Departamento de Estadística e I.O.

Máster en Estadística Aplicada

UNIVERSIDAD
DE GRANADA

MODELOS DE RESPUESTA DISCRETA APLICACIONES BIOSANITARIAS

Tema 2
Modelos logit con variables explicativas categóricas

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias. Tema 3: Modelos logit con variables explicativas categóricas

Índice general

Capítulo 3

Modelos logit con variables explicativas categóricas

En este capítulo consideraremos el caso de variables explicativas categóricas que son aquellas cuyos valores no son susceptibles de medida sino que son un conjunto de cualidades exhaustivas y excluyentes. La metodología para construir el modelo de regresión logística con este tipo de variables consiste en asociarles variables cuantitativas que reciben el nombre de variables del diseño y construir el modelo tomando como variables explicativas dichas variables del diseño. De este modo la estimación de parámetros y la inferencia sobre ellos se lleva a cabo mediante las técnicas introducidas en el capítulo anterior para variables explicativas cuantitativas.

3.1. Variables del diseño

Asociadas a una variable cualitativa A con categorías denotadas por $A_i (i = 1, \ldots, I)$, se definen un total de $(I - 1)$ variables del diseño o variables ficticias. La razón para definir una variable ficticia menos que el número de categorías es que la matriz de diseño resultante del modelo de regresión logística múltiple correspondiente sea invertible (tenga columnas linealmente independientes).

A continuación se presentan distintas formas de codificación de las variables del diseño.

3.1.1. Método parcial

Recibe también el nombre de codificación respecto a un grupo de referencia (*reference cell coding*). Consiste en elegir una categoría de referencia, de

modo que todas las variables del diseño asignan el valor 0 a dicha categoría de referencia. Asociada a cada una de las restantes categorías se define una variable del diseño binaria que toma el valor 1 para su categoría asociada y el valor 0 para todas las demás. En estos apuntes tomaremos como categoría de referencia la de menor código (primera). Con el software libre R que se usará en las prácticas se puede cambiar la categoría de referencia y elegir la que se considere más adecuada.

De acuerdo con lo expuesto anteriormente, la m-ésima variable del diseño va asociada con la categoría A_m, y se define en la siguiente forma:

$$X_{im}^A = X_m^A(A = A_i) = \begin{cases} 1 & i = m \\ 0 & i \neq m \end{cases} \qquad \forall m = 2, \ldots, I; i = 1, \ldots, I.$$

3.1.2. Método marginal

Se llama también codificación mediante desviación respecto a la media (*deviation from mean coding*). De nuevo se toma una categoría de referencia a la que todas las variables del diseño asignan el mismo valor que en este caso es -1. Asociada a cada una de las restantes categorías se define una variable del diseño que toma el valor 1 para su categoría asociada y el valor 0 para todas las demás excepto para la de referencia (primera en BMDP) a la que todas las variables ficticias asignan el valor -1.

Por lo tanto, la m-ésima variable del diseño está asociada con la categoría A_m y se define como

$$X_{im}^A = X_m^A(A = A_i) = \begin{cases} 1 & i = m \\ -1 & i = 1 \\ 0 & i \neq m, 1 \end{cases} \qquad \forall m = 2, \ldots, I; i = 1, \ldots, I.$$

Observaciones

1. El tipo de codificación elegido dependerá de los objetivos perseguidos en el análisis estadístico.

2. El método usual es el de codificación con respecto a un grupo de referencia debido a que los parámetros tienen una interpretación sencilla en términos de cocientes de ventajas y suele ser de interés práctico estimar el riesgo de desarrollar una enfermedad para un grupo de exposición relativo a un grupo control no expuesto que se toma de referencia.

3. El método de desviación respecto a la media es el usado en Análisis de la Varianza para analizar la desviación de la media de la variable respuesta en cada nivel de la variable cualitativa y su media global. En

el caso de regresión logística corresponderá al análisis de la desviación entre el logit en cada nivel de la variable cualitativa y la media de los logit.

3.1.3. Codificación de variables ordinales

En el caso de una variable explicativa ordinal se puede proceder también de varias formas para construir el modelo de regresión logística

1. Se puede considerar como una variable nominal, definiendo a partir de ella las correspondientes variables del diseño con el método parcial o el marginal.

2. Codificar la variable ordinal asignando puntuaciones monótonas z_i a cada una de sus categorías A_i $(i = 1, \ldots, I)$. La codificación puede ser arbitraria, asignando generalmente puntuaciones igualmente espaciadas, o basada en polinomios ortogonales.

   Si se opta por la codificación de variables ordinales, éstas se introducirán en el modelo de regresión logística como variables cuantitativas cuyos valores son los códigos asignados a cada categoría.

3.2. Una variable explicativa categórica

Denotemos por L_i al logit de respuesta $Y = 1$ en la categoría $A = A_i$ de la variable explicativa cualitativa A. Es decir,

$$L_i = \ln\left[\frac{p_i}{1 - p_i}\right],$$

donde p_i es la probabilidad de respuesta $Y = 1$ para un individuo clasificado en la categoría A_i de A

$$p_i = P[Y = 1/A = A_i] \quad \forall i = 1, \ldots, I.$$

El modelo de regresión logística para explicar la variable aleatoria binaria Y en términos de la variable explicativa categórica A se construye como un modelo de regresión logística múltiple para Y en términos de las $I - 1$ variables del diseño asociadas a A. La expresión general de dicho modelo es de la forma

$$L_i = \beta_0 + \sum_{m=2}^{I} \tau_m^A X_{im}^A.$$

A continuación se obtiene la formulación particular de este modelo y la interpretación de sus parámetros para los dos métodos de codificación de las variables del diseño considerados.

3.2.1. Método parcial

Sin más que utilizar la definición de las variables ficticias en este caso se tiene

$$\begin{aligned} L_1 &= \beta_0 \\ L_i &= \beta_0 + \tau_i^A \quad i = 2, \ldots, I, \end{aligned}$$

que puede resumirse como

$$L_i = \beta_0 + \tau_i^A \quad i = 1, \ldots, I, \tag{3.1}$$

con $\tau_1^A = 0$.

Interpretación de parámetros

Observemos que la exponencial del parámetro β_0 del modelo 3.1 es la ventaja de respuesta $Y = 1$ para la categoría de referencia A_1.

Denotemos por θ_{i1} al cociente de ventajas de respuesta $Y = 1$ de la categoría A_i respecto de la categoría de referencia A_1, definido por

$$\theta_{i1} = \frac{\dfrac{p_i}{1-p_i}}{\dfrac{p_1}{1-p_1}}.$$

Sustituyendo la expresión anterior por el modelo 3.1 se obtiene

$$\theta_{i1} = \exp\left(\tau_i^A\right) \quad \forall i = 2, \ldots, I,$$

que proporciona una interpretación sencilla de los parámetros del modelo.

3.2.2. Método marginal

La definición de las variables del diseño para este caso lleva a la siguiente expresión:

$$\begin{aligned} L_1 &= \beta_0 - \sum_{i=2}^{I} \tau_i^A \\ L_i &= \beta_0 + \tau_i^A \quad i = 2, \ldots, I. \end{aligned} \tag{3.2}$$

Observemos que este último modelo puede expresarse en la forma general

$$L_i = \beta_0 + \tau_i^A \quad i = 1, \ldots, I,$$

sin más que definir $\tau_1^A = -\sum_{i=2}^{I} \tau_i^A$.

Interpretación de parámetros

Con este tipo de codificación de las variables ficticias el parámetro β_0 es claramente la media de las transformaciones logit de cada categoría de la variable explicativa dada por

$$\beta_0 = \bar{L} = \frac{1}{I} \sum_{i=1}^{I} L_i.$$

Por lo tanto, cada uno de los restantes parámetros es la desviación del logit de la categoría que lleva asociada con respecto a la media de todos los logit

$$\tau_i^A = L_i - \bar{L} \quad \forall i = 1, \ldots, I.$$

Como consecuencia se obtiene que la exponencial de cada parámetro es el cociente entre la ventaja de respuesta $Y = 1$ para su categoría asociada y la media geométrica de todas las ventajas de respuesta $Y = 1$

$$\begin{aligned} \exp(\tau_i^A) &= \frac{\exp(L_i)}{\left(\prod_{i=1}^{I} \exp L_i\right)^{\frac{1}{I}}} \\ &= \frac{\frac{p_i}{1-p_i}}{\left(\prod_{i=1}^{I} \frac{p_i}{1-p_i}\right)^{\frac{1}{I}}}. \end{aligned}$$

Con el método marginal los cocientes de ventajas de cada categoría con respecto a la de referencia se pueden calcular a partir de los parámetros del modelo en la siguiente forma:

$$\theta_{i1} = \exp(\tau_i^A + \sum_{i=2}^{I} \tau_i^A) \quad \forall i = 2, \ldots, I.$$

3.2.3. Estimación directa por máxima verosimilitud

Para la estimación del modelo de regresión logística con una única variable explicativa categórica A, dispondremos como es usual de una muestra de tamaño N, de modo que a cada categoría A_i corresponderán n_i observaciones de entre las que y_i serán unos y el resto, $n_i - y_i$ serán ceros.

Observemos que este modelo de regresión logística es saturado porque el número de transformaciones logit muestrales (tantas como categorías distintas A_i) coincide con el número de parámetros independientes del modelo que son en total I. Esto llevará como veremos a continuación a que las frecuencias esperadas de respuesta uno en cada categoría coinciden con el número observado de unos en dicha categoría $\hat{m}_i = n_i\hat{p}_i = y_i$.

Se puede demostrar que efectivamente las ecuaciones de verosimilitud tienen solución exacta. En el caso del método parcial los estimadores MV son de la forma

$$\begin{aligned} \hat{\beta}_0 &= \ln\left(\frac{y_1}{n_1 - y_1}\right). \\ \hat{\tau}_i^A &= \ln\left(\frac{\frac{y_i}{n_i - y_i}}{\frac{y_1}{n_1 - y_1}}\right) \quad \forall i = 2, \ldots, I. \end{aligned}$$

Por otro lado, los estimadores MV de los parámetros del método marginal son

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{I}\sum_{i=1}^{I} \ln\left(\frac{y_i}{n_i - y_i}\right). \\ \hat{\tau}_i^A &= \ln\left(\frac{y_i}{n_i - y_i}\right) - \hat{\beta}_0 \quad \forall i = 2, \ldots, I. \end{aligned}$$

3.3. Dos variables explicativas categóricas

Consideremos ahora dos variables explicativas categóricas $A : A_1, \ldots, A_I$ con variables del diseño asociadas $X_2^A, \ldots, X_I^A$, y $B : B_1, \ldots, B_J$ con variables del diseño $X_2^B, \ldots, X_J^B$.

Si la probabilidad de respuesta $Y = 1$ en cada combinación de niveles de las dos variables cualitativas es

$$P[Y = 1/A = A_i, B = B_j] = p_{ij},$$

el modelo de regresión logística se construye en la siguiente forma tomando

como variables explicativas las $(I + J - 2)$ variables del diseño

$$L_{ij} = \ln[\frac{p_{ij}}{1 - p_{ij}}] = \beta_0 + \sum_{m=2}^{I} \tau_m^A X_{im}^A + \sum_{m=2}^{J} \tau_m^B X_{jm}^B \quad i = 1, \ldots, I; j = 1, \ldots, J.$$

3.3.1. Método parcial

Utilizando la definición binaria de las variables del diseño se obtiene la siguiente expresión para el modelo de regresión logística:

$$\begin{aligned} L_{11} &= \beta_0 \\ L_{i1} &= \beta_0 + \tau_i^A \quad i = 2, \ldots, I \\ L_{1j} &= \beta_0 + \tau_j^B \quad j = 2, \ldots, J \\ L_{ij} &= \beta_0 + \tau_i^A + \tau_j^B \quad i = 2, \ldots, I; j = 2, \ldots, J, \end{aligned}$$

que puede expresarse de forma global como

$$L_{ij} = \beta_0 + \tau_i^A + \tau_j^B \quad i = 1, \ldots, I; j = 1, \ldots, J,$$

con $\tau_1^A = \tau_1^B = 0$.

Interpretación de parámetros

La exponencial del parámetro β_0 es claramente la ventaja de respuesta $Y = 1$ para un individuo que pertenece a las categorías de referencia de ambas variables A_1 y B_1.

Para las exponenciales del resto de los parámetros se tiene lo siguiente:

$$\exp[\tau_i^A] = \frac{\frac{p_{ij}}{1 - p_{ij}}}{\frac{p_{1j}}{1 - p_{1j}}} = \theta_{i1/B}^A \quad \forall i = 2, \ldots, I, j = 1, \ldots, J,$$

donde $\theta_{i1/B}^A$ representa el cociente de ventajas de respuesta $Y = 1$ de la categoría A_i respecto de la categoría de referencia A_1 controlando fija la categoría B_j de B.

Análogamente, se tiene

$$\exp[\tau_j^B] = \frac{\frac{p_{ij}}{1 - p_{ij}}}{\frac{p_{i1}}{1 - p_{i1}}} = \theta_{j1/A}^B \quad \forall j = 2, \ldots, J; i = 1, \ldots, I,$$

donde $\theta^B_{j1/A}$ representa el cociente de ventajas de respuesta $Y = 1$ de la categoría B_j respecto de la categoría de referencia B_1 controlando fija la categoría A_i de A.

Observemos que los cocientes de ventajas anteriores no dependen de la variable que se controla de modo que la asociación entre cada variable respuesta y la variable explicativa es la misma en todos los niveles de la otra variable explicativa. Esta propiedad se conoce con el nombre de ausencia de interacción entre las dos variables explicativas.

3.3.2. Método marginal

A partir de la definición de las variables del diseño se obtiene de nuevo expresión para el modelo de regresión logística, que en este caso es

$$\begin{aligned} L_{11} &= \beta_0 - \sum_{i=2}^{I} \tau_i^A - \sum_{j=2}^{J} \tau_j^B \\ L_{i1} &= \beta_0 + \tau_i^A - \sum_{j=2}^{J} \tau_j^B \quad i = 2, \ldots, I \\ L_{1j} &= \beta_0 + \tau_j^B - \sum_{i=2}^{I} \tau_i^A \quad j = 2, \ldots, J \\ L_{ij} &= \beta_0 + \tau_i^A + \tau_j^B \quad i = 2, \ldots, I; j = 2, \ldots, J, \end{aligned}$$

y puede representarse globalmente por la expresión

$$L_{ij} = \beta_0 + \tau_i^A + \tau_j^B \quad i = 1, \ldots I; j = 1, \ldots, J,$$

definiendo $\tau_1^A = -\sum_{i=2}^{I} \tau_i^A$ y $\tau_1^B = -\sum_{j=2}^{J} \tau_j^B$.

Interpretación de parámetros

El parámetro β_0 es de nuevo la media de todas las transformaciones logit

$$\beta_0 = \bar{L} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} L_{ij}.$$

Los parámetros asociados a la variable A se interpretan en la forma

$$\tau_i^A = L_{i\bullet} - \bar{L} \quad i = 2, \ldots, I,$$

donde $L_{i\bullet}$ es la media marginal de los logit en la categoría A_i dada por

$$L_{i\bullet} = \frac{1}{J}\sum_{j=1}^{J} L_{ij}.$$

Análogamente, los parámetros asociados a la variable B se interpretan en la forma

$$\tau_j^B = L_{\bullet j} - \bar{L} \quad j = 2, \ldots, J,$$

donde $L_{\bullet j}$ es la media marginal de los logit en la categoría B_j dada por

$$L_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} L_{ij}.$$

Resumiendo, hemos obtenido que cada parámetro es la desviación de la media marginal de los logit en su categoría asociada respecto de la media global de todas las transformaciones logit. En este caso, la exponencial de cada parámetro no es un cociente de ventajas sino un cociente de medias geométricas de ventajas, como se pone de manifiesto a continuación:

$$\exp(\tau_i^A) = \frac{\left(\prod_{j=1}^{J} \frac{p_{ij}}{1-p_{ij}}\right)^{\frac{1}{J}}}{\left(\prod_{i=1}^{I}\prod_{j=1}^{J} \frac{p_{ij}}{1-p_{ij}}\right)^{\frac{1}{IJ}}} \quad i = 2, \ldots, I.$$

$$\exp(\tau_j^B) = \frac{\left(\prod_{i=1}^{I} \frac{p_{ij}}{1-p_{ij}}\right)^{\frac{1}{I}}}{\left(\prod_{i=1}^{I}\prod_{j=1}^{J} \frac{p_{ij}}{1-p_{ij}}\right)^{\frac{1}{IJ}}} \quad j = 2, \ldots, J.$$

Finalmente, el método marginal permite también calcular los cocientes de ventajas de respuesta $Y = 1$ para cada categoría de cada una de las variables respecto de la de referencia, controlando la otra variable. Estos cocientes de ventajas no dependen de la variable que se controla (ausencia de interacción) y se obtienen como la siguiente transformación de los parámetros asociados a la variable correspondiente

$$\theta^A_{i1/B} = \frac{\frac{p_{ij}}{1-p_{ij}}}{\frac{p_{1j}}{1-p_{1j}}} = \exp\left(\tau_i^A + \sum_{i=2}^{I} \tau_i^A\right) \quad \forall j = 1, \ldots, J, \quad i = 2, \ldots, I.$$

$$\theta^B_{j1/A} = \frac{\dfrac{p_{ij}}{1-p_{ij}}}{\dfrac{p_{i1}}{1-p_{i1}}} = \exp\left(\tau^B_j + \sum_{j=2}^{J} \tau^B_j\right) \quad \forall i = 1, \ldots, I, \quad j = 2, \ldots, J.$$

3.4. Variables explicativas cuantitativas y cualitativas

En la práctica suele ocurrir que las variables explicativas son tanto cuantitativas como cualitativas. En este caso el modelo de regresión logística se construye con las variables explicativas cuantitativas y las variables del diseño asociadas a las variables explicativas cualitativas.

Comencemos por el caso más simple de una única variable explicativa cuantitativa X y una variable explicativa categórica A con categorías $A_1, \ldots, A_I$.

Si denotamos $p_i(x) = P[Y = 1/X = x, A = A_i]$, el modelo de regresión logística es de la forma

$$\ln\left[\frac{p_i(x)}{1-p_i(x)}\right] = \beta_0 + \sum_{m=2}^{I} \tau^A_m X^A_{im} + \beta_1 x.$$

De la definición de las variables del diseño se obtiene la expresión general de este modelo

$$L_i(x) = \ln\left[\frac{p_i(x)}{1-p_i(x)}\right] = \beta_0 + \tau^A_i + \beta_1 x \quad i = 1, \ldots, I, \tag{3.3}$$

donde $\tau^A_1 = 0$ si se utiliza el método parcial de codificación de las variables del diseño y $\tau^A_1 = -\sum_{i=2}^{I} \tau^A_i$ si se utiliza el método marginal.

Para la estimación por máxima verosimilitud de los parámetros del modelo (3.3) dispondremos de una muestra aleatoria simple de tamaño N con Q observaciones diferentes $x_q \quad (q = 1, \ldots, Q)$ de la variable X, de modo que se tendrán un total de $I \times Q$ transformaciones logit muestrales dados por

$$L_{iq} = \ln\left[\frac{p_{iq}}{1-p_{iq}}\right] = \beta_0 + \sum_{m=2}^{I} \tau^A_m X^A_{im} + \beta_1 x_q,$$

siendo $p_{iq} = P[Y = 1/A = A_i, X = x_q]$.

Veamos a continuación que los parámetros del método parcial se interpretan en términos de cocientes de ventajas.

La exponencial del parámetro β_1 es el cociente de ventajas de respuesta $Y = 1$ cuando se incrementa en una unidad la variable cuantitativa X y se controla la categoría de la variable cualitativa A. Es decir,

$$\theta(\triangle X = 1/A = A_i) = \frac{\dfrac{p_i(x+1)}{1-p_i(x+1)}}{\dfrac{p_i(x)}{1-p_i(x)}} = \exp(\beta_1) = \theta(\triangle X = 1/A).$$

Por otro lado, la exponencial de cada uno de los parámetros asociados a cada categoría de la variable cualitativa es el cociente de ventajas de respuesta $Y = 1$ para su categoría asociada con respecto a la primera categoría controlando la variable cuantitativa X. Es decir,

$$\theta^A_{i1/X=x} = \frac{\dfrac{p_i(x)}{1-p_i(x)}}{\dfrac{p_1(x)}{1-p_1(x)}} = \exp(\tau_i^A) = \theta^A_{i1/X} \quad i = 2, \ldots, I.$$

Los cocientes de ventajas anteriores se pueden calcular también a partir de los parámetros del método marginal.

La exponencial del parámetro de la variable cuantitativa tiene la misma interpretación que con el método parcial

$$\theta(\triangle X = 1/A) = \exp(\beta_1).$$

Si denotamos por $\bar{L}(x)$ a la media de las transformaciones logit para cada x dada por

$$\bar{L}(x) = \frac{1}{I}\sum_{i=1}^{I} L_i(x) = \frac{1}{I}\sum_{i=1}^{I}(\beta_0 + \tau_i^A + \beta_1 x) = \beta_0 + \beta_1 x,$$

despejando en el modelo 3.3, se tiene que los parámetros asociados a la variable cualitativa son las desviaciones del logit en su categoría asociada respecto de la media de todos los logit, controlando la variable X

$$\tau_i^A = L_i(x) - \bar{L}(x) \quad i = 2, \ldots, I.$$

Además, los cocientes de ventajas de respuesta $Y = 1$ de cada categoría respecto de la primera, para cada valor fijo de X, se obtienen a partir de los parámetros como sigue:

$$\theta^A_{i1/X} = \frac{\exp\left(\beta_0 + \tau_i^A + \beta_1 x\right)}{\exp\left(\beta_0 + \tau_1^A + \beta_1 x\right)} = \exp\left(\tau_i^A + \sum_{i=2}^{I} \tau_i^A\right).$$

Observemos que estos cocientes de ventajas son iguales para todos los valores de X debido a que estamos considerando un modelo logístico sin interacción entre las variables A y X.

Consideremos ahora el caso general de R variables explicativas cuantitativas, representadas por $X = (X_1, \ldots, X_r, \ldots, X_R)'$, y varias variables cualitativas $A, B, C, \ldots,$, para las que se dispone de una muestra de tamaño N en la que se han observado un total de Q combinaciones distintas de valores de las variables explicativas denotadas por $x_q = (x_{q1}, \ldots, x_{qr}, \ldots, x_{qR})' \quad \forall q = 1, \ldots, Q$.

Construyendo el modelo con las variables del diseño asociadas a tres variables cualitativas A, B y C de categorías $A_1, \ldots, A_I$; $B_1, \ldots, B_J$ y $C_1, \ldots, C_K$, respectivamente, se tiene la siguiente expresión general para cada una de las $I \times J \times K \times Q$ transformaciones logit muestrales:

$$L_{ijkq} = \beta_0 + \sum_{m=2}^{I} \tau_m^A X_{im}^A + \sum_{m=2}^{J} \tau_m^B X_{jm}^B + \sum_{m=2}^{K} \tau_m^C X_{km}^C + \sum_{r=1}^{R} \beta_r x_{qr},$$

definiendo

$$L_{ijkq} = \ln \left[\frac{p_{ijkq}}{1 - p_{ijkq}} \right]$$

con $p_{ijkq} = P[Y = 1/A = A_i, B = B_j, C = C_k, X = x_q]$.

A partir de la definición de las variables del diseño se obtiene la forma general del modelo muestral

$$L_{ijkq} = \beta_0 + \tau_i^A + \tau_j^B + \tau_k^C + \sum_{r=1}^{R} \beta_r x_{qr},$$

verificando que $\tau_1^A = \tau_1^B = \tau_1^C = 0$ para el método parcial o bien $\sum_{i=1}^{I} \tau_i^A = \sum_{j=1}^{J} \tau_j^B = \sum_{k=1}^{K} \tau_k^C = 0$ para el método marginal.

3.5. Modelos con interacción

Recordemos que la interacción entre dos variables cuantitativas se incluye en el modelo de regresión logística múltiple como producto de ambas variables.

La interacción entre dos variables cualitativas (interacción de orden uno) se incluye en el modelo en forma de combinación lineal de todos los posibles

productos cruzados entre sus variables del diseño. Por ejemplo, la interacción entre dos variables cualitativas A y B se representa en la parte lineal del modelo logit mediante

$$\sum_{l=2}^{I}\sum_{m=2}^{J}\tau_{lm}^{AB}X_l^A X_m^B.$$

Finalmente, la interacción entre una variable cuantitativa X y una cualitativa A se incluye en el modelo mediante una combinación lineal de los productos entre la variable cuantitativa y cada una de las variables del diseño asociadas a la variable cualitativa, en la siguiente forma:

$$\sum_{m=2}^{I}\tau_m^{AX}X_m^A X.$$