LA CONSTRUCCION DE UN MODELO DE REGRESION

Eduardo Granados F.

Centro Internacional de Agricultura Tropical

- C I A T -

La construcción de un modelo de regresión es un proceso iterativo entre las etapas de identificación, estimación y chequeo del modelo. Se presentan varios conceptos que contribuyen a la apropiada formulación y escogencia de la ecuación de regresión que mejor ajuste a un conjunto de datos.

1. Introducción

El análisis de Regresión consiste en el ajuste de una ecuación a un conjunto de datos de tal manera que una variable, de respuesta o dependiente Y, se expresa como una función de una o màs variables predictoras o independientes (X_1, X_2,...,X_p), (p un número entero). La contribución de cada una de estas variables independientes en la respuesta se determina a través de coeficientes constantes desconocidos, denominados parámetros (β_0, β_1,...β_p). El término ajuste se refiere al proceso de búsqueda de la función de regresión que tenga mejor capacidad de predicción de la respuesta esperada según alguna medida que indique que tan buena o mala es la ecuación. Uno de los métodos que permiten encontrar la ecuación de mejor ajuste es el de Mínimos Cuadrados, que tiene como objetivo fundamental minimizar la suma de las discrepancias o errores cuadráticos, Σe^2_i, entre los datos base de la construcción del ajuste y los valores que se pronosticarian con la ecuación. La forma general de esta ecuación se plantea mediante la expresión:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ki} + e_i \qquad [1]$$

denominada Modelo de Regresión y sobre la cual es necesario hacer varios supuestos:

- Se observan n sujetos o unidades
- La respuesta Y es medida controlando todos los factores X_k en un ambiente experimental o al menos suponer que son fijos respecto de la respuesta (K = 1,2, ..., p < n)
- El valor esperado de los errores es cero, E(e)=0; su varianza es constante, $V(e_i)=\sigma^2$ sin depender del instante en que se hizo la observación, ni de la observación misma.

- No existe dependencia entre los errores de observación a observación: no correlacionados.
- Para efectos de inferencia estadística los errores se distribuyen normalmente.
- La forma del modelo es la correcta.

Este último supuesto se refiere a la adecuada expresión matemática que represente la relación y acción de los predictores en la respuesta. Los factores X_k son cuantitativos o al menos corresponden a una variable indicadora DUMMY con valores uno o cero, para señalar la presencia o ausencia de alguna característica que afecta la respuesta.

2. El Modelo de Regresión

Debido a que no necesariamente se conoce de antemano la forma correcta del modelo de regresión, se considera en un sentido amplio, el modelo Lineal en los parámetros:

$$T(Y_i) = \sum_{j=0}^{q} g_j(X_{1i}, X_{2i},\ldots,X_{pi}).\beta_j + r_i \qquad [2]$$

donde $T(Y_i)$ es una transformación de la variable de respuesta, que necesariamente afecta los supuestos distribucionales de los errores. Las funciones $g_j(X_{1i}, X_{2i},\ldots,X_{pi})$ son modificaciones o nuevas expresiones de las variables predictoras que son de interés encontrar y que mejoran la capacidad predictiva de la respuesta. Por convención se define $g_0(\ )$ como cero o uno, según se desee incluír o no en el modelo el coeficiente β_0 denominado intercepto. Los β_j son coeficientes de regresión, desconocidos y uno de los propósitos del análisis de regresión es su estimación; "q", es el tamaño del modelo, o número de variables y sus transformaciones que afectan la respuesta; r_i es una discrepancia específica a la observación i entre la respuesta observada,

o su transformación y la respuesta esperada. Para fines de dócimas de hipótesis, se deben hacer supuestos acerca de la distribución de los errores.

Para simplificar la notación, en adelante Z_j indicará cualquier función no nula $g_j(X_1, X_2,\ldots,X_p)$, $j = 1,2,\ldots,q < n$; T_i es la correspondiente transformación $T(Y_i)$; por lo tanto la expresión del modelo planteada en [2] será equivalente a:

$$T_i = \beta_0 + \sum_{j=1}^{q} \beta_j Z_{ji} + r_i \qquad [3]$$

Cuando no se requiere transformación para la respuesta, T = Y, es la identidad.

Algunas formas especiales de modelos de regresión son las siguientes:

Lineal simple	$Y = \beta_0 + \beta_1 X_1 + e$
Cuadrático	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$
Polinomial de orden p	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p + e$
Superficie de respuesta de orden 2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 . X_2 + e$
Semilogarítmico	$Y = \beta_0 + \beta_1 \text{ LOG } X_1 + e$
Exponencial	$\text{LOG } Y = \beta_0 + \beta_1 X_1 + r$
Logarítmico	$\text{LOG } Y = \beta_0 + \beta_1 \text{ LOG } X_1 + r$
Bilineal	$Y = (\beta_0+\beta_1 X_1)+(\beta_2+\beta_3 X_1) X_2 + e$
Inverso	$Y = \beta_0 + \beta_1 / X_1 + e$
Geométrico	$\text{LOG } Y = \beta_0 + \beta^*_1 X_1 + r;\ \beta^*_1 = \log \beta_1$
Lineal múltiple	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e$
Senoidal	$Y = \beta_0 + \beta_1 \text{ sen } X_1 + \beta_2 \cos X_1 + e$
Multiplicativo	$\text{LOG } Y = \beta_0 + \beta_1 \text{ LOG } X_1 + \ldots + \beta_p \text{ LOG } X_p + r$

Recíproco $1/Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + r$

Inverso general $Y = \beta_0 + \beta_1/X_1 + \ldots + \beta_p/X_p + e$

Nótese que las funciones $g_j(\ )$ toman distintas formas según la expresión del modelo a analizar y las transformaciones en la respuesta modifican los supuestos de distribución de probabilidad de la misma y de los errores o residuos.

2.1 Características de un buen modelo

El modelo de regresión que se escoja debe minimizar alguna medida de discrepancia entre los datos observados y los estimados, por ejemplo,

$$\Sigma_i (T_i - \hat{T}_i)^2 \qquad [4]$$

Es importante también, la simplicidad y parsimonia del modelo, en el sentido de retener sólo el menor número de predictores que garanticen una buena estimación de la respuesta sin que ello implique eliminación de predictores importantes. La buena formulación del modelo en el caso de regresión polinomial debe tener en cuenta el concepto jerarquia, es decir, al incluir en el modelo el término $X_{k1}^{s1} X_{k2}^{s2} \ldots X_{km}^{sm}$ con s1, s2,...,sm exponentes enteros, se deben incluir todos los términos que contengan las combinaciones de los productos de X_{k1}, $X_{k2}, \ldots, X_{km}$ con su correspondiente potencia de igual y de menor orden, Peixoto (1987). Por ejemplo, los modelos que incluyan los términos $[1, X_1, X_2, X_1X_2, X^2_1]$ ó $[1, X_1, X^2_1, X^3_1]$ ó $[1, X_1, X_2, X^2_2, X^3_2, X_1X^2_2, X_1X_2]$ son jerárquicamente bien formulados, mientras que el modelo $[1, X_2, X^2_1, X^2_1X_2]$ no lo es, porque no incluye los términos X_1 y X_1X_2. Bajo condiciones muy generales, el espacio de estimación de un modelo de regresión polinomial es invariante bajo transformaciones de escala, si el modelo está jerárquicamente bien formulado, Peixoto (1990).

McCullagh y Nelder (1983) y Peixoto (1987) puntualizan que la no inclusión de términos con potencia de menor orden puede ser aceptable como parsimonioso, cuando el modelo sea usado únicamente con fines predictivos y no se esperan reparametrizaciones o cambios de escala, y no son erróneos cuando el modelo describe leyes exactas de fenómenos físicos y químicos Peixoto (1990).

Todas las variables predictoras que se piense incluir en el modelo deben ser linealmente independientes, es decir no se pueden encontrar constantes C_j, $j = 1,2, \ldots, q$, no todas cero que cumplan

$$\sum_{j=1}^{q} C_j Z_j = 0 \qquad [5]$$

Hay ciertas relaciones entre variables que pueden producir el efecto "ALIASING" que identificadas permiten establecer un mejor modelo en el sentido de parsimonia. McCullagh y Nelder (1983) ilustran este efecto mediante el siguiente ejemplo:

Nominando Z_1=LOG(LARGO), Z_2=LOG(ANCHO), Z_3=LOG(AREA), C=LOG(CONSTANTE), Si AREA=CONSTANTE X LARGO X ANCHO entonces el modelo

$$T = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + e \qquad [6]$$

se puede reescribir, debido a que $Z_3 = C + Z_1 + Z_2$, como

$$T = (\beta_0+\beta_3 C) + (\beta_1+\beta_3) Z_1 + (\beta_2+\beta_3) Z_2 + e \qquad [7]$$

lo que implica estimar sólo tres parámetros y no cuatro como inicialmente se planteó. El buen juicio del analista de datos permitirá identificar en cada caso éste efecto. Del modelo se deben eliminar las variables redundantes: Flury (1989) define Z_j como una variable redundante en un modelo de regresión múltiple si la correlación múltiple

entre T y Z_1, Z_2,...,Z_q no decrece al eliminar Z_j. Este criterio de redundancia a veces se entiende con la existencia de correlación entre variables pedictoras, pero Hamilton (1987) mostró que este hecho no siempre implica redundancia.

El modelo de regresión debe tener especificado su campo de acción o dominio de los predictores, sobre los cuales da buenas predicciones. Si el dominio de los predictores es suficientemente amplio, se puede asemejar esta situación con invarianza paramétrica, es decir, si con otro conjunto de datos relacionados al mismo fenómeno se estiman los parámetros para el mismo modelo, las nuevas estimaciones no cambian sustancialmente. Brooks, et al. (1988) definen el dominio de un modelo de regresión en términos de superficies convexas, que comparativamente a otras definiciones es más conservadora y presentan un programa IML/SAS para determinar si una predicción proviene o no del dominio del modelo y cae dentro del rango válido.

Finalmente, para propósitos de inferencia estadística el modelo no debe violar los supuestos acerca de independencia y distribución de probabilidad de los errores.

2.2 Proceso de Construcción del Modelo

La construcción del modelo, Box, et al. (1978), Box y Jenkins (1976), McCullagh y Nelder (1983), se asemeja a un proceso iterativo entre las etapas: Especificación o identificación, Selección y ajuste y chequeo de diagnósticos. La primera se desarrolla informalmente con la ayuda de gráficos y un análisis preliminar de los datos buscando las

características del modelo para llegar a una clase de modelos candidatos. La segunda, conlleva estimación de parámetros y discriminación entre los modelos candidatos para seleccionar el mejor o los mejores modelos y en la última se analiza que tan adecuado o inadecuado es el modelo seleccinado; los gráficos de los residuos indicarán que modificaciones se deben hacer para los siguientes ciclos del proceso o si es suficientemente bueno el modelo como para terminar el proceso. Box et al. (1978) ilustran mediante un diagrama de flujo las posibles acciones a tomar en estudios conducentes a plantear modelos de superficies de respuesta.

3. Consecuencias de la incorrecta especificación del Modelo de Regresión

Hocking (1976) establece la incorrecta especificaciòn del modelo en los casos en que: se incluyen variables irrelevantes o se descartan variables que son relevantes. Rao (1971) para los modelos de regresión donde todas las variables predictoras no son estocásticas y los términos de error tienen varianza constante e independientes, considera que: la omisión de una variable relevante introduce sesgo en todas las estimaciones de los parámetros por mínimos cuadrados y arroja menor varianza de los mismos. La inclusión de una variable irrelevante no introduce sesgo en las estimaciones de los parámetros, pero incrementa la varianza de todas estimaciones.

La especificación del modelo incluyendo variables irrelevantes sacrifica el tamaño reducido o parsimonia del modelo e incrementa el riesgo de establecer fluctuaciones aleatorias que produzcan efectos engañosos o falsos.

4. Medidas de buen ajuste del modelo

La escogencia del mejor modelo que ajuste a los datos se hace con base en alguna medida que indique que tan bueno es el ajuste, la cual puede ser segun Hocking (1976): Cuadrado Medio de Residuos, R-cuadrado, Suma de Cuadrados de la Predicción PRESS, Suma de Cuadrados Residuales Estandarizados, RSS_p, Cuadrado Medio de Error de Predicción Promedio S_p, el Error Cuadrático Total C_p, Varianza de una Predicción Promedio. Se discuten a continuación los de uso más frencuente.

a. Cuadrado Medio de Residuos:

Esta medida sugiere cual es el tamaño apropiado del modelo, debido a que tiende a estabilizarse cuando se presenta ajuste con inclusión progresiva de más predictores que los necesarios, Draper y Smith (1981). La escogencia entre varios modelos candidatos se inclinará por aquel modelo que dé menor cuadrado medio de residuos. Su desventaja radica en la comparación de modelos que tengan distintas transformaciones para la respuesta.

b. Error Cuadrático total C_p, definido por

$$C_p = \frac{SCR_p}{S^2} - (n-2p) \qquad [8]$$

donde SCR_p es la Suma de Cuadrados de residuos de un modelo que contiene p parámetros, incluído el intercepto, S^2 es el cuadrado medio de residuos del modelo de mayor tamaño propuesto que contenga todos los predictores, tomado como estimación de la varianza de error. Se escoge p como mejor tamaño de subconjunto de variables a retener en el modelo cuando C_p sea igual o próximo a p. Draper y Smith (1981) sugieren intentar estrategias de construcción de modelos más racionales que la de

examinar todos los posibles subconjuntos de variables a retener en el modelo.

c. R-Cuadrado:

Es la medida de buen ajuste de un modelo de regresión. Kvalseth (1985) hace una revisión de las distintas formas de cálculo de R^2 y analiza las circunstancias bajo las cuales una u otra forma puede ser mal empleada. Chang y Afifi (1987) complementan la lista con una modificación de R^2 cuando se tienen repeticiones para las respuestas. Willett y Singer (1988) definen R^2 adecuado cuando se estiman los parámetros por el método de mínimos cuadrados ponderados. Theil y Chung (1988) sugieren otra corrección por número de paràmetros en el modelo como factor en exponente, obviando asi situaciones en los que R^2 podrìa ser negativo. El cuadro 1 presenta las distintas definiciones de R^2.

La apropiada escogencia de R^2, Kwalseth (1985), depende: del tipo de modelo, presencia de β_0 o uso de transformaciones, de la técnica de modelaje utilizada, del propósito con el cual se usa R^2 y de las propiedades que se consideren deseables para R^2.

Una situación frecuente es calcular R^2 para un modelo linealizado con $[T, \hat{T}]$ y compararlo con el R^2 para un modelo lineal con $[Y, \hat{Y}]$. El primero brinda una medida del buen ajuste para el modelo linealizado y no para el modelo no lineal en los parámetros. Para hacer comparables las medidas de ajuste de los dos modelos se debe calcular R^2 para el modelo linealizado usando $[Y, \hat{Y}=T^{-1}(\hat{T})]$ donde $T^{-1}(\hat{T})$ es la transformación inversa de T( ) que opera sobre los valores de predicción del modelo linealizado ajustado.

Cuadro 1. Definiciones de R^2 como medida de buen ajuste de una regresión.

No.	Definición	Observaciones	Referencia
R^2_1	$1-\Sigma(Y-\hat{Y})^2/\Sigma(Y-\bar{Y})^2$	Recomendada para modelos lineales o linealizados con o sin β_0.	Kvalseth (1988) def. R^2_1 a R^2_{11}
R^2_2	$\Sigma(\hat{Y}-\bar{Y})^2/\Sigma(Y-\bar{Y})^2$		
R^2_3	$\Sigma(\hat{Y}-\tilde{\bar{Y}})^2/\Sigma(Y-\bar{Y})^2$	$\tilde{\bar{Y}}$ = media de $\hat{Y}$	
R^2_4	$1-\Sigma(e-\bar{e})^2/\Sigma(Y-\bar{Y})^2$	No recomendada si β_0 se excluye o si $T(Y)\neq Y$	
R^2_5	$R^2_{T,Z1,\ldots,Zq}$	T es la transformación de Y y Z de los predictores	
R^2_6	$r^2_{Y\hat{Y}}$	$\hat{Y}$ valor de predicción después de hacer la transformación inversa de T(Y) sobre $\hat{T}$.	
R^2_7	$1-\Sigma(Y-\hat{Y})^2/\Sigma Y^2$	Recomendada si β_0 no está en el modelo	
R^2_8	$\Sigma\hat{Y}^2/\Sigma Y^2$	Recomendada si β_0 no está en el modelo	
R^2_9	$1-Me^2[\lvert Y-\hat{Y}\rvert]/Me^2[\lvert Y-\tilde{Y}\rvert]$	Me=Mediana, Resistente a datos muy alejados, aquí $\tilde{Y}$ se podría reemplazar por la mediana.	
R^2_{10}	$1-(1-R^2_1).a$	a = (n-1)/(n-q-1) ó a = (n-1)/(n-q) si β_0 no está en el modelo. q = número de predictores.	
R^2_{11}	$1-(1-R^2_9).a$	a = Como en Def. 10	
R^2_{12}	$1-(SCFA)/(SCR+SCFA)$	SCFA = Suma de Cuadrados debido a falta de ajuste SCR = Suma de Cuadrados debido a la regresión	Chang y Afifi (1987)
R^2_{13}	$1-(1-R^2_{12})a$	a = como en Def.10	
R^2_{14}	$1-\Sigma(Y-Y^*)^2/\Sigma(Y-\bar{Y})^2$	Y^* = Predicciones con β estimados por mínimos Cuadrados Ponderados	Willet y Singer (1988)
R^2_{15}	$1-(1-R^2_1)^a$	a = n/(n+q)	Theil y Chung (1988)

Nota: en (5): Cuadrado de la correlación Múltiple entre T y $Z_1, Z_2, \ldots, Z_q$
en (6): Cudrado de la correlación simple entre Y, $\hat{Y}$.

Otro error frecuente es utilizar R'_7 ó R'_8 para modelos que no incluyen β_0 y compararlo con R^2 para modelos que sí lo incluyen. En el primer caso por razones teóricas se fuerza el modelo a no tener intercepto, así los datos sugieran que para el dominio del modelo, β_0 debe ser incluído, Kwalseth (1985). Se recomienda usar modelos sin β_0 cuando tanto las consideraciones teóricas como el dominio del modelo y los datos lo sugieren.

5. La construcción del modelo

La selección del modelo sería más fácil de realizar si se tuviera de antemano un conocimiento exacto de la varianza de los residuos. Como esta varianza es desconocida y un estimador es el cuadrado medio de residuos, surge la incertidumbre sobre si su magnitud es debida a falta de ajuste del modelo o si es natural en los datos. Hay varios métodos expuestos en la literatura sobre la escogencia de la mejor ecuación de regresión, pero varias son las dificultades a resolver en este proceso, siendo algunas de ellas el establecer: Cuál es la transformación más adecuada de la respuesta?, son todos los predictores observados o controlados los que afectan de una manera importante la respuesta?, Cuál es el número de predictores ideal para el modelo?, los predictores afectan linealmente la respuesta?, hay interacciones lineales o de mayor orden entre los distintos predictores?

5.1 Transformación en la respuesta

Cook y Weisberg (1982) identifican tres clases de situaciones que motivan transformaciones de la respuesta.

a. Las respuestas son independientes y provienen de una población con distribución no normal y el objetivo de la transformación es buscar que la nueva forma de la respuesta y sus errores sea suficientemente cercana a la distribución normal con la finalidad de aplicar los métodos basados en esa distribución de probabilidad. McCullagh y Nelder (1983), dentro de los modelos lineales generalizados nominan ésta transformación como funciones de enlace y dependen del supuesto inicial de la distribución de probabilidad de la respuesta y sus errores. (Cuadro 2).

b. Las respuestas esperadas estan relacionadas con las variables predictoras por una función conocida y no lineal en los parámetros y el objetivo de la transformación es linealizar la función de la respuesta, y así poder usar el método de mínimos cuadrados usuales; por ejemplo la ecuación del modelo de regresión exponencial. Aqui es necesario revisar los supuestos acerca de la apropiada estructura de los errores en el modelo no linealizado: aditivos, multiplicativos u otra composición.

c. Tanto la forma funcional de la relación entre la respuesta y sus predictores como la distribución de probabilidad de los errores no son conocidas exactamente. Lo ideal seria obtener una transformación de la respuesta que conduzca a un modelo donde los errores se distribuyan normalmente con media cero y varianza constante. En el Cuadro No. 3 se presentan varias familias de transformaciones de la respuesta sugeridas por Box y Cox, John y Draper, Mosteller y Tukey. La determinación de las constantes L que definen la transformación se hace por el método de máxima verosimilitud, suponiendo que los errores de la respuesta transformada son independientes y se distribuyen normalmente con media cero y varianza constante. Algunas de estas transformaciones buscan

Cuadro 2.: Distribución de probabilidad de la respuesta y su posible transformación.

Distribución	Transformación	Nombre
Normal	$T = Y$	Identidad
Poisson	$T = LOG(Y)$, $Y \geq 0$	Logaritmica
	$T = \sqrt{Y}$, $Y \geq 0$	Raiz cuadrada
Binomial	$T = LOG(Y/(1-Y))$, $0<Y<1$	Logit
	$T = \Phi^{-1}(Y)$, $0 \leq Y \leq 1$	Probit
	$T = Log(-Log(1-Y))$, $0 \leq Y \leq 1$	Log-log complementaria
	$T =$ arcoseno $(\sqrt{Y})$, $0 \leq Y \leq 1$	Arco Seno
Gamma	$T = 1/Y$, $Y \geq 0$	Inversa
Gauss Inversa	$T = 1/Y^2$, $Y \neq 0$	Inversa Cuadrática
Binomial Negativa	$T = \sqrt{(1-Y)} - \sqrt{(1-Y)^3}/3$, $0 \leq y \leq 1$	

Nota: $\Phi^{-1}( )$ función inversa de la Distribución Normal acumulada.

estabilizar la varianza de la respuesta y así dejarla independiente de su valor medio.

Con el objeto de establecer relaciones aproximadamente lineales entre la respuesta y los predictores, Mosteller y Tukey (1977) sugieren algunas técnicas prácticas para encontrar transformaciones apropiadas para las respuestas según sean cantidades y conteos, usando $Log(Y+C)$ o $(Y+C)^d$ con d y C constantes. Para determinar C, Allen y Cady (1982) sugieren para un predictor, tomar 3 puntos suficientemente amplios y equidistantes x, con sus correspondientes valores Y, por ejemplo Y_1, Y_2, Y_3, escogiéndose C tal que cumpla con: $Log(Y_2+C)-Log(Y_1+C) = Log(Y_3+C)-Log(Y_2+C)$
o, definiendo previamente d, $(Y_2+C)^d - (Y_1+C)^d = (Y_3+C)^d - (Y_2+C)^d$
Para varios valores C y d, se pueden comparar distintos ajustes y escoger el que satisfaga algún criterio; por ejemplo mayor R^2.

Cuadro 3.: Familias de transformaciones

Familia de Transformaciones	Transformación		
Potencia	$(Y^{L}-1)/L$	,	con $Y > 0$, $L\neq 0$
	$Log(Y)$		$L=0$
Potencia Extendida	$((Y+L_2)^{L1}-1)/L1$	,	con $Y+L_2 > 0$, $L_1\neq 0$
	$Log(Y+L_2)$	,	$L_1=0$
Módulo	$Signo(Y)[(\|Y\|+1)^{L}-1]/L$		$L\neq 0$
	$Signo(Y)[Log(\|Y\|+1)]$		$L=0$
Potencia Cruzada	$[Y^{L}-(b-Y)^{L}]/L$		con $0 \leq Y \leq b$ $L\neq 0$
	$Log[Y/(b-Y)]$		$L=0$

5.2 Predictores a incluir en el modelo

En Draper y Smith (1981) se discuten los distintos métodos tradicionalmente recomendados para la escogencia de subconjuntos de predictores que ajustan el modelo, pero la mayoría suponen que la relación entre los predictores y las respuesta es lineal. Algunos de ellos han perdido aceptación; por ejemplo, inclusión de variables por pasos (Forward) o el método de todas las regresiones posibles que se considera ineficiente en la medida en que el número de predictores es grande y formas no lineales se presumen.

Otros procedimientos como exclusión de variables por pasos, máximo R^2 y el criterio de la estadística C_p de Mallow, son útiles cuando la forma de cada uno de los predictores es la correcta según, está expresado en el modelo general.

5.3 Análisis de residuos

Recientemente se han desarrollado métodos que mediante graficación y la identificación de patrones en el análisis de residuos permiten decidir

sobre el adecuado ajuste y la necesidad de incluir al modelo nuevos predictores o nuevas re-expresiones de los predictores.

Cook y Weisberg (1982) exponen ampliamente las bases teóricas del análisis de residuos. Para la construcción del modelo resultan de utilidad los residuos ordinarios, los residuos de la variable adicionada, los residuos parciales y los residuos recursivos.

5.3.1 Gráficas de los residuos ordinarios

Estas gráficas consisten en ubicar en un plano con ordenada los residuos $(Y-\hat{Y})$; en la abcisa se intercambian tanto la respuesta esperada $\hat{Y}$ como cada uno de los predictores X_k. La primera gráfica $[e^*\hat{Y}]$, permite determinar si existe falta de ajuste en el modelo al detectarse que los puntos estan siguiendo algun patrón o secuencia. En caso contrario, si los puntos sugieren que están aleatoriamente distribuídos en el gráfico, el análisis podría terminarse porque el ajuste logrado hasta ahora no se puede mejorar o porque las variables utilizadas no tienen capacidad predictiva de la respuesta. Si se determina algún patrón en la gráfica $[e^*X_k]$ se sugiere adicionar al modelo términos X_k con potencia o alguna reexpresión de este predictor. Estos gràficos no son muy adecuados para detectar posibles interacciones entre los predictores.

5.3.2 Gráficas de residuos de variable adicionada

Otro método para determinar la adecuada expresión del predictor X_k en el modelo se logra graficando $[(Y-\hat{Y}_k) * (X_k-\hat{X}_k)]$, donde $\hat{Y}_k = \hat{\beta}^*_0 + \Sigma\hat{\beta}^*_j X_j$, con $j\neq k$, es el valor de predicción de la respuesta excluído el predictor X_k del modelo, $\hat{X}_k=\hat{\alpha}_0+\Sigma\hat{\alpha}_j X_j$, con $j\neq k$, es el valor de predicción de la

variable X_k en un modelo de regresión en términos de los restantes predictores.El coeficiente de la regresión lineal entre $(Y-\hat{Y}_k)$ y $(X_k-\hat{X}_k)$ coincide con $\beta_{k'}$ el coeficiente de X_k en el modelo completo. El análisis del gráfico muestra si la variable X_k está relacionada linealmente con la respuesta: los puntos estarían distribuidos irregularmente alrededor de la recta $(Y-\hat{Y}_k)=\hat{\beta}_k(X_k-\hat{X}_k)$. Alguna curvatura o separación sistemática de la mencionada recta indica que la variable X_k se debe reexpresar o se deben adicionar otros términos X_k con potencia. Esta gráfica es la que el programa REG/SAS imprime con la opción PARTIAL.

5.3.3 Gráficas de Residuos Parciales.

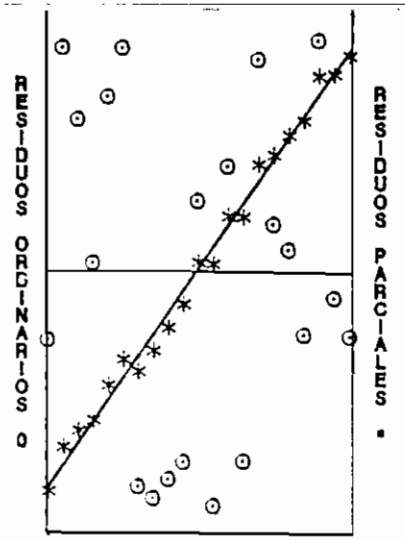
Un sustituto del gráfico de variable adicionada es el gráfico de residuos parciales; ambos utilizan los mismos residuos pero su apariencia puede ser diferente. Consiste en gráficar $[(e+\hat{\beta}_k X_k)*X_k]$, donde $e=Y-\hat{Y}$ son los residuos ordinarios, $\hat{\beta}_k$ es la estimación del coeficiente de regresión de X_k en el modelo completo [1]. Una particularidad importante de $e+\hat{\beta}_k X_k$ es que adiciona la parte irregular del modelo e, con la parte sistemática debida a X_k.

Mansfield y Conerly (1987) muestran la bondad del análisis de residuos parciales simultáneo con los residuos ordinarios; por ejemplo hay situaciones en las que el patrón de los residuos ordinarios sugiere adicionar X_k con término cuadrático y el patrón de los residuos parciales muestra que podría ser más adecuada la transformación logarítmica o una función inversa. En otras ocasiones ambas gráficas sugieren la misma transformación. Si las dos gráficas no presentan

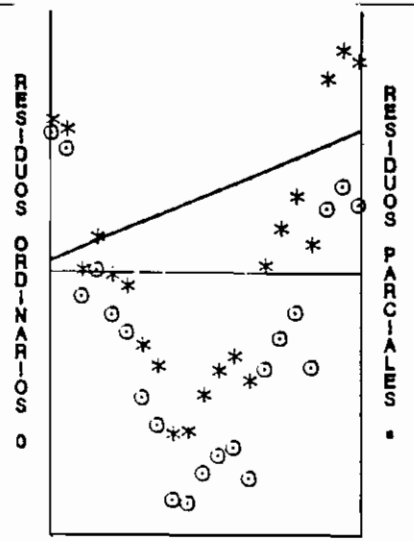
algún patrón por su irregularidad en la distribución de los puntos, se concluye que esa variable o alguna re-expresión de ella no contribuye como buena predictora de la respuesta, por lo tanto $\beta_k=0$, o que las otras variables predictoras distintas a X_k no permiten describir la adecuada forma de X_k.

La excepción de la utilidad de los residuos parciales se presenta cuando existe colinealidad extrema entre una predictora mal especificada en el modelo y otras variables, mostrando el mismo patrón o curvatura en la gráfica de los residuos parciales de todas las variables involucradas, Mansfield y Conerly (1987).

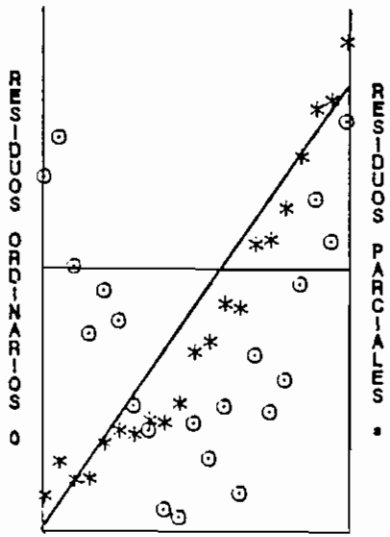
Las gráficas 1 a 6 muestran algunos patrones que en el análisis de residuos y de residuos parciales sugieren transformar el correspondiente predictor. Las zonas extremas de cada gráfica muestran generalmente la curvatura y es de utilidad en el residuo parcial dibujar la parte sistemática debida al predictor X_k, como una línea de referencia. Si en los residuos los puntos estan aleatoriamente distribuídos y en los residuos parciales muestran tendencia lineal, entonces el predictor está bien especificado en el modelo; si en ésta gráfica muestra un patrón distinto, entonces debe ser transformado. Finalmente, si se presume la existencia de interacciones entre los predictores, se pueden incorporar al modelo los productos de esos predictores y hacer el análisis de residuos.



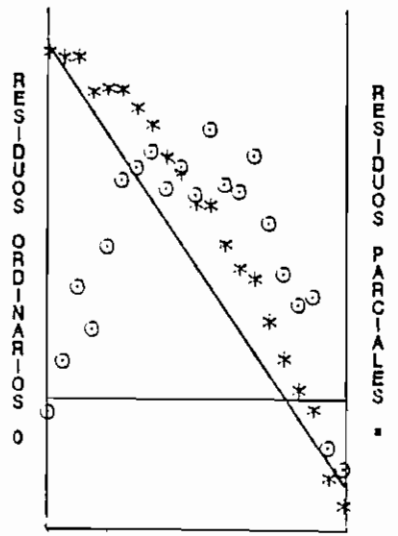
GRAFICA 1.
TRANSFORMACION: IDENTIDAD



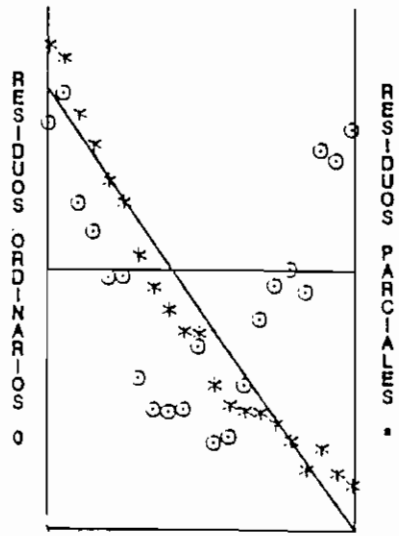
GRAFICA 2.
TRANSFORMACION: Z=(X-C)**2



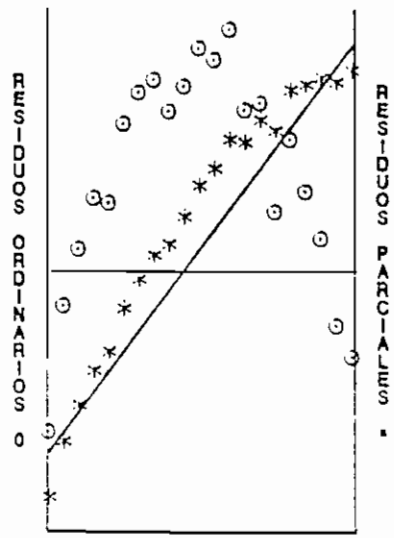
GRAFICA 3.
TRANSFORMACION: Z=(X+C)**2



GRAFICA 4.
TRANSFORMACION: Z=(X+C)**2



GRAFICA 5.
TRANSFORMACION: Z=1/(X+C)



GRAFICA 6.
TRANSFORMACION: Z=LOG(X+C)

BIBLIOGRAFIA

Allen, D.M. y Cady, F.B. (1982). Analyzing Experimental Data by Regression. Lifetime Learning Publications. Belmont, CA

Box, G., Hunter, W. y Hunter, J. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. John Wiley, New York.

Brooks, D.G., Carroll, S.S. y Verdini, W. A. (1988). Characterizing the Domain of a Regression Model. The American Statistician 42: 187-190.

Cook, R.D. y Weisberg, S. (1982). Residuals and Influence in Regression. Chapman & Hall, New York.

Flury, B. (1989). Understanding Partial Statistics and Redundancy of Variables in Regression and Discriminant Analysis. The American Statistician 43: 27-31.

Chang, P.C. y Afifi, A.A. (1987). Goodness-of-Fit Statistics for General Linear Regression Equations in the Presence of Replicated Responses. The American Statistician, 41, 195-199.

Drapper, N.R. y Smith, H. (1981). Applied Regression Analysis (2da. ed.). John Wiley, New York.

Hamilton, D. (1987). Sometimes $R^2 > r^2_{yx1} + r^2_{yx2}$. Correlated Variables Are Not Always Redundant. The American Statistician. 41: 129-132.

Healy, M.J. (1984). The Use of R^2 as a Measure of Goodness of Fit. Journal of the Royal Statistical Society. A. 147: 608-609.

Hocking, R.R. (1976). The Analysis and Selection of Variables in Linear Regression. Biometrics 32: 1-49.

Kvalseth, T. (1985). Cautionary Note about R^2. The American Statistician, 39: 279-285.

Mansfield, E. y Conerly, M.D.(1987). Diagnostic Value of Residual and Partial Residual plots. The American Statistician 41: 107-116.

McCullagh, P. y Nelder, J.A. (1983). Generalized Linear Models. Chapman & Hall, London.

Mosteller, F. y Tukey, J.W. (1977). Data Analysis and Regression Adisson-Wesley, Reading, Mass.

Peixoto, J.L. (1990). A property of Well-Formulated Polynomial Regression Models. The American Statistician 44: 26-30.

Peixoto, J.L. (1987). Hierarchical Variable Selection in Polynomial Regression Models. The American Statistician, 41: 311-313.

Rao, P. (1971). Some notes on Misspecification in Multiple Regressions. The American Statistician, 37-39.

Theil, H. y Chung, Ch. (1988). Information - Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions. The American Statistician, 42: 249-252.

Willett, J.B. y Singer, J.D. (1988). Another Cautionary Note about R^2: Its use in Weighted Least-Squares Regression Analysis. The American Statistician, 42: 236-238.