ellos mismos si así lo desean. Pero las instituciones tienen la obligación moral de servir también a la agricultura y a la sociedad, y para cumplir honorablemente con esa obligación deben contribuir a la educación de los investigadores y de los lideres científicos cuya motivación principal es servir a la humanidad.

Bajo este contexto entonces, es importante que las instituciones agrícolas como la nuestra, preparen profesionales capaces de generar investigación y a la vez puedan interpretar los resultados de la misma, con el propósito de transmitir los nuevos conocimientos en las áreas donde se genera la producción. Lo anterior requiere de herramientas técnicas como el conocimiento del método científico y su relación con la Estadística y los Diseños Experimentales.

En el estudio de fenómenos aleatorios en agronomía, la Estadística auxilia al método científico, en la colección de los hechos por observación o experimentación, en la comprobación de los mismos a través de la formulación de hipótesis en términos de causa efecto, en la aplicación de herramientas de medición como los Diseños Experimentales y por ultimo en la interpretación de resultados que surgen bajo ciertas condiciones y que posteriormente deberán ser verificados mediante nuevas observaciones.

Por otra parte cuando hablamos de la adecuación de la regresión lineal para muchas situaciones de estudio de fenómenos en las ciencia agrícola, encontramos

que algunas variables no se conectan entre sí por una relación tan simple. El descubrir una descripción precisa de la relación entre dos o más cantidades es uno de los problemas de ajuste de curva que se conoce como regresión curvilínea. Desde este punto de vista general, el ajuste de la recta no es mas que un caso especial, que es el mas sencillo de todos y en realidad el de mayor utilidad. Son varios los motivos para ajustar curvas a datos no lineales. Algunas veces una buena estimada de la variable dependiente es la que se busca, correspondiente a cualquier valor particular de la independiente. Esto puede comprender el pulir datos irregulares y la interpolación de las Y estimadas para valores de X que no estén dentro de la serie observada. Algunas veces la idea es probar alguna ley que relacione las variables, como una curva de crecimiento que haya sido propuesta por alguna investigación anterior, o del análisis matemático del mecanismo que conecte las variables.

El análisis de la Covarianza trata de dos o más variables medidas y donde cualquier variable independiente no se encuentra a niveles predeterminados. Los usos mas importantes del análisis de la Covarianza son: controlar el error y aumentar la precisión, ajustar medias de tratamientos de la variable dependiente de las diferencias en conjunto de valores de variables independientes correspondientes, ayudar en la interpretación de datos, especialmente en lo concerniente a la naturaleza de los efectos de los tratamientos, y por ultimo seccionar una Covarianza total o suma de productos cruzados en componentes.

El presente trabajo, sin pretender ser un compendio sobre los temas que aquí se abordan, trata de una recopilación sobre Covarianza y Regresión y su utilización en las ciencias agropecuarias, esto debido principalmente a que la literatura común utilizada durante nuestra formación académica, no contempla la aplicación de la Estadística en esta especialidad. Así mismo este trabajo independientemente de complementar mi formación profesional, servirá como requisito para mi titilación como Ingeniero Agrónomo Fitotecnista en la modalidad de Monografía, que se contempla en el Articulo 85° - IV, del reglamento académico para alumnos de licenciatura de la Universidad Autónoma Agraria Antonio Narro.

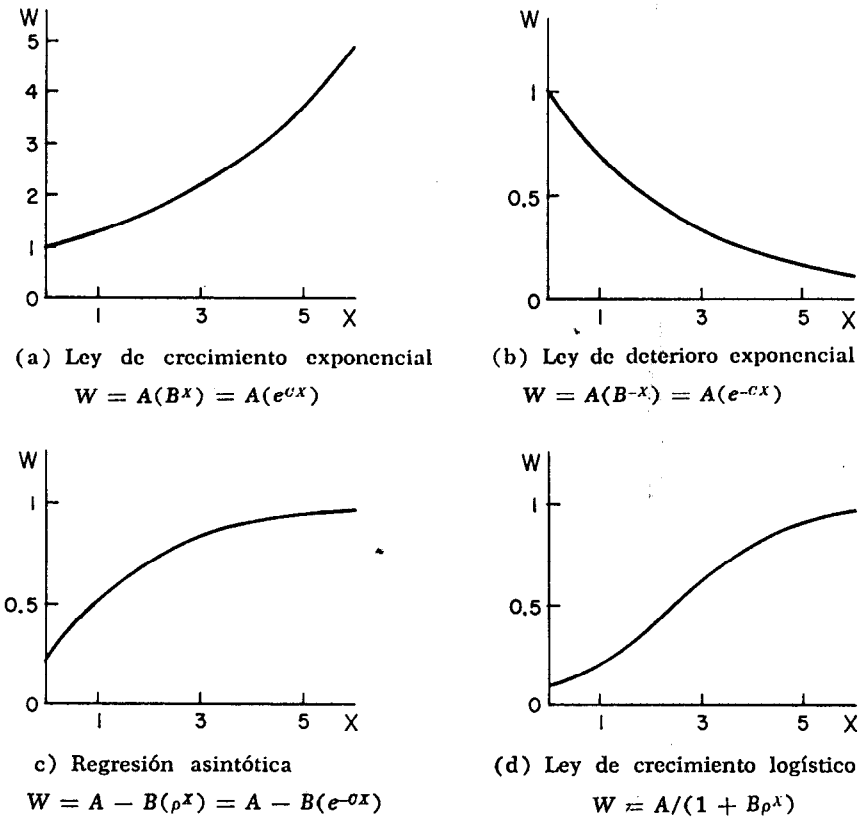
Regresión curvilínea

Introducción

Aunque la regresión lineal es adecuada para muchas situaciones, algunas variables no se conectan entre sí por una relación tan simple. El descubrir una descripcion precisa de la relación entre dos o más cantidades es uno de los problemas de ajuste de curva que se conoce como regresión curvilínea. Desde este punto de vista general, el ajuste d la recta no es mas que un caso especial, que es el mas sencillo de todos y en realidad el de mayor utilidad.

Son varios los motivos para ajustar curvas a datos no lineales. Algunas veces una buena estimada de la variable dependiente es la que se busca, correspondiente a cualquier valor particular de la independiente. Esto puede comprender el pulir datos irregulares y la interpolación de las Y estimadas para valores de X que no estén dentro de la serie observada. Algunas veces la idea es probar alguna ley que relacione las variables, como una curva de crecimiento que haya siso propuesta por alguna investigacion anterior, o del análisis matemático del mecanismo que conecte las variables. Otras veces, la forma en si de la relación es de poco interés; siendo la meta la mera eliminación de impresiones que pueda introducir la no linealidad de la regresión en un coeficiente de correlación o en un error experimental.

La figura siguiente muestra cuatro relaciones comunes no lineales. La parte (a) es la ley de interés compuesto o curva de crecimiento exponencial $W = A (B^x)$, en la que hemos utilizado W en lugar de Y acostumbrada. Si B =1 + i, donde i es la tasa de interés anual, W nos da la cantidad que alcanzara una suma de dinero A, si se le deja X años a interés compuesto. Como ya veremos, esta curva también representa la forma en que algunos organismos crecen durante ciertas etapas. La curva que se muestra en la parte (a) tiene A = 1.



(a) Ley de crecimiento exponencial $W = A(B^X) = A(e^{cX})$

(b) Ley de deterioro exponencial $W = A(B^{-X}) = A(e^{-cX})$

c) Regresión asintótica $W = A - B(\rho^X) = A - B(e^{-cX})$

(d) Ley de crecimiento logístico $W = A/(1 + B\rho^X)$

FIG. 15.1.1. Cuatro curvas comunes no lineales

Si B es menor que 1, la curva toma la forma que se muestra en (b). A menudo se le designa curva exponencial de decadencia, decreciendo el valor de W a cero, desde su valor inicial A, conforme aumenta X. El deterioro por emisiones de un elemento radiactivo se apega a esta curva.

La curva en © es $W = A - B\rho^X$, con $0 < \rho < 1$. Esta curva surge del valor (A-B) cuando X es 0, y continuamente se aproxima a un máximo valor de A, designada la asintota, conforme que X se va haciendo grande. La curva se conoce con diversos nombres. En agricultura se le ha conocido como la ley de Mitscherlich, por un quimico alemán quien la utilizo para representar la relación entre rendimiento W de un cultivo (en macetas) y la cantidad de fertilizante X añadido al suelo en las macetas. En química, algunas veces se le designa curva de reaccion de primer orden. También se ha utilizado el nombre de regresión asintotica.

La curva (d), la ley de crecimiento logístico, ha desempeñado un papel prominente en el estudio de poblaciones humanos. Esta curva de un ajuste notablemente bueno al crecimiento de población en Estados Unidos, de acuerdo con los datos de los censos decenales de 1970 a 1040.

En este capitulo vamos ilustrar como se ajustan tres tipos de curvas: (1) ciertas curvas no lineales, como en (a) y en (b) de la figura anterior que pueden ser reducidas a líneas rectas por medio de una transformación de la escala de W o de X; (2) el polinomio en X, que muchas veces sirve como buena aproximación; (3) curvas no lineales, como (c) y (d), también de la figura anterior que requieren metodos más complejos para el ajuste.

Ejemplo: El ajuste de la curva logística del censo poblacional de los Estados Unidos de Norteamérica (excluyendo Hawai y Alaska) para un periodo de 150 años de 1790 a 1940 es un ejemplo interesante tanto por la notable precesión en ajuste, como también por su tremenda falla al extrapolar para predecir poblaciones para 1950 y 1960. La curva ajustada por Pearl y Reed es:

$$W = \frac{184}{1 + (66.69)\,(10^{-0.1398x})}$$

donde X = 1 en 1970, y una unidad en X representa 10 años, de suerte que X = 16 en 1940. La tabla dada a continuación muestra la población real del censo, la población estimada por la logística y el error de estimación.

	Población		
Año	Real	Estimada	A-E
1970	3.9	3.7	+ 0.2
1800	5.3	5.1	+ 0.2
1810	7.2	7.0	+ 0.2
1820	9.6	9.5	+ 0.1
1830	12.9	12.8	+ 0.1
1840	17.1	17.3	-0.2
1850	23.2	23	+ 0.2

Continuación tabla anterior.

1860	31.4	30.3	+ 1.1
1870	38.6	39.3	- 0.7
1880	50.2	50.2	0.0
1890	62.9	62.8	+ 0.1
1900	76	76.7	- 0.7
1910	92	91.4	+ 0.6
1920	105.7	106.1	- 0.4
1930	122.8	120.1	+ 2.7
1940	131.4	132.8	- 1.4
1950	150.7	143.8	+ 6.9
1960	178.5	153	+ 25.5

Nótese cuan malas estaban las predicciones para 1950 y 960. La predicción por la curva decía que la población de Estados Unidos nunca excedería los 184 millones; la población real en 1966 ya rebasaba bien los 190 millones. Dos de los factores responsables son el auge de bebes de la posguerra y los servicios de salubridad publica mejorados.

Curva de crecimiento exponencial.

Una característica de algunos de los fenómenos mas sencillos de crecimiento es que el aumento a cualquier momento es proporcional al tamaño ya alcanzado. Durante una fase en el crecimiento de un cultivo bacteriano, el numero de organismos se ajusta a esa ley. La relación queda bien ilustrada por el peso seco de embriones de pollitos de 6 a 16 días anotados en la tabla 15.2.1. La gráfica de pesos de la figura 15.2.1 asciende con mayor rapidez conforme aumenta la edad, siendo la forma de la ecuación de regresión

$W = (A)\ (B^X)$,

Donde A y B son constantes que hay que estimar. Tomando logaritmos a la ecuación, tenemos

$$\log W = \log A + (\log B) X$$

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

Donde Y = log W, α = log A, y β = log B. Esto significa que si log W en lugar de W, se gráfica contra X, la gráfica será lineal. La técnica de utilizar el logaritmo, en lugar de la cantidad misma, se designa rectificación de datos.

Tabla 15.2.1. PESO SECO DE EMBRIONES DE POLLO DE EDAD DE 6 A 16 DIAS, Y SUS LOGARITMOS COMUNES.

Edad en días X	Peso seco, W (gramos)	Logaritmos comunes del peso Y
6	0.029	-1.538
7	0.052	-1.284
8	0.079	-1.102
9	0.125	-0.903
10	0.181	-0.742
11	0.261	-0.583
12	0.425	-0.372
13	0.738	-0.132
14	1.130	0.053
15	1.882	0.275
16	2.812	0.449

- De la tabla de logaritmos, se puede leer log 0.029 = log 2.9 – log 100 = 0.462 – 2 = -1.538.

Los valores de Y = log W se encuentran en la ultima columna de la tabla, y estan graficados al lado opuesto de X en el grabado. La ecuación de regresión calculada en la forma acostumbrada de las columnas X y Y de la tabla es

Y = 0.1959X – 2.689

La línea de regresión se ajusta a los puntos de los datos con una fidelidad extraordinaria, siendo la correlación de Y sobre X, 0.9992. la conclusión a que se

llega es que los embriones de pollo, cuando se miden por peso seco, crecen de acuerdo con la ley exponencial, y el logaritmo del peso seco aumenta a ritmo estimado uniforme de 0.1959 por día.

Muchas veces la meta trazada es saber si los datos siguen la ley exponencial. La gráfica de log W contra X sirve para formar un juicio inicial acerca de este punto, y puede bastar para liquidarlo. Si así fuere, el uso de papel semilogarítmico evita la necesidad de buscar los logaritmos de W. las líneas horizontales en el papel de gráfica estan trazadas en una escala, de suerte que la gráfica de los datos originales de una recta, siempre que los datos sigan la ley

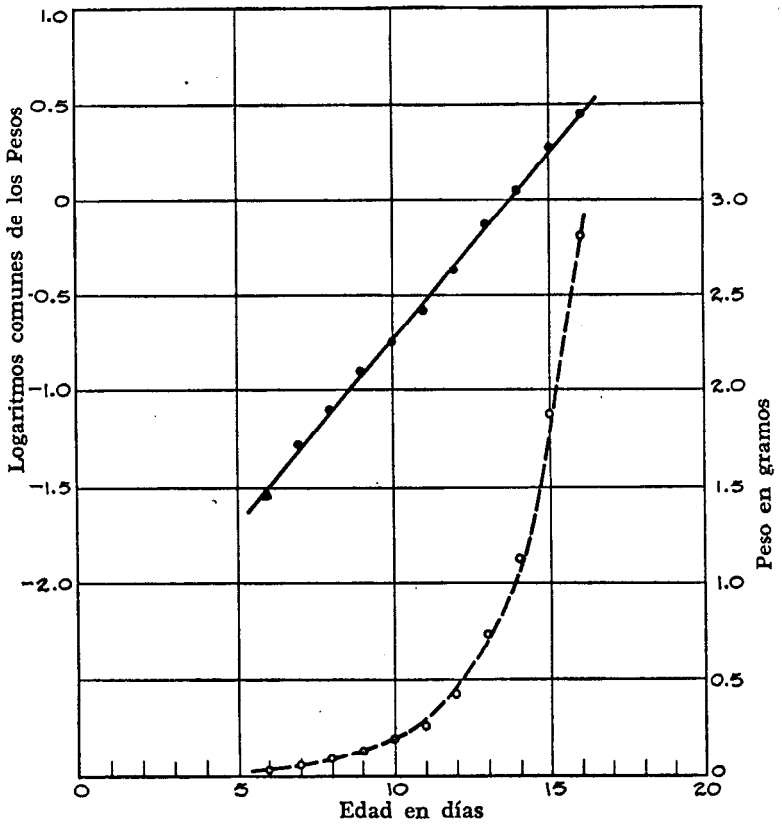


FIG. 15.2.1. Peso seco de embriones de pollo a la edad de 6 a 16 días, con curvas ajustadas. Escala uniforme: $W = 0.002046(1.57)^X$. Escala logarítmica: $Y = 0.1959\ X - 2.689$

exponencial de crecimiento. El papel semilogarítmico se puede adquirir en cualquier comercio de artículos de papelería. Si se necesitase un método mas detallado para la prueba de si la relación entre log W y X es lineal, véase la parte final de la sec. 15.3.

Para aquellas personas que tienen conocimientos de calculo, la ley que dice que el ritmo de aumento a cualquier etapa es proporcional al tamaño ya alcanzado, queda descrito matemáticamente por la ecuación

$\frac{dW}{dX} = cW,$

donde c es el ritmo relativo constante de aumento. Esta ecuación lleva a la relación

$$\log_e W = \log_e A + cX,$$

$$W = Ae^{cX},$$

Donde e = 2.718, es la base del sistema de logaritmos naturales. La relación de la ecuación anterior es exactamente la misma que nuestra relación previa

$$\text{Log}_{10} W = \alpha + \beta X$$

Salvo en que esta expresada en log a base e, en lugar de a base 10.

Como $\log_e W = (\text{Log}_{10} W)(\log_e 10) = 2.3026\ \text{Log}_{10} W$, se desprende que c = 2.3026 β. Para los embriones de pollo, el ritmo relativo de crecimiento es (2.3026)(0.1959) = 0.451 g por día por gramo. Esta claro que la tasa de crecimiento relativo puede calcularse por logaritmos ya sea comunes o naturales.

Para convertir la ecuación $\log W = (0.00205)(1.57)^X$

Donde 0.00205 = antilog (-2.689) = antilog (0.311 – 3) = 2.05/1000 = 0.00205. igualmente, 1.57 = antilog (0.1959). en la forma exponencial

$$W = (0.00205)e^{0.451X}$$

Siendo el exponente 0.451 la tasa relativa.

Las otras relaciones que pueden ser ajustadas por una simple transformación de la W o de la variable X son W = 1/X, W = α + β log X. Que la ley propuesta sea aplicable o no, eso se habrá de examinar gráficamente primero. Si los datos parecen caer en una recta, en la escala revelante transformada, entonces se puede proseguir con los cálculos de regresión. Para la ultima de las relaciones arriba mencionadas hay papel logarítmico disponible, con el rayado tanto vertical como horizontal en escala logarítmica.

La transformación de una relación no lineal para que se convierta en una línea recta, consiste en un método sencillo de ajuste; pero ello comprende algunas

suposiciones que debemos mencionar. Para la curva de crecimiento exponencial estamos suponiendo que la relación de población tiene la forma

$$\mathbf{Y = \log W = \alpha + \beta X + e,}$$

Donde las residuales e son independientes, y tienen medias cero y varianza constante. Ademas, si usamos las pruebas ordinarias de significación en α y β, esto implica la suposición de que las e estan normalmente distribuidas. Algunas veces aparece mas realistico, del conocimiento de la naturaleza del proceso o de las mediciones, suponer que las recidualidades son normales y que tienen varianza constante en la escala original W. Esto significa que estamos postulando una relación de población

$$W = (A)\,(B^X) + d$$

Donde A,B representan ahora los parámetros poblacionales, y las residuales d son N $(0, \sigma^2)$.

Si la ecuación anterior es valida, puede probarse que en la ecuación

$Y = \log W = \alpha + \beta X + e$, las e no serán normales, y que sus varianzas cambian conforme cambia X. Dado el modelo $W = (A)\,(B^X) + d$, el método eficaz para ajustar consisten estimar A y B por la minimizaron de

$$\Sigma(W - AB^X)^2$$

Tomada sobre los valores muéstrales. Esto produce ecuaciones no lineales en A y B, que habran de resolverse por aproximaciones sucesivas. Un método general de ajuste de tales ecuaciones esta dado en la sec. 15.7.

Ejemplo 15.2.1. J.W.C. Price contaron el numero de lesiones de virus de mosaico del Aucuba, que se desarrolla después de exposición a rayos X, de diversos lapsos de tiempo (datos proporcionados por cortesía de los investigadores).

Minutos de exposición	0	3	7.5	15	30	45	60

Conteo en cientos	271	226	209	108	59	29	12

Graficar el conteo como ordenada, luego graficar sus logaritmos. Derivar la regresión Y = 2.432 – 0.02227 X, donde Y es el logaritmo del conteo y X los minutos de exposición.

Ejemplo 15.2.2. repetir los ajustes del ultimo ejemplo utilizando logaritmos naturales. Comprobar que el ritmo de disminución en los centenares de lesiones por minuto por ciento es (2.3026) (0.02227= = 0.05128.

Ejemplo 15.2.3. si el significado de la tasa relativa no esta muy claro, pruebe este método aproximado de cálculos. El aumento en peso de los embriones de pollo durante el treceavo día es 1.130 – 0.738 = 0.392 g, esto es, el ritmo promedio durante este periodo es (1.130 + 0.738)/2 = 0.934 g. El ritmo relativo, o sea el ritmo de aumento de cada gramo es, por lo tanto, 0.392/0.934 = 0.42 g/día/g. Esto difiere del promedio obtenido durante todo el intervalo de 6 a 16 días, 0.451, en parte porque el peso promedio, así como el aumento en peso en el treceavo día sufrieron cierta variación por muestreo, y en parte porque la tasa relativa correcta esta basada en peso, y en aumento en peso, en cualquier instante del tiempo, y no en los promedios de cada día.

15.3. Polinomio de segundo grado. Si nos encontramos frente a una regresión no lineal, a veces no contamos con el conocimiento sobre cual ecuación teórica se habrá de utilizar. En muchos casos el polinomio de segundo grado

$$Y = a + bX + cX^2,$$

Se encontrara satisfactorio para ajustar los datos. La gráfica es una parábola cuyo eje es vertical, pero por lo general aparecen únicamente pequeños segmentos de esa parábola en el proceso de ajuste. En lugar de rectificar los datos, agregamos una tercera variante, el cuadrado de X. Con esto introducimos los metodos de regresión múltiple. Los cálculos prosiguen exactamente siendo X y X^2 las dos

variables independientes. Solamente falta hacer notar que raíz cuadrada de X, log X, o 1/X podrán haberse agregado en lugar de X^2, si los datos así lo hubieran requerido.

Para ilustrar este método y algunas de sus aplicaciones, presentamos los datos sobre rendimiento de trigo, y contenido en proteína de la tabla 15.3.1 y de la figura 15.3.1. el investigador deseaba estimar la significación del contenido de proteína en diversos rendimientos. También probaremos la significación de un alejamiento de la linealidad.

La segunda columna de la tabla consigna los cuadrados de los rendimientos de la columna 1. Los cuadrados se tratan en todos sentidos como una tercera variable en la regresión múltiple. La ecuación de regresión, calculada como de costumbre,

$$Y = 17.703 – 0.3415X + 0.004075X^2,$$

TABLA 15.3.1. PORCENTAJES DE CONTENIDO DE PROTEINA (Y) Y RENDIMIENTO (X) DE TRIGO DE 91 PARCELAS.

Rendimiento, quintales por acre X	Cuadrado X^2	Porcentaje proteína Y	Rendimiento, quintales por acre X	Cuadrado X^2	Porcentaje proteína Y
43	1849	10.7	19	361	13.9
42	1764	10.8	19	361	13.2
39	1521	10.8	19	361	13.8
39	1521	10.2	18	324	10.6
38	1444	10.3	18	324	13.0
38	1444	9.8	18	324	13.4
37	1369	10.1	18	324	13.7
37	1369	10.4	18	324	13.0
36	1296	10.3	17	289	13.4
36	1296	11.0	17	289	13.5
36	1296	12.2	17	289	10.8

35	1225	10.9	17	289	12.5
35	1225	12.1	17	289	12.7
34	1156	10.4	17	289	13.0
34	1156	10.8	17	289	13.8
34	1156	10.9	16	256	14.3
34	1156	12.6	16	256	13.6
33	1089	10.2	16	256	12.3
32	1024	11.8	16	256	13.0
32	1024	10.3	16	256	13.7
Continuación					
32	1024	10.4	15	225	13.3
31	961	12.3	15	225	12.9
31	961	9.6	14	196	14.2
31	961	11.9	14	196	13.2
31	961	11.4	12	144	15.5
30	900	9.8	12	144	13.1
30	900	10.7	12	144	16.3
29	841	10.3	11	121	13.7
28	784	9.8	11	121	18.3
27	729	13.1	11	121	14.7
26	676	11.0	11	121	13.8
26	676	11.0	11	121	14.8
25	625	12.8	10	100	15.6
25	625	11.8	10	100	14.6
24	576	9.9	9	81	14.0
24	576	11.6	9	81	16.2
24	576	11.8	9	81	15.8
24	576	12.3	8	64	15.5
22	484	11.3	8	64	14.2
22	484	10.4	8	64	13.5
22	484	12.6	7	49	13.8

21	441	13.0	7	49	14.2
21	441	14.7	6	36	16.2
21	441	11.5	5	25	16.2
21	441	11.0			
20	400	12.8			
20	400	13.0			

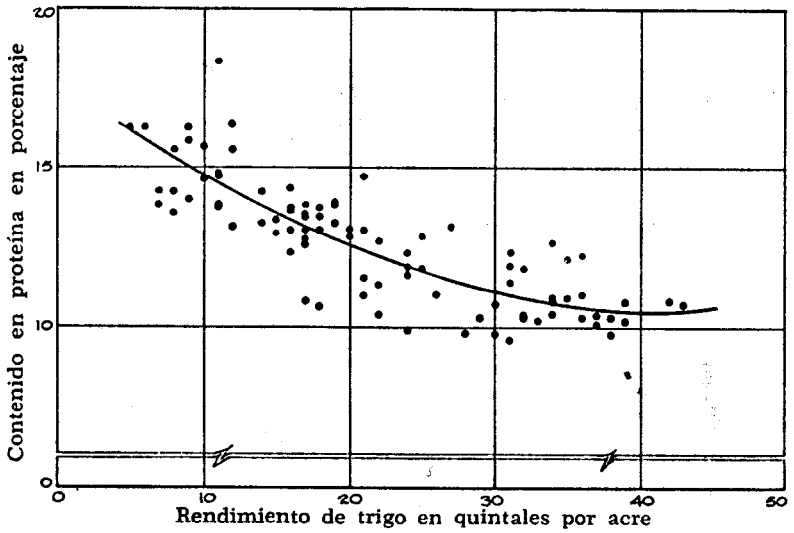


FIG. 15.3.1. Regresión del contenido en proteína sobre rendimiento de trigo, en 91 parcelas. $Y = 17.703 - 0.3415\,X + 0.004075\,X^2$

esta graficada en la figura. A pequeños valores de rendimiento el termino de segundo grado con su pequeño coeficiente apenas es perceptible, y la gráfica se desliza casi como una línea recta. Hacia la derecha, sin embargo, el termino X^2 ha doblegado la curva hasta una dirección casi horizontal.

El análisis de varianza y la prueba de significación estan presentados en la tabla 15.3.2. la regresión ajustada en X y en X^2 da una suma de cuadrados de desviaciones, 97.53, con 88 g.l. la suma de cuadrados de desviaciones de la regresión lineal, sumatoria y^2 – (sumatoria xy)^2 / sumatoria X^2, es 110.48, con 89 g.l. la reducción en suma de cuadrados, probada contra la media cuadrada restante después de la regresión curvilínea, prueba ser significativa. La hipótesis de regresión lineal se abandona; existe una significativa curvilinealidad en la regresión.

En la tabla 15.4.1, muchos de los valores de X (esto es, X = 39) tienen dos o más valores de Y. Con esos datos, la suma de cuadrados de desviaciones de la regresión curva (88 g.l.) puede dividirse en dos partes para proporcionar una prueba mas critica del ajuste de la cuadratica. La tecnica se describe en la siguiente sección. En el ejemplo considerado, esta tecnica apoya el ajuste de la cuadratica.

La ecuación de regresión es útil también para estimar e interpolar.

Como siempre ocurre en regresión, ya sea lineal o curva, habremos de ser cautelosos con la extrapolación. Los datos pueden resultar incompetentes para proporcionar evidencia de tendencia mas allá de su propia amplitud. Observando la fig. 15.2.1 podríamos sentirnos tentados por el excelente ajuste a suponer la misma tasa de crecimiento antes del sexto día y después del décimo sexto. Sin embargo, el hecho es que hubo bruscos cambios en ritmo de crecimiento en ambos de estos días. Para que sea de utilidad la extrapolación se requiere un extenso conocimiento así como un pensamiento muy penetrante.

EJEMPLO 15.3.1. la prueba de significación de alejamiento de la regresión lineal de la tabla 15.3.2 también puede utilizarse para examinar si una transformación de rectificación del tipo que presentemos en la sección 15.2 ha producido una relación de línea recta. Aplique esta prueba a los datos sobre embriones de pollo de la tabla 15.2.1, ajustando una parábola en X a los logaritmos de los pesos Y. Comprobar que la parábola es

$$Y = -2.783162 + 0.214503X – 0.000846\ X^2$$

Y que la prueba de resultados como se muestra a continuación:

	Grados de libertad	Suma de cuadrados	Media cuadrada

Desviaciones de la regresión lineal	9	0.007094	
Desviaciones de la regresión cuadratica	8	0.006480	0.000810
Curvilinealidad de la regresión	1	0.000614	0.000614

F = 0.76, 1 y 8 g.l. cuando las X estan igualmente espaciadas, como en este ejemplo, en la sec. 15.6 damos una forma mas rápida para calcular la prueba.

15.4. datos que tienen varias Y para cada valor de X. Si se han medido varios valores de Y para cada valor de X, lo adecuado del polinomio ajustado puede probarse mas concienzudamente. Supongamos que para cada X hay disponible un grupo de n valores de Y. Para ilustrar esto a un modelo lineal, si Y_{ij} representa al j-esimo miembro del i-esimo grupo, el modelo lineal es

$$Y_{ij} = \alpha + \beta Xi + Eij,$$

Donde Eij se ajusta a $N(0,\sigma^2)$. Se infiere que las medias del grupo Yi. estan relacionadas a las Xi por la relación lineal

$$Yi. = \alpha + \beta Xi + Ei.$$

(1) Ajustando una regresión cuadratica de las Yi. sobre las Xi, la prueba por curvatura de la tabla 15.3.2 puede utilizarse como hicimos anteriormente. Como es importante en lo que diremos mas adelante, nótese que las residuales Ei. Tienen σ^2/n, puesto que cada Ei. Es la media de n residuales independientes de la relación 15.4.1.

(2) El nuevo aspecto es que las desviaciones de Yij de sus medias de grupo Yi. proporcionan una estimada independiente de σ^2 . la estimada global es

$$S^2 = \sum_{i=1}^{K} (Yij - yi.)^2 / k(n-1)$$

Con K (n-1) g.l. si multiplicamos las medias cuadradas del análisis por n, para hacer las partes (1) (2) comparables, tendremos el análisis de varianza de la Tabla 15.4.1.

TABLA. 15.4.1. ANALISIS DE VARIANZA PARA PRUEBAS DE REGRESION LINEAL

Fuente de variación	Grados de libertad	Media cuadrada
Regresión lineal de Yi en Xi	1	S_1^2
Regresión cuadratica de Y en Xi	1	S_2^2
Desviaciones de Y de la cuadratica	k-3	S_d^2
global dentro de grupos	K(n-1)	S^2
Total	Kn -1	

Los siguientes resultados son básicos en la interpretación de esta tabla. Si la regresión poblacional es lineal, la media cuadrada S_2^2 es una estimada no perjuiciada de σ^2; si la regresión de población es curva, S_2^2 tiende a volverse grande. Si la regresión de población de es o lineal o cuadratica, S_d^2 es estimada no perjudiciada de σ^2. ¿ Cuándo tiende S_d^2 a volverse mucho mas grande que σ^2? O bien, si la regresión de población es no lineal, pero que no este adecuadamente representada por una cuadratica; por ejemplo, podría ser una curva de tercer grado, o una con algún aspecto político: o si hay fuentes de variación que sean constantes dentro de cualquier grupo, pero que varían de grupo a grupo. Esto podría ocurrir si las mediciones en diferentes grupos se tomaran en tiempos desiguales, o de hospitales distintos o de arbustos diferentes. La varianza S^2 global dentro de grupos es estimada no prejuiciada de σ^2 no importa la forma de la relación entre Yi. y Xi.

En consecuencia, calcular primero la razón F, S_d^2 / S^2, con (k –3) y k (n-1) g.l. si esto es significativo, hay que observar la gráfica de Y contra X para ver si un polinomio de grado mas alto o un tipo diferente de relación matemática esta indicada. Es de gran utilidad el examen de las desviaciones de Yi. de la cuadratica ajustada, buscando también huellas de una tendencia sistemática. Si no hay

tendencia sistemática, la explicación mas indicada es que alguna fuente extra de variación entre grupos se ha colocado a los datos.

TABLA 15.4.2. DOSIS LETAL (MENOS 50 UNIDADES) DE UABAINA ESTANDAR DE ESTADOS UNIDOS, POR INYECCION INTRA VENOSA LENTA EN GATOS, HASTA ALTO TOTAL DEL CORAZON

Xi = velocidad de inyección en (mg/kg/min)/1045.75					Total
	1	2	4	8	
	5	3	34	51	
	9	6	34	56	
	11	22	38	62	
	13	27	40	63	
	14	27	46	70	
	16	28	58	73	
	17	28	60	76	
	20	37	60	89	
	22	40	65	92	
	28	42			
	31	50			
	31				
Σ Yij = yi.	217	310	435	632	1594
ni	12	11	9	9	41
Yi.	18.1	28.2	48.3	70.2	
ΣYij^2	4727	10788	22261	45940	83716

Si S_d^2 / S^2 claramente no es significativa, formase la media cuadrada global de S_d^2 y S^2. Designe esto con (kn – 3) g.l. luego pruebe la $F = S^2 / S_p^2$ con 1 y (km – 3) g.l. como prueba de curvatura de la relación.

El procedimiento esta ilustrado por los datos de la tabla 15.4.2, que fue proporcionada por la cortesía de B.J. Vos y W.T. Dawson. El punto de interés es si hay relación lineal entre la dosis letal de uabaina, inyectada a gatos, y el ritmo de inyección. Se utilizaron cuatro ritmos, cada uno siendo el doble del anterior.

Primero, la suma total de cuadrados de las dosis letales 21 744 se analiza "entre ritmos" 16093, y "dentro de grupos de ritmos" 5651. Nótese que el numero de gatos n_i difiere ligeramente de grupo a grupo.

La desigualdad de n_i habrá de tomarse en cuenta al elaborar ecuaciones para la regresión de Yi. sobre X_i y X_i^2. Calcular:

$$\Sigma n_i X_i = 12(1) + 11(2) + 9(4) + 9(8) = 142$$
$$\Sigma n_i X_i^2 = 12(1) + 11(4) + 9(16) + 9(64) = 776$$
$$\Sigma n_i X_i^3 = 12(1) + 11(8) + 9(64) + 9(512) = 5284$$

y en la misma forma $\Sigma n_i X_i^4 = 39356$. También necesitamos

$$\Sigma n_i X_i Y_{i.} = \Sigma X_i Y_{i.} = 1(217) + 2(310) + 4(435) + 8(632) = 7633$$
y $\Sigma X^2 Y_{i.} = 48865$.

Cada cantidad es corregida por la media en la forma acostumbrada. Por ejemplo,

$$\Sigma n_i (Xi^2 - X^2)^2 = \Sigma n_i X_i^4 - (\Sigma n_i X_i^2)^2 / \Sigma n_i = 3956 - (776)^2/41$$
$$= 24668.8$$
$$\Sigma n_i (X_i - X)(Yi. - Y..) = \Sigma X_i Y_i - (\Sigma n_i X_i)(\Sigma Y_{i.})/\Sigma n_i$$
$$= 7633 - (142)(1594)/41 = 2112.3$$

para completar las cantidades requeridas para las ecuaciones normales, podemos comprobar que

$$\Sigma n_i (X_i - X)^2 = 284.2, \Sigma n_i (X_i - X)(Xi^2 - X^2)^2 = 2596.4,$$
$$\Sigma n_i (Xi^2 - X^2)^2 (Yi. - Y.) = 18695.6$$

las ecuaciones normales para b_1 y b_2 son:

$$284.2\, b_1 + 2596.4\, b_2 = 2112.3$$

$2596.4\ b_1 + 24668.8\ b_2 = 18695.6$

En la forma acostumbrada, la reducción de la suma de cuadrados de Y debido a la regresión sobre b_1 y b_2 se encuentra que es 16082, en tanto que para la regresión lineal la reducción es 15700. El análisis final de la varianza aparece en la tabla 15.4.3.

La media cuadrada 11 para las desviaciones de la cuadratica es mucho mas baja que la media cuadrada dentro de grupos, aunque no tan extremadamente para tan solo 1 g.l. el promedio global de estas dos medias cuadradas es 149, con 38 g.l. para la prueba de curvatura, F = 382/149 = 2.56, con 1 y 38 g.l. que yace entre el nivel de 25% y el de 10%. Llegamos a la conclusión de que los resultados son consistentes en una relación lineal en la población.

TABLA 15.4.3 PRUEBA DE DESVIACIONES DE REGRESIONES LINEAL Y CUADRATICA

Fuente de variación	Grados de libertad	Suma de cuadrados	Media cuadrada
Regresión lineal sobre X	1	15700	15700
Regresión cuadratica sobre X	1	382	382
Desviaciones de la cuadratica	1	11	11
Global dentro de grupos	37	5651	153
total	40	21744	

Ejemplo 15.4.1 los siguientes datos tomados de Swanson y Smith (4) para servirnos de ejemplo, con n iguales, muestra el total de contenido de nitrógeno Y (gramos por 100 cc de plasma) de plasma sanguínea de ratas a nueve edades X (en días).

Edad de la rata	25	37	50	60	80	100	130	180	360
	0.83	0.98	1.07	1.09	0.97	1.14	1.22	1.20	1.16

	0.77	0.84	1.01	1.03	1.08	1.04	1.07	1.19	1.29
	0.88	0.99	1.06	1.06	1.16	1.00	1.09	1.33	1.25
	0.94	0.87	0.96	1.08	1.11	1.08	1.15	1.21	1.43
	0.89	0.90	0.88	0.94	1.03	0.89	1.14	1.20	1.20
	0.83	0.82	1.01	1.01	1.17	1.03	1.19	1.07	1.06
	5.14	5.40	5.99	6.21	6.52	6.18	6.86	7.20	7.39

Una gráfica de los totales Y contra X muestra que (i) los valores de Y para X = 100 son anormalmente bajos y requieren una investigación especial, (ii) la relación es claramente curva. Omítanse los datos para X = 100 y pruébense las desviaciones de una regresión parabólica, contra la media cuadrada de Dentro de grupos. Resp. F = 1.4.

15.5. Prueba de alejamiento de la regresión lineal en el análisis de covarianza. Igual que en trabajos de correlación y regresión, es necesario en covarianza estar seguros de que la regresión sea lineal Se recordará que en los diseños de tipo estándar, clasificaciones de un sentido, de dos sentidos (blocks aleatorios) y cuadros Latinos, la regresión de Y sobre X se calcula del renglón de Residuales o de Error del análisis de varianza. Un método gráfico para comprobar la linealidad, que muchas veces resulta suficiente, consiste en graficar los residuales de Y del modelo del análisis de varianza contra las correspondientes residuales de X, buscando siempre señales de curvatura.

El método numérico de comprobación consiste en sumar un término en X^2 al modelo. Anotando $X_1 = X$, $X_2 = X^2$ se desarrollan las sumas de cuadrados de residuales o de error Y, X_1 y X_2 de las sumas de errores de productos de X, X_2, YX_1, y YX_2 como se mostró en la Sec. 14.8 para clasificaciones de un sentido. De estos datos, calcular la prueba de significación de alejamiento de la regresión lineal como en la Tabla 15.3.2.

Si se encuentra que la regresión es curva, las medias de tratamiento se ajustan para la regresión parabólica. Los cálculos se apegan al método presentado en la Sec. 14.8.

15.6. Polinomios ortogonales. Si los valores de X están igualmente espaciados, el ajuste del polinomio

$$Y= b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \ldots$$

se acelera, utilizando tablas de polinomios ortogonales. El paso esencial consiste en sustituir X^1 (i = 1, 2, 3...) por un polinomio de grado i en X, que vamos a designar Xi. Los coeficientes en estos polinomios se seleccionan de modo que '

$$\Sigma x_i = 0 : \Sigma x_i x_j = 0 \qquad (i \neq j)$$

donde las sumas son sobre todos los n valores de X en la muestra. Los diferentes polinomios son ortogonales entre sí. En esta sección, damos fórmulas explícitas para estos polinomios.

En lugar de calcular la regresión polinomia de Y sobre X, en la forma anterior, la calculamos en la forma

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 +$$

que puede demostrarse que da el mismo polinomio ajustado. Debido a la ortorgonalidad de las Xi tenemos los resultados

$$B_0 = Y : B_i = \Sigma X_i Y / X_i^2 \qquad (i = 1, 2, 3...)$$

Los valores de Xi y de ΣX_i^2 se encuentran en las tablas, lo que hace más sencillo los cálculos de Bi. Además, las reducciones en $\Sigma (Y - \acute{Y})^2$ debido a términos sucesivos en el polinomio están dados por

$(\Sigma X_1 Y)^2 (\Sigma X_1^2)$; $(\Sigma X_2 Y)^2 / (\Sigma X_2^2)$; $(\Sigma X_3 Y)^2 / (\Sigma X_3^2)$; así sucesivamente. Entonces resulta fácil comprobar que la adición de una potencia más alta de X al polinomio produce una marcada reducción en la suma de cuadrados de residuales. Para ahorrar tiempo, los polinomios ortogonales son más efectivos cuando los cálculos se hacen en una calculadora de escritorio. Con calculadora electrónica, los programas de rutina para el ajuste de regresión múltiple pueden utilizarse para ajustar la ecuación en su forma original. La mayoría de los programas también proporcionan reducciones en suma de cuadrados debido a cada potencia sucesiva.

Como ilustración, un polinomio se va a ajustar a los datos sobre embriones de pollo, aunque como lo vimos en la sección 15.2, estos datos son mas ajustables a una curva de crecimiento exponencial.

La tabla 15.6.1. muestra los pesos (Y) y los valores de X_1, X_2, X_3 , X_4 , X_5 para n =11, tomados en la tabla A 17. Para ahorrar espacio, la mayoría de las tablas dan los valores Xi sólo para la mitad superior de los valores de X. En nuestro ejemplo éstos son los valores de X = 11 a X = 16. El método de anotar las Xi para la mitad inferior de la muestra se puede ver en la Tabla 15.6. l. Para los términos de grado impar, X_1 X_3 y X_5 los signos se cambian en la mitad inferior.
; para términos de grado par, X_2 y X_4 los signos siguen igual.

TABLA 15.6.1. AJUSTE DE UN POLINOMIO DE CUARTO GRADO AL PESO DE EMBRIONES DE POLLO.

Edad X	Peso seco y	X_1	X_2	X_3	X_4	X_5	Y_4
(días)	(g)						
6	0.029	-5	15	-30	6	-3	0.026

7	0.052	-4	6	6	-6	6	0.056
8	0.079	-3	-1	22	-6	1	0.086
9	0.125	-2	-6	23	-1	-4	0.119
10	0..181	-1	-9	14	4	-4	0.171
11	0.261	0	-10	0	6	0	0.265
12	0.425	1	-9	-14	4	4	0.434
13	0.738	2	-6	-23	-1	4	0.718
14	1.130	3	-1	-22	-6	-1	1.169
15	1.882	4	6	-6	-6	-6	1.847
16	2.812	5	15	30	6	3	2.822
$\sum X^2_i$		110	858	4290	286	156	
λ_i		1	1	15/6	1/12	1/40	
$\sum X_i Y$	7.714	25.858	39.768	31.873	1.315	-0.254	
B_i	0.701273	0.235073	0.046349	0.007430	0.004598		

Vamos a suponer que la meta es encontrar el polinomio de grado más bajo que parezca adecuado para ajustar. En consecuencia, la reducción en suma de cuadrados se comprobará para cada término sucesivo que se añade. A cada etapa, calcular

$$\Sigma x_i\, Y,\ Bi = \Sigma Xi\, Y/\Sigma X_i^{\,2}$$

(se muestra al final de la Tabla 15.6.1) y la reducción en la suma de cuadrados $(\Sigma X_i\ Y)^2/\Sigma X_i^{\,2}$, anotado en la Tabla 15.6.2. Para el término lineal, el valor de F es (6.078511)1(0.232177) = 26.2. Los valores posteriores de F, para los términos cuadráticos y cúbicos, son aún más grandes, 59.9 y 173.4. Para X_4 (cuártica) F es 10.3, significativo a nivel de 5%, pero no a nivel de 1%. Sin embargo, el término de quinto grado tiene una F menor que 1. Como medida de precaución, también debemos comprobar el término de sexto grado, pero para este ejemplo, aquí nos vamos a detener y concluimos que un polinomio de cuarto grado es ajuste satisfactorio.

TABLA 15.6.2. REDUCCIONES EN LA SUMA DE CUADRADOS DEBIDO A TERMINOS SUCESIVOS

Fuente	Grados de libertad	Suma de cuadrados	Media cuadrada	F
Total: Σ (Y – Y)	10	8.168108	0.232177	26.2
Reducción a lineal	1	3.078511		
Desviaciones de lineal	9	2.089597		
Reducción a cuadratica	1	1.843233	0.030796	59.9
Desviaciones a cuadratica	8	0.246364		
Reducción a cubica	1	0.236803	0.001366	173.4
Desviaciones de cubica	7	0.009561		
Reducción a cuadratica	1	0.006046	0.000586	10.3
Desviaciones de cuadratica	6	0.003515		
Reducción a quintica	1	0.000414	0.000620	0.7
Desviaciones de quintica	5	0.003101		

Para graficar el polinomio, los valores estimados de Y, para cada valor de X, son calculados fácilmente de la Tabla 15.6.1:

$$\mathbf{Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4}$$

Nótese que B_0 = Y= 0.701273. At X = 6,

$$Y = 0.701273 - 5(0.235073) + 15(0.046349) - 30(0.007430)$$
$$+ 6(0.004598) = 0.026,$$

y así sucesivamente. La Fíg. 15.6.1 muestra el ajuste por una recta, obviamente malo, el polinomio de segundo grado considerablemente mejor y el polinomio de cuarto grado.

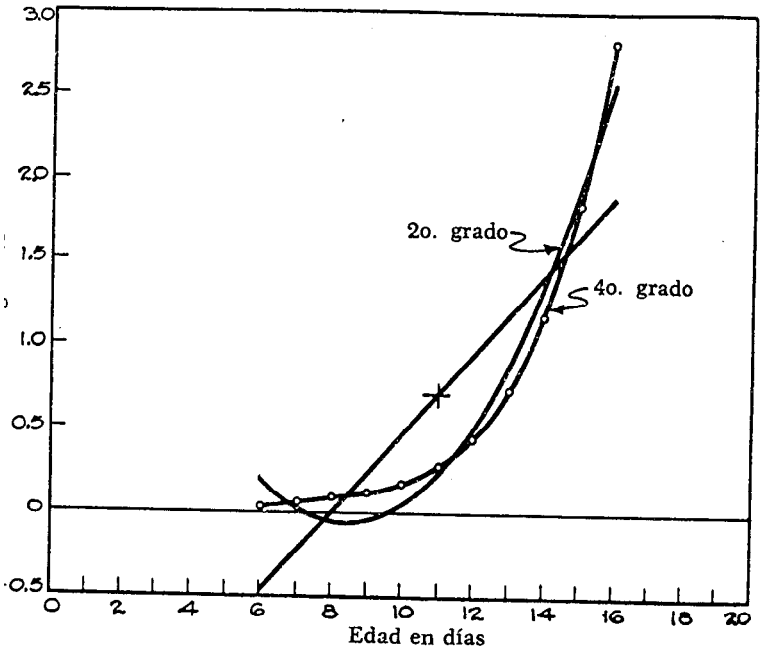


Fig. 15.6.1 gráfica de polinomios de primero, segundo y cuarto grados ajustados a los datos sobre embriones de pollo a que se refiere la Tabla 15.6.1

Para expresar el polinomio como una ecuación en las variables originales X, es mucho más tedioso. Para esto, necesitamos fórmulas que den Xi en términos de X y sus potencias. En el método estándar desarrollado por Fisher, por el cual fueron calculadas las tablas de polinomios, él comenzó con un conjunto ligeramente diferente de polinomios ξ_i , que satisfacían las relaciones recurrentes

$$\xi_0 = 1 \quad : \quad \xi_1 = X - X \text{ media} \quad : \xi_{i+1} = \xi_1 \xi_i - i^2(n^2 - i^2)/\, 4(4_i^2 - 1) \; * \; \xi_{i-1}$$

Estos polinomios son ortogonales, pero cuando sus valores son tabulados para cada miembro de la muestra, estos valores no siempre son números enteros. En consecuencia, Fisher encontró por inspección el multiplicador λ_i que hace $X_i = \lambda_i \xi_i$ el conjunto más pequeño de enteros. Esto facilita los cálculos para la persona que

los usa. Los valores de las λ_i se muestran en la Tabla 15.6.1, y bajo cada polinomio de la Tabla A 17, así corno también en las Reís. (5 y 6).

Ahora veamos los cálculos del ejemplo. El primer paso es multiplicar cada B_i por la correspondiente λ_i. Esto da

$$\mathbf{B_i' = 0.235073;\ B_2' = 0.046349;\ B_2' = 0.0061921;\ B_4' = 0.0003832}$$

Estos son los coeficientes para la regresión de Y sobre las ξ_i de modo que

$$\hat{Y} = Y \text{ media} + B_{1'}\, \xi_1 + B_2'\xi_2 + B_3\, \xi_3 + B_4\{\, \xi_4 \qquad (15.6.1)$$

Las ecuaciones generales que conectan las ξ_i, con X son como sigue:

$\xi_i = X - X \text{ media} = x$

$\xi_2 = x^2 - n^2 - 1/12$

$\xi_3 = x^3 - 3n^2 - 7/20\ \ (x)$

$\xi_4 = x^4 - (3n^2 - 13)/14\ \ (x^2) + 3\,(n^2 - 1)\,(n^2 - 9)$

$\xi_5 = x^5 - 5(n^2 - 7)/18\ \ (x^3) + (15\, n^4 - 230n^2 + 407)/1008\ \ (x)$

Por sustitución en la fórmula (15.6.1). la $\hat{Y}$ queda expresada como un polinomio en $x = X - \bar{X}$. Si es conveniente detenerse en esta etapa, hay dos ventajas. Se ahorran más cálculos, y hay menor pérdida de precisión decimal. Sin embargo, para completar el ejemplo, notamos que n = 11 y $\bar{X} = 11$. Entonces, en términos de X,

$\xi_1 = X - 11$

$\xi_2 = (X - 11)^2 - 10 = X^2 - 22X + 111$

$\xi_3 = (X - 11)^3 - 17.8\,(X-11) = X^3 - 33\,X^2 + 345.2X - 1135.2$

$\xi_4 = (X - 11)^4 - 25\,(X-11)^2 + 72$

$= X^4 - 44X^3 + 701X^2 - 4774X + 11688$

De donde, finalmente, utilizando la fórmula (15.6.1),

$\hat{Y} = 0.701273 + 0.235073\,\xi_1 + 0.046349\,\xi_2 + 0.006192\,\xi_3 + 0.0003832\,\xi_4$

$= 0.701273 + 0.235073(X - 11) + 0.046349(X^2 - 22X + 111)$

$+ 0.006192(X^3 - 33X^2 + 345.2X - 1,135.2)$

$+ 0.0003832(X^4 - 44X^3 + 701\,X^2 - 4,774X + 11,688)$

$= 0.7099 - 0.47652X + 0.110636X^2 - 0.010669X^3 + 0.0003832X^4$

En la tabla 15.6.1 todavía hay una manera de abreviar que no utilizamos. Al calcular ΣX_1 ,Y, las Y, a los dos extremos de la muestra, digamos Y_n y Y_1, están multiplicadas por 5 y por -5. Y_{n-1} y Y_2 están multiplicadas por 4 y - 4. Si establecemos las diferencias $Y_n - Y_1$, $Y_{n-1} - Y_2$,y así en forma sucesiva, únicamente el conjunto de multiplicadores 5, 4, 3, 2, 1 necesitan usarse. Esta técnica surte efecto para, cualquier ΣX_i Y, en que i sea impar. Con i par, formamos las sumas $Y_n + Y_1$, $Y_{n-1} + Y_2$, etc. El método está elaborado para estos datos en el Ej. 15.6.1.

Ejemplo 15.6.1. En la Tabla 15.6.1 fórmense las sumas y las diferencias de pares de valores de Y, operando desde fuera. Compruébese que estos valores dan los resultados que se muestran posteriormente y que los valores de ΣX_i Y están en acuerdo con los dados en la Tabla 15.6.1.

Sumas	X_2	X_4	Diferencias	X_1	X_3
0.261	-10	6	0.26	0	0
0.606	-9	4	0.224	1	-14
0.863	-6	-1	0.613	2	-23
1.209	-1	-6	1.051	3	-22
1.934	6	-6	1.830	4	-6
2.841	15	6	2.783	5	30

EJEMPLO 15.6.2. Aquí estan seis puntos de la cubica $y = 9x - 6x^2 + x^3$. (0,0), (1,4), (2,2), (3,0), (4,4), (5,20). Llévense a cabo los cálculos para regresión lineal, cuadratica y cubica. Compruébese que no hay suma residual de cuadrados, después de ajustar la cubica, y que los valores del polinomio, en esa etapa, son exactamente las Y.

Ejemplo 15.6.3. el método para construir polinomios ortogonales puede ilustrarse encontrando X_1 y X_2 cuando n = 6.

(1)	(2)	(3)	(4)	(5)
X	$\xi_1 = X - \bar{X}$	$X_1 = 2\xi_1$	ξ_2	$X_2 = 2/3\ \xi_2$
1	-5/2	-5	10/3	5
2	-3/2	-3	-2/3	-1
3	-1/2	-1	-8/3	-4

Continuación

4	½	1	-8/3	-4
5	3/2	3	-2/3	-1

6	5/2	5	10/3	5

Comenzar con X = 1, 2, 3, 4, 5, 6 con $\bar{X}$ = 7/2. Compruebe que los valores de ξ_1 = = x = X - X son corno se muestra en la columna (2). Corno las ξ_1 no son números enteros, tomamos $\square_1$ = 2, lo que da $X_1 = 2\xi_1$, columna (3). Para encontrar ξ_2 anotamos

$$\xi_2 = \xi_1^2 - b\,\xi_1 - c$$

Esta es una cuadrática en X. Querernos $\Sigma\xi_2$ = o. Esto da

$$\Sigma\xi_1^2 - b\Sigma\xi_1 - nc = 0 \quad \text{: i.e., } 35/2 - 6c = 0 \quad : \quad c = 35/12$$

Además, deseamos $\Sigma\xi_1\xi_2 = 0$, lo que da

$$\Sigma\xi_1^3 - b\,\Sigma\xi_1^2 - c\Sigma\xi_1 = 0 \text{ : ¡.e., } b\,\Sigma\xi_1^2 = 0 \text{ : } b = 0$$

Entonces, $\xi_2 = \xi_1^2$ - 35/12. Compruebe los valores de ξ_2, en la columna (4)
Para convertir éstos a enteros multiplique por $\square_2$ = 3/2.

15.7. Método general para ajustar regresiones no lineales. Suponiendo que la relación de población entre Y y X tiene la forma

$$Y_i = f(\alpha, \beta, \gamma, X_i) + ei \qquad (i = 1, 2, \ldots n)$$

donde f es una función de regresión que contiene X_i, y los parámetros α, β, γ. (Puede haber más de una variable X.) Si las residuales e_i tienen medias cero y varianza constante, el método de cuadrados mininos para ajuste de regresión consiste en estimar los valores de α, β, γ, por minimizacíón de

$$\Sigma(Y_i - f(\alpha, \beta, \gamma, X_i))$$

Esta sección presenta un método general de llevar a cabo los cálculos. Los detalles requieren cierto conocimiento de diferenciales parciales, pero el ataque no deja de ser sencillo.

La dificultad surge, no por la no linealidad en X_i, sino por la no linealídad en una o más de los parámetros α β γ . La parábola (α, βX, γX^2) se ajusta por los métodos ordinarios de regresión múltiple lineal, porque es lineal en α, β, γ. Considere la regresión asintótica α, - $\beta(\gamma^x)$. Si el valor de γ fuera conocido de antemano, podríamos anotar $X_1 = \gamma^x$. Las estimadas de cuadrados mínimos de α y β estarían entonces dados por ajuste de una regresión lineal ordinaria de Y sobre X_1. Sin embargo, cuando γ tiene que ser estimada de los datos, los métodos de regresión lineal no se pueden aplicar.

El primer paso del método general es obtener buenas estimadas iniciales para a_1, b_1 c_1 de las estimadas finales de mínimas cuadradas α, β, γ.ˆ Para los tipos comunes de funciones no lineales, se han desarrollado varias técnicas para hacer esto, algunas veces gráficas, otras por estudios especiales del problema. Luego utilizamos el teorema de Taylor. Este enuncia que si la f (α, β, γ, X) es continua en α, β, y γ, y si (α - a_1) (β - b_1) (γ -c_1), son pequeñas,

$$f(\alpha, \beta, \gamma, X_i) = f(a_1, b_1\ c_1\ X_i) + (\alpha - a)\, f_a + (\beta - b_1)\, f_b(\gamma - c_1)\, f_c$$

El símbolo □ significa que "es aproximadamente igual a ”Los símbolos f_a f_b, f_c significan las derivadas parciales de f con' respecto a α, β, y γ, y, respectivamente, evaluadas en los puntos a_1, b_1 c_1. Por ejemplo, en la regresión asintótica

$$f(\alpha, \beta, \gamma, X_i) = \alpha - \beta(\gamma^{xi})$$

tenemos

$$f_a = 1;\ f_b = -c_1^{xi};\ f_c = -b_1 X_i (c_1^{xi-1})$$

Como a_1, b_1 y c_1, son conocidas, los valores, f_a ,f_b ,f_c se pueden calcular para cada miembro de la muestra, donde hemos anotado f para f (a_1, b_1 c_1 X_i)Por el teorema de Taylor, la relación original de regresión

$Y_i = f(\alpha, \beta, \gamma, X_i)\, e_i$

puede entonces escribirse aproximadamente,

$$Y_i \approx f + (\alpha - a_1) f_a + (\beta - b_1) f_b + (\gamma - c_1) f_c + e_i \qquad (15.7.1)$$

Ahora escribimos

$$Y_{res} = Y - f;\ X_1 = f_a;\quad X_2 = f_b;\quad X_3 = f_c$$

De la Ec. 15.7.1

$$Y_{res} \approx (\alpha - a_1) X_1 + (\beta - b_1) X_2 + (\gamma - c_1) X_3 + e_i \qquad (15.7.2)$$

La variante Y_{res} es la residual de Y de la primera aproximación. La relación (15.7.2) representa una regresión lineal ordinaria de Y_{res} sobre las variantes X_1, X_2, X_3, siendo los coeficientes de regresión (α - a_1), (β - b_1) y (γ - c_1). Si la relación (15.7.2) fuera válida con exactitud y no en forma aproximada, los cálculos de la regresión de muestra de Y_{res} sobre X_1, X_2, X_3, darían los coeficientes de regresión (α - a_1), (β - b_1) y (γ - c_1)de donde las estimadas correctas de mínimos cuadrados α ,β, y γ se hubieran obtenido de inmediato.

Como la relación (15.7.2) es aproximada, el ajuste de esta regresión da segundas aproximaciones a_2, b_2 y c_2, para α ,β, y γ respectivamente. Recalculamos luego, f_a ,f_b ,f_c para este punto a_2, b_2 y c_2,, encontrando una nueva Y_{res}, y nuevas variantes X_1, X_2, X_3. La regresión de muestra de esta Y_{res}, sobre X_1, X_2, X_3 da los coeficientes de regresión ($a_3 - a_2$), ($b_3 - b_2$) y ($c_3 - c_2$) de donde

se encuentran las terceras aproximaciones a_3, b_3, c_3 para α ,β, y γ y así sucesivamente.

Si el proceso es efectivo, la suma de cuadrados de las residuales ΣY_{res}^2 habrá de decrecer continuamente a cada etapa, y las disminuciones se hacen más pequeñas, conforme la solución de mínimos cuadrados se va alcanzando. En la práctica, los cálculos se suspenden cuando la disminución en ΣY_{res}^2 los cambios en a, b y c se consideran pequeños, lo suficiente como para despreciarlos. La residual de medias cuadradas es

$$S^2\ \Sigma Y_{res}^2/(n-k),$$

donde k es el número de parámetros que han sido estimados (en nuestro ejemplo, k = 3). Con regresiones no lineales, s^2 no es una estimada no prejuiciada de σ^2 , aunque tienda a hacerse no prejuiciada conforme la n se hace grande.

Los errores aproximados estándar de las estimadas α ,β, y γ se obtienen en la forma acostumbrada, de los multiplicadores de Gauss en la regresión múltiple final que fue calculada. Así,

$$\text{s.e. } (\alpha) \square\ \sqrt{c_{11}};\quad \text{s.e. } (\beta) \square^{S}\sqrt{c_{22}}\quad \text{s.e}(\gamma) \square^{S}\sqrt{c_{33}}$$

Los límites de confianza aproximados para α están dados por (α+-√ts donde t tiene (n - 3) g.l.

Si se hacen necesarias varias etapas en la aproximación, los cálculos se vuelven tediosos usando una máquina de escritorio, ya que la regresión múltiple ha de trabajarse para cada etapa. Con las relaciones no lineales más comunes, sin embargo, los cálculos se prestan fácilmente para una programación en

máquina electrónica. Los investigadores que tienen acceso a un centro de computaciones se les recomienda investigar si está disponible un programa, o si es factible construir uno. Sí hay que hacer el trabajo en una calculadora de escritorio, es obvia la importancia de una primera aproximación que sea buena.

15.8. Ajuste de una regresión asintótica. La función de regresión de población se anotará (utilizando el símbolo ρ en lugar de γ)

$$f(\alpha, \beta, \rho, X) = \alpha + \beta(\rho^X) \qquad (15.8.1)$$

Sí $0 < \rho < 1$ y β es negativa, esta curva tiene la forma que se muestra en el grabado 15.1.1 (c), Pág. 546, subiendo del valor $(\alpha + \beta)$ a X = 0, al asíntota α, conforme X se hace grande. Si $0 < \rho < 1$ y β es positiva, la curva declina del valor $(\alpha + \beta)$ a X = 0, al asíntota α cuando X es grande.

Como la función no es lineal, únicamente en lo que respecta al parámetro ρ**,** se simplifica algo el método de aproximaciones sucesivas descrito en la sección anterior. Deje que r_1 sea una primera aproximación para ρ. Por el teorema de Taylor

$$\alpha + \beta(\rho^X) \approx \alpha + \beta(r_1^X) + \beta(\rho - r_1)(Xr_1^{X-1})$$

Anótese $X_0 = 1$, $X_1 = r_1^X$, $X_2 = Xr_1^{X-1}$. Sí ajustamos la regresión de muestra.

$$\hat{Y} = aX_0 + bX_1 + cX_2 \qquad (15.8.2)$$

se infiere que a, b son segundas aproximaciones de las estimadas de cuadrados mínimos, α , β de $\alpha + \beta$ en (15.8.1), en tanto que

$$c = b(r_2 - r_1),$$

de suerte que

$$r_2 = r_1 + c/b$$

es la segunda aproximación para ρ.

El caso más común es aquel en que los valores de X cambian por la unidad (v.g., X = 0, 1, 2... o X = 5, 6, 7 ...) o se le puede codificar para que responda así. Designamos los correspondientes valores de Y por $Y_0, Y_1, Y_2, \ldots, Y_{n-1}$.

Nótese que el valor de X correspondiente a Y_0. no tiene que ser O. Para n = 4, 5, 6 y 7 buenas primeras aproximaciones para ρ según Patterson (7) son las siguientes:

n = 4. $r_1 = (4Y_3 + Y_2 - 5\ Y_1)/(4\ Y_2 + Y_1 - 5\ Y_0)$

n = 5. $r_1 = (4\ Y_4 + 3\ Y_3 – Y_2 - 6\ Y_1)/(4\ Y_3 + 3\ Y_2 – Y_1 - 6\ Y_0)$

n=6. $r_1 =(4\ Y_5 + 4\ Y_4 + 2\ Y_3 - 3\ Y_2 - 7\ Y_1)/(4\ Y_4 +4\ Y_3+ 2\ Y_2 - 3\ Y_1 - 7\ Y_0)$

n = 7. $r_1\ (Y_6 + Y_5 + Y_4 – Y_2 - 2\ Y_1) / (Y_5 + Y_4 + Y_3 – Y_1 - 2\ Y_0)$

En un trabajo posterior (8), Patterson da primeras aproximaciones mejores para muestras de tamaño n = 4 a n = 12. El valor de r_1 que se obtiene por la solución de una ecuación cuadrática, es notablemente buena de acuerdo con nuestra experiencia.

En un ejemplo dado por Stevens (9), la Tabla 15.8.1 muestra seis lecturas consecutivas de un termómetro a intervalos de medio minuto después de sumergirlo en una mezcla refrigerante.

TABLA 15.8.1 DATOS PARA AJUSTAR UNA REGRESION ASINTOTICA

X TIEMPO (1/2 min)	Y TEMP. °F	$X_1 = (0.55^x)$	$X_2 = X\ (0.55^{x-1})$	$\hat{Y}_2$	$Y_{res} = Y - \hat{Y}_2$
0	57.5	1.00000	0	57.544	-0.044
1	45.7	0.55000	1.00000	45.525	+0.175
2	38.7	0.30250	1.10000	38.892	-0.193
3	35.3	0.16638	0.90750	25.231	+0.069
4	33.1	0.09151	0.66550	33.211	-0.111
5	32.2	0.05033	0.45753	32.096	+0.104
Total	242.5	2.16072	4.13053		+0.001

Según la fórmula de Patterson (anterior) para n = 6, encontramos r_1, = 10.42/-18.86 = 0.552. Tomando r_1 = 0.55 calculamos 105 valores de muestra de X_1 y X_2 y los insertamos en la Tabla 15.8.1.

La matriz de las sumas de cuadrados y productos de las tres variantes X_i es como sigue:

$\Sigma X_0^2 = 6$ $\Sigma X_0X_1 = 2.16072$ $\Sigma X_0X_2 = 4.13053$

$\Sigma X_0X_1 = 2.16072$ $\Sigma X_1^2 = 1.43260$ $\Sigma X_1X_2 = 1.11767$

$\Sigma X_0X_2 = 4.13053$ $\Sigma X_1X_2 = 1.11767$ $\Sigma X_2^2 = 3.68578$

(Alternativamente, podríamos utilizar el método de las Secs. 13.2 - 13.4 (Pág. 470), y obtener una matriz de 2 X 2 de la ΣX_i ΣX_j, pero al final poco es el tiernpo que se ahorra con esto.)

La matriz inversa de los multiplicadores de Gauss se calcula luego. Cada hilera de esta matriz es multiplicada a su vez por los valores de ΣX_iY (colocados en la columna del lado derecho).

Matriz inversa			ΣX_iY
$C_{11} = 1.62101$	$C_{12} = -1.34608$	$C_{13} = -1.40843$	242.5
$C_{12} = -1.34608$	$C_{22} = 2.03212$	$C_{23} = 0.89229$	104.86457
$C_{13} = -1.40843$	$C_{23} = 0.89229$	$C_{33} = 1.57912$	157.06527

Estas multiplicaciones dan

$$a = 30.723; \quad b = 26.821 ; \quad c = b(r_2 - r_1) = 0.05024 \qquad (15.8.4)$$

De donde,

$$r_2 = r_1 + clb = 0.55 + 0.05024/26.821 = 0.55187$$

La segunda aproximación a la curva es

$$\hat{Y} = 30.723 + 26.821(0.55187)^x \qquad (15.8.5)$$

Para juzgar si la segunda aproximación está bastante cercana a la solución de mínimos cuadrados, encontramos ΣY_{res}^2 para las primeras dos aproximaciones. La primera aproximación es

$$\hat{Y}_1 = a_1 + b_1 (0.55^x) = a_1 + b_1 X_1 \qquad (15.8.6)$$

donde a_1, b_1 están dadas por la regresión lineal de Y sobre X_1. En los cálculos anteriores a_1, b_1 no fueron calculadas, puesto que no se necesitan para encontrar la segunda aproximación. Sin embargo, por las reglas comunes para la regresión lineal ΣY_{res}^2 de la primera aproximación está dada por

$$\Sigma Y^2 - (\Sigma Y)^2/n - (\Sigma y x_1)^2 / \Sigma X_1^2 , \qquad 15.8.7)$$

donde, como de costumbre, $x_1 = X_1 - \bar{X}_1$. Cuando la curva ajusta estrechamente, corno en este ejemplo, hay que llevar muchos decimales en los cálculos, tal como Stevens (9) lo ha previsto. Alternativamente, podemos calcular a_1 y b_1 en (15.8.6) y de alli $Y - \hat{Y}_1$, obteniendo así la suma residual de cuadrados directamente. Con el numero de decimales que hemos llevado, obtenemos 0.0988 por la formula 15.8.7 y 0.0990 por el método directo, siendo la primera cifra la mas exacta.

Para la segunda aproximación, calculamos las potencias de $r_2 = 0.55187$, y de allí encontramos $\hat{Y}_2$ por la (15.8.5). los valores de $\hat{Y}_2$ y de $Y - \hat{Y}_2$ estan mostrados en la tabla 15.8.1. la suma de cuadrados residuales es 0.973. la disminución de la primera aproximación (0.0988 a 0.0973) es tan pequeña, que bien podemos detenernos en la segunda aproximación. Mas aproximaciones conducen a un mínimo de 0.0972.

La media cuadrada residual para la segunda aproximación es $s^2 = 0.0973/3 = 0.0324$, con n- 3 = 3 g.l. los errores estándar aproximados son (utilizando la matriz inversa):

e.e. $(a_2) = s\sqrt{c_{11}} = +- 0.23$; e.e. $(b_2) = s\sqrt{c_{22}} = +-0.26$;

e.e.$(r_2) = s\sqrt{c_{33}}/b_2 = 0.226/26.82 = +-0.0084$

estrictamente hablando, los valores de c_{ii} habrán de calcularse para r = 0.55187 en lugar de r = 0.55, pero los resultados anteriores son bastante justos. Ademas, como $r_2 – r_1 = c/b$, una mejor aproximación para el error estándar r_2 esta dado por la formula para el error estándar de una razón.

$$e.e.(r_2) = sc/b\{c_{33}/c^2 + c_{22}/b^2 – 2c_{23}/bc\}$$

En casi todos los casos, el termino c_{33}/c^2 en la raíz cuadrada domina, lo que reduce el resultado a $s\sqrt{c_{33}}/b$.

Cuando X tiene valores 0,1,2,..., (n-1), los cálculos en maquina de escritorio, de la segunda aproximación se abrevian considerablemente por las tablas auxiliares. Las c_{ii} y c_{ij} en la matriz inversa 3 x 3, que hemos de calcular para cada etapa, depende únicamente de n y r. Stevens (9) tabulo estos valores para n = 5, 6, 7. Con estas tablas, el que las usa encuentra la primera aproximación r_1, y calcula los valores muéstrales de X_1 y X_2, así como las cantidades ΣY, $\Sigma X_1 Y$, $\Sigma X_2 Y$. Los valores de $c_{i\,j}$ correspondientes a r_1 no se leen de las tablas de Stevens, y las segundas aproximaciones se obtienen rápidamente como en la tabla (15.8.4) anterior. Hiorns (10) ha tabulado la matriz inversa para r avanzando por 0.01, de 0.1 a 0.9 y para tamaños de muestra de 5 a 50.

EJEMPLO 15.8.1. En un experimento sobre trigo en Australia, se aplicaron fertilizantes a una serie de niveles, con estos resultados en rendimientos:

Nivel	X	0	10	20	30	40
Rendimiento	Y	26.2	30.4	36.3	37.8	38.6

Ajustar la ecuación de Mitscherlich. Resp. La formula de Patterson da r_1=0.40. La segunda aproximación es r_2 = 0.40026, pero la suma residual de cuadrados es prácticamente la misma que para la primera aproximación, que es $y= 38.679-12.425 (0.4)^{x}$.

EJEMPLO 15.8.2. En una reacción química, la cantidad de pentoxido de nitrógeno que se descompone a diversos tiempos, después de iniciarse la reacción, fue como sigue:

Tiempo (T)	2	3	4	5	6	7
Cantidad descompuesta (Y)	18.6	22.6	25.1	27.2	29.1	30.1

Ajústese una regresión asintotica. Obtenemos $\hat{Y} = 33.802 - 26.698 (0.753)^{T}$, con S.C residual = 1.105.

Ejemplo 15.8.3. Stevens (9) ha hecho notar que cuando ρ esta entre 0.7 y1, la curva de regresión asintotica se aproxima mucho a un polinomio de segundo grado. La ecuación asintotica $Y = 1 - 0.9 (0.8)^{x}$ toma los siguientes valores:

X	0	1	2	3	4	5	6
Y	0.100	0.280	0.424	0.539	0.631	0.705	0.764

Ajústese la parábola por polinomios ortogonales y obsérvese que tan bien concuerdan los valores de Y.

Introducción

El tipo más común y sencillo de ajuste de curvas es de la línea recta. Sin embargo, cuando se representan pares de observaciones, éstas suelen quedar sobre una línea curva; las teorías biológicas y de algún otro tipo hasta pueden exigir una curva de forma especificada. Este capítulo considera tal regresión. Además, hay una breve exposición de la construcción y uso de polinomios ortogonales.

Regresión no lineal

Una relación entre dos variables puede ser aproximadamente lineal cuando se estudia en un intervalo limitado, pero puede ser marcadamente curvilínea si se amplía el intervalo. Por ejemplo, la relación entre madurez y rendimiento en arvejas para enlatar usualmente es lineal sobre el intervalo de madurez aceptable para la industria de enlatado. Pero, el aumentar la madurez, la tasa de aumento de¡ rendimiento se disminuye, esto es, se vuelve curvilínea. Análogamente, la tasa de aumento en rendimiento tiende a mejorar en las etapas de inmadurez. Así,

pues, para describir la relación en todo el intervalo es inadecuada una ecuación lineal.

Además de usar una curva apropiadamente descriptiva, es procedimiento acertado quitar del error toda componente que mida la regresión curvilínea. Así, si una observación se describe apropiadamente con $Y = \beta_0, + \beta_1 X + \beta_2 X^2 + e$, y usamos $Y = \beta_0, + \beta X + e$ como modelo, entonces se asigna a la medida del error una parte de la variación entre medias de población, o sea, la asociada con $\beta^2 X^2$. Es claro que ello exagera nuestra medida del error.

La selección de la forma de la ecuación de regresión que mejor expresa una relación curvilínea no siempre es problema simple. Prácticamente no hay límite en cuanto al numero de tipos de curvas que pueden expresarse por ecuaciones matemáticas. Entre las ecuaciones posibles, puede haber muchas igualmente buenas para minimizar la SC(residuos). Por tanto, al escoger la forma de la curva, es deseable tener alguna teoría dada por especialistas que trabajen en el campo de la materia del tema, además, también puede ser bueno considerar la labor que entra en el ajuste de la regresión y si se cumples los supuestos acostumbrados necesarios para la validez de la estimación y los procedimientos de prueba.

Tales consideraciones nos llevan a clasificar las relaciones curvilíneas en dos tipos:

lineales y no lineales en los parámetros. Los modelos que son lineales en los parámetros son aquellos para los cuales se dispone de técnicas de regresión múltiple, entre ellos los modelos polinomiales. Los modelos que no son lineales en los parámetros son intrínsecamente lineales si los hace lineales una transformación. Ejemplos típicos de esta transformación son las curvas logarítmica y exponencial. Modelos que no se pueden linealizar mediante una transformación son intrínsicamente no lineales y los análisis correspondientes se llaman regresiones no lineales. Este problema no se estudia aquí, pero en Draper

y Smith (19.3, cap. 10) se da una introducción y Gallant (19.6) presenta un artículo expositivo sobre el tema.

Las transformaciones tienen por objeto proporcionar un procedimiento más fácil de ajuste y/o procedimientos validos de estimación y prueba. Por ejemplo, podernos convenir en que la ecuación $E(Y) = \beta_0 X^{b1}$ se basa en un sólido razonamiento biológico. Entonces $\log E(Y) = \log \beta_0 + \beta_1, \log X$ es una ecuación si el par de observaciones se considera como (log Y, log X). Los procedimientos de los caps. 10 y 17 son aplicables. Con datos como esos, no es infrecuente encontrar que los supuestos que se refieren a la normalidad casi sean más apropiados en escala transformada que en la original.

Ahora consideramos dos tipos generales de curvas: polinomiales y exponencial o logarítmica. He aquí algunos ejemplos, para los cuales se muestran en la fig. 19.1 las formas generales, Para las ecuaciones exponenciales, e puede reemplazarse por cualquiera otra constante sin que se afecte la forma de la curva que se ajusta. Para la ecuación $E(Y) = \beta_0 X^{b1}$ los valores enteros del exponente β_1 dan casos especiales de polinomios. Sin embargo, es más probable que se use este tipo de curva cuando se desea un experimento fraccionado como ocurre a menudo en el campo de la economía.

Polinomial	Exponencial	Logarítmica
Lineal $E(Y) = \beta_0 + \beta_1 X$	$e^{E(Y)} = \beta_0 X^{B1}$	$E(Y) = \beta_0 + \beta_1 \log X$
Cuadratica $E(Y) = \beta_0 + \beta_1 + \beta_2 X^2$	$E(Y) = \beta_0 X_1^{x}$	$\text{Log } E(Y) = \beta'_0 + \beta'_1 X$
Cubica $E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 \beta_3 X^3$	$E(Y) = \beta_0 X^{B1}$	$\text{Log } E(Y) = \beta'_0 + \beta'_1 \log X$

Los polinomios pueden tener picos y depresiones cuyo numero a lo mas es uno menos que el exponente mas alto. Por ejemplo, la ilustración de la parte inferior izquierda de la fig. 19.1 tiene un pico y una depresión, o sea dos de tales puntos en una curva en que el exponente mas elevado es 3. A los picos se les llama máximos y a las depresiones se les llama mínimos. Al ajustar curvas polinomiales, el investigador se interesa usualmente en un dado intervalo total representado por la ecuación.

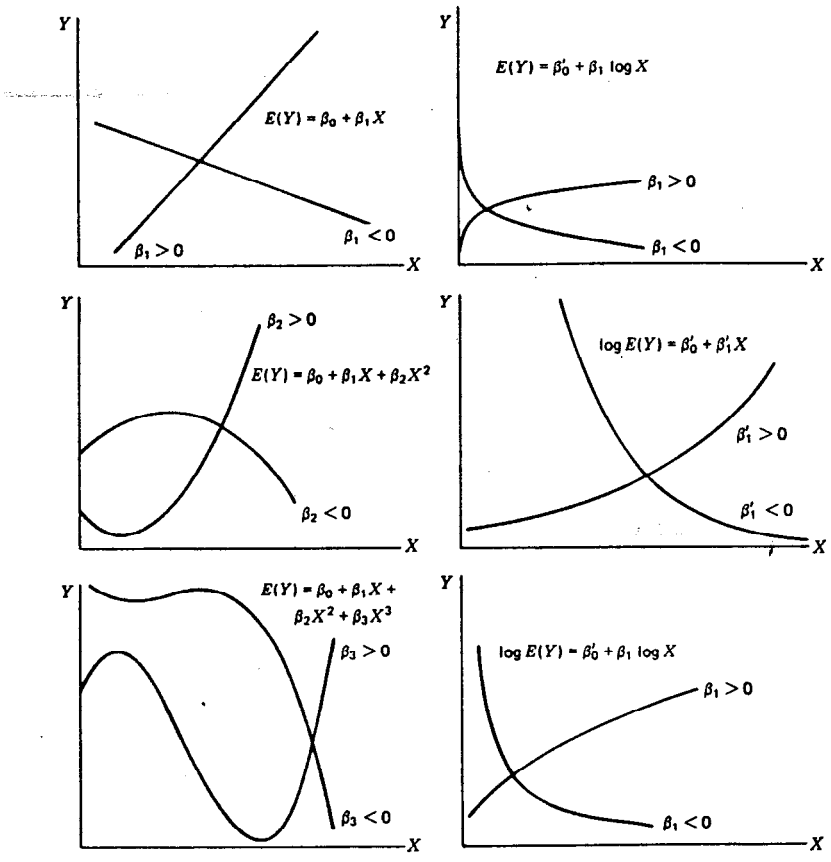


Figura 19.1 Tipos generales de curvas

Las curvas exponenciales o logarítmicas, excepto las de la forma Log E (Y) = β'_0 + β'_1 log X se caracterizan por un aplanamiento hacia un extremo del intervalo. Por ejemplo, la curva E(Y) = log X se aproxima más y más a X = 0 a medida que Y toma valores negativos numéricamente más y más grandes; pero esta curva nunca cruza la recta vertical X = 0. Los números negativos no tienen logaritmos reales.

19.3 Curvas logarítmicas o exponenciales

Las curvas logarítmicas, o simplemente log, son lineales cuando se representan en papel logarítmico, apropiado. Refiriéndonos a la fig. 19.1 (lado derecho, de arriba a abajo) tenemos:

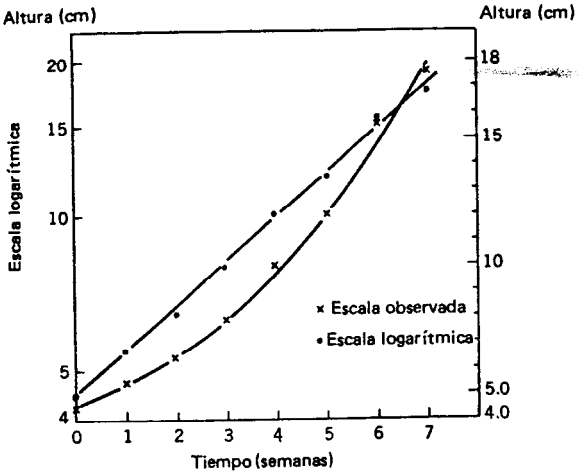


Figura 19.2 Puntos observados representados en escalas de intervalo fijo y logarítmica.

1. $e^y = \beta_0 X^{B1}$ o $Y = \beta'_0 \log X$. Representando en papel semilogarítmico, los puntos (X -Y) dan lugar a una recta, en la que Y se representa en la escala de intervalos iguales, y X, en la escala logarítmica. En esencia, el papel semilogarítmico encuentra y representa los logs de X.

2. $Y = \beta_0 \beta^x_1$ o $\log Y = \beta'_0 \beta'_0 X$. Los puntos (X, Y) dan lugar a una recta al representarlos en papel semilogarítmico, en la que Y va en escala logarítmica y X en la escala de intervalos iguales (ver también fig, 19.2).

3. $Y = \beta_0 X^{B1}$ o $\log Y = \beta'_0$, log X. Esta se comporta como una recta cuando se representa en doblemente logarítmico. En el que ambas escalas son logarítmicas.

(Ver también fig. 19.3).

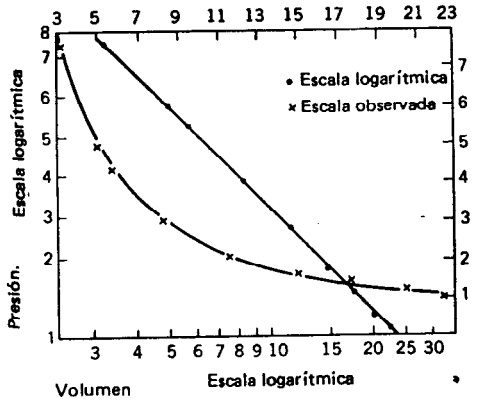


Figura 19.3 Puntos observados representados en escalas de intervalo fijo y logarítmica.

Para determinar si una curva logarítmica puede describir datos, suele ser suficiente con representar los datos en papel logarítmico. Una vez tomada la decisión respecto al tipo de curva logarítmica, se transforman los valores observados de X o Y o de ambos a logaritmos antes de realizar los cálculos. Los datos transformados se tratan luego por los métodos de los caps. 10 y 17. Los supuestos usuales se aplican a los datos transformados en lugar de a los originales.

Para ilustrar lo anterior, W. J. Drapala, Escuela Superior del Estado de Misisipi, ha proporcionado dos conjuntos de datos. En la tabla 19.1 se presentan los datos de la altura en centímetros por encima de los cotiledones, tomadas semanalmente, de repollos Golden Acre. Los datos aparecen representados en la fig. 19.2, con X en la escala de intervalo fijo y Y en la escala logarítmica, y también en la escala de intervalo fijo. Obsérvese como el conjunto de puntos toma la forma lineal cuando se representan en papel semilogarítmico.

Ahora procedemos con los cálculos para los datos transformados, como se hizo en los caps. 10 y 13. 0 bien un programa de computador provee información, tal como en la tabla 19.2.

La ecuación de regresión es

$$\widehat{\log Y} = 0.6497 + 0.0866X$$

La media de los valores de los Y es 0.9528.

Los datos de la tabla 19.3 son presiones, en atmósferas de gas oxígeno a 25°C al ocupar varios volúmenes, medidos en litros. Los datos se comportan en forma curvilínea cuando se representan en una escala de intervalos iguales, y en forma casi lineal en la escala logarítmica doble (ver fig. 19.3). Se. ajusta la relación, $\log Y = b_0 + b_1 \log X$. También son apropiados los procedimientos de los caps. 10 y 13. La tabla 19.4 es parte de una salida impresa del computador del análisis de datos. La ecuación de regresión es

$$\widehat{\log Y} = 1.3790 - 1.0022 \log X$$

Tabla 19.1 Altura por encima de los cotiledones de repollo Golden Acre media a intervalos semanales

Semanas después de la primera observación X	Altura cm Y	Logaritmo decimal de altura Y
0	4.5	0.653
1	5.5	0.740
2	6.5	0.813
3	8.0	0.903
4	10.0	1.000
5	12.0	1.079
6	15.5	1.190
7	17.5	1.243

Tabla 19.2. Análisis de los datos de cotiledones mediante SAS

PROCEDIMIENTO GENERAL DE MODELOS LINEALES

VARIABLE DEPENDIENTE		SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR>F	R-CUADRADO
FUENTE	G.L					
MODELO	1	0.31497610	0.31497610	2347.62	0.0001	0.997451
ERROR	6	0.00080501	0.00013417			
TOTAL CORREGIDO	7	0.31578111				
DESV. EST 0.01158310		LOG Y MEDIA 0.95276619				

FUENTE	G.L		TIPO ISC	VALOR F	PR>F	G.L	TIPO IV SC	VALOR F	PR>F
X	1		0.31497610	2347.6	0.0001	1	0.31497610	2347.62	0.0001
PARAMETRO	**ESTIMACION**		T PARA Ho: PARAME =0	PR> T	ERROR ESTANDAR DE LA ESTIMACION				
INTERCEPTO	**0.64966880**		86.89	0.0001	0.00747686				
X	**0.08659926**		48.45	0.0001	0.00178731				

Tabla 19.3 presion de oxigeno a 25 °C al ocupar varios volúmenes

		Logaritmo natural	
Volumen Litros X	Presion Atm, Y	Volumen X	Presion Y
3.25	7.34	0.512	0.866
5.00	4.77	0.699	0.679
5.71	4.18	0.757	0.621
8.27	2.88	0.918	0.459
11.50	2.07	1.061	0.316
14.95	1.59	1.175	0.201
17.49	1.36	1.243	0.134
20.35	1.17	1.309	0.068
22.40	1.06	1.350	0.025

Obsérvese que el valor de F

deberá comenzar con 4. Pero el numero es demasiado grande para el espacio disponible, así, todos los 9 se imprimen por convención de programación.

Ejercicio 19.3.1 Brockington et al. (19.2) recolectaron los datos que se presentan en la tabla adjunta, sobre la relación entre contenido de humedad y la humedad relativa interstical de semillas de maíz entero.

	Contenido de humedad		Humedad relativa
Muestra No.	Brown-Duval	Horno de dos etapas	Equilibrio, Porcentaje
1	7.0	9.4	40.0
2	7.5	9.9	40.0
3	11.6	12.9	59.0
4	11.8	12.6	63.5
5	12.9	14.1	71.5
6	13.2	14.7	71.0
7	14.0	15.2	76.5
8	14.2	14.6	75.5
9	14.6	15.2	79.0
10	14.8	15.8	79.0
11	15.7	15.8	82.0
12	17.3	17.2	85.5
13	17.4	17.0	85.0
14	17.8	18.2	87.5
15	18	18.5	86.5
16	18.8	18.2	88.0
17	18.9	19.4	90.0
18	20.0	20.3	90.5
19	20.7	19.9	88.5
20	22.4	19.5	89.5
21	22.5	19.8	91.0
22	26.8	22.6	92.0

Tabla 19.4. Análisis de los datos de presión mediante SAS

PROCEDIMIENTO GENERAL DE MODELOS LINEALES

VARIABLE DEPENDIENTE		SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR>F	R-CUADRADO
FUENTE	G.L					
MODELO	1	0.70897285	0.0.70897285	99999.99	0.0000	0.999998
ERROR	7	0.00000109	0.00000016			
TOTAL CORREGIDO	8	0.70897394				
DESV. EST 0.00039463		LOG Y MEDIA 0..37435348				

FUENTE	G.L		TIPO SC	VALOR F	PR>F	G.L	TIPO IV SC	VALOR F	PR>F
LOG X	1		0.70897285	99999.99	0.0000	1	0.70897285	99999.99	0.0000
PARAMETRO	**ESTIMACION**		T PARA Ho: PARAME =0	PR> T	ERROR ESTANDAR DE LA ESTIMACION				
INTERCEPTO	**0.1.37899089**		2820.73	0.0001	0.00048888				
LOG X	**-1.00219470**		-2133.68	0.0001	0.00046970				

Representar Y = humedad relativa de equilibrio respecto de X = contenido de humedad (cualquiera de las dos medidas) con X en escala ordinaria y en escala logarítmica. ¿Cuál parece ser la escala apropiada para obtener una recta? Calcular la regresión lineal de la humedad relativa de equilibrio respecto al, log del contenido de humedad. ¿Qué porcentaje de la SC(total) queda aplicado por la regresión lineal?

Ejercicio 19.3.2 En un estudio sobre el tamaño y forma óptimos de parcela, Weber y Homer (19.9) consideraron la longitud de la parcela y la varianza de las medias de parcelas por unidad básica de rendimiento en gramos y el porcentaje de proteína (entre otras formas y Características de parcelas). Obtuvieron los datos que se indican en la tabla adjunta.

		Varianza	
Forma	Número de unidades	Rendimiento, g	Proteína, porcentaje
8 x 1	1	949	.116
16 x 1	2	669	.080
24 x 1	3	540	.053
32 x 1	4	477	.048

Representar los dos conjuntos de varianzas con respecto al número de unidades después de transformar las tres variables a escala logarítmica. ¿Los datos resultantes guardan alguna relación razonablemente lineal?

19.4 El polinomio de segundo grado

En la sección 14.8 se, trato en forma breve de los modelos polinomiales. Aquí hacemos una ilustración sobre el polinomio de grado 2.

Sea la ecuación (19.1) la descripción matemática de una observación. Es claro que la linealidad se aplica a los parámetros que deben estimarse y no a los observables.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 x^2_i + e_i \qquad (19.1)$$

En esta ecuación, β_1 y β_2 son coeficientes de regresión parcial, tal como se expuso en el cap. 14, pero no pueden interpretarse sino en espacio que no sea de más de dos dimensiones.

Para estimaciones por mínimos cuadrados de los parámetros, ha de cumplirse la ec. (19.2).

$$\Sigma (Y - \hat{Y})^2 = \Sigma (Y_i - \beta_1 X_1 - \beta_2 X_i^2)^2 = \min \qquad (19.2)$$

Las ecuaciones normales o por mínimos cuadrados son

$$X'X\beta_1 = X'Y \qquad (19.3)$$

Tabla 19.5. rendimiento Y, en libras por parcelas, y lectura del tenderometro X, de arvejas Alaska, cultivadas en Madison, Wisconsin, 1953.

Y rendimiento, libras por parcela:	24.0	22.0	26.5	22.0	25.0	37.5	36.0	39.5	32.0	26.5	55.5	49.5	56.0	55.5
X lectura del tenderometro:	76.2	76.8	77.3	79.2	80.0	87.8	93.2	93.5	94.3	96.8	97.5	99.5	104.2	106.3

Continuación

Y rendimiento, libras por parcela:	58.0	61.5	69.0	71.5	73.0	76.5	78.5	74.0	71.5	77.0	85.5		
X Lectura del tenderometro:	106.7	119.0	119.7	119.8	119.8	123.5	141.0	142.3	145.5	149.0	150.0		
$\sum X = 2698.9$	$\sum X^2 =$ 305148.35			$\sum X^3 =$ 36045287.22			$\sum X^4 =$ 4429289685.23						
$\sum Y = 1303.5$	$\sum Y^2 =$ 78797.25			$\sum XY =$ 152129.55			$\sum X^2Y =$ 18424791.12						

Donde

$X'X = (n, \Sigma X_i, \Sigma X^2_i),(\Sigma X_i, \Sigma X^2_i, \Sigma X^3_i),(\Sigma X^2_i, \Sigma X^3_i, \Sigma X^4_i)$ y $X'Y = \Sigma Y_i, \Sigma X_i Y_i, \Sigma X_i Y^2_i$.

Como ilustración se usan los datos de la tabla 19.5. Las entradas numéricas de **X'X y X'Y** se dan allí. La tabla 19.6 proporciona un análisis de la varianza,

pruebas de hipótesis y estimaciones de los parámetros necesarios. Aproximadamente el 91 por ciento (= 100 R^2) de la variación, medida por la suma de cuadrados total, puede explicarse por una ecuación cuadrática en X. La ecuación de regresión muestral es

$$\hat{Y} = - 138.5068 + 2.7133X – 0.0084 X^2$$

La figura 19.4 da tanto la ecuación lineal como la cuadrática. Los cálculos para la regresión lineal proceden de la salida impresa del computador, pero aquí no se presentan.

Ejercicio 19.4.1 A la tabla 10.5 agregar el par de observaciones (1,949, 1,976). Calcular la regresión cuadrática del número de caballos por año. ¿Sería deseable incluir X^2 en la ecuación de regresión? ¿Por que?

19.5 Polinomios ortogonales

Cuando los aumentos entre niveles sucesivos de X son iguales y los valores de Y tienen una varianza común, pueden usarse tablas de valores de polinomios ortogonales en los cálculos que lleven a las pruebas de hipótesis respecto a la bondad de ajuste de polinomios de diversos grados. En la sec. 15.7 se ilustran ambos procedimientos.

Tabla 19.6. Análisis de los datos de arveja Alaska mediante SAS

PROCEDIMIENTO GENERAL DE MODELOS LINEALES

VARIABLE DEPENDIENTE: Y		SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR>F	R-CUADRADO
FUENTE	G.L					
MODELO	2	9888.85463514	4944.42731757	115.24	0.0001	0.912866
ERROR	22	943.90536486	42.90478931			
TOTAL	24	10832.760000				

CORREGIDO		00				
DESV. EST 6.55017475		Y MEDIA 52.14000000				

FUENTE	G.L		TIPO ISC	VALOR F	PR>F	G.L	TIPO IV SC	VALOR F	PR>F
X	1		9441.75392063	220.006	0.0001	1	917.28607392	21.38	0.0001
X*X	**1**		447.10071451	10.42		1	447.10071451	10.42	0.0039
PARAMETRO	**ESTIMACION**		T PARA H0 PARAMETRO =0	PR>T	ERROR ESTANDAR DE LA ESTIMACION				
INTERCEPTO	**-138.50680942**		-4.33	0.0003	31.98150177				
X	**2.71328314**		4.62	0.0001	0.58680743				
X*X	**-0.00837858**		-3.23	0.0039	0.00259550				

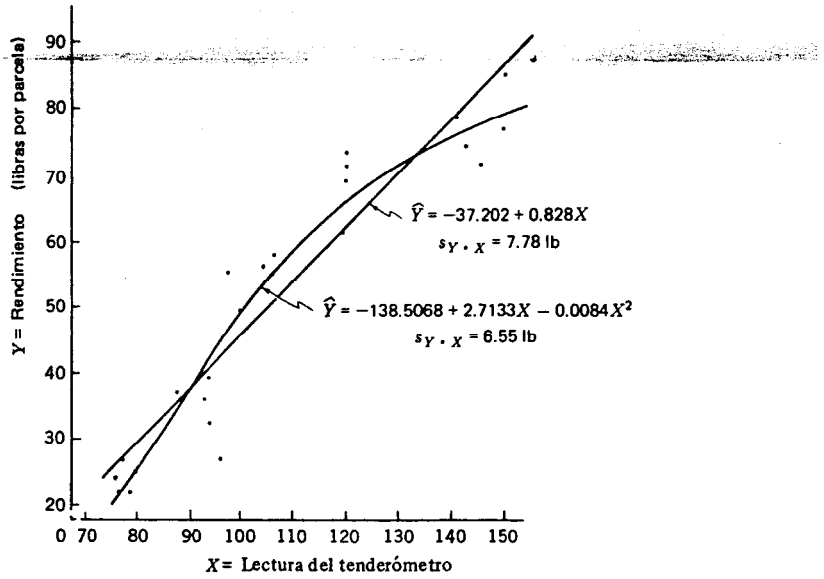


Figura 19.4 Relación entre rendimiento y lectura del tenderómetro para los datos de la tabla 19.5

Los polinomios ortogonales pueden usarse en situaciones en que las X no están igualmente espaciadas o las observaciones tienen varianzas desiguales pero conocidas. Los polinomios sucesivos son independientes unos de otros y nos permiten calcular otras sumas de cuadrados atribuibles a las varias potencias de X.

Para X desigualmente espaciadas o medias Y basadas en números desiguales, o ambas, Robson (19.8) usa lo que esencialmente es una razón de Fisher (19.4, 19.5) para obtener una formula recurrente para calcular coeficientes. (Una formula recurrente es la que se aplica una y otra vez, utilizando en cada nueva aplicación la aplicación anterior). Robson procede como sigue:

$$\hat{Y}_i = b_0 f_0 (X_i) + b_1 f_1(X_i) + ...+ b_r f_r(X_i) \qquad (19.4)$$

I = 1,.....,n>r, donde $\hat{Y}_i$ es la media de población estimada de las Y cuando $X = X_i$

El polinomio $f_j(X_i)$ es de grado j y proporcionara los valores de los polinomios ortogonales, uno para cada X, necesario para determinar los coeficientes de regresión. Los coeficientes son

$$B_j = \Sigma Y_i f_j (X_i) \qquad (19.5)$$

Cada coeficiente de regresión representa una comparación, tal como se define en la sec. 8.3. estas comparaciones son ortogonales y dan otras sumas de cuadrados atribuibles a la inclusión de X a la potencia j-esima en la ecuación de regresión. Así, b_0 es el grado cero o efecto medio (no el principal), b_1 el primer grado o efecto lineal, y así sucesivamente. Finalmente

$$\Sigma Y^2_i = b^2_0 + \ldots + b^2_{n-1}$$

Si efectuamos de r a n-1 de tal manera que todos los gl = n hayan siso tenidos en cuenta individualmente.

La formula de recurrencia dada por Robson es la ec. (19.6) con c_h definido mediante la ec. (19.7).

$$f_h(X_i) = 1/c_h \left(X^h_i \Sigma f_j(X_i) \Sigma X^h_g f_j(X_g) \right) \qquad (19.6)$$

$$c^2_h = \Sigma \left(X^h_i \Sigma f_j(X_i) \Sigma X^h_g f_j(X_g) \right)^2 \qquad (19.7)$$

obsérvese que c^2_h es la suma de cuadrados de cantidades, tales como las entre corchetes de la ec. (19.6).

ahora vamos a ilustrar el uso de la formula de recurrencia. Supóngase que tenemos medias del porcentaje de digestión de celulosa a 6, 12, 18, 24, 36, 48, y 72 h. Por la aplicación de las ecs. (19.6) y (19.7) se tiene

$$f_0(X_i) = 1/c_o (X^0_i) = 1/c_0$$

$$c^2_0 = X^0_1 + \ldots + X^0_n = n$$

(Una cantidad elevada a la potencia cero es igual a uno) ahora bien,

$$f_0(X_1) = 1/\sqrt{n} = f_0(X_2) = \ldots = f_0(X_n)$$

de la ecuación (19.5),

$$b_0 = \Sigma\, Y_i 1/\sqrt{n}$$

y la reducción atribuible a este polinomio es

$$c^2_0 = (\Sigma Y)^2/n$$

Este es el termino de corrección, tal como pudimos haberlo esperado. Los restantes n – 1 grados de libertad para medias de tratamientos.

En seguida

$$f_1(X_i) = 1/c_1 \left(X_i - f_0(X_i)\, \Sigma\, X_g\, f_0(X_g) \right)$$

$$= 1/c_1 \left(X_i - 1/\sqrt{n}\, \Sigma X_g\, 1/\sqrt{n} \right)$$

$$= 1/c_1 (X_i - \bar{X})$$

$$c^2_1 = \Sigma(X_i - \bar{X})^2$$

Ahora

$$\bar{X} = 216/7 = 30.86 \qquad \text{y } \Sigma(X_i - \bar{X})^2 = 3{,}198.86$$

$$f_1(X_i) = (6 - 30.86)/\sqrt{3{,}198.86}, \ldots\ldots f_1(X_n) = (72 - 30.86)/\sqrt{3{,}198.86}$$

Finalmente

$$B_1 = \Sigma Y_i (X_i - \bar{X}) / \sqrt{\Sigma (X_i - \bar{X})^2}$$

Y la reducción atribuible a regresión lineal es

$$b^2_1 \left(\Sigma (X_i - \bar{X}) \right)^2 / \Sigma (X_i - \bar{X})^2$$

También se esperaba este resultado. La cantidad entre corchetes del numerador usualmente se escribe en la forma $\Sigma(Y_i - \bar{Y})(X_i - \bar{X})$, pero las dos formas son equivalente. Obsérvese que para r = 1, la ec. (19.4) se convierte en

$$\hat{Y} = (\Sigma Y_i)/\sqrt{n} \; 1/\sqrt{n} + \Sigma Y_i (X_i - \bar{X}) / \sqrt{\Sigma (X_i - \bar{X})^2}$$

$$= \hat{Y} + \Sigma (X_i - \bar{X}) (Y_i - \bar{Y}) / \Sigma (X_i - \bar{X})^2 \quad (X - \bar{X})$$

para la ecuación de segundo grado,

$$f_2 (X_i) = 1/c_2 \left(X^2_1 - 1/\sqrt{n} (\Sigma X^2_g 1/\sqrt{n}) - X_i - \bar{X} / \sqrt{\Sigma (X_i - \bar{X})^2} \; \Sigma X^2_g (X_g - \bar{X}) / \sqrt{\Sigma (X_i - \bar{X})^2} \right)$$

$$= 1/c_2 \left(X^2_1 - 1/n \, \Sigma (X^2_g - (X - \bar{X}) \Sigma X^2_g (X_g - \bar{X}) / \Sigma (X_i - \bar{X})^2 \right)$$

de nuevo, debemos encontrar c^2_2 mediante la ec.(19.7). Obsérvese que $f_2 (X_i)$ es un polinomio de grado dos; tiene un X^2_i un X_i como un $X_i - \bar{X}$ multiplicado por una constante, y un termino constante, que es $- \Sigma X^2_g / n$. El hecho de que exista un X_i indica que la ecuación cuadratica completa tendrá un coeficiente de X que difieren del coeficiente de X en la ecuación lineal; el coeficiente de X_i en $f_2 (X_i)$ provee el ajuste.

Obsérvese que hemos calculado las funciones ortogonales para el caso general. Entonces, para nuestro ejemplo, hemos usado las X del experimento para obtener los coeficientes de las Y para la función lineal de las Y que da la reducción adicional atribuible a la mas alta potencia de X introducida.

Ejercicio 19.5.1 A. Van Tienhoven, Universidad de Cornell, llevo a cabo un experimento factorial 5 x 5 para estudiar los efectos de la hormona tiroide (HT) y una hormona estimulante de la tiroide (HET) sobre las alturas del epitelio folicular (entre otras respuestas) en pollos. La TH se mide en unidades γ , HET en unidades Junkmann-Schoeller, y la respuesta en unidades de micrómetro. En la tabla adjunta se presentan los totales de tratamiento. Cada total proviene de cinco observaciones.

	HET					
		.00	.03	.09	.27	.81
HT	.00	3.42	6.21	11.21	14.40	19.40
	.04	5.64	5.85	9.16	18.30	19.65
	.16	5.13	8.39	12.74	15.20	15.07
	.64	5.37	5.24	9.14	17.66	16.30
	2.56	4.54	6.49	8.37	14.23	16.90

Fuente : datos no publicados y usados con permiso de A. Van Tienhoven, Universidad de Cornell, Ithaca, Nueva York.

Se supone que la respuesta medida en alturas del epitelio sigue una curva logarítmica. Los primeros niveles de HT y HET no se encuentran en secuencias logarítmicas igualmente espaciadas con los otros niveles. Pero se quería obtener información para este tratamiento particular.

Un análisis preliminar es como sigue:

Omitir los tratamientos HET = 0.00 y HT = 0.00. representar la respuesta a HT, para cada nivel de HET, en papel logarítmico con tratamiento en la escala logarítmica. ¿ en esta escala aparece lineal la respuesta? Repítase lo anterior para la respuesta a HET.

Emplear polinomios ortogonales para calcular sumas de cuadrados para las respuestas lineal, cuadratica y cubica globales a la HET para los cuatro niveles diferentes de cero. Tomar la escala logarítmica para los tratamientos de modo que se pueda usar la tabla de coeficientes directamente. Hacer una prueba de significancia para cada una de las respuestas.

Considerar como sé probaría la homogeneidad de estas varias respuestas para los niveles de HT. Probar la homogeneidad de las respuestas cubicas, ya que miden desviaciones con respecto a las respuestas cuadraticas y pueden considerarse como candidatos para un termino de error.

Calcular las sumas de cuadrados para las formas lineal, cuadratica y cubica globales de las respuestas a la HT.

¿ como se encontrarían coeficientes para medir la HET (lineal) por HT (lineal), HET (lineal) por HT (cuadratica), y HET (cuadratica) por HT(lineal)? Hallar y probar estas componentes.

Análisis preliminar

Fuente	gl	SC
Total	123	1378.85
TSH	4	540.33
TH	4	48.09
TSH x TH	16	154.67
Error	99	635.76

* Falto una observación

BIBLIOGRAFIA

Steel. 1986. Bioestadistica. Segunda edición. Ed. McGraw-Hill. México.

Snedecor, G.W. 1979. Metodos Estadísticos. Sexta impresión. Ed continental.

Little. T.M. y F.J. Hill. 1989. Métodos Estadísticos para la investigación en la agricultura. 2ª edicio. Ed. Trillas.

EL EMPLEO DE LOS DISEÑOS EXPERIMENTALES EN LA INVESTIGACIÓN AGROPECUARIA Y SU RELACION CON OTRAS CIENCIAS.

Introducción

Un experimento se ha definido, en términos muy generales como “un curso de acción destinado a responder una o más preguntas debidamente enmarcadas”, para nosotros, sin embargo, nuestro interés es más restringido pues se trata de experimentos en los que el experimentador escoge ciertos factores para su estudio y deliberadamente modifica estos factores en forma controlada y observa los efectos de esta acción.

El objetivo es lograr una descripción, explicación y predicción de los procesos naturales. La descripción es la etapa inicial del estudio. La explicación consiste en encontrar las leyes naturales que rigen el fenómeno. La predicción se logra mediante el uso adecuado de las leyes naturales.

Importancia de los diseños experimentales en la ciencia

Como una referencia histórica podemos señalar que las investigaciones que se llevan a cabo con animales no es nuevo ni desconocido ya que por ejemplo en medicina se llevan a cabo experimentos con animales de laboratorio como los hámster, muchos científicos utilizan los experimentos animales y la experiencia clínica a la vez, es casi imposible hoy de determinar de qué método proceden los descubrimientos. Seguramente, los experimentadores y sus partidarios atribuirían cualquier éxito en el tratamiento y curación de enfermedades a la investigación basada en experimentos animales.

El uso de animales como modelos de experimentación viene desde muy antiguo. Toda la descripción de la circulación sanguínea, por ejemplo, ha sido con animales. Ha sido la manera de poder ir viendo cómo se comporta y cómo funciona el cuerpo humano, de qué se compone, mediante la disección en

cadáveres, pero, realmente, para llegar al funcionamiento era imprescindible el uso de animales.

Por otro lado, es evidente que las ciencias con su orientación hacia los experimentos animales, es fruto de su época, una época en la que todo parece ser posible técnicamente, desde la exploración lunar hasta el transplante de órganos. La experimentación animal como un medio de las ciencias médicas por ejemplo sí encaja en el siglo presente, totalmente materializado: *lo que no se puede medir no existe.*

En el campo de la zootecnia de la misma manera, estudiosos del área efectúan investigaciones que van desde el mejoramiento genético enfocados por ejemplo a obtener mejor material genético donde el procedimiento es realizar cruzas y retrocruzas hasta obtener el material que ellos deseen así por ejemplo buscar en el caso de pollos broiler (para carne) una estirpe o línea con mejor conversión alimenticia y para ello lógicamente se llevan a cabo dichas investigaciones empleando modelos estadísticos y/o experimentales que los lleve a predecir y respaldar sus resultados.

Los criadores de ganado mayor específicamente bovinos para carne como el Charolee por mencionar alguno, llevan a cabo pruebas de comportamiento que van desde incremento de peso con tal o cual programa de alimentación, empleando x o y ingredientes en la formulación de la dieta y emplean como herramienta básica los diseños experimentales adoptando el modelo que justifiquen sus objetivos de evaluación y que involucren las variables que de una u otra manera afectan en tal estudio.

La estadística y la relación con el método científico

La estadística en su funcionamiento interno hace uso del método científico, cuando el método científico se aplica a fenómenos aleatorios, es la estadística la que auxilia al método científico.

El método científico consta de cuatro etapas fundamentales que ocurren sucesivamente.

Etapa 1. Colección de hechos por observación o experimentación.

Etapa 2. Formulación de hipótesis que explique los hechos en términos de relaciones de causa efecto. Es un supuesto que necesite ser comprobado.

Etapa 3. Por deducción, se determina que resultados surgen bajo ciertas condiciones.

Etapa 4. Verificación de las deducciones mediante nueva observación.

Si en la etapa cuarta no hay contradicción de la hipótesis, se efectúan nuevas deducciones bajo otras condiciones y se verifican por observación. Si una hipótesis que no ha entrado en contradicción con una extensa gama de hechos observados en condiciones variadas, se le denomina ley natural.

Si en la etapa cuarta existe contradicción entre la hipótesis y los nuevos hechos, la hipótesis se rechaza.

Se puede decir que el método científico sirve para rechazar hipótesis, pero no para aceptarlas. Algunos ejemplos del método científico: la teoría de la evolución de las especies de Darwin, las leyes de la herencia de Mendel, etc.

A continuación se discuten las contribuciones de la estadística en la aplicación de las cuatro etapas del método científico a los fenómenos aleatorios.

En la etapa 1 la estadística es un auxiliar valioso debido a tres aspectos.

a. Permite determinar la cantidad más conveniente de observaciones para estudiar el fenómeno. Se procura conciliar aspectos prácticos de la toma de observaciones con aspectos teóricos de los modelos. Además es posible tomar en cuenta limitaciones de recursos en el proceso. Dos ramas de la estadística tocan estos aspectos, muestreo y diseño de experimentos.
b. Una vez efectuadas las mediciones, la estadística permite los aspectos principales del conjunto de números o nombres que se han medido, facilitando así la comprensión y la comunicación del conjunto de observaciones.
c. Determina el grado de confiabilidad de las observaciones al evaluar los rangos de variación permisibles de acuerdo a las fluctuaciones aleatorias.

Durante la etapa 2 la estadística casi no aporta nada, la descripción de relaciones entre mediciones de varias modalidades del fenómeno, quizá ayuden al científico a plantear su hipótesis.

En la etapa 3 no hay tampoco una ayuda apreciable de la estadística. Una vez planteada la hipótesis es la lógica deductiva la que lleva a determinar que se espera y en que condiciones.

En la etapa 4 la hipótesis no se rechaza, si los hechos son observados concuerdan con los esperados según la deducción

En esta última etapa del método científico es muy importante el hecho de que las hipótesis son abstracciones e idealizaciones de la realidad, entonces no se espera una concordancia perfecta entre los hechos observados y los esperados por deducciones de la hipótesis.

Es de este modo como la estadística auxilia el método científico cuando se estudian fenómenos con un grado de aleatoriedad considerable.

De esta manera la estadística ha proporcionado metodologías útiles en muchas áreas de la ciencia y tecnología.

Para un uso efectivo de la estadística, esta debe conjugarse con los conocimientos de la materia a la cual se aplica.

Medición de fenómenos

Los hechos señalan una o varias propiedades de los fenómenos, determinando las modalidades que presentan dichas propiedades relevantes de los fenómenos. Este es el proceso de medición, puede presentar cinco alternativas que se reflejan en las denominadas escalas de medición.

1. Escala nominal.- Únicamente se les dan nombres a las diferentes modalidades de una propiedad que presenta el fenómeno bajo estudio.

2. Escala ordinal.- También se dan nombres a las modalidades que presentan las propiedades del fenómeno, lo que se debe reflejar en relaciones de orden.

3. Escala de intervalo.- A las modalidades en que se presenta la propiedad se le asignan números de modo que además de establecer un orden, también se tenga en orden entre las diferencias de las modalidades.

4. Escala de razón.- Son escalas para medir propiedades donde hay un estado o grado donde la propiedad no existe, a este estado se le asigna el numero cero.. Es el caso de alturas, pesos, distancias, etc.

5. Escalas absolutas.- Son las escalas donde no se puede cambiar nada; por ejemplo los conteos.

En el proceso de medición es importa distinguir dos tipos: a) medición fundamental – en la cual una propiedad de un objeto o fenómeno se cuantifica en la misma propiedad de otro objeto o fenómeno. B) medición derivada- la cual se basa en alguna ley natural acepta que relacione la propiedad por medirse con otra propiedad.

En todo estudio científico es necesario recurrir a alguna escala de medición.

Fenómenos aleatorios

Los modelos matemáticos constituyen abstracciones de la realidad, ya que solo presentan las propiedades relevantes para la hipótesis planteada, ignorando el resto de las propiedades presentes en los fenómenos. Además, estos modelos representan en forma aproximada a la realidad. Existen fenómenos donde no es posible establecer modelos matemáticos con una representación satisfactoria de la realidad; dichos fenómenos se denominan aleatorios.

Los fenómenos aleatorios se caracterizan porque no se puede determinar con suficiente precisión su estado final, sino que se tiene cierto grado de incertidumbre. Los fenómenos aleatorios pueden presentarse por una o varias de las razones siguientes.

1. No se puede conocer el estado inicial del fenómeno con suficiente precisión para determinar con base en leyes naturales conocidas, el estado final del mismo. Esto se puede deber a que:

a. Una variación muy pequeña en el estado inicial produce un cambio muy grande en el estado final. Es el principio en los juegos de azar.
b. El estado inicial es muy complicado y es impractico medirlo; por ejemplo situación microscópica de un gas.
c. Existe la posibilidad teórica de medir el estado inicial.

1. Con relación a las leyes naturales relativas al fenómeno se puede tener que:

a. Las leyes son conocidas pero complicadas por lo que en la practica no se aplican.

b. Las leyes naturales involucradas no se conocen lo suficientemente para hacer la predicción. Este es el caso de la mayoría de los fenómenos biológicos y sociales.

Todo fenómeno es aleatorio; pero si en la practica sé esta satisfecho con el grado de aproximación que producen las leyes conocidas, y su aplicabilidad es factible a ciertos fenómenos, estos no se consideran aleatorios. Existe un gran numero de fenómenos que son aleatorios, por ello es importante contar con alguna ayuda metodológica para el estudio de dichos fenómenos, esta ayuda metodológica es la estadística, que basándose en ciertos postulados probabilísticas, permite hacer descripciones, optimizaciones y predicciones en los fenómenos aleatorios.

Naturaleza básica de la probabilidad y estadística

Los fenómenos aleatorios, cuando se estudian un gran numero de veces es condiciones similares, presentan cierta regularidad denominada estadística.

La probabilidad es un modelo matemático que se usa para representar las proporciones a las que tienden las diferentes modalidades con las que ocurren los fenómenos aleatorios.

Con base a la estructura matemática de la probabilidad se construyen las llamadas funciones de distribución, las que permiten determinar con que probabilidad pueden ocurrir los diferentes valores o nombres con los que se midió una propiedad del fenómeno aleatorio.

La estadística esta encaminada básicamente a determinar cual es la función de distribución mas adecuada para representar un fenómeno aleatorio particular y comparar las distribuciones para algunas variantes del fenómeno.

La estadística puede asumir que por conocimientos previos sobre un fenómeno particular, se conoce la forma general de la función de distribución, solo se deben determinar los denominados parámetros para especificar totalmente esa función. A esto se le llama estadística paramétrica.

Cuando se conoce la forma de la función de distribución, se recurre a las llamadas pruebas de bondad de ajuste, donde se investiga si la información sobre el fenómeno se pueden considerar bien representadas por una función de distribución particular.

Los usos prácticos más importantes de la estadística radican en la comparación de algunas variantes de un mismo fenómeno. También se puede investigar como cambian algunas características de las funciones de distribución en relación a cambios cuantifícales en los factores que intervienen en el fenómeno.

El razonamiento básico de la estadística es considerar a un cierto modelo que pueda especificar relaciones entre mediciones de fenómenos aleatorios como una hipótesis.

La estadística en la mayoría de sus aplicaciones es de carácter empírico; esto es: No pretende determinar relaciones causa – efecto, únicamente describe, usando modelos estadísticos, o probabilísticas, las características de las modalidades de los fenómenos aleatorios y sus relaciones.

Hipótesis y decisiones

La difusión de la metodología estadística, así como la amplia disponibilidad de ordenadores y sofisticados programas de análisis de datos, permiten "producir" un número creciente de resultados, y sin embargo conviene que en algún momento nos detengamos a pensar en las características, y por qué no las debilidades, del método que estamos empleando. En opinión de muchos autores la omnipresencia

de los niveles de significación en la literatura clínica tanto humana como en veterinaria es totalmente desafortunada, dado que la práctica clínica se basará más en la magnitud de la diferencia observada que en el nivel de probabilidad, por lo que se aconseja el uso de los ***intervalos de confianza***, (Molinero, 2001)

En este artículo se revisa el concepto de prueba estadística de contraste de hipótesis, así como los errores asociados, justificando mediante ejemplos la utilidad de los intervalos de confianza. Asimismo se plantea la problemática de los estudios en los que se busca confirmar una hipótesis de igualdad y no su rechazo.

Pruebas estadísticas de contraste de hipótesis

Es habitual que en la investigación científica se utilicen hipótesis simples para explicar la realidad, debido a que son más fáciles de contrastar de manera empírica y de esa forma determinar su validez o deficiencia. En un ensayo típico el objetivo es estudiar la respuesta al tratamiento y compararla con la respuesta a una alternativa, ya sea ésta el tratamiento habitual o un placebo. La respuesta al tratamiento se determinará, en cada caso, en base a una medida numérica, por ejemplo incremento de peso en corderos de acuerdo a diferentes niveles de proteína contenidos en la ración proporcionada. La diferencia observada en la respuesta (medias para variables cuantitativas, proporciones cuando se trate de variables cualitativas) entre el grupo de tratamiento y el grupo control constituye una estimación de la efectividad del tratamiento.

Aunque en realidad la efectividad del nuevo tratamiento y del tratamiento control fuera la misma, es previsible que la diferencia observada no sea exactamente cero, aunque sí próxima a cero, siendo más improbables (más raros) los resultados a medida que se alejen de ese valor.

En el procedimiento clásico de contraste de hipótesis se denomina *hipótesis nula* a la que considera que ambos tratamientos son iguales y si, en el supuesto

de que sea cierta, la probabibilidad de que se observe una diferencia tan grande o mayor que la obtenida en nuestro estudio es muy baja (típicamente 0.05, aunque hay autores más exigentes que sugieren otros valores como 0.01) se rechaza dicha hipótesis y se acepta la contraria, denominada *hipótesis alternativa*, que establece que ambos tratamientos son diferentes.

A pesar de la amplia aceptación en las publicaciones científicas de esta metodología desde sus orígenes, existe un gran debate respecto a su validez metodológica y sobre todo respecto a su aplicación rutinaria, controversia que va calando, aunque de forma muy lenta, en la comunidad investigadora frente a la inercia de su aplicación como meras recetas para la toma de decisiones.

Estudios de equivalencia

En primer lugar hay que destacar que la hipótesis nula nunca puede ser demostrada o establecida, siendo posible sólo, con esta metodología, refutarla mediante un experimento. Nos encontramos entonces en una situación paradójica, ya que en algunos estudios lo que se busca es precisamente demostrar una igualdad, Molinero, 2001.

Por ejemplo cuando se pretende demostrar que un tratamiento menos agresivo que el tradicional es igualmente eficaz, o cuando se pretende demostrar por ejemplo que un garrapaticida tradicional es igualmente eficaz a uno nuevo en el mercado con la misma sustancia activa al primero.

Probabilidad de error en un contraste de hipótesis

Es evidente que en el contraste estadístico de hipótesis se pueden dar dos posibles errores. El denominado **error tipo I o error alfa**, que es el que se produce cuando se rechaza la hipótesis nula y en realidad es cierta. La

probabilidad de cometer este error se fija de antemano por el investigador cuando sitúa el nivel de rechazo, habitualmente 0.05.

Si no se rechaza la hipótesis nula, cuando el valor de probabilidad es inferior al nivel fijado, también se corre el riesgo de cometer un error que se denomina **error tipo II o error beta ß**. Ahora las cosas no son tan sencillas, la probabilidad de cometer un error tipo II no es un valor único como la que corresponde al error tipo I. La probabilidad de un error tipo I se calcula suponiendo que la hipótesis nula (no existen diferencias) es correcta, mientras que la probabilidad de un error tipo II se tiene que calcular cuando ésta es falsa, es decir cuando existen diferencias entre los tratamientos. Pero la magnitud **D** de esa diferencia puede tomar en principio cualquier valor y la probabilidad de error dependerá de esa magnitud. Hay que fijar pues la diferencia para la que se desea acotar ese error. Habitualmente se utilizará un valor a partir del cual se puede considerar como diferencia relevante en términos del proceso en estudio.

Inconvenientes de las pruebas de contraste de hipótesis

El principal problema de las pruebas estadísticas de contraste de hipótesis radica en que la decisión de rechazar o no la hipótesis de igualdad descansa en el tamaño de muestra, ya que una diferencia muy pequeña puede ser estadísticamente significativa si la muestra es suficientemente grande y por contra, una diferencia de magnitud relevante puede no llegar a ser estadísticamente significativa si la muestra es pequeña. Aquí es donde entra en juego el concepto de significancía, para las áreas biológicas o el término que corresponda en cada campo de aplicación) frente al de significación estadística, que pone de relieve que lo importante no es demostrar que existen diferencias.

Otro problema a subrayar es la evidente arbitrariedad a la hora de fijar el punto de corte y el planteamiento maximalista del todo o nada.

Veamos un ejemplo. En la siguiente tabla se resumen los datos obtenidos en un estudio con becerros de una semana de edad, criado en diferentes condiciones de becerreras.

	Media	Desv. típica	Tamaño muestra
Tratamiento	128	9	20
Control	123	8	20

Como vemos la diferencia entre las medias en ambos grupos es de 5. Supongamos que una diferencia de **4** o más se considera de importancia. Existe pues una diferencia importante entre los dos grupos.

Si se realiza un contraste de hipótesis basado en la *t de Student*, se obtiene un valor de probabilidad para una diferencia igual o mayor que la observada de 0.07, que es inferior al nivel de rechazo universalmente aceptado de 0.05. Por lo que aplicando la "doctrina" estadística no se rechaza la hipótesis de igualdad de los tratamientos.

Sin embargo se trata de una diferencia de importancia y es cierto que también una probabilidad de 7 entre 100 es bastante baja.

Intervalos de confianza

A pesar de la inercia intelectual que nos conduce la aceptación universal del uso de pruebas de hipótesis como un asunto de todo o nada, hay otras alternativas que no se quedan en la mera receta: el uso de los intervalos de confianza. Esta alternativa ha sido defendida vehementemente por diferentes autores y secundada por diferentes editores de revistas médicas. El intervalo de

confianza nos da el margen de valores en los que es previsible esperar que se encuentre la verdadera diferencia entre los tratamientos, para una probabilidad dada (habitualmente el 95 %). En realidad se sustenta sobre la misma teoría, pero el enfoque es mucho más expresivo, proporcionando información y no sólo documentando una mera decisión, como es el caso del contraste de hipótesis.

En nuestro ejemplo, el intervalo de confianza va aproximadamente desde -0.5 a 10.5, lo que revela que es insuficiente para descartar el 0 (igualdad de tratamientos) con toda certidumbre, pero se encuentra muy escorado hacia el lado positivo llegando a valores de gran importancia, por lo que de acuerdo con otros datos (por ejemplo de costo, efectos secundarios, etc.) proporcionan una mayor información a la hora de determinar el interés del nuevo tratamiento.

Es frecuente manejar un concepto relacionado con el error tipo II o ß, el de potencia de una prueba estadística. Para una diferencia D dada, la probabilidad de no cometer un error de tipo II, es decir la probabilidad de detectar una diferencia D en el estudio es

1 - ß

Es evidente que el ejemplo anterior no puede ser aceptado como prueba de que no existen diferencias entre los tratamientos, por el mero hecho de que no se encontrase diferencias estadísticamente significativas. Si con un estudio mucho mayor los resultados hubieran sido:

	Media	**Desv. típica**	**Tamaño muestra**
Tratamiento	128	9	200
Control	127.5	8	200

El intervalo de confianza es ahora (-1.2 , 2.2), que engloba el cero y que no incluye valores relevantes, lo que nos aporta más evidencia para aceptar una eficacia similar de ambos tratamientos.

Si se calcula la potencia de la prueba en el primer ejemplo con dos muestras de tamaño 20, desviaciones típicas 9 y 8, para detectar una diferencia de 4, veríamos que es 0.297, a todas luces insuficiente.

Por el contrario, la potencia de la prueba para los mismos valores anteriores salvo el tamaño que sería ahora de 200 es de 0.9967.

Y finalmente, para complicar las cosas supongamos que el resultado observado es el de la siguiente tabla:

	Media	**Desv. típica**	**Tamaño muestra**
Tratamiento	128	9	200
Control	126	8	200

Se puede comprobar que un contraste mediante la *t de Student* resulta estadísticamente significativo con p = 0.019. Sin embargo la magnitud de la diferencia 2 puede no ser de importancia, dado que comentábamos que se consideraban relevantes diferencias iguales o superiores a 4. Una vez más el intervalo de confianza puede ser más ilustrativo que el simple valor de p, en este caso es 0.3, 3.7, no llegando a incluir el valor considerado relevante, aunque se aproxima.

La importancia, en un sentido o en otro, de estos resultados solo se podrá considerar a la luz del proceso de manejo concreto en que se analiza y no con meros números como aquí se ha expuesto, lo que recalca el absurdo de la utilización cómoda y simplista de las "recetas estadísticas", siendo esta metodología una herramienta más para ayudar en la investigación y no estando exenta ella misma de controversia.

METODOLOGÍA DE LA INVESTIGACIÓN

1.-**Diseños experimentales.**

Características de los diseños experimentales: validez interna, externa, de constructo y estadística. Diseños experimentales con grupos de sujetos distintos. Diseños experimentales con los mismos sujetos. Diseños factoriales.

Características de la metodología experimental

Definición:

Un experimento es un estudio en el que al menos una variable es manipulada y las unidades son aleatoriamente asignadas a los distintos niveles o categorías de las variables manipuladas. (Pedhazur y Pedhazur, 1991).

Por ejemplo si pretendo obtener la proporción de crecimiento compensatorio que se ha de presentar en pollos de engorda después de manipularlo mediante un estrés causado por restricción al acceso al alimento en horas en las primeras cuatro semanas la variable que estoy manipulando es el estrés causado por la restricción del alimento.

Características del diseño experimental:

1. Manipulación: es la intervención deliberada del investigador para provocar cambios en la v. dependiente.

2. Aleatorización: mayor tamaño de los efectos frente a la equiparación.

Todos los diseños experimentales se caracterizan por la manipulación, pero pueden ser clasificados atendiendo a la aleatorización en:

- Auténticamente experimentales.
- Cuasiexperimentales.

En los diseños experimentales la aleatorización es como se distribuyen los sujetos en los diferentes grupos que forman parte del estudio. El primer ensayo clínico aleatorizado se efectuó en 1947 por Sir Austin Bradford Hill y lo llevó a cabo sobre el efecto de la Estreptomicina en la Tuberculosis, es el primer estudio realizado con un diseño experimental, hasta ese momento el diseño investigador que se realizaba era el “estudio de casos”, estudios observacionales simples.

La aleatorización mide y reduce el error

En las Ciencias biológicas como es tan importante estudiar los efectos que produce una variable, sus consecuencias y la relación causa-efecto que se puede producir, es muy importante conocer el error y reducirlo en todo lo posible, por ello los estudios de investigación deben ser y deben reunir la característica de la aleatorización, por ello deben utilizarse diseños experimentales.

Ejemplo: Estudio de incidencia de **Ascítico** en pollos de engorda. Para llevarlo a cabo se tomarían dos grupos de aves (parvadas) que deberán reunir idénticas características en cuanto al mismo número de individuos que lo componen, criados a la misma altura sobre el nivel del mar, y con la misma cantidad y calidad de alimento proporcionado, edad en que se evalúa, etc., posteriormente se procedería a la comparación e investigación sobre la presencia de fluidos en la parte del abdomen y saco vitelina (patología característica de ascitis).

Ventajas del diseño experimental

1. Se elimina el efecto de las variables perturbadoras o extrañas, mediante el efecto de la aleatorización, por ejemplo al evaluar un suplemento alimenticio en corderos se ha de distribuir en corraletas con el mismo numero de animales para cada tratamiento y repetición y de la misma edad y peso de tal manera que tengan la misma oportunidad de acceder al alimento.

2. El control y manipulación de las variables predictorias clarifican la dirección y naturaleza de la causa.

2. Flexibilidad, eficiencia, simetría y manipulación estadística.

Viabilidad de los diseños experimentales

1. Imposibilidad de manipular algunas variables.

2. Cuestiones éticas.

3. Practicabilidad.

Inconvenientes del diseño experimental.

1. Dificultad de elegibilidad y manejo de las variables de control.

2. Dificultad de disponer de muestras representativas.

3. Falta de realismo.

Calidad del diseño experimental

1. Validez Interna.
2. Validez Externa.
3. Validez Ecológica.
4. Validez de Constructo.

1.- Validez interna.

Es el grado en que los cambios observados se pueden atribuir a la manipulación experimental. Estudia hasta que punto una causa puede ser atribuida a un efecto. Ej.: Ensayo clínico: tiene el máximo grado de validez interna.

Teniendo en cuenta la validez interna de mayor a menor grado los diseños los podemos clasificar en los siguientes grupos:

1. **Experimentales auténticos:** Verdaderos, puros, pues no tienen problemas de validez interna (True Desing).
2. **Cuasiexperimentales:** No se pueden descartar la presencia de variables confundidoras, pues no es posible eliminarlas todas. El investigador sabe que A es causa de B, pero no está seguro que A también pueda ser causa de otros factores como C ó D.
3. **No experimentales:** Están cerca de los anteriores en cuanto a validez interna, aunque presentan más variables confundidoras, pueden ser:

    3.1 Longitudinales: (Prospectivo / Retrospectivo)

    3.2 Transversales.

Cuántas más variables entran en un diseño van restando validez interna.

Las variables confundidoras afectan al diseño, forman parte de las **AMENAZAS** a la validez interna.

Ej. : Incremento de peso en becerros a los 3 meses, evaluando dos grupos:

1.- Con becerros a los que se les proporciono suplementación.

2.- Con becerros a los que no se les dio suplemento.

A unos les proporciona suplemento y a otros no.

Si posteriormente analizamos cuales presenta mayor incremento de peso y si nuestra hipótesis es que aquellos suplementados reflejaran mayor incremento de peso, también deberemos tener en cuenta que hay variables que han influido, por ejemplo genética de los animales, calidad del alimento proporcionado, etc.

- Cuando una variable hace que un grupo de partida sea diferente (por ej. Calidad genetica de los animales) se les llama AMENAZA.

Amenaza a la validez interna.

1. Historia. Hay amenaza de historia, cuando hay acontecimientos externos que ocurren simultáneamente con éste y que pueden alterar o influir.

2. Selección. Cuando los grupos de estudio son diferentes. Ej. Raza, sexo.

3. Maduración. Son los cambios producidos por evolución natural. Tiene relevancia en salud y confunde el efecto del cambio de la variable con el de la causa. Ej. : Herida mejora hagamos o no hagamos nada, pero

¿cuánto depende la mejoría de la herida de lo que hemos hecho sobre ella?.

4. Efectos relativos al pretest. Es la influencia que produce el pre-test.
5. Mortalidad (o astricción) El que desaparezcan sujetos de los grupos de comparación. No sabemos que sujetos se pierden.
6. Instrumentación. Uso de instrumentos no fiables ni validos.
7. Regresión estadística. Los sujetos seleccionados representan situaciones o puntuaciones en alguna variable. Cuando se usan sujetos extremos. Sucede cuando para probar los efectos algo se escogen a los sujetos más extremos. Ej.: Para probar los efectos de una dieta seleccionamos a los más gordos.

El Tamaño muestral afecta a la validez interna.

2.- Validez Externa.

Es el grado en que los resultados de un estudio pueden ser generalizados a muestras o condiciones espacio-temporales diferentes. Ej. "A" causa "B", pero seguiría causando "B" con otros:

- Sujetos.
- Contexto ---- validez ecológica.
- Momentos.

Los estudios descriptivos (encuestas) son los que más se preocupan por la validez externa.

La validez externa está afectada por los siguientes aspectos:

- Por la variable independiente. Es el nivel de operacionalización del v. Independiente.

- "Efecto Rosenthal": es el efecto derivado de las expectativas, es decir, el efecto derivado de que se presupone o se espera que ocurra, cuando algo se espera un efecto favorece que se produzca. Afecta tanto a la variable interna como a la v. externa.

- "Efecto Hawthorne": son las expectativas que el individuo se le tiene contemplado.

En el Efecto Rosenthal las expectativas se reflejan en el otro individuo, mientras que el Efecto Hawthorne es el producido por las expectativas del individuo sobre sí mismo.

3.- Validez Ecológica.

Es aquella que se puede aplicar en distintos contextos. Ej. La conversión alimenticia de pollos de engorda de la misma línea (Ross) evaluados durante 6 semanas en que se sacan al mercado criados en diferentes regiones no son iguales, por lo tanto lo que allí es válido puede no serlo aquí.

4.- Validez de constructo.

Alude a la relación existente entre la v. independiente que se manipula y el constructo teórico que se supone se manipula.

Representa principalmente dos amenazas:

1. Problemas en la definición operacional del constructo.

8. Poco desarrollo teórico del constructo.

Tipos de diseños Experimentales.

En todos hay manipulación, luego la clasificación se llevará a cabo en relación al grado de aleatorización. Se pueden distinguir dos grandes grupos:

1. Experimentales AUTENTICOS. Hay manipulación y aleatorización. Hay dos tipos básicos: * Con realización de medición “pre-test” y * Sin realización de medición “pre-test”.

9. Cuasiexperimentales o pre-experimentales. Hay manipulación pero no hay aleatorización.

Diseño Experimental Autentico.

Presenta dos características importantes:

- Manipulación: es la intervención deliberada del investigador para provocar cambios en la v. dependiente.

- Aleatorización: mayor tamaño de los efectos frente a la equiparación.

Es aquel en el cuanto más aleatorización haya mejor.

El efecto del azar: cuando la muestra aleatoria es grande, el tamaño del efecto es alto.

Ej.: - Si tenemos 30 animales y queremos distribuirlos al azar, se podría hacer sorteándolos con una moneda, pero si tenemos 100 animales, habría más probabilidades de que la muestra sea al azar.

Hay cálculos y sistemas para conocer el número de estudios que se necesitan, para poder afirmar que es una muestra aleatoria.

Clasificación de los diseños

En función de la variable independiente:

- Diseños simples.
- Diseños factoriales.

En función de la aplicación:

- Diseños experimentales con grupos de sujetos distintos.
- Diseños experimentales con los mismos sujetos.

En función de las variables dependientes:

- Diseños de medidas repetidas.

Diseño experimental con grupos de sujetos distintos.

1.- Diseños de grupos aleatorios o independientes.

1.1 Diseño de dos grupos elegidos al azar con medidas en el post-test.

1.2 Diseño de dos grupos elegidos al azar con medidas en el pre-test y en el post-test.

Sólo interfiere la variable independiente.

En las investigaciones con animales la variable independiente (X) suele ser un estimulo que provocará una respuesta.

Hay una variable endógena que hay que controlar.

2.- Diseños de grupos aleatorios por bloques.

Intenta controlar una variable contaminadora, esto lo hace incluyéndola en todos los grupos y convirtiendo la variable contaminadora en una constante.

Modalidades:

2.1. Varios sujetos por nivel y bloque.

2.2. Un sujeto por nivel y bloque.

2.3. Diseños con “camadas”.

2.4. Diseño con misma raza, de la misma línea .

3. - Diseños especiales:

3.1. Control con placebo (ciego).

3.2. Estudio del doble ciego.

Diseño experimental con los mismos sujetos.

Son llamados también “Diseños Infrasujeto o de medidas repetidas”.

El grupo control y experimental están formados por los mismos sujetos. Los mismos sujetos son control e investigados.

“Todos los sujetos pasan por todas las condiciones experimentales”.

Presentan principalmente dos ventajas:

- Economía de individuos.
- Pequeño número de animales del grupo en estudio .

Amenazas a la validez interna:

- Efectos de la práctica sobre la variable dependiente.
- Efecto de orden en la presentación de tratamientos y medición de resultados.
- Efectos de la fatiga.
- Efectos de la motivación.

Procedimientos de control:

1. Aleatorización simple y por bloques.
2. Equilibrado o reequilibrado. Ej.: V.I con dos niveles A y B.
   - Equilibrado: ABBA (A afecta a B, B afecta A).
   - Alternativa (Dos grupos): grupo 1: orden: AB y grupo 2: orden: BA.
3. Cuadrado latino: sustituye al equilibrado cuando la variable independiente tiene más de dos niveles. Ej. 3 niveles: ABC, ACB, BAC, BCA, CAB, CBA (Permutaciones de 3 elementos: 3 X 2 X 1 = 6)

   Definición: Elección de permutaciones al azar sin que se repita posición del nivel de la VI.

   1. – ABC 10 sujetos al azar (con n = 30)

2. - ACB

3. – BAC 10 sujetos al azar.

4. - BCA

5. – CAB 10 sujetos al azar.

6. - CBA

No exclusivamente intrasujeto, sino también Inter.-grupo.

Es un diseño idóneo para el estudio de enfermedades crónicas y enfermedades que afectan a grandes grupos de población. Muy utilizado en los ensayos clínicos.

Diseños complejos o diseños factoriales.

Es un tipo de diseño experimental en el que hay más de una variable independiente. Cada variable recibe el nombre de factor. Su principal acción es que sirven para valorar el efecto de la interacción, es decir, saber el efecto combinado de las distintas variables. Cada variable recibe el nombre de factor y el número indica los niveles de cada variable.

Ejemplo: 2X2 (dos variables independientes con dos niveles cada una)

2X2X3 (tres variables independientes, dos de ellas con dos niveles y una con tres).

Ejemplo de un diseño complejo o factorial:

Hipótesis: **Las personas que son distraídas, frente a las que no lo son, aguantan más el dolor. (Meter la mano en agua helada).**

Tenemos 2 v. independientes que tienen dos niveles:

- **Distracción (se consigue mediante la lectura de un cuento).- Con cuento (distracción) y - Sin cuento (sin distracción).**
- **Sexo del investigador:- Hombre.- Mujer.**

Se forman dos grupos: uno experimental y otro de control.

Y se plantea realizar una estrategia distractiva: leer un cuento mientras se realiza la prueba.

Diseño del experimento: 2 grupos, uno con distracción y al otro sin distracción (no lectura del cuento), y mido el tiempo que aguanta cada uno con la mano sumergida en agua helada.

La investigadora pensó que ella misma podía ser un elemento de distracción y entonces añadió una variable de confusión que era el "atractivo" de la propia investigadora, pasando el estudio a ser de dos variables y por lo tanto se necesitaban cuatro grupos.

Las dos variables tenían efecto sobre el dolor, tanto con el entretenimiento como con la presencia del investigador.

No hubo efecto combinado de potenciación entre las variables.

CARACTERÍSTICAS DE LOS DISEÑOS COMPLEJOS O FACTORIALES.

1. Un diseño complejo es mejor que dos diseños simples, ya que es el único que permite observar el comportamiento de una variable bajo todas las condiciones. PERMITE VALORAR EL EFECTO DE INTERACCIÓN (el efecto combinado de ambas variables), es decir, permite saber el efecto principal de A, el de B y el efecto combinado de ambos).

2. A más niveles en variables mejor se rastrea la relación causal, pero presenta el Inconveniente de necesitar más sujetos. Cuanto más aumenta el nivel de las variables, más aumenta la cantidad de sujetos que se necesitan.

Para garantizar un buen resultado hay que tener por lo menos 10 sujetos por grupo, si tengo un diseño de: 7. x 9, necesitaré 630 sujetos.

Experimentos o ensayos secuenciales

(Todos los vistos anteriormente pueden serlo)

Características:

Es un estudio en que los sujetos se asignan a los grupos poco a poco hasta tener todos los efectos que esperamos.

En el mundo de la química no hay mas remedio que tener un grupo control y otro experimental. Esto puede plantear problemas éticos.

En estos estudios de antemano no asignamos los individuos a los grupos, cuando el tratamiento es efectivo, se para el estudio para poder aplicar dicho tratamiento al resto de pacientes, no hace falta acabar el estudio. Así mismo si vemos que el tratamiento no es efectivo y está perjudicando al grupo experimental, se para no perjudicar más al grupo.

Por lo tanto podemos afirmar que los estudios están condicionado a los resultados que se van obteniendo.

En estos experimentos, estudios o ensayos la continuidad no está asegurada, sino que depende y está condicionada por los resultados que van apareciendo.

Utilización e Interpretación de las Técnicas de Regresión y Correlación

Estos métodos se emplean para conocer las relaciones y significación entre series de datos.

Cuando, simultáneamente, contemplamos dos variables continuas, aunque por extensión se pueden emplear para variables discretas cuantitativas, surgen preguntas y problemas específicos. Esencialmente, se emplearán estadísticos descriptivos y técnicas de estimación para contestar esas preguntas, y técnicas de contraste de hipótesis específicos para resolver dichos problemas. La mayoría de estos métodos están encuadrados en las técnicas regresión y correlación. En este artículo comentaremos las técnicas bivariantes lineales.

Si se parte de un modelo en el cual una de las dos variables continuas es dependiente o respuesta (y) y la otra es independiente o explicativa (x), surgen nuevos estadísticos para describir los datos.

La nube de puntos, o el diagrama de dispersión, resultante de la representación gráfica de los datos está "concentrada" en la recta de regresión de mejor ajuste obtenida por el método de mínimos cuadrados. Una condición previa, en las técnicas lineales, es que la nube de puntos debe tender a la linealidad (en sentido rectilíneo, se entiende). Los coeficientes de la regresión lineal, la ordenada en el origen (a) y la pendiente de la recta (b), son estadísticos muestrales. Se suelen presentar de la forma $y' = a + bx$.

La dispersión de los puntos alrededor de la recta de mejor ajuste es una característica de los datos bidimensionales que merece cuantificarse. El estadístico correspondiente es la desviación típica de los residuos. Es posible obtener la distribución de los residuos. Estos son las distancias en vertical de cada punto a la recta de regresión. Su medida es cero (esta propiedad es compartida por otras muchas rectas de ajuste, además de por la de mejor ajuste, que es la nuestra), y su desviación típica es el estadístico de elección para describir la dispersión alrededor de la recta. Sus unidades son las de la variable dependiente (y).

Es posible, que estudiando una variable bidimensional, no se desee establecer ninguna relación de subordinación de una variable con respecto a la otra. En este supuesto, se intenta cuantificar la asociación entre las dos características.

Entramos en las técnicas de correlación lineal. Es posible definir otro estadístico muestral a partir del las dos pendientes teóricas de las dos posibles rectas de regresión (y) sobre(x) y de (x) sobre (y). Este estadístico es el coeficiente de correlación r. Su cuadrado r^2 es el coeficiente de determinación y da una medida entre 0 y 1 de la cantidad de información compartida por dos características o variables continuas en los datos muestrales.

La magnitud de la asociación entre dos variables continuas está en relación con la dispersión de la nube de puntos. Se puede establecer una relación matemática perfecta entre la desviación típica de los residuos y el coeficiente de determinación.

El hecho de que dos variables estén correlacionadas, e incluso que lo estén con valores muy cercanos a 1, no implica que exista una relación de causalidad entre ellas. Se pueden producir correlaciones espurias (causales) entre dos variables, por estar ambas relacionadas con otra tercera variable continua y anterior en el tiempo.

Los nuevos estadísticos generados en la regresión y correlación lineal se emplean como estimadores de los correspondiente parámetros poblacionales. Para que los coeficientes de la regresión y correlación sean estimadores adecuados (centrados y de mínima varianza) de sus correspondientes parámetros poblacionales, es necesario que se asuman ciertas condiciones en la población de origen, referidas fundamentalmente a las distribuciones de los residuos:

a).- Que la medida de los residuos sea cero.

b).- Que su varianza sea similar (homogénea) a lo largo de la variable (x): homocedasticidad.

c).- Que estén normalmente distribuidos y que sean incorelacionados.

Un buen estimador de la desviación típica de los residuos es la cuasi-cuasidesviación típica de los residuos muestrales Sn-2 , e.

Asimismo, este estadístico juega un papel relevante en la construcción de los intervalos de confianza de los coeficientes de la regresión y correlación poblacionales.

Es posible también, formularse hipótesis relativas a estos parámetros, cuya resolución se basa en la construcción de estadísticos de contraste que sigan funciones de probabilidad conocidas, como son la T-Student y la Chi-cuadrado.

Por ejemplo se sabe que la calidad del forraje en Zacate Guinea (*Panicum maximum Jacq.*) que es una gramínea tropical, esta influenciada por la frecuencia de corte por lo que para evaluar la relación y correlación que este tiene con la calidad de la misma expresada en contenido de PC (%) el modelo factible es la Regresión lineal ya que esta nos lleva a obtener hasta el grado de predecir la frecuencia de corte (intervalo en días) optima para cosechar un buen forraje. Lo anterior suena lógico ya que como a medida que la planta crece y envejece por llamarlo de alguna manera, esta tiende a lignificar y y existe correlación negativa entre la lignificación de la planta y el contenido de proteína, es decir a mayor lignificación, menor contenido de proteína y viceversa.

Diseño de bloques completamente al azar (DBCA)

Este tipo de diseño se puede utilizar cuando una fuente identificable de variación y las unidades experimentales se puedan agrupar de acuerdo a esta fuente. El objetivo de agrupar es el de mantener las unidades en un bloque lo mas uniforme posible, de manera que las diferencias que surjan se deban en gran parte a los tratamientos. Las fuentes de variación, se deberán identificar y distribuir los bloques de acuerdo a esta variación. Los bloques, también llamados repeticiones, deberán consistir de un grupo de localidades homogéneas.

Distribución

Por ejemplo, al evaluar el incremento de peso en corderos alimentados con una ración con aporte de cuatro diferentes cantidades de proteína (14, 16 y 18 % de PC) y que serán los tratamientos, la distribución de los animales se hará al azar, es decir si por ejemplo la población de corderos a evaluar es un total de 60 animales de un mismo sexo, de la misma edad y con el mismo peso y se quiere hacer un arreglo con cuatro repeticiones para tener una mayor veracidad se repartirá sin contemplaciones en localidades de 5 animales, la aleatorización es con el propósito de que todos tenga la misma posibilidad de acceder al alimento en evaluación y reducir al máximo el sesgo siempre latente en un estudio.

Covarianza

La covarianza entre dos variables es un estadístico resumen indicador de si las puntuaciones están relacionadas entre sí. La formulación clásica, se simboliza por la letra griega sigma (σ_{xy}) cuando ha sido calculada en la población. Si se obtiene sobre una muestra, se designa por la letra "**S_{xy}**".

Este tipo de estadístico puede utilizarse para medir el grado de relación de dos variables si ambas utilizan una escala de medida a nivel de intervalo / razón (variables cuantitativas).

Ejemplificando, al evaluar el incremento de peso en becerros cuando la dieta es la misma es decir con los mismos ingredientes pero con diferentes niveles energéticos (2300 y 2400 kcal) respectivamente pero a la vez con dos diferentes aportes de proteína (16 y 18 % de PC) y diferentes digestibilidades, el modelo factible a emplear es el de Covarianza, ya que con ello puedo evaluar con un solo modelo las diferentes variables que intervienen o que inciden en mi objetivo que en este caso es la ganancia de peso.

La formula suele aparecer expresada como:

$$\hat{S}_{xy} = \frac{\sum_{i=1}^{n} x_i\, y_i}{n-1} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})\,(Y_i - \bar{Y})}{n-1}$$

La expresión se resuelve promediando el producto de las puntuaciones diferenciales por su tamaño muestral (n pares de puntuaciones, n-1 en su forma insesgada).

Este estadístico, refleja la relación lineal que existe entre dos variables. El resultado numérico fluctúa entre los rangos de +infinito a -infinito. Al no tener unos límites establecidos no puede determinarse el grado de relación lineal que existe entre las dos variables, solo es posible ver la tendencia.

1. Una covarianza positiva significa que existe una relación lineal positiva entre las dos variables. Es decir, las puntuaciones bajas de la primera variable (X) se asocian con las puntuaciones bajas de la segunda variable (Y), mientras las puntuaciones altas de X se asocian con los valores altos de la variable Y.
2. Una covarianza de negativa significa que existe una relación lineal inversa perfecta (negativa) entre las dos variables. Lo que significa que las puntuaciones bajas en X se asocian con los valores altos en Y, mientras las puntuaciones altas en X se asocian con los valores bajos en Y.
3. Una covarianza 0 se interpreta como la no existencia de una relación lineal entre las dos variables estudiadas.

BIBLIOGRAFÍA

Molina, L. M. 2001. Alce ingeniería. http://www.seh.lelh-lelhha.org/stat1.htm.2001

Martínez, O. 2001. Eficiencia de los diseños experimentales usados en algodón. 2002. Agronomía Tropical. 20 (2): 81-95.

http://www. redpav-fpolar-info.ve/agrotrop/v20_2/v2

Diseño Investigación II. 2002.

http:/perso.wanado.es/aniorte_nic/apunt_meto_invesigac4_5ttm

http://www.ur.mx/ur/veritas/dolores.htm

http://www.udec.cl/panorama/p385/p18.htm

http://www.unanleon.edu.ni/~vitro/sigatoka.htm

http://www.uniovi.es/UniOvi/Apartados/Departamento/Psicologia/metodos/tutor.6/fcova.html

ANALISIS DE COVARIANZA

Introduccion

Covarianza significa variación simultánea de dos variables correlacionadas; su valor se expresa por:

$$\text{Covarianza} = \frac{\sum xy}{n-1} = \frac{\sum (Xi - \overline{xi})(Yi - \overline{yi})}{G.L.}$$

El análisis consiste en separar las diversas causas de variación de cada variable y de la variación conjunta. Fundamentalmente, se dan los siguientes pasos.

1. Análisis de varianza para la variable *X.*
2. Analisis de varianza para la variable *Y.*
3. Cálculo de *b* y *x.*
4. Obtención de la ecuación de regresión y ajustes de los promedios de la variable dependiente *Y.*

Cuando se aplica para corregir por un diferente número de plantas, el ajuste se hace como si todas las unidades experimentales contaran con igual número. La técnica también se puede aplicar en aquellos experimentos agropecuarios en que los animales tienen un peso inicial diferente, siendo la variable independiente *X* dicho peso inicial y la ganancia en peso, después de aplicar los tratamientos, la variable dependiente *Y:*

La teoría es compleja, pero su aplicación es relativamente sencilla y la mayor eficacia para ajustar valores de los casos citados anteriormente.

Análisis de covarianza para una distribución en bloques al azar

X = número de plantas; *Y* = producción

i = 1 ... a tratamientos

j = 1 ... n repeticiones

tabla 0.1

	Repeticiones							
Tratamientos	*I*		*II*		*III*			
	X	*Y*	*X*	*Y*	*X*	*Y*	*Xi*	*Yi*
1	*Xi*	*Yij*	*Xij*	*Yij*	*Xij*	*Yij*	*X1*	*Y1*
2								
3								
a								
Xi: Yi	*X.*1	*Y.1*	*X.2*	*Y.2*	*X.3*	*Y.3*	*X ..*	*Y..*

1. F.C. = $\frac{X^2..}{an}$
2. S.C.$_{total}$ = $\sum Xij^2 - F.C.$

3. S.C. $_{bloques}$ = $\frac{\sum X.j^2}{a}$ - F.C.

4. S.C. $_{tratamientos}$ = $\frac{\sum Xi.^2}{n}$ - F.C.

5. S.C. $_{error}$ = S.C. $_{total}$ - (S.C. $_{tratamientos}$ + S.C. $_{bloques}$)

Procedimiento para calcular la S.C. para *Y*

1. F.C. = $\frac{Y^2..}{an}$

2. S.C. $_{total}$ = $\sum Y^2ij - F.C. = \sum y^2$

3. S.C. $_{bloques}$ = $\frac{\sum Yj^2}{n}$ - F.C = $\sum y^2B$

4. S.C. $_{tratamientos}$ = $\frac{\sum Yi.^2}{n}$ - F.C.= $\sum y^2T$

5. S.C. $_{error}$ = S.C. $_{total}$ - (S.C. $_{tratamientos}$ + S.C. $_{bloques}$)= $\sum y^2E$

Procedimiento para calcular la $\sum xy$

1. F.C. = $\frac{X^2..X^2..}{an}$

2. $\sum xy$ $_{total}$ = $\sum XijYij - F.C.$

3. $\sum xy$ $_{bloques}$ = $\frac{\sum XjYj}{a}$ - F.C

4. $\sum xy$ $_{tratamientos}$ = $\frac{\sum Xi.Yi.}{n}$ - F.C.

5. $\sum xy$ $_{error}$ = $\sum xy$ $_{total}$ - ($\sum xy$ $_{tratamientos}$ + $\sum xy$ $_{bloques}$)

Cálculo del coeficiente de regresión

$$b_{xy} = \frac{\sum xy_{error}}{\sum x^2_{error}} = \frac{\sum xy_E}{S.C._E}$$

Hacer prueba de *F* para la variable *X* , para la variable *Y* y para la covarianza

Causas	*G.L*	*S.C.x*	$\sum xy$	*S.C.y*	*Valores ajustados* G.L.	S.C.y	C.M.
Total	*(an-1)*	*S.C. total*	$\sum xy_{total}$	$\sum y^{2`}$			
Bloques	*(n-1)*	*S.C. bloques*	$\sum xy_{total}$	$\sum xy^2_{bloques}$			
Tratamientos	*(a-1)*	*S.C. total*	$\sum xy_{total}$	$\sum xy^2_{total}$			
Error	*a-1)(n-1)*	*S.C. error*	$\sum xy_{errorl}$	$\sum xy_{error}$	*(a-1)(n-1)-1*	$\sum xy^2 - \frac{(\sum xy_{error})^2}{S.C._{ERROR}}$	S^2_{yx}
T + E	*n (a-1)*	*S.C. TE*	$\sum xy_{TE}$	$\sum xy^2_{TE}$	*n(a-1) – 1*	$\sum y^2_{TE} - \frac{(\sum xy_{TE})^2}{s.c._{TE}}$	
Tratamientos ajustados					*a – 1*	*S.C.*	*C.M..*

$$S.C._{tratamientosajustados} = \left[\sum y^2_{TE} - \frac{(\sum xy_{TE})^2}{S.C._{TE}}\right] - \left[\sum y^2_{TE} - \frac{(\sum xy_{error})^2}{S.C._{error}}\right]$$

$$S^2_{yx} = \frac{\sum y^2_{error} \frac{(\sum xy_{error})^2}{S.C._{error}}}{(a-1)(n-1)-1}$$

1.1 Introducción.

El análisis de la covarianza trata de dos o más variables medidas y donde cualquier variable independiente mediable no se encuentra a niveles predeterminados.

1.2 Usos del análisis de la covarianza.

Los usos mas importantes del análisis de la covarianza son:

1. Controlar el error y aumentar la precisión.
2. Ajustar medias de tratamientos de la variable dependiente de las diferencias en conjunto de valores de variables independientes correspondientes.
3. Ayudar en la interpretación de datos, especialmente en lo concerniente a la naturaleza de los efectos de los tratamientos.
4. Particionar una covarianza total o suma de productos cruzados en componentes.
5. Estimar datos faltantes.

1. Control del error. La varianza de una media de tratamiento es $\sigma_{\bar{Y}}^2 = \sigma^2 / n$. Así para disminuir esta varianza, sólo tenemos dos enfoques: el aumento del tamaño de la muestra o el control de la varianza en una población muestreada.

El control de σ^2 se logra mediante el diseño experimental o mediante el uso de una o más covariables. Ambos métodos pueden usarse simultáneamente. Cuando se usa la covarianza, como método para reducir el error, esto es, de controlar σ^2, se hace reconociendo el hecho de que la variación observada de la variable dependiente *Y* es parcialmente atribuible a la variación de la variable independiente.

El uso de la covarianza para controlar el error es un medio de aumentar la precisión con la cual los efectos de los tratamientos pueden medirse eliminando, por regresión, ciertos efectos reconocidos que no pueden ser o no han sido

controlados efectivamente por el diseño experimental. Por ejemplo, en un experimento de nutrición animal para comparar el efecto de varias raciones en el momento de peso, los animales asignados a un bloque varían en un peso inicial.

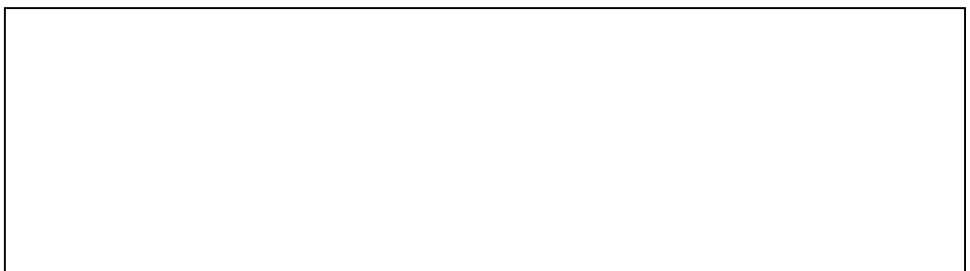
Ahora, si el peso inicial esta correlacionado con la ganancia de peso, una porción de error experimental en la ganancia puede deberse a diferencias de peso inicial. Mediante el análisis de la covarianza esta porción, una contribución que puede atribuirse a diferencias del peso inicial puede calcularse y eliminarse del error experimenta para ganancia.

2.- ajuste de medias de tratamientos. Cuando la variación observada *Y* puede atribuirse parcialmente a la variación en *X* , la variación entre la $\overline{Y}$ de los tratamientos también debe afectarse por la $\overline{X}$ de los tratamientos. Para que sean comparables, las $\overline{Y}$ de los tratamientos deberán ajustarse para hacer de ellas las mejores estimaciones de lo que hubieran sido si todas las $\overline{X}$ de los tratamientos hubiesen sido iguales.

Si el objeto principal de la covarianza es ajustar las $\overline{Y}$ de tratamiento, es también el reconocimiento de una situación de regresión que se exige el correspondiente ajuste del error en todo caso, es necesario medir la regresión apropiada independientemente de otras fuentes de variación que pueden invocarse en el modelo.

La idea general es evidente en la figura 1.1 para dos tratamientos. Por cada tratamiento, se ve la variación de X contribuye ala variación de Y. Así pues se vela necesidad de controlar la varianza del error mediante el uso de la covariable.

Al mismo tiempo la distancia entre $\overline{X1}$ y $\overline{X2}$ puede contribuir a la diferencia entre $\overline{Y1}$ y $\overline{Y2}$. Si las Y de los tratamientos se han observado a partir de una $\overline{X}$ común, digamos X 0, entonces serán comparables. Así pues ajustar las medias de los tratamientos es evidente.



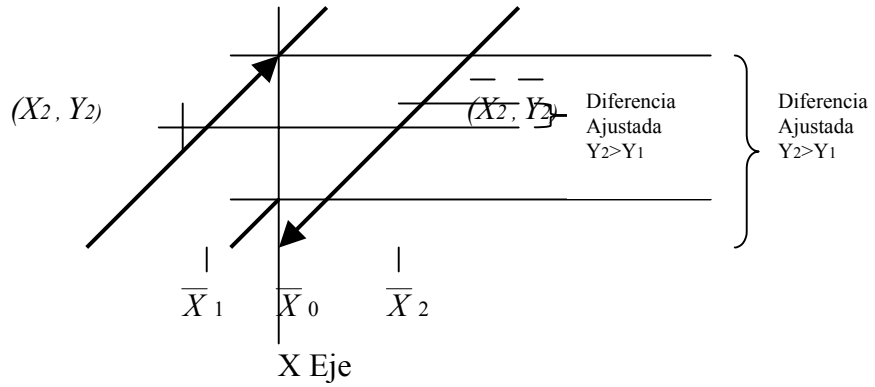


Figura 1.1 Control del error y ajuste a las medidas de tratamiento mediante la covarianza.

Es experimentos de nutrición animal las diferencias de medias de tratamientos no ajustadas puede obedecer a las diferencias en el valor nutritivo de las raciones, a diferencia de las cantidades consumidas, o ambas. Si las diferencias entre ganancias medias de peso para las diferentes raciones se ajustan a un consomo común de alimento, las medias ajustadas indican si las raciones difieren o no en valor nutritivo. Aquí al proveer información de base sobre la forma como las tratamientos producen los efectos la covarianza toca los principios fundamentales de los resultados de la investigación.

3.- Interpretación de datos. Todo procedimiento aritmético y técnica estadística asociada se proponen contribuir a la interpretación de datos. Así, los usos 1 y 2 definitivamente tienen que ver con interpretación de datos.

Sin embargo, se piensa que el uso 3 sea más especifico en cuanto que el análisis de la covarianza a menudo ayuda al experimentador entender los principios que fundamentan los resultados de una investigación.

Por ejemplo: puede ser bien sabido que ciertos tratamientos producen efectos tanto en variable dependiente como en las independientes. La covarianza, como medio de controlar el error y ajustar medias de tratamiento, se usa primordialmente, cuando la variable independiente mide efectos ambientales y en ella no influyen los tratamientos. Pero si ocurre así, la interpretación de los datos cambia. Esto es así porque las medias de tratamientos para la variable independiente son las mismas. El ajuste elimina parte de los efectos de

tratamiento cuando las medias de la variable independiente están afectadas por los tratamientos. La varianza debe usarse con precaución.

En un ensayo de fertilización en remolacha azucarera, los tratamientos pueden producir diferencias en densidad. Cuando la densidad, la variable independiente, está afectada por los tratamientos, el análisis del rendimiento ajustado a las diferencias elimina parte del efecto del tratamiento y entonces el experimentador puede desorientarse en la interpretación de los datos. Aun así, el análisis de covarianza puede proporcionar información útil.

El análisis de covarianza proporciona un método más apropiado y satisfactorio de ajuste de los datos experimentales. En situaciones en las que presentación diferencias reales entre los tratamientos para la variable independiente, pero que no son el efecto directo de los tratamientos , se justifica el ajuste. Por ejemplo, considérese un ensayo variental para el cual se ha producido semilla de diferentes variedades en diferentes áreas. Tal semilla puede tener una germinación muy diferente, no debido a causas intrínsecas de las mismas, sino como resultados de las condiciones ambientales de las cuales crecieron. En consecuencia, en densidad pueden ocurrir aun cuando se controlen la tasa de siembra.

En esta situación, esta justificada el uso de la covarianza tanto para control del error como para ajuste de rendimiento.

4.- La aparición de una covarianza total. Una covarianza proveniente de un experimento replicado se particiona cuando queremos determinar la relación entre dos o más variables medias cuando en la relación no influyen otras fuentes de variación.

5.- Estimación de datos faltantes. Las formulas dadas anteriormente para estimular datos faltantes dan lugar a una suma de cuadrados residual mínima. Pero la suma de cuadrados de tratamientos presentan un sesgo hacia arriba. El uso de la covarianza para estimar las parcelas faltantes lleva a una suma residual de cuadrados mínima más una de cuadrados de tratamientos no sesgada.

1.3 El modelo y los supuestos para la covarianza.

En el primer caso, hacemos énfasis en los aspectos del diseño experimental del problema. Deseamos efectuar un análisis de la varianza de los valores que han sido ajustados a la regresión respecto de una variable independiente. Se está destacando el uso 2 (sec. 1.2) aunque obviamente tenemos en mente los usos de 1 y 3.

Los supuestos necesarios para el uso válido de la covarianza son :

1.- Los X son fijos, medidos sin error, e independientes de los tratamientos.

2.- La regresión de Y respecto de X, después de eliminar diferencias de bloques tratamientos, es lineal e independiente de tratamientos y bloques.

3.- Los residuos se distribuyen normal e independientemente con media cero y varianza común.

El supuesto 1 establece que los X son fijos. Esto quiere decir que, en la obtención de los valores sesgados, se repite el mismo, conjunto X. A su turno, las interferencias sólo se aplican al conjunto de los X realmente observados. Mientras que los X no se seleccionen exactamente o se visualicen como idénticos en muestreo real repetido, entonces las inferencias se harán para valores interpolados en vez de extrapolados. También los X deben medirse exactamente de tal forma que las medias de población se identifiquen con propiedad. En realidad, la medición del error va a ser simplemente trivial en relación con la variación observada. El supuesto 1 también aplica, al igual que lo establece, que para el uso normal de la covarianza, los tratamientos no afectarán a los valores X, porque al fijarlos, pueden escogerse o restringirse por razones de comodidad . ya se ha indicado que la covarianza puede usarse cuando los valores X están así afectados, pero debe usarse cuando los valores X están así afectados, pero debe usarse con prudencia.

El supuesto 2 establece que el efecto de X sobre Y es aumentar o disminuir todo Y, en promedio, en un múltiple constante a la desviación del correspondiente X respecto de la $\overline{X}$.. para todo el experimento, es decir, en β (X_{ij} - $\overline{X}$..), se supone que la regresión es estable u homogénea. Así, no se requiere subíndice en β para relacionarlo con bloques o tratamientos. Un caso de tal relación se expone en la sec. 1.9

El supuesto 3 es aquel del cual depende la validez de las pruebas usuales, t y F. Un análisis, tal como lo determina el modelo, da una estimación válida de la varianza común cuando se ha aleatorizado los tratamientos dentro de los bloques. El supuesto de normalidad no es necesario para estimar las componentes de la varianza en Y; pero es necesaria la aleatorización.

La varianza residual se estima con base en los estimadores mínimos cuadrados de μ , los τ_i, los ρ_j y β indicados con acentos circunflejos. La ec. (1.2) cumple.

$$\sum_{i,j} Y_{ij} - \mu - \tau_i - \rho_j - \beta (X_{ij} - \overline{X}..)^2 = \text{mínimo}$$

Aquí ajustamos el modelo completo, al cual se apela cuando todas las hipótesis alternas aplicadas son verdaderas. Para las estimaciones por mínimos cuadrados, es correcta la ec. 1.3

$$\sum_{i,j}\left(Y_{ij} - \mu - \tau_i - \rho_j - \beta (X_{ij} - \overline{X}..) \right) = 0$$

La suma de todas las desviaciones en cero. No es necesario obtener y usar las estimaciones de los parámetros en la ec. (1.2) como tampoco para el análisis de bloques completos al azar sin covarianza. Sin embargo, las ecs. (1.4) a (1.6) definen las aceptaciones y dan la varianza residual.

$$\mu = \overline{Y}$$

$$\tau_i = t_i = \overline{Y}_i - \overline{Y}_{..} -$$

$$\rho_j = r_j = \overline{Y}_i - \overline{Y}_{..} - b\,(\overline{X}_j - \overline{X}..) \qquad (1.4)$$

$$\beta = b = \frac{Exy}{Exx}$$

$$\sigma \quad \frac{2}{y}.x = s\,\frac{2}{y}.x = \frac{Eyy - (Exy)^2 / Exx}{f_e} \qquad (1.5)$$

Donde *Exx, Exy y Eyy* son sumas de productos ajustadas al error; por ejemplo *Exx* es la suma de cuadrados del error para *X* y f_e los grados de error. Puede verse en la segunda de las ecs. (1.4) que para estimar el efecto de tratamiento σ_i para que estimar el efecto de tratamiento σ_i la desviación de toda media de tratamiento respecto de la media general debe ajustarse en la cantidad $b(\overline{X}_i - \overline{X}..)$. Este ajuste elimina todo efecto atribuible a la variable *X.* Son las medias de tratamientos ajustadas las que son comparables. Se deberán introducir variables ficticias para μ, los τ y los *p* para que la analogía sea completa. La ec. (1.6) es la fórmula de cálculo para lo que corresponde a la fórmula de definición. Dada como ec. (1.2) (1.6)

1.4 Prueba de medias de tratamientos ajustadas.

La tabla 1.1 da el análisis de la covarianza para un diseño de bloques completos al azar y, al mismo tiempo, ilustra el procedimiento general. Obsérvese la nueva notación con letras mayúsculas y subíndices pareados para indicar sumas de productos; un cuadrado es un tipo de producto particular.

La lógica del procedimiento depende del ajuste de modelos mediante técnicas de regresión múltiple. Una analogía estricta exige la inclusión de $r - 1$ y $t - 1$ variables ficticias para efectos de bloques y tratamientos, respectivamente, lo mismo que todas las covariables medidas. Nuestro interés se centra en el aspecto de la regresión donde SC (total, ajustada) se particiona en componentes atribuibles a la regresión y al error de residuo. Esto debe hacerse para el *modelo completo* , el que está dentro de H_1 , y de nuevo para el *modelo reducido,* que esta dentro de H_0 . La reducción adicional debido a la introducción, en el modelo, del conjunto de parámetros que se prueban mediante :

SC (regresión | H_1) – SC (regresión | H_0)

.=SC(residuos | H_0) –SC (residuos | H_1)

recuérdese que :

SC (regresión | H_1) – SC (τ_i , p_i , β)

$\{\ \}\ \{\ \}$

La " reducción adicional " en la forma de cuadrado medio se prueba con CM (residuos $|H_1$). La tabla 1.1 esquematiza el proceso. Primero se ajusta el modelo completo . Se emplea un proceso secuencial, tal como se expuso en la sec. 1.7 con las salidas de computador. La secuencia seguida ajusta efectos de bloques y tratamientos en primer lugar. Esto es cómodo debido a la ortogonalidad. La columna *Y, Y* se calcula con el residuo.

E_{yy} = SC (total, ajustado) – SC ($\{\tau_i\}, \{p_i\},$)

Aún sin ajustar por la regresión con respecto a *X.* Para esto también se necesitan los residuos *Exx* y *Exy* . La idea se discutió en la sec. 1.7 cuando se consideró la segunda parte de la tabla 1.5. Ahora úsese la línea del "Error " para calcular la contribución atribuible a la regresión ajustada para bloques y tratamientos, es decir,

$$\frac{E^2xy}{Exx} = \text{SC}\ (\beta \mid \{\tau_i\}, \{p_j\}).$$

Tabla 1.1 Prueba de medias de tratamiento ajustadas.

		Sumas de productos de				
Fuente	gl	*X,X*	*X,Y*	*Y,Y*	*gl*	$\sum (Y - Y)^2$ *CM*
Total	rt – 1	$\sum X - \overline{X})^2$	$\sum (X - \overline{X})(Y - \overline{Y})^2$	$\sum (Y - \overline{Y})^2$		
Bloques	*r – 1*	*R XX*	*RXY*	*Rxx*		
Tratamientos	*t-1*	*Txx*	*Txy*	*Tyy*		
Error	*(r – 1) (t-1)*	*Exx*	*Exy*	*Eyy*	*(r-1) (t-1) –1*	$\left(Err - \frac{(Exy)^2}{Exx}\right)$ $s\frac{2}{y-x}$
Tratamientos + error	*r(t-1)*	*Sxx*	*Sxy*	*Syy*	*r(t-1) –1*	$Syy - \frac{(Sxy)^2}{Sxx}$
Tratamientos ajustados					*t – 1*	$Syy - \frac{(Sxy)^2}{Sxx}$ *CM(T, Ajustados)*
					-	$Err - \frac{(Exy)^2}{Exx}$

Finalmente, tenemos CME = $sy^2 .x$ para el modelo completo. El coeficiente de regresión parcial de Y con respecto a X esta dado por $b = Exy / Exx$; estima el β de la ec. 1.1. según de ajusta el modelo reducido. Aquí, no hay efectos de tratamiento; en el modelo solo hay efectos de bloques y con respecto a X.

comiéncese por ajustar por bloques, pero tratamientos, ya que los últimos ya no se encuentran en el modelo; se necesitan Ryy=SC ($\{\rho_j\}$) y una nueva SC (residuos) = $\sum (Y-\bar{Y})^2$ - *Ryy* *C*omo hemos visto, *Rxx, Rxy* y el correspondiente SC (residuos) nuevo sera necesario cuando se ajusta β. Claramente SC (residuos) para *Y* Tyy + *Eyy* = *Syy* y lo será independientemente del diseño experimental usado. Los otros residuos serán similares. Ahora ajuste β y obténgase SC (β | $\{\rho_j\}$) = $S\frac{2}{xy}$ */ S xx.* A su turno, calcúlese.

$$\sum (Y - Y)^2 = \text{SC (Total ajustado)-SC } (\{\rho + \frac{S^2xy}{Sxx},$$

donde

$$\text{SC } (\{\rho_j\}, \beta) = Rxx + \frac{S^2xy}{Sxx},$$

Finalmente la diferencia entre las sumas residuales de cuadrados para los dos modelos es la cantidad atribuible a la inclusión de los efectos de los tratamientos como el ultimo conjunto de componentes en el modelos, o sea, luego de los efectos de los bloques y el termino de la regresión. Es SC ($\{\tau_i\} \mid \{\rho_j\}, \beta$). Hemos hecho lo equivalente a encontrar una suma de cuadrados atribuible al ajuste de coeficientes de regresión parcial para los efectos de los tratamientos.

Para aprobar el cuadrado medio de tratamientos ajustado, el cuadrado medio del error apropiado es $s\frac{2}{y}$*.x.* Cuando se desea hacer varias pruebas como cuando se particiona una suma de cuadrados de tratamientos en componentes, las pruebas exactas exigen un calculo separado de tratamientos más error y tratamientos ajustados para comparación.

COVARIANZA MÚLTIPLE.

Es posible que sobre cada unidad experimental se observen los valores de *p* variables compañeras, simultáneamente con el valor observado con las características en estudio; si estas variables se miden prácticamente sin error, y el investigador sospecha que pueden ejercer alguna influencia en el valor de las característica bajo estudio, entonces la extensión del modelo que podemos escribir en la forma:

$$Yij = \mu + \beta\, i + \tau\, j + \gamma\, 1\ X_{ij}^{(1)} + \gamma\, 2\, X_{ij}^{(2)} + \ldots + \gamma\, p\, X_{ij}^{(p)} + eij \qquad (2.1)$$

Seria apropiada para interpretar la información experimental, supuesto el experimento de bloques completos al azar (los errores de *eij* tienen las propiedades usuales, es decir son no correlacionados, con media 0 y varianza σ^2, y distribuidos normalmente; sobre la unidad experimental (*i,j)* se observarían, además los valores de *yij,* los valores de $X_{ij}^{(1)}$, $X_{ij}^{(2)}$, ..., $X_{ij}^{(p)}$ de *p* variables como siendo γ *1,* γ *2, ...,* γ *p* los coeficientes de varianza de manera similar a la covarianza simple para estimar los coeficientes de varianza primero sería necesario construir los análisis de varianza y de productos cruzados, de las observaciones yij y las p variables, es decir, tendrían que realizarse *p + 1* análisis de varianza y *p(p+1) / 2* análisis de productos cruzados. El objetivo fundamental de tales análisis sería obtener los términos de error, ya que de ellos depende la estimación de γ *1,* γ *2, ...,* γ *p .*

Denotemos que por *Eyy, E11, E22, ... , Epp,* respectivamente, los términos del error en los análisis de varianza de *yij* , $X_{ij}^{(1)}$, $X_{ij}^{(2)}$, ..., $X_{ij}^{(p)}$. sean *E1y, E2y, ...,Epy,* respectivamente la suma de productos de términos de error en los análisis de productos cruzados de las $X_{ij}^{(1)}$*,por las* $X_{ij}^{(2)}$, *E13* la suma de productos del termino de error, en el análisis de productos cruzados de las $X_{ij}^{(1)}$ por las $X_{ij}^{(3)}$, etc. Los datos anteriores se

consignan en una tabla de productos cruzados, que en forma condensada tendría la lectura siguiente :

	$X^{(1)} X^{(1)}$	$X^{(1)} X^{(2)}$	$X^{(p)} X^{(p)}$	$X^{(1)}.Y$	....*Y.Y*
Bloques	*B 11*	*B 12*	*…..B pp*	*B 1 y*	*… B yy*
Tratamientos	*T 11*	*T 12*	*.....T pp*	*T 1 y*	*…T yy*
Error	*E 11*	*E 12*	*.....E pp*	*E 1 y*	*...E yy*

Donde la anotación es análoga a la covarianza simple utilizando los términos de error de la tabla anterior, el sistema de ecuaciones normales reducidas para las covariables, eliminando los elementos de clasificación en (2.1) esta dado por :

$$\left\{\begin{array}{l} \gamma\ 1\ E11 + \gamma\ 2\ E12 + \ldots + \gamma\ p\ E1p = E1y, \\ \gamma\ 1E21 + \gamma\ 2\ E22 + \ldots + \gamma\ p\ E2p = E2y, \\ . \qquad . \qquad . \qquad . \\ . \qquad . \qquad . \qquad . \\ . \qquad . \qquad . \qquad . \\ \gamma\ 1\ Ep1 + \gamma\ 2\ Ep2 + \ldots + \gamma\ p\ Epp = Epy, \end{array}\right.$$

figura 2.2

Si ahora definimos las cantidades de *E ´11 = E 11 + T11, E ´12 + T 12, ..., E1p = E1p + T1p, E´1y + T 1y, ... , E ´yy = E yy + Tyy,* entonces la solucion al sistema de la ecuación

$$\left\{\begin{array}{l} \gamma\ 1\ E'11 + \gamma\ 2\ E'12 + \ldots + \gamma\ p\ E'1p = E'1y, \\ \gamma\ 1E'21 + \gamma\ 2\ E'22 + \ldots + \gamma\ p\ E'2p = E'2y, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \gamma\ 1\ E'p1 + \gamma\ 2\ E'p2 + \ldots + \gamma\ p\ E'pp = E'py, \end{array}\right.$$

figura 2.8

Solución que también puede obtenerse por los métodos de inversión de matices, produce los estimadores minicuadraticos de $\hat{\gamma}1$, $\hat{\gamma}2$, ... $\hat{\gamma}p$ de los coeficientes de varianza en el modelo 2.1 bajo la hipótesis de Ho; $\tau 1=\tau 2 = ... = \tau$ t.-, es decir los estimadores de los parámetros $\gamma 1$, $\gamma 2$, ... γp del modelo reducido:

$$Yij = \mu + \beta i + \tau j + \gamma 1 \; X_{ij}^{(1)} + \gamma 2 \; X_{ij}^{(2)} + \ldots + \gamma p \; X_{ij}^{(p)} + eij \qquad (2.9)$$

Procediendo de manera idéntica al caso del modelo completo, la suma de cuadrado de los errores del modelo (2.9), SCE´, esta dada por:

$$SCE' = E'yy - \{ \hat{\gamma}\mathbf{1}\, E'1y + \hat{\gamma}\mathbf{2}\, E'2y + ... + \hat{\gamma}p\, E'py\} \qquad \textit{figura 2.10}$$

Por consiguiente la suma de cuadrados debido a la hipótesis Ho; $\tau 1= \tau 2 = ... = \tau$ t, SC (TA), es :

$$\begin{aligned} SC\,(TA) &= SCE' - SCE \\ &= \{ E'yy - \hat{\gamma}\mathbf{1}\, E'1y + \hat{\gamma}\mathbf{2}\, E'2y + ... + \} + \hat{\gamma}p\, E'py \\ &= \{ Eyy - \hat{\gamma}\mathbf{1}\, E1y + \hat{\gamma}\mathbf{2}\, E2y + ... + \} + \hat{\gamma}p\, Epy \qquad (2.11) \end{aligned}$$

De aquí que para probar la hipótesis Ho; $\tau 1= \tau 2 = ... = \tau$ t, vs H1; τ tj $\neq \tau$ t´j, al menos para un par de índices j y ´j , con j diferente a ´j se calcula la estadística

$$F=\frac{CM(TA)}{s^2} \qquad (2.12)$$

Donde CM (TA), el cuadrado medio debido a tratamientos ajustados por covarianza, es la suma de cuadrados correspondiente SC (TA) , dividida en t –1 , y s^2 es el cuadrado medio del error bajo el modelo completo, dado completo por la relación 2.7 si la hipótesis Ho es cierta, (2.12) se distribuye como una *F*con t – 1 y (r-1) (t-1) – p grados de libertad, siendo este elk criterio que se emplea para realizar la prueba.

Si se desea comparar la hipótesis Ho: $\gamma \mathbf{1} = \gamma \mathbf{2} = \ldots = \gamma \mathbf{p} = \mathbf{0}$ **vs** H ´1 : al menos una $\gamma \mathbf{j} \neq \mathbf{0}$, $j = 1, 2, \ldots, p$ se calcula la estadística:

$$F = \frac{\hat{\gamma}\,\mathbf{1}\,E1y + \hat{\gamma}\,\mathbf{2}\,E2y + \ldots + \} + \hat{\gamma}\,p\,E1p}{sp^2} \qquad (2.13)$$

Que cuando H ´o es cierta , se disminuye como F con p y (r-1) (t-1) – p grados de libertad, siendo esta la distribución que se emplea para realizar la prueba.
El arreglo en p hileras y p columnas de los términos de error genera la matriz p X p siguiente:

$$\begin{pmatrix} E11 & E12 & \ldots & E1p \\ E21 & E22 & \ldots & E2p \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ Ep1 & Ep2 & \ldots & Epp \end{pmatrix} \qquad (2.3)$$

Donde por simetría *E12 = E21, E13= E31.*

Que es no singular (existe su inversa que es otra matriz de dimensiones *p X p,* la cual puede calcularse por el método que se escribe en la sección de regresión múltiple. Sea :

$$\begin{pmatrix} E^{11} & E^{12} & \ldots & E^{1}p \\ E^{21} & E^{22} & \ldots & E^{2}p \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ Ep^{1} & Ep^{2} & \ldots & Epp \end{pmatrix} \qquad (2.4)$$

La inversa de la matriz en 2.3; en terminos de elementos de la matriz en 2.6 la solucion al sistema de las ecuaciones normales reducidas para las covariables (2.2) se obtiene como sigue:

$$\hat{\gamma}1 = E^{11}\,E1y + E^{12}\,E2y + \ldots + E^{1}p\,Epy$$
$$\hat{\gamma}\,2 = E^{21}\,E1y + E^{22}\,E2y + \ldots + E^{2}p\,Epy$$
. . . .
. . . .
. . . .
$$\hat{\gamma}\ \ p = Ep^{1}\,E1y + Ep^{2}\,E2y + \ldots + Epp\,Epy$$

Asi, la reducción en la suma de cuadrados de los errores debida exclusivamente al ajuste de las covariables esta dada por la cantidad:

$$\hat{\gamma}\,1 = \hat{\gamma}\,2\ E2y + \ldots + \hat{\gamma}\,p\ Epy$$

Por consiguiente, la suma de cuadrados de los errores de vida al ajuste del modelo 2.1 es :

$$SCE = Eyy - \{\ \hat{\gamma}\ 1\ E1y + \hat{\gamma}\ \ 2\ E2y + \ldots + \hat{\gamma}\ \ p\ Epy\ \} \qquad (2.6)$$

El cuadrado del error bajo el modelo (2.1), *CME,* se obtiene dividiendo *SCE* entre (r-1) (t-1) – p, grados de libertad, los grados de libertad asociados con el error bajo el modelo completo, es decir :

$$CME = \frac{SCE}{(r-1)(t-1)\ p} = \frac{Eyy - \{\hat{\gamma}1E1y + \hat{\gamma}2E2y + \ldots + \hat{\gamma}pEpy\}}{(r-1)(t-1)\ p} = s^2 \qquad (2.7)$$

El mejor estimador línea insesgado de un contraste de la forma $\sum_{j=1}^{t} \lambda j \tau j$, con esto $\sum_{j=1}^{t} \lambda j = 0$, viene dado por :

$$\sum_{j=1}^{t} \lambda j \tau j = \sum_{j=1}^{t} \lambda j \{\hat{y}\,j - \hat{y}\,1\ \overline{X}\,j\ ^{(1)} - \hat{y}\,2\ \overline{X}\,j\ ^{(2)} - \ldots - \hat{y}\,p\ \overline{X}\,j\ ^{(p)} - \} \qquad (2.14)$$

donde $\hat{y}\,j$, $\overline{X}\,j\ ^{(1)}, \overline{X}\,j\ ^{(2)}, \ldots, \overline{X}\,j\ ^{(p)}$ son las medias de la característica y de las covariables en estudio, correspondiente al tratamiento *j.* La varianza de $\sum_{j=1}^{t} \lambda j \tau j$ es:

$$\text{Var}\ (\sum_{j=1}^{t}\ \lambda j \hat{\tau} j) = \frac{\sigma^2}{r} \sum_{j=1}^{t} \lambda j^2 + \sigma^2 \sum_{j=1}^{p} \sum_{j=1}^{p}\ \alpha i \alpha i\,' Eii'$$ es el elemento en la hilera *i,* columna i´de la matriz (2.4).

ANÁLISIS DE COVARIANZA

VARIABLE EXPERIMENTAL : CONDICIÓN DE PASTIZAL

Tabla de datos de la variable y (biomasa final)

Tratamientos	I	II	III	IV	
1	317	311	274	321	1223
2	305	316	321	385	1327
3	300	335	309	305	1249
4	298	292	258	297	1145
5	290	298	273	270	1131

Tabla de datos de la variable x (cobertura inicial)

Tratamientos	I	II	III	IV	
1	240	235	207	242	924
2	210	216	221	264	911
3	218	242	224	216	900
4	242	237	210	241	930
5	230	236	216	214	896

Hipótesis a probar:

Ho: Ti = Tj Y LA Ha: Ti ≠ Tj

PORQUÉ SE UTILIZO ESTE DISEÑO ?

Pueden planearse experimentos de modo que ciertos tipos de efectos ambientales sean eliminados de las estimaciones de los efectos de tratamientos, con el resultado de que estas estimaciones se hacen con mayor exactitud. muy frecuentemente sucede que algunas fuentes de variación que no pueden ser controladas por el diseño, en cambio pueden ser medidas tomando observaciones adicionales. Cuando esto ocurra el análisis de covarianza puede ser utilizado frecuentemente con gran ventaja. una precaución es que, ya que las observaciones adicionales miden los efectos ambientales, no deben estar afectadas por los tratamientos.

MODELO ESTADÍSTICO.

$Yij = \mu + Ti + \beta (\overline{xi.} - \overline{x ..}) + Eij$

i = 1,2,3,4,5, TRATAMIENTOS Y j= 1,2,3,4, REPETICIONES

PORQUE DEL MODELO ?

En el modelo se hace una introducción de β (xi. – x ..) para describir el efecto de la cobertura inicial se ha supuesto que el efecto es lineal, esto es una constante múltiple β de la cantidad por lo cual la cobertura inicial xij en la parcela difiere del promedio inicial x para todo experimento.

TOTALES DE TRATAMIENTOS

Xi.	$\overline{Yi.}$
924	1223
911	1327
900	1249
931	1145
896	1131

X..=4561 Y..=6075

SUMAS DE CUADRADOS Y PRODUCTOS

$$Txx = \sum_{i=1}^{t}\sum_{j=1}^{r} Xij^2 - \frac{X..^2}{tr} = 240^2 + 210^2 + \ldots 214^2 - \frac{4561^2}{5x4}$$

$Txx = 4276.95$

$$Txx = \sum_{i=1}^{t} \frac{Xi.^2}{r} - \frac{X..^2}{tr} = \frac{924^2 + \ldots 896^2}{4} - \frac{4561^2}{5 \times 4}$$

$Txx = 217.2$

$$Exx = Txx - Txx = 4276.95 - 217.2$$

$Sxx = 4276.95$

$Sxx = Exx - txx = 4059.75 + 217.2$

$Sxx = 4276.95$

$t \quad r$

$$Txy = \sum_{i=1}^{t}\sum_{j=1}^{r} Xij\ Yij - \frac{X..\ Y..}{tr} = (240 \times 317) + \ldots + (896 \times 270) - \frac{4561x6075}{5x4}$$

$Txy = 5476.25$

$$Txy = \sum_{i=1}^{t} Xij\ Yij - \frac{X..Y..}{tr} = \frac{(924 \times 1223) + \ldots + (896 \times 1131)}{4} - \frac{(4561 \times 6075)}{5x4}$$

$try = -85$

$Exy = Txy - txy = 5467.25 - (-85)$

$Exy = 5552.25$

CONTINUACIÓN.

$Sxy = Exy + txy = 5552.25 + (-85)$

$Sxy = 5467.25$

$t \quad r$ $\qquad Y..^2 \qquad 6075^2$

$Tyy = \sum\sum Yij^2 = \_\_\_\_\_ = 317^2 + 305^2 + \ldots + 297^2 + 270^2 - \_\_\_\_\_$

$Tyy = 14097.75$

$$Tyy = 6430.0$$

$$Eyy = \sum_{i=1}^{t} \frac{Yi.^2}{r} - \frac{Y..^2}{tr} = \frac{1223^2 + \ldots + 1131^2}{4} - \frac{6075^2}{5x4}$$

Tyy = *6430.0*

$Eyy = Tyy - tyy = 14097 - 6430.0$

$Eyy = 7667.75$

$Syy = Eyy + tyy = 7667.75 + 6430$

$SYY = 14097.75$

CUADRO DE SUMAS DE CUADRADOS Y PRODUCTOS

F.V.	GL	SXX	SXY	SYY
TRATAMIENTOS	t – 1	txx	txy	tyy
	5 – 1= 4	217.2	-85	6430
E.Exp.	t (r – 1)	Exx	Exy	Eyy
	5 (4 – 1)=15	4059.75	5552.25	7667.75
TOTAL	tr – 1	Sxx	Sxy	Syy
	5x4 – 1 = 19	4276.95	5467.25	14097.75

ANALISIS DE VARIANZA

ANALISIS DE REGRESION

Hipótesis a probar:

Ho: $\beta = 0$ Y LA Ha: $\beta \neq 0$

F.V.	GL	SC	CM	FC	$F\alpha^{1\ 14}$
REGRESION	1	7596.442	7593.442	1430.835	0.05=4.60
RESIDUAL	t (r - 1) – 1 5(4 -1) – 1=14	74.307	5.307		3.01=8.86
TOTAL	tr – 1 5x4 – 1 = 19	7667.75			

SUMAS DE CUADRADOS

$$Sc\ \text{Re}\ gresion = \frac{(Exy)^2}{Exx} = \frac{5552.25^2}{4059.75}$$

Sc Reg. = 7593.442

Sc Total = *Eyy* = 7667.75

Sc Re *sidual*= *Sc Total* – *Sc* Re *gresion* = 7667.75 – 7593.442

Sc Res. = 74.307

$$\frac{Sc\text{REG}}{} \quad \frac{7593.442}{CM} = =$$

$_{REG}$ **CM = 7593.442**

$$\frac{Sc\text{REG}}{} \quad \frac{74.307}{CM} = =$$

$_{REG}$ **CM = *5.307***

$$\frac{CM_{REG}}{} \quad \frac{7593.442}{} = \mathbf{Fc} \qquad =$$

$$CM_{REG} \qquad 5.307$$

Fc = *1430.835*

CONCLUSIÓN PRELIMINAR

Como f calculada es mayor que f tabulada al ($p \leq 0.01$) rechazo Ho: la biomasa inicial afecta significativamente la biomasa final.

ANÁLISIS DE VARIANZA AJUSTADO

F.V.	GL	SC	CM	FC	$F_{a}^{4}{}_{14}$
TRATAMIENTOS	t - 1 5 – 1 = 4	7034.625	1758.656	331.344	0.05 = 3.11 0.01 = 5.04
E. Exp.	t (r - 1) – 1 5(4 -1) – 1=14	74.307	5.307		
	tr – 2	7108.932			
TOTAL	5x4 – 2 = 18				

SUMAS DE CUADRADOS

$$Sc\ Total = Syy - \frac{Sxy^2}{Sxx} = \quad - \frac{5467.25^2}{4276.95}$$

$$Sc\ Total. = 7108.932 \qquad 14097.75$$

$$ScE.Exp. = Eyy - \frac{Exy^2}{Exx} = 7667.75 - \frac{5552.25}{4059.75}$$

$$ScE.\ Exp. = 74.307$$

$$ScTratamientos = Sc\ Total - ScE.Exp. = 7108.923 - 74.307$$

$$ScTrats.. = 7034.625$$

$$\frac{ScTrats.}{CM} = \frac{7593.442}{} =$$

$$_{TRATS}\mathbf{CM = 1758.656}$$

$$ScE.Exp. \qquad 74.307$$

CM = =

$_{E.EXP.}$**CM = *5.307***

$$\frac{CMtrats}{CM_{E.Exp}} \qquad \frac{1758.656}{5.307} = \qquad =$$

Fc = *331.344*

Como f calculada es mayor que f tabulada al ($p \leq 0.01$) rechazo Ho; por lo tanto la producción de biomasa en los sitios es muy diferente.

AJUSTE DE MEDIAS DE TRATAMIENTO

FORMA ORIGINAL DE AJUSTE

$\overline{Yi.} = \overline{Yi.} - \beta (\overline{Xi.} - \overline{X..})$

$\beta = \frac{Exy}{Exx} = \frac{5552.2}{4059.75} = 1.367$

Y1. = 305.75 – 1.367 (231 – 228.05) = 301.717

Y2. = 331.75 – 1.367 (227.75 – 228.05) = 332.160

Y3. = 312.25 – 1.367 (225 – 228.05) = 316.419

Y4. = 286.25 – 1.367 (232.5 – 228.05) = 280.166

Y5. = 282.75 – 1.367 (224 – 228.05) = 228.268

POR MEDIO DE LA PRUEBA DE SCHEFF HAY QUE VER ENTRE

TRATAMIENTOS SON DIFERENTES.

$$Fo = \sqrt{(t-1)\,S^2\,Fa\left(\frac{1}{ri}+\frac{1}{rj}\right)} = \sqrt{(5-1)\,(5.307)\,(3.11)\left(\frac{1}{20}+\frac{1}{20}\right)}$$

$Fo = 2.569$

t = no. DE TRATAMIENTOS = 5

S^2 = CM DEL E. Exp. = 5.307

F 0.05 = a^{4}_{14} = 3.11

ri,rj = no. DE TRATAMIENTOS = 5

Yi.	MEDIA	Yi.
T1	301.717	75.429 c
T2	332.16	83.04 a
T3	316.419	79.104 b
T4	280.166	70.041 d
T5	288.286	72.071 d

SHEFFE AL ($p \leq 0.05$)

	T2	T3	T1	T5	T4
	83.04	79.104	75.429	72.071	70.041
T4 70.041	12.999**	9.063 **	5.388**	2.03NS	0
T5 72.071	10.969**	7.033**	3.358**	0	
T1 75.429	7.611**	3.675**	0		
T3 79.104	3.936 **	0			
T2 83.04	0				

CONCLUSIÓN

Como nos interesa que haya mayor productor de biomasa por sitio concluimos en forma vertical al realizar la prueba de scheffe y podemos decir que donde se obtuvo la mayor producción de biomasa fue en el del tratamiento 2, seguida del 3, luego del 1, y por ultimo los tratamientos 4 y 5 que resultaron el mismo efecto.

```
Title <P.17> ANALISIS DE COVARIANZA MULTIPLE:  FACTORIAL DE 4X2 EN
DISEÑO B.A.;
      * Steel y Torrie: Capitulo 15  ;
      * Tabulación de datos: (Experimento en nutrición animal);
      *            Sin abono            Con Abono                  ;
      *            X1  X2  Y            X1  X2  Y
      *Bloque 1      -  -  -            -  -  -       SUELO TIPO  1 ;
      *Bloque 2      -  -  -            -  -  -       SUELO TIPO  1 ;
      *Bloque 3      -  -  -            -  -  -       SUELO TIPO  1 ;
      *..............................  etc.  ..................................... ;
      *Bloque 1      -  -  -            -  -  -       SUELO TIPO  4 ;
      *Bloque 2      -  -  -            -  -  -       SUELO TIPO  4 ;
      *Bloque 3      -  -  -            -  -  -       SUELO TIPO  4 ;
      *..................................................................... ;


      * X1 = Peso inicial: X2 = Pasto consumido: Y = Ganancia en peso  ;
      Data Acovmult;
      do Tipsuelo = 1 to 4;
      do Bloque = 1 to 3;
      do Abono = 1 to 2;
      input x1 x2 y @@ ;
      output;
      end;
      input;
      end;
      end;
      cards;
      220    1155    224    222    1326    237
      246    1423    289    268    1559    265
      262    1576    280    314    1528    256
      198    1092    118    205    1154    82
      266    1703    191    236    1250    117
      335    1546    115    268    1667    117
      213    1573    242    188    1381    184
      236    1730    270    259    1363    129
      288    1593    198    300    1564    212
      256    1532    241    202    1375    239
      278    1220    185    216    1170    207
      283    1232    185    225    1273    227
      ;

      * proc print ; run;
      proc glm ;
      class bloque abono tipsuelo;
```

```
model y = x1 x2 bloque abono tipsuelo abono*tipsuelo /solution;
lsmeans abono tipsuelo abono*tipsuelo /stderr pdiff;
run;
```

<P.17> ANALISIS DE COVARIANZA MULTIPLE: FACTORIAL DE 4X2 EN DISEÑO B 78

13:01 Saturday, March 13, 1999

General Linear Models Procedure

Class Level Information

Class	Levels	Values
BLOQUE	3	1 2 3
ABONO	2	1 2
TIPSUELO	4	1 2 3 4

Number of observations in data set = 24

<P.17> ANALISIS DE COVARIANZA MULTIPLE: FACTORIAL DE 4X2 EN DISEÑO B 79

13:01 Saturday, March 13, 1999

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	76758.297	6978.027	16.31	0.0001
Error	12	5135.536	427.961		
Corrected Total	23	81893.833			

R-Square	C.V.	Root MSE	Y Mean
0.937290	10.32211	20.687	200.42

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	3.993	3.993	0.01	0.9246
X2	1	13219.292	13219.292	30.89	0.0001
BLOQUE	2	332.048	166.024	0.39	0.6867
ABONO	1	2290.828	2290.828	5.35	0.0392
TIPSUELO	3	59775.554	19925.185	46.56	0.0001
ABONO*TIPSUELO	3	1136.582	378.861	0.89	0.4764

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	1341.589	1341.589	3.13	0.1020
X2	1	10585.053	10585.053	24.73	0.0003
BLOQUE	2	395.401	197.701	0.46	0.6408
ABONO	1	1850.625	1850.625	4.32	0.0597
TIPSUELO	3	59216.443	19738.814	46.12	0.0001
ABONO*TIPSUELO	3	1136.582	378.861	0.89	0.4764

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		131.5064653 B	1.88	0.0850	70.05431175
X1		-0.4943035	-1.77	0.1020	0.27918123
X2		0.1583657	4.97	0.0003	0.03184321
BLOQUE	1	-10.6200784 B	-0.49	0.6302	21.49557117
	2	2.2976508 B	0.17	0.8693	13.67121479
	3	0.0000000 B	.	.	.
ABONO	1	-0.7599656 B	-0.03	0.9743	23.06124772
	2	0.0000000 B	.	.	.
TIPSUELO	1	23.4517602 B	1.06	0.3090	22.07630728
	2	-121.4808291 B	-6.80	0.0001	17.87124714
	3	-58.0638738 B	-3.01	0.0109	19.31008202
	4	0.0000000 B	.	.	.
ABONO*TIPSUELO	1 1	13.5765144 B	0.42	0.6845	32.60547285
	1 2	37.3361588 B	1.48	0.1653	25.27099103
	1 3	29.7392803 B	0.97	0.3498	30.56777125
	1 4	0.0000000 B	.	.	.
	2 1	0.0000000 B	.	.	.
	2 2	0.0000000 B	.	.	.
	2 3	0.0000000 B	.	.	.
	2 4	0.0000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique

estimators of the parameters.

<P.17> ANALISIS DE COVARIANZA MULTIPLE: FACTORIAL DE 4X2 EN DISEÑO B 80

13:01 Saturday, March 13, 1999

General Linear Models Procedure
Least Squares Means

ABONO	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	Pr > |T| H0: LSMEAN1=LSMEAN2
1	210.118178	6.292618	0.0001	0.0597
2	190.715155	6.292618	0.0001	

TIPSUELO	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number
1	259.598426	8.594161	0.0001	1
2	126.545658	8.485458	0.0001	2
3	186.164174	9.334271	0.0001	3
4	229.358408	9.145719	0.0001	4

Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j	1	2	3	4
1	.	0.0001	0.0001	0.0347
2	0.0001	.	0.0006	0.0001
3	0.0001	0.0006	.	0.0091
4	0.0347	0.0001	0.0091	.

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

ABONO	TIPSUELO	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0	LSMEAN Number

1	1	266.006700	12.078556	0.0001	1
1	2	144.833755	12.769879	0.0001	2
1	3	200.653832	13.986703	0.0001	3
1	4	228.978425	14.268837	0.0001	4
2	1	253.190151	12.917782	0.0001	5
2	2	108.257562	12.444568	0.0001	6
2	3	171.674517	11.962663	0.0001	7
2	4	229.738391	15.152531	0.0001	8

Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j	1	2	3	4	5	6	7	8
1	.	0.0001	0.0042	0.0772	0.4940	0.0001	0.0001	0.0735
2	0.0001	.	0.0138	0.0004	0.0001	0.0735	0.1519	0.0021
3	0.0042	0.0138	.	0.2213	0.0184	0.0004	0.1361	0.1846
4	0.0772	0.0004	0.2213	.	0.1950	0.0001	0.0102	0.9743
5	0.4940	0.0001	0.0184	0.1950	.	0.0001	0.0006	0.3090
6	0.0001	0.0735	0.0004	0.0001	0.0001	.	0.0032	0.0001
7	0.0001	0.1519	0.1361	0.0102	0.0006	0.0032	.	0.0109
8	0.0735	0.0021	0.1846	0.9743	0.3090	0.0001	0.0109	.

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

BIBLÍOGRAFIA

William G. Cochran y Gertrude M. Cox, 1974. Diseños Experimentales, Editorial Trillas México.

Martínez.G.A. 1997 Diseños Experimentales Métodos y Elementos de Teoría Editorial Trillas. México. P.p. 281 – 284

Steel, R. G. D. Y Torrie, J.H. 1997. Bioestadistica Principios y Procedimientos Segunda Edición.
Editorial. Mcgraw - Hill. Pp. 293 - 424.

Reyes, C. P. 1982. Diseños de Experimentos Aplicados.
Editorial Trillas. México. Pp. 286 - 290.

Martínez, A. G 1996. Diseños Experimentales Métodos y elementos de teoría. Editorial Trillas. P.p 281 - 284.