Revista Colombiana de Estadística
Vols. 21-22, 1990

SUPERFICIES DE RESPUESTA PARA ANALISIS DE DATOS CATEGORICOS

Juan Ramos
Profesor Asociado
Universidad Nacional

Beatriz García Peña
MS. Univ. Nacional de Col.

Resumen. El presente artículo describe como se pueden construir superficies de respuesta para datos categóricos suceptibles de ser analizados con la metodología de Grizzle, Starmer y Koch.

La concepción de Superficies de Respuesta para análisis de datos categóricos se justifica en su aplicabilidad en el tratamiento de información cualitativa.

1. Marco conceptual.

Las superficies de respuesta para análisis de datos categóricos se enmarcan dentro de la Metodología de Grizzle, Starmer y Koch y la Metodología de Superficies de Respuesta.

A. Metodología de Grizzle, Starmer y Koch.

Grizzle, Starmer y Koch enmarcan su metodología dentro de los mínimos cuadrados ponderados y los modelos de regresión.

Su análisis es apropiado cuando:

- Las variables estudiadas son nominales, ordinales o continuas agrupadas, es decir, categóricas.

- Las variables se diferencian en variables factor y variables dependientes.

- Las observaciones objeto de estudio son clasificadas en tablas de contingencia.

El esquema de muestreo considerado en la metodología G.S.K. es muestreo multinomial con parámetros n_i y las probabilidades poblacionales correspondientes a la fila. La distribución de las diferentes filas es independiente.

El objetivo de esta metodología es hacer inferencia acerca de la estructura subyacente de la tabla de contingencia; especificada por las proporciones poblacionales por celda desconocidas $\{\pi_{ij}\}$.

En términos generales esta metodología incluye dos pasos:

- La construcción de una función de las proporciones observadas objeto de investigación y la cual se calcula por una serie de operaciones matriciales junto con transformaciones logarítimicas y exponenciales. Esta función muestra algunos aspectos de la relación entre la distribución y la naturaleza de las subpoblaciones.

- La estimación de los parámetros de un modelo y la construcción de pruebas estadísticas que involucran esa función por medio del cálculo de mínimos cuadrados ponderados.

La matriz de pesos utiliza en esta metodología es la inversa de la matriz de covarianzas de F. Con este procedimiento se da mayor peso a los elementos en F que tengan varianzas pequeñas.

B. Superficies de Respuesta.

El objetivo de esta metodología es encontrar las condiciones óptimas de operación para una variable respuesta en función de los factores que intervienen en ella.

Esta metodología se lleva a cabo en dos pasos:

- Construcción de la función respuesta.

- Optimización de la misma.

La superficie de respuesta es un modelo de regresión estándar, el cual es lineal en los parámetros pero no necesariamente lineal en las variables factor. Los ajustes más conocidos de superficies de respuesta son a polinomios de primer y segundo grado en los diferentes factores.

La superficie de respuesta es estudiada utilizando la siguiente ecuación de regresión:

$$F = X\beta + E$$

donde F es el vector de observaciones de la variable respuesta, X es la matriz diseño la cual corresponde a la representación codificada del espacio de factores, β es el vector de los coeficientes a estimar y E se distribuye normal multivariada con parámetros θ y $\sigma^2 I$.

β es estimado mediante los métodos tradicionales de regresión múltiple con la siguiente expresión:

$$b = (X'X)^{-1}X'F \ .$$

La siguiente etapa de esta metodología consiste en hallar los niveles de los factores que hacen la respuesta máxima. Para este fin se calculan e interpretan las primeras y segundas derivadas parciales, cuando este procedimiento no arroja resultados se pueden utilizar herramientas como la forma canónica.

Otra forma de representación de una función respuesta es:

$$y = \beta_o + \sum\beta_h x_h + \sum\beta_{hh}x_h^2 + \sum\sum\beta_{hk}x_h x_k + e$$

equivalente a:

$$y = \beta_o + X'\beta^* + X'\beta 1X + E$$

donde $\beta^{*'} = (\beta_1,\beta_2,\beta_3,\ldots,\beta_k)$

$$\beta 1 = \begin{bmatrix} \beta_{12} & \frac{1}{2}\beta_{12} & \cdots & \frac{1}{2}\beta_{1k} \\ \cdot & \beta_{12} & \cdots & \frac{1}{2}\beta_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \beta_{kk} \end{bmatrix}$$

La forma general de los puntos estacionarios bajo la suposición de que el modelo es correcto es la siguiente:

$$x = -\tfrac{1}{2}(\beta 1^{-1}\beta^*)$$

La respuesta en el punto estacionario es dada por:

$$y = \beta_0 - \tfrac{1}{2}(\beta^{*'} \beta 1^{-1} \beta^{*}) .$$

2. Superficies de respuesta para análisis de datos categóricos.

Se parte de una tabla de contingencia que reune las condiciones necesaria para ser estudiada con la metodología G.S.K.

Se desarrollan funciones de respuesta de primer y segundo orden para las proporciones lineales y logit de una respuesta procedente de una tabla de contingencia. Estas se determinan partiendo del supuesto que los elementos del vector respuesta tienen la misma varianza y que el tamaño de la muestra es suficientemente grande como para garantizar que e se distribuya normal.

Para ilustrar la construcción de las funciones de respuesta se desarrolla en detalle la función de respuesta logit para un modelo de segundo orden:

Dos factores con tres niveles es el esquema mínimo requerido para poder construir un modelo de segundo orden. Los niveles de los factores se codifican con +1, 0, -1.

La respuesta a ajustar es:

$$y = \begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \\ y_{9i} \end{bmatrix} = \begin{bmatrix} LN(P_{11}) - LN(P_{12}) \\ LN(P_{21}) - LN(P_{22}) \\ LN(P_{31}) - LN(P_{32}) \\ LN(P_{41}) - LN(P_{42}) \\ LN(P_{51}) - LN(P_{52}) \\ LN(P_{61}) - LN(P_{62}) \\ LN(P_{71}) - LN(P_{72}) \\ LN(P_{81}) - LN(P_{82}) \\ LN(P_{91}) - LN(P_{92}) \end{bmatrix}$$

Los coeficientes a estimar son:

$$\beta = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{12} \\ \beta_{22} \end{bmatrix}$$

donde, β_o: Intercepto

β_1: Razón de cambio del factor A

β_2: Razón de cambio del factor B

β_{11}: Razón de cambio del factor A cuadrático

β_{22}: Razón de cambio del factor B cuadrático

β_{12}: Razón de cambio de la interacción de A*B.

La matriz diseño a utilizar es:

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

La superficie de respuesta es estudiado con el uso de la siguiente ecuación de regresión:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e$$

donde e se distribuye $N(0,\sigma^2)$ y σ^2 es la varianza experimental. La solución es:

$$b = W*y$$

donde,

$$W = \begin{bmatrix} -0.111 & 0.222 & -0.111 & 0.222 & 0.555 & 0.222 & -0.111 & 0.222 & -0.111 \\ -0.167 & -0.167 & -0.167 & 0.000 & 0.000 & 0.000 & 0.167 & 0.167 & 0.167 \\ -0.167 & 0.000 & 0.167 & -0.167 & 0.000 & 0.167 & -0.167 & 0.000 & 0.167 \\ 0.167 & 0.167 & 0.167 & -0.333 & -0.333 & -0.333 & 0.167 & 0.167 & 0.167 \\ 0.167 & -0.333 & 0.167 & 0.167 & -0.333 & 0.167 & 0.167 & -0.333 & 0.167 \\ 0.150 & 0.000 & -0.250 & 0.000 & 0.000 & 0.000 & -0.250 & 0.000 & 0.250 \end{bmatrix}$$

Dado que los b's son correlacionados la varianza de la superficie de respuesta en cualquier punto (x_1, x_2) es dado por:

$$V(y) = V(b_o + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2$$

$$= 0.555\sigma^2 + 0.166\sigma^2x_1^2 + 0.5\sigma^2x_2^2 + 0.5\sigma^2x_1^4 + 0.5\sigma^2x_2^4$$
$$+ 0.75\sigma^2x_1^2x_2^2 - 0.666\sigma^2x_1^2 - 0.666\sigma^2x_2^2 + \sigma^2x_1x_2$$

Cuando se considera la varianza de los elementos del vector respuesta diferente y la muestra suficientemente grande de los coeficientes de la función de respuesta son estimados por medio de la siguiente expresión:

$$b = (X' \; V_b^{-1} \; X)^{-1} \; X' V_b^{-1} \; y$$

tal como se ilustra en la metodología de Grizzle, Starmer y Koch.

Las funciones de respuesta de primer orden corresponden a los modelos de efectos principales y las de segundo orden corresponden a los modelos donde existen interacciones y efectos cuadráticos de los factores.

El interpretar los modelos lineales procedentes de la aplicación de la metodología G.S.K. como funciones de respuesta permite considerar además de las transformaciones lineales y logit, las exponenciales y logarítmicas.

En este punto se recomienda considerar bajo que condiciones el desarrollo de las funciones de respuesta con mí-

nimos cuadrados ponderados y ordinarios concide ya que la tesis alcanzó a vislumbrar que para la medida de asociación gama estas son muy próximas cuando la muestra es suficiente mente grande.

Procediendo con la metodología de Superficies de Respuesta se requiere encontrar la configuración del espacio de factores que contribuye a optimizar la función respuesta.

Al respecto se concluye:

- El máximo matemático se consigue al evaluar la función respuesta en el espacio de factores codificado y ordenar las respuestas así obtenidas después se determina la correspondiente configuración del espacio de factores.

Este procedimiento se puede realizar porque el dominio de definición de la función respuesta es finito y discreto.

- Recomendamos utilizar los métodos tradicionales de diferenciación para optimizar la función respuesta cuando esta involucra variables continuas agrupadas u ordinales.

- Sugerimos adaptar las primeras diferencias finitas para optimizar funciones de respuesta procedentes de tablas que incluyan variables nominales ya que de esta manera se considera la naturaleza discreta del espacio de factores.

En síntesis, en la creación de las superficies de respuestas se utilizó una metodología cualitativa cuando determinamos por medio de la metodología G.S.K. los factores significativos en la respuesta estudiada y una metodología cuan

titativa cuando se optimiza la función respuesta.

3. Evaluación de las pérdidas no técnicas en la compañía de electricidad y gas Cundinamarca S.A.

A continuación se describen los resultados de aplicar la metodología G.S.K. y de Superficies de Respuesta para datos categóricos, al análisis de las pérdidas no técnicas; sancionadas de conformidad con el decreto-ley 1303 durante el período comprendido entre junio de 1989 y julio de 1990 en ejecución del programa de control y detección de pérdidas en Celgac S.A.

El objetivo de aplicar estas metodologías a las pérdidas no técnicas fue detectar factores asociados a los usuarios que los hacen más propensos al fraude y para encontrar la combinación de factores que originan una respuesta máxima de pérdidas. La estrategia estadística propuesta también se puede utilizar para encontrar el punto de equilibrio entre beneficios y costos en la detección de pérdidas no técnicas.

Se estudiaron las pérdidas no técnicas, es decir, las originadas en robos de energía y deficiencia en la medición y facturación.

Específicamente se construyeron modelos y se optimizaron las funciones de respuesta para los siguientes tipos de pérdidas:

1. Cambio fraudulento de uso del servicio eléctrico contratado (causal 16A).

2. Aumento ilítico de la carga instalada (Causal 16D).

3. Alteración del equipo de medida (Causal 16F).

4. Alteración del equipo de medida (Causal 16E).

El cambio fraudulento de uso del servicio eléctrico, se consideró con dos perfiles de respuesta cambio de residencial a comercial y cambio de residencial a industrial.

Se relacionó el cambio fraudulento de uso con el distrito y la carga contratada.

El problema inicialmente planteado fue analizar la probabilidad del cambio fraudulento de residencial a comercial para ello se investigó el efecto del distrito, de la carga contratada y el efecto medio.

Se concluyó:

1. El factor principal que explica el cambio fraudulento analizado es la carga contratada menor a 5 kilovatios.

Asociada a los usuarios que tienen una carga contratada menor a 5 kilovatios existe una probabilidad del 24% de que cambien fraudulentamente de residencial a comercial.

2. El cambio ilícito de uso, residencial a comercial no está asociado a los distritos.

3. 32% es la probabilidad estimada de cambio fraudulento residencial-comercial sin considerar el distrito y la carga contratada.

4. Se asocia a la carga contratada menor o igual a 5 kilova-

tios la presencia de respuesta máxima para esta causal.

El aumento no autorizado de la carga instalada fue investigado utilizando la variable carga fraudulenta, diferencia entre la carga contratada y la encontrada en el momento de la inspección. Esta variable se relacionó con el distrito y los usos comercial, industrial y hotelero.

Se analizó la probabilidad de aumentar sin autorización en más de 5 kilovatios la carga instalada. Para ello se investigó el efecto del distrito, de la modalidad de uso y el efecto medio.

Se concluyó:

1. Los factores principales que explican la frecuencia de cargas fraudulentas iguales o superiores a 5 kilovatios son el distrito y la modalidad de uso.

Se asoció a los distritos de Girardot, La Mesa, Pacho y Choachí una probabilidad del 15%, 11%, 9% y 3% respectivamente de que un usuario aumente ilícitamente en 5 kilovatios o más la carga contratada.

Se asocia al sector industrial y hotelero una probabilidad del 7% y 20% de incurrir en el fraude de aumentar la carga instalada en 5 o más kilovatios.

2. Las proporciones de carga fraudulenta mayor o igual a 5 kilovatios se maximiza en Girardot Sector Industrial y se minimiza en Choachí Sector Comercial.

Posteriormente se estudió simultáneamente la causal

16A y 16D ya que se quería probar estadísticamente su relación. Se concluyó que por distrito la rata de crecimiento de las causales analizadas son iguales.

Por último se relacionó la causal 16E y 16F en un modelo logit en 2 dimensiones (es decir alterar los sellos del equipo de medida y el equipo de medida) y se concluyó que la rata de crecimiento por distrito para estas dos causales son estadísticamente iguales.

Se sugiere:

1. Utilizar las técnicas de muestreo para seleccionar los usuarios a revisar en ejecución del programa de control y detección de pérdidas.

2. En el momento de la inspección o posteriormente identificar en el usuario fraudulento:

- Estrato.
- El tipo de actividad industrial.
- La marca del equipo de medida.

Esto con el objeto de poder relacionar los tipos de fraude con estos factores.

3. Adelantar un estudio de beneficios y costos por tipo de fraude, lo cual nos indicará la combinación óptima de recursos a invertir en ejecución de programas tendientes a detectar pérdidas no técnicas y el máximo beneficio susceptible de obtener para Cegac S.A.

4. Orientar la detección de las pérdidas no técnicas a los

usuarios que tienen mayor propensión, es decir, según las conclusiones de los modelos desarrollados.

*

BIBLIOGRAFIA

Agresti Alan. *Analysis of Ordinal Categorical Data.* John Wiley & Sons. 1984.

Belz Maurice. *Statistical Methods in the Process Industries.* Lowe & Brydone Ltda. 1973.

Bishop, Fienberg, Holland. *Discrete Multivariate Analysis: Theory and Practice.* Mit Press. Sixth Printing. 1980.

Cochran William, Cox Gertrude. *Diseños Experimentales.* Editorial Trillas, México. 1971.

Forthofer Ronald Lehnen Robert. *Public Programa Analysis.* Lifetime Learning Publications. California. 1981.

Forthofer Ronald. "An Analysis for Compounded for Categorical Data", **Biometrics**, March 1973.

Forthofer Ronald N. Koch Gary. "An Analysis for Compounded Functions of Categorical Data". **Biometrics**, March 1973.

Grizzle James. "Multivariate Logit Analysis". **Biometrics**. 1971.

Grizzle J. Starmer F. y Koch G. "Analysis of Categorical Data by Linear Models". **Biometrics**. September 1969.

Johnson William y Koch Gary. "A note on the Weighted Least Squares Analysis of the Ries-Smith Contigency

Table Data". **Technometrics**. March 1972.

Landis Richard y otros. *A Computer for the Generalized Chi-square Analysis of Categorical Data Using Weighted Least Squares*. North Holland Publishing Company.

Mood Alexander, Graybill Franklin, Boes Duane. *Introduction to the theory of Statistics*. McGraw-Hill, Third Edition.

SAS Institude Inc. *User's Guide Statistics*. Version 5 Edition. Cary. NC: SAS Institute Inc. 1985.

* *