REGRESIÓN LOGÍSTICA CON PENALIDAD *RIDGE* APLICADA A DATOS DE EXPRESIÓN GENÉTICA

Por

Karen A. Prieto Castellanos

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

Diciembre, 2005

Aprobada por:

Tokuji Saito, Ph.D Miembro, Comité Graduado	Fecha
Edgardo Lorenzo, Ph.D Miembro, Comité Graduado	Fecha
Edgar Acuña, Ph.D Presidente, Comité Graduado	Fecha
Noel Artiles León, Ph.D Representante de Estudios Graduados	Fecha
Pedro Vásquez, D.Sc. Director del Departamento de Matemáticas	Fecha

Abstract of Disertación Presented to the Graduate Studies Office
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

LOGISTIC REGRESSION WITH RIDGE PENALTY APPLYING TO GENETIC EXPRESSION DATA

By

Karen A. Prieto Castellanos

December 2005

Chair: Edgar Acuña, Ph.D.
Major Department: Mathematics Department

Logistic regression analysis is used in classification to find out which group an individual belong from a predictor variables set. In classification sometimes we work with data sets with more variables than observations. This is the case of microarray data sets, where there are a relatively small number of observations, generally less than one hundred, and a huge number of features, usually thousands. The following problems in the parameters estimation of logistic regression may occur: overfitting, unstability and multicollineality. This work explores the logistic regression with *Ridge* penalty as an alternative to deal with that sort of data sets. It stabilizes the statistical problem, eliminates the numeric degeneracy due to multicollineality and gets low error rates of classification when it is compared with others methods.

Resumen de Disertación Presentado a Oficina de Estudios Graduados
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

REGRESIÓN LOGÍSTICA CON PENALIDAD *RIDGE* APLICADA A DATOS DE EXPRESIÓN GENÉTICA

Por

Karen A. Prieto Castellanos

Diciembre 2005

Consejero: Edgar Acuña, Ph.D.
Departamento: Departamento de Matemáticas

El análisis de regresión logística es utilizado en clasificación para determinar la clase a la que pertenece un individuo a partir de un conjunto de variables predictoras. En clasificación algunas veces se trabaja con bases de datos que tienen más variables que observaciones. Este es el caso de las bases de datos de microarreglos, que consisten de un número relativamente pequeño de observaciones, generalmente menos de 100, y una gran cantidad de variables, usualmente miles. Esto genera que al estimar los parámetros de la regresión logística se presenten problemas como: sobreajuste, inestabilidad y multicolinealidad. Este trabajo explora la regresión logística con penalidad "ridgeçomo una alternativa para tratar con este tipo de datos. Esta estabiliza el problema estadístico, elimina la degeneración numérica debida a la multicolinealidad y obtiene bajas tasas de error de clasificación en comparación con otros métodos.

TABLA DE CONTENIDO

página

LISTA DE TABLAS

Tabla página

LISTA DE FIGURAS

Figura página

LISTA DE ABREVIATURAS

KKT	Karush-Kuhn-Tucker.
MCO	Mínimos Cuadrados Ordinarios.
MCRI	Mínimos Cuadrados Reponderados Iterativamente.
MCP	Mínimos Cuadrados Ponderados.
MLG	Modelos Lineales Generalizados.
RLP	Regresión Logística Penalizada
SMO	Algoritmo de Optimización Mínima Secuencial.

Capítulo 1
INTRODUCCIÓN

El análisis de regresión logística puede ser usado para desarrollar modelos predictivos en casos en que la variable respuesta es de tipo categórico. Es utilizado en clasificación para determinar la clase a la que pertenece un individuo a partir de un conjunto de variables predictoras. En clasificación algunas veces se trabaja con bases de datos que tienen más variables que observaciones. Este es el caso de las bases de datos de microarreglos, que consisten de un número n relativamente pequeño de observaciones, generalmente menos de cien, y una gran cantidad p de variables, usualmente miles. La regresión logística clásica no trabaja bien con ese tipo de datos. Un primer problema es que al ser $n < p$ hay más parámetros desconocidos que ecuaciones, por lo tanto puede haber infinitas soluciones. Otro problema consiste en el sobre ajuste, una ecuación que ajusta muy bien a los datos pero tiene demasiados parámetros no dará una buena predicción de nuevos datos debido a que el ruido del primer conjunto de datos no se ha filtrado aún. El tercer problema es la multicolinealidad, en microarreglos es muy probable que haya genes con patrones de expresión casi idénticos. Cuando la multicolinealidad está presente, algunos de los coeficientes de regresión (o todos) son muy grandes e inestables, un pequeño cambio en los datos puede producir coeficientes estimados muy diferentes.

Uno de los métodos para tratar con estos inconvenientes es la selección de conjuntos de predictoras, la cual proporciona modelos fácilmente interpretables, pero puede ser extremadamente variable debido a que es un proceso discreto. Pequeños

cambios en los datos pueden resultar en modelos muy diferentes y esto reduce la exactitud de la predicción [27].

Otro método disponible es aplicar una penalidad *Ridge* al logaritmo de la función de verosimilitud. Esta restricción permite que solo los coeficientes de regresión que son realmente relevantes sean grandes. La penalidad *Ridge* estabiliza el problema estadístico y resuelve el problema de multicolinealidad en las predictoras. Este trabajo pretende evaluar el desempeño de la regresión logística con penalidad *Ridge*, en la clasificación supervisada de ocho bases de datos de microarreglos disponibles en varios sitios de la Internet.

En el primer capítulo se presenta el marco teórico de la regresión logística, regresión logística multinomial y los métodos de regresión lineal que involucran penalidad: *Ridge* o lasso. En el segundo capítulo se habla de la existencia y unicidad de los estimadores en regresión logística y se describen dos métodos para encontrar los estimadores cuando éstos existen. El tercer capítulo explica la regresión logística penalizada para el caso binomial y multinomial y para la estimación de parámetros expone el algoritmo de optimización mínima secuencial (SMO, por sus siglas en inglés) para el caso multinomial. Por último, en el cuarto capítulo se presentan los resultados obtenidos al aplicar regresión logística con penalidad *Ridge* en las ocho bases de datos y se compara con otros métodos.

Capítulo 2
REVISIÓN DE LITERATURA

2.1. Regresión logística

Suponga que se hace un estudio en el cual se desea predecir la probabilidad de que una mujer desarrolle cáncer de cuello uterino, y para ello se cuenta con ciertas variables como: edad, número de abortos, existencia de un tumor en el cuello uterino, tamaño del tumor, etc. En este tipo de estudios la variable respuesta es binaria, toma el valor 1 si el evento ocurre, o 0 si el evento no ocurre. Ajustar un modelo de regresión lineal tiene los siguientes inconvenientes:

- Debido a que la variable respuesta Y toma solo dos valores (0 y 1), los errores no son normalmente distribuidos, así se viola un supuesto necesario para hacer inferencia.
- Los errores son heterocedásticos: $var(error) = \pi(x)(1 - \pi(x))$, donde π es la probabilidad del evento. Como π depende de la matriz de datos X, se viola otro de los supuestos de la regresión lineal.
- Los valores predichos pueden ser mayores a 1 o menores a 0, lo cual puede ser un problema si los valores son usados para análisis posteriores.

Como una solución a estos inconvenientes se utilizan los modelos lineales generalizados (MLG), los cuales extienden los modelos de regresión ordinarios para abarcar distribuciones de respuesta no normales y modelar funciones de la media. Los MLG constan de tres componentes [2]:

1. Un *componente aleatorio* que identifica la variable respuesta Y y su distribución de probabilidad, perteneciente a la familia exponencial:

$$f(y_i, \theta_i) = a(\theta_i)b(y_i)exp[y_iQ(\theta_i)]$$

donde θ_i y y_i denotan: el parámetro del modelo y la variable respuesta para la i-ésima observación respectivamente y a, b y Q son funciones de valor real.

2. Un *componente sistemático* que especifica las variables explicativas usadas en una función lineal. Esto es, un vector $(\eta_1, \ldots, \eta_n)$ de variables explicativas a través de un modelo lineal. Dado x_{ij} el valor de la predictora j para el sujeto i entonces

$$\eta_i = \sum_{j=0}^{p} \beta_j x_{ij} \tag{2.1}$$

Con $x_{i0} = 1$ para todo i.

3. Una *función de enlace* que conecta los componentes aleatorio y sistemático. Dada $\mu_i = E(Y_i), i = 1, \ldots, n$. El modelo enlaza a μ_i con η_i por $\eta_i = g(\mu_i)$ donde la función de enlace g es monótona y diferenciable. Así, g enlaza $E(Y_i)$ a las variables explicativas a través de la fórmula

$$g(\mu_i) = \eta_i$$

En el modelo de regresión clásico el objetivo es estimar la media condicional de Y, así, la función de enlace es la función identidad $g(\mu_i) = \mu_i$ y el componente sistemático es una función lineal de las variables x. Esto es:

$$\mu_i = \sum_{j=0}^{p} \beta_j x_{ij}$$

En el modelo de regresión logística se tienen n variables aleatorias binomiales $y_i, i = 1, \ldots, n$. Para cada y_i se conoce la cantidad de ensayos m_i y un vector de p predictoras asociado $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. La probabilidad de éxito $\pi(\mathbf{x}_i)$ depende de $\mathbf{x}_i$, entonces $y_i|\mathbf{x}_i \sim Bin(m_i, \pi(\mathbf{x}_i)), i = 1, \ldots, n$. La media y varianza para la variable

aleatoria y_i están dadas por[6]:

$$\mu_i = E(y_i|\mathbf{x}_i) = m_i\pi(\mathbf{x}_i) \tag{2.2}$$

$$V_i = V(y_i|\mathbf{x}_i) = m_i\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \tag{2.3}$$

En la mayoría de los casos $m_i = 1$, es decir, y_i es una variable Bernoulli que toma el valor 1 si la característica de interés está presente y 0 si no. Para estimar $\pi(\mathbf{x}_i)$ se utiliza una función kernel M aplicada a η_i. La función kernel para la regresión logística debe estar entre 0 y 1. La más frecuentemente utilizada es la función logística dada por:

$$\pi(\mathbf{x}_i) = M(\eta_i) = \frac{exp(\eta_i)}{1+exp(\eta_i)} = \frac{1}{1+exp(-\eta_i)} \tag{2.4}$$

Con η_i dado por (2.1). Al resolver la ecuación (2.4) para η_i se obtiene:

$$ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \eta_i \tag{2.5}$$

Entonces, para el modelo de regresión logística el componente sistemático está dado por: $\eta_i = \sum_j \beta_j x_{ij}$ y la función de enlace (también denominada *logit*) que hace al modelo lineal es:

$$g\left(\pi(\mathbf{x}_i)\right) = ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right). \tag{2.6}$$

La razón $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$ es llamada *razón de apuestas* y representa la oportunidad de éxito. Por ejemplo, si la probabilidad de éxito es .25, la razón de apuestas es .25/(1-.25)=1/3, es decir, un éxito por cada tres fallas.

En el modelo de regresión logística simple la probabilidad $\pi(x)$ está dada por:

$$\pi(x) = \frac{1}{1+exp(-\eta)} = \frac{1}{1+exp(-(\beta_0+\beta_1 x))}.$$

Como se observa en la Figura 2–1, cuando $x \to \infty$, $\pi(x) \to 0$ si $\beta_1 < 0$ y $\pi(x) \to 1$ si $\beta_1 > 0$. Además, a medida que $|\beta_1|$ crece, la tasa de incremento o descenso de $\pi(x)$

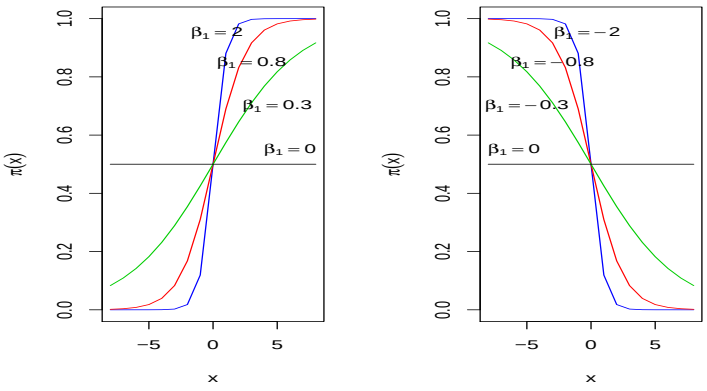

Figura 2–1: Comportamiento de la función logística

aumenta, y a medida que β_1 se acerca a 0, la curva se va aplanando hasta convertirse en una línea horizontal (cuando $\beta_1 = 0$). Exponenciando ambos lados de (2.5) se obtiene:

$$\frac{\pi(x)}{1-\pi(x)} = exp(\eta) = exp(\beta_0 + \beta_1 x) = exp(\beta_0)exp(\beta_1 x)$$

Esto es útil para la interpretación de β_1, pues de aquí se ve que la razón de apuestas se incrementa multiplicativamente en e^{β_1} por cada unidad de incremento en x. En el caso de la regresión logística múltiple:

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

La razón de apuestas se incrementa multiplicativamente en e^{β_i} por cada unidad de incremento en x_i a niveles fijos de las otras x_j.

2.2. Regresión logística multinomial

Aunque el modelo de regresión logística es más frecuentemente utilizado cuando la variable respuesta es binomial, también puede extenderse a variables respuesta multinomiales. Suponga que la variable respuesta Y tiene J categorías: $0, 1, 2, \ldots, J-1$. De manera similar al caso binomial, en el cual la variable respuesta es parametrizada en términos del logit de $Y = 1$ vs $Y = 0$, en el caso multinomial se tienen $J-1$ funciones logit comparando: $Y = 1$ vs $Y = 0$, $Y = 2$ vs $Y = 0$, $\ldots$, $Y = J-1$ vs $Y = 0$; los demás logit pueden derivarse de combinaciones de los anteriores, por

ejemplo el logit de $Y = 2$ vs $Y = 1$ se obtiene de la diferencia entre el logit de $Y = 2$ vs $Y = 0$ y el logit de $Y = 1$ vs $Y = 0$. Dado $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ el vector de p predictoras asociado a y_i, y dado $\pi_{ji} = \pi_j(\mathbf{x}_i) = p(Y = j|\mathbf{x}_i)$ para $j = 0, 1, \ldots, J-1$, entonces las $J-1$ funciones logit son:

$$
\begin{aligned}
g_1(\mathbf{x}_i) &= ln\left(\frac{\pi_{1i}}{\pi_{0i}}\right) = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \cdots + \beta_{1p}x_{ip} \\
g_2(\mathbf{x}_i) &= ln\left(\frac{\pi_{2i}}{\pi_{0i}}\right) = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \cdots + \beta_{2p}x_{ip} \\
&\vdots \\
g_{J-1}(\mathbf{x}_i) &= ln\left(\frac{\pi_{(J-1)i}}{\pi_{0i}}\right) = \beta_{(J-1)0} + \beta_{(J-1)1}x_{i1} + \beta_{(J-1)2}x_{i2} + \cdots + \beta_{(J-1)p}x_{ip}
\end{aligned}
$$

Ahora, sabemos que $\pi_{0i} + \pi_{1i} + \cdots + \pi_{(J-1)i} = 1$, entonces,

$$1 + \frac{\pi_{1i}}{\pi_{0i}} + \cdots + \frac{\pi_{(J-1)i}}{\pi_{0i}} = \frac{1}{\pi_{0i}}$$

Lo que es equivalente a:

$$1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)} = \frac{1}{\pi_{0i}}. \tag{2.7}$$

Despejando para π_{0i} en (2.7) se obtiene:

$$\pi_{0i} = \frac{1}{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}}$$

Con el fin de obtener una expresión para π_{1i}, multiplicamos por π_{1i} en ambos lados de (2.7):

$$\pi_{1i}\{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}\} = \frac{\pi_{1i}}{\pi_{0i}}$$

De lo cual se obtiene:

$$\pi_{1i} = \frac{e^{g_1(\mathbf{x}_i)}}{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}}$$

De manera similar, multiplicando por π_{ji} en ambos lados de (2.7) resulta:

$$\pi_{ji} = \frac{e^{g_j(\mathbf{x}_i)}}{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}}, j = 1, \ldots, J-1.$$

2.3. Regresión *Ridge*

Cuando se hace un análisis estadístico por regresión lineal, se cuenta con datos $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, donde $\mathbf{x}_i$ es el vector de variables predictoras y y_i la variable respuesta para la i-ésima observación. Usualmente se obtienen los coeficientes estimados mediante mínimos cuadrados ordinarios (MCO), los cuales minimizan la suma de cuadrados de los errores. Los estimadores MCO tienen dos inconvenientes:

- Son insesgados pero pueden tener varianza grande cuando hay problemas de multicolinealidad lo cual afecta la exactitud de las estimaciones.
- Son difíciles de interpretar cuando existen muchas predictoras. En este caso sería mejor determinar un conjunto más pequeño con las variables que tienen mayor efecto sobre la respuesta.

Uno de los métodos para tratar con estos inconvenientes es la selección de subconjuntos de predictoras, la cual proporciona modelos fácilmente interpretables, pero puede ser extremadamente variable: pequeños cambios en los datos pueden resultar en modelos muy diferentes y esto reduce la exactitud de la predicción [27].

Otro método disponible es la regresión *Ridge*, cuyo objetivo es encontrar un estimador $\widetilde{\beta}$ que aunque sea sesgado sea más corto que el estimador mínimo cuadrático $\widehat{\beta}$, es decir, $\widetilde{\beta}'\widetilde{\beta} < \widehat{\beta}'\widehat{\beta}$. El estimador *Ridge* es obtenido tratando de encoger el estimador mínimo cuadrático hacia el origen.

La regresión *Ridge* es uno de los métodos utilizados para tratar con el problema de multicolinealidad en las predictoras. Es un proceso continuo que encoge los coeficientes y así, es más estable que la selección de subconjuntos, sin embargo, no lleva ningún coeficiente a cero, y así, no da modelos fácilmente interpretables [27].

2.3.1. Estimación de parámetros

La función a minimizar en regresión *Ridge* está dada por:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2, \text{ sujeto a } \sum_j \beta_j^2 \leq t, \text{ con } N < p$$

o, equivalente

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2$$

Así, en términos matriciales la función a minimizar es

$$\begin{aligned} Q(\alpha, \beta) &= (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \\ &= (Y' - \beta'X')(Y - X\beta) + \lambda\beta'\beta \\ &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta + \lambda\beta'\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta + \lambda\beta'\beta \\ &= Y'Y - 2\beta'X'Y + \beta'(X'X + \lambda I)\beta \end{aligned}$$

Ahora, derivando con respecto a β e igualando a cero se obtiene:

$$\begin{aligned} \frac{dQ}{\beta} &= -2X'Y + 2(X'X + \lambda I)\beta \\ -2X'Y + 2(X'X + \lambda I)\widetilde{\beta} &= 0 \\ (X'X + \lambda I)\widetilde{\beta} &= X'Y \end{aligned}$$

$$\widetilde{\beta} = (X'X + \lambda I)^{-1}X'Y \tag{2.8}$$

El estimador (2.8) fue propuesto por Hoerl y Kennard en 1970 [13] para valores positivos del escalar λ. El parámetro de encogimiento λ (por lo general, $0 < \lambda < 1$) debe ser estimado de los datos tomados. Si $\lambda = 0$ se obtiene el estimador mínimo cuadrático y a medida que λ aumenta el estimador se aleja del estimador mínimo cuadrático y se hace más sesgado. La regresión *Ridge* aumenta el sesgo para ganar reducción en la varianza. Para valores pequeños de λ el sesgo puede ser muy pequeño mientras que la reducción en varianza puede ser bastante sustancial.

Para la elección de λ se hace un gráfico de $\widetilde{\beta}$ para varios valores de λ, este gráfico es llamado "traza *Ridge* ". La "traza *Ridge* "permite encontrar algunos coeficientes de regresión estables, es decir, que no exhiben grandes cambios, y otros que decrecen rápidamente o cambian de signo, las variables correspondientes a esos coeficientes serán del eliminadas del modelo. Desde el punto de vista computacional es recomendable trabajar con variables estandarizadas.

Para el caso ortonormal, $X'X = I$, así soluciones *Ridge* de (2.8) son:

$$\widetilde{\beta} = (I + \lambda I)^{-1} X'Y = \frac{1}{1+\lambda}\widehat{\beta}_0 \tag{2.9}$$

Con $\hat{\beta}_0$ el estimador obtenido por mínimos cuadrados ordinarios. De aquí puede observarse que para el caso ortonormal la regresión *Ridge* escala los coeficientes $\hat{\beta}_0$ por un factor constante. Entre más grande es λ más pequeños son los coeficientes $\tilde{\beta}$.

2.3.2. Geometría de los estimadores *Ridge*

Como se mencionó anteriormente, la función a minimizar en regresión *Ridge* es:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2, \text{ sujeto a } \sum_j \beta_j^2 \leq t$$

Matricialmente, la primera expresión es equivalente a:

$$\begin{aligned}(Y - X\beta)'(Y - X\beta) &= (Y - X\beta - X\hat{\beta}_0 + X\hat{\beta}_0)'(Y - X\beta - X\hat{\beta}_0 + X\hat{\beta}_0) \\ &= (Y - X\hat{\beta}_0 - X(\beta - \hat{\beta}_0))'(Y - X\hat{\beta}_0 - X(\beta - \hat{\beta}_0)) \\ &= (Y - HY - X(\beta - \hat{\beta}_0))'(Y - HY - X(\beta - \hat{\beta}_0)) \\ &= \Big((I - H)Y - X(\beta - \hat{\beta}_0)\Big)' \Big((I - H)Y - X(\beta - \hat{\beta}_0)\Big) \\ &= \Big(Y'(I - H) - (\beta - \hat{\beta}_0)'X'\Big) \Big((I - H)Y - X(\beta - \hat{\beta}_0)\Big)\end{aligned}$$

Pero,

$$\begin{aligned}Y'(I - H)X(\beta - \hat{\beta}_0) &= Y'(I - X(X'X)^{-1}X')X(\beta - \hat{\beta}_0) \\ &= Y'(X - X)(\beta - \hat{\beta}_0) = 0\end{aligned}$$

$$\begin{aligned}(\beta - \hat{\beta}_0)'X'(I - H)Y &= (\beta - \hat{\beta}_0)'X'(I - X(X'X)^{-1}X')Y \\ &= (\beta - \hat{\beta}_0)'(X' - X')Y = 0\end{aligned}$$

Por consiguiente,

$$(Y - X\beta)'(Y - X\beta) = (\beta - \hat{\beta}_0)'X'X(\beta - \hat{\beta}_0) + Y'(I - H)Y$$

Esta es una función cuadrática de los β. Para el caso $p = 2$ variables predictoras, las curvas de nivel $(\beta - \hat{\beta}_0)'X'X(\beta - \hat{\beta}_0) + Y'(I - H)Y = c$ determinan una familia de elipses centradas en el estimador obtenido por mínimos cuadrados ordinarios $\hat{\beta}_0$.

Se busca obtener el valor de β que minimiza esta función sujeto a la restricción $\sum_j \beta_j^2 \leq t$. Como se muestra en la Figura 2–2, esto corresponde a la mínima elipse

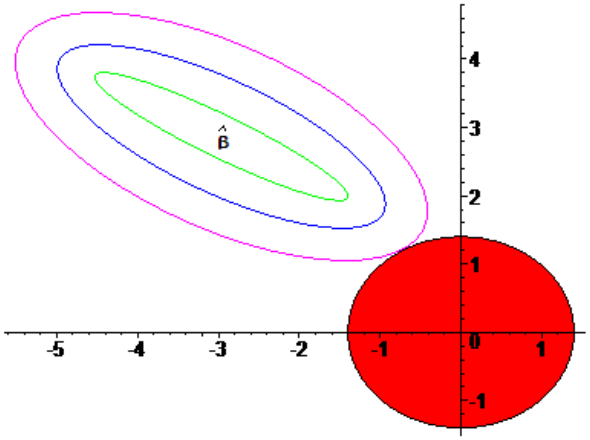

Figura 2–2: Interpretación gráfica de la regresión *Ridge*

que toca la región de restricción, determinada por el círculo sombreado.

2.3.3. Error estándar de los estimadores *Ridge*

Considere la descomposición en valores singulares de X:

$$X_{n\times k} = U_{n\times k} D_{k\times k} V_{k\times k} \tag{2.10}$$

con $n > k$, donde las columnas de U y V son ortogonales y normalizadas, es decir, $U'U = I$, $V'V = I$ y D es la matriz diagonal de los valores singulares[19]. Entonces

$$X'X = (UDV)'(UDV) = (V'DU')(UDV) = V'D^2V \tag{2.11}$$

Así,

$$\begin{aligned}
\widetilde{\beta} &= (X'X + \lambda I)^{-1}X'y = (V'D^2V + \lambda I)^{-1}V'DU'y \\
&= (V'D^2V + \lambda V'V)^{-1}V'DU'y = [V'(D^2 + \lambda I)V]^{-1}V'DU'y \\
&= V^{-1}(D^2 + \lambda I)^{-1}(V')^{-1}V'DU'y = V^{-1}(D^2 + \lambda I)^{-1}DU'y \\
&= V'(D^2 + \lambda I)^{-1}DU'y
\end{aligned}$$

Para una matriz dada A y un vector z se tiene: $Var(Az) = AVar(z)A'$. Utilizando esto, la matriz de covarianza de $\widetilde{\beta}$ está dada por:

$$\begin{aligned} V(\widetilde{\beta}) &= Var[V'(D^2+\lambda I)^{-1}DU'y] \\ &= V'(D^2+\lambda I)^{-1}DU'Var(y)UD(D^2+\lambda I)^{-1}V \\ &= V'(D^2+\lambda I)^{-1}DU'UD(D^2+\lambda I)^{-1}V\sigma^2 \\ &= V'(D^2+\lambda I)^{-1}D^2(D^2+\lambda I)^{-1}V\sigma^2 \\ &= V'(D^2+\lambda I)^{-2}D^2V\sigma^2 \\ &= \sigma^2V'diag\{\frac{d_i^2}{(d_i^2+\lambda)^2}\}V \end{aligned}$$

donde $d_i, i = 1, 2, \ldots, k$ son los valores singulares. De aquí puede observarse que los valores singulares de X más pequeños dominan la varianza, pero sumar λ reduce su contribución.

2.4. Encogimiento absoluto mínimo y selección de operador *(lasso)*

Cuando se desea analizar un conjunto de datos por medio de una regresión lineal usualmente se utiliza el método de mínimos cuadrados ordinarios (MCO) para la estimación de parámetros. En la Sección (2.3) se mencionaron algunos inconvenientes del MCO y se presentó la regresión *Ridge* como un método alternativo. En esta Sección se explica otro método denominado: least absolute shrinkage and selection operator(lasso, por sus siglas en inglés), el cual, al igual que la regresión *Ridge*, es utilizado para tratar con los inconvenientes del MCO. Esta aproximación impone un límite sobre la suma de valores absolutos de los coeficientes de regresión y va bajando este límite hasta que alguna clase de valor óptimo es alcanzado.

Suponga que para un conjunto de datos en particular, se tienen los coeficientes de regresión estimados por mínimos cuadrados, algunos con valores entre cero y uno y otros con valores alrededor de 100. Un cambio de 1 en el segundo conjunto de

coeficientes tendrá poco efecto sobre el ajuste, mientras que un cambio de la misma magnitud en el primer conjunto tendrá un gran efecto. Lasso, realiza un ajuste de mínimos cuadrados sujeto a una restricción sobre la suma de los valores absolutos de los coeficientes de regresión, así, los coeficientes más grandes son encogidos sustancialmente, mientras que los más pequeños quedarán casi iguales. Por la naturaleza de dicha restricción esta tiende a producir algunos coeficientes que son exactamente cero y así da modelos interpretables. Lasso disfruta de las propiedades favorables de la selección de subconjuntos y *Ridge*. Produce modelos interpretables como la selección de subconjuntos y exhibe la estabilidad de la regresión *Ridge*.

2.4.1. Estimación de parámetros - caso ortonormal

Dada $X_{n\times p}$ la matriz de variables predictoras, la función a minimizar en lasso está dada por:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2, \text{ sujeto } \sum_j |\beta_j| \leq t$$

Utilizando multiplicadores de Langrange se tiene:

$$L(\beta_j, \lambda) = \sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \left(\sum_j |\beta_j| - t\right)$$

O equivalentemente,

$$\begin{aligned} L(\beta, \lambda) &= (Y - X\beta)'(Y - X\beta) + \lambda(1'_{1\times p}|\beta| - t) \\ &= (Y' - \beta' X')(Y - X\beta) + \lambda(1'_{1\times p}|\beta| - t) \\ &= Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta + \lambda(1'_{1\times p}|\beta| - t) \end{aligned}$$

En caso ortonormal, $X'X = I$, así, $L(\beta, \lambda)$ se convierte en:

$$L(\beta, \lambda) = Y'Y - 2\beta' X'Y + \beta'\beta + \lambda(1'_{1\times p}|\beta| - t)$$

$$\frac{\partial L}{\partial \beta} = -2X'Y + 2\beta + \lambda \text{ signo}(\beta)$$
$$= -2\hat{\beta^0} + 2\beta + \lambda \text{ signo}(\hat{\beta}_0)$$

Donde $\hat{\beta}_0 = X'Y$ es el estimador obtenido por (MCO). Igualando a cero se obtiene:

$$2\hat{\beta} = 2\hat{\beta}_0 - \lambda \text{signo}(\hat{\beta}_0)$$
$$\hat{\beta} = \hat{\beta}_0 - \frac{\lambda}{2} \text{ signo}(\hat{\beta}_0)$$

$$\hat{\beta} = \text{signo}(\hat{\beta}_0)\left(|\hat{\beta}_0| - \frac{\lambda}{2}\right) \tag{2.12}$$

Ahora, derivando con respecto a λ e igualando a cero se obtiene:

$$\frac{\partial L}{\partial \lambda} = 1'_{1\times p}|\beta| - t = 0$$

O equivalentemente $\sum_j |\hat{\beta}_j| = t$. Donde $t \geq 0$ es un parámetro que varía.

De (2.9) y (2.12)puede notarse para el caso ortonormal, que la regresión *Ridge* escala los coeficientes por un factor constante, mientras que lasso los traslada hacia cero por un factor constante.

2.4.2. Geometría de los estimadores lasso

La función a minimizar en lasso es:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2, \text{ sujeto a } \sum_j |\beta_j| \leq t$$

En la Sección (2.3.2) se mostró que la primera expresión es equivalente a:

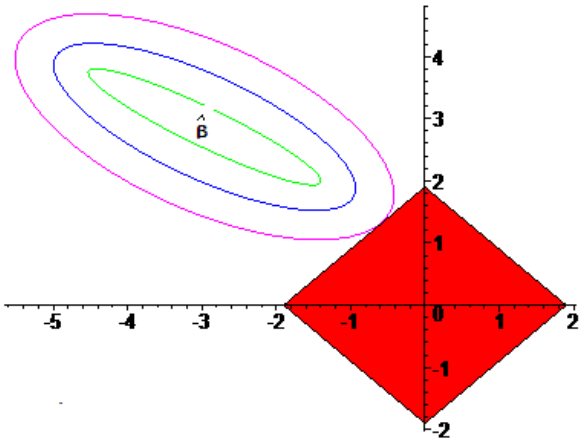

Figura 2–3: Interpretación gráfica de lasso

$$(Y - X\beta)'(Y - X\beta) = (\beta - \hat{\beta}_0)'X'X(\beta - \hat{\beta}_0) + Y'(I - H)Y$$

Esta es una función cuadrática de los β cuyas curvas de nivel $(\beta-\hat{\beta}_0)'X'X(\beta-\hat{\beta}_0)+Y'(I-H)Y = c$, para el caso $p = 2$ variables predictoras, determinan una familia de elipses centradas en los estimativos obtenidos por mínimos cuadrados ordinarios $\hat{\beta}_0$.

Se busca obtener el valor de β que minimiza esta función sujeto a la restricción $\sum_j |\beta_j| \leq t$. Como se muestra en la Figura 2–3, esto corresponde a la mínima elipse que toca la región de restricción, determinada por el rombo sombreado.

En la Figura 2–3 se puede observar que el punto en el cual se la elipse toca la región de restricción puede coincidir con una esquina, en este caso la solución lasso será cero para alguno de los coeficientes. Esto no ocurre en *Ridge*, en donde la región de restricción es un circulo (ver Figura 2–2).

2.4.3. Error estándar de los estimadores lasso

Debido a que el estimador lasso es una función no lineal y no diferenciable de la variable respuesta, incluso para un valor fijo t, es difícil obtener un estimador exacto de su error estándar. Una aproximación propuesta por Tibshirani [27], es escribir la penalidad lasso $\sum |\beta_j|$ como $\sum \beta_j^2/|\beta_j|$. Así, el estimador $\hat{\beta}$ puede ser aproximado con una regresión *Ridge* $\beta^* = (X'X + \lambda W^-)^{-1}X'y$, donde W es una matriz diagonal con elementos $|\beta_j|$, W^- es la inversa generalizada de W y λ es escogido de tal forma que $\sum |\beta_j|^* = t$.

La matriz de covarianzas puede ser aproximada por:

$$(X'X + \lambda W^-)^{-1}X'X(X'X + \lambda X^-)^{-1}\hat{\sigma}^2 \tag{2.13}$$

Donde, $\hat{\sigma}^2$ es una estimación de la varianza del error. Esta aproximación sugiere un algoritmo iterativo de regresión *Ridge* para calcular el estimador lasso, lo cual resulta bastante ineficiente.

Otra aproximación es utilizar bootstrap, t puede ser fijo u optimizado para cada muestra bootstrap. Fijar t es equivalente a seleccionar el mejor subconjunto y usar el error estándar de ese subconjunto.

2.4.4. Algoritmo lasso

Existen varios algoritmos para calcular las estimaciones por lasso. El algoritmo utilizado por la librería lasso2 del paquete estadístico R [18] trata a lasso como un problema de programación convexa y deriva su dual para encontrar la solución óptima. Este algoritmo fue propuesto por Osborne en 1999 [21].

2.4.5. Estimación del parámetro t

Como se mencionó en la Sección 2.4.3, el estimador $\hat{\beta}$ puede ser aproximado con una regresión *Ridge*:

$$\beta^* = (X'X + \lambda W^-)^{-1} X'y$$

Donde W es una matriz diagonal con elementos $|\beta_j|$, W^- es la inversa generalizada de W y λ es escogido de tal forma que $\sum |\beta_j|^* = t$.

Así, el número de parámetros en el modelo restringido lasso puede ser aproximado por:

$$p(t) = tr\{X(X'X + \lambda W^-)^{-1} X'\}$$

Dado SSE, la suma de cuadrados de los errores del modelo restringido. El error por validación cruzada generalizada (VCG) está dado por:

$$VCG(t) = \frac{1}{N} \frac{SSE}{\{1 - p(t)/N\}^2}$$

Se escoge el valor de t que hace que VCG sea mínimo.

2.4.6. Ejemplo - Cáncer de próstata

El conjunto de datos "prostate" corresponde a un estudio realizado por Stamey et al [26] que pretende examinar la correlación entre el nivel de un antígeno específico de próstata y algunas medidas clínicas, en hombres que reciben una prostatectomía radical. Los datos están disponibles en la librería lasso2 del programa estadístico R.

La base de datos consta de 97 observaciones y las siguientes 8 variables:

lcavol: logaritmo del volumen del cáncer

lweight: logaritmo del ancho de la próstata

age: edad del paciente

lbph: cantidad inicial de hiperplasia prostática

svi: invasión vesicula seminal

lcp: logaritmo penetración capsular

gleason: puntaje Gleason

pgg45: porcentaje de puntajes Gleason 4 o 5

Para llevar a cabo la rutina lasso, la matriz de variables predictoras fue primero centrada y escalada. Luego se estimó el valor óptimo 't'de la restricción, por medio de validación cruzada generalizada (VCG).

Programa:

```
> library(lasso2)
> data(Prostate)
> datos = Prostate
> p = dim(datos)[2]
> nobs = dim(datos)[1]
> d.mean = apply(datos, 2, mean)
> datos2 = sweep(datos, 2, d.mean, "-")
> d.std = apply(datos2, 2, var)
> datos2 = sweep(datos2, 2, sqrt(d.std), "/")
> datos2[, p] = datos[, p]
> nombres = colnames(datos)
> f1 = as.formula(paste(nombres[p], ".-1", sep = "~"))
> b = l1ce(f1, datos2, bound = seq(0, 1, length = 10), sweep.out = NULL)
> g = gcv(b, type = c("Tibshirani"), gen.inverse.diag = 1e+11)
> infopt = g[which(g[, 4] == min(g[, 4])), ]
> opt = infopt[1]
```

```
> f2 = as.formula(paste(nombres[p], ".", sep = "~"))
> plot(g[, 1], g[, 4], type = "l", xlab = "Restriccion relativa",
+     ylab = "VCG")
> datos2 = as.data.frame(datos2)
> res = l1ce(f2, datos2, bound = (1:40)/40)
> plres <- plot(res, plot = F)
> matplot(plres$bounds[, "rel.bound"], plres$mat.of.coef[, -1],
+     type = "l", xlim = c(0, 1.1), xlab = "Restriccion relativa",
+     ylab = "Coeficientes estimados")
> text(cbind(1.03, coef(res[40])[-1]), labels(res), adj = 0, cex = 0.8)
> abline(v = opt, col = "red")
```

En la Figura 2–4 se muestra el valor del estadístico VCG para diferentes valores de t. Como se observa en el gráfico, el valor relativo de la restricción en el cual se minimiza el VCG es .44. Este valor corresponde a $t/\sum|\hat{b}_j^0|$, donde $\hat{b}_j^0$ son los betas estimados por mínimos cuadrados ordinarios.

En la Figura 2–5 se muestran los coeficientes estimados por lasso para diferentes valores de t. La línea roja corresponde al valor óptimo $t/\sum|\hat{b}_j^0| = .44$.

A medida que aumenta el valor de la restricción, se van introduciendo mas variables al modelo, hasta llegar a los estimativos de mínimos cuadrados ordinarios (restricción relativa=1, es decir, sin restricción). Tomando como restricción .44 se obtiene un modelo con tres variables: lcavol, lweight y svi. Los coeficientes estimados por lasso son:

```
Call:
 l1ce(formula = f2, data = datos2, bound = opt)
Residuals:
```

Min	1Q	Median	3Q	Max
-1.72948	-0.32845	0.08286	0.42348	1.94585

Coefficients:

	Value	Std. Error	t	score Pr(>|t|)
(Intercept)	2.4784	0.0794	31.2138	0.0000
lcavol	0.5608	0.1113	5.0372	0.0000
lweight	0.1005	0.0898	1.1193	0.2630
age	0.0000	0.0870	0.0000	1.0000
lbph	0.0000	0.0883	0.0000	1.0000
svi	0.1583	0.1070	1.4790	0.1391
lcp	0.0000	0.1373	0.0000	1.0000
gleason	0.0000	0.1254	0.0000	1.0000
pgg45	0.0000	0.1353	0.0000	1.0000

Residual standard error:	0.782 on 88.41 degrees of freedom
The relative L1 bound was:	0.4444444
The absolute L1 bound was:	0.8195489
The Lagrangian for the bound is:	17.49261

Correlation of Coefficients:

	(*Intercept*)	*lcavol*	*lweight*	*age*	*lbph*	*svi*	*lcp*	*gleason*
lcavol	0,0000							
lweight	0,0000	−0,2215						
age	0,0000	−0,0682	−0,0946					
lbph	0,0000	−0,0187	−0,5291	−0,1519				
svi	0,0000	−0,2360	−0,1638	0,0554	0,0626			
lcp	0,0000	−0,4186	0,0613	0,0882	0,0465	−0,3819		
gleason	0,0000	−0,1998	0,1172	−0,0817	−0,0575	0,1102	−0,0277	
*pgg*45	0,0000	0,0984	−0,0415	−0,0789	−0,0992	−0,1908	−0,3012	−0,6470

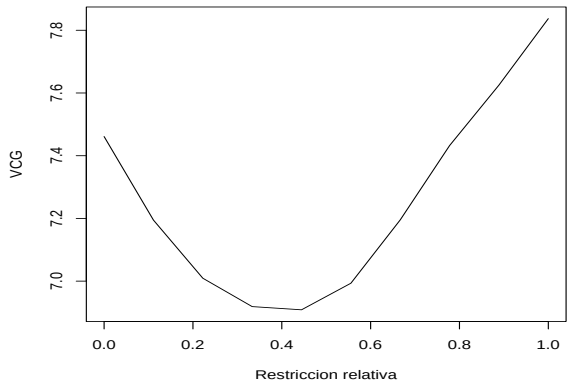

Figura 2–4: Validación cruzada generalizada para diferentes valores de t.

Los resultados obtenidos con regresión *Ridge* son:

	Value	Std. Error	t	score Pr(>|t|)
(Intercept)	0.66940	1.29638	0.52	0.6069
lcavol	0.58702	0.08792	6.68	<.0001
lweight	0.45446	0.17001	2.67	0.0090
age	-0.01964	0.01117	-1.76	0.0823
lbph	0.10705	0.05845	1.83	0.0704
svi	0.76616	0.24431	3.14	0.0023
lcp	-0.10547	0.09101	-1.16	0.2496
gleason	0.04514	0.15746	0.29	0.7751
pgg45	0.00453	0.00442	1.02	0.3089

Residual standard error:	0.708 on 88 degrees of freedom

Ambos métodos coinciden en escoger como las variables significativas: lcavol, lweight y svi. Puede notarse la tendencia de lasso a dar coeficientes estimados más pequeños.

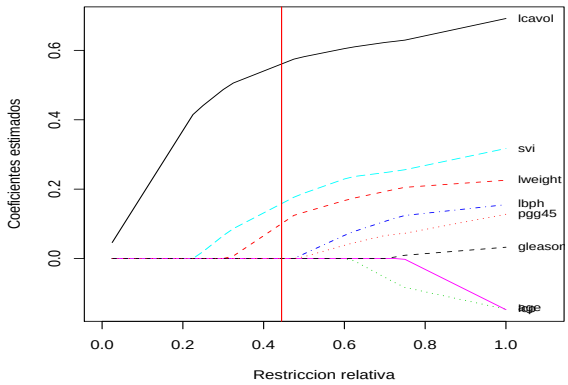

Figura 2–5: Coeficientes estimados por lasso para diferentes valores de t.

Capítulo 3
ESTIMACIÓN DE PARÁMETROS EN REGRESIÓN LOGÍSTICA Y PROBLEMAS ASOCIADOS

3.1. Estimación de parámetros en regresión logística

En regresión lineal el método más usado para estimación de parámetros es mínimos cuadrados. El método consiste en encontrar los valores de β que minimizan la suma de cuadrados de las desviaciones de los valores observados de Y a los valores predichos por el modelo.

Bajo los supuestos usuales de la regresión lineal, el método de mínimos cuadrados proporciona estimadores con algunas propiedades estadísticas deseables. Como se mencionó anteriormente, el modelo de regresión logística no cumple los supuestos de la regresión lineal, por lo tanto el método de mínimos cuadrados no produce estimaciones eficientes de la regresión logística.

Otro método frecuentemente utilizado es el método de máxima verosimilitud, el cual proporciona los valores de los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observado. El procedimiento consta de los siguientes pasos [14]:

1. Primero se construye la función de verosimilitud, la cual expresa la probabilidad de los datos observados en función del vector de parámetros desconocidos:

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)' \tag{3.1}$$

Para el caso de la regresión logística, si Y es codificada como cero o uno, entonces la expresión para $\pi(\mathbf{x})$ dada en (2.4) da la probabilidad condicional de que Y sea igual a 1 dado $\mathbf{x}$ y la cantidad $1-\pi(\mathbf{x})$ da la probabilidad condicional de que Y sea igual a 0 dado $\mathbf{x}$. Así, para los pares $(\mathbf{x}_i, y_i)$ en los cuales $y_i = 1$ la contribución a la función de verosimilitud es $\pi(\mathbf{x}_i)$ y para los pares en los que $y_i = 0$ la contribución a la función de verosimilitud es $1-\pi(\mathbf{x}_i)$. Por lo tanto, la contribución del par $(\mathbf{x}_i, y_i)$ a la función de verosimilitud es $\pi(\mathbf{x}_i)^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}$. Como las observaciones se asumen independientes, la función de verosimilitud es obtenida como el producto de los n términos:

$$l(\beta) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}$$

$$L(\beta) = ln(l(\beta)) = \sum_{i=1}^{n} y_i ln(\pi(\mathbf{x}_i)) + (1-y_i) ln(1-\pi(\mathbf{x}_i))$$

2. Se encuentra el vector de derivadas parciales de $L(\beta)$ con respecto a β,

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta} \tag{3.2}$$

con

$$u_k = \frac{\partial L(\beta)}{\partial \beta_k}\ ,\ k = 0, \ldots, p$$

donde p es la cantidad de parámetros. Cada una de las $p+1$ ecuaciones se iguala a cero y se resuelve para β_k, lo cual forma un conjunto de $p+1$ ecuaciones en $p+1$ incógnitas.

Tomando logaritmo natural en (2.4) se tiene:

$$ln[\pi(\mathbf{x}_i)] = \eta_i - ln(1 + e^{\eta_i})$$

$$ln[1 - \pi(\mathbf{x}_i)] = -ln(1 + e^{\eta_i})$$

$$\frac{\partial ln[\pi(\mathbf{x}_i)]}{\partial \beta_0} = 1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} = 1 - \pi(\mathbf{x}_i)$$

$$\frac{\partial ln[\pi(\mathbf{x}_i)]}{\partial \beta_j} = x_{ij} - \frac{x_{ij}e^{\eta_i}}{1 + e^{\eta_i}} = x_{ij}\left[1 - \pi(\mathbf{x}_i)\right]$$

$$\frac{\partial ln[1 - \pi(\mathbf{x}_i)]}{\partial \beta_0} = -\frac{e^{\eta_i}}{1 + e^{\eta_i}} = -\pi(\mathbf{x}_i)$$

$$\frac{\partial ln[1 - \pi(\mathbf{x}_i)]}{\partial \beta_j} = -\frac{x_{ij}e^{\eta_i}}{1 + e^{\eta_i}} = -x_{ij}\pi(\mathbf{x}_i)$$

Por lo tanto,

$$\begin{aligned}\frac{\partial L(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n}\{y_i(1 - \pi(\mathbf{x}_i)) + (1 - y_i)(-\pi(\mathbf{x}_i))\} \\ &= \sum_{i=1}^{n}\{y_i - \pi(\mathbf{x}_i)\}\end{aligned}$$

$$\begin{aligned}\frac{\partial L(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n}\{y_i x_{ij}(1 - \pi(\mathbf{x}_i)) + (1 - y_i)(-x_{ij}\pi(\mathbf{x}_i))\} \\ &= \sum_{i=1}^{n}\{x_{ij}[y_i - \pi(\mathbf{x}_i)]\}\end{aligned}$$

Luego,

$$U(\beta) = \left(\sum_{i=1}^{n}\{y_i - \pi(\mathbf{x}_i)\}, \sum_{i=1}^{n}\{x_{i1}[y_i - \pi(\mathbf{x}_i)]\}, \ldots, \sum_{i=1}^{n}\{x_{ip}[y_i - \pi(\mathbf{x}_i)]\}\right)' \tag{3.3}$$

$$= X'(Y - \Pi) \tag{3.4}$$

Con $Y = (y_1, \ldots, y_n)'$, $\Pi = (\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_n))'$ y la matriz de datos X dada por:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1' \\ 1 & \mathbf{x}_2' \\ \vdots & \vdots \\ 1 & \mathbf{x}_n' \end{pmatrix} \tag{3.5}$$

Así, las ecuaciones de verosimilitud son:

$$\sum_{i=1}^{n} \{y_i - \pi(\mathbf{x}_i)\} = 0 \tag{3.6}$$

$$\sum_{i=1}^{n} \{x_{ij}[y_i - \pi(\mathbf{x}_i)]\} = 0 \tag{3.7}$$

3. Para asegurar que los valores encontrados en el paso anterior identifican un máximo, la matriz de segundas derivadas parciales de $L(\beta)$

$$H(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_{j'}} \tag{3.8}$$

debe ser definida negativa. Para el caso de regresión logística:

$$H(\beta) = -\begin{pmatrix} \sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) & \cdots & \sum_{i=1}^{n} x_{ip}\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_{ip}\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) & \cdots & \sum_{i=1}^{n} x_{ip}^2\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \end{pmatrix} \tag{3.9}$$

$$= -X'WX \tag{3.10}$$

Con $W_{n\times n} = diag(w_i)$ y $w_i = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$. Debido a que las ecuaciones de verosimilitud (3.6) y (3.7) son no lineales en β y la matriz Hessiana (3.9) involucra los parámetros desconocidos, se requieren métodos especiales para obtener los estimadores de máxima verosimilitud, los cuales serán descritos en la Sección 3.3.

3.2. Estimación de parámetros en regresión logística multinomial

Suponga que la variable respuesta Y tiene J categorías: $0, 1, 2, \ldots, J-1$. Sea $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ el vector de p predictoras asociado a y_i, y sea $\pi_{ji} = \pi_j(\mathbf{x}_i) = p(Y = j|\mathbf{x}_i)$ para $j = 0, 1, \ldots, J-1$. Para construir la función de verosimilitud es conveniente crear J variables binarias codificadas como 0 o 1 indicando a que grupo pertenece la observación. Si $Y = j$ entonces $Y_j = 1$ y $Y_k = 0, \forall k \neq j$, $j = 0, \ldots, J-1$. La función condicional de verosimilitud para una muestra de n observaciones independientes es:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_{0i}^{y_{0i}} \pi_{1i}^{y_{1i}} \ldots \pi_{(J-1)i}^{y_{(J-1)i}}$$

Con $\boldsymbol{\beta}' = (\boldsymbol{\beta}_\mathbf{1}', \boldsymbol{\beta}_\mathbf{2}', \ldots, \boldsymbol{\beta}_{\mathbf{J-1}}')$ y $\boldsymbol{\beta}_\mathbf{j}' = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})$. Utilizando el hecho de que $\sum_j y_{ji} = 1$ para cada i, el logaritmo de la función de verosimilitud queda determinado por:

$$\begin{aligned}
L(\boldsymbol{\beta}) = ln(l(\boldsymbol{\beta})) &= \sum_{i=1}^{n} \left(y_{0i} ln(\pi_{0i}) + y_{1i} ln(\pi_{1i}) + \cdots + y_{(J-1)i} ln(\pi_{(J-1)i}) \right) \\
&= \sum_{i=1}^{n} [ln(\pi_{0i}) \left(y_{0i} + y_{1i} + \cdots + y_{(J-1)i} \right) + y_{1i} ln(\pi_{1i}) - y_{1i} ln(\pi_{0i}) + \ldots \\
&+ y_{(J-1)i} ln(\pi_{(J-1)i}) - y_{(J-1)i} ln(\pi_{0i})] \\
&= \sum_{i=1}^{n} [ln(\pi_{0i}) + y_{1i} g_1(\mathbf{x}_i) + \cdots + y_{(J-1)i} g_{J-1}(\mathbf{x}_i)] \\
&= \sum_{i=1}^{n} [y_{1i} g_1(\mathbf{x}_i) + \cdots + y_{(J-1)i} g_{J-1}(\mathbf{x}_i) - ln \left(1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)} \right)]
\end{aligned}$$

Las ecuaciones de verosimilitud se encuentran tomando la primera derivada parcial de $L(\boldsymbol{\beta})$ con respecto a cada uno de los $(J-1)(p+1)$ parámetros desconocidos:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} &= \sum_{i=1}^{n} \left(y_{ij} x_{ik} - \frac{x_{ik} e^{g_j(\mathbf{x}_i)}}{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}} \right) \\
&= \sum_{i=1}^{n} x_{ik} \left(y_{ij} - \pi_{ji} \right)
\end{aligned}$$

para $j = 1, \ldots, J-1$, $k = 0, 1, \ldots, p$ y $x_{0i} = 1$ para todo i. El estimador de máxima verosimilitud para β es obtenido igualando a cero las ecuaciones de verosimilitud y despejando para β. La solución requiere del mismo tipo de proceso iterativo del caso binomial. La matriz de información se obtiene a partir de la matriz de segundas derivadas parciales:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_{i=1}^{n} x_{ik'} x_{ik} \pi_{ji} (1 - \pi_{ji}) \tag{3.11}$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk} \partial \beta_{j'k'}} = \sum_{i=1}^{n} x_{ik'} x_{ik} \pi_{ji} \pi_{j'i} \tag{3.12}$$

para $j = 1, \ldots, J-1$, $j' = 1, \ldots, J-1$, $k = 0, 1, \ldots, p$ y $k' = 0, 1, \ldots, p$. La matriz de información $I(\boldsymbol{\beta})$ es $2(p+1) \times 2(p+1)$ y sus elementos son los negativos de los valores esperados de (3.11) y (3.12). La matriz de covarianza asintótica del estimador de máxima verosimilitud es: $\sum(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})$. Los estimadores de las matrices de información y covarianza se obtienen reemplazando los parámetros desconocidos por sus estimadores de máxima verosimilitud.

3.3. Métodos para obtener los estimadores de máxima verosimilitud

3.3.1. Algoritmo de Newton-Raphson

El algoritmo de Newton-Raphson es un método iterativo que sirve para resolver ecuaciones no lineales. Comienza con la estimación de mínimos cuadrados ordinarios (MCO) como punto inicial. Luego aproxima la función a ser maximizada por un polinomio de segundo grado en una vecindad del punto inicial, la segunda estimación será el valor máximo de ese polinomio. El siguiente paso es aproximar la función a ser maximizada por otro polinomio de segundo grado en una vecindad de la segunda estimación y el valor máximo de ese polinomio será la tercera estimación. De esta manera el método genera una sucesión de estimaciones que converge al máximo de la función cuando este existe [2]. La Figura 3–1 ilustra el proceso.

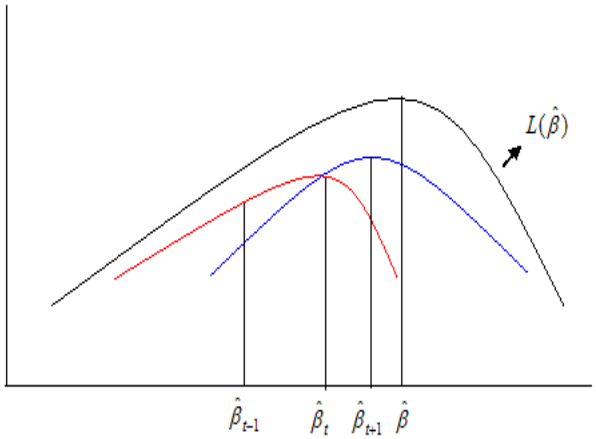

Figura 3–1: Método de Newton-Raphson

Formalmente, el algoritmo determina el valor $\hat{\beta}$ para el cual la función $L(\beta)$ es máxima. Dados $U^{(t)}$ y $H^{(t)}$ el vector (3.4) y la matriz (3.10) evaluados en la t-ésima estimación $\beta^{(t)}$. El paso t del proceso ($t = 0, 1, 2, \dots$) aproxima $L(\beta)$ en la vecindad de $\beta^{(t)}$ por la expansión de Taylor de segundo orden:

$$L(\beta) \approx L(\beta^{(t)}) + U'^{(t)}(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})' H^{(t)}(\beta - \beta^{(t)}) \tag{3.13}$$

Luego se busca el máximo del polinomio (3.13) resolviendo:

$$\frac{\partial L(\beta)}{\partial \beta} \approx U^{(t)} + H^{(t)}(\beta - \beta^{(t)}) = 0 \tag{3.14}$$

Despejando (3.14) para β se obtiene la estimación del paso $t + 1$:

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} U^{(t)}$$

Asumiendo que $H^{(t)}$ es no singular. En el modelo de regresión logística H y U están dados por (3.10) y (3.4), así $\beta^{(t+1)}$ se convierte en:

$$\beta^{(t+1)} = \beta^{(t)} + (X' W^{(t)} X)^{-1} X'(Y - \Pi^{(t)}) \tag{3.15}$$

El proceso iterativo continua hasta que los cambios en $L(\beta^{(t)})$ de una iteración a otra sean suficientemente pequeños.

3.3.2. Mínimos cuadrados reponderados iterativamente

Este método es utilizado por varios paquetes estadísticos para la estimación de parámetros en MLG. Para el caso de la regresión logística este método es equivalente a la estimación de máxima verosimilitud.

Note que la ecuación (3.15) puede ser reescrita como:

$$\begin{aligned} X'W^{(t)}X\beta^{(t+1)} &= X'W^{(t)}X\beta^{(t)} + X'(Y - \Pi^{(t)}) \\ &= X'W^{(t)}X\beta^{(t)} + X'W^{(t)}M \\ &= X'W^{(t)}(X\beta^{(t)} + M) \\ &= X'W^{(t)}\mathbf{z}^{(t)} \end{aligned}$$

Con $M = (\frac{y_1-\pi(\mathbf{x}_1)}{w_1}, \ldots, \frac{y_n-\pi(\mathbf{x}_n)}{w_n})'$, $w_i = \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))$ y $\mathbf{z}^{(t)} = (X\beta^{(t)} + M)$ con elementos:

$$\begin{aligned} z_i^{(t)} &= \sum_j x_{ij}\beta_j^{(t)} + \frac{y_i - \pi(\mathbf{x}_i)}{w_i} \\ &= \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial g(\mu_i)^{(t)}}{\partial \mu_i^{(t)}} \end{aligned}$$

Donde η_i, μ_i y g están dados por (2.1),(2.2) y (2.6). De esta forma, la ecuación (3.15) se convierte en:

$$\beta^{(t+1)} = (X'W^{(t)}X)^{-1}X'W^{(t)}\mathbf{z}^{(t)} \tag{3.16}$$

La expresión (3.16) corresponde a la estimación por mínimos cuadrados reponderados iterativamente (MCRI). Su nombre proviene del método de mínimos cuadrados ponderados (MCP), cuya solución para un modelo de la forma: $\mathbf{y} = X\beta + \epsilon$, con matriz de covarianza V, es:

$$(X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y} \tag{3.17}$$

En el MCRI los elementos de $W^{(t)}$ son $w_i = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i)) = V(y_i|\mathbf{x}_i)$ y el vector $\mathbf{z}$ es la linealización de la función de enlace g evaluada en $\mathbf{y}$.

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = \eta_i + (y_i - \mu_i)(\partial g(\mu_i)/\partial \mu_i) = z_i.$$

Todos los términos deben ser estimados y actualizados iterativamente hasta que las estimaciones se estabilicen. En cada paso, la variable dependiente ajustada $\hat{z}_i = \hat{g}(y_i)$ está dada por:

$$\hat{z}_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{\partial \hat{g}(\mu_i)}{\partial \hat{\mu}_i}$$

Los estimados por MCRI se obtienen de aplicar repetidas veces la ecuación(3.16), actualizando W y $\mathbf{z}$ hasta que la diferencia de las estimaciones de una iteración a otra sea insignificante. El proceso converge a los estimadores de máxima verosimilitud en pocas iteraciones, cuando estos existen [23].

3.4. Existencia y unicidad de los estimadores de máxima verosimilitud

Cuando se trabaja con conjuntos de datos pequeños en regresión logística, algunas veces ocurre que aunque la función de verosimilitud converge, por lo menos uno de los parámetros estimados es infinito. Casos como este ocurren cuando se presenta cierto patrón en los datos. Albert y Anderson [1] definen tres tipos de conFiguración de los datos: completamente separados, cuasicompletamente separados y traslapados.

3.4.1. Separación completa

Existe separación completa de los datos si la presencia o ausencia de la característica de interés puede ser perfectamente separada por una sola variable predictora o combinación lineal de variables predictoras. Esto es, si existe un vector b tal que:

$$\begin{cases} \mathbf{b}'\mathbf{x}_i > 0 & y_i = 1 \\ \mathbf{b}'\mathbf{x}_i < 0 & y_i = 0. \end{cases} \tag{3.18}$$

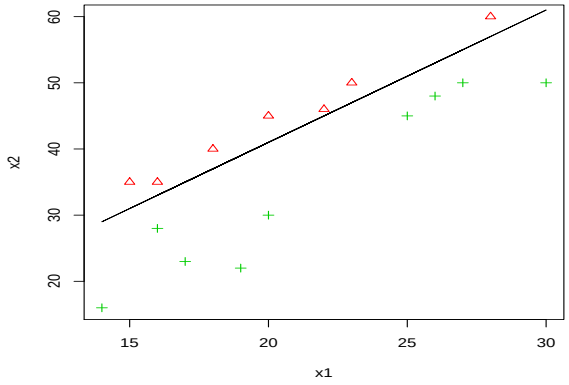

Figura 3–2: Separación completa

Por ejemplo, los datos de la Figura (3–2) pueden ser totalmente separados en sus grupos de respuesta, por la linea $x_2 = 2x_1 + 1$. En estos casos, el logaritmo de la función de verosimilitud converge a cero pero el estimador de máxima verosimilitud no existe, tiende a infinito.

3.4.2. Separación cuasicompleta

En la separación cuasicompleta los datos no están completamente separados, existe un vector b tal que:

$$\begin{cases} \mathbf{b}'\mathbf{x}_i \geq 0 & y_i = 1 \\ \mathbf{b}'\mathbf{x}_i \leq 0 & y_i = 0 \end{cases} \tag{3.19}$$

y la igualdad se cumple por lo menos para un individuo de cada grupo de respuesta. Un ejemplo de esto se muestra en la Figura (3–3), no existe una línea que separe completamente los datos. En este caso, el logaritmo de la función de verosimilitud no tiende a cero, puede converger a un valor diferente de cero. El estimador de máxima verosimilitud y la matriz de dispersión de los parámetros estimados tienden a infinito a medida que el número de iteraciones crece. Varianzas grandes de los pseudoestimados son típicas de una separación cuasicompleta de los datos [30].

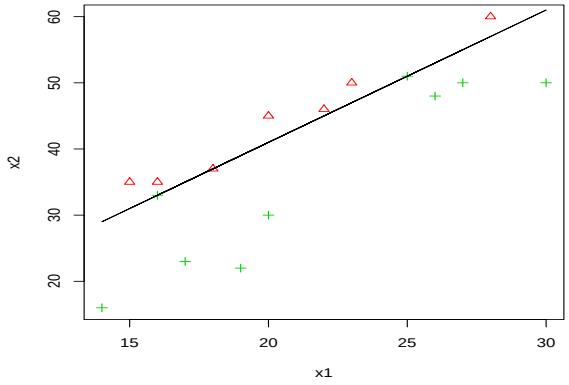

Figura 3–3: Separación cuasicompleta

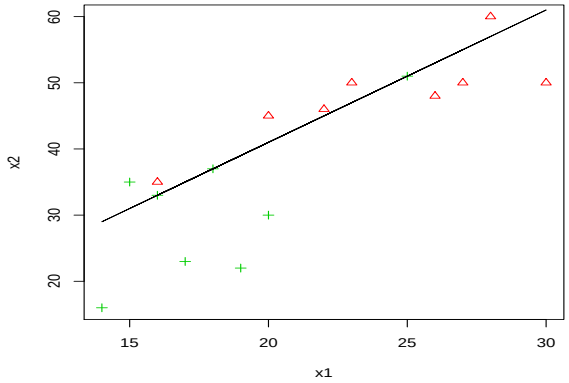

Figura 3–4: Datos traslapados

3.4.3. Traslape

No existe separación completa ni cuasicompleta, los datos están traslapados. Un ejemplo de esto se muestra en la Figura (3–4). El estimador de máxima verosimilitud existe y es único.

Capítulo 4
REGRESIÓN LOGÍSTICA PENALIZADA

4.1. Introducción

En clasificación algunas veces se trabaja con bases de datos que tienen más variables que observaciones. Este es el caso de las bases de datos de microarreglos, que consisten de un número relativamente pequeño de observaciones, generalmente menos de 100, y una gran cantidad de variables, usualmente miles. Esto genera que al estimar los parámetros de la regresión logística se presenten inconvenientes como: sobreajuste, inestabilidad y multicolinealidad.

El problema de sobre ajuste se refiere a que se obtiene una ecuación que ajusta muy bien a los datos pero tiene demasiados parámetros y así, no dará una buena predicción de nuevos datos debido a que el ruido del primer conjunto de datos no se ha filtrado aún. El segundo inconveniente, la inestabilidad, ocurre porque al ser $n < p$ hay más parámetros desconocidos que ecuaciones, por lo tanto puede haber infinitas soluciones. El tercer inconveniente, la multicolinealidad, es muy común en datos de microarreglos pues hay una alta probabilidad de encontrar genes con patrones de expresión casi idénticos; adicionalmente puede presentarse multicolinealidad accidental debido a que los niveles de expresión se miden con precisión limitada, hay muchas fuentes de error en el proceso que lleva del patrón fluorescente del microarreglo a los números que se utilizan para el modelamiento estadístico [8]. Cuando la multicolinealidad está presente, algunos de los coeficientes de regresión (o todos) son muy grandes e inestables, un pequeño cambio en los datos puede producir coeficientes

estimados muy diferentes. Una de las formas de tratar con estos inconvenientes es aplicar una penalidad *Ridge* al logaritmo de la función de verosimilitud. Esta restricción permite que solo los coeficientes de regresión que son realmente relevantes sean grandes. La penalidad *Ridge* estabiliza el problema estadístico y elimina la degeneranción numérica debida a la multicolinealidad [8]. La regresión logística penalizada no es una técnica de reducción de la dimensionalidad, todas las variables predictoras se permiten en el modelo de regresión.

El uso de penalidades en regresión logística fue propuesto por Cessie y Houwelingen [4] en 1992 para el caso en que la variable respuesta es de tipo ordinal. El método fue aplicado a un problema de sobrevivencia. En 2002, Eilers et al [8] extendieron el resultado a variables respuesta es de tipo nominal con dos clases. Para la estimación de parámetros propusieron utilizar el método de Newton Raphson reemplazando la matriz de predictoras X por su descomposición en valores singulares, en cuyo caso se trabaja con una matriz de tamaño $n \times n$ en vez de la matriz X de tamaño $n \times p$, esto permite solucionar el sistema rápidamente cuando n es pequeño. El método fue aplicado a la base de datos leukemia. En 2004, Zhu y Hastie [32] generalizaron el método de Eilers al caso en que la variable respuesta es de tipo multinomial. En cuanto a la estimación de parámetros mencionaron que el método de Newton Raphson requiere invertir matrices de tamaño $Jn \times Jn$ (donde J es la cantidad de clases y n la cantidad de observaciones) en cada iteración, lo cual es computacionalmente muy pesado. Ellos propusieron utilizar el algoritmo: Optimización Minima Secuencial (SMO), el cual fue creado por Platt [22] en 1998 para la estimación de parámetros en *"Support Vector Machines"* en el caso en que la variable respuesta tiene dos categorías. Posteriormente Keerthi et al [15] lo aplicaron a regresión logística penalizada con dos clases y Zhu y Hastie lo extendieron al caso multinomial. El método fue aplicado a tres bases de datos públicas: leukemia, SRBCT y Ramaswamy. Antes de hacer la

regresión logística penalizada, ellos efectuaron selección de variables predictoras.

En este capítulo se presenta la regresión logística con penalidad *Ridge* para el caso binomial y luego se generaliza al caso multinomial. También se describe el algoritmo SMO propuesto por Zhu [32] como método para encontrar los estimadores de máxima verosimilitud de la regresión logística penalizada. Con el fin de comprender totalmente el algoritmo SMO se introducen primero algunos conceptos necesarios sobre optimización no lineal.

4.2. Conceptos preliminares

4.2.1. Parámetros no identificables

Dado $\mathbf{x}$ un vector de variables aleatorias observadas, con $\mathbf{x}$ en el espacio muestral Ω. Sea f la función de distribución de probabilidad para un modelo completamente especificado por los parámetros θ. Si existen $\theta_1 \neq \theta_2$ para los cuales $f(\mathbf{x}|\theta_1) = f(\mathbf{x}|\theta_2)$ para todo $\mathbf{x} \in \Omega$, entonces los parámetros del modelo son no identificables. Esto produce que todos los posibles conjuntos de observaciones tengan probabilidades idénticas para dos diferentes conjuntos de parámetros [3].

Ejemplo:

Suponga un modelo de regresión lineal simple $\mathbf{y} = \alpha + \beta X$, en el cual se tiene solo una observación para determinar la línea de regresión (ver Figura 4–1). En este caso los parámetros del modelo: α y β son no identificables pues existen infinitas rectas que pasan por el punto y así habrá infinitas líneas de regresión que ajustan el modelo.

4.2.2. Funciones convexas y cóncavas

Función convexa:

Una función $f(x)$ es *convexa* si (Ver Figura 4–2):

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

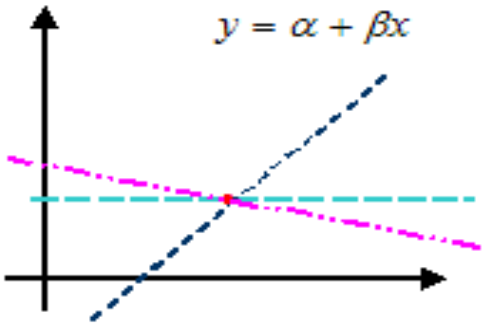

Figura 4–1: Regresión lineal simple. No identificabilidad de los parámetros α y β.

para todo x y y y para todo $\lambda \in [0, 1]$.

Función estrictamente convexa:

Una función $f(x)$ es *estrictamente convexa* si:

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y)$$

para todo x y y y para todo $\lambda \in [0, 1]$, $y \neq x$.

Función cóncava:

Una función $f(x)$ es *cóncava* si (Ver Figura 4–3):

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$$

para todo x y y y para todo $\lambda \in [0, 1]$.

Función estrictamente cóncava:

Una función $f(x)$ es *estrictamente cóncava* si:

$$f(\lambda x + (1 - \lambda)y) > \lambda f(x) + (1 - \lambda)f(y)$$

para todo x y y y para todo $\lambda \in [0, 1]$, $y \neq x$.

4.2.3. Optimización no lineal: problemas primal y dual

Considere la función Lagrangiana:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

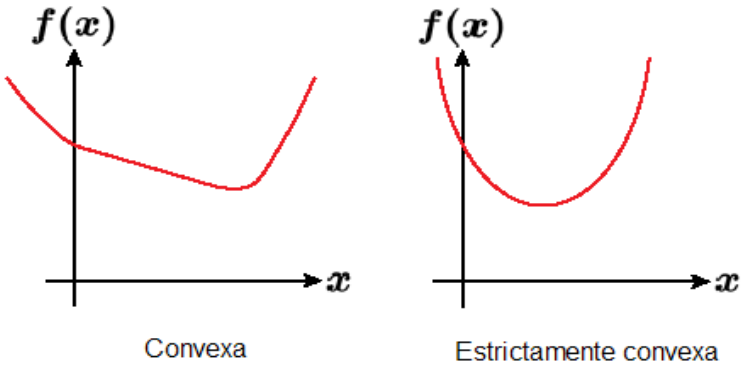

Figura 4–2: Funciones convexas

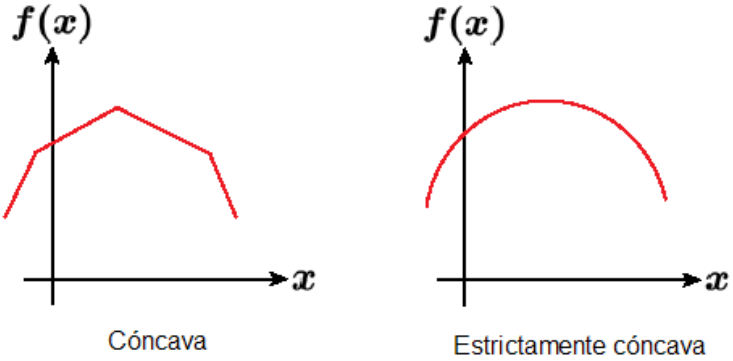

Figura 4–3: Funciones cóncavas

Con $x \in X \subset R^n$ y $\lambda \in \Lambda = \{\lambda = (\lambda_1, \ldots, \lambda_m)' | \lambda_i \geq 0, i = 1, \ldots, m\}$. Se definen las funciones primal y dual como [31]:

$$\textit{Función primal: } L_*(x) = \min_{\lambda \in \Lambda} L(x, \lambda)$$

$$\textit{Función dual : } L^*(\lambda) = \max_{x \in X} L(x, \lambda)$$

Estas funciones generan los problemas primal y dual. El problema *primal* es determinar el valor óptimo x^0 tal que:

$$L_*(x^0) = \max_{x \in X} L_*(x)$$

El correspondiente problema *dual* es calcular un λ^0 óptimo tal que:

$$L^*(\lambda^0) = \min_{\lambda \in \Lambda} L^*(\lambda)$$

El problema primal consiste entonces en la maximización de la función primal:

$$L_*(x) = \min_{\lambda}\{f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)\} = \begin{cases} f(x) & \text{si todas las } g_i(x) \geq 0, \\ -\infty & \text{si existe una } g_i(x) < 0. \end{cases} \tag{4.1}$$

Como $\lambda_i \geq 0$ entonces si $g_i(x) \geq 0$, el valor de λ_i que minimiza (4.1) es $\lambda_i = 0$. Sin embargo, si algún $g_i(x) < 0$, hacer $\lambda_i \longrightarrow +\infty$ minimizará (4.1). El objetivo del problema primal es maximizar la función primal $L_*(x)$, así que la porción en la cual $L_*(x) = -\infty$ puede ser descartada. El problema primal se convierte en:

$$\text{máx } f(x)$$

$$\text{sujeto a } g_i(x) \geq 0, i = 1, \ldots, m$$

Análogamente, el problema dual consiste en la minimización de la función dual: $L^*(\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$ sujeto a:

$$f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = \max_{x \in X} f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \tag{4.2}$$

Pero, si $X = R^n$ y todas las funciones son cóncavas entonces:

$$\nabla L(x, \lambda) = 0 \text{ si y solo si } L(x, \lambda) = \max_{x \in R^n} L(x, \lambda) \tag{4.3}$$

Entonces el problema dual, con función objetivo y restricciones cóncavas, se convierte en:

$$\text{mín } f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

$$\text{sujeto a: } \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) = 0$$

$$\lambda_i \geq 0, i = 1, \ldots, m$$

Ejemplo:

Suponga el problema de programación lineal:

$$\text{máx } q'x$$

$$\text{sujeto a: } Ax \leq b$$

El correspondiente problema dual es;

$$\text{mín } q'x + \lambda'(b - Ax)$$

$$\text{sujeto a: } q' - \lambda'A = 0$$

$$\lambda \geq 0$$

Pero, $q' - \lambda'A = 0$ implica $q'x - \lambda'Ax = 0$, entonces el problema dual se convierte en:

$$\text{mín } \lambda'b$$

$$\text{sujeto a: } A'\lambda = q$$

$$\lambda \geq 0$$

Problema dual de Wolfe

Como se ilustró en el ejemplo anterior, dependiendo de la forma del problema primal, el problema dual puede reescribirse de una manera más simple. Cuando el problema primal consiste en:

$$\text{mín } f(\mathbf{w}, \mathbf{z}) = \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{2\lambda}\sum_{i=1}^{l} z_i^2$$

$$\text{sujeto a: } y_i - \mathbf{x}_i\mathbf{w} = z_i, i = 1, \ldots, l.$$

El dual se convierte en el denominado problema dual de Wolfe:

$$\text{máx } L(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{2\lambda}\sum_{i=1}^{l} z_i^2 + \sum_{i=1}^{l} \alpha_i(y_i - \mathbf{x}_i\mathbf{w} - z_i)$$

$$\text{sujeto a: } \nabla_{\mathbf{w}} L(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{l} \alpha_i \mathbf{x}_i = 0.$$

$$\nabla_{\mathbf{z}} L(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha}) = \frac{1}{\lambda}\mathbf{z} - \boldsymbol{\alpha} = 0.$$

4.2.4. Teorema KKT (Karush-Kuhn-Tucker)

Modelo de optimización convexa

CP: minimizar $f(x)$

sujeto a:

$$g_1(x) \leq 0,$$
$$\vdots$$
$$g_m(x) \leq 0,$$
$$Ax = b,$$
$$x \in \mathbb{R}^n,$$

CP es llamado problema de optimización convexa si $f(x), g_1(x), \ldots, g_m(x)$ son funciones convexas.

Teorema KKT

Suponga que $f(x), g_1(x), \ldots, g_m(x)$ son todas funciones convexas. Entonces $\bar{x}$ resuelve CP si y solo si existen $\bar{y}_i \geq 0, i = 1, \ldots, m$, tal que [31]:

i. $\nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{y}_i \nabla g_i(\bar{x}) = 0$

ii. $g_i(\bar{x}) - b_i \leq 0$

iii. $\bar{y}_i(b_i - g_i(\bar{x})) = 0$

Ejemplo: Considere el problema de minimizar $6(x_1 - 10)^2 + 4(x_2 - 12{,}5)^2$ sujeto a:

$x_1^2 + (x_2 - 5)^2 \leq 50$

$x_1^2 + 3x_2^2 \leq 200$

$(x_1 - 6)^2 + x_2^2 \leq 37$

y verifique si $\bar{x} = (7, 6)$ es una solución óptima de este problema.

Solución:

$f(x) = 6(x_1 - 10)^2 + 4(x_2 - 12{,}5)^2$

$g_1(x) = x_1^2 + (x_2 - 5)^2 - 50$

$g_2(x) = x_1^2 + 3x_2^2 - 200$

$g_3(x) = (x_1 - 6)^2 + x_2^2 - 37$

Para verificar (ii), se evalúa: $g_1(\bar{x}) = 0$, $g_2(\bar{x}) = -43 < 0$ y $g_3(\bar{x}) = 0$. Como $g_2(\bar{x}) \neq 0$ entonces y_2 debe ser cero para que se cumpla (iii). Para verificar las condiciones de optimalidad se evalúan los gradientes en $\bar{x}$:

$$\nabla f(\bar{x}) = \begin{pmatrix} 12(\bar{x}_1 - 10) \\ 8(\bar{x}_2 - 12{,}5) \end{pmatrix} = \begin{pmatrix} -36 \\ -52 \end{pmatrix}, \ \nabla g_1(\bar{x}) = \begin{pmatrix} 2\bar{x}_1 \\ 2(\bar{x}_2 - 5) \end{pmatrix} = \begin{pmatrix} 14 \\ 2 \end{pmatrix},$$

$$\nabla g_2(\bar{x}) = \begin{pmatrix} 2\bar{x}_1 \\ 6\bar{x}_2 \end{pmatrix} = \begin{pmatrix} 14 \\ 36 \end{pmatrix} \text{ y } \nabla g_3(\bar{x}) = \begin{pmatrix} 2(\bar{x}_1 - 6) \\ 2\bar{x}_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 12 \end{pmatrix}$$

y se resuelve el sistema:

$$\begin{pmatrix} -36 \\ -52 \end{pmatrix} + \begin{pmatrix} 14 \\ 2 \end{pmatrix} y_1 + \begin{pmatrix} 14 \\ 36 \end{pmatrix} y_2 + \begin{pmatrix} 2 \\ 12 \end{pmatrix} y_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

para $y_1 \geq 0$, $y_2 = 0$ y $y_3 \geq 0$. La solución del sistema es: $\bar{y} = (\bar{y}_1, \bar{y}_2, \bar{y}_3) = (2, 0, 4)$, la cual cumple las condiciones (i) a (iii), por lo tanto $\bar{x}$ es una solución óptima del problema.

4.3. Regresión logística binomial penalizada

La regresión logística binomial penalizada fue propuesta por Cessie y Houwelingen en 1992 [4]. Dada la variable respuesta $Y = (y_1, \ldots, y_n)'$ y dado $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$

el vector de predictoras asociado a y_i, el modelo de regresión logística es:

$$g\left(\pi(\mathbf{x}_i)\right) = ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right)$$

Con $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij}$, $x_{0i} = 1$ para todo i y $\pi(\mathbf{x}_i) = p(y_i = 1|\mathbf{x}_i)$. Como se mencionó en la Sección 3.1, el logaritmo de la función de verosimilitud está dado por:

$$L(\beta) = \sum_{i=1}^{n} y_i ln(\pi(\mathbf{x}_i)) + (1-y_i) ln(1-\pi(\mathbf{x}_i))$$

Y la función penalizada está dada por:

$$L^*(\beta) = L(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2$$

Se debe notar que β_0 no está involucrado en la penalidad. En la Sección 2.3.1 se ilustró cómo el parámetro λ regula la penalidad: entre más grande es λ más pequeños son los coeficientes $\beta's$.

Con el fin de encontrar el valor de β que maximiza la función $L^*(\beta)$, se hallan las derivadas parciales:

$$U^*(\beta) = \frac{\partial L^*(\beta)}{\partial \beta_j} = \begin{cases} \frac{\partial L(\beta)}{\partial \beta_j} & j = 0 \\ \frac{\partial L(\beta)}{\partial \beta_j} - \lambda \beta_j & j = 1, \ldots, p. \end{cases}$$

$$H^*(\beta) = \begin{cases} \frac{\partial^2 L^*(\beta)}{\partial \beta_j \partial \beta_{j'}} = \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_{j'}} & j = 0, \ldots, p \\ \frac{\partial^2 L^*(\beta)}{\partial \beta_j^2} = \frac{\partial^2 L(\beta)}{\partial \beta_j^2} & j = 0 \\ \frac{\partial^2 L^*(\beta)}{\partial \beta_j^2} = \frac{\partial^2 L(\beta)}{\partial \beta_j^2} - \lambda & j = 1, \ldots, p \end{cases}$$

Dada $O_{(n+1)\times(n+1)}$ la matriz identidad con $o_{11} = 0$ y dados β, X, $U(\beta)$ y $H(\beta)$ como en (3.1), (3.5), (3.4) y (3.10) respectivamente, se puede expresar $U^*(\beta)$ y $H^*(\beta)$

matricialmente como:

$$U^*(\beta) = U(\beta) - \lambda O\beta = X'(Y - \Pi) - \lambda O\beta$$
$$H^*(\beta) = H(\beta) - \lambda O = -X'WX - \lambda O$$

Debido a que $U^*(\beta)$ y $H^*(\beta)$ no son lineales en β se debe recurrir a los métodos vistos en la Sección 3.3 para obtener los estimadores de máxima verosimilitud.

El algoritmo de Newton Raphson, descrito en la Sección 3.3.1, aplicado a la función penalizada $L^*(\beta)$ da una estimación de paso $t+1$ dada por:

$$\beta^{(t+1)} = \beta^{(t)} - (H^{*(t)})^{-1}U^{*(t)} \tag{4.4}$$
$$= \beta^{(t)} + (X'W^{(t)}X + \lambda O)^{-1}[X'(Y - \Pi^{(t)}) - \lambda O\beta^{(t)}] \tag{4.5}$$

Lo cual es equivalente a:

$$\begin{aligned}(X'W^{(t)}X + \lambda O)\beta^{(t+1)} &= (X'W^{(t)}X + \lambda O)\beta^{(t)} + X'(Y - \Pi^{(t)}) - \lambda O\beta^{(t)} \\ &= X'W^{(t)}X\beta^{(t)} + \lambda O\beta^{(t)} + X'(Y - \Pi^{(t)}) - \lambda O\beta^{(t)} \\ &= X'W^{(t)}X\beta^{(t)} + X'(Y - \Pi^{(t)}) \\ &= X'W^{(t)}(X\beta^{(t)} + M) \\ &= X'W^{(t)}\mathbf{z}^{(t)}\end{aligned}$$

Con $M = (\frac{y_1 - \pi(\mathbf{x}_1)}{w_1}, \ldots, \frac{y_n - \pi(\mathbf{x}_n)}{w_n})'$, $w_i = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$ y $\mathbf{z}^{(t)} = (X\beta^{(t)} + M)$ con elementos:

$$z_i^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + \frac{y_i - \pi(\mathbf{x}_i)}{w_i}$$

De esta forma la ecuación (4.5) se convierte en:

$$\beta^{(t+1)} = (X'W^{(t)}X + \lambda O)^{-1}X'W^{(t)}\mathbf{z}^{(t)} \tag{4.6}$$

La expresión (4.6) corresponde a la estimación por el método MCRI descrito en la Sección 3.3.2. Los estimados por MCRI se obtienen de aplicar repetidas veces la ecuación (4.6), actualizando W y $\mathbf{z}$ hasta que la diferencia de las estimaciones de una iteración a otra sea insignificante.

La ecuación (4.6) genera una gran cantidad de ecuaciones con un gran número de parámetros desconocidos, lo cual es computacionalmente muy pesado con bases de datos grandes. Eilers et al [8] recomiendan utilizar la descomposición de X en valores singulares para resolver el sistema y usar como valores iniciales $\beta_0 = log[\bar{y}/(1-\bar{y})]$ con $\bar{y} = \sum_{i=1}^{n} y_i/n$ y $\beta_j = 0$ para $j = 1, \ldots, p$.

4.4. Regresión logística multinomial penalizada

Este método fue propuesto por Zhu y Hastie en 2004 [32]. Suponga que se tiene una variable respuesta $Y = (y_1, ..., y_n)'$ con J categorías, es decir que cada $y_i, i = 1, \ldots, n$ toma uno de los J valores: $0, 1, 2, \ldots, J-1$. Suponga además que se crean J variables binarias codificadas como 0 o 1 indicando a que grupo pertenece cada observación, es decir: si $Y = j$ entonces $Y_j = 1$ y $Y_k = 0, \forall k \neq j$, $j = 0, 1, \ldots, J-1$. Dado $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ el vector de p predictoras asociado a y_i, $\pi_{ji} = \pi_j(\mathbf{x}_i) = p(\mathbf{y} = j|\mathbf{x}_i)$ para $j = 0, 1, \ldots, J-1$, $\beta' = (\beta_1', \beta_2', \ldots, \beta_{J-1}')$, $\beta_j' = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})$ y $\mathbf{b}_j' = (\beta_{j1}, \ldots, \beta_{jp})$ en la Sección 2.2 se mostró que:

$$\pi_{ji} = \frac{e^{g_j(\mathbf{x}_i)}}{e^{g_0(\mathbf{x}_i)} + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}}, j = 0, \ldots, J-1$$

con $g_0(\mathbf{x}_i) = 1$ y

$$g_j(\mathbf{x}_i) = ln\left(\frac{\pi_{ji}}{\pi_{0i}}\right) = \beta_{j0} + \beta_{j1}x_{1i} + \beta_{j2}x_{2i} + \cdots + \beta_{jp}x_{pi}$$
$$= \beta_{j0} + \mathbf{b}_j'\mathbf{x}_i, j = 1, \ldots, J-1.$$

Note que $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{J-1}(\mathbf{x})$ son no identificables en este modelo, pues si se agrega un $\beta_0 + \sum_{j=1}^{p}\beta_j x_j$ común a cada componente de $\boldsymbol{g}(\mathbf{x}) = (g_0(\mathbf{x}), \ldots, g_{J-1}(\mathbf{x}))$, se tiene:

$$\boldsymbol{g}'(\mathbf{x}) = \left(g_0(\mathbf{x}) + \beta_0 + \sum_{j=1}^{p}\beta_j x_j, \ldots, g_{J-1}(\mathbf{x}) + \beta_0 + \sum_{j=1}^{p}\beta_j x_j\right) \neq \boldsymbol{g}(\mathbf{x})$$

y sin embargo para $j = 0, \ldots, J-1$ se tiene:

$$\pi_j(\boldsymbol{g}'(\mathbf{x})) = \frac{e^{g_j(\mathbf{x}_i)+\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{e^{g_0(\mathbf{x}_i)+\beta_0+\sum_{j=1}^{p}\beta_j x_j} + \cdots + e^{g_{J-1}(\mathbf{x}_i)+\beta_0+\sum_{j=1}^{p}\beta_j x_j}}$$
$$= \frac{e^{g_j(\mathbf{x}_i)}e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}\left(e^{g_0(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}\right)}$$
$$= \pi_j(\boldsymbol{g}(\mathbf{x}))$$

Para hacer $g_j(\mathbf{x})$ identificable, se toma la restricción simétrica $\sum_{i=0}^{J-1} g_j(\mathbf{x}) = 0$. Así, dado $\tilde{\boldsymbol{g}}(\mathbf{x}) = (\tilde{g}_0(\mathbf{x}), \ldots, \tilde{g}_{J-1}(\mathbf{x}))$ arbitrario, entonces $\pi_j(\boldsymbol{g}(\mathbf{x})) = \pi_j(\tilde{\boldsymbol{g}}(\mathbf{x})), j = 0, \ldots, J-1$ implica que:

$$\frac{e^{g_0(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} = \frac{e^{\tilde{g}_0(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}}$$
$$\frac{e^{g_1(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} = \frac{e^{\tilde{g}_1(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}}$$
$$\vdots$$
$$\frac{e^{g_{J-1}(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} = \frac{e^{\tilde{g}_{J-1}(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}}$$

Multiplicando todos los $\pi_j, j = 1, \ldots, J-1$ se tiene:

$$\frac{e^{g_0(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} \frac{e^{g_1(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} \cdots \frac{e^{g_{J-1}(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}} = \frac{e^{\tilde{g}_0(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}} \frac{e^{\tilde{g}_1(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}} \cdots \frac{e^{\tilde{g}_{J-1}(\mathbf{x})}}{\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}}$$

$$\frac{e^{g_0(\mathbf{x})+g_1(\mathbf{x})+\cdots+g_{J-1}(\mathbf{x})}}{\left(\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}\right)^n} = \frac{e^{\tilde{g}_0(\mathbf{x})+\tilde{g}_1(\mathbf{x})+\cdots+\tilde{g}_{J-1}(\mathbf{x})}}{\left(\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}\right)^n}$$

Pero $\sum_{i=0}^{J-1} g_j(\mathbf{x}) = 0$ y $\sum_{i=0}^{J-1} \tilde{g}_j(\mathbf{x}) = 0$, entonces:

$$\frac{1}{\left(\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})}\right)^n} = \frac{1}{\left(\sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}\right)^n}$$

$$\sum_{j=0}^{J-1} e^{g_j(\mathbf{x})} = \sum_{j=0}^{J-1} e^{\tilde{g}_j(\mathbf{x})}$$

Por lo tanto, de los π_j y de la monotonicidad de la función exponencial se tiene que:

$$g_1(\mathbf{x}) = \tilde{g}_1(\mathbf{x})$$

$$g_2(\mathbf{x}) = \tilde{g}_2(\mathbf{x})$$

$$\vdots$$

$$g_{J-1}(\mathbf{x}) = \tilde{g}_{J-1}(\mathbf{x})$$

Entonces $\boldsymbol{g}(\mathbf{x}) = \tilde{\boldsymbol{g}}(\mathbf{x})$. Así, la restricción simétrica:

$$\sum_{i=0}^{J-1} g_j(\mathbf{x}) = 0 \tag{4.7}$$

hace que los parámetros $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{J-1}(\mathbf{x})$ sean identificables en el modelo.

En la Sección 2.2 se mostró además que el logaritmo de la función de verosimilitud está dado por:

$$L(\beta) = \sum_{i}^{n} [y_{1i} g_1(\mathbf{x}_i) + \cdots + y_{(J-1)i} g_{J-1}(\mathbf{x}_i) - ln\left(1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}\right)]$$

Aplicando la penalidad *Ridge* a la función $L(\beta)$ se obtiene:

$$L(\beta) - \frac{\lambda}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \tag{4.8}$$

La maximización de la función cóncava (4.8) es equivalente a la minimización de la función convexa:

$$L^*(\beta) = -\sum_{i}^{n}[y_{1i}g_1(\mathbf{x}_i) + \cdots + y_{(J-1)i}g_{J-1}(\mathbf{x}_i) - ln\left(1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}\right)] + \frac{\lambda}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \tag{4.9}$$

Con el fin de encontrar el valor de β que minimiza la función $L^*(\beta)$ se hallan las derivadas parciales:

$$\frac{\partial L^*(\beta)}{\partial \mathbf{b}_j} = \lambda \mathbf{b}_j - \sum_{i=1}^{n}\left[y_{ji}\frac{\partial g_j(\mathbf{x}_i)}{\partial \mathbf{b}_j} - \frac{e^{g_j(\mathbf{x}_i)}}{1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}}\frac{\partial g_j(\mathbf{x}_i)}{\partial \mathbf{b}_j}\right] \tag{4.10}$$

$$= \lambda \mathbf{b}_j - \sum_{i=1}^{n}\mathbf{x}_i(y_{ji} - \pi_{ji}) \tag{4.11}$$

$$= \lambda \mathbf{b}_j - X'(Y_j - \Pi_j) \tag{4.12}$$

para $j = 0, \ldots, J-1$, $X' = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $Y_j = (y_{j1}, \ldots, y_{jn})'$ y $\Pi_j = (\pi_{j1}, \ldots, \pi_{jn})'$. Igualar a cero la ecuación (4.12) indica que $\mathbf{b}_j$ está en el espacio fila de X, por lo tanto $\mathbf{b}_j$ puede ser escrito como:

$$\mathbf{b}_j = \sum_{i=1}^{n} a_{ij}\mathbf{x}_i, \;\; j = 0, \ldots, J-1 \tag{4.13}$$

con $a_{ij} \in \mathbb{R}$. Dado el producto interno Euclidiano:

$$< \mathbf{x}, \mathbf{x}^* >= \sum_{j=1}^{p} x_j x_j^*$$

Entonces $g_j(\mathbf{x})$ tiene la forma:

$$\begin{aligned} g_j(\mathbf{x}) &= \beta_{j0} + \mathbf{b}_j'\mathbf{x} \\ &= \beta_{j0} + \sum_{i=1}^{n} a_{ij}\mathbf{x}_i'\mathbf{x} \\ &= \beta_{j0} + \sum_{i=1}^{n} a_{ij} < \mathbf{x}, \mathbf{x}_i > \end{aligned}$$

y

$$\begin{aligned} \| \mathbf{b}_j \|^2 = \sum_{k=1}^{p} b_{jk}^2 = \mathbf{b}_j'\mathbf{b}_j &= \sum_{i=1}^{n} a_{ij}\mathbf{x}_i' \sum_{i=1}^{n} a_{ij}\mathbf{x}_i \\ &= (a_{1j}\mathbf{x}_1' + a_{2j}\mathbf{x}_2' + \cdots + a_{nj}\mathbf{x}_n')(a_{1j}\mathbf{x}_1 + a_{2j}\mathbf{x}_2 + \cdots + a_{nj}\mathbf{x}_n) \\ &= \sum_{i,i'} a_{ij}a_{i'j} < \mathbf{x}_i, \mathbf{x}_{i'} > \end{aligned}$$

Utilizando esto, la ecuación (4.9) se convierte en:

$$\begin{aligned} L^*(\beta) = -\sum_{i}^{n} [y_{1i}g_1(\mathbf{x}_i) + \cdots + y_{(J-1)i}g_{J-1}(\mathbf{x}_i) - ln\left(1 + e^{g_1(\mathbf{x}_i)} + \cdots + e^{g_{J-1}(\mathbf{x}_i)}\right)] \\ + \frac{\lambda}{2}\sum_{j=0}^{J-1}\sum_{i,i'} a_{ij}a_{i'j} < \mathbf{x}_i, \mathbf{x}_{i'} > \quad (4.14) \end{aligned}$$

El modelo de regresión logística penalizada se ajusta encontrando los a_{ij}'s que minimizan (4.14). Los b_{jk}'s se obtienen utilizando (4.13). Esto reduce la cantidad de parámetros de $(p+1)J$ en (4.9) (pJ b_{jk}'s y J b_{j0}'s) a $(n+1)J$ en (4.14) (nJ a_{ij}'s y J b_{j0}'s), haciendo los cálculos factibles para n razonablemente pequeño.

Si se desea utilizar el método de Newton Raphson para resolver (4.14) es necesario invertir matrices de $nJ \times nJ$ en cada iteración, lo cual tiene un alto costo computacional cuando n o J son grandes. Zhu y Hastie [32] proponen utilizar el algoritmo Optimización Minima Secuencial (SMO, por sus siglas en inglés). Este algoritmo fue aplicado por Keerthi et al [15] a problemas de clasificación de dos clases, pero Zhu y Hastie [32] lo generalizaron al caso multiclase.

4.5. Algoritmo Optimización Minima Secuencial (SMO)

El algoritmo que se presenta a continuación es la generalización al caso multi-clase realizada por Zhu y Hastie [32] para resolver el problema de minimización de (4.9).

Dividiendo por λ, el problema de minimización de la ecuación (4.9) sujeto a (4.7) puede ser reescrito como:

$$min\ P = C\sum_{i=1}^{n} g(\boldsymbol{\xi}_i) + \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2, \tag{4.15}$$

$$\text{sujeto a: } \xi_{ij} = \beta_{j0} + \mathbf{b}_j'\mathbf{x}_i, \forall i, j, \tag{4.16}$$

$$\sum_{j=0}^{J-1} \beta_{j0} = 0 \tag{4.17}$$

donde $C = 1/\lambda$ es el parámetro de regularización, $\sum_{j=0}^{J-1} \beta_{j0} = 0$ por la restricción (4.7) y

$$g(\boldsymbol{\xi}_i) = -(y_{1i}\xi_{1i} + \cdots + y_{(J-1)i}\xi_{(J-1)i}) + ln\left(1 + e^{\xi_{1i}} + \cdots + e^{\xi_{(J-1)i}}\right)$$

El Lagrangiano para este problema está dado por:

$$L = C\sum_{i}^{n} g(\xi_i) + \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 + \sum_{j=0}^{J-1}\sum_{i=1}^{n} a_{ij}[\xi_{ij} - \beta_{j0} - \mathbf{b}_j'\mathbf{x}_i] + a_0\sum_{j=0}^{J-1}\beta_{j0}, \tag{4.18}$$

donde los a_{ij}'s y a_0 son los multiplicadores de Lagrange. Las condiciones de optimalidad de KKT (ver Sección 4.2.4) están dadas por:

$$\frac{\partial L}{\partial \beta_{j0}} = a_0 - \sum_{i=1}^{n} a_{ij} = 0, \forall j \tag{4.19}$$

$$\frac{\partial L}{\partial \mathbf{b}_j} = \mathbf{b}_j - \sum_{i=1}^{n} a_{ij}\mathbf{x}_i = 0, \forall j \tag{4.20}$$

$$\frac{\partial L}{\partial \xi_{ij}} = C\left(-y_{ij} + \frac{e^{\xi_{ij}}}{1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}}}\right) + a_{ij} = 0, \forall i, j. \tag{4.21}$$

Estas ecuaciones permiten expresar a $\mathbf{b}_j$ y ξ_{ij} como función de los a_{ij}'s. De (4.20):

$$\mathbf{b}_j = \sum_{i=1}^{n} a_{ij}\mathbf{x}_i, \forall j \tag{4.22}$$

Despejando $e^{\xi_{ij}}$ en (4.21) se obtiene:

$$e^{\xi_{ij}} = \left(y_{ij} - \frac{a_{ij}}{C}\right)\left(1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}}\right), \forall i, j. \tag{4.23}$$

Dado un j_1 arbitrario de (4.23) se tiene que:

$$1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}} = \frac{e^{\xi_{ij_1}}}{y_{ij_1} - \frac{a_{ij_1}}{C}}, \forall i. \tag{4.24}$$

Utilizando (4.23) y (4.24) se obtiene:

$$e^{\xi_{ij}} = \left(y_{ij} - \frac{a_{ij}}{C}\right)\frac{e^{\xi_{ij_1}}}{y_{ij_1} - \frac{a_{ij_1}}{C}}, \forall i, j. \tag{4.25}$$

$$\xi_{ij} = ln\left(y_{ij} - \frac{a_{ij}}{C}\right) + \xi_{ij_1} - ln\left(y_{ij_1} - \frac{a_{ij_1}}{C}\right), \forall i, j. \tag{4.26}$$

Ahora, sumando (4.26) sobre todos los j's y utilizando el echo de que $\sum_{j'=0}^{J-1} \xi_{ij} = 0$ (por la restricción (4.7)), se tiene:

$$0 = \sum_{j'=0}^{J-1} ln\left(y_{ij'} - \frac{a_{ij'}}{C}\right) + J\xi_{ij_1} - Jln\left(y_{ij_1} - \frac{a_{ij_1}}{C}\right)$$

De donde:

$$\xi_{ij_1} = ln\left(y_{ij_1} - \frac{a_{ij_1}}{C}\right) - \frac{1}{J}\sum_{j'=0}^{J-1} ln\left(y_{ij'} - \frac{a_{ij'}}{C}\right), \forall i$$

Como j_1 se tomo arbitrario, entonces

$$\xi_{ij} = ln\left(y_{ij} - \frac{a_{ij}}{C}\right) - \frac{1}{J}\sum_{j'=0}^{J-1} ln\left(y_{ij'} - \frac{a_{ij'}}{C}\right), \forall i, j. \tag{4.27}$$

Notar también que si sumamos (4.21) sobre todos los j's y utilizando el hecho de que $\sum_{j=0}^{J-1} y_{ij} = 1$ se tiene:

$$\begin{aligned}
0 &= C\sum_{j=0}^{J-1}\left(-y_{ij} + \frac{e^{\xi_{ij}}}{1+\sum_{j'=1}^{J-1} e^{\xi_{ij'}}}\right) + \sum_{j=0}^{J-1} a_{ij} \\
0 &= C\left(-\sum_{j=0}^{J-1} y_{ij} + \frac{\sum_{j=0}^{J-1} e^{\xi_{ij}}}{1+\sum_{j'=1}^{J-1} e^{\xi_{ij'}}}\right) + \sum_{j=0}^{J-1} a_{ij} \\
0 &= \sum_{j=0}^{J-1} a_{ij}
\end{aligned}$$

y de (4.19) se obtiene:

$$\begin{aligned}
a_0 &= \sum_{i=1}^{n} a_{ij}, \forall j, \\
\sum_{j=0}^{J-1} a_0 &= \sum_{j=0}^{J-1}\sum_{i=1}^{n} a_{ij}, \\
Ja_0 &= \sum_{i=1}^{n}\sum_{j=0}^{J-1} a_{ij}, \\
Ja_0 &= 0
\end{aligned}$$

Así, $a_0 = 0$ y $\sum_{i=1}^{n} a_{ij} = 0, \forall j$ y $\sum_{j=0}^{J-1} a_{ij} = 0, \forall i$. Utilizando esto, el lagrangiano (4.18) se convierte en:

$$\begin{aligned} L &= C\sum_{i}^{n} g(\boldsymbol{\xi}_i) + \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 + \sum_{j=0}^{J-1}\sum_{i=1}^{n} a_{ij}\xi_{ij} - \sum_{j=0}^{J-1}\beta_{j0}\sum_{i=1}^{n} a_{ij} - \sum_{j=0}^{J-1}\mathbf{b}_j'\sum_{i=1}^{n} a_{ij}\mathbf{x}_i \\ &= C\sum_{i}^{n} g(\boldsymbol{\xi}_i) + \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 + \sum_{j=0}^{J-1}\sum_{i=1}^{n} a_{ij}\xi_{ij} - \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\ &= C\sum_{i}^{n} g(\boldsymbol{\xi}_i) - \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 + \sum_{j=0}^{J-1}\sum_{i=1}^{n} a_{ij}\xi_{ij} \end{aligned}$$

Pero

$$g(\boldsymbol{\xi}_i) = -\sum_{j=1}^{J-1} y_{ij}\xi_{ij} + ln\left(1 + \sum_{j=1}^{J-1} e^{\xi_{ij}}\right)$$

entonces

$$L = -C\sum_{i}^{n}\sum_{j=1}^{J-1} y_{ij}\xi_{ij} + C\sum_{i}^{n} ln\left(1 + \sum_{j=1}^{J-1} e^{\xi_{ij}}\right) + \sum_{i=1}^{n}\sum_{j=0}^{J-1} a_{ij}\xi_{ij} - \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2$$

De la ecuación (4.24):

$$\frac{e^{\xi_{ij}}}{1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}}} = y_{ij} - \frac{a_{ij}}{C}, \forall i.$$

Tomando logaritmos:

$$\begin{aligned} \xi_{ij} - ln\left(1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}}\right) &= ln\left(y_{ij} - \frac{a_{ij}}{C}\right) \\ \xi_{ij} - ln\left(y_{ij} - \frac{a_{ij}}{C}\right) &= ln\left(1 + \sum_{j'=1}^{J-1} e^{\xi_{ij'}}\right) \end{aligned}$$

Sustituyendo en L y utilizando (4.27) se tiene:

$$
\begin{aligned}
L =& - C \left[\sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) \xi_{ij} \right] + C \sum_{i}^{n} \left[\xi_{ij} - ln \left(y_{ij} - \frac{a_{ij}}{C} \right) \right] - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \left[\sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) \xi_{ij} \right] - \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) \left[ln \left(y_{ij} - \frac{a_{ij}}{C} \right) - \frac{1}{J} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) \right] \\
& - \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) ln \left(y_{ij} - \frac{a_{ij}}{C} \right) + \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) \\
& - \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) ln \left(y_{ij} - \frac{a_{ij}}{C} \right) + \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) \\
& - \frac{C}{J} \sum_{i}^{n} \sum_{j'=0}^{J-1} ln \left(y_{ij'} - \frac{a_{ij'}}{C} \right) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \sum_{i}^{n} \sum_{j=1}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) ln \left(y_{ij} - \frac{a_{ij}}{C} \right) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2 \\
=& - C \sum_{i=1}^{n} G(\frac{\boldsymbol{a}_i}{C}) - \frac{1}{2} \sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2
\end{aligned}
$$

donde

$$
G(\frac{\boldsymbol{a}_i}{C}) = \sum_{j=0}^{J-1} \left(y_{ij} - \frac{a_{ij}}{C} \right) ln \left(y_{ij} - \frac{a_{ij}}{C} \right)
$$

con $\boldsymbol{a}_i = (a_{i0}, \ldots, a_{i(J-1)})'$. El dual de Wolfe (ver Sección 4.2.3) de (4.15) sujeto a (4.16) y (4.17) corresponde a la maximización de L sujeto a (4.19), (4.20) y (4.21).

Lo cual es equivalente a:

$$\text{min: } D = C\sum_{i=1}^{n} G(\frac{\boldsymbol{a}_i}{C}) + \frac{1}{2}\sum_{j=0}^{J-1} \parallel \mathbf{b}_j \parallel^2, \tag{4.28}$$

$$\text{sujeto a: } \sum_{j=0}^{J-1} a_{ij} = 0, \forall i, \tag{4.29}$$

$$\sum_{i=1}^{n} a_{ij} = 0, \forall j \tag{4.30}$$

Después de resolver el problema para los a_{ij}'s, las variables primarias $\mathbf{b}_j$'s y ξ_{ij}'s se obtienen usando (4.22) y (4.27).

Ahora, reemplazando $a_{i(J-1)}$ con $-\sum_{j=0}^{J-2} a_{ij}$, el problema de minimización de (4.28) se convierte en:

$$\text{min: } D = C\sum_{i=1}^{n} G(\frac{\mathbf{a}_i}{C}) + \frac{1}{2}\sum_{j=0}^{J-2} \parallel \mathbf{b}_j \parallel^2 + \frac{1}{2} \parallel \mathbf{b}_{J-1} \parallel^2, \tag{4.31}$$

$$\text{sujeto a: } \sum_{i=1}^{n} a_{ij} = 0, j = 0, \ldots, J-2, \tag{4.32}$$

donde $\mathbf{b}_{J-1} = \sum_{i=1}^{n} a_{i(J-1)}\mathbf{x}_i = -\sum_{i=1}^{n}\left(\sum_{j=0}^{J-2} a_{ij}\right)\mathbf{x}_i$ y

$$\begin{aligned} G(\frac{\mathbf{a}_i}{C}) &= \sum_{j=0}^{J-2}\left(y_{ij} - \frac{a_{ij}}{C}\right) ln\left(y_{ij} - \frac{a_{ij}}{C}\right) + \left(y_{i(J-1)} - \frac{a_{i(J-1)}}{C}\right) ln\left(y_{i(J-1)} - \frac{a_{i(J-1)}}{C}\right) \\ &= \sum_{j=0}^{J-2}\left(y_{ij} - \frac{a_{ij}}{C}\right) ln\left(y_{ij} - \frac{a_{ij}}{C}\right) \\ &+ \left[1 - \sum_{j=0}^{J-2}\left(y_{i(J-1)} - \frac{a_{i(J-1)}}{C}\right)\right] ln\left[1 - \sum_{j=0}^{J-2}\left(y_{i(J-1)} - \frac{a_{i(J-1)}}{C}\right)\right], \end{aligned}$$

utilizando el hecho de que $y_{i(J-1)} = 1 - \sum_{j=0}^{J-2} y_i j$. El lagrangiano para (4.31) es:

$$\tilde{L} = C\sum_{i=1}^{n} G(\frac{\mathbf{a}_i}{C}) + \frac{1}{2}\sum_{j=0}^{J-2} \parallel \mathbf{b}_j \parallel^2 + \frac{1}{2} \parallel \mathbf{b}_{J-1} \parallel^2 - \sum_{j=0}^{J-1}\left[c_j \sum_{i=1}^{n} a_{ij}\right]$$

Ahora,

$$\begin{aligned}\frac{\partial G}{\partial a_{ij}} &= -\frac{1}{C}ln\left(y_{ij}-\frac{a_{ij}}{C}\right)-\frac{1}{C}\frac{\left(y_{ij}-\frac{a_{ij}}{C}\right)}{\left(y_{ij}-\frac{a_{ij}}{C}\right)}+\frac{1}{C}ln\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]\\ &+\frac{1}{C}\frac{\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]}{\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]}\\ &= -\frac{1}{C}ln\left(y_{ij}-\frac{a_{ij}}{C}\right)+\frac{1}{C}ln\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]\end{aligned}$$

Así, las condiciones KKT están dadas por:

$$\begin{aligned}\frac{\partial \tilde{L}}{\partial a_{ij}} &= C\frac{\partial G}{\partial a_{ij}}+\mathbf{b}'_j\mathbf{x}_i-\mathbf{b}'_{J-1}\mathbf{x}_i-\beta_j\\ &= (\mathbf{b}'_j-\mathbf{b}'_{J-1})\mathbf{x}_i-ln\left(y_{ij}-\frac{a_{ij}}{C}\right)+ln\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]-\beta_j\\ &= F_{ij}-\beta_j=0, \quad i=1,\ldots,n; j=0,\ldots,J-2\end{aligned}$$

con $F_{ij}=(\mathbf{b}'_j-\mathbf{b}'_{J-1})\mathbf{x}_i-ln\left(y_{ij}-\frac{a_{ij}}{C}\right)+ln\left[1-\sum_{j=0}^{J-2}\left(y_{ij}-\frac{a_{ij}}{C}\right)\right]$. Dado:

$$i_{up}(j)=argmax_iF_{ij}, j=0,\ldots,J-2$$

$$i_{low}(j)=argmin_iF_{ij}, j=0,\ldots,J-2$$

las condiciones de optimalidad KKT se cumplen en un a_{ij} dado si y solo si:

$$F_{i_{up}(j),j}=F_{i_{low}(j),j} \tag{4.33}$$

Para que los F_{ij}'s estén bien definidos se requiere:

$$\begin{cases}0<a_{ij}<C & \text{si } y_{ij}=1,\\ -C<a_{ij}<0 & \text{si } y_{ij}=0,\end{cases} \tag{4.34}$$

y

$$0 < \sum_{j=0}^{J-2} \left(y_{ij} - \frac{a_{ij}}{C} \right) < 1$$

Si existe un par de índices (i, i') tal que:

$$F_{ij} \neq F_{i'j}$$

para algún j, entonces se pueden ajustar a_{ij} y $a_{i'j}$ manteniendo la restricción (4.32). Para ello, se define:

$$\tilde{a}_{ij}(t) = a_{ij} + t,$$

$$\tilde{a}_{i'j}(t) = a_{i'j} - t,$$

$$\tilde{a}_{i'j''}(t) = a_{i'j''}, \text{ para todos los demás.}$$

Se puede verificar que:

$$\frac{\partial D}{\partial t} = F_{ij} - F_{i'j},$$

donde F_{ij} y $F_{i'j}$ se evalúan en t. Como $F_{ij} \neq F_{i'j}$ en $t = 0$, es posible decrecer D escogiendo t adecuadamente alejado de cero. El algoritmo SMO se presenta a continuación.

Algoritmo SMO:

1. Escoja $a^{(0)}$ que satisfaga las condiciones (4.32) y (4.34).Tome $r = 1$.
2. Si $a^{(r)}$ satisface (4.33) deténgase. Si no, haga:

$$a_{i_{low},j*}(t) = a^{(r)}_{i_{low},j*} + t, \tag{4.35}$$

$$a_{i_{up},j*}(t) = a^{(r)}_{i_{up},j*} - t, \tag{4.36}$$

$$a_{i,j}(t) = a^{(r)}_{ij}, \text{ para los demás } i, j. \tag{4.37}$$

Encuentre el valor de t que minimiza D (4.31).

3. Actualice $a^{(r+1)}$ de acuerdo a (4.35), (4.36)y (4.37). Regrese al paso 2.

Note que el paso 2 es un problema de optimización univariada.

Capítulo 5
METODOLOGÍA Y RESULTADOS

En este capítulo se pretende evaluar el desempeño de la regresión logística con penalidad *Ridge* (RLP) en ocho bases de datos de expresión genética: leukemia, breastcc, colon, prostate, SRBCT, lymphoma, brain y carcinoma. En todas las bases de datos se trabajó con todos los genes. La descripción de las bases de datos utilizadas se presenta a continuación.

5.1. Bases de datos

Leukemia: Esta base de datos contiene los niveles de expresión de n=72 pacientes con dos tipos de leucemia: leucemia linfoblástica aguda (47 casos) y leucemia mielóide (25 casos). Los datos fueron obtenidos con microarreglos oligonucleotidos de Affymetrix. En el preprocesamiento los datos fueron filtrados, estándarizados y tranformados logarítmicamente. Finalmente la base de datos quedo conformada por los valores de expresión de $p = 3571$ genes [7]. Los datos están disponibles en [28].

Breast cancer (BRCA): Esta base de datos contiene los niveles de expresión de 3227 genes de pacientes con cancer de pecho, con uno de tres tipos de tumores: sporadic, BRCA1 and BRCA2 [12]. Los datos preprocesados están disponibles en [7].

Colon: Esta base de datos está compuesta por los niveles de expresión de 40 tejidos de colón con tumor y 22 tejidos de colón normales, medidos en 2000 genes humanos usando la tecnología Affymetrix. Al igual que en las demás bases, los datos fueron estandarizados y transformados logarítmicamente [7]. Los datos están disponibles en

[5].

Prostate: Los datos comprenden los niveles de expresión de 52 muestras de prostata con tumor y 50 muestras de prostata sin tumor, obtenidos de $p = 6033$ genes usando la tecnología Affymetrix. Los datos fueron normalizados, estándarizados y transformados logarítmicamente [7]. Los datos están disponibles en [28].

SRBCT: Esta base de datos contiene la expresión genética para clasificar tumores de pequeñas células azules de niñez (SRBCT) en cuatro clases: neuroblastoma, rhabdomyosarcoma, non-Hodgkin linfoma y la familia de tumores Ewing. Se obtuvieron 63 muestras de tejidos SRBCT utilizando microarreglos cDNA. Cada muestra de tejido está asociada con un perfil de expresión de $p = 2308$ genes, previamente estándarizados con media cero y varianza uno [7]. Los datos están disponibles en [20]

Lymphoma: Esta base de datos contine niveles de expresión de los $k = 3$ más prevalentes linfomas malignos: 42 muestras de gran linfoma difuso Bcell, 9 observaciones de linfoma folicular y 11 casos de leucemia linfotica crónica. El tamaño total de muestras es $n = 62$ y $p = 4026$ genes. Los datos fueron estandarizados y se imputaron los datos faltantes [7]. Los datos están disponibles en [17].

Brain: Esta base de datos contiene $n = 42$ microarreglos con perfiles de expresión de genes de $k = 5$ tumores diferentes del sistema nervioso central: 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/ rhabdoid tumors (AT/RTs), 8 primitive neuro-ectodermal tumors (PNETs) and 4 human cerebella. Los datos fueron obtenidos usando la tecnología Affymetrix. En el preprocesamiento los datos fueron filtrados, estandarizados y transformados logarítmicamente. la base de datos contiene $p = 5597$ genes [7]. Los datos están disponibles en [28].

Carcinoma: Esta base de datos comprende los niveles de expresión de 7463 genes para 18 tejidos normales y 18 carcinomas. Cada arreglo fue estandarizado para tener media cero y varianza uno.

5.2. Procedimientos para estimar el error

Estimación de parámetros: la estimación de parámetros se hizo mediante el algoritmo SMO. La programación del algoritmo fue realizada en el programa estadístico R y fue proporcionada por el profesor del departamento de estadística de la Universidad de Michigan: J.Zhu.

Estimación del error: con el fin de determinar la tasa de error de clasificación del método cada base de datos se dividió aleatoriamente en tres partes de una manera proporcional entre las clases. De esas tres partes, 2/3 se utilizaron como muestra de entrenamiento y 1/3 como muestra de prueba. Este procedimiento se repitió 200 veces y la tasa de error de clasificación consistió en el promedio de las 200 repeticiones. El proceso se realizó para 25 valores de λ (penalidad) variando entre $1/2^{24}$ y $1/2^0 = 1$ y se tomó el valor de λ que reportó el mínimo error de clasificación (ver Figuras 5–1 y 5–2). Las tasas de error de clasificación y sus desviaciones estándar se muestran en la Tabla 5–1.

Inicialmente la tasa de error de clasificación del método se estimó mediante 50 repeticiones observandose tasas de error más bajas, esto es debido a la variabilidad del método de estimación del error, por esta razón es más conveniente realizar un número grande de repeticiones para que los resultados no se vean afectados por la variabilidad.

En la Tabla 5–2 se comparan las tasas de error de clasificación de RLP con las

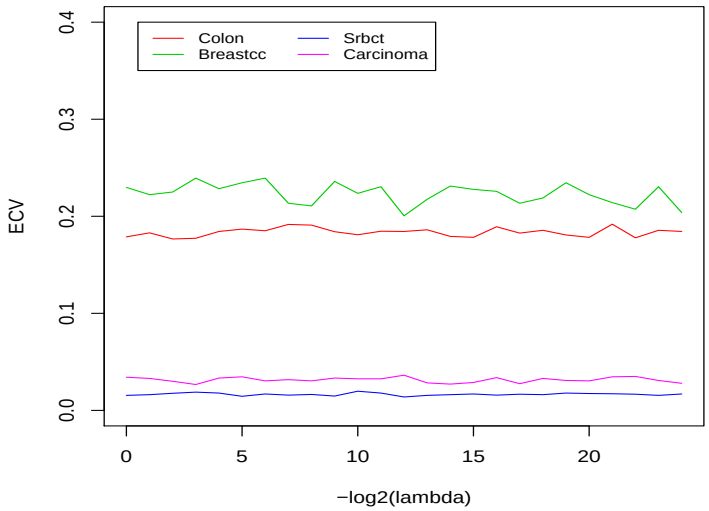

Figura 5–1: Error de clasificación por $2/3 - 1/3$, $0 \leq \lambda \leq 1$

obtenidas con otros métodos. En términos generales RLP tiene un buen desempeño. Para prostate, lymphoma y brain obtiene los más bajos errores de clasificación y para leukemia presenta el segundo error más bajo.

El procedimiento realizado por Zhu y Hastie, RLP con preselección de variables predictoras, exhibe menores errores de clasificación para SRBCT que RLP sin preselección de variables, sin embargo presenta mayores errores de clasificación para leukemia. Es decir, que en algunos casos combinar la RLP con selección de variables mejora el desempeño del clasificador, pero no siempre sucede.

Aunque se ha hecho un numero grande de repeticiones para estimar la tasa de error de clasificación, los programas corren relativamente rápido. Entre mayor sea la cantidad de genes que contenga la base de datos, mayor es el tiempo que tarda el programa en reportar resultados.

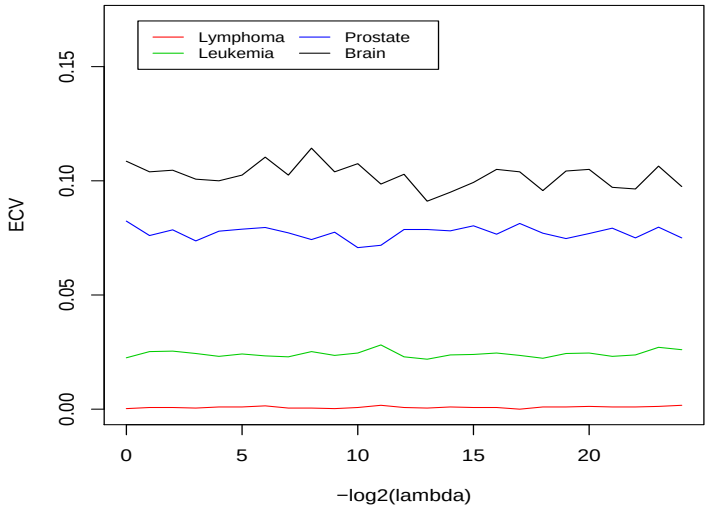

Figura 5–2: Error de clasificación por $2/3 - 1/3$, $0 \leq \lambda \leq 1$

Base de datos	Filas	Columnas	Obs por clase					% mal clasificados	Varianza 200 repet	λ
			1	2	3	4	5			
Colon	62	2000	22	40				17.66	0.0043	.250000
Leukemia	72	3571	47	25				2.19	0.0005	.000122
Prostate	102	6033	50	52				7.07	0.0013	.009765
Carcinoma	36	7457	18	18				2.66	0.0015	.125000
BRCA	22	3226	7	8	7			20.04	0.0165	.000242
Lymphoma	62	4026	42	9	11			0.00	0.0000	7.6e-06
SRBCT	63	2308	23	20	12	8		1.38	0.0006	.000244
Brain	42	5597	10	10	10	4	8	9.11	0.0045	.000122

Tabla 5–1: Porcentaje de mala clasificación de RLP, 200 repeticiones y $0 \leq \lambda \leq 1$

Base de datos	RLP	Ridge-PLS		RLP-Zhu & Hastie		PDA	SVM	k-NN
		k=1	k=2	RLP-UR	RLP-RFE			
Colon	17.66	14.30	15.00	-	-	14.0	13.7	18.2
Leukemia	2.19	5.23	5.18	8.82	2.94	3.3	2.0	3.7
Prostate	7.07	-	-	-	-	-	8.6	11.2
Lymphoma	0.00	-	-	-	-	0.17	1.07	1.12
SRBCT	1.38	-	-	0.00	0.00	1.7	2.3	0.9
Brain	9.11	-	-	-	-	-	22.5	23.0

Tabla 5–2: Porcentaje de mala clasificación de RLP vs Ridge-PLS (ridge-partial least square)[9], RLP-Zhu (RLP con selección de variables)[32], PDA (penalized discriminant analysis)[10, 11], SVM (support vector machines)[24] y k-NN[16]

Capítulo 6
CONCLUSIONES

La regresión logística con penalidad *Ridge* es una buena alternativa cuando se trabaja con bases de datos que tienen más variables que observaciones. Esta estabiliza el problema estadístico, elimina la degeneranción numérica debida a la multicolinealidad y obtiene bajas tasas de error de clasificación en comparación con otros métodos.

El algoritmo SMO modificado por Keerthi y generalizado por Zhu al caso multiclase, permite estimar los parámetros de la RLP de una forma relativamente rápida.

Comparando los resultados obtenidos en este trabajo con los resultados obtenidos por Zhu y Hastie [32] puede verse que aunque RLP no es un método de reducción de la dimensionalidad, y la idea original del método es trabajar con todas las variables, algunas veces la preselección de variables predictoras mejora el desempeño del clasificador.

Como un trabajo futuro podría explorarse la regresión logística con penalidad lasso. Lasso exhibe la estabilidad de la regresión *Ridge* y debido a su tendencia a producir algunos coeficientes que son exactamente cero, da modelos más fácilmente interpretables.

Bibliografía

[1] Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, Vol 71,1–10.

[2] Agresti, A (2002). *Categorical Data Analysis*. 2nd ed. John Wiley & Songs: New Jersey.

[3] Bruce, R. (2002). Identifiability of parameters in MCMC bayesian inference of phylogeny. *Systematic Biology*, 51(5),754–760.

[4] Cessie, N. (1992) Ridge estimators in logistic regression. *Applied Satistics*, 41:191–201.

[5] Colorectal Cancer Microarray Research. [http://microarray.princeton.edu/oncology/]

[6] Cook R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*. John Wiley & Songs: New York.

[7] Dettling, M. and Bühlmann, P(2002) Supervised clustering of genes. *Genome Biology*, 3(12):research0069.1–0069.15.

[8] Eilers, P., Boer, J., Jan van Ommen, G. and Houwelinhen, H (2002). Classification of microarray data with penalized logistic regression. Preprint.

[9] Fort, G and Lambert-Lacroix, S. (2003). Clasification using partial least squares with penalized logistic regression. *IAP Statistics Network*.

[10] Hastie, T and Tibshirani, R.(1995). Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102.

[11] Hastie, T, Tibshirani, R. and Buja, A(1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270.

[12] Hedenfalk, I. et al (2001). Gene expression profiles in hereditary breast cancer. *New Engl J Med*, vol 344, No 8:539–548.

[13] Hoerl, A. E. and Kennard R.W. (1970). Ridge regression application to orthogonal problems. *Technometrics*, 12:69–82.

[14] Hosmer, D. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Songs.

[15] Keerthi, S.S., Duan, K., Sherade, S.K. and Poo, A. N. (2002). A fast dual algorithm for kernel logistic regression. 19^{th} *International Conference on Machine Learning*.

[16] Li, L., Darden, T. A., Weinberg, C. R., Levine, A. J. and Pedersen, L. G. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/knearest neighbor method. *Combinatorial Chemistry & High Throughput Screening*, 4(8):727–739.

[17] Lymphoma/Leukemia Molecular Profiling Project Gateway. [http://llmpp.nih.gov/lymphoma/]

[18] Lokhorst, J. and Turlach, B.A. (1999). Lasso2:An S-plus library to solve regression problems while imposing an L1 constraint on the parameters.

[19] Miller, A. (2002). *Subset Selection in Regression*. 2nd ed. Chapman & Hall: Boca Raton, Florida.

[20] National Human Genome Research Institute: microarray project. [http://www.nhgri.nih.gov/DIR/Microarray/Supplement]

[21] Osborne, M.R. (1999). On the LASSO and its dual. *Journal of the Computational and Graphical Statistics*.

[22] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14. *Microsoft Research*.

[23] Powers, D. (1999). *Statistical Methods for Categorical Data Analysis*. Academic Press.

[24] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S. and Golub, T. R.(2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science*, 98(26):15149–15154.

[25] Seber, G.A.F. (1977). Linear regression analysis. *New York: Willey.*

[26] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients. *J. Urol.*, 16,1076–1083.

[27] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58:267–288.

[28] Whitehead Institute Center for Genomic Research: cancer genomics. [http://www-genome.wi.mit.edu/cancer]

[29] Wenjiang, J. Fu. (1998). Penalized regressions: the bridge versus lasso. *Journal of Computational and graphical Statistics*, Vol 7, Num 3:397–416.

[30] Ying, S. A Tutorial on Logistic Regression. [http://www.ats.ucla.edu/stat/sas/library/logistic.pdf].

[31] Zangwill, W (1969). *Nonlinear Programming: A Unified Approach.* PRENTICE-HALL,INC.,ENGLEWOOD CLIFFS, N. J..

[32] Zhu, J. and Hastie, T. (2004). Classification of gene microarray by penalized logistic regression. *Biostat*, 5:427–443.