UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS MATEMATICAS

FACULTAD DE CIENCIAS

Algunos modelos para el análisis de respuesta ordinal múltiple en la medición de la densidad mineral ósea.

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRA EN CIENCIAS (MATEMATICAS)

P R E S E N T A

ANA LEONOR TOUSSAINT MARTINEZ DE CASTRO

DIRECTORA DE TESIS:

DRA. SILVIA RUIZ VELASCO ACOSTA

MEXICO, D. F. OCTUBRE, 2005

Gracias a mis padres, compañeros, profesores y amigos por apoyar y creer siempre en mi.

Gracias a Juan Manuel, María Fernanda y Mariana que siempre me acompañaron en este deseo de continuar.

Gracias a los profesores del departamento de estadística del iimas por darme la oportunidad de estudiar, compartir conocimientos y ser motivación para no claudicar.

Gracias a Salvador por su paciencia a mi entendimiento. Este trabajo no existiría sin su apoyo brindado en todo momento.

Gracias a Silvia por su apoyo y comprensión incondicional

Gracias a Alma por compartir su investigación y conocimiento.

ÍNDICE

INTRODUCCIÓN

En los últimos cincuenta años los estadísticos y los bioestadísticos han incrementado su interés en el desarrollo de métodos de regresión para el análisis de variables categóricas. Áreas como salud pública y biomédicas se han beneficiado de los métodos desarrollados para este tipo de variables. Sin embargo, existen otras áreas como ciencias sociales, ecología, educación, mercadotecnia y control de calidad, que también han incorporado estos métodos en el análisis de la información.

En muchas de estas áreas la variable explicativa de interés no siempre es una variable cuantitativa, es más, la mayoría de las veces la variable es cualitativa o categórica.

Las variables categóricas se clasifican:

Nominal → Las categorías no tienen ningún valor cuantitativo, ni de orden; por ejemplo: nombres de países y ciudades.

Ordinal → Las categorías están ordenadas y no existe una distancia predeterminada entre las categorías; por ejemplo: bueno, regular y malo.

Las variables categóricas también se clasifican según los valores que puedan dar como resultado, es decir

Respuesta Binaria ó Dicotómica → Las variables categóricas tienen dos posibles valores, por ejemplo: si y no, vivo y muerto.

Respuesta Múltiple ó Politómica → Las variables categóricas tienen más de dos posibles valores, por ejemplo: partido político (PRI, PAN, PRD); o de opinión (parcialmente de acuerdo, de acuerdo, totalmente de acuerdo).

Es importante hacer notar que cuando se tienen variables de interés categóricas, muchas veces los niveles de dicha variable ya están determinadas (femenino y masculino). Si los niveles son subjetivos y dependen del experimentador, no deben crearse al azar o arbitrariamente, sino que deben crearse bajo un análisis que concuerde con el interés propio de la investigación. Esto es debido a que cualquier modificación que sufran (eliminación de niveles o unión de niveles adyacentes) afectará a la información disponible, los parámetros estimados, la significancia del modelo y la bondad de ajuste.

Los modelos de regresión para el análisis de variables categóricas fueron establecidos principalmente por Cox (1970), Fienberg (1980) y Plackett (1981). En donde se plantean los modelos de regresión bajo una liga para variables de respuesta categórica, denotada por Y, dado un conjunto de variables explicativas $X's$ denotadas por $x_1, x_2,..., x_p$, que pueden ser cuantitativas o cualitativas. Si la variable de respuesta Y es politómica, sus k categorías posibles de respuesta se denotan por $y_1, y_2,...,y_k$.

En particular, cuando a los investigadores se les presentan análisis de variables de respuesta categóricas, comúnmente ajustan los datos dependiendo del interés del análisis, es decir, realizan alguno de los siguientes análisis

- Generan dos categorías convirtiendo la variable en dicotómica y ajustan el modelo de regresión logística.
- Agrupan las variables en intervalos de clase y ajustan el modelo de regresión lineal.
- Ignoran el orden de las variables y ajustan el modelo de regresión logística de respuesta politómica, el cual se presentará en el capítulo 1.
- Utilizan modelos específicos para variables de respuesta categóricas ordinales.

La finalidad de la presente tesis es mostrar tres de los modelos que se utilizan para variables con respuesta ordinal, es decir, en donde la variable de respuesta tiene k posibles valores ordenados. Estos modelos son:

El Modelo de Momios Proporcionales
El Modelo de Cocientes Continuos
El Modelo Estereotipo

El Modelo de Momios Proporcionales, McCullagh (1980) y el Modelo de Cocientes Continuos, Fienberg (1980), asumen una relación entre las variables explicativas y la variable de respuesta, la cual toma en cuenta el orden de los niveles de las categorías. Si la variable de respuesta no presenta orden en sus categorías, es conveniente ajustar el modelo logístico de respuesta politómica ya que no asume ninguna relación de orden en la variable de respuesta. Sin embargo una desventaja del modelo es que el número de parámetros a ajustar es mayor con respecto al número de parámetros de los modelos de momios proporcionales y de cocientes continuos, por lo cual la interpretación del modelo puede complicarse. Además puede existir inestabilidad en la estimación si el tamaño de muestra es pequeño.

Un modelo alternativo para modelos con respuesta ordinal es el modelo estereotipo, Anderson (1984). Este modelo es menos conocido que los dos modelos anteriores debido, principalmente, a la dificultad que se presenta al ajustar el modelo. Sin embargo, actualmente es posible ajustarlo gracias al desarrollo reciente de la paquetería estadística.

Aunque el objetivo es mostrar los modelos para respuesta ordinal, se iniciará el desarrollo con los modelos para respuesta binaria y politómica, ya que se considera que ilustran adecuadamente algunos de los procesos estadísticos de los modelos para respuesta ordinal.

CAPÍTULO 1

MODELO LOGÍSTICO

El modelo logístico se ha convertido en una parte importante del análisis en investigaciones cuando se busca describir la relación entre una variable de respuesta categórica y sus variables explicativas. Al igual que en muchos otros modelos estadísticos, el investigador al utilizar el modelo logístico busca encontrar la mejor descripción y explicación a las observaciones de una investigación.

1.1 Modelo logístico para respuesta binaria

Los modelos para respuesta categórica ordinal o nominal, tienen en común que si las categorías son sólo dos, reproducen al modelo logístico para respuesta binaria, es por esta razón que se presenta dicho modelo.

El modelo logístico, también conocido como regresión logística se basa en una variable de respuesta binaria Y con dos posibles valores representados por *0* y *1*. Debido a lo anterior, la variable de respuesta Y se asume con distribución Bernoulli.

Entonces el valor esperado de Y está dado por:

$$E(Y) = P(Y = 1) = \pi(x)$$

con

$$V(Y) = \pi(x)\,(1 - \pi(x))$$

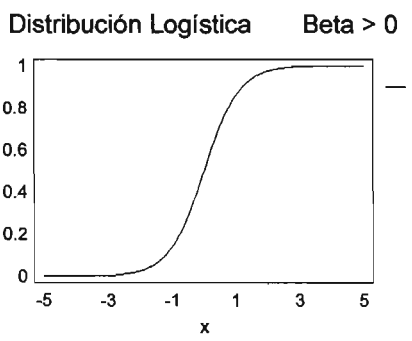
Un primer intento para modelar esta probabilidad, es a través del modelo de regresión lineal usual, es decir,

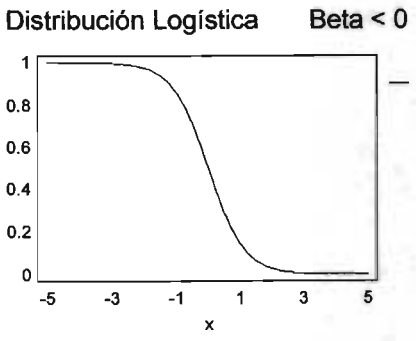
$$E(Y|x) = \pi(x) = \alpha + \beta x$$

Esta manera de modelar presenta algunos problemas estructurales importantes. Las probabilidades deben tener valores entre *0* y *1*, mientras que las variables explicativas pueden tomar valores en todos los $\Re$. Por lo que el modelo puede predecir valores imposibles de $\pi(x) < 0$ y $\pi(x) > 1$, cuando los valores de x_i son suficientemente grandes o suficientemente pequeños.

También se presentan dificultades para ajustar el modelo por mínimos cuadrados debido a que las condiciones de optimización no se satisfacen. La varianza no es constante $\pi(x)(1-\pi(x))$ y si $\pi(x)$ se acerca al valor *0* o al valor *1*, la distribución condicional de *Y* se concentra en un punto, y la varianza se acerca a cero. Debido a lo anterior los estimadores no son de mínima varianza dentro de los estimadores lineales insesgados.

Como consecuencia a lo anterior, es conveniente utilizar un modelo que permita una relación curvilínea entre $\pi(X)$ y la variable explicativa *X*, por lo que se recomienda aplicar una transformación que tenga la forma natural de S como la que tiene la función logística.





Aplicando dicha transformación obtenemos:

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}$$

La expresión anterior es la *Distribución Logística*. Esta función es monótona, con $\pi(x) \downarrow 0$ ó $\pi(x) \uparrow 1$ cuando $x \uparrow \infty$, dependiendo del valor de $\beta < 0$ o' $\beta > 0$ respectivamente. Si $\beta \downarrow 0$ la curva se aplana en una línea recta horizontal. Si $\beta = 0$ la respuesta es independiente de la variable X y la función tiene el valor de $\pi(x) = ½$ cuando $x = -\alpha/\beta$.

Una medida de asociación común cuando la variable de respuesta es categórica, es el momio, que compara la probabilidad de observar la respuesta dadas las variables explicativas contra la probabilidad de no observarla. Así para este modelo, se puede construir el momio

$$\frac{\pi(x)}{1-\pi(x)} = exp(\alpha + \beta x)$$

Aplicando la función logarítmica obtenemos

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

El logaritmo de momios nos permite tener un modelo lineal. Esta transformación es conocida como logit, así

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

Esta transformación nos proporciona una interpretación básica de β, misma que depende de la escala de medición de X, de modo que

- Si X es dicotómica, con niveles "0" y "1" entonces el cociente de los momios, se incrementa en e^{β} cuando x cambia de "0" a "1".

- Si X es politómica, existen dos casos:
  - ✓ Si X es politómica ordinal, entonces el cociente de momios, se incrementa en e^{β} por cada cambio de categoría con respecto a la categoría de referencia de la variable X.
  - ✓ Si X es politómica no ordinal, entonces el cociente de los momios, se incrementa en e^{β_j} por cada cambio de categoría con respecto a una categoría de referencia de la variable X.

  Por lo general, esta categoría de referencia es el nivel más bajo, pero se recomienda investigar cuál es la categoría de referencia del paquete estadístico que se vaya a utilizar..

- Si X es continua, entonces el cociente de los momios, se incrementa en e^{β} por cada unidad que se incremente la variable X. O bien, como $e^{r\beta}$ si la variable se incrementa r unidades.

Generalizando la expresión del modelo logístico para p variables explicativas se obtiene:

$$logit(\pi(\underline{x})) = log\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Entonces la probabilidad de la respuesta es:

$$\pi(\underline{x}) = \frac{exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_j\right)}{\left[1 + exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_j\right)\right]}$$

El modelo logístico fue formulado para variables de respuesta dicotómica, sin embargo, se puede extender para variables de respuesta politómica. El parámetro β_j se interpreta de la misma manera que en el caso de una variable explicativa cuando x_j cambia manteniendo las demás variables explicativas constantes.

1.2 Modelo logístico para respuesta politómica no ordinal

El modelo logístico para respuesta politómica, a diferencia del modelo de respuesta binaria, tiene una variable de respuesta con más de dos opciones. Existen varias propuestas, pero solo se presenta la más común.

Sea Y_i la variable de respuesta, con $i = 1,2,\ldots, k$, entonces el modelo se expresa como

$$P[Y = y_i|\underline{x}] = \pi_i(\underline{x}) = \frac{exp(\alpha_i + \sum_{j=1}^{p} \beta_{ij}^t x_{hj})}{\sum_{i=1}^{k} exp(\alpha_i + \sum_{j=1}^{p} \beta_{ij}^t x_{hj})}$$

donde $\pi_i(\underline{x})$ denota la probabilidad de la respuesta Y_i y además

$\alpha_i \rightarrow$ Son las constantes para cada respuesta i, con $i = 1,\ldots,k$.

$\beta_{ij} \rightarrow$ Es el parámetro asociado a la respuesta i en la covariable j, con $j = 1,\ldots p$.

$x_{hj} \rightarrow$ Es la covariable j del individuo h, con $h = 1,\ldots,n$.

1.2.1 Estimación del modelo

El método general de estimación es a través de máxima verosimilitud, la cual estima los parámetros desconocidos que maximizan la probabilidad de los valores observados. La función de Máxima Verosimilitud se basa en que la variable de respuesta Y tiene una distribución multinomial de n observaciones independientes, entonces la verosimilitud se expresa

$$L(\alpha_1,\ldots,\alpha_k,\underline{\beta}_1,\ldots,\underline{\beta}_p,X,Y)=\prod_{h=1}^{n}\pi_1(\underline{x})^{y_{h1}}\pi_2(\underline{x})^{y_{h2}}\cdots\pi_k(\underline{x})^{y_{hk}}$$

donde Y_{hi} representa las respuestas de los individuos

$$Y_{hi}=\begin{cases}1 & \rightarrow si \quad el \quad individuo \quad h \quad presenta \quad la \quad respuesta \quad i\\ 0 & \rightarrow en \quad otro \quad caso\end{cases}$$

En lugar de maximizar la verosimilitud se acostumbra maximizar la log-verosimilitud. Por simplicidad, manejaremos la expresión $\ell(\underline{\theta})$.

$$\ell(\underline{\theta})=\sum_{h=1}^{n}\left[y_{h1}ln(\pi_1(\underline{x}))+y_{h2}ln(\pi_2(\underline{x}))+\cdots+y_{hk}ln(\pi_k(\underline{x}))\right]$$

Para encontrar los valores de los parámetros desconocidos α's y β's que maximizan $\ell(\underline{\theta})$, se debe derivar la función con respecto a cada uno de los parámetros. Estas derivadas son:

$$\frac{\partial}{\partial\beta_{ij}}\ell(\underline{\theta})=\sum_{h=1}^{n}x_{hj}\left(y_{hi}-\pi_{hi}(\underline{x})\right)$$

para $j = 1,2,...,p$.

Los estimadores máximo verosímiles de las β´s, se obtienen resolviendo las ecuaciones de las derivadas parciales igualadas a cero. Las ecuaciones obtenidas no tienen una solución cerrada con respecto a las β´s, por lo que se debe recurrir a métodos iterativos para su solución, como el método de Newton – Raphson.

Las soluciones son los estimadores máximo verosímiles y se denotan por $\hat{\alpha}$´s y $\hat{\beta}$´s, que se utilizan para encontrar los valores estimados de $\hat{\pi}_i(\underline{x})$.

La matriz de información de Fisher, denotada por $I(\underline{\theta})$, es la matriz que nos proporciona las varianzas y covarianzas de los parámetros estimados.

$$I(\underline{\theta}) = -E\left[\frac{\partial^2 \ell(\underline{\theta})}{\partial \beta_{ij} \partial \beta_{il}}\right]$$

Con $j, l = 1,2,...,p$ y $i = 1,2,...,k$.

El cálculo de la esperanza de los elementos en la matriz, puede ser muy **complejo**, por lo que una buena aproximación a esta matriz se encuentra calculando las segundas derivadas parciales de $\ell(\underline{\theta})$ y evaluarlas en los estimadores máximo verosímiles.

$$I(\hat{\underline{\theta}}) = \left[\frac{\partial^2 \ell(\underline{\theta})}{\partial \beta_{ij} \partial \beta_{il}}\right]_{(\beta_{ij}, \beta_{il})}$$

La matriz obtenida se llama matriz de información observada de Fisher y se representa por $I(\hat{\underline{\theta}})$ y es una matriz de $(p+1)(k-1)\times(p+1)(k-1)$. Los estimadores de la matriz de covarianzas de los estimadores máximo verosímiles se calculan con la inversa de la matriz observada.

$$Cov(\hat{\underline{\theta}}) = I(\hat{\underline{\theta}})^{-1}$$

Los elementos de la diagonal de esta matriz simétrica son los estimadores de las varianzas y los elementos fuera de la diagonal son los estimadores de las covarianzas.

1.2.2 Pruebas de Hipótesis

Las pruebas de hipótesis son una parte fundamental del análisis del modelo. La primera prueba de hipótesis es determinar la significancia del modelo, es decir

$$\text{Ho: } \underline{\beta_1} = \underline{\beta_2} = \ldots = \underline{\beta_k} = 0 \quad \text{vs} \quad H_1: \underline{\beta_j} \neq 0 \quad \text{para alguna } i = 1,2,\ldots k$$

donde $\underline{\beta_j} = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})$

Para obtener la significancia del modelo, se calcula la prueba estadística llamada *Devianza*, basada en una transformación del cociente de verosimilitudes

$$Devianza = -2ln\left[\frac{verosimilitud \quad del \quad modelo \quad propuesto}{verosimilitud \quad del \quad modelo \quad saturado}\right]$$

La *Devianza* es una estadística de prueba que se aproxima asintoticamente a la distribución χ^2 con $(n-p)$ grados de libertad, donde n es el número de observaciones y p es el número de parámetros asociados a las variables explicativas.

Esta aproximación es válida si n es suficientemente grande. El nivel de significancia (valor p) asociado a la distribución χ^2 se utiliza para determinar si el modelo es o no significativo.

Si se concluye que el modelo es significativo, entonces la siguiente prueba de hipótesis es determinar si cada una de las variables explicativas por separado contribuyen o no significativamente al ajuste del modelo.

Entonces la prueba de hipótesis es:

$$Ho: \beta_{ij} = 0 \quad vs \quad H_1: \beta_{ij} \neq 0$$

Con $i = 1,2,...,k$ y $j = 1,2,...,p$. Esta prueba se utiliza para determinar si una variable explicativa es una variable que contribuye o no contribuye significativamente al ajuste del modelo y una manera de determinarlo es calcular la estadística de Wald que es el cociente del estimador máximo verosímil y el error estándar de dicho estimador

$$Z = \frac{\hat{\beta}_{ij}}{\sqrt{\hat{Var}(\hat{\beta}_{ij})}} \sim^{a} N(0,1)$$

El nivel de significancia asociado a la distribución Normal estandarizada nos determinará si la variable explicativa es o no significativa.

Si se desea hacer una comparación entre dos modelos anidados, para determinar si un grupo o conjunto de dos o más variables explicativas, contribuyen o no contribuyen significativamente al ajuste del modelo, entonces sean los modelos M1 con ***r*** y M2 con ***s*** parámetros, con r < s. Así la prueba de hipótesis se plantea

$$Ho: \beta_{ir+1} = \ldots = \beta_{is} = 0 \quad vs \quad H_1: \beta_{it} \neq 0$$

para toda i y para algún $t \, \varepsilon \, (r+1,\ldots,s)$.

Para comprobar la hipótesis se realiza la resta de devianzas de los modelos, donde dicha resta se distribuye asintóticamente como una Ji-cuadrada con *(s – r)* grados de libertad.

Existen varias generalizaciones del modelo logístico cuando la variable de respuesta es ordinal, y la elección de cuál modelo aplicar y ajustar depende de las finalidades de la investigación en cuestión, tipo de las variables explicativas, etc. En el siguiente capítulo se mostrarán tres de las generalizaciones de los modelos logísticos para respuesta politómica ordinal.

CAPÍTULO 2

ALGUNOS MODELOS DE RESPUESTA POLITÓMICA ORDINAL

Una manera natural de construir los modelos de regresión logística para respuesta ordinal, es suponer que existe una variable latente continua, que da lugar a las categorías de la respuesta ordinal a través de algún proceso de discretización. La discretización de una variable continua es un proceso muy utilizado por los investigadores; variables que se asumen como categóricas ordinales, como nivel socio-económico, estado de salud, estado nutricional, entre otras, son variables que provienen de discretizar variables latentes continuas.

2.1 Discretización de variables continuas

Entonces, dado un valor Y_i^*, no observado, de la variable latente continua, considérese el modelo lineal

$$Y_i^* = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Sea η_i el componente sistemático del modelo anterior, i.e.

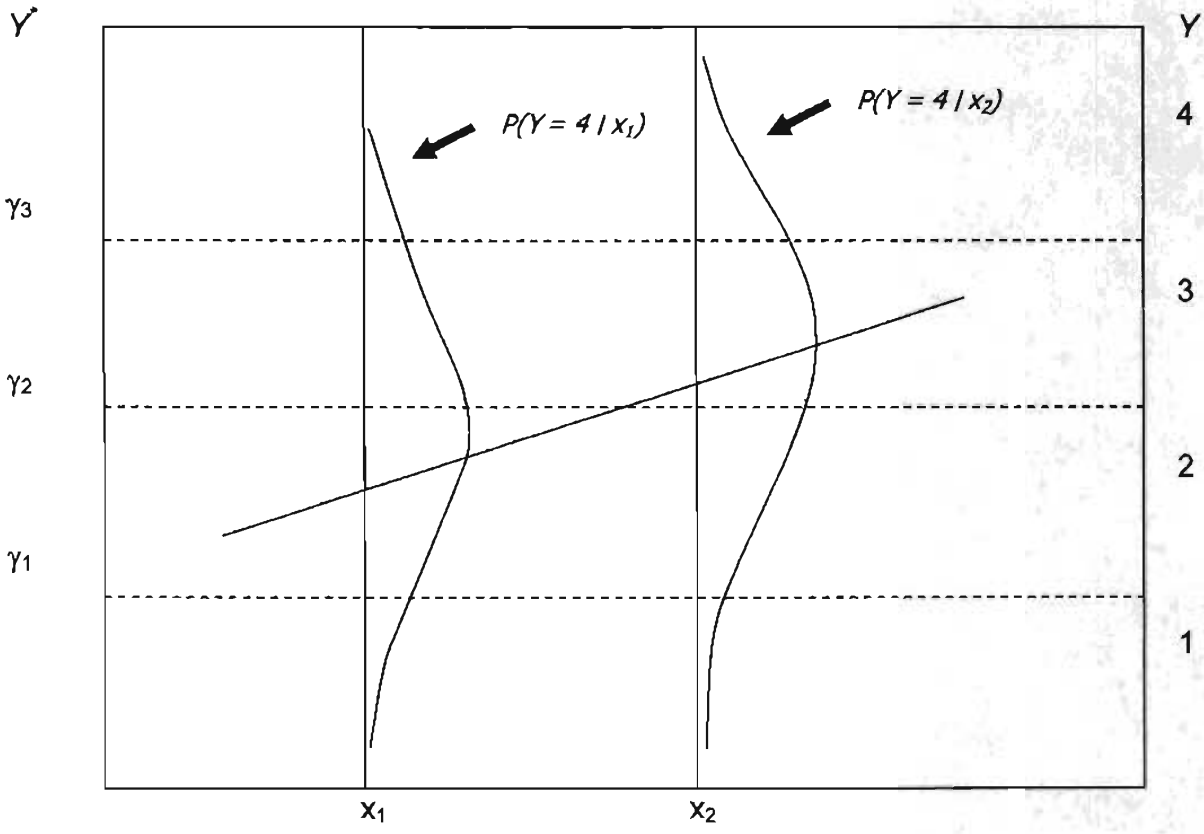
$$\eta_i(\underline{x}) = \underline{\beta}' \underline{x}_i$$

La variable de respuesta que realmente observamos es Y_i definida como

$$\begin{aligned} Y_i &= 1 \quad si \quad \gamma_0 < Y_i^* \leq \gamma_1 \\ Y_i &= 2 \quad si \quad \gamma_1 < Y_i^* \leq \gamma_2 \\ &\vdots \\ Y_i &= k \quad si \quad \gamma_{k-1} < Y_i^* \leq \gamma_k \end{aligned}$$

con $\gamma_0 = -\infty$ y $\gamma_k = \infty$

Así, las probabilidades ordinales se observarían como se muestra en el ejemplo



Entonces

$$
\begin{aligned}
P[Y_i = k] &= P[\gamma_{k-1} < Y_i^* \le \gamma_k] \\
&= P[\gamma_{k-1} < \eta_i + \varepsilon_i \le \gamma_k] \\
&= P[\eta_i + \varepsilon_i < \gamma_k] - P[\eta_i + \varepsilon_i < \gamma_{k-1}] \\
&= P[\varepsilon_i < \gamma_k - \eta_i] - P[\varepsilon_i < \gamma_{k-1} - \eta_i] \\
&= F(\gamma_k - \eta_i) - F(\gamma_{k-1} - \eta_i)
\end{aligned}
$$

donde F es la distribución acumulativa de ε.

La log-verosimilitud para los n individuos es

$$\ell(\beta,\gamma,Y) = \sum_{i=1}^{n}\sum_{k=1}^{K} log\{P[Y_i = k]\}I(Y_i = k)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} log\{F(\gamma_k - \eta_i) - F(\gamma_{k-1} - \eta_i)\}I(Y_i = k)$$

Derivando la log-verosimilitud con respecto a β y γ, obtenemos

$$\frac{\partial}{\partial\beta_j}\ell(\beta,\gamma,Y) = \sum_{i=1}^{n} x_{ij}\sum_{k=1}^{K}\left\{\frac{-f(\gamma_k - \eta_i) + f(\gamma_{k-1} - \eta_i)}{F(\gamma_k - \eta_i) - F(\gamma_{k-1} - \eta_i)}\right\}I(Y_i = k)$$

y

$$\frac{\partial}{\partial\gamma_j}\ell(\beta,\gamma,Y) = \sum_{i=1}^{n}\left\{\frac{f(\gamma_j - \eta_i)}{F(\gamma_j - \eta_i) - F(\gamma_{j-1} - \eta_i)}I(Y_i = j) - \frac{f(\gamma_j - \eta_i)}{F(\gamma_{j+1} - \eta_i) - F(\gamma_j - \eta_i)}I(Y_i = j+1)\right\}$$

Los estimadores máximos verosímiles se obtienen resolviendo las primeras derivadas igualadas a cero.

El desarrollo anterior es general para cualquier distribución que se le asigne al término de error ε. El modelo de momios proporcionales que se presenta en este capítulo, supone que la distribución del error es una *logística estándar* con una liga logit para las probabilidades acumuladas de respuesta, mientras que el modelo de cocientes continuos puede deducirse suponiendo la misma distribución para el error y una liga *log-log* para la probabilidad de respuesta.

2.2 Modelo de Momios Proporcionales

El Modelo de Momios proporcionales para regresión logística ordinal fue propuesto por McCullagh (1980) y permite una extensión del uso de modelos logísticos de respuesta politómica en los casos donde la variable respuesta es ordinal, es decir, sus categorías están ordenadas.

El modelo se representa por medio de regresiones logísticas para variables binarias dependientes, con parámetros de regresión comunes que reflejan el supuesto de proporcionalidad. Este supuesto se basa en que debe existir cierta proporcionalidad entre los momios creados por el modelo. Debido a este supuesto es por lo que el modelo recibe el nombre de *momios proporcionales*.

2.2.1 Modelo

Suponiendo que tenemos n observaciones independientes de la variable de respuesta ordinal Y, con k categorías ordenadas y que $\underline{X}$ es un vector de variables explicativas de dimensión p. El Modelo de Momios Proporcionales se basa en la distribución acumulativa de probabilidad

$$\gamma_l = P[Y \leq y_l | \underline{x}]$$

y sea

$$\pi_l(\underline{x}) = P[Y = y_l | \underline{x}]$$

entonces

$$logit(\gamma_l) = log\left[\frac{P(Y \leq y_l|\underline{x})}{1 - P(Y \leq y_l|\underline{x})}\right]$$

$$= log\left[\frac{\gamma_l}{1-\gamma_l}\right]$$

$$= log\left[\frac{\pi_1(\underline{x}) + \pi_2(\underline{x}) + \cdots + \pi_i(\underline{x})}{1 - [\pi_1(\underline{x}) + \pi_2(\underline{x}) + \cdots + \pi_i(\underline{x})]}\right]$$

$$= \alpha_i + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$= \alpha_i + \underline{\beta}^t \underline{x}$$

con $i = 1, 2, \ldots, k-1$.

El vector $\underline{\beta}$ es de dimensión p y $\alpha_1 < \alpha_2 < \ldots < \alpha_{k-1}$ representan parámetros desconocidos llamados puntos de corte.

El modelo de momios proporcionales asume que el efecto de las variables en los momios de respuesta bajo la categoría i es el mismo efecto para toda i, si se satisface

$$logit(\gamma_i \mid \underline{x}^{(1)}) - logit(\gamma_i \mid \underline{x}^{(2)}) = log\left[\frac{P(Y \le y_i \mid \underline{x}^{(1)}) / (1 - P(Y \le y_i) \mid \underline{x}^{(1)})}{P(Y \le y_i \mid \underline{x}^{(2)}) / (1 - P(Y \le y_i) \mid \underline{x}^{(2)})}\right]$$

$$= log\left[\frac{\dfrac{\pi_1(\underline{x}^{(1)}) + \pi_2(\underline{x}^{(1)}) + \cdots + \pi_i(\underline{x}^{(1)})}{1 - [\pi_1(\underline{x}^{(1)}) + \pi_2(\underline{x}^{(1)}) + \cdots + \pi_i(\underline{x}^{(1)})]}}{\dfrac{\pi_1(\underline{x}^{(2)}) + \pi_2(\underline{x}^{(2)}) + \cdots + \pi_i(\underline{x}^{(2)})}{1 - [\pi_1(\underline{x}^{(2)}) + \pi_2(\underline{x}^{(2)}) + \cdots + \pi_i(\underline{x}^{(2)})]}}\right]$$

$$= (\alpha_i + \beta_1 x_1^{(1)} + \cdots + \beta_p x_p^{(1)}) - (\alpha_i + \beta_1 x_1^{(2)} + \cdots + \beta_p x_p^{(2)})$$

$$= \left(\alpha_i + \beta^t \underline{x}^{(1)}\right) - \left(\alpha_i + \beta^t \underline{x}^{(2)}\right)$$

$$= \beta^t (\underline{x}^{(1)} - \underline{x}^{(2)})$$

donde $\underline{x}^{(1)}$ y $\underline{x}^{(2)}$ son dos subconjuntos diferentes de las variables explicativas. La diferencia anterior es el cociente de momios de probabilidades acumulativas, que también recibe el nombre del *cociente de momios acumulativo*.

El logaritmo del cociente de momios acumulativo es proporcional a la distancia entre los valores de las variables explicativas, teniendo la misma constante de proporcionalidad. La constante de proporcionalidad es independiente del valor de los puntos de corte.

La interpretación del cociente de momios acumulativo es respecto a cuando $\underline{x} = \underline{x}^{(1)}$ el momio de obtener una respuesta es menor o igual a i es $exp\,[\underline{\beta}^t(\underline{x}^{(1)} - \underline{x}^{(2)})]$ veces más grande que cuando $\underline{x} = \underline{x}^{(2)}$.

Es importante notar que cuando $\beta_i > 0$, para toda i, el modelo de la ecuación

$$logit(\gamma_i) = log\left[\frac{P(Y \le y_i|\underline{x})}{1-P(Y \le y_i|\underline{x})}\right] = log\left[\frac{\gamma_i}{1-\gamma_i}\right] = \alpha_i + \beta^t \underline{x}$$

crece mientras x crece, así que la probabilidad acumulativa crece. Esto significa que relativamente más probabilidad cae por debajo del final de la escala de Y, es decir, Y tiende a ser más pequeño a mayores valores de x_i. Para que cada $\beta_i > 0$ tenga mejor interpretación con respecto a cuando se incremente Y a mayores valores de x_i, remplazamos β por $-\beta$.

Para esta parametrización tenemos

$$\text{logit}(\gamma_i) = log\left[\frac{P(Y \le y_i|\underline{x})}{1-P(Y \le y_i|\underline{x})}\right] = log\left[\frac{\gamma_i}{1-\gamma_i}\right] = \alpha_i - \beta^t \underline{x} \qquad (2)$$

con $i = 1,...,k-1$.

Es importante mencionar que el modelo de momios proporcionales no presenta cambios significativos en las interpretaciones de los parámetros cuando se presenta la unión de categorías adyacentes, como también no presenta cambios significativos a la reversibilidad del orden de la variable de respuesta. Lo anterior es debido a que el modelo goza de invarianza palindrómica.

2.2.2 Estimación y Pruebas de Hipótesis

La estimación del modelo se basa en la función verosimilitud, y debido a que la variable de respuesta tiene una distribución Multinomial, se expresa

$$\ell(\alpha_1, \alpha_2, \ldots, \alpha_k, \beta_1, \ldots, \beta_p, \underline{X}, \underline{Y}) = \prod_{h=1}^{n} \pi_1(\underline{x})^{y_{h1}} \pi_2(\underline{x})^{y_{h2}} \cdots \pi_k(\underline{x})^{y_{hk}}$$

Los estimadores máximo verosímiles se encuentran al derivar la función de verosimilitud con respecto a cada uno de los parámetros desconocidos β's, e igualando cada una de las ecuaciones a cero. Estas ecuaciones no presentan soluciones cerradas, por lo que sus soluciones se encuentran a través de métodos iterativos. Los estimadores de la matriz de varianzas y covarianzas de los parámetros estimados, se obtienen al evaluar la inversa de la matriz de información de Fisher, en los estimadores máximo verosímiles, $\hat{\beta}$'s , dando lugar a la matriz de información observada de Fisher, $I(\hat{\underline{\beta}})^{-1}$.

Las pruebas de hipótesis proporcionan información sobre la significancia del modelo, además de la significancia de cada una de las variables explicativas. Las pruebas utilizadas para este modelo son semejantes a las utilizadas en el modelo logístico de respuesta politómica, tales como el cociente de verosimilitudes, la estadística de Wald, *etc.*

2.2.3 Supuesto de momios proporcionales

Es importante comprobar el supuesto de momios proporcionales. Sea una Y_i la variable de respuesta, con $i = 1,2,\ldots,k$; y *las variables explicativas* X_j, *con* $j = 1,\ldots,p$.

Entonces bajo la hipótesis de momios proporcionales, la hipótesis nula es

$$\text{Ho: } \beta_{1j} = \beta_{2j} = \ldots = \beta_{kj} = \beta_j \qquad \forall j = 1,2,\ldots,p.$$

Es decir, existe un solo valor común β_j, para j =1,2,...,p .

Sean $\beta_1,\beta_2,\ldots,\beta_k$ estos valores comunes. Sean $\hat{\alpha}_1,..,\hat{\alpha}_k$ y $\hat{\beta}_1,..,\hat{\beta}_p$ los estimadores máximo verosímiles bajo esta hipótesis.

Denotemos por $\theta_0 = (\theta_1',\ldots,\theta_k')$ con $\theta_i = (\hat{\alpha}_i,\hat{\beta}_1,\ldots,\hat{\beta}_p)'$, con $i = 1,\ldots,k$. Así la prueba de puntajes o "score" es

$$U^* = U'(\theta_0)I(\theta_0)^{-1}U(\theta_0)$$

donde

$U(\hat{\theta}_0) \rightarrow$ es el vector de primeras derivadas de la log-verosimilitud en θ_0.

$I(\hat{\theta}_0)^{-1} \rightarrow$ es la inversa de la matriz de información observada de Fisher.

$U^* \rightarrow$ se distribuye asintóticamente como una $\chi^2_{p(k-1)}$.

Un valor de U^* mayor a $\chi^2_{(p(k-1),\alpha)}$ será indicativo de que al menos una de las variables explicativas introducidas dentro del modelo, no cumple el supuesto de momios proporcionales.

2.3 Modelo de Cocientes Continuos

El modelo de Cocientes Continuos fue creado por Fienberg (1980). Este modelo es muy semejante al Modelo de Momios Proporcionales, y también es una extensión de los modelos logísticos de respuesta ordinal.

A diferencia del modelo de momios proporcionales, el modelo de cocientes continuos modela la probabilidad de obtener la respuesta *i* dada una respuesta mayor o igual que *i*, o la probabilidad de obtener una respuesta *i* dada una respuesta menor o igual que *i*.

2.3.1 Modelo

Para definir el modelo, es necesario definir la probabilidad de obtener la respuesta i, entonces

$$\pi_i(\underline{x}) = P(Y = y_i|\underline{x})$$

Y definiendo las probabilidades acumulativas

$$\gamma_i(\underline{x}) = P(Y > y_i|\underline{x}) \quad \text{ó} \quad \textit{alternativamente} \quad \gamma_i(\underline{x}) = P(Y \leq y_i|\underline{x})$$

entonces el Modelo de Cocientes Continuos se puede definir de la siguiente manera

$$logit\,(\gamma_i|\underline{x}) = log\left[\frac{P(Y = y_i|\underline{x})}{P(Y > y_i|\underline{x})}\right] = log\left[\frac{\pi_i(\underline{x})}{\pi_{i+1}(\underline{x}) + \cdots + \pi_k(\underline{x})}\right] = \alpha_i - \beta' \underline{x} \qquad i = 1,...,k\text{-}1$$

ó

$$logit(\gamma_i|\underline{x}) = log\left[\frac{P(Y = y_{i+1}|\underline{x})}{P(Y \leq y_i|\underline{x})}\right] = log\left[\frac{\pi_{i+1}(\underline{x})}{\pi_1(\underline{x}) + \cdots + \pi_i(\underline{x})}\right] = \alpha_i - \beta' \underline{x} \qquad i = 1,...,k\text{-}1$$

Se puede crear el cociente de momios de cocientes continuos (cociente de momios acumulativo), para la comparación de dos subpoblaciones $\underline{x}^{(1)}$ y $\underline{x}^{(2)}$.

$$logit(\gamma_i|\underline{x}^{(1)}) - logit(\gamma_i|\underline{x}^{(2)}) = log\left[\frac{P(Y = y_i|\underline{x}^{(1)})/P(Y > y_i|\underline{x}^{(1)})}{P(Y = y_i|\underline{x}^{(2)})/P(Y > y_i|\underline{x}^{(2)})}\right] = \beta^t(\underline{x}^{(1)} - \underline{x}^{(2)})$$

La interpretación del cociente de momios acumulativo es respecto a cuando $\underline{x} = \underline{x}^{(1)}$ el momio de obtener una respuesta $\leq i$ es $exp\,[\underline{\beta}^t(\underline{x}^{(1)} - \underline{x}^{(2)})]$ veces más grande que cuando $\underline{x} = \underline{x}^{(2)}$.

El cociente de momios de cocientes continuos para las respuestas menores o iguales a *i* se crea de igual manera.

Sea $\rho_i(\underline{x})$ la probabilidad de respuesta *i* dividido entre la probabilidad de la respuesta mayor o igual a *i*, entonces

$$\rho_i(\underline{x}) = \left[\frac{P(Y = y_i|\underline{x})}{P(Y \geq y_i|\underline{x})}\right] = \left[\frac{\pi_i(\underline{x})}{\pi_i(\underline{x}) + \cdots + \pi_k(\underline{x})}\right] \qquad i = 1,\ldots,k\text{-}1$$

Es posible utilizar logits ordinarios de probabilidades condicionales, para definir el logit de cocientes continuos, basándonos en la probabilidad $\rho_i(\underline{x})$, entonces

$$logit(\rho_i) = log\left[\frac{\rho_i(\underline{x})}{1 - \rho_i(\underline{x})}\right] = log\left[\frac{\dfrac{P(Y = y_i|\underline{x})}{P(Y \geq y_i|\underline{x})}}{1 - \dfrac{P(Y = y_i|\underline{x})}{P(Y \geq y_i|\underline{x})}}\right]$$

$$= log\left[\frac{\frac{\pi_l(\underline{x})}{\pi_i(\underline{x})+\cdots+\pi_k(\underline{x})}}{1-\frac{\pi_l(\underline{x})}{\pi_i(\underline{x})+\cdots+\pi_k(\underline{x})}}\right]$$

$$= log\left[\frac{\frac{\pi_l(\underline{x})}{\pi_l(\underline{x})+\cdots+\pi_k(\underline{x})}}{\frac{\pi_{l+1}(\underline{x})+\cdots+\pi_k(\underline{x})}{\pi_l(x)+\cdots+\pi_k(x)}}\right]$$

$$= log\left[\frac{\pi_i(\underline{x})}{\pi_{i+1}(\underline{x})+\cdots+\pi_k(\underline{x})}\right]$$

$$= log\left[\frac{P(Y=y_i|\underline{x})}{P(Y>y_i|\underline{x})}\right]$$

$$= \alpha_i - \beta' \underline{x}$$

con $i = 1,...,k-1$.

Después de algo de álgebra observamos que el *logit* (ρ_i) se puede simplificar de tal forma que su expresión es equivalente al Modelo de Cocientes Continuos, por lo que el *logit* (ρ_i) se utiliza con mayor frecuencia como la expresión del modelo de cocientes continuos, debido a la estructura del momio de ρ_i.

La transformación de probabilidades llamada cocientes continuos presenta cambios significativos en sus parámetros e interpretaciones al tratar de unir categorías adyacentes, por lo que el modelo de Fienberg es aconsejable bajo circunstancias donde las categorías de respuesta no sean grupos creados arbitrariamente, sino que dichas categorías de respuesta sean de interés. El modelo también presenta cambios significativos a la reversibilidad del orden de la variable de respuesta.

2.3.2 Estimación y Pruebas de Hipótesis

La estimación del modelo se realiza por máxima verosimilitud, en donde se asume que los datos provienen de una muestra aleatoria simple con n valores. Se utilizará el cambio de variable de respuesta $Y_l(\underline{x})$ por $n_l(\underline{x})$ debido a la familiarización de la notación en la distribución Multinomial y Binomial.

La función multinomial para un conjunto de valores de $\underline{x}$, para $n_i(\underline{x})$ las respuestas obtenidas con $i = 1,...,k$ y con $n(\underline{x}) = \sum_l n_l(\underline{x})$, es

$$f(n_i(\underline{x})) = \frac{n!}{\prod_{i=1}^{k} n_l!} \prod_{i=1}^{k} p_i^{n_l}$$

Se puede demostrar que la función multinomial para $\{n_l(\underline{x}),\ i = 1,...,k\}$ tiene factorización

$$b[n(\underline{x}),n_1(\underline{x});\rho_1(\underline{x})]\, b[n(\underline{x})-n_1(\underline{x}),n_2(\underline{x});\rho_2(\underline{x})]\cdots b[n(\underline{x})-n_1(\underline{x})-\cdots-n_{k-2}(\underline{x}),n_{k-1}(\underline{x});\rho_{k-1}(\underline{x})]$$

donde $b\ (n\ ,n_l(\underline{x});\rho_i)$ denota la probabilidad binomial de *"éxito"* en n ensayos (individuos), cuando la probabilidad de éxito es ρ_i para cada ensayo. La verosimilitud del modelo completo es el producto de funciones binomiales para diferentes valores de $\underline{x}$.

Al aplicar la función logaritmo a la verosimilitud obtenemos una suma de términos, con diferentes ρ_i's. Los parámetros del modelo se pueden encontrar maximizando cada término por separado.

Los estimadores de las varianzas y covarianzas de los parámetros estimados, se encuentran calculando la inversa de la matriz de información observada de Fisher.

El ajuste separado de modelos para diferentes logits de cocientes continuos, permite obtener el mismo ajuste que si se ajustase el modelo completo. Es decir, se puede sumar las *k-1* pruebas estadísticas de restas de cocientes de verosimilitud para obtener la significancia de ajuste del cociente de verosimilitud del modelo completo.

La significancia de cada una de las variables explicativas en el modelo, se determinan a través del cociente entre el estimador y el error estándar del estimador frente al nivel de significancia predeterminado, o si es de un conjunto de variables explicativas, a través de la resta de devianzas de los modelos pertinentes, la que se aproxima asintóticamente a una $\chi^2_{(p(k-1),\alpha)}$.

En el modelo de cocientes continuos, el supuesto de momios proporcionales debe comprobarse y se comprueba de la misma forma que en el modelo de momios proporcionales.

2.4 Modelo Estereotipo

El Modelo Estereotipo fue propuesto por Anderson (1984) para datos categóricos. Este modelo es diferente a los anteriores con respecto al efecto de las variables explicativas x_j bajo la liga log, la cual permite variación en los parámetros β_j para cada una de las posibilidades de respuesta. Los modelos de momios proporcionales y el de cocientes continuos asumen que no varía dicho efecto de las variables explicativas, mientras que en el modelo estereotipo se asume que la variable explicativa x_j puede influir a través de la función liga en diferente manera de una categoría a otra de la respuesta.

El modelo de Anderson es similar al modelo logístico de respuesta politómica basándose en que las $k-1$ regresiones logísticas se relacionan con sus p vectores de regresión β_j, con $j = 1,...,p$. Sin embargo, la diferencia radica en que el modelo de Anderson permite adjuntar las β_j, lo cual permite la variedad de regresiones para ajustar con diferentes restricciones y el modelo logístico de respuesta politómica no permite adjuntar en ninguna forma las β_j.

2.4.1 Modelo

Existen dos conceptos importantes dentro de la definición de este modelo

Dimensionalidad

Un supuesto fundamental en los modelos presentados con anterioridad (Momios Proporcionales y Cocientes Continuos) es que, dado un conjunto de variables explicativas x_{hj}, $h = 1,...,n$ y $j = 1,...,p$, la combinación de variables $\sum_{j=1}^{p} x_{hj}\beta_j$ puede utilizarse para distinguir entre todos los niveles de la respuesta, por lo que los modelos Momios Proporcionales y Cocientes Continuos, son bajo este concepto, unidimensionales. Sin embargo, si una combinación puede distinguir entre los niveles de la respuesta *1* y *2*, y otra combinación distinta se requiere para distinguir entre los niveles *2* y *3*, el modelo es bidimensional, y así sucesivamente. El modelo logístico politómico tiene dimensión *k-1*, debido a que cada respuesta es identificada por una combinación particular.

Un ejemplo de la categorización en una enfermedad como: no presenta, moderada y severa; donde puede existir un factor relacionado con la

susceptibilidad a la enfermedad, en otras palabras, el cambio de no presentar la enfermedad a moderada o severa. Sin embargo, otros factores distintos pueden determinar la severidad de la enfermedad, es decir, el cambio de moderada a severa. En esta situación el supuesto de que el efecto de cada predictor es el mismo en todos los niveles de la respuesta, es difícil de sostener.

Distinguibilidad

Dos respuestas categóricas son *indistinguibles* con respecto a una variable x_j si dicha variable, no diferencía estas categorías. El modelo estereotipo identifica cuándo dos categorías adyacentes son esencialmente iguales con respecto a las covariables $X's$ y cuándo el orden de las relaciones es relevante. Si el orden es relevante, entonces el modelo ordena a las $\beta_j's$ en lugar de ordenar los momios o la función liga. El orden está directamente relacionado a los efectos de las variables explicativas.

El modelo estereotipo permite hasta $d = min\ (k-1,\ p)$ restricciones correspondientes a $k-1$ categorías de respuesta y p variables explicativas, por lo que Anderson determinó que el modelo se clasificaría según el número de las restricciones, llamándolos unidimensional, bidimensional, tridimensional, etc...

Así, si el modelo es unidimensional, entonces el modelo lineal sobre las β_{ij} es de la forma

$$\beta_{ij} = \varphi_i \beta_j$$

con $i = 1, \ldots, k;\ \varphi_k \equiv 0$ cuando $\beta_k \equiv 0$.

Entonces, sustituyendo el valor anterior en la probabilidad de Y en el modelo logístico de respuesta politómica, obtenemos

$$\pi_i(\underline{x}) = P(Y = y_i|\underline{x}) = \frac{exp(\alpha_i - \varphi_i \sum_{j=1}^{p} \beta_j x_{hj})}{\sum_{i=1}^{k} exp(\alpha_i - \varphi_i \sum_{j=1}^{p} \beta_j x_{hj})}$$

En el modelo unidimensional, el orden de los momios es fácil de verificar. Ya que supusimos que $\varphi_1 = 1 > \varphi_2 > \cdots > \varphi_k = 0$ y si $\beta > 0$, entonces los cocientes de momios forman una secuencia decreciente

$$e^{\beta} > e^{\varphi_2 \beta} > \cdots > e^{\varphi_k \beta} = 1$$

Es decir, el efecto de las covariables sobre el primer momio es mayor que el efecto sobre el segundo momio y así sucesivamente.

Si el modelo es bidimensional, entonces β_{ij} dependerá de dos parámetros desconocidos φ_i y ϕ_i, los cuales se tendrán que estimar. Sean

$$\beta_{ij} = \varphi_i \beta_j + \phi_i \gamma_j$$

con $i = 1,...,k$, $j = 1,...,p$. Entonces, sustituyendo el valor anterior en la probabilidad de Y en el modelo logístico de respuesta politómica, obtenemos

$$\pi(\underline{x}) = P(Y = y_i|\underline{x}) = \frac{exp\left(\alpha_i - \sum_{j=1}^{p} (\varphi_i \beta_j + \phi_i \gamma_j) x_{hj}\right)}{\sum_{i=1}^{k} exp\left(\alpha_i - \sum_{j=1}^{p} (\varphi_i \beta_j + \phi_i \gamma_j) x_{hj}\right)}$$

donde $\phi_k \equiv 0$, $\varphi_k \equiv 1$.

Debido a la identificabilidad de los parámetros es necesario imponer restricciones adicionales, que deberán ser sugeridas por el contexto del análisis, por ejemplo: $\phi_1 = 1$, $\varphi_1 = 0$. Así sucesivamente se van creando los modelos de mayor dimensionalidad. Sin embargo, el modelo se va complicando según aumenta la dimensionalidad.

Cuando la dimensionalidad es mayor a uno, Anderson supone un orden de las φ_i´s o de las ϕ_i´s según sea el caso; sin embargo Luna comenta que la evidencia de los datos puede mostrarnos que dicho orden puede ser no presentarse.

Anderson sugiere un proceso de tres pasos al ajustar el modelo estereotipo a los datos:

1. Decidir la dimensionalidad del modelo *d*, determinado por

$$d \leq mínimo(k-1, n^{\circ}\ de\ covariables)$$

Debido a lo anterior se sugiere ajustar los modelos

$$\beta_{ij} = 0$$
$$\beta_{ij} = \varphi_i \beta_j$$
$$\beta_{ij} = \varphi_i \beta_j + \phi_i \gamma_j$$

y así sucesivamente hasta las *d* dimensiones. El mejor modelo se escogerá con base a las diferencias de devianzas. El modelo de menor dimensionalidad es preferible a un modelo de mayor dimensionalidad debido a que tiene menor número de parámetros a estimar.

2. Realizar las pruebas de hipótesis $\phi_i = \phi_{i+1}$ para alguna *i*, es decir, verificar si una variable explicativa *x* afecta de igual manera a las categorías *i* e *i +1* de la variable de respuesta.

3. Disminuir el número de parámetros del modelo a través del paso 1 y del paso 2. Si el modelo final es unidimensional, entonces se deberá discutir si las ϕ_i´s están ordenadas.

Es importante notar que el modelo ajustado no especifica la necesidad del ordenamiento de las ϕ_i´s, las conclusiones con respecto al orden son de acuerdo a la evidencia empírica proveniente de los datos.

2.4.2 Estimación y Pruebas de hipótesis

Como mencionamos en la introducción, el modelo estereotipo se construye con base en el modelo logístico para respuesta politómica; de hecho, varios autores llaman a este modelo, modelo politómico con restricciones. Por lo tanto, la estimación de los parámetros, así como las pruebas de hipótesis pertinentes, se realizan de manera análoga a las del modelo logístico politómico, con sus debidas adecuaciones.

2.5 Bondad de Ajuste para modelos con respuesta ordinal

En todos los modelos de regresión es muy importante la calidad del ajuste del modelo, ya que si el modelo no ajusta adecuadamente a los datos, las inferencias que se desprendan de él, son cuestionables.

En el caso del modelo de regresión logística, la falta de ajuste puede deberse a la ausencia de alguna interacción, a que la función liga no es la apropiada o a que no cumple el supuesto de momios proporcionales.

Generalmente la bondad de ajuste en los modelos logísticos se realiza a través de la prueba ji-cuadrada y la devianza. Cuando las variables explicativas son datos agrupados, estas pruebas comparan el ajuste del modelo propuesto contra el modelo saturado y se encuentran en casi todos los paquetes de cómputo estadísticos.

Sin embargo, existe documentación donde se muestra que la utilización de las pruebas ji-cuadrada y devianza no son adecuadas para juzgar la bondad de ajuste en modelos de regresión logística binaria, cuando existen datos desagrupados. Hosmer-Lemeshow (2000) propusieron una prueba de bondad de ajuste para subsanar este problema.

Pulkstenis (2004) propone modificar la ji-cuadrada y la devianza para modelos de regresión logística con respuesta ordinal en los casos donde existan variables explicativas continuas y categóricas. Esta propuesta de bondad de ajuste parece ser muy poderosa para detectar interacciones omitidas en el ajuste o para detectar cuando no se cumple el supuesto de momios proporcionales o para modelar la estructura inadecuada de una variable continua. También proporciona información inmediata de cuando el modelo no tiene un buen ajuste.

La propuesta generaliza la metodología de la regresión logística para respuesta ordinal propuesta por Pulkstenis y Robinson (2002). La esencia de la misma consiste en agregar un discretización extra a los datos, a partir de los llamados puntajes ordinales (que se definen más adelante).

Como sabemos nuestra variable de respuesta ordinal Y, tiene K categorías ordenadas y $\underline{X}$ es un vector de variables explicativas de dimensión p.

A partir de los datos, se construye una tabla de valores observados con los L posibles patrones de covariables[1] de la siguiente manera

Respuesta					
Patrón de covariables.	Y=1	Y=2	...	Y=K	
X_1	O_{11}	O_{12}	...	O_{1k}	N_1
X_2	O_{21}	O_{22}	...	O_{2K}	N_2
⋮	⋮	⋮	⋮	⋮	⋮
X_L	O_{L1}	O_{L2}	...	O_{LK}	N_L

El patrón de covariables es el conjunto de individuos que tienen los mismos valores observados en todas las covariables incluidas en el modelo.

Entonces sea $E_{lk} = \sum_{l=1}^{N_L} \hat{p}_{lk}$, donde $\hat{p}_{lk}$ es la probabilidad de cada individuo en el renglón l con la respuesta categórica *k*.

Utilizando la información de la tabla podemos definir

$$\chi^2 = \sum_{l=1}^{L}\sum_{k=1}^{K}\frac{(O_{lk}-E_{lk})^2}{E_{lk}} \qquad \text{y} \qquad D^2 = 2\sum_{l=1}^{L}\sum_{k=1}^{K} O_{lk} log\frac{O_{lk}}{E_{lk}}$$

donde se comparan los datos observados contra los datos esperados. Estas pruebas se aproximan a una distribución ji-cuadrada con *(L-1)*(K-1)-p* grados de libertad.

Pulkstenis considera que dichas pruebas no son buenas aproximaciones asintóticas a la ji-cuadrada cuando existen variables explicativas continuas o cuando el número de patrones L tiende al número de individuos *N*.

[1] Covariables es un término equivalente a variables explicativas

Debido a lo anterior propone subdividir en dos los patrones de covariables de la tabla basándose en los puntajes ordinales de los patrones de las observaciones. La división es con respecto a la mediana de los puntajes estimados de cada renglón *L*.

Los puntajes ordinales los define de manera similar a Lipsitz (1996).

$$puntajes = \hat{p}_1 + 2\hat{p}_2 + \cdots + K\hat{p}_K$$

donde $\hat{p}_K$ es $\hat{P}[Y = K]$.

Al subdividir los L patrones de covariables y al incluir los puntajes en las pruebas son

$$\chi^{*2} = \sum_{l=1}^{L}\sum_{h=1}^{2}\sum_{k=1}^{K}\frac{(O_{lhk} - E_{lhk})^2}{E_{lhk}} \quad \text{y} \quad D^{*2} = 2\sum_{l=1}^{L}\sum_{h=1}^{2}\sum_{k=1}^{K}O_{lhk}\log\frac{O_{lhk}}{E_{lhk}}$$

donde *l* denota los patrones de covariables, *h* indica la subdivisión realizada por los puntajes ordinales, y *k* corresponde a la respuesta categórica. Los grados de libertad a las pruebas modificadas son *(2L -1)*(K-1)-p-1*. Esta propuesta está sujeta a las mismas restricciones sobre los valores esperados, que se tienen en las pruebas de ji-cuadrada.

Con respecto al modelo estereotipo el ajuste del modelo es complicado y su complejidad aumenta según aumenta la dimensionalidad. Para este modelo, no existen pruebas de bondad de ajuste desarrolladas.

CAPÍTULO 3

UNA APLICACIÓN DE LOS MODELOS PRESENTADOS

Factores asociados a densidad mineral ósea en
Trabajadoras del IMSS Morelos

3.1 Presentación de la investigación

El estudio se realizó en el estado de Morelos, mediante un diseño transversal, de octubre de 1998 a marzo del 2000. Se midieron los niveles de densidad mineral ósea (DMO), así como factores asociados a dichos niveles en 1491 mujeres activas y jubiladas entre los 20 y 80 años de edad, pertenecientes a todas las categorías laborales del IMSS en dicha entidad.

La finalidad del estudio fue investigar sobre los niveles de densidad ósea prevalecientes en la población elegida, para encontrar factores determinantes que conducen a algún nivel de osteoporosis.

La información de interés se recolectó mediante cuestionarios autoaplicables que fueron contestados por las participantes del estudio, y que se entregaron a las trabajadoras en su sitio de trabajo o al momento de acudir a realizar su medición de DMO. La medición de DMO se realizó en el antebrazo con un densitómetro portátil. Esta variable se define como la cantidad de calcio, en gramos, contenida en un centímetro cuadrado de hueso (gr/cm^2).

Las variables que se midieron fueron:

- Densidad mineral ósea
- Edad
- Peso
- Talla
- Edad de la primera menstruación
- Menopausia
- Edad en que se presenta la menopausia
- Terapia Sustitutiva Hormonal (TSH)
- Actividad física
- Calcio en la dieta
- Calorías en la dieta
- Vitamina D en la dieta

La densidad mineral ósea es la variable de respuesta continua que se discretiza en niveles ordinales

- Normal (0)→ para valores mayores a -1 desviaciones estándar[1], que corresponde a niveles normales de densidad mineral ósea.

- Osteopenia (1) → osteoporosis leve, para valores entre -1 y -2.5 desviaciones estándar, que corresponde a niveles bajos de densidad de mineral ósea.

- Osteoporosis (2)→ para valores por debajo de -2.5 desviaciones estándar, que corresponde a niveles muy bajos de densidad mineral ósea.

Las variables edad, peso, talla, edad de la primera menstruación, edad en que se presenta la menopausia, etc; son variables explicativas continuas. La única variable explicativa que es categórica es menopausia.

Debido a que la variable menopausia es una variable dicotómica que representa la existencia o la no existencia de la característica, los investigadores crearon dos diferentes modelos.

3.2 Modelos para osteoporosis sin la presencia de la menopausia

El número de mujeres que no presentan la menopausia son 1102 y las variables a incluir en el modelo son:

	Nombre	Tipo de variable	Medición
Osteoporosis	Osteop0	ordinal	0,1,2
Edad	edad	continua	años
Ind. Masa Corporal[2]	IMC	continua	kg/m^2
Actividad Física	actfis	continua	mets[3]
Edad de la primera menstruación	edadmena	continua	años
Vitamina D	vitd	continua	UI[4]
Calorías	calor	continua	kcal[5]

[1] La población de referencia es la misma población participante del estudio, mujeres sanas de 20 a 35 años. Lo anterior fue propuesto por la OMS en 1994.

[2] Índice de Masa Corporal se calcula como peso/ talla^2.
IMC < 19 → Bajo peso
19 < IMC < 25 → Normal
25 < IMC < 30 → Sobrepeso

[3] Consumo de energía. Semanal en equivalentes metabólicos (mets)

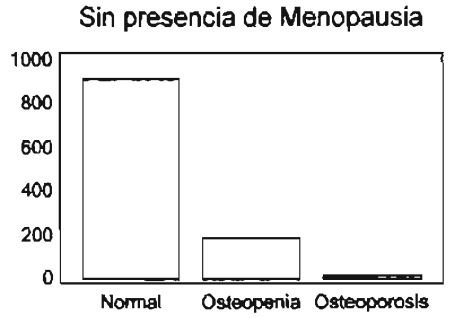
[4] Unidades Internacionales, consumo diario.

[5] Consumo diario de calorías

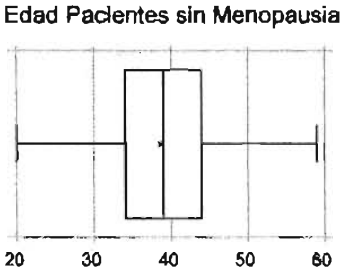
El análisis descriptivo de las variables explicativas es:

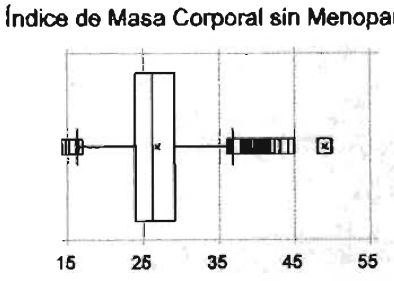
	Media	Desviación estándar
Edad	38.6851	6.4156
Ind. Masa Corporal	26.8218	4.3375
Actividad Física	8.4716	12.7996
Edad menstruación	12.6071	1.3669
Vitamina D	203.6761	121.0040
Calorías	2293.4800	732.4620

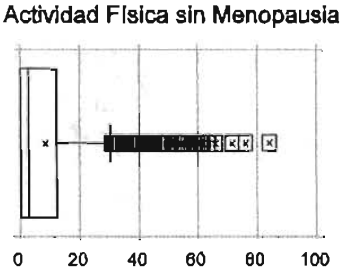
Las frecuencias de los datos observados son:

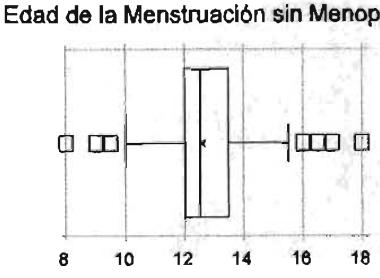


Normal	911
Osteopenia	182
Osteoporosis	9

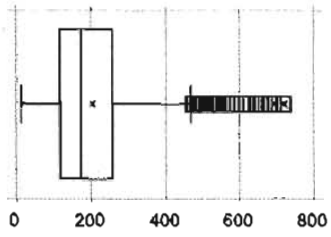




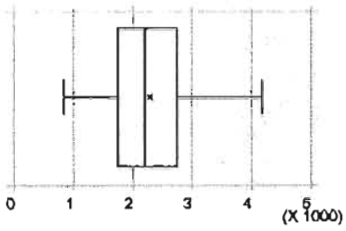




Vitamina D sin Menopausia



Calorías Promedio sin Menopausia



El análisis se realizará a través de los tres modelos propuestos y discutidos en el capítulo 2. Los análisis se realizaron en los paquetes estadísticos Stata 8.0, R 2.1.0 y SAS.

- Modelo de Momios Proporcionales

```
. ologit  osteop0 edad IMC actfis edadmena vitd calor if meno==0

Iteration 0:  log likelihood = -544.42714
Iteration 1:  log likelihood = -526.66063
Iteration 2:  log likelihood = -526.00824
Iteration 3:  log likelihood = -526.00601
Iteration 4:  log likelihood = -526.00601

Ordered logit estimates                           Number of obs   =       1102
                                                  LR chi2(6)      =      36.84
                                                  Prob > chi2     =     0.0000
Log likelihood = -526.00601                       Pseudo R2       =     0.0338

------------------------------------------------------------------------------
    osteop0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
       edad |   .0242928   .0131173     1.85   0.064    -.0014166    .0500022
        IMC |  -.1034378   .0223456    -4.63   0.000    -.1472345   -.0596412
     actfis |  -.0000252   .0064206     0.00   0.997    -.0126094    .0125591
   edadmena |   .0915072   .0595980     1.54   0.125    -.0253027    .2083172
       vitd |  -.0022179   .0008429    -2.63   0.009    -.0038700   -.0005658
      calor |   .0001792   .0001277     1.40   0.161    -.0000711    .0004294
------------+-----------------------------------------------------------------
      _cut1 |    .927359   1.109099    (Ancillary parameters)
      _cut2 |   4.204912   1.155429
------------------------------------------------------------------------------

Approximate likelihood-ratio test of proportionality of odds across response categories:
                chi2(6) =      1.35
                Prob > chi2 =   0.9689
```

En el análisis anterior observamos que las variables IMC y vitd son variables significativas a un nivel de significancia del 5%. Las variables edad, actfis, edadmena y calor son variables que no son significativas al mismo nivel de significancia.

La variable edad proporciona una significancia observada de 0.064, por lo que en la literatura médica, se le considera como marginalmente significativa.

La prueba del supuesto de momios proporcionales proporciona un valor de significancia descriptivo (valor p) de 0.9689, por lo que no se rechaza este supuesto a un nivel de significancia del 5%. Debido a lo anterior se dice que las proporciones entre los momios son estadísticamente iguales.

Los investigadores del área recomiendan que las variables como actfis, edadmena y calor son variables importantes y por lo tanto no deben eliminarse del análisis a pesar de que no sean estadísticamente significativas.

Debido a lo anterior, ninguna variable será eliminada del modelo y el ajuste del modelo implica lo siguiente con respecto a los coeficientes obtenidos

Momio para edad = exp (0.0242) = 0.024590

Si el coeficiente es negativo, el momio se calcula igual, pero se interpreta con el complemento a 1.

Momio para IMC = exp (-0.1034) = 0.9017
Complemento a 1 = (1 – 0.9017) = 0.09826

Cambios porcentuales en los momios de momios proporcionales sin menopausia

Variable	Cambio en una unidad	Cambio en 3 unidades	Cambio en 5 unidades	Cambio en 10 unidades
Edad	2.4590	7.5599	12.9148	27.4976
IMC	-9.8267	-26.6782	-40.3805	
Actfis	-0.0025	-0.0075	-0.0125	-0.0251
Edadmena	9.5824	31.5900	58.0175	
Vitd (×100)	-19.8916	-48.5916	-67.0094	
Calor (×100)	1.8081	5.5231	9.3736	

Los números de la tabla se interpretan de la siguiente manera:

- Para la variable edad, la estimación indica que existe un 2.45% de incremento en el momio para un menor nivel de densidad mineral ósea, es decir, pasar del nivel normal al de osteopenia, o de osteopenia a osteoporosis, por cada incremento en un año de edad. A pesar de que el porcentaje puede ser pequeño, si el incremento en edad se calcula cada 10 años, implica un incremento en el momio de 27.49% para disminuir de nivel de densidad mineral ósea por cada incremento de 10 años.

- Para la variable IMC, la estimación indica que existe un 9.82% de disminución en el momio para un menor nivel de DMO[1] (pasar de un nivel a otro), por cada incremento de unidad en el IMC. Si el incremento en IMC se calcula con 3 unidades de kg/m^2, entonces existe un 26.67% de disminución en el momio en un menor nivel de DMO.

- En la variable actfis, la estimación no presenta cambios significativos en los momios, aunque la tendencia es de disminución en los momios por cada incremento en el consumo de energía (mets).

- En la variable edadmena, la estimación indica que existe un 9.58% de incremento en el momio para un menor nivel de DMO, por cada incremento de un año en que una mujer tuvo su primera menstruación. Si el incremento en la edadmena se calcula con 3 años, entonces existe un incremento de 31.59% en el momio para un menor nivel de DMO.

- Para la variable vitd, la estimación indica que existe una disminución del 19.89% en el momio para un menor nivel de DMO por cada incremento de 100 UI[3]. Si el incremento es de 300 UI[3], entonces existe una disminución del 48.59% en el momio para un menor nivel de DMO.

- En la variable calor, la estimación no presenta cambios significativos en los momios, aunque la tendencia es de incremento en los momios para un menor nivel de DMO, cuando se incrementa el consumo de calorías.

Para las interpretaciones es importante recordar que los términos del momio de momios proporcionales son probabilidades acumuladas.

- Modelo de Cocientes continuos

```
. ocratio osteop0 edad IMC actfis edadmena vitd calor if meno==0

Continuation-ratio logit Estimates                Number of obs =   1293
                                                  chi2(6)       =  34.33
                                                  Prob > chi2   =  0.0000
Log Likelihood = -527.2627                        Pseudo R2     =  0.0315

------------------------------------------------------------------------------
     osteop0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0234323   .0127228     1.84   0.066    -.0015040    .0483687
         IMC |  -.0974789   .0215541    -4.52   0.000    -.1397242   -.0552335
      actfis |  -.0006862   .0062886    -0.11   0.913    -.0130117    .0116393
    edadmena |   .0862601   .0581950     1.48   0.138    -.0277999    .2003201
        vitd |  -.0020180   .0008156    -2.47   0.013    -.0036166   -.0004194
       calor |   .0001557   .0001243     1.25   0.210    -.0000879    .0003993
-------------+----------------------------------------------------------------
       _cut1 |   .9583012   1.075631     (Ancillary parameters)
       _cut2 |   2.6022210  1.125377
------------------------------------------------------------------------------
```

[1] Abreviatura de Densidad Mineral Ósea.

Approximate likelihood-ratio test of proportionality of odds across response categories[1]:
Prob > chi2 = 0.68319

En el ajuste del modelo encontramos que las significancias de las variables son muy similares a las obtenidas en el ajuste del modelo de momios proporcionales. Las variables estadísticamente significativas son IMC y vitd.

Es importante hacer notar que el número de observaciones que presenta el ajuste está modificado por las rutinas internas del paquete y que no representan el número de datos observados; Debido a lo anterior también se realizó el ajuste en otra paquetería estadística (R 2.1.0), en donde se encontró que los resultados del ajuste obtenido son iguales a los anteriores obtenidos con la paquetería STATA 8.0.

La variable edad es una variable marginalmente significativa y las variables actfis, edadmena y calor son variables no significativas. En este ajuste tampoco existe evidencia para rechazar el supuesto de momios proporcionales a un nivel de significancia del 5% (valor p = 0.68319).

El ajuste del modelo implica lo siguiente con respecto a los coeficientes obtenidos

Cambios porcentuales en los momios de cocientes continuos sin menopausia

Variable	Cambio en una unidad	Cambio en 3 unidades	Cambio en 5 unidades	Cambio en 10 unidades
Edad	2.3708	7.2826	12.4300	26.4052
IMC	-9.2878	-25.3557	-38.5775	
Actfis	-0.0685	-0.2056	-0.3425	-0.6838
Edadmena	9.0089	29.5349	53.9258	
Vitd (×100)	-18.1416	-45.1484	-63.2451	
Calor(×100)	1.5691	4.7818	8.0960	

Los cambios porcentuales obtenidos son muy similares a los cambios porcentuales en el modelo de momios proporcionales. Las interpretaciones de los cambios cambian ligeramente con respecto a las interpretaciones de los cambios del modelo de momios proporcionales. Lo anterior es debido a como se construyeron los momios, el momio del modelo de cocientes continuos contiene una probabilidad puntual y una probabilidad acumulativa.

[1] La prueba de momios proporcionales se realizó en R utilizando la librería VGAM.

Los cambios más importantes son:

- Para la variable edad, encontramos que existe un 2.37% de aumento en el momio para un menor nivel de DMO, para cada incremento de un año de edad.

- Para la variable IMC, el análisis indica que se disminuye un 9.28% el momio para un menor nivel de DMO, por cada unidad de IMC que se incremente.

- En la variable edadmena indica que existe un incremento del 9% en el momio para un menor nivel de DMO, por cada incremento de un año. Si el incremento se calcula con 5 años, encontramos que existe un aumento del 53.92% en el momio para un menor nivel de DMO.

- Para la variable vitd encontramos que existe una disminución del 18.14% en el momio para un menor nivel de DMO, por cada incremento del consumo de 100 UI^3.Si el incremento se calcula con 500 unidades, encontramos que la disminución es del 45.14% para un menor nivel de DMO.

- Modelo Estereotipo

```
. soreg osteop0 edad IMC actfis edadmena vitd calor if meno==0

iteration 0: Log Likelihood = -525.3721
iteration 1: Log Likelihood = -527.1806
iteration 2: Log Likelihood = -525.6863
iteration 3: Log Likelihood = -525.7493
iteration 4: Log Likelihood = -525.7757
iteration 5: Log Likelihood = -525.7719
iteration 6: Log Likelihood = -525.7719
iteration 7: Log Likelihood = -525.7719

Stereotype Logistic Regression                  Number of obs =   1102
Comparison to null model                        LR Chi2(7)    =  37.31
                                                Prob > chi2   =  0.0000
Comparison to full model                        LR Chi2(5)    =   1.17
                                                Prob > chi2   =  0.9476

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
phi11        |           1          .        .       .            .           .
phi21        |   -.4508689   1.452524    -0.31   0.756    -3.297764    2.396026
phi31        |  (dropped)
beta11       |   -.0165746   .0186398    -0.89   0.374    -.0531080    .0199588
beta21       |    .0721663   .0722376     1.00   0.318    -.0694169    .2137495
beta31       |   -.0003953   .0044957    -0.09   0.930    -.0092068    .0084162
beta41       |   -.0669249   .0777881    -0.86   0.390    -.2193866    .0855369
beta51       |    .0015894   .0016632     0.96   0.339    -.0016703    .0048491
beta61       |   -.0001358   .0001603    -0.85   0.397    -.0004500    .0001784
------------------------------------------------------------------------------
```

beta1 = edad
beta2 = IMC
beta3 = actfis
beta4 = edadmena
beta5 = vitd
beta6 = calor

En el análisis del modelo, observamos que ninguna variable es significativa a un nivel de significancia del 5%. Sin embargo se observa que el parámetro phi21 es estadísticamente igual a cero. Como el modelo nos permite proponer restricciones sobre estos parámetros (capítulo 2), lo ajustaremos con la restricción adicional de phi21 = 0. La paquetería de Stata 8.0 da la restricción natural de phi11 = 1. Entonces

```
. constraint define 1 phi11=1

. constraint define 2 phi21=0

.  soreg  osteop0 edad IMC actfis edadmena vitd calor if meno==0, c(1,2)

( 1)  phi11 = 1
( 2)  phi21 = 0

iteration 0: Log Likelihood = -525.3722
iteration 1: Log Likelihood = -525.6447
iteration 2: Log Likelihood = -525.7838
iteration 3: Log Likelihood = -525.7722
iteration 4: Log Likelihood = -525.7719
iteration 5: Log Likelihood = -525.7719

Stereotype Logistic Regression                  Number of obs =     1102
Comparison to null model                        LR Chi2(7)    =    37.31
                                                Prob > chi2   =   0.0000
Comparison to full model                        LR Chi2(5)    =    1.17
                                                Prob > chi2   =   0.9476
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
phi11	1	.	.	.	.	
phi21	(dropped)					
phi31	.3107575	.6900027	0.45	0.652	-1.041671	1.663186
beta11	-.0240469	.0133153	-1.81	0.071	-.0501444	.0020506
beta21	.1047033	.0227617	4.60	0.000	.0600912	.1493153
beta31	-.0005741	.0064992	-0.09	0.930	-.0133124	.0121641
beta41	-.0971006	.0605720	-1.60	0.109	-.2158196	.0216183
beta51	.0023061	.0008598	2.68	0.007	.0006208	.0039913
beta61	-.0001971	.0001298	-1.52	0.129	-.0004515	.0000574

beta1 = edad
beta2 = IMC
beta3 = actfis
beta4 = edadmena
beta5 = vitd
beta6 = calor

En este análisis observamos que al fijar una restricción adicional, el modelo se estereotipo se convirtió en un modelo unidimensional, el cuál es equivalente al modelo logístico para respuesta politómica no ordinal (capítulo 1). Debido a lo anterior el análisis obtenido es muy similar a los obtenidos con los modelos de momios proporcionales y de cocientes continuos. De hecho, el modelo estereotipo en este caso, es una reparametrización del modelo de momios proporcionales.

El ajuste del modelo implica lo siguiente con respecto a los coeficientes obtenidos

Cambios porcentuales en los momios de estereotipo sin menopausia

Variable	Cambio en una unidad	Cambio en 3 unidades	Cambio en 5 unidades	Cambio en 10 unidades
Edad	-2.3760	-6.9600	-11.3287	-21.3740
IMC	11.0381	36.9040	68.7952	
Actfis	-0.0573	-0.1720	-0.2866	-0.5724
Edadmena	-9.2535	-25.2709	-38.4612	
Vitd (×100)	25.9367	99.7367	216.7840	
Calor (×100)	-1.9517	-5.7415	-9.3849	

Las interpretaciones de la tabla son similares que las obtenidas en el modelo de momios proporcionales. El cambio de signo de las significancias es debido al cambio de signo en los parámetros $\beta's$ de los modelos que utiliza cada paquetería. Por lo que concluimos que los ajustes de los tres modelos son similares en este conjunto de datos.

3.3 Modelos para osteoporosis con la presencia de la menopausia

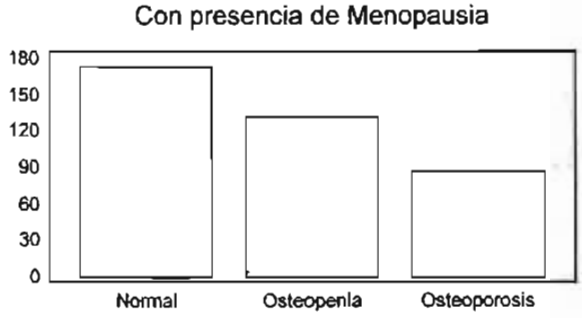
El número de mujeres que presentan la menopausia es 389 y las variables a incluir en el modelo son:

	Nombre	Tipo de variable	Medición
Osteoporosis	Osteop0	Ordinal	0,1,2
Edad	edad	continua	años
Ind. Masa Corporal	IMC	continua	kg/m^2
Actividad Física	actfis	continua	mets
Edad menopausia	edadmeno	continua	años
T. sustitutiva hormonal	TSH	continua	años
Calcio	calclac	continua	mg/día
Calorías	calor	continua	kcal

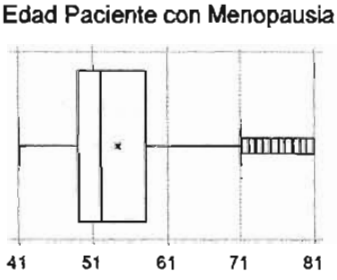
El análisis descriptivo de las variables es:

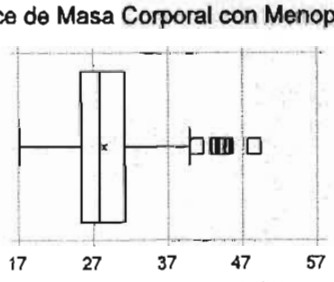
	Media	Desviación estándar
Edad	54.2365	7.2967
Ind. Masa Corporal	28.4091	4.8653
Actividad Física	6.4024	10.1702
Edad menopausia	46.4331	3.9378
T. sustitutiva hormonal	0.6693	1.8225
Calcio	150.494	131.154
Calorías	2277.6823	702.6532

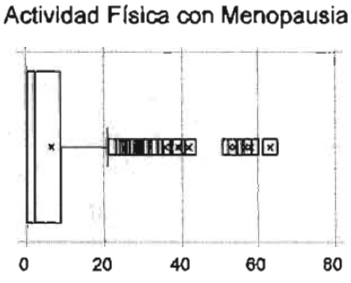
Las frecuencias de los datos observados son:

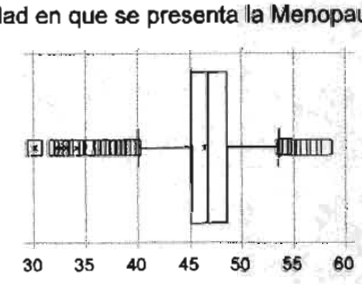


Normal	173
Osteopenia	131
Osteoporosis	65

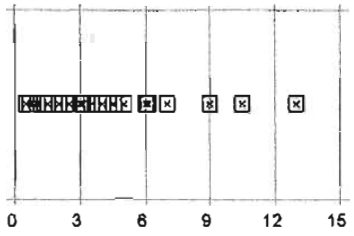




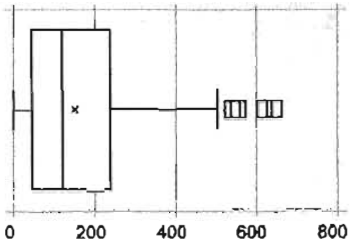




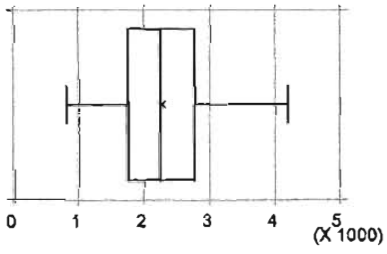
Terapia Sustitutiva Hormonal con Menopausia



Calcio en Dieta con Menopausia



Calorías Promedio con Menopausia



En este caso, los modelos ajustados son:

- Modelo de momios proporcionales

```
. ologit osteop0  edad IMC actfis edadmeno tsh calclac calor if meno==1

Iteration 0:  log likelihood = -412.03671
Iteration 1:  log likelihood = -342.56787
Iteration 2:  log likelihood = -341.00244
Iteration 3:  log likelihood = -340.98357
Iteration 4:  log likelihood = -340.98357

Ordered logit estimates                           Number of obs   =        389
                                                  LR chi2(7)      =     142.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -340.98357                       Pseudo R2       =     0.1724

------------------------------------------------------------------------------
     osteop0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |   .1647949   .0180479     9.13   0.000     .1294217    .2001680
         IMC |  -.0993574   .0225074    -4.41   0.000    -.1434710   -.0552438
      actfis |  -.0140413   .0106604    -1.32   0.188    -.0349353    .0068528
   edadmeno |  -.0961809   .0284006    -3.39   0.001    -.1518450   -.0405167
         tsh |  -.0648901   .0592411    -1.10   0.273    -.1810004    .0512203
    calclac |  -.0026786   .0008947    -2.99   0.003    -.0044321   -.0009251
      calor |   .0002321   .0001593     1.46   0.145    -.0000801    .0005443
-------------+----------------------------------------------------------------
       _cut1 |   1.250804   1.507427     (Ancillary parameters)
       _cut2 |   3.273813   1.517566
------------------------------------------------------------------------------
```

Approximate likelihood-ratio test of proportionality of odds across response categories:
chi2(7) = 11.76
Prob > chi2 = 0.1086

En este análisis observamos que las variables edad, IMC, edadmeno y calclac son estadísticamente significativas a un nivel de significancia del 5%. Las variables de actfis, tsh y calor son variables que no son estadísticamente significativas. A diferencia de los análisis sin la presencia de la menopausia encontramos que la edad pasa de ser marginalmente significativa a ser significativa.

El valor de significancia descriptivo de la prueba para el supuesto de momios proporcionales es igual a 0.1084, por lo que no se rechaza este supuesto a un nivel de significancia del 5%, lo cual significa que las proporciones entre los momios son estadísticamente iguales.

Los investigadores del área recomiendan que las variables como actfis, tsh y calor son variables importantes y por lo tanto no deben eliminarse del análisis a pesar de que no sean estadísticamente significativas.

Debido a lo anterior, ninguna variable será eliminada del modelo y el ajuste del modelo implica lo siguiente con respecto a los coeficientes obtenidos

Cambios porcentuales en los momios de momios proporcionales con menopausia

Variable	Cambio en una unidad	Cambio en 3 unidades	Cambio en 5 unidades	Cambio en 10 unidades
Edad	17.9151	63.9489	127.9541	419.6311
IMC	-9.4580	-25.7752	-39.1517	-62.9748
Actfis	-1.3943	-4.1249	-6.7798	-13.1000
Edadmeno	-9.1700	-25.0645	-38.1776	-61.7799
TSH	-6.2829	-17.6894	-27.7075	-47.7380
Calclac (×100)	-2.6430	-7.7214	-12.5348	
Calor (×100)	2.3481	7.2111	12.3052	

Los números de la tabla se interpretan de la siguiente manera:

- Para la variable edad, la estimación indica que existe un 17.91% de incremento en el momio para un menor nivel de densidad mineral ósea, por cada incremento en un año de edad. Si el incremento en edad se calcula cada 5 años, implica un incremento de más del doble en el momio (127.49%) para disminuir de nivel de DMO por cada incremento de 5 años.

- Para la variable IMC, la estimación indica que existe un 9.45% de disminución en el momio para un menor nivel de DMO, por cada incremento de unidad en el IMC. Si el incremento en IMC se calcula con 5 unidades, entonces existe un 39.15% de disminución en el momio en un menor nivel de DMO, por cada incremento de 5 kg/m^2.

- En la variable actfis a pesar de no ser una variable estadísticamente significativa, encontramos una tendencia de disminución de los momios por cada unidad de incremento en el consumo de energía (mets).

- En la variable edadmeno, la estimación indica que existe un 9.17% de disminución en el momio para un menor nivel de DMO, por cada incremento de un año en que una mujer tuvo su primera menstruación. Si el incremento en la edadmeno se calcula con 5 años, entonces existe un incremento de 38.17% en el momio para un menor nivel de DMO, por cada incremento de 5 años.

- Para la variable TSH, se indica que existe un 6.28% de disminución en el momio para un menor nivel de DMO, por cada incremento de un año. Si el incremento en TSH se calcula con 10 años, entonces existe una disminución de 47.73% en el momio para un nivel menor de DMO, por cada incremento de 10 años.

- Las variables como calclac y calor no presentan cambios porcentuales significativos en los momios. La primera variable presenta una tendencia de disminución en los momios por cada incremento en los mg/día, en la segunda variable se encontró que la tendencia es de incremento en los momios por cada incremento en el consumo diario de calorías.

- Modelo de cocientes continuos

```
. ocratio osteop0  edad IMC actfis edadmeno tsh calclac calor if meno==1

Continuation-ratio logit Estimates                Number of obs =   605
                                                  chi2(7)       =  138.95
                                                  Prob > chi2   =  0.0000
Log Likelihood = -343.1447                        Pseudo R2     =  0.1686

------------------------------------------------------------------------------
     osteop0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Edad |    .1429408   .0165401     8.64   0.000     .1105228    .1753588
         IMC |   -.0946837    .020492    -4.62   0.000    -.1348473   -.0545202
      actfis |   -.0121946   .0095417    -1.28   0.201    -.0308959    .0065067
    edadmeno |   -.0876159   .0258432    -3.39   0.001    -.1382677   -.0369642
         tsh |   -.0652869   .0536874    -1.22   0.224    -.1705122    .0399384
     calclac |   -.0022775   .0007934    -2.87   0.004    -.0038326   -.0007225
       calor |    .0002150   .0001423     1.51   0.131    -.0000640    .0004940
-------------+----------------------------------------------------------------
       _cut1 |    .6688728   1.343825          (Ancillary parameters)
       _cut2 |    2.003114   1.357131
------------------------------------------------------------------------------
```

Approximate likelihood-ratio test of proportionality of odds across response categories:
Prob > chi2 = 0.016858

El análisis para este modelo es muy semejante al que se obtuvo en momios proporcionales; sin embargo, se rechaza el supuesto de momios proporcionales a un nivel de significancia del 5%. Este hecho implica que el uso del modelo de cocientes continuos para analizar esta población, es inadecuado.

En los resultados observamos que al igual que en la población de mujeres sin menopausia, el número de observaciones es mayor que el número de datos observados, sin embargo la diferencia es producida por las rutinas internas de la paquetería.

Una alternativa, no considerada en este trabajo, es determinar cuál o cuáles de las variables no cumplen este supuesto y ajustar un modelo de cocientes continuos parciales (Ananth and Kleinbaum, 1997)

- Modelo estereotipo

. soreg osteop0 edad IMC actfis edadmeno tsh calclac calc if meno==1

iteration 0: Log Likelihood = -342.0016
iteration 1: Log Likelihood = -341.5631
iteration 2: Log Likelihood = -341.4842
iteration 3: Log Likelihood = -341.4718
iteration 4: Log Likelihood = -341.4699
iteration 5: Log Likelihood = -341.4696
iteration 6: Log Likelihood = -341.4696
iteration 7: Log Likelihood = -341.4696
iteration 8: Log Likelihood = -341.4696

Stereotype Logistic Regression — Number of obs = 389

Comparison to null model — LR Chi2(8) = 141.13
Prob > chi2 = 0.0000
Comparison to full model — LR Chi2(6) = 13.40
Prob > chi2 = 0.0370

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
phi11	1	.	.	.	.	.
phi21	.4324763	.0672802	6.43	0.000	.3006095	.5643431
phi31	(dropped)					
beta11	-.2521362	.0305208	-8.26	0.000	-.3119558	-.1923165
beta21	.1435464	.0347149	4.14	0.000	.0755065	.2115863
beta31	.0180853	.0154161	1.17	0.241	-.0121297	.0483003
beta41	.1557905	.0441506	3.53	0.000	.0692569	.2423241
beta51	.0630308	.0911451	0.69	0.489	-.1156102	.2416719
beta61	.0038066	.0013308	2.86	0.004	.0011983	.0064149
beta71	-.0003358	.0002353	-1.43	0.154	-.0007969	.0001253

beta1 = edad
beta2 = IMC
beta3 = actfis
beta4 = edadmeno
beta5 = tsh
beta6 = calclac
beta7 = calor

El análisis que se obtiene es muy semejante a los anteriores, por lo que los ajustes son equivalentes.

Debido a la importancia de las variables para los investigadores, ninguna variable será eliminada del modelo y el ajuste del modelo implica lo siguiente con respecto a los coeficientes obtenidos

Cambios porcentuales en los momios de estereotipo con menopausia

Variable	Cambio en una unidad	Cambio en 3 unidades	Cambio en 5 unidades	Cambio en 10 unidades
Edad	-22.2861	-53.0650	-71.6539	-91.9649
IMC	15.4360	53.8240	104.9779	320.1594
Actfis	1.8249	5.5754	9.4641	19.8239
Edadmeno	16.8581	59.5794	117.9188	374.8861
TSH	6.5059	20.8152	37.0470	87.8188
Calclac (×100)	3.8799	12.0974	20.9648	
Calor (×100)	-3.3022	-9.5831	-15.4561	

El cambio de signo de las significancias es debido al cambio de signo en el modelo de los parámetros $\beta's$ en los modelos que utiliza cada paquetería. Por lo que los cambios porcentuales de la tabla anterior (cambiando los signos) se interpretan de la siguiente manera:

- Para la variable edad, la estimación indica que existe un 22.28% de incremento en el momio para un menor nivel de DMO, por cada incremento en un año de edad. Si el incremento en edad se calcula cada 10 años, implica un incremento en el momio de 91.96% para disminuir de nivel de DMO por cada incremento de 10 años.

- Para la variable IMC, la estimación indica que existe un 15.43% de disminución en el momio para un menor nivel de DMO, por cada incremento de kg/m^2 en el IMC. Si el incremento en IMC se calcula con 5 unidades, entonces existe el doble (104.97%) de disminución en el momio en un menor nivel de DMO, por cada incremento de 5 kg/m^2

- En la variable edadmeno, la estimación indica que existe un 16.85% de disminución en el momio para un menor nivel de DMO, por cada incremento de un año en que una mujer tuvo su primera menstruación. Si el incremento en la edadmeno se calcula con 5 años, entonces existe un incremento de más del doble (117.91%) en el momio para un menor nivel de DMO, por cada incremento de 5 años.

- Para la variable TSH, se indica que existe un 6.50% de disminución en el momio para un menor nivel de TSH, por cada incremento de un año. Si el incremento en TSH se calcula con 10 años, entonces existe una disminución de 87.81% en el momio para un nivel menor de DMO, por cada incremento de 10 años.

- Las variables como actfis, calclac no son variables estadísticamente significativas, pero presentan una tendencia de disminución en los momios, cuando se incrementan las unidades de cada una de las variables. En la variable calor se observa una tendencia de incremento en los momios cuando se incrementa el consumo de calorías, a pesar de no ser una variable estadísticamente significativa.

3.4 Modelos para osteoporosis con categorización en orden inverso

La codificación que realizaron los investigadores para la variable de respuesta "osteoporosis" fué una categórica ordinal, sin embargo la asociación a la categorización creemos que debe ser de forma inversa, es decir, a valores normales de densidad mineral ósea se le asocie una categoría alta, a valores bajos o muy bajos de densidad mineral ósea, se le asocie valores bajos o muy bajos en las categorías.

Debido a lo anterior recodificamos la variable de respuesta de la siguiente manera:

- Normal (2)→ para valores mayores a -1[1], que corresponde a niveles normales de densidad mineral ósea.
- Osteopenia (1) → osteoporosis leve, para valores entre -1 y -2.5, que corresponde a niveles bajos de densidad de mineral ósea.
- Osteoporosis (0)→ para valores por debajo de -2.5, que corresponde a niveles muy bajos de densidad mineral ósea.

[1] La población de referencia es la misma población participante del estudio, mujeres sanas de 20 a 35 años. Lo anterior fue propuesto por la OMS en 1994.

Teniendo esta nueva codificación, la variable de respuesta se denomina *alrevés* y se realiza otra vez el análisis.

Como se mencionó en el capítulo 2, los modelos de momios proporcionales y estereótipo son invariantes a los cambios en el orden de las categorías (con sus debidas adecuaciones en las restricciones). Sin embargo en el modelo de cocientes continuos si se realiza un cambio en el orden de las categorías, entonces habrá cambios significativos en el ajuste del modelo.

- Modelo de Cocientes Continuos sin la presencia de la menopausia

```
. ocratio  alreves edad IMC actfis edadmena vitd calor if meno==0

Continuation-ratio logit Estimates              Number of obs =   2195
                                                chi2(6)       =  36.88
                                                Prob > chi2   = 0.0000
Log Likelihood = -525.987                       Pseudo R2     = 0.0339

------------------------------------------------------------------------------
alreves |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
edad    |  -.0241201   .0130184    -1.85   0.064    -.0496357    .0013955
IMC     |   .1026981   .0221822     4.63   0.000     .0592217    .1461745
actfis  |   .0000147   .0063727     0.00   0.998    -.0124755    .0125049
edadmena|  -.0906134   .0590199    -1.54   0.125    -.2062903    .0250634
vitd    |   .0022074   .0008378     2.63   0.008     .0005654    .0038495
calor   |  -.0001791   .0001267    -1.41   0.157    -.0004273    .0000691
--------+---------------------------------------------------------------------
_cut1   |  -4.205086   1.146887          (Ancillary parameters)
_cut2   |  -.9780208   1.100568
------------------------------------------------------------------------------

Approximate likelihood-ratio test of proportionality of odds across response categories:
                  Prob > chi2 =   0.952998
```

- Modelo de Cocientes Continuos con la presencia de la menopausia

```
. ocratio  alreves edad IMC actfis edadmeno tsh calclac calor if meno==1

Continuation-ratio logit Estimates              Number of obs =   693
                                                chi2(7)       =  139.82
                                                Prob > chi2   = 0.0000
Log Likelihood = -342.1288                      Pseudo R2     = 0.1697
```

alreves	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	-.1516922	.0166324	-9.12	0.000	-.1842911	-.1190933
IMC	.0879692	.0210519	4.18	0.000	.0467083	.1292301
actfis	.0119390	.0097285	1.23	0.220	-.0071286	.0310066
edadmeno	.0852437	.0254575	3.35	0.001	.0353480	.1351394
tsh	.0446062	.0552226	0.81	0.419	-.0636282	.1528406
calclac	.0024558	.0008316	2.95	0.003	.0008260	.0040857
calor	-.0002001	.0001455	-1.37	0.169	-.0004852	.0000851
_cut1	-3.334121	1.392791	(Ancillary parameters)			
_cut2	-1.709027	1.380975				

Approximate likelihood-ratio test of proportionality of odds across response categories:
Prob > chi2 = 0.099096

Las diferencias aunque no modifican las significancias de los modelos completos, si modifican las significancias de las variables explicativas. En particular el supuesto de momios proporcionales es diferente. En el modelo ajustado con la presencia de la menopausia y la codificación original, se rechaza el supuesto de momios proporcionales, y en el ajuste anterior, encontramos que es contrario. Así se confirma que un cambio en las categorías de la variable de respuesta, cambia el ajuste y las significancias estadísticas de las variables.

Es importante recordar que el número de observaciones que nos presentan los ajustes de los modelos en la paquetería de STATA 8.0, no es el número de datos observados. Este resultado esta modificado por rutinas internas y que no modifican, ni alteran los resultados obtenidos.

3.5 Conclusiones de la investigación

Los modelos que se analizaron y se discutieron con los investigadores en todo momento estuvieron de acuerdo en las variables incluidas. Sin embargo, por la naturaleza de los datos creemos que se pudieron utilizar otros análisis más adecuados y representativos para la investigación.

Los modelos se ajustaron razonablemente bien en forma global, mas al realizar la comparación de los datos observados frente a los esperados, observamos que el modelo de momios proporcionales predice perfectamente la categoría de nivel normal de DMO, las otras dos categorías las predice como si los datos fueran de nivel normal de DMO. En este estudio, el modelo estereotipo predice exactamente igual que momios proporcionales

. table predict1 osteop0 if meno==0

predict1	osteop0 0	1	2
0	911	182	9

. table predict1 osteop0 if meno==1

predict1	osteop0 0	1	2
0	135	63	18
1	35	53	27
2	3	15	40

En el modelo de cocientes continuos el porcentaje de datos que el modele predice de manera adecuada, es similar que en el de momios proporcionales.

. table predict2 osteop0 if meno==0

predict2	osteop0 0	1	2
0	911	182	9

. table predict2 osteop0 if meno==1

predict2	osteop0 0	1	2
0	138	65	18
1	29	48	23
2	6	18	44

Al implementar las pruebas de bondad de ajuste que se mencionan en el capítulo 2, se encontraron los siguientes inconvenientes:

Primeramente, todas las variables se consideraron continuas desde su creación, a pesar de que algunas pudieron considerarse como variables discretas. Para los modelos aquí analizados era muy importante la existencia de variables discretas, por lo que, con la finalidad de ejemplificar estas pruebas de bondad de ajuste, se realizó un ajuste en SAS discretizando algunas variables; lamentablemente se obtuvo que muchísimos valores esperados fueron menores a 5, lo que, como comentan los autores, invalida su uso, ya que no se cumple la convergencia de estas estadísticas a la χ^2. Debido a lo anterior, no se continuaron las pruebas de bondad de ajuste de los modelos con las variables discretizadas.

Entre los ajustes de los diferentes modelos no se encontraron diferencias significativas entre ellos. En relación a los datos obtenidos y recordando que solo presentamos tendencias de cada una de las variables explicativas, manteniendo a las demás variables explicativas constantes, podemos comentar lo siguiente:

- Al aumentar la edad, se aumenta la propensión de tener menor densidad mineral ósea. La propensión es bastante mayor cuando la mujer presenta la menopausia, que cuando no la presenta. Lo anterior es explicable debido a que durante la menopausia se presenta un menor nivel de absorción y retención del calcio por la falta de hormonas.

- Al aumentar el Índice de Masa Corporal, se disminuye la propensión de tener menor densidad ósea. Lo anterior es debido a que a más peso sobre la talla (IMC) se tiende a perder menos calcio. Se encontró similitud entre la propensión de las mujeres que presentan la menopausia a las mujeres que no la presentan.
- Si las mujeres aumentan la actividad física, entonces se disminuye la propensión de tener menor densidad ósea. La actividad física contribuye a la mayor absorción, depósito y retención de calcio.
- La propensión de tener una menor densidad ósea es bastante mayor cuando las mujeres han tenido su primera menstruación a edades tardías (aproximadamente 15 años) que cuando han tenido su primera menstruación a edades tempranas (aproximadamente 9 años).
- En las mujeres que no presentan la menopausia se tiene bastante menor propensión de tener menor densidad ósea al aumentar la cantidad de consumo de vitamina D. Lo anterior es debido a la vitamina D, promueve la absorción del calcio a nivel intestinal.
- Al aumentar la cantidad de consumo de calorías, se aumenta la propensión de tener menor densidad mineral ósea. Sin embargo se recomienda ajustar la cantidad de ingestión de calcio, según la edad e índice de masa corporal. Se encontró similitud entre la propensión de las mujeres que presentan la menopausia y las mujeres que no la presentan.
- En las mujeres que presentan la menopausia se tiene menor propensión a tener menor densidad ósea, cuando la edad a las que se les presenta la menopausia es mayor. Esto es explicable ya que la edad promedio para la presencia de la menopausia es de 54 años, y la literatura explica que a partir de esa edad se presenta una menor producción de hormonas, que deriva una menor densidad ósea.
- Las mujeres que presenta la menopausia tienen menor propensión de tener menor densidad ósea, cuando aumentan el tiempo de consumo de la terapia sustitutiva hormonal.
- Si el consumo de calcio aumenta, las mujeres que presentan la menopausia tienen menor propensión de tener menor densidad de mineral ósea.

Los investigadores comentan que las tendencias son relativas debido a que son modificadas dependiendo de cada individuo, de sus condiciones, costumbres, y hábitos.

BIBLIOGRAFÍA

- Ananth, Cande V. and Kleinbaum, David G, (1997). Regression Models for Ordinal Responses. A review of Methods and Aplications. Vol26, N°6.
- Agresti, Alan (1984). Analysis of Ordinal Categorical Data. Wiley
- Agresti, Alan (1990). Categorical Data Analysis. Wiley
- Anderson, J.A. (1984). Regression and Ordered Categorical Variables. Journal of the Royal Statistical Society, Series B 46 No. 1 1-30
- Brant, Rollin (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. Biometrics 46, 1171-1178
- Greenland, Sander (1994). Alternative Models for Ordinal logistic Regression. Statistics in Medicine. Vol. 13 1665-1677
- Greenwood, C and Farewell, V (1989). A comparison of regression models for ordinal data in analysis of transplanted kidney function. Canadian Journal of Statistics 16, 325- 336.
- Hosmer, David W. and Lemeshow, Stanley (2000). Applied Logistic Regression. John Wiley & Sons, Inc.
- López Caudana Alma Ethelia (2001), Factores Asociados a Densidad Mineral Ósea en Trabajadoras del IMSS Morelos. Unidad de Investigación Epidemiológica y en Servicios de Salud. "Predictors of bone mineral density en female workers in Morelos state, Mexico", Archives of Medical Research, abril 2004.
- Lunt, Mark. (Junio 2001). Stereotype Ordinal Regression. ARC Epidemiology Unit, Universidad de Manchester.
- Lipsitz S, Fitzmaurice G, Molenberghs G,. Goodness of fit test for ordinal response regression models. Applied statistics 1996. Vol 45, 175-190.
- Lunt, Mark. (Junio 2001). Predicting an Ordinal Outcome: Options and Assumptions. ARC Epidemiology Unit. Universidad de Manchester.
- McCullagh, Peter (1989). Generalized Linear Models. Chapman and Hall.

- Mood, Alexander M (1974). Introduction to the Theory of Statistics. McGrall Hill.
- Pulkstenis, E., RobinsonT.(2002). Goodness of fit test for logistic regression models with continuos covariates. Statistics in medicine; Vol 21, 79-93.
- Pulkstenis, Erik. (2004). Goodness of fit tests for ordinal response regression models; Statistics in medicine. Vol 23, 999-1014.
- SAS, Manual, Chapter 9, Logistic Regression II: Polytomous Response
- STAT, Manual, Cap XIII, D. Gillen, Regression Models for Ordinal Data
- VGAM, Thomas W. Yee (2004), Family Functions for Categorical Data, Department of Statistics, University of Auckland, New Zealand.