TRABAJO METODOLÓGICO

Revista Investigación Educativa - Vol. 8 - n.º 15 - 1990 (P. 79-92)

ANÁLISIS DE LAS CONDICIONES DE APLICACIÓN DE LA REGRESIÓN HACIENDO USO DE LOS PAQUETES DE PROGRAMAS

por
Juan Etxeberría Murguiondo
Universidad del País Vasco

INTRODUCCIÓN

Uno de los métodos más clásicos de los empleados en Estadística Aplicada a las Ciencias Humanas y que, con la aparición de las computadoras y los paquetes de programas ha pasado a ser «habitual» en la mayor parte de las investigaciones es el método de Regresión, tanto simple como múltiple. Dicho método se ha mostrado compacto y ha sufrido muy pocas variaciones con la aparición en el mercado de los distintos programas que lo desarrollan.

La facilidad de manejo de los distintos paquetes de programas ha traído consigo un uso y a veces abuso de los distintos subprogramas sin poner una excesiva atención en las condiciones de aplicación que en mayor o menor medida deben cumplir las matrices de datos para poder ser aplicado un determinado análisis estadístico. Este problema queda claramente evidenciado en el modelo de regresión.

Hoy día, los distintos paquetes de programas nos permiten detectar los posibles errores o desviaciones a las condiciones de aplicación de los modelos, que en el caso de la regresión pueden distorsionar fuertemente los resultados del ajuste.

Surge implícito con el empleo de la regresión el problema de la causalidad. Es ya por todos admitido que, aunque al hablar de la regresión estamos hablando en términos de causalidad, no estamos hablando en términos condicionales. Los coeficientes de correlación y de determinación sólo nos dan información del grado de asociación entre las variables. Es imposible determinar empíricamente las leyes de la causalidad. Por ello quizá fuera más correcto hablar en términos correlacionales. Bartolomé, M. (1978, pág. 10) relaciona los estudios correlacionales y los predictivos y afirma: «Prácticamente, y a nivel de comprensión estadística, la relación entre

ambos índices —correlación y regresión— es tan profunda que un estudio nos lleva con facilidad a otro». Más adelante afirma: «En realidad es el conocimiento lo que nos permite realizar una predicción relativamente correcta en un momento determinado».

Sin embargo a la hora de analizar las condiciones de aplicación de la regresión debemos distinguir los dos planos: el correlacional o descriptivo y el inferencial.

1. CONDICIONES DE APLICACIÓN

Las hipótesis que se deben verificar para utilizar cualquiera de los dos modelos antedichos son:

i) LINEALIDAD: La relación entre las variables independientes y la variable dependiente es lineal.

ii) ESPECIFICACIÓN DEL MODELO: El modelo general está correctamente especificado, es decir, no se han excluido variables relevantes ni se han incluido en el modelo variables irrelevantes.

iii) INDEPENDENCIA DE LOS ERRORES: Las variables independientes se miden sin error, esto es, las distintas observaciones de las variables son valores exactos.

iv) INSESGADEZ de modelo.

Las hipótesis adicionales necesarias para considerar la regresión desde el punto de vista inferencial son las siguientes:

v) HOMOCEDASTICIDAD: Las varianzas de las distribuciones de la variable dependiente Y_i ligadas a los distintos valores X_i de las variables independientes son todas iguales. Esto es, las distribuciones condicionales de los residuos tienen igual varianza.

vi) INDEPENDENCIA: Existe independencia si el valor observado en un sujeto no está, en ningún sentido, influenciado por valores de esta variable observados en otros sujetos. Esto es: las variables Y_i e Y_j ligadas a cualquier par de valores X_i y X_j son estocásticamente independientes.

vii) AUSENCIA DE MULTICOLINEALIDAD: Las variables explicativas deben ser independientes entre sí. Se produciría multicolinealidad si existe alguna variable independiente que esté fuertemente correlacionada con otra o con un conjunto de variables independientes.

viii) NORMALIDAD: La distribución de la variable formada por los residuos debe seguir una distribución normal.

Es en base a estas hipótesis, como se trabaja y a partir de las que luego se realiza el estudio de la bondad del ajuste, lo que normalmente equivale a comprobar el poder de predicción.

Anscombe, ya en el año 1973, plantea la necesidad del estudio de las hipótesis o condiciones de aplicación de la regresión y el posterior estudio de los residuos. Planteó los siguientes cuatro conjuntos de datos, cada uno de ellos formado por once pares de puntos (X_i, Y_i) (tabla 1) con análogos valores de medias, varianzas,

coeficientes de correlación y regresión, y que sin embargo obedecen, como podemos ver, a nubes de puntos absolutamente distintas.

TABLA 1

N	X1	Y1	Y2	Y3	X2	Y4
1	10	8.04	9.14	7. 46	8	6.58
2	8	6.95	8. 14	6. 77	8	5. 76
3	13	7.58	8. 74	12.74	8	7.71
4	9	8.81	8.77	7.11	8	8.84
5	11	8.33	9.26	7.81	8	8.47
6	14	9.96	8.10	8.84	8	7.04
7	6	7.24	6.13	6.08	8	5.25
8	4	4.26	3.10	5.39	19	12.50
9	12	10.84	9.13	8.15	8	5.56
10	7	4.82	7.26	6.42	8	7.91
11	5	5.68	4.74	5.73	8	6.89

(Fuente: Anscombe. 1973)

Las representaciones gráficas de estos conjuntos de datos son las siguientes:

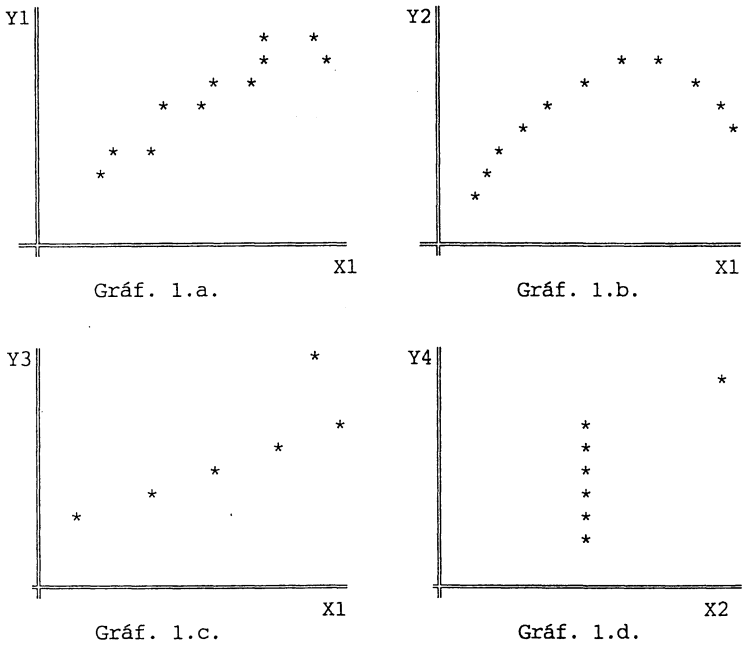

GRÁFICO 1 (Fuente: Anscombe. 1973)

Sin embargo el más simple análisis de los casos demuestra que no son igualmente válidos los resultados de las distintas regresiones.

El primer conjunto de datos (Gráf. 1. a) es el único que responde a un modelo apropiado para poder aplicar la regresión.

El segundo conjunto de datos (Gráf. 1. b) sugiere evidentemente una regresión polinomial.

El gráfico 1. c, nos indica que la regresión simple podría ser adecuada, para la mayor parte de los datos, pero uno de ellos evidentemente distorsiona el ajuste. Es de los llamados datos aberrantes (outlier). Si eliminamos este dato la ecuación de regresión pasaría a ser $Y' = 4 + 0.\ 346\ X$ con un coeficiente de determinación: $R^2 = 1$. Sin un estudio contextual de los datos no podríamos juzgar cuál de los dos ajustes es el correcto.

Por último, el gráfico 1. d, es distinto de los anteriores. En este caso el modelo parece correcto. Pero podemos darnos cuenta que estos resultados no son satisfactorios. Si se suprimiera la puntuación del sujeto octavo, no sería posible hacer los cálculos, la varianza es nula.

Evidentemente estos son casos extremos, que en la práctica no se encuentran de forma tan neta. En cualquier caso, deberemos poner los medios para detectar los posibles problemas, pues si el incumplimiento de las hipótesis no es muy significativo, las conclusiones pueden seguir siendo válidas, debido a la robustez del método, pero si el incumplimiento de las hipótesis o las desviaciones al modelo son más importantes, las conclusiones pueden llegar a perder toda su validez.

Hasta aquí hemos visto algunos de los problemas que puede plantear el modelo general de la regresión. Vamos a analizar algunos de los más habituales en Estadística de las Ciencias Humanas, abordando la forma de detectarlos y posibles soluciones a los mismos.

Tal como indican Cohen & Cohen (1983) el análisis de los residuos de los valores estimados en la regresión, nos proporciona la base para asegurar la adecuación del modelo de regresión.

Para realizar dichos análisis los métodos gráficos son muy usados, y, debido a ello los diversos paquetes de programas estadísticos ofrecen diversas posibilidades para realizar representaciones gráficas de los residuos.

Detallaremos a continuación los aspectos más importantes que se deben tener en cuenta al hacer uso del método de regresión.

2. ESPECIFICACIÓN DEL MODELO-SELECCIÓN DE VARIABLES

A la hora de realizar cualquier investigación a nadie se le escapa la importancia que tiene el diseño del mismo. Si en general es muy importante, lo es más cuando usamos el método de la regresión.

Está unánimemente admitido que el análisis de la secuencia de entrada de las variables explicativas en el modelo nos proporciona una información muy impor-

tante. Es éste un tema del que se ha escrito mucho, hay importantes estudios realizados al respecto. Pardoux (1982), en un trabajo sobre el tema indica al respecto que el problema de la selección de entrada de variables no debe ser abordado más que después de un análisis serio de los datos. El estadístico no debe aceptar los resultados dados por un método como la solución definitiva. Debemos citar al respecto, el trabajo presentado en el n° 4 de esta misma revista por Juan Mateo y Sebastián Rodríguez.

Los diversos métodos de selección de variables, tienen como objetivo el buscar el mejor subconjunto de variables según diversos criterios de «calidad» de la predicción.

Entre los métodos más usuales podemos realizar la siguiente clasificación:

a) Métodos por etapas: consisten en buscar un subconjunto de variables explicativas, incorporando o eliminando las mismas de la ecuación (según sea ascendente o descendente) de una en una, atendiendo a diversos criterios. Entre estos métodos podemos destacar.

1. *Forward*: Criterio que se emplea en los métodos por etapas ascendentes. En cada paso, las variables que todavía no han entrado a formar parte de la ecuación son examinadas. De éstas entra a formar parte de la ecuación aquella cuya F, sea más significativa, caso de que supere el límite previamente establecido.

2. *Backward*: Criterio que se emplea en los métodos descendentes. En cada paso se examinan todas las variables incluidas en la ecuación. La variable cuya F sea más baja, si no llega a cumplir el requisito preestablecido de significatividad es eliminada.

3. *Stepwise*: Procedimiento que podríamos calificar como de ascendente mixto. Es un procedimiento que se emplea en los métodos ascendentes. Las dos primeras variables son introducidas por el procedimiento Forward. A continuación en dicha ecuación se emplea el método Backward. Caso de no haber sido eliminada ninguna variable se introduce una nueva variable por el criterio Forward, y así sucesivamente hasta llegar a una ecuación estable, en el sentido de que no pueda entrar ninguna variable por no ser significativa su aportación al coeficiente de determinación, ni pueda ser eliminada ninguna variable previamente incluida en la ecuación.

Resulta importante señalar que los distintos métodos por etapas no dan siempre los mismos resultados cuando se ponen en práctica sobre los mismos datos. Ninguno de ellos garantiza la obtención del mejor subconjunto de p variables explicativas. Esto sí que lo garantiza el siguiente procedimiento:

b) Métodos que tratan de buscar el mejor subconjunto de variables en base a estudiar todas las posibles combinaciones de variables. Estos métodos tratan, mediante algoritmos pertinentes, de determinar el número de variables y la elección de las mismas, generalmente en base a minimizar la suma de cuadrados de los residuos debidos a la regresión, en términos «ajustados». En principio son métodos más laboriosos, hay que tener en cuenta que con «r» variables es posible realizar $2^r - 1$ subconjuntos de variables.

El B. M. D. P. en su programa P9R ofrece la posibilidad, aunque reducida, de trabajar con todos los posibles subconjuntos de la regresión.

c) Método completo. Incluye todas las variables a la vez. Es un método que prácticamente ha caído en desuso porque tiene la desventaja, con respecto de los métodos por etapas, que no nos proporciona la valiosa información que nos aporta la secuencia de entrada o de «caída» de las variables.

Nosotros nos hemos inclinado por el método por etapas «Stepwise» por considerarlo el más completo de los que disponemos ya que no hemos podido conseguir el P9R antes mencionado ni el MAXR que ha sido incorporado al S. A. S., y del que últimamente hay bastante literatura.

Es preciso indicar que no siempre al introducir nuevas variables en el modelo, obtendremos mejores resultados. Diferentes autores indican que a partir de la cuarta o quinta variable explicativa (independiente) el incremento del coeficiente de determinación raramente es significativo, siendo a veces negativo si trabajamos con coeficientes ajustados.

Sería deseable que se incluyeran en el diseño todas las variables que tuvieran alguna relación con la/s variable/s dependiente/s, y no se incluyeran las variables irrelevantes. Pero como dice Fox D. J. (1981) (pág. 511):

> «En la situación ideal, los experimentos se diseñan de tal modo que existe una relación directa entre las variables independientes y dependientes (relación directa en el sentido de que sea razonable pensar que las diferencias en las variables dependientes, después de hacer el experimento, se puedan atribuir al efecto de las variables independientes). En la mayoría de las investigaciones educativas este diseño ideal no es realizable ya que entre las variables independientes y las dependientes hay otras, las variables «extrañas» que, en el caso más sencillo, se pueden definir como el conjunto de condiciones que impiden atribuir todas las diferencias observadas en las variables dependientes a las variables independientes».

En algunas ocasiones es posible, a través del análisis de los residuos, detectar la ausencia de variables relevantes. Nos interesa que toda variable independiente que hayamos introducido en la regresión tenga un coeficiente de correlación cero con los residuos. Por otra parte, si una variable no incluida en el modelo tiene una alta correlación con los residuos, evidentemente, al introducirla en el modelo como variable independiente reducirá el error, aumentando consecuentemente el coeficiente de determinación.

Vemos que el problema de la especificación del modelo es bastante arduo en Estadística de las Ciencias Humanas, aunque se debe intentar afinar lo máximo posible. En cualquier caso, conviene dejar bien claro que, en última instancia, es el interés del investigador y el modelo teórico que se proponga, el criterio fundamental a tener en cuenta.

3. LINEALIDAD DEL MODELO

El punto de partida general del modelo es la linealidad del mismo. Esto es, la relación entre la/s variable/s independiente/s y la variable dependiente es lineal. Es una suposición cómoda y que presenta numerosas ventajas. Sin embargo, no siempre las relaciones son lineales o, cuando menos, se podrían conseguir mejores aproximaciones por métodos no lineales.

Para poder darnos cuenta de la no linealidad del modelo, aparte de la, a veces, muy clarificadora representación gráfica de los datos, podemos hacer uso de los siguientes métodos:

a) Coeficientes de correlación no lineales.

b) La regresión polinómica: La regresión polinómica también podemos considerarla como regresión lineal, dado que existen fáciles transformaciones de las variables que las linealizan.

Para poder trabajar con la regresión polinómica, los diversos paquetes de programas tienen ya desarrollados distintos métodos.

El B. M. D. P. (1985) lo desarrolla a través de su programa P5R.

El S. P. S. S. (1986) no desarrolla en esta última versión un programa específico, pero a través de los COMPUTE podemos fácilmente generar nuevas variables con las cuales podemos realizar una regresión polinómica.

El A. D. D. A. D. (1986), versión para micros del S. P. A. D., tampoco tiene un programa específico de regresión polinómica, pero al igual que en el S. P. S. S., podemos generar fácilmente nuevas variables para poder realizar dicha regresión.

c) Representación gráfica de los residuos. El gráfico 2 es un ejemplo claro de no linealidad de los residuos. Sobre todo, cuando estamos trabajando con varias variables independientes, es conveniente realizar la representación gráfica de los residuos en función de las predicciones del modelo.

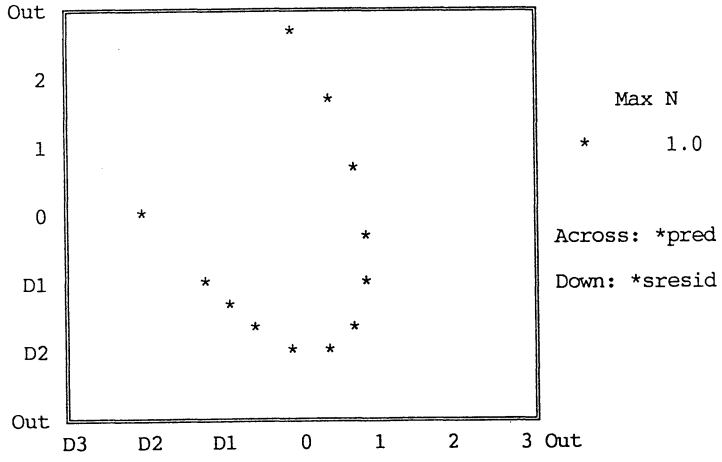

GRÁFICO 2. RESIDUOS NO LINEALES. SALIDA DE S.P.S.S.

El S. P. S. S., por ejemplo, nos ofrece varias alternativas de representación que nos permiten detectar la no linealidad, como pueden ser:

— representación de los residuos studentizados en función de las predicciones del modelo: (*pred, *sresid).

— representación de los residuos studentizados en función de los valores de la variable dependiente o de cualquiera de las variables independientes: (variable, *sresid).

Recordemos que, para que se cumpla la condición de linealidad, la distribución de los puntos en la gráfica debe ser aleatoria. Podemos darnos cuenta que en la que presentamos, esto no se verifica.

Caso de conocer previamente la no linealidad del mismo y la función que relaciona las variables, podemos ver en la tabla 2 distintas transformaciones muy usuales, para poder linealizar el modelo. En cualquier caso, siempre podemos realizar desarrollos en serie de las funciones matemáticas que nos transforman las mismas en polinomios.

TABLA 2 : FUNCIONES LINEALIZABLES

Forma inicial	*Forma linealizada*
1 $\Gamma = \beta_0 x^{\beta_1}$	$\log(\Gamma) = \delta_0 + \delta_1 \log x$
2 $\Gamma = \beta_0 \exp(-\beta_1 x)$	$\log(\Gamma) = \delta_0 + \delta_1 x$
2 $\Gamma = \beta_0 \exp(-\beta_1 x) + \beta_2$	$\log(\Gamma - \beta_2) = \delta_0 + \delta_1 x$
4 $\Gamma = \beta_0 \exp(-\beta_1 x)^{\beta_2}\ x > 0$	$\log[-\log(\Gamma/\beta_0)] = \delta_0 + \delta_1 \log(x)$

Por último indicar que el Cálculo Numérico ha desarrollado numerosas técnicas de aproximación de funciones que nos podrían servir para determinar la función que relaciona óptimamente las variables.

Hay modelos que no se pueden linealizar. Por ejemplo es bastante usual, sobre todo en Estadística aplicada a la Ingeniería o a la Física, intentar el ajuste a la nube de puntos por medio de ecuaciones del tipo:

$$Y = \frac{\beta_0 + \beta_1 \beta_2 e^{\beta_2 (\beta_3 - \beta_4) X}}{1 + \beta_2 e^{\beta_3 X}}$$

Este tipo de ecuaciones se ajustan muy bien sobre todo a modelos de crecimiento.

Podría ser una tentación el intentar trasplantar este tipo de modelos a la Estadística de las Ciencias Humanas y en concreto a las variables educativas, pero así como en Física o Ingeniería, existe una justificación teórica de la aplicación de estos modelos, en la disciplina que nos ocupa no podemos partir de este supuesto, con lo que estamos obligados a abordar el problema desde otros puntos de vista. Única-

mente indicar que a las matrices de datos que me ha tocado analizar no he obtenido resultados satisfactorios con esta última transformación.

En cualquier caso, debemos tener cuidado a la hora de realizar transformaciones que nos permitan linealizar el modelo, pues hay que tener en cuenta que todo modelo tiene una parte aleatoria o de error y, al efectuar las transformaciones, los errores también pueden resultar profundamente alterados. Recordemos que la distribución de los errores debe ser Normal.

4. HETEROCEDASTICIDAD

Hemos indicado como condición de aplicación que las distribuciones condicionales de los residuos deben tener igual varianza. Si observamos el gráfico 3 podremos apreciar que la variabilidad de los residuos no es constante a lo largo del campo de existencia de la predicción de la varible dependiente. Es evidente pues que la varianza no es constante. En el mismo gráfico podemos comprobar que la variabilidad del residuo aumenta a medida que aumenta el valor de la predicción. Evidentemente este hecho afectará a la predicción que posteriormente vayamos a realizar, siendo alterados los intervalos de confianza.

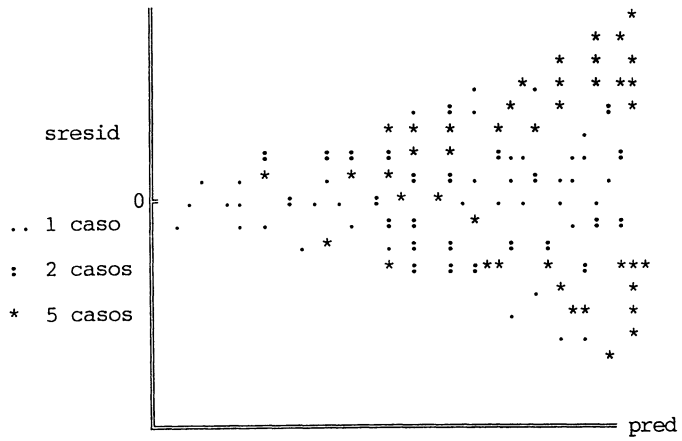

GRÁFICO 3. RESIDUOS HETEROCEDASTICIDAD. SALIDA S. P. S. S.

Frecuentemente este problema de heterocedasticidad viene provocado por la no linealidad del modelo.

Para poder detectar el problema podemos partir de las mismas representaciones gráficas que hemos realizado para comprobar la linealidad del modelo. Si observáramos cualquier posible irregularidad, en el sentido que tratamos, deberíamos realizar un análisis de comparación de varianzas por medio del test F para comparar las mismas en los distintos rangos de la variable predecida, o en su caso, de las variables independientes. En caso de que estemos trabajando con una matriz de datos no excesivamente extensa, podemos detectar el problema por medio de los residuos studentizados. En el S. P. S. S lo podemos hacer con el (*PRED, *SRESID).

Podemos también detectar este error calculando el coeficiente de correlación entre la variable predecida y los residuos. Para subsanar este problema Cohen & Cohen (1983) indican que la solución del mismo, gran parte de las veces, puede venir dada por una transformación no lineal de la variable dependiente, indicando a su vez varias transformaciones que habitualmente resuelven el problema.

El paquete B. M. D. P. en su programa 7D, nos permite trabajar con la prueba de Levene, cuyo desarrollo se puede ver en el artículo de Bisquerra, R. publicado en el número 9, de esta revista.

5. AUTOCORRELACIÓN

Una de las desviaciones que con más frecuencia se da, al aplicar en Estadística de las Ciencias Humanas el modelo de regresión, es éste de autocorrelación.

Debido posiblemente a los problemas de medición que existen a veces en la disciplina que nos ocupa, en muchas ocasiones hemos podido detectar esta situación. La autocorrelación trata de estudiar los problemas de medición «sistemáticos». Si, por ejemplo, pasamos un test colectivo en diversas aulas, y, en alguna de ellas se ha cometido algún error en la pasación de la prueba, este error nos producirá un problema de autocorrelación.

Para poder detectar la existencia del problema, como en los casos anteriores podremos hacerlo por medio de representaciones gráficas. En este caso nos interesaría tener la gráfica de los residuos de los distintos sujetos según el orden de recogida de los mismos, para poder detectar posibles fallos metodológicos en la recogida de los datos. El S. P. S. S. nos da estos resultados por medio del CASEWISE=ZRESID, si nos interesan los residuos standarizados. El inconveniente que puede tener este procedimiento es el número de elementos de que dispongamos, porque su análisis debe de ser sujeto a sujeto, con lo que el mismo puede resultar tedioso.

Es posible detectar la existencia del problema si comprobamos que existen series de sujetos cuyo residuo está sesgado en el mismo sentido.

El test de hipótesis que nos permite decidir sobre la existencia o no del problema es el contraste de Durbin-Watson, que es el test que realiza el SPSS bajo el subcomando DURBIN.

Para solucionar esta cuestión deberíamos analizar las circunstancias de la recogida de estos datos y, una vez analizadas las posibles causas del sesgo, mediante factores de corrección rectificar en lo posible los mismos. En caso contrario no nos quedaría otra alternativa que eliminar dichos datos.

6. MULTICOLINEALIDAD

Cuando hemos analizado la especificación del método hemos indicado que muchas veces no existe una relación directa entre el número de variables del modelo y la bondad del mismo. Una de las causas de la no relación mencionada es el problema de la multicolinealidad.

Una vez conseguida la ecuación de regresión, a la hora de interpretar los distintos coeficientes de la misma, debemos partir siempre del siguiente supuesto: Las distintas variables que forman parte del conjunto de las independientes son incorreladas entre sí. Esto vendría a significar que si una variable independiente cualquiera aumenta en una unidad, permaneciendo costantes el resto de las variables, la variable dependiente debe aumentar o disminuir en su caso la cantidad indicada por el coeficiente de regresión de la variable tratada. Ahora bien, en el supuesto que dos de las variables independientes estuvieran correlacionadas, al aumentar en una unidad una de ellas, la otra no podría permanecer constante, con lo que resultaría imposible estudiar los efectos individuales de dichas variables. Por otra parte es importante simplificar, en la medida de lo posible, el modelo de regresión para una más fácil interpretación del mismo, con lo que la introducción en el modelo de variables que explican la misma parte de varianza que ya estaba previamente explicada por otras variables ya presentes en el mismo, no nos conduce más que a una complicación innecesaria del modelo.

Para poder explorar la multicolinealidad de las variables los procedimientos más simples pueden ser los cálculos de los diferentes coeficientes de correlación, tanto parciales como semiparciales.

Si contamos con un gran número de variables independientes, puede ser interesante, y a veces imprescindible, el realizar un análisis factorial previo de las mismas, para poder elegir entre ellas las variables más ortogonales posibles.

Los paquetes de programas, cuando realizan la regresión paso a paso (STEPWISE), tienen en cuenta el supuesto anterior a la hora de introducir nuevas variables en la ecuación.

El S. P. S. S. en el momento de introducir nuevas variables en la ecuación tiene en cuenta, entre otras cosas, el coeficiente de correlación parcial y la tolerancia, definida como la proporción de la varianza de la variable independiente que no es explicada por las variables previamente introducidas en la ecuación de regresión.

7. NORMALIDAD DE LOS RESIDUOS

La última de las hipótesis o condiciones de aplicación, y, una de las más importantes desde el punto de vista inferencial, que se debe cumplir para una correcta aplicación del método de regresión es, como hemos indicado al comienzo de este artículo, el que la distribución de los residuos siga una distribución normal.

La distribución normal de los residuos es necesaria para poder justificar el empleo de los test F de Snedeckor y la t de Student, así como para construir los distintos intervalos de confianza.

El problema de la No-normalidad de los residuos lo podemos detectar mediante varios procedimientos. Un primer procedimiento puede ser la representación gráfica del histograma de los residuos standarizados. Un segundo procedimiento puede ser el hacer uso de los tests no paramétricos para comprobar la normalidad. Los test más habituales son los de la ji-2 y el de Kolmogorov-Smirnov

También en este caso los distintos paquetes de programas disponen de métodos que nos permiten detectar esta anomalía. Así, por ejemplo, el S. P. S. S. nos posibilita obtener la representación gráfica de los residuos tipificados y realizada su aproximación a la probabilidad teórica mediante el NORMPROB (ZRESID). Los distintos test no parámetricos los podemos encontrar entre los NPAR-TESTS.

8. ELEMENTOS ABERRANTES-OUTLIERS

A pesar de no ser un problema de condiciones de aplicación en el sentido estricto del término, por los problemas que para la aplicación del modelo pueden tener estos elementos, he incluido este último apartado en el artículo.

Debemos empezar intentando definir un elemento como aberrante o «outlier». En términos de regresión parece que se está llegando a un criterio unánime, si bien todavía hay definiciones tan vagas como la que da Mickey, M. R. en el manual de B. M. D. P. (en Dixon (1985) pág. 698) cuando los define como «aquellos casos para los que el valor de la variable dependiente está relativamente poco explicada por la apropiada ecuación de regresión». Estoy más de acuerdo en definir un elemento como aberrante en el sentido que lo considera el S. P. S. S., «aquellos elementos cuyo residuo studentizado/estandarizado es mayor que tres unidades».

De lo que no cabe ninguna duda es, de los efectos negativos que para la bondad del ajuste producen los citados elementos. En primer lugar como hemos podido apreciar en las figuras l. c y l. d nos distorsionan, a veces gravemente, el ajuste. En segundo lugar el peso o contribución a la ecuación de regresión es significativamente mayor que la del resto de los datos. Debemos recordar que el método de los mínimos cuadrados minimiza el cuadrado de los residuos, y estos datos nos aumentan la varianza de los mismos reduciendo consecuentemente el coeficiente de determinación.

La forma de detectar la existencia de elementos aberrantes es diferente en el SPSS. y el BMDP. Parten de dos definiciones distintas de elemento aberrante. El

método seguido por el BMDP está desarrollado por Gaviria, J. L. (1987) en un artículo en el que analiza la problemática de los outliers en los análisis multivariantes.

El SPSS, empleando el criterio antes expresado, los indica como OUT, cuando le solicitamos el RESIDUALS=HISTOGRAM.

En cualquier caso, usando el criterio del SPSS, nos bastaría realizar una representación gráfica de los residuos normalizados, y considerar como aberrantes aquellos que quedan más distantes que tres desviaciones típicas de la media.

Hay otras formas de detección en base a las distancias de Mahalanobis o de Cook que vienen desarrolladas en Weisberg (1980).

¿Cómo solucionar el problema de los elementos aberrantes?

Tampoco hay criterios claros a la hora de atacar este problema. Parece evidente que el estudio de los mismos debe ser absolutamente particularizado. En este estudio deberíamos analizar las causas de la aberración.

Hay tres posibles razones por las que un elemento puede ser aberrante.

a) Es un problema de medición y la causa que justifica esta desviación no ha influido más que en este o estos casos aisladamente.
b) El modelo matemático no es exacto.
c) La probabilidad de que tal evento ocurra suponiendo que el modelo es exacto es muy pequeña.

Si la razón es la primera, la exclusión de dicho/s elemento/s no nos causa ningún problema estadístico. En el segundo caso es evidente que debemos replantearnos el modelo empleado. Lógicamente la polémica surge en el tercero de los casos.

En este último caso y ante la tentación de eliminarlos, la pregunta es ¿qué ganamos eliminando dichos elementos? Es cierto que conseguiremos un mayor coeficiente de determinación, pero artificialmente conseguido. Por otra parte ¿cuándo una probabilidad es pequeña?

Parece claro que eliminarlos no nos aporta nada positivo. Hay autores que indican la posibilidad de sustituir los valores que toman dichos elementos por valores que se desvían tres desviaciones típicas de la media en aquellas variables que los sobrepasen. Pienso que es una solución muy arbitraria, máxime pensando que un elemento no es aberrante por el valor que toma en una variable sino por las interrelaciones que se dan entre las distintas variables del modelo.

Considerando que estamos hablando en términos inferenciales, si lo que perseguimos es minimizar el error máximo, una solución sería realizar el modelo de regresión en función del criterio Minimax.

Otra posible solución seria el realizar dos regresiones una con todos los datos y la otra eliminando de la matriz los elementos aberrantes y obteniendo la ecuación de regresión definitiva en base a la ponderación de las dos ecuaciones obtenidas.

Para terminar sólo quiero indicar que el BMDP y el SAS pueden todavía ofrecer otras alternativas a la hora de detectar posibles desviaciones al teórico modelo matemático necesario para una correcta aplicación de la regresión. El no contar con el paquete

SAS, y carecer de algunos de los subprogramas del BMDP hace que forzosamente la anterior exposición no sea todo lo exhaustiva como hubiera sido de desear.

Quiero finalizar el presente trabajo reseñando lo que indican Mateo y Rodríguez (1984, pág. 121) al referirse a la regresión «. . . dicho análisis, si bien constituye uno de los instrumentos de análisis más potentes que poseemos, debemos usarlo con precisión y precaución, basándonos en un auténtico conocimiento de sus posibilidades y limitaciones».

BIBLIOGRAFÍA

ANSCOMBE, F. J. (1973): Graphs in statistical analysis. American *Statiscian,* n. º 27, 17-21.

BARTOLOMÉ, M. (1978): *Estudios correlacionales y predictivos en la investigación pedagógica.* Universidad de Barcelona. Barcelona.

BISQUERRA, R. (1987): La prueba de Levene para la homogeneidad de varianzas en el BMDP. *Revista de Investigación Educativa.* N. º 9, 79-85.

COHEN, J. & COHEN, P. (1983): *Applied Multiple Regression/Correlation analysis for the behavioral* Sciencies Hillsdale. Nueva York.

COOK, R. D. & WEISBER G, S. (1986): *Residuals and influence in regression.* Chapman and Hall. Nueva York.

DIXON, W. J. et alii *(1983) BMDP Statistical Software.* University of Calififormia Press. Los Ángeles.

DRAPER, N. & SMITH, H. (1981): *Applied regression analysis.* John Wiley & sons. Nueva York.

FOX, D. J. (1981): *El proceso de investigación en educacion.* Eunasa. Pamplona.

GAVIRIA, J. L. (1987): *La detección de outliers en los análisis multivariantes.* Comunicación presentada en II Congreso Mundial Vasco. Bilbao.

LEBEAUX, M. O. (1986): *ADDAD: Logiciel d' analyse des données. Manuel de reference.* A. D. D. A. D. Paris.

MATEO, J. RODRÍGUEZ, S. (1984): Precisiones y limitaciones explicativas en los métodos correlacionales. Alternativas pedagógicas. *Revista de Investigación Educativa,* n.º 4, 103-132.

MICKEY, M. R. (1985): *Detecting outliers with Stepwise Regression.* En DIXON, W. J. et alii (Op. cit.).

NORUSSIS, M. J. (1986): S. P. S. S. / P. C. (+) SPSS. Inc. Chicago.

PARDOUX, C. (1982): Sur la selection de variables en regression multiple. *Cahiers du bureau universitaire de recherche opérationelle.* Tiré a part n.º 40. Paris.

PEDHAZUR, E. J. (1982): *Multiple regression in behavioral research.* Holt. Nueva York.

ROUSSEEUW, P. J. & LEROY, A. M. (1987): *Robust regression & outlier detection.* John Wiley & Sons. Nueva York.

SUICH, R. & DERRINGER, G. C. (1977): Is the regression equation adequate? One criterion. Technometrics, n.º 19, 213-216.

SUICH, R. & DERRINGER, G. C. (1977): Is the regression equation adequate? A further note. *Technometrics* n.º. 22, 125-128.

TOMASSONE, R. LESQUOY, E. & MILLIER, C. (1983): *La regression. Nouveaux regards sur une ancienne méthode statistique.* Masson. Paris.

VINOD, H. & ULLAH, A. (1981): *Recent advances in regression methods.* Marcel Dekker. Nueva York.