Regresión ordinal y sus aplicaciones

UNIVERSIDAD Ð SEVILLA

FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Dirigido por: Juan Manuel Muñoz Pichardo

Miguel Arias Benítez

Junio 2018

"It only ends once. Anything that happens before that is just progress."
Jacob, Lost.

Índice general

Resumen

El objetivo de este trabajo es la construcción de una base teórica con la que formular y desarrollar la forma general de algunos modelos de regresión ordinal, a su vez que motivar el estudio con una serie de ejemplos reales donde se emplearon los métodos descritos.

Para ello, a modo de introducción en el primer capítulo, se exponen una serie de conceptos teóricos sobre la idea estadística de la regresión; desde la definición y algunos conceptos de la distribución multinomial, hasta una descripción breve del modelo que tomaremos como base, el *Modelo Lineal Generalizado*.

En el capítulo segundo se presenta el *Modelo de Odds Proporcionales*, también conocido como *Modelo Logit Acumulado*. Primero se estudia e ilustra el caso bivariante para posteriormente su extensión a k-variables, donde se presenta la función de probabilidad y sus residuos, a su vez que se estiman los parámetros y se construyen los intervalos de confianza.

En el tercer capítulo se trabajará sobre un conjunto de datos en R para ilustrar la construcción y los resultados al aplicar las técnicas de regresión descritas en estos. Los datos empleados representan un experimento sobre ciertos factores que determinan la acidez en el vino; en nuestro caso, la temperatura y el contacto entre el zumo y las pieles de las uvas cuando se extrae de ellas.

Por último, en el capítulo cuarto, se describen una serie de estudios reales de diversas índoles sobre los que se aplicaron métodos de regresión ordinales con el objetivo de extraer una serie de conclusiones, por ejemplo, para la identificación de factores de riesgo en el cáncer de cuello uterino.

Summary

The main objective of this work is the construction of a theoretical basis with which to formulate and develop the general form of some models of ordinal regression, at the same time as motivating the study with some real examples where the described methods were used.

For this, as an introduction, in the first chapter some theoretical concepts on the statistical idea of regression are exposed; from the definition and some concepts of the multinomial distribution, to a brief description of the model that we will take as a base, the *Generalized Linear Model*.

In the second chapter we present the *Proportional Odds Model*, also known as the *Accumulated Logit Model*. First, the bivariate case is studied and illustrated for its extension to k variables, where the probability function and its residuals are presented, while the parameters are estimated and confidence intervals are constructed.

In the third chapter we will work on a set of data in R to illustrate the construction and the results when applying the regression techniques described. The data used represent an experiment on certain factors that determine the bitterness in the wine; in our case, the temperature and the contact between the juice and the skins of the grapes when it is extracted from them.

Finally, in the fourth chapter, we describe some real studies of various types on which ordinal regression methods were applied in order to dig out a series of conclusions, for example, for the identification of risk factors in cancer of the cervix.

Capítulo 1

Introducción a la Regresión Categórica

Mas allá de la regresión clásica binaria, existen otros casos en los que la variable objetivo cualitativa toma valores en diferentes grupos o modalidades. Cuando estos estén configurados de forma ordinal surge el problema que se tratará en este trabajo.

Este tipo de variables categóricas las encontraremos en muchos problemas reales; por ejemplo, los ciudadanos que votan a una serie de partidos políticos y a su vez valoran a sus líderes. En el primer caso no hay orden entre las categorías creadas y en el segundo, en cambio, sí lo hay.

1.1. Distribución Multinomial

En cuanto la regresión, en el caso de tener más de dos variables X explicativas y una variable objetivo no ordinal, nos será de capital importancia el uso de la distribución multinomial; ya que esta aportará una generalización natural del modelo logístico binario hacia otro donde la variable objetivo sea multicategórica.

Sea $y = (y_1, ..., y_n)'$ un vector de observaciones dado, con $y_i \in \{0, 1\}$ variables binarias y $\pi = (\pi_1, ..., \pi_n)'$ el vector de probabilidades asociadas, la función de probabilidad para un experimento aleatorio que se repite m veces viene dada por:

$$f(y|\pi) = \frac{m!}{y_1! \cdot ... \cdot y_n!(m - \sum_{i=1}^{n} y_i)!} \pi_1^{y_1} \cdot ... \cdot \pi_n^{y_n} (1 - \sum_{i=1}^{n} \pi_i)^{m - \sum_{i=1}^{n} y_i}. \tag{1.1}$$

Una variable aleatoria N-dimensional $Y = (Y_1, ..., Y_n)$ se dice que sigue una distribución multinomial de parámetros m y $\pi_1, ..., \pi_n$ si su función de probabilidad viene dada por la expresión 1.1 y se denotará por:

$$y \sim \mathcal{M}(m, \pi)$$

De donde se deduce:

$$\mathbb{E}(y) = m\pi = \begin{pmatrix} m\pi_1 \\ \vdots \\ m\pi_c \end{pmatrix}, \quad Cov(y) = m \begin{pmatrix} \pi_1(1-\pi_1) & \cdots & -\pi_1\pi_c \\ \vdots & \ddots & \vdots \\ -\pi_c\pi_1 & \cdots & \pi_c(1-\pi_c) \end{pmatrix}.$$

1.1.1. Familia exponencial

La distribución de una variable aleatoria Y, caracterizada por los parámetros θ y ϕ pertecene a la familia exponencial si presenta la forma:

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

$f()$ denota la función de probabilidad o de densidad en el caso en el que Y sea discreta o continua respectivamente. θ es el parámetro canónico, ϕ el parámetro de escala y $a(\phi)$,$b(\theta)$ y $c(y,\phi)$ son funciones específicas de cada elemento de la familia. La función $a(\phi)$ es comúnmente escrita como $a(\phi) = \phi/\omega$, donde ω es una ponderación para cada observación.
Se verifica:

$$E(Y) = \mu = b'(\theta) = \frac{\partial b(\theta)}{\partial\theta}; \quad Var(Y) = \sigma^2 = a(\phi)\frac{\partial^2 b(\theta)}{\partial\theta^2} = a(\phi)V(\mu).$$

$b'(\theta)$ y $b''(\theta)$ son, respectivamente, la primera y segunda derivadas de $b(\theta)$ y donde $V(\mu)$ se denomina **función de varianza**. Esta función relaciona $E(y)$ y $Var(y)$.

A continuación se mostrará una tabla donde se resumen los elementos principales que caracterizan a algunas de las distribuciones más usadas de la familia exponencial:

Distribuciones	Rango de Y	θ	$a(\phi)$	$b(\theta)$	$V(\mu)$
Binomial:$B(n,p)$	$\{0,n\}$	$ln\left(\frac{p}{1-p}\right)$	1	$nln(1+exp(\theta))$	$np(1-p)$
Gamma:$G(\mu,v)$	$(0,\infty)$	$-1/\mu$	$1/v$	$-ln(-\theta)$	μ^2
Normal:$N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	σ^2	$\theta^2/2$	1
Poisson:$P(\mu)$	$Ent[0,\infty)$	$ln(\mu)$	1	$exp(\theta)$	μ

1.2. Modelo lineal generalizado

La unificación de varios modelos estadísticos como el lineal, el logístico y el de Poisson fue realizada por Nelder y Wedderburn (1972) usando la idea de un modelo lineal generalizado. Como se describe en Dobson y Barnet [6] y en Nelder y Baker [17] este modelo está definido en términos de un conjunto de variable aleatorias independientes $Y_1, ...Y_n$, cada una de ellas con una distribución de la familia exponencial y con las siguientes propiedades:

1. La distribución de cada una de las Y_i tiene la forma estándar y depende de un único parámetro θ_i (los $\theta_i's$ no tienen por qué ser iguales); entonces tenemos:

$$f(y_i;\theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)].$$

2. La distribución de todas las $Y_i's$ son de la misma forma; por ejemplo, todas normales o todas binomiales, es por esto que los subíndices en b,c y d no son necesarios.

   Entonces tenemos, la función de densidad conjunta de las variables $Y_1, ...Y_n$ es:

$$\begin{aligned} f(y_1, ..., y_n; \theta_1, ..., \theta_n) &= \prod_{i=1}^{n} exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] \\ &= exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right]. \end{aligned}$$

Los parámetros θ_i no son de interés a menos que sean distintos para cada una de las observaciones. Para modelar estaremos interesados en un pequeño conjunto de parametros $\beta_1, ..., \beta_p$ (donde $p < n$). Suponiendo que $E(Y_i) = \mu_i$, donde μ_i es una función de θ_i. Para cada modelo lineal generalizado hay una transformación de μ_i tal que:

$$g(\mu_i) = x_i^T \beta.$$

En esta ecuación:

- g es una función llamada **función "enlace" o "link"**, que es continua, monótona en función de los valores de μ_i.
- El vector x_i es un (p x 1) vector de variables explicativas (covariables y variables dummy para distintos niveles),

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \quad entonces \quad x_i^T = [x_{i1} \cdots x_{ip}]$$

  y

- β es el (p x 1) vector de parámetros $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$. El vector x_i^T es la i-ésima fila de la matriz diseño X.

Entonces, un modelo lineal generalizado tiene 3 componentes:

1. Las variables respuesta $Y_1, ..., Y_n$, sobre las cuales se supone que comparten la misma distribución de la familia exponencial.

2. Un conjunto de parámetros β y de variables explicativas

$$X = \begin{bmatrix} x_1^t \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix};$$

3. Una función link monótona g tal que

$$g(\mu_i) = x_i^T \beta,$$

donde

$$\mu_i = E(Y_i).$$

Si queremos modelar una variable respuesta categórica, Y, de categorías $y_1, ..., y_c$ con un conjunto de variables explicativas (factores o covariables) $X = (X_1, ..., X_n)$, mediante un modelo lineal general, podemos plantearnos las opciones siguientes:

c	¿Y ordinal?	Regresión	Modelo
2	No importa	Logística	$f(P(Y = y_2\|X)) = \alpha + \beta' X$
$\geqslant 3$	No	Multinomial	$f(P(Y = y_j\|X_i)) = \alpha_i + \beta_j' X_i$ $j = 2, ..., c$; $i = 1, ..., n$
$\geqslant 3$	Sí	Ordinal	$f(\gamma_j(X)) = f(P(Y \leq y_j\|X)) = \alpha_j + \beta' X$ $j = 1, ..., c-1$

donde $f()$ es la función de enlace (usualmente Logit, Log-Log o Probit), $\alpha_j + \beta' X$ es el predictor lineal y α_j y $\beta = (\beta_1, ..., \beta_n)'$ parámetros a estimar.

1.2.1. Estimación de los parámetros

Dos de los métodos más comunes en la estimación estadística son el método de **Mínimos Cuadrados Ordinarios** y el **Método de Máxima Verosimilitud**. Usaremos este último ya que nos proporcionará las propiedades de consistencia y eficiencia asintótica (Vasconcellos y otros [19]) y procedemos como se describe en los primeros capítulos de Hardin y otros [8]

Sea la muestra $y_1, ..., y_n$ junto con las covariantes $x_1, ..., x_n$ maximizaremos la verosimilitud para obtener un estimador del vector de parámetros desconocidos β en el modelo:

$$E[Y_i|X_i = x_i] = \mu_i = h(x_i, \beta)$$

Suponemos que el parámetro de escala ϕ es conocido y dado que aparece como factor en la verosimilitud, puede considerarse $\phi = 1$, sin pérdida de generalidad. Posteriormente obtendremos un estimador de dicho parámetro mediante el método de los momentos.

Asumiendo que las distribuciones de cada componente de Y provienen de la familia exponencial de la forma denotada anteriormente, escribimos la función de verosimilitud como:

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta) \quad \text{con } y = (y_1, ..., y_n)'$$

Dado que las observaciones son independientes, la función *log-verosimilitud* viene dada por:

$$l(\theta, \phi, y) = \sum_{i=1}^{n} l_i(\theta_i, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

La función $c(y_i, \phi)$ que no depende de ϕ_i ha sido omitida. Añadiendo la relación $\theta_i = \theta(\mu_i)$ entre el parámetro natural y la esperanza de la i-ésima observación,

$$l(\mu_i, \phi, y) = \sum_{i=1}^{n} l_i(\beta, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i(\mu_i) - b(\beta_i(\mu_i))}{a(\phi)} \right\}$$

Dada la relación entre la esperanza y el vector de parámetros $\mu_i = h(x_i\beta)$, se tiene:

$$l(\beta, \phi, y) = \sum_{i=1}^{n} l_i(\beta, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\beta_i(h(x_i^t\beta)) - b(\theta_i(h(x_i^t\beta))}{a(\phi)} \right\}$$

Su primera derivada es la denominada **función score** o **función marcador**:

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_i s_i(\beta)$$

Las contribuciones individuales a la función marcador son:

$$s_i(\beta) = x_i D_i(\beta) \sigma_i^{-2}(\beta)[y_i - \mu_i(\beta)]$$

donde

$$\begin{cases} \mu_i(\beta) = h(x_i^t\beta) \\ \sigma_i^2(\beta) = a(\phi)v(h(x_i^t\beta)) \\ V(\mu) = \partial^2 b(\theta)/\partial\sigma^2 \\ D_i(\beta) = \partial h(x_i^t\beta)/\partial\eta \quad \text{con } \eta_i = x_i^t\beta \end{cases}$$

Otros conceptos importantes a tener en cuenta en la estimación máximo-verosímil del vector de parámetros son:

- **Matriz de información de Fisher esperada:**

$$F(\beta) = Cov\, s(\beta) = \sum_i F_i(\beta)$$

$$F_i(\beta) = x_i x_i^t w_i(\beta) \quad w_i(\beta) = D_i^2(\beta)\sigma_i^{-2}(\beta)$$

- **Matriz de Fisher observada:**

$$F_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t}$$

  verificándose que $F(\beta) = E(F_{obs}(\beta))$

Para las funciones de enlace naturales $\sigma(\mu_i) = x_i^t\beta$, las matrices se simplifican de la forma:

$$s(\beta) = \frac{1}{a(\phi)}\sum_i x_i[y_i - \mu_i(\beta)]$$

$$F(\beta) = \frac{1}{a(\phi)}\sum_i V(\mu_i(\beta))x_i x_i^t \quad F(\beta) = F_{obs}(\beta)$$

La obtención de la estimación de máxima-verosmilitud se plantea generalmente como las soluciones de la ecuación de verosimiltud $s(\hat{\beta}) = 0$ lo que corresponde a un máximo local, es decir, con la matriz de segundas derivadas $F_{obs}(\hat{\beta})$ definida positiva. Las ecuaciones resultantes no suelen ser lineales y para resolverlas necesitaremos de métodos numéricos iterativos como el de *Fisher Scoring* o el de *Mínimos Cuadrados Ponderados Iterativos*, cuyas iteraciones se definen a partir de un estimador inicial $\hat{\beta}^0$ por:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)})s(\hat{\beta}^{(k)}) \quad con : k = 0, 1, 2, ...$$

Obteniendo a través de estos métodos las estimaciones de los parámetros del modelo $\hat{\beta}$, estas estimaciones máximo-verosímiles tienen las propiedades de consistencia, eficiencia asintótica y distribución normal asintótica.

En el caso en el que el parámetro de dispersión sea desconocido, se puede considerar el siguiente estimador consistente:

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{[y_i - \mu_i(\hat{\beta})]^2}{v(\mu_i(\hat{\beta}))}$$

1.2.2. Bondad de ajuste

Una vez estimados los parámetros debemos valorar cuan bueno es nuestro modelo, es decir valorar la discrepancia entre los datos observados y los datos esperados.

De esta manera, determinar cuantos términos son necesarios en la estructura lineal para una descripción óptima de los datos intentando no saturarlo de variables explicativas que harán un modelo bien ajustado pero de dificil comprensión ni, en caso contrario, un defecto de variables que harán un modelo de fácil interpretación pero de pobre ajuste es otro de los problemas que nos plantearemos.

Trataremos de constuir un modelo intermedio entre el *modelo saturado* y el *modelo nulo*, donde el primero se refiere al modelo en el que el número de parámetros es igual al número de observaciones (ninguna simplificación) y el segundo que es el modelo más simple en el que solo se usa el parámetro μ, el valor esperado para todas las observaciones (simplificación total, asume efecto nulo de las variables explicativas)

En el *modelo lineal generalizado*, la bondad de ajuste se puede evaluar de distintas formas, entre ellas destacan:

- **La función o estadístico desviación**

$$D(y;\mu) = 2\{l(y;y) - l(\hat{\mu};y)\}$$

Es la distancia entre el logaritmo de la función verosimilitud del modelo saturado y el modelo con el que se está trabajando.
Un valor pequeño de la desviación indica que para un número menor de parámetros, se obtiene un ajuste tan bueno como cuando se ajusta el modelo saturado.

Si el modelo es correcto el estadístico se distribuye asintóticamente segun una χ^2_{n-p} con $n-p$ grados de libertad [15].

$$D(y,\hat{\mu}) \sim \chi^2_{n-p}$$

- **Coeficiente de determinación R^2:**
  La medida R^2 se define como la proporción de la varianza total de la variable explicada por la regresión. El R^2, también llamado *coeficiente de determinación*, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.
  Este coeficiente viene dado por:

$$R^2 = 1 - \frac{D(y, \hat{\mu})}{D(y, \hat{\mu}_0)}$$

  donde $D(y, \hat{\mu})$ y $D(y, \hat{\mu}_0)$ son las funciones de desviación del modelo ajustado y nulo respectivamente. Se verifica que $0 \leq R^2 \leq 1$

- **Estadístico Chi-cuadrado de Pearson:**

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i}$$

  donde $V(\hat{\mu})$ es la función varianza estimada para la distribución de la variable objetivo.

En cuanto a lo que a procesos de seleción de modelos se refiere mencionaremos el **Criterio de información Akaike (AIC)**, en el caso general el AIC es:

$$AIC = k - 2ln(\hat{L})$$

donde k es el número de parámetros y $\hat{L}$ es el máximo valor de la función de verosmilitud para el modelo estimado. El modelo se optimiza minimizando el valor de AIC, este modelo recompensa la bondad de ajuste y penaliza el aumento de la cantidad de parámetros estimados.

1.2.3. Regiones de confianza para β

A continuación construiremos un intervalo de confianza realizando inferencias sobre el vector de parámetros desconocidos β de dimensión p, la mayoria de cuestiones consiguen reformularse a traves de una hipótesis lineal de la forma $C\beta$, siendo C una matriz de rango total $s \leq p$ y ξ un vector de constantes conocido de dimensión s

$$H_o : C\beta = \xi$$
$$H_1 : C\beta \neq \xi$$

Para este procedimiento se puede usar el **Estadístico de Wald** entre otros. Éste se basa en la distribución asintótica del vector $\hat{\beta}$ y está definido por:

$$\xi_W = [C\hat{\beta} - \xi]^T [CF^{-1}\hat{\beta}(C)'][C\hat{\beta} - \xi]$$

determina la distancia ponderada entre el estimador $C\hat{\beta}$ y su valor determinado por la hipótesis nula.
$F^{-1}(\hat{\beta})$ denota la estimación de la matriz de información de Fisher de $\hat{\beta}$

Asintóticamente y bajo hipótesis nula este estadístico se distribuye como una distribución Chi-cuadrado con s grados de libertad χ^2_s

Usando este estadístico, una región de confianza para β con un nivel de confianza del $100(1-\alpha)\,\%$ viene dada por:

$$\{\beta \in \mathbb{R}^p \mid (\hat{\beta}-\beta)^T[Var(\hat{\beta})]^{-1}(\hat{\beta}-\beta) < \chi^2_{p,1-\alpha}\}$$

1.2.4. Residuos

En la práctica podemos encontrar el problema de que aún habiendo escogido cuidadosamente un modelo, al ajustarlo a un conjunto de datos el resultado sea insatisfactorio.

Las desviaciones sistemáticas se originan por haber escogido inadecuadamente la función de enlace o las variables explicativas. Las diferencias aisladas pueden darse por puntos extremos o porque estos realmente sean erróneos, éstos se conocen como *outliers*. La comprobación de la adecuación del modelo es un requisito fundamental que se realiza sobre el conjunto de datos para encontrar posibles fallos en las suposiciones hechas por el modelo, así como los *outliers* que puedan interferir desproporcionadamente en los resultados del ajuste.

Como en la regresión lineal, los residuos son los utilizados para verificar esta adecuación del modelo. Expresan la discrepancia entre una observacción y su valor ajustado y también pueden indicar la presencia de valores anómalos que puedan requerir un estudio más concreto. Entre otros residuos los más destacados son:

- **El residuo básico:**

  Definido como la diferecnia entre le valor observado, y_i, de la variable respuesta y el valor ajustado, $\hat{y}_i$, por el modelo.

  $$r_i^b = y_i - \hat{y}_i \quad con\, i = 1, ..., n$$

- **El residuo de Pearson:**

  Es la contribución individual al estadístico χ^2 de Pearson, se define como:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\, Var(\hat{\mu}_i)}} \quad con\, i = 1, ..., n$$

  siendo $\hat{\phi}$ un estimador consistente del parámetro escala ϕ.

  Y su versión studentizada viene dada por:

$$r_{s_i}^{\ p} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\, Var(\hat{\mu}_i)(1 - h_i)}} \quad con\, i = 1, ..., n$$

  siendo h_i el elemento diagonal de la matriz H, donde:

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$$

  con W una matriz diagonal cuyos elementos de la diagonal principal son:

$$w_i = \frac{1}{Var(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta} \right)^2$$

  La ventaja de este residuo studentizado frente al anterior reside en que la captación de la variabilidad de los datos es mejor debido a que usa el valor h_i, este es útil para medir la influencia de cada observación.

- **El residuo desviación:**

  Se define como:

$$r_i^D = sign(y_i - \hat{\mu}_i \sqrt{d_i} \quad i = 1, ..., n$$

  d_i es el llamado *componente desviación*, $d_i = 2(l(y_i, y_i) - l(\mu_i\hat{,} y_i)$

  Y su versión studentizada:

$$r_{s_i}^{\ D'} = \frac{r_i^D}{\sqrt{\hat{\phi}(1 - h_i)}}$$

  donde h_i es el i-ésimo elemento de la diagonal de la matriz H y $\hat{\phi}$ es la estimación del parámetro de escala ϕ.

1.3. Modelos logit para respuestas nominales

Al igual que se describe en Marín [14], se denota C como el número de categorías de la variable Y y $\{\pi_1, ..., \pi_j\}$ las probabilidades de cada respuesta, satisfaciendo $\sum_j \pi_j = 1$.

Se parte de n observaciones independientes extraidas. La distribución de probabilidad del número de observaciones de las C categorías sigue una distribución multinomial tal y como nombramos anteriormente. Esta modeliza la probabilidad de cada una de las posibles maneras en que n observaciones pueden repartirse entre C categorías.
Al ser la respuesta nominal, el orden entre las categorías es irrelevante; este es el problema central que trataremos más adelante.

Se toma una categoría como respuesta **base**, por ejemplo la última categoría (C), y se define un modelo logit con respecto a ella:

$$log\left(\frac{\pi_c}{\pi_C}\right) = \alpha_c + \beta_c x$$

donde $c = 1, ..., C-1$.

El modelo tiene $C-1$ ecuaciones con sus propios parámetros, y los efectos varían con respecto la categoría que se ha tomado como base.

Cuando $C = 2$, el modelo equivale a una única ecuación $log(\pi_1/\pi_2) = logit(\pi_1)$ y se obtiene el modelo de regresión logística estándar.
La ecuación general logit con respecto a la categoría base C determina también los logits para cualquier pareja de categorías. Considerando c_1 y c_2 dos categorías cualesquiera tenemos:

$$\begin{aligned} log\left(\frac{\pi_{c_1}}{\pi_{c_2}}\right) &= log\left(\frac{\pi_{c_1}/\pi_C}{\pi_{c_2}/\pi_C}\right) = log\left(\frac{\pi_{c_1}}{\pi_C}\right) - log\left(\frac{\pi_{c_2}}{\pi_C}\right) \\ &= (\alpha_{c_1} - \beta_{c_1} x) - (\alpha_{c_2} - \beta_{c_2} x) \\ &= (\alpha_{c_1} - \alpha_{c_2}) + (\beta_{c_1} - \beta_{c_2})x. \end{aligned}$$

De este modo, la ecuación para las categorías c_1 y c_2 tiene también la forma $\alpha + \beta x$ donde $\alpha = (\alpha = (\alpha_{c_1} - \alpha_{c_2})$ y $\beta = (\beta_{c_1} - \beta_{c_2})x$

Capítulo 2

Modelos de regresión ordinal

En este capítulo, la regresión logística estándar será extendida para soportar variables respuesta que tengan más de dos categorías ordenadas. Cuando las categorías de la variable respuesta tengan un orden natural la regresión logística ordinal será la elegida como forma más óptima de estudiar estos datos.

Se estudiará la forma matemática general del modelo de regresión logística ordinal, así como se desarrollará su interpretación, las fórmulas para la *odds ratio*, intervalos de confianza, técnicas para tests de hipótesis y para el estudio de la significación de la variable objetivo.

Para el desarrollo de este capítulo se ha utilizado diversa bibliografía recogida al final de la memoria, especialmente Harrell [9], Kleinbaum y otros [12] y Kleinbaum y Klein[11]

2.1. Modelo de Odds Proporcionales

El modelo logístico ordinal que se va a desarrollar es el llamado de *Odds Proporcionales* o *modelo de odds proporcionales*, también conocido como el *Modelo Logit Acumulado*.

Para ilustrar la idea del *modelo odds proporcional* asumimos que tenemos una variable respuesta con cinco categorías y consideramos las cuatro posibles formas de dividirlas en sólo dos categorías respetando el orden natural. Por ejemplo, todas estas divisiones serían posibles respetando el orden:

0 |	1	2	3	4
0	1 |	2	3	4
0	1	2 |	3	4
0	1	2	3 |	4

Generalmente, si una variable respuesta ordinal D tiene G categorías ($D = 0, 1, 2, ..., G-1$), entonces hay $G-1$ formas de dicotomizar la respuesta: ($D \geq 1$ ó $D < 1$; $D \geq 2$ ó $D < 2, ..., D \geq G-1$ ó $D < G-1$). Para un suceso aleatorio S, se define su "odds" o "ventaja" como la razón entre la probabilidad de ocurrencia y la probabilidad de no ocurrencia.
Con la categorización de D, se puede definir la "odds" o "ventaja" de que $D \geq g$ dividida por la probabilidad de que $D < g$, i.e.

$$odds\,(D \geq g) = \frac{\mathbf{P}(D \geq g)}{\mathbf{P}(D < g)} \quad donde\ g = 1, 2, 3, ..., G-1$$

El *modelo odds proporcional* hace una importante suposición. Bajo este modelo, el *odds ratio* que evalúa el efecto de una variable explicativa para cualquiera de las divisiones o categorizaciones anteriores será el mismo independientemente de donde se realice el punto de corte sobre las categorias.

Suponemos que tenemos una variable respuesta con cinco niveles y una variable explicativa dicotómica ($E = 0$, $E = 1$). Entonces, bajo la suposición de *odds proporcionales*, el *odds ratio* que compara categorías iguales o mayores que 1 y categorías menores que 1 es el mismo que el que compara categorías mayores o iguales a 4 con categorías menores que 4. Formalmente:

$$OR(D \geq 1) = \frac{odds\,[(D \geq 1)|E = 1]}{odds\,[(D \geq 1)|E = 0]} = \frac{odds\,[(D \geq 4)|E = 1]}{odds\,[(D \geq 4)|E = 0]} = OR(D \geq 4)$$

En otras palabras, el *odds ratio* es invariante al punto utilizado para la dicotomización.
Esto implica que si hay G categorías en la respuesta, solo hay un parámetro (β) para cada una de las variables predictoras o explicativas. Sin embargo sigue habiendo constantes separadas (α_g) para cada una de las $G-1$ comparaciones.

Esto contrasta con la regresión logística politómica[1], donde hay $G-1$ parámetros para cada variable predictora, así como constantes separadas para cada una de las $G-1$ comparaciones. En resumen:

Variable	Parámetro
Constante	$\alpha_1, \alpha_2, ..., \alpha_{G-1}$
X_1	β_1

Cuadro 2.1: Ordinal

Variable	Parámetro
Constante	$\alpha_1, \alpha_2, ..., \alpha_{G-1}$
X_1	$\beta_{11}, \beta_{21}, ..., \beta_{G-1}$

Cuadro 2.2: Politómica

La hipótesis de invarianza del *odds ratio* en cuanto a los puntos de corte no es la misma que suponer que el *odds* dado para un patrón de exposición es invariante. Usando

[1] No solo dicotómica, multi-categórica

el ejemplo anterior, para una realización dada de $E\,(e.j.\,, E = 0)$, el *odds* que compara categorías mayores o iguales a 1 con las menores **no es igual** al *odds* que compara categorías mayores o iguales a 4 con las menores.

$$odds\,(D \geq 1) \neq odds\,(D \geq 4)$$

donde, para $E = 0$

$$odds_{(D \geq 1)} = \frac{\mathbf{P}\,(D \geq 1)|E = 0)}{\mathbf{P}\,(D \geq 1)|E = 0)} \neq \frac{\mathbf{P}\,(D \geq 4)|E = 0)}{\mathbf{P}\,(D \geq 4)|E = 0)} = odds_{(D \geq 4)}$$

pero

$$OR\,(D \geq 1) = OR\,(D \geq 4)$$

2.1.1. Presentación del modelo

Procedemos ahora a presentar la forma del *modelo odds proporcional* con una respuesta D de G niveles $(D = 0, 1, 2, ..., G - 1)$ y una variable explicativa X_1.El modelo expresa la probabilidad de que la variable respuesta esté es una categoría igual o superior a g en función de la variable explicativa X_1 como sigue:

$$P(D \geq g\,|X_1) = \frac{1}{1 + exp[-(\alpha_g + \beta_1 X_1)]}, \quad g = 1, 2, ..., G - 1$$

Por tanto, la probabilidad de que la variable respuesta esté en una categoría *inferior* a g es:

$$P(D < g\,|X_1) = \frac{exp[-(\alpha_g + \beta_1 X_1)]}{1 + exp[-(\alpha_g + \beta_1 X_1)]}$$

El modelo puede ser definido equivalentemente en términos del odds de una desigualdad. Si sustituimos la formula $P(D \geq |X_1)$ por la expresión para el *odds* entonces:

$$odds(D \geq g\,|X_1) = \frac{P(D \geq g\,|X_1)}{1 - P(D \geq g\,|X_1)} = \frac{P(D \geq g\,|X_1)}{P(D < g\,|X_1)} =$$
$$= exp(\alpha_g + \beta_1 X_1) = e^{\alpha_g} \cdot e^{\beta_1 X_1}$$

El *modelo de odds proporcional* está escrito en términos diferentes al *modelo logístico estándar*. El modelo se formula como la probabilidad de una desigualdad, esto es, que la variable respuesta D sea mayor o igual a g.

Modelo Odds Proporcional	Modelo Logístico Estándar
$P(D \geq g \,\vert X)$	$P(D = g \vert X)$

2.1.2. Odds Ratios e Intervalos de Confianza

Primero consideraremos el caso especial donde la variable explicativa X_1 es la única variable independiente y es dicotómica ($X_1 = 0$ ó $X_1 = 1$). Según lo recogido en el apartado anterior, el *odds* que compara $D \geq g$ con $D < g$ es $exp(\alpha_g + \beta_1 X_1)$. Para evaluar el efecto de la variable explicativa sobre la variable respuesta formulamos el llamado *odds ratio* de $D \geq g$ para comparar $X_1 = 0$ y $X_1 = 1$ (i.e., el *odds ratio* para $X_1 = 0$ vs. $X_1 = 1$).

$$OR(D \geq g \,\vert X_1) = \frac{odds(D \geq g \,\vert X_1 = 1)}{odds(D \geq g \,\vert X_1 = 0)} = \frac{exp(\alpha_g + \beta_1)}{exp(\alpha_g)} = e^{\beta_1}$$

Es decir, la odds ratio es constante para cualquier punto de corte g considerado. Además, el coeficiente β_1 es:

$$\beta_1 = log\, OR(D \geq g \,\vert X_1) \quad \forall g$$

Análogamente, en el caso de X_1 variable cuantitativa, la comparación entre dos valores X_1 y X_1^* de la misma:

$$OR(D \geq g \,\vert X_1, X_1^*) = exp\big\{\beta_1(X_1^* - X_1)\big\}$$

El cálculo del intervalo de confianza es equivalente al cálculo descrito en el Capítulo 1. Así, la fórmula general para un intervalo de confianza al 95 % de confianza para dos cualesquiera niveles de la variable independiente X_1 y X_1^* es el siguiente:

$$IC\,95\,\% = exp\left[\hat{\beta}_1(X_1^* - X_1) \pm 1{,}96(X_1^* - X_1)s_{\hat{\beta}_1}\right]$$

siento $\hat{\beta}$ el estimador de máxima-verosimilitud del modelo y $s_{\hat{\beta}}$ el error de estimación del mismo.

2.1.3. Extensión del Modelo Ordinal a k variables

Expandir el modelo para añadir más variables explicativas se obtiene de forma directa, basta expandir el predictor lineal.
Reprsentando por $\underline{X}$ el vector aleatorio de variables explicativas, el modelo se puede expresar por:

$$P(D \geq g\,|\underline{X}) = \frac{1}{1 + exp[-(\alpha_g + \sum_{i=1}^{k} \beta_i X_i)]}, \quad g = 1, 2, 3, ..., G-1$$

El *odds* para la respuesta mayor o igual al nivel g sería el siguiente:

$$odds(D \geq g\,|\underline{X}) = \frac{P(D \geq g\,|\underline{X})}{P(D < g\,|\underline{X})} = exp\,(\alpha_g + \sum_{i=1}^{k} \beta_i X_j)$$

Como en la *regresión logística estándar*, el uso de múltiples variables independientes permite la estimación del *odds ratio* para una variable controlando los efectos de las demás variables explicativas del modelo.

$$OR = exp\,(\beta_i) \qquad X_i \in \{0, 1\}$$

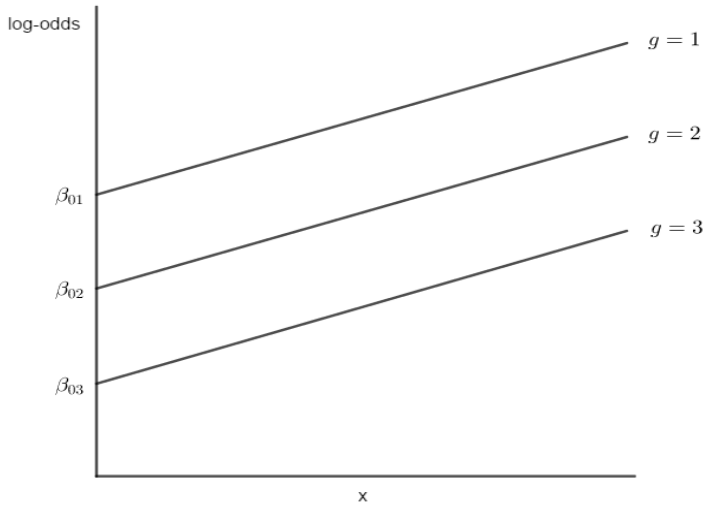

Figura 2.1: Modelo de odds proporcionales sobre escala odds

La figura 2.1 muestra el modelo para $G = 3$ categorías y una variable continua explicativa X

2.1.4. Función de Probabilidad para el Modelo Ordinal

A continuación, se recoge brevemente la deducción de la función de probabilidad para el *modelo de odds proporcional*. Para formularla, necesitamos las probabilidades observadas en las respuestas para cada individuo. Una expresión para estas probabilidades en términos de los parámetros que rigen el modelo puede ser obtenida por la relación siguiente:

$$P = odds/(odds + 1), \quad \text{o la expresión equivalente} \quad P = 1/[1 + (1/odds)]$$

dado que

$$odds = \frac{P}{1-P} \quad \Rightarrow \quad P = \frac{odds}{odds+1} = \frac{1}{1+(\frac{1}{odds})}.$$

En el *modelo de odds proporcional*, modelamos la probabilidad de que $D \geq g$. Para obtener una expresión para la probabilidad de que $D = g$, podemos usar la relación:

$$P(D = g) = P(D \geq g) - P(D \geq (g+1))$$

De esta forma, podemos calcular la probabilidad de que un individuo esté es una categoría específica para unas variables explicativas X_i dadas.

La **Función de Probabilidad (L)** se calcula tomando el producto de las contribuciones individuales como sigue:

$$L = \prod_{j=i}^{n} \prod_{g=0}^{G-1} P(D = g\,|X)^{y_{jg}} \qquad y_{jg} = \begin{cases} 1 & \text{si el j-ésimo sujeto cumple D = g} \\ 0 & \text{c.c.} \end{cases}$$

2.1.5. Estimación de parámetros

Se puede estimar los parámetros del modelo por máxima verosimilitud, maximizando la función de verosimilitud:

$$L(\alpha, \beta|Y, X) = \cdots = \prod_{i=1}^{n} \prod_{j=2}^{g-1} \left[\frac{1}{1+e^{-(\alpha_1+\beta' X_j)}}\right]^{\delta_{j1}} \left[\frac{1}{1+e^{-(\alpha_j+\beta' X_j)}} - \frac{1}{1+e^{-(\alpha_{j-a}+\beta' X_j)}}\right]^{\delta_{ij}}$$

donde:

$$\delta_{ij} = \begin{cases} 1 & \text{si el i-ésimo individuo muestra } Y = y_j \\ 0 & \text{en caso contrario} \end{cases}$$

De aquí por las propiedades de los estimadores de MV,[2]

[2]F es la matriz de información de Fisher

$$\hat{\theta}_{k,MV} \overset{asint.}{\sim} N\left(\theta_k, \sqrt{\hat{F}_{kk}^{-1}}\right)$$

Análogamente al *Modelo Lineal Generalizado*, se puede realizar la *prueba de Wald* para resolver el contraste de hipótesis

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

con el estadístico de contraste

$$\frac{\hat{\beta}_k}{\sqrt{\hat{F}_{kk}^{-1}}} \overset{H_0}{\sim} N(0,1), \quad \acute{o}\, equivalentemente \quad \frac{\hat{\beta}_k^2}{\hat{F}_{kk}^{-1}} \overset{H_0}{\sim} \chi_1^2$$

2.1.6. Residuos

Para el *modelo de odds proporcional* se puede analizar la contribución individual de cada sujeto a la primera derivada de la función de log-probabilidad respecto a β_m, promediándolos por separado según los niveles de Y, la variable objetivo, y examinando las tendencias en las gráficos de los residuos. Este método es complejo ya que las gráficas de los residuos no suelen ser fáciles de intepretar.

Los residuos parciales para el i-ésimo sujeto y la m-ésima variable explicativa se definen como sigue

$$r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)},$$

donde

$$\hat{P}_i = \frac{1}{1 + exp\left[-(\alpha + X_i\hat{\beta})\right]}$$

Una gráfica más suave[3] de X_{im} y r_{im} proporciona una estimación no paramétrica de cuanto influye X_m al log-odds relativo, que asume $Y = 1\,|X_m$. Para una Y ordinal, necesitamos simplemente repetir en cada corte de nivel g,

$$r_i m = \hat{\beta}_m X_{im} + \frac{[Y_i \geq g] - \hat{P}_{ig}}{\hat{P}_{ig}(1 - \hat{P}_{ig})}$$

después se debe hacer una gráfica para cada m mostrando una curva suave para cada g y buscar formas o pendientes similares para cada g con una variable predictora fija, cada

[3]Más regular en el sentido de diferenciabilidad

curva da una estimación de cuanto influye X_m al log-odds relativo tal que $Y \geq g$. Dado que los residuos parciales permiten el estudio de las transformaciones en las variables predictoras (linealidad), al mismo tiempo que permiten el estudio sobre el *modelo de odds proporcional*(paralelismo), generalmente se prefieren las gráficas de residuos parciales en lugar de las llamadas *gráficas de residuos score*, para modelos ordinales.

2.2. Otros modelos ordinales de interés

A continuación se expondrán brevemente otra serie de modelos ordinales a tener en cuenta y que serán usados en el caso de que las condiciones y los datos sean idóneos para éstos.

2.2.1. Modelo de Ratios Continuados

Al contrario que el modelo de odds proporcional, el cual está basado en probabilidades acumuladas, el modelo de *Ratios Continuados (CR)* está basado en las probabilidades condicionadas. El modelo CR queda determinado para $Y = 0, ..., k$ de la siguiente forma:

$$\begin{aligned} P(Y = g \,|Y \geq g, X) &= \frac{1}{1 + exp\,[-(\theta_j + X_\gamma)]} \\ logit\,(Y = 0 \,|Y \geq 0, X) &= logit\,(Y = 0 \,|X) \\ &= \theta_0 + X_\gamma \\ logit\,(Y = 1 \,|Y \geq 1, X) &= \theta_1 + X_\gamma \\ &\cdots \\ &= \theta_{k-1} + X_\gamma \end{aligned}$$

donde γ es el vector de coeficientes de regresión

Se suele decir que el modelo CR ajusta las respuestas ordinales cuando los individuos tienen que *"pasar a través de"* una categoría para alcanzar la siguiente.[4]

2.2.2. Modelo Logit de Categorías Adyacentes

Una alternativa al modelo de odds acumulado es considerar los ratios de las probabilidades para sucesivas categorías, por ejemplo

$$\frac{\pi_1}{\pi_2}, \frac{\pi_2}{\pi_3}, ..., \frac{\pi_{G-1}}{\pi_G}$$

[4]El modelo CR es una versión discreta del *Modelo de Riesgos Proporcionales de Cox*

El modelo de categorías adyacentes es

$$log\left(\frac{\pi_g}{\pi_{g+1}}\right) = x_g^T \beta_g$$

Si se simplifica como

$$log\left(\frac{\pi_g}{\pi_{g+1}}\right) = \beta_{0g} + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1}$$

donde $x = x_1, ..., x_p$ es el vector formado por las variables explicativas que conforman el modelo y $\beta_1, ..., \beta_p$ los coeficientes asociados a cada una de ellas.

Se asume que los efectos de cada variable explicativa sobre cada par de categorías adyacentes es el mismo. Los coeficientes β_i son usualmente interpretados como odds-ratios usando la expresión previamente expuesta

$$OR = exp\,(\beta_i)$$

Capítulo 3

Modelado en R

En este capítulo se aborda la aplicación del modelo objeto del presente trabajo desde dos aspectos distintos:

- La aplicación de los métodos de inferencia a través de la librería "ordinal" de R.
- Una ilustración de su aplicación e interpretación de resultados.

3.1. Librería "ordinal"

Para el modelaje se usará la librería "ordinal" [4], su objetivo consiste en la implementación del modelo de odds proporcionales [2.1] y otros modelos ordinales.

Las funciones máss relevantes para la regresión ordinal son:

- *clm*

  Ajusta mediante modelos acumulados como el modelo de odds proporcional. El modelo permite varias funciones de enlace y umbrales estructurados que restringen los puntos de corte equidistante o simetricamente dispuestos alrededor de los umbrales centrales. Se usa una modificación del algoritmo de *Newton* para optimizar la función de máxima verosimilitud.

  Se pueden añadir estructuras determinadas para las distribución de los puntos de corte.

  A través de esta función podemos obtener los vectores de los coeficientes de regresión, los vectores de las constantes α, las probabilidades ya ajustadas, etc.

- *anova.clm*

  Comparación de modelos acumulados a través de contrastes de razón de verosimilitudes.

- *confint*

  Calcula intervalos de confianza a partir de la función de máxima verosimilitud de uno o más parámetros.

  Obtenemos una matriz donde sus columnas proporcionan los intervalos para cada parámetro. También tiene la opción de crear la gráfica en el caso de que queramos estudiarla en busca de problemas de linealidad.

- *convergence*

  Verifica la precisión de las estimaciones de los parámetros de los modelos acumulados. El número correcto decimales y número de dígitos significativos se da para las estimaciones de máxima verosimilitud de los parámetros en un modelo de enlace acumulado creado con la función *clm*.

  Se obtiene información sobre la convergencia, errores de estimación de los parámetros estimados.

- *predict.clm*

  Se obtienen los valores esperados para un modelo previamente creado con la orden *clm*. Requiere del modelo "clm" y de un "data frame" donde buscar las variables con las que predecir.

  Devuelve una lista con las predicciones o valores ajustados en el caso de que no se le introduzca un "data frame" con las variables explicativas sobre las que apoyarse. Ofrece la opción de que se muestren los intervalos de confianza y los errores.

3.2. Ejemplo en R

3.2.1. Datos wine

Vamos a considerar los datos de acidez de ciertos vinos y los procesos seguidos en Thompson [18], Fahrmeir y otros [7], Christensen [3] y basandonos en el ejemplo expuesto en Christensen [5] se ilustrará en R el modelaje. Los datos de *Randall (1989)*, disponibles en el paquete "ordinal" de R, están presentados en el cuadro 3.1, disponible como el conjunto de datos *wine* en el paquete *ordinal*.

```
library("ordinal")
data(wine)
head(wine)

##   response rating temp contact bottle judge
## 1       36      2 cold      no      1     1
## 2       48      3 cold      no      2     1
## 3       47      3 cold     yes      3     1
## 4       67      4 cold     yes      4     1
## 5       77      4 warm      no      5     1
## 6       60      4 warm      no      6     1

str(wine)

## 'data.frame': 72 obs. of  6 variables:
##  $ response: num  36 48 47 67 77 60 83 90 17 22 ...
##  $ rating  : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 2 3 3 4 4 4 5 5 1 2 ...
##  $ temp    : Factor w/ 2 levels "cold","warm": 1 1 1 1 2 2 2 2 1 1 ...
##  $ contact : Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 2 2 1 1 ...
##  $ bottle  : Factor w/ 8 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 1 2 ...
##  $ judge   : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
```

El conjunto de datos representa un experimento sobre ciertos factores que determinan la amargura del vino; donde 1 = "lo menos amargo" y 5 = "lo más amargo", dos factores de tratamiento (temperatura y contacto) cada una con dos niveles, la temperatura y el contacto entre el zumo y las pieles de las uvas cuando se extrae de ellas. Nueve jueces evaluaron cada vino de dos botellas para cada una de las cuatro condiciones de tratamiento, por lo tanto, hay 72 observaciones en total. La variable objetivo Y a estudiar será la variable "rating" $\in \{1, ..., 5\}$ que es una categorización de la variable "response" la cual califica la acidez de los vinos.

Vamos a ajustar el siguiente modelo acumulado para los datos *wine*:

$$\begin{gathered} logit(P(Y_i \leq j)) = \theta_j - \beta_1(temperatura_i) - \beta_2(contacto_i) \\ i = 1, ..., n \quad j = 1, ..., J-1 \end{gathered} \tag{3.1}$$

			Juez								
Temperatura	Contacto	Botella	1	2	3	4	5	6	7	8	9
frío	no	1	2	1	2	3	2	3	1	2	1
frío	no	2	3	2	3	2	3	2	1	2	2
frío	sí	3	3	1	3	3	4	3	2	2	3
frío	sí	4	4	3	2	2	3	2	2	3	2
templado	no	5	4	2	5	3	3	2	2	3	3
templado	no	6	4	3	5	2	3	4	3	3	2
templado	sí	7	5	5	4	5	3	5	2	3	4
templado	sí	8	5	4	4	3	3	4	3	4	4

Cuadro 3.1: Calificaciones de la amargura de algunos vinos blancos. Los datos han sido tomados de Randall (1989).

Este es un modelo para la probabilidad acumulada de que la calificación i-ésima caiga sobre la categoría j-ésima o superior, donde i indica cada observación ($n = 72$) y los índices $j = 1, ..., J$ reflejan la categoría respuesta ($J = 5$).
El parámetro θ_j es el punto de corte para el j-ésimo modelo acumulado, $logit(P(Y_i \leq j))$.

Este modelo es el *modelo de odds proporcional* descrito en la sección 2.1 de este trabajo.

Con el comando clm modelizamos a partir del modelo de odds proporcionales previamente desarrollado, ajustado mediante el método de máxima verosmilitud

```
fm1<-clm(rating ~ temp + contact, data=wine)
fm1

## formula: rating ~ temp + contact
## data:    wine
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  72   -86.49 184.98 6(0)  4.02e-12 2.7e+01
##
## Coefficients:
##   tempwarm contactyes
##      2.503      1.528
##
## Threshold coefficients:
##    1|2    2|3    3|4    4|5
## -1.344  1.251  3.467  5.006
```

Podemos obtener información adicional mediante *summary*

```
summary(fm1)

## formula: rating ~ temp + contact
## data:    wine
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  72   -86.49 184.98 6(0)  4.02e-12 2.7e+01
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## tempwarm     2.5031     0.5287   4.735 2.19e-06 ***
## contactyes   1.5278     0.4766   3.205  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 1|2  -1.3444     0.5171  -2.600
## 2|3   1.2508     0.4379   2.857
## 3|4   3.4669     0.5978   5.800
## 4|5   5.0064     0.7309   6.850
```

El primer resultado es la tabla de coeficientes con estimaciones de parámetros, errores estándar y los p-valores basados en el método de Wald. Las estimaciones mediante el método de máxima verosmilitud para los parámetros son:

$$\hat{\beta}_1 = 2{,}50,\ \hat{\beta}_2 = 1{,}53,\ \{\hat{\theta}_j\} = \{-1{,}34, 1{,}25, 3{,}47, 5{,}01\}$$

El número de iteraciones Newton-Raphson se da a continuación de *niter*. Tenemos *max.grad* que es el gradiente absoluto máximo de la función de log-verosimilitud con respecto a los parámetros. Un gradiente absoluto pequeño es una condición necesaria para la convergencia del modelo. El procedimiento iterativo indicará convergencia siempre que el gradiente absoluto máximo esté por debajo de

```
clm.control()$gradTol

## [1] 1e-06
```

Los coeficientes para la temperatura y el contacto son positivos, lo que indica que una temperatura más alta y más contacto aumenta la amargura del vino, es decir, la calificación en categorías superiores es más probable.

La odds ratio del suceso $Y \geq j$ es $exp(\beta_{tratamiento})$, por lo que la odds ratio de acidez que clasifica en la categoría j o superior a temperaturas templadas frente a las frías es

```
exp(coef(fm1)[5])

## tempwarm
## 12.22034
```

Los p-valores para los coeficientes de ubicación dados por el *summary* se basan en el estadístico de Wald. Las pruebas de razón de verosimilitud proporcionan pruebas más precisas. Estas se puede obtener con el método anova, por ejemplo, la prueba de ratios de máxima verosimilitud del contacto es

```
fm2 <- clm(rating ~ temp, data=wine)
anova(fm2,fm1)

## Likelihood ratio tests of cumulative link models:
##
##     formula:                link: threshold:
## fm2 rating ~ temp           logit flexible
## fm1 rating ~ temp + contact logit flexible
##
##     no.par    AIC  logLik LR.stat df Pr(>Chisq)
## fm2      5 194.03 -92.013
## fm1      6 184.98 -86.492  11.043  1  0.0008902 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

que en este caso produce un p-valor ligeramente menor. De manera equivalente, podemos usar *Drop1* para obtener pruebas de razón de verosimilitud de las variables explicativas mientras se controlan el resto variables:

```
drop1(fm1,test="Chi")

## Single term deletions
##
## Model:
## rating ~ temp + contact
##         Df    AIC    LRT  Pr(>Chi)
## <none>     184.98
## temp     1 209.91 26.928 2.112e-07 ***
## contact  1 194.03 11.043 0.0008902 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las pruebas de razón de verosimilitud de las variables explicativas ignorando el resto vienen dadas por el método *add1*:

```
fm0 <- clm(rating ~ 1,data=wine)
add1(fm0, scope = ~ temp + contact, test = "Chi")

## Single term additions
##
## Model:
## rating ~ 1
##         Df    AIC     LRT  Pr(>Chi)
## <none>     215.44
## temp     1 194.03 23.4113 1.308e-06 ***
## contact  1 209.91  7.5263   0.00608 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, estas últimas pruebas no son tan fuertes como las pruebas que controlan la otra variable.
Los intervalos de confianza se obtienen a partir del comando *confint* como sigue

```
confint(fm1)

##                2.5 %   97.5 %
## tempwarm   1.5097627 3.595225
## contactyes 0.6157925 2.492404
```

Estos se basan en los perfiles de la funcion de verosimilitud y generalmente son bastante precisos. Menos preciso, pero más simple y con intervalos de confianza simétricos basados en los errores estándar de los parámetros (también llamados, intervalos de confianza de Wald) se obtienen con:

```
confint(fm1, type="Wald")

##                 2.5 %    97.5 %
## 1|2        -2.3578848 -0.330882
## 2|3         0.3925794  2.109038
## 3|4         2.2952980  4.638476
## 4|5         3.5738541  6.438954
## tempwarm    1.4669081  3.539296
## contactyes  0.5936345  2.461961
```

Además del enlace logit, que es el predeterminado, el probit, el log-log, el log-log complementario y el de cauchit también se pusen usar para el ajuste. Por ejemplo, un modelo de odds proporcionales como el anterior pero con el enlace log-log sería:

```
fm.cll <- clm(rating ~ contact + temp, data=wine, link="cloglog")
```

En resumen, el primer modelo que creamos supone que los $\{\beta_j\}$ son constantes para todos los valores de las variables explicativas restantes, en nuestro ejemplo **temperatura** y **contacto**. Esto se conoce como la *suposición de odds proporcionales* o *suposición de pendientes iguales*.

3.2.2. Predicciones sobre wine

Los valores ajustados se obtienen con la orden *fitted(fm1)* y producen probabilidades ajustadas, es decir, la i-ésima probabilidad ajustada sería la probabilidad de que la i-ésima observación caiga sobre la categoría prevista. Las predicciones sobre qué categoría respuesta tiene mas probabilidad sobre las que caer la i-ésima observación son:

```
pfm1=predict(fm1,type="class")
pfm1$fit[1:15]

##  [1] 2 2 3 3 3 3 4 4 2 2 3 3 3 3 4
## Levels: 1 2 3 4 5
```

Digamos que solo queremos las predicciones para las cuatro posibles combinaciones entre temperatura y contacto, entonces sería:

```
combinaciones <- expand.grid(temp=levels(wine$temp),
contact=levels(wine$contact))
cbind(combinaciones, predict(fm1, newdata=combinaciones)$fit)

##   temp contact           1          2         3          4          5
## 1 cold      no 0.206790132 0.57064970 0.1922909 0.02361882 0.00665041
## 2 warm      no 0.020887709 0.20141572 0.5015755 0.20049402 0.07562701
## 3 cold     yes 0.053546010 0.37764614 0.4430599 0.09582084 0.02992711
## 4 warm     yes 0.004608274 0.05380128 0.3042099 0.36359581 0.27378469
```

Los errores estándar y los intervalos de confianza para las predicciones también se pueden calcular. Por ejemplo, para las primeras cuatro observaciones; las predicciones, los errores estándar y los intervalos de confianza al 95 % serían:

```
h=head(do.call("cbind", predict(fm1, se.fit=TRUE, interval=TRUE)))
h[1:4,]

##             fit     se.fit        lwr       upr
## [1,] 0.57064970 0.08683884 0.39887109 0.7269447
## [2,] 0.19229094 0.06388672 0.09609419 0.3477399
## [3,] 0.44305990 0.07939754 0.29746543 0.5991420
## [4,] 0.09582084 0.04257593 0.03887676 0.2173139
```

Capítulo 4

Ejemplos Reales

En esta sección se nombrarán y describirán brevemente algunas situaciones reales que fueron estudiadas con modelos ordinales.

4.1. Regresión Logística Ordinal Aplicada a la Identificación de Factores de Riesgo para Cáncer de Cuello Uterino

Este estudio fue realizado por *Evaristo Navarro*, *Aníbal Verbel*, *Delia Robles* y *Kennedy Hurtado*, en Barranquilla, Colombia y publicado el 25 de Agosto de 2014. (Navarro y otros [16])

Según los autores, la identificación de factores de riesgo para cáncer de cuello uterino es determinante a la hora de establecer diagnósticos efectivos que, en un momento dado, pueden ser determinantes para salvar vidas. Desde esta perspectiva se realizó este estudio sobre una muestra constituida por 105 pacientes. En el estudio fue considerada como variable objetivo el Cáncer de cuello uterino (CCU) y como variables explicativas los factores relacionados con la paridad (Edad (ED), Número de Hijos Nacidos Vivos (NHV), Número de Hijos Nacidos Muertos (NHM), tipo de parto (TP) y tipo de embarazo (TE)). También se incluyeron las características de la conducta sexual (Enfermedades venéreas (EV)). De manera general se observa que el riesgo de tener cáncer de cuello uterino es mayor cuando aumenta el número de hijos en partos por cesárea y se ha perdido un hijo.

4.2. Predicción del rendimiento en una asignatura empleando la regresión logística ordinal

Este estudio fue realizado por *Jobany J. Heredia*, *Aida G. Rodríguez* y *José A. Vilalta*, llevado por el departamento de Ingeniería Industrial, Facultad de Ingeniería Industrial del Instituto Superior Politécnico "José Antonio Echeverría" en La Habana, Cuba. Año de publicación, 2014. (Heredia y otros [10])

En las asignaturas donde el índice de fracaso es considerable, es fundamental que el profesor posea información relevante sobre sus alumnos para desarrollar un tratamiento específico para cada uno de ellos. En el trabajo se emplea la regresión logística ordinal para construir una ecuación que relacione la puntuación en la asignatura Modelos Probabilísticos de los Procesos (MPP), la cual se imparte en segundo año de la carrera de Ingeniería Industrial, con sus resultados en primer año. Con los datos de 274 estudiantes pertenecientes a dos cursos académicos distintos, se obtuvo como mejor modelo el que relaciona la evaluación en MPP con la media del alumno en las asignaturas de ciencia que recibe en primer año. Las probabilidades estimadas de este modelo se usaron como base para el desarrollo de un método que permitió mejorar la experiencia general y califcación de los alumnos del curso posterior.

4.3. El impacto de las relaciones interpersonales en la satisfacción laboral general

Este estudio fue realizado por *Rodrigo Yañez*, *Mallén Arenas* y *Miguel Ripoll* en la Universidad de Concepción, Concepción, Chile. Publicado en el año 2010. (Yañez y otros [20])

Los autores evaluaron el impacto de las relaciones interpersonales en el trabajo en la satisfacción laboral general. Primero, se construyó una escala para evaluar la satisfacción con las relaciones interpersonales en el trabajo y se aplicó a 209 trabajadores de un hospital. Un análisis factorial obtuvo una solución de dos factores y una adecuada consistencia interna de los ítems. Posteriormente, se aplicó la escala a 321 trabajadores de 7 centros de salud. Utilizando una regresión logística ordinal se obtuvo que las relaciones interpersonales en el trabajo tienen un impacto significativo en la satisfacción laboral general, especialmente, las relaciones con los jefes.

4.4. Las percepciones de riesgo de los consumidores en alimentos lácteos: aplicación de una regresión logística ordinal

Este estudio fue realizado por *Beatriz Lupín*, *Maria Victoria Lacaze* y *Elsa Mirta M. Rodríguez*, presentado en la XII Reunión Científica del Grupo Argentino de Biometría y I Encuentro Argentino-Chileno de Biometría, San Martín de los Andes, Argentina en Octubre de 2007. (Lupín y otros [13])

Según los autores de este trabajo, la creciente preocupación por la calidad de los alimentos manifestada por los consumidores se relaciona con la percepción de riesgos reales o potenciales asociados a los métodos y a las tecnologías empleados en la producción y en el procesamiento de los mismos. Dicha preocupación se ve influenciada por la información a la que acceden los consumidores, constituyendo un factor crítico de las

decisiones de compra. El objetivo de este trabajo fue aplicar un método estadístico de estimación que incorpora la naturaleza ordinal de la variable objetivo, a fin de analizar la incidencia de los factores asociados a las percepciones de riesgo de los consumidores en el caso de los alimentos lácteos. Las percepciones de riesgo para la salud derivadas del contenido de conservantes en los productos lácteos (variable objetivo con tres niveles de riesgo: alto, medio y bajo), interviniendo como variables explicativas las relacionadas con la información sobre la calidad de los alimentos, los sistemas de regulación vigentes y aspectos socio-demográficos de los consumidores. Los datos provienen de una encuesta realizada a 301 consumidores, captados en la Ciudad de Buenos Aires, durante abril de 2005.

Bibliografía

[1] A. Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] M. Alcaide. Modelo de regresión binominal negativa. 2015.

[3] R. Christensen. Analysis of ordinal data with cumulative link models—estimation with the ordinal package. *R-package version*, 13:9–13, 2011.

[4] R. Christensen. ordinal—regression models for ordinal data. *R package version*, 28:2015–06, 2015.

[5] R. Christensen. A tutorial on fitting cumulative link models with the ordinal package, 2015.

[6] A. J. Dobson and A. Barnett. *An introduction to generalized linear models*. CRC press, 2008.

[7] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

[8] J.W. Hardin, J.M. Hilbe, and J. Hilbe. *Generalized linear models and extensions*. Stata press, 2007.

[9] F. E. Harrell. Ordinal logistic regression. In *Regression modeling strategies*, pages 331–343. Springer, 2001.

[10] J. Heredia, A. Rodríguez, and J. Vilalta. Predicción del rendimiento en una asignatura empleando la regresión logística ordinal. *Estudios pedagógicos (Valdivia)*, 40(1):145–162, 2014.

[11] D. Kleinbaum and M. Klein. *Survival analysis*, volume 3. Springer, 2010.

[12] D. Kleinbaum, M. Klein, and ER. Pryor. Logistic regression: a self-learning text. 2002.

[13] B. Lupín, M. Lacaze, and E. Rodríguez. Las percepciones de riesgo de los consumidores en alimentos lácteos: Aplicación de una regresión logística ordinal. 2007.

[14] J.M. Marín. Regresión logística multinomial. `http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema5Cate.pdf`.

[15] P. McCullagh et al. Generalized linear models. *CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York*, 1989.

[16] E. Navarro, A. Verbel, D. Robles, and KR. Hurtado. Regresión logística ordinal aplicada a la identificación de factores de riesgo para cáncer de cuello uterino. *Ingeniare*, 9(17):87–105, 2014.

[17] J.A. Nelder and R. J. Baker. *Generalized linear models*. Wiley Online Library, 1972.

[18] L. A Thompson. S-plus (and r) manual to accompany agresti's categorical data analysis (2002). 2009.

[19] K. Vasconcellos, G. M. Cordeiro, and L. Barroso. Improved estimation for robust econometric regression models. *Brazilian Journal of Probability and Statistics*, pages 141–157, 2000.

[20] R. Yañez, M. Arenas, and M. Ripoll. El impacto de las relaciones interpersonales en la satisfacción laboral general. *Liberabit*, 16(2):193–202, 2010.