CAPÍTULO 5

REGRESIÓN CON VARIABLES CUALITATIVAS

5.1 Regresión con variables predictoras cualitativas.

Frecuentemente se considera que entre las variables predictoras, que explican el comportamiento de la variable de respuesta, hay algunas que son cualitativas o categóricas. Por ejemplo, si en una empresa se trata de explicar el salario de un empleado hay muchas variables predictoras a considerar algunas de ellas cuantitativas y otras cualitativas. Entre las variables cuantitativas estarán años de experiencia en la empresa, años de educación, edad, etc. y entre las variables cualitativas estarán el sexo del empleado, estado civil, jerarquía del empleado, etc.
Cuando una variable cualitativa asume solamente dos valores es llamada variable indicadora, variable binaria o variable "dummy". Estas variables son codificadas numéricamente con 0's y 1's.
Algunas veces la variable cualitativa puede asumir más de dos valores. Por ejemplo, la variable Opinión: "A favor", "Indeciso", "En contra". Se podría codifucar los valores como 0, 1 y 2 pero esto estaría implicando una suposición de ordenamiento y además implicaría que el efecto de cambiar de "A favor" a "Indeciso" es lo mismo que cambiar de "Indeciso" a "En contra" (o sea se está suponiendo igual espaciamiento). Ambas suposiciones no son justificables. Una mejor alternativa es definir dos variables indicadoras

A_1=1 "A favor", 0 en otro caso
A_2= 1 "En contra", 0 en otro caso

La combinación A_1=1 y A_2=0 representaría que la variable Opinión asume el valor "A favor", la combinación A_1=0 y A_2=1 representaría que la variable Opinión asume el valor "En contra". Usar una tercera variable es redundante, puesto que los indecisos pueden ser representados por A_1=A_2=0. Estas variables cualitativas, donde el orden es irrelevante, son llamadas más propiamente variables nominales. Si la variable nominal asume k valores distintos habría que usar k+1 variables indicadoras para representar todos sus valores. Es decir, el número de variables predictoras en el modelo se incrementaría en k. Si se tuviera un gran número de variables predictoras nominales el modelo de regresión se volvería bastante complejo ya que tendría un gran número de parámetros que estimar.
Las variables cualitativas donde el orden si interesa se le conoce como variables ordinales y en ese caso es más frecuente codificar la variable como una secuencia ordenada de números enteros.

En un problema de regresión debe haber por lo menos una variable predictora cuantitativa. Si todas las variables predictoras fueran cualitativas entonces el problema se convierte en uno de **diseños experimentales.**

5.1.1 Regresión con una sola variable predictora cualitativa

Consideremos un modelo de regresión con una sola variable cualitativa A y una variable cuantitativa X. Es decir,

$$Y=\beta_0+\beta_1X+\beta_2A+\varepsilon \qquad (5.1)$$

Notar que si A=0 se obtiene el modelo lineal simple

$$Y=\beta_0+\beta_1 X + \varepsilon \qquad (5.2)$$

Y que si A=1 se obtiene el modelo

$$Y=(\beta_0+\beta_2)+\beta_1 X+\varepsilon \qquad (5.3)$$

Notar que las lineas estimadas de los modelos (5.2) y (5.3) serán paralelas (igual pendiente). El valor estimado de β_2 representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable "dummy".

Ejemplo 1. En el conjunto de datos **bajopeso**, disponible en la página de internet del texto, se trata de relacionar el peso de lo recién nacidos con los pesos de sus madres y la condición de fumar de las mismas. El conjunto de datos contiene 189 observaciones y será tratado en forma más completa más adelante.

Solución. Considerando que la variable fumar asume el valor 0 si la persona no fuma y 1 si la persona fuma se obtiene los siguientes resultados

```
> l2<-lm(pbebe~pmama+fuma)
> summary(l2)

Call:
lm(formula = pbebe ~ pmama + fuma)

Residuals:
     Min       1Q   Median       3Q      Max
-2030.90  -445.69    29.16   521.76  1967.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835   <2e-16 ***
pmama          4.237      1.690   2.507   0.0130 *
fuma        -272.081    105.591  -2.577   0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 707.8 on 186 degrees of freedom
Multiple R-Squared: 0.06777,    Adjusted R-squared: 0.05775
F-statistic: 6.761 on 2 and 186 DF,  p-value: 0.001464
```

Notar que el R^2 es bajísimo. El coeficiente de regresión estimado de fuma es –272 y significa que si la mama fuma en promedio el peso del bebe disminuirá en 272 gramos.

Podemos hacer la regresión por grupos. Es decir, una regresión para los madres que no fuman y otras para las que si fuman. Se obtienen los siguientes resultados.

Para madres no fumadoras:

```
> l3<-lm(pbebe0~pmama0)
> summary(l3)
```

Call:
lm(formula = pbebe0 ~ pmama0)

Residuals:
    Min   1Q Median   3Q   Max
-2029.87 -550.86   28.23 551.78 1976.84

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578   326.583   7.197 7.21e-11 ***
pmama0   5.387   2.439   2.209   0.0292 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 740.2 on 113 degrees of freedom
Multiple R-Squared: 0.04139,   Adjusted R-squared: 0.03291
F-statistic: 4.88 on 1 and 113 DF, p-value: 0.02919

Para madres fumadoras:

```
> summary(l4)

Call:
lm(formula = pbebe1 ~ pmama1)

Residuals:
    Min      1Q  Median      3Q     Max
-2038.8  -414.4    35.0   473.6  1490.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2391.962    301.229   7.941 1.98e-11 ***
pmama1         2.965      2.274   1.304    0.196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 656.5 on 72 degrees of freedom
Multiple R-Squared: 0.02307,    Adjusted R-squared: 0.0095
F-statistic:   1.7 on 1 and 72 DF,  p-value: 0.1964
```

Notar que ambos casos los R^2 son más bajos que el R^2 anterior. Los plots son mostrados en la figura 5.1

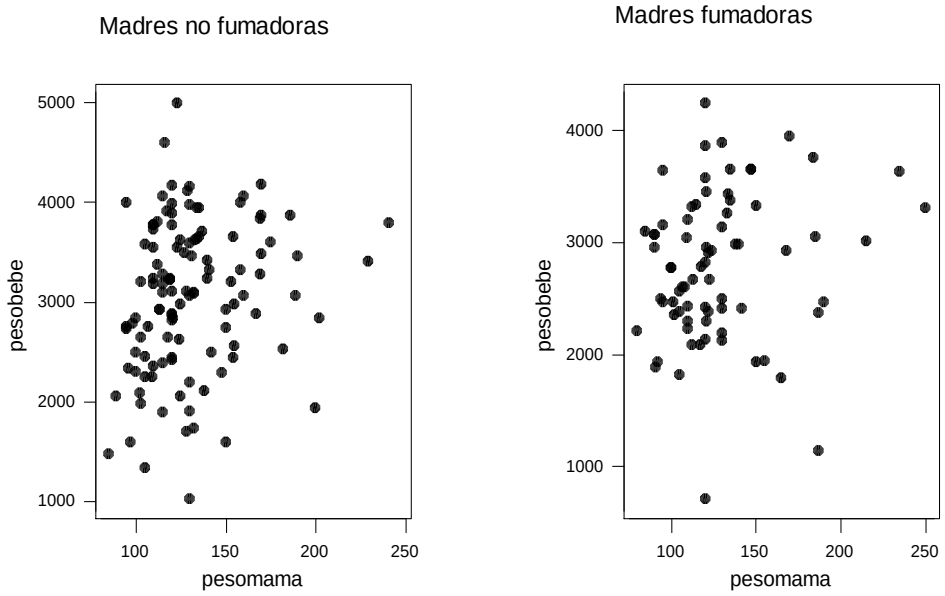

Figura 5.1. Plots de la relación pesobebe versus pesomama según la condición de fumar de la madre

En ambos plots se puede ver que no parece haber relación entre el peso del bebe y peso de la madre aunque esto es más evidente para las madres fumadoras. Los "outliers" parecen afectar más la regresión del peso bebe versus peso mama entre las madres no fumadoras.

Si se desea comparar las pendientes de las línea de regresión de los dos grupos se puede usar una prueba de t similar a la prueba de comparación de dos medias y asumiendo que hay homogeneidad de varianza. También se puede usar una prueba de F parcial o de t para probar la hipótesis Ho: $\beta_3=0$ en el siguiente modelo

$$Y=\beta_0+\beta_1 A+\beta_2 X+\beta_3 AX +e$$

Cuando la hipótesis nula no es rechazada se concluye que la pendiente de regresión de ambos grupos son iguales y el uso del modelo 5.1 seria adecuado.
Si no hubiera igualdad de varianza de los dos grupos, habría que usar una prueba de t aproximada similar al problema de Behrens-Fisher. La prueba está definida por

$$t=\frac{\hat{\beta}_1-\hat{\beta}_2}{\sqrt{\frac{s_1^2}{Sxx_1}+\frac{s_2^2}{Sxx_2}}}$$

Donde, $\hat{\beta}_1$ and $\hat{\beta}_2$ son los pendientes estimadas de cada línea de regresión y s_1^2 y s_2^2 son las estimaciones de la varianza del error en cada modelo.
Los grados de libertad de la prueba t se aproximan por

$$gl = \frac{(c_1 + c_2)^2}{\frac{c_1^2}{m-1} + \frac{c_2^2}{n-1}}$$

con $c_1 = \frac{s_1^2}{m}$ y $c_2 = \frac{s_2^2}{n}$.

Donde m y n son los grados de libertad de la suma de cuadrados del error en cada modelo.

5.1.2 Comparando las líneas de regresión de más de dos grupos.

Supongamos que se tiene una variable predictora continua X para explicar el comportamiento de Y en tres grupos. Luego hay tres modelos de regresión que se pueden comparar. Estos son:

i) $Y=\beta_{01}+\beta_{11}X+\varepsilon$

ii) $Y=\beta_{02}+\beta_{12}X+\varepsilon$

iii) $Y=\beta_{03}+\beta_{13}X+\varepsilon$

Para relacionar las lineas de regresión hay que introducir 3 variables "dummy" para identificar los grupos G_1, G_2, y G_3 y 3 variables adicionales $Z_1=G_1X$, $Z_2=G_2X$, y $Z_3=G_3X$. Otra alternativa sería usar solo dos variables "dummy". Hay 4 posibilidades que podrían ocurrir:
a) Que las lineas se intersecten en un punto cualquiera, ya que tendrían diferente intercepto y pendiente. En este caso se ajusta el modelo $Y=\beta_{01}G_1+\beta_{11}Z_1+ \beta_{02}G_2+\beta_{12}Z_2+ \beta_{03}G_3+\beta_{13}Z_3+\varepsilon$.
Usando dos variables "dummy" este modelo sería $Y=\beta_0+\beta_1X+ \beta_{01}G_1+ \beta_{02}G_2+\beta_{11}Z_1+\beta_{12}Z_2+\varepsilon$.
b) Que las lineas sean paralelas (homogeneidad de pendientes). En este caso se ajusta el modelo $Y=\beta_{01}G_1+ \beta_{02}G_2+ \beta_{03}G_3+\beta X+\varepsilon$
c) Que las lineas tengan el mismo intercepto con el eje Y pero distintas pendientes (homogeneidad de interceptos). En este caso se ajusta el modelo $Y=\beta_0+\beta_{11}Z_1+\beta_{12}Z_2+\beta_{13}Z_3+\varepsilon$
d) Que las tres lineas coincidan. En este caso se ajusta el modelo $Y=\alpha+\beta X+\varepsilon$

Para probar la hipótesis Ho: el modelo satisface b) o c) o d) versus
Ha: el modelo satisface a)

Se usa una prueba de F parcial dada por

$$F_m=[(SSE_m-SSE_a)/(gl_m-gl_a)]/[SSE_a/gl_a]$$

Donde m, representa los modelos b,c, o d, y gl_m y gl_a representan los grados de libertad del error del modelo m y del modelo a, respectivamente. La F parcial se distribuye como una f con (gl_m-gl_a, gl_a) grados de libertad.

5.2 Regresión Logística

Consideraremos ahora que la variable de respuesta, Y, es una del tipo binario y que se tiene p variables predictoras x's, las cuales son consideradas aleatorias. Es decir, que el conjunto de

datos consiste de una muestra de tamaño $n=n_1+n_2$, donde n_1 observaciones son de una clase C_1 y n_2 son de una clase C_2 . Así, para cualquier observación $\mathbf{x}_j$ la variable de respuesta Y es igual a 1 si $\mathbf{x}_j$ es de la clase C_1 , que contiene las observaciones donde el evento que estamos interesados se cumple. Mientras que Y es igual a 0 si $\mathbf{x}_j$ pertenece a la clase C_2 .

Ejemplo 2: El conjunto de datos **bajopeso** contiene los pesos de 189 bebes recién nacidos. Para determinar si el niño es de bajo peso (menos de 2500 gramos) o no lo es, o sea **bajopeso**=1 si peso bebe<2500 y **bajopeso**=0 en otro caso, se han medido las siguientes variables predictoras

Edad: edad de la madre
Pesomama: peso de la madre en su último período muestral
Raza: raza de la madre:1=blanca, 2=negra, 3=otro
Fuma*:* 0 ssi la madre no fuma, 1 si lo hace.
Prematur: número de partos prematuros de la madre
Hiperten: 0 si la madre no sufre de hipertensión, 1 si sufre
Uterirrit: *0 si no tiene utero irritado, 1 si lo tiene.*
Chequeos: número de visitas al médico en los tres primeros meses del embarazo.

Suponiendo que **pesomama** es la variable predictora más importante, podriamos explorar su relación con **bajopeso**. Haciendo un plot de **bajopeso** versus **pesomama** y ajustando una línea de regresión lineal simple se obtiene la siguiente figura

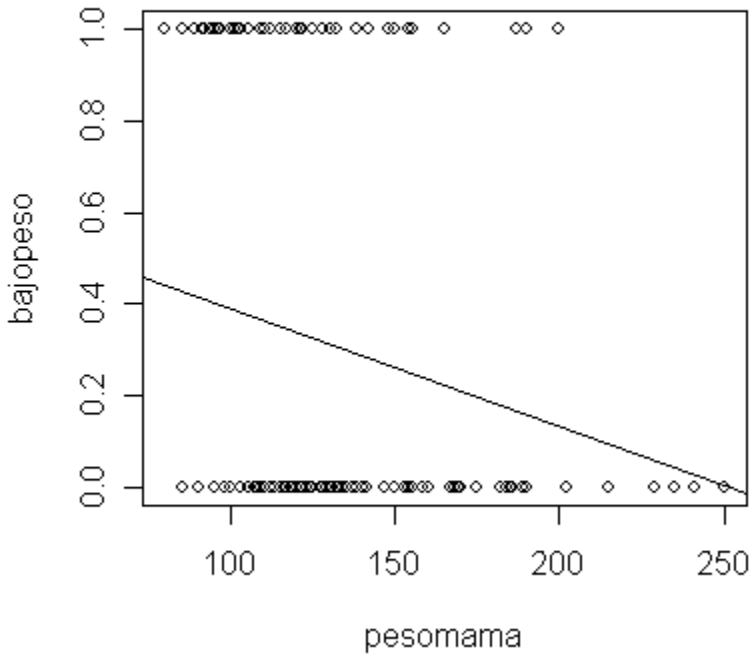

Figura 5.2. Regresión lineal estimada de los datos del ejemplo 2.

Como se puede ver es imposible que la linea de regresión represente la tendencia de los puntos. Además, la linea de regresión puede predecir valores de **bajopeso** que no son necesariamente 0 y

1, lo cual es totalmente ilógico. Asímismo, la suposición de varianza constante para la variable de respuesta no se cumple, como lo muestra el plot de residuales de la siguiente figura.

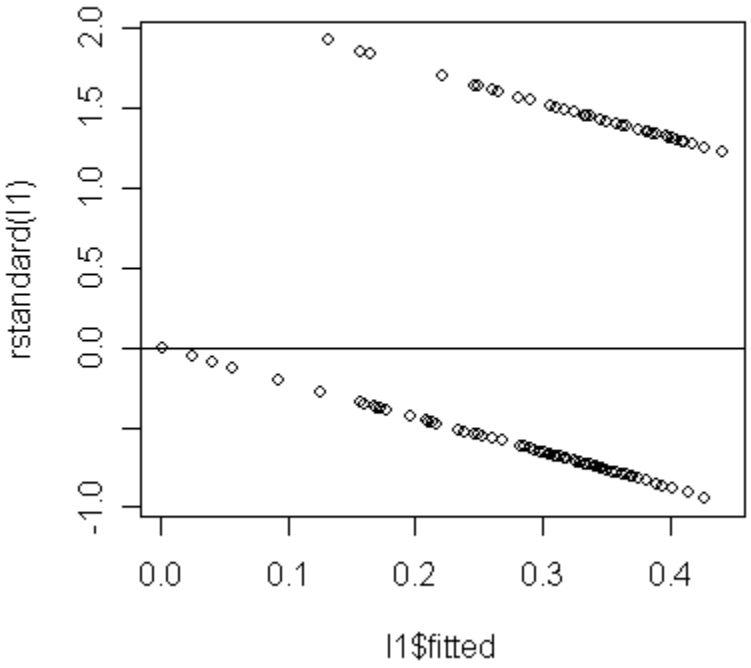

Figura 5.3. Plot de los residuales correspondiente a la regresión estimada del ejemplo 2.

Se podría usar mínimos cuadrados reponderados para remediar esta situación pero aún asi conseguir predicciones 0 y 1 usando el modelo lineal sería imposible. En lugar del modelo de regresión lineal es más conveniente modelar la probabilidad de que la variable de respuesta asuma los valores 0 y 1 basado en las mediciones de las variables predictoras.

Notar que una curva en forma de S ajustaría bien los datos. Por otro lado, existe un modelo bien conocido en crecimiento poblacional cuya curva tiene esta forma y este modelo es llamado el modelo logístico y el cual se muestra en la figura 5.4. Propiamente, se ha graficado $f(x) = \frac{1}{1+e^{-x}}$, -10<x<10. Esta curva también puede ser considerada como la gráfica de la distribución acumulada correspondiente a la densidad logística $p(x) = \frac{e^{-x}}{(1+e^{-x})^2}$

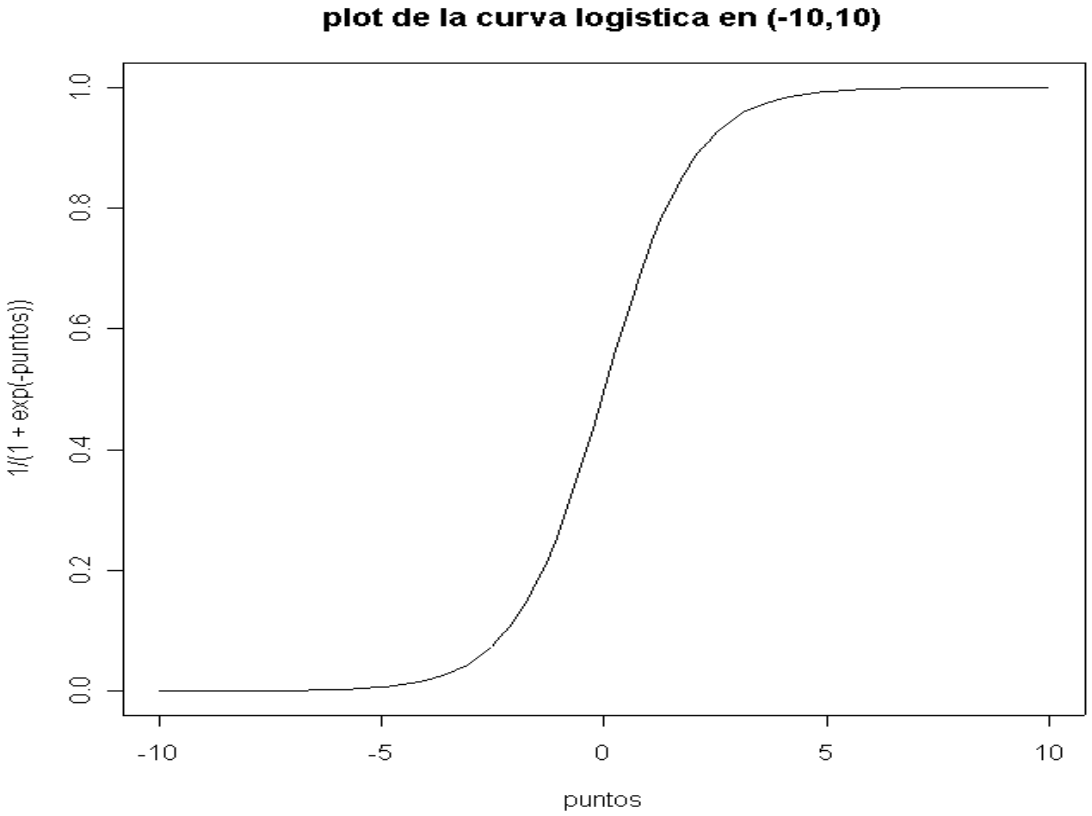

Figura 5.4. Plot de la función de distribución logistica acumulada en el intervalo (-10,10).

Algunas veces la variable de respuesta viene dada en forma agrupada en g grupos, con n_i observaciones en el i-ésimo grupo, como en el siguiente ejemplo.

Ejemplo 3. Se trata de relacionar la variable Y: el estudiante aprueba o no aprueba un examen con la variable X:horas de estudio para el examen. Para ello se recolecta información de los resultados de 430 estudiantes que tomaron el examen. La siguiente tabla resume dichos resultados.

X:Horas de estudio para el examen	Número de estudiantes	Estudiantes que aprueban el examen	Proporcion que aprueban
0	20	1	.05
1	50	5	.10
2	80	25	.3125
3	120	40	.3333
4	90	45	.5000
5	40	30	.7500
6	20	17	.8500
7	10	9	.9000

En total 172 son de la clase 1: aprobar el examen y 258 de la clase 0: fracasar el examen. Los estudiantes se agruparon en 8 grupos de acuerdo a sus horas de estudio.

Sea $f(\mathbf{x}/C_i)$ (i=1,2) la función de densidad del vector aleatorio p-dimensional **x** en la clase C_i, en el modelo logístico se asume que

$$\log(\frac{f(\mathbf{x}/C_1)}{f\ (\mathbf{x}/C_2)}) = \alpha + \boldsymbol{\beta}'\mathbf{x} \qquad (5.4)$$

Aquí $\boldsymbol{\beta}$ es un vector de p parámetros y α representa el intercepto.

Notar que si las variables **x** en cada clase se distribuyen normalmente con igual matriz de covarianza Σ entonces se satisface la suposición (5.4) ya que

$$\log(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}) = (\mathbf{u_1}-\mathbf{u_2})'\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-1/2(\mathbf{u_1}+\mathbf{u_2})) \qquad (5.5)$$

En este caso $\alpha = -(\mathbf{u_1}-\mathbf{u_2})'\boldsymbol{\Sigma}^{-1}(\mathbf{u_1}+\mathbf{u_2})/2$ y $\boldsymbol{\beta} = (\mathbf{u_1}-\mathbf{u_2})'\boldsymbol{\Sigma}^{-1}$.
La suposición (5.4) se cumple también para otros tipos de distribuciones distintas de la normal multivariada tales como distribuciones de Bernoulli, y mezclas de éstas.
Por otro lado, sea p=P(Y=1/**x**) la probabilidad a posteriori de que Y sea igual a 1 para un valor observado de **x**, entonces haciendo uso de probabilidad condicional se tiene que:

$$\frac{p}{1-p} = \frac{\frac{P\{Y=1\}f(\mathbf{x}/y=1)}{f(\mathbf{x})}}{\frac{P\{Y=0\}f(\mathbf{x}/y=0)}{f(\mathbf{x})}} = \frac{\pi_1 f\ (\mathbf{x}/C_1)}{\pi_2 f\ (\mathbf{x}/C_2)} \qquad (5.6)$$

donde π_i representa la *probabilidad a priori* de que **x** pertenezca a la clase C_i.. La expresión $\frac{p}{1-p}$ es llamado la razón de apuestas (***odds ratio***). Tomando logaritmos en ambos lados de (5.6) se obtiene

$$\log(\frac{p}{1-p}) = \log(\frac{\pi_1}{\pi_2}) + \log\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}$$

Usando la suposición (5.4), la ecuación anterior puede ser escrita como

$$\log(\frac{p}{1-p}) = \alpha + \beta'\mathbf{x} \qquad (5.7)$$

y $\log(\frac{p}{1-p})$ es llamado la transformacion ***logit***.
Despejando p de la expresión anterior se obtiene

$$p = \frac{\exp(\alpha+\beta'\mathbf{x})}{1+\exp(\alpha+\beta'\mathbf{x})} \qquad (5.8)$$

La ecuación (5.8) representa el modelo de la regresión logística, que fue introducida en 1944 por J. Berkson.

Un coeficiente b_k en el modelo de regresión logística estimado representa el cambio promedio de la logit función cuando la variable X_k cambia en una unidad adicional asumiendo que las otras variables permanecen constantes.
También se puede considerar que $\exp(b_k)$ es una razón de cambio de la razón de apuestas cuando X_k varia en una unidad adicional. Si X_k es binaria entonces $\exp(b_k)$ es el cambio en la razón de apuestas cuando ella asume el valor 1.
Cuando el modelo tiene una sola variable predictora, que además es binaria entonces existe una relación entre la regresión logística y el análisis de una tabla de contingencia 2 X 2.

5.2.1 Estimación del modelo logístico.

El método más usado para estimar α y $\boldsymbol{\beta}$ es el método de máxima verosimilitud. Dada una observación **x**, las probabilidades de que ésta pertenezca a las clases C_1 y C_2 son :

$$P(C_1/\mathbf{x}) = \frac{\exp(\alpha + \beta'\mathbf{x})}{1+\exp(\alpha + \beta'\mathbf{x})} \qquad (5.9)$$

$$P(C_2/\mathbf{x}) = 1 - P(C_1/\mathbf{x}) = \frac{1}{1+\exp(\alpha + \beta'\mathbf{x})} \qquad (5.10)$$

respectivamente.

Considerando una muestra de tamaño $n=n_1+n_2$ y un parámetro binomial p igual a $\exp(\alpha+\boldsymbol{\beta}'\mathbf{x})/(1+\exp(\alpha+\boldsymbol{\beta}'\mathbf{x})$ la función de verosimilitud es de la forma

$$L(\alpha,\boldsymbol{\beta}) = \prod_{i=1}^{n_1} \frac{\exp(a+\mathbf{x_i}'\boldsymbol{\beta})}{1+\exp(\alpha+\mathbf{x_i}'\boldsymbol{\beta})} \cdot \prod_{j=n_1+1}^{n} \frac{1}{1+\exp(\alpha+\mathbf{x_j}'\boldsymbol{\beta})} \qquad (5.11)$$

asumiendo que las primeras n_1 observaciones son de la clase C_1 y las restantes son de la clase C_2. Los estimados $\tilde{\alpha}$ y $\tilde{\boldsymbol{\beta}}$ son aquellos que maximizan la función anterior y son encontrados aplicando métodos iterativos tales como Newton-Raphson (SAS) o minimos cuadrados reponderados iterativos (MINITAB, R/S-Plus).

La solución de la ecuación de verosimilitud puede no ser única si existe una marcada separación entre las dos clases.

Otra forma de hacer la estimación es como sigue: Los parámetros α y $\boldsymbol{\beta}$ pueden ser estimados haciendo la regresión lineal múltiple de logit($\hat{p}$) versus $x_1, x_2, \ldots x_p$. Usando los resultados de la sección 4.2 para aproximación de la varianza de una transformación se tiene que

$$Var[\ln(\frac{\hat{p}}{1-\hat{p}})] \cong [\frac{1}{p(1-p)}]^2 \frac{p(1-p)}{n_1} = \frac{1}{n_1 p(1-p)} \qquad (5.12)$$

Como p=p(**x**) se llega a un problema donde la varianza no es constante y se puede usar mínimos cuadrados ponderados con pesos $w_i(\mathbf{x})=n_1 \hat{p}(\mathbf{x})(1-\hat{p}(\mathbf{x}))$ para estimar los parámetros α y $\boldsymbol{\beta}$ del modelo logístico.

La regresión logística es un caso particular de los modelos lineales generalizados (GLM) propuesto por Nelder y Wedderburn (1972). Los modelos lineales generalizados extienden los modelos lineales en dos sentidos: Primero, con la especificación de una **función link** que relaciona el esperado de la variable de respuesta con las predictoras lineales y segundo con la especificación de una función de distribución de los errores que es distinta de la Gaussiana. La forma de un modelo lineal generalizado es

$$\eta[E(y)]= \alpha+\boldsymbol{\beta}\mathbf{x}$$

donde $\eta(.)$ es la función link. En un modelo de regresión clásico, $\eta(t)=t$ y la distribución de los errores es Gaussiana o Normal. En la regresión logística $\eta(t)=\log(t/(1-t))$ y la distribución de los errores es binomial. El modelo de regresión viene dado por $E(y)= \eta^{-1}(\alpha+\boldsymbol{\beta}\mathbf{x})$

En **MINITAB** el menú de **Regresión** contiene tres tipos de regresión logística: regresión logística binaria (aplicada a dos clases), regresión logística ordinal (si hay mas de dos clases) y regresión logística nominal (si hay mas de dos clases no ordenadas). Para ajustar un modelo logístico en SAS se usa el procedimiento **LOGISTIC,** mientras que en **R** se usa el procedimiento **glm** (modelos lineales generalizados) con la opción family=binomial. Aquí **family** representa el tipo de distribución de los errores.

5.2.2 Medidas de Confiabilidad del Modelo

Las siguientes son unas medidas que cuantifican el nivel de ajuste del modelo logístico al conjunto de datos:

a) La Devianza Residual: Es similar a la suma de cuadrados del error de la regresión lineal y se define como el negativo de dos veces la función de verosimilitud maximizada. Para los casos cuando la variable de respuesta Y no está agrupada se tiene que:

$$D = -2\{\sum_{i:y_i=1}^{n} \log(\hat{p}_i) + \sum_{i:y_i=0}^{n} \log(1-\hat{p}_i)\}$$

Donde $\hat{p}_i$ es el valor estimado de la ecuación (5.8). D es equivalente a la prueba de razón de verosimilitud para probar la validez del modelo logístico. El estadístico D se distribuye como una Ji-Cuadrado con n-p-1 grados de libertad, donde p es el número de variables predictoras. Si D es mayor que una Ji-Cuadrado con n-p-1 grados de libertad para un nivel de significación dado, entonces el modelo logístico no es confiable.

b) El Pseudo-R^2. Se han propuesto versiones similares al R^2 de la regresión lineal. Aqui definimos el propuesto por McFadden

$$Pseudo - R^2 = 1 - \frac{Devianza\,\mathrm{Re}\,sidual}{Devianza.Nula}$$

donde la Devianza Nula es la Devianza considerando solamente el intercepto y que se distribuye como una Ji-Cuadrado con n-1 grados de libertad. Para hallar la Devianza Nula se hace una

regresión logística considerando que hay una sola variable predictora cuyos valores son todos unos. Un Pseudo-R^2 mayor de .3 es considerado como aceptable.

c) El Criterio de Información de Akaike (AIC): Se define por

$$AIC=D+2(p+1)$$

Donde p es el número de variables predictoras. Un modelo es mejor que otro si su AIC es más pequeño.

d) La Prueba de Bondad de Ajuste de Hosmer-Lemeshov. En esta prueba los valores ajustados son agrupados en g grupos. La prueba es una del tipo χ^2 y se define por

$$C = \sum_{i=1}^{g} \frac{(O_i - n'_i\ \overline{p}_i)^2}{n'_i\ \overline{p}_i(1-\overline{p}_i)}$$

donde g es el número de grupos de los valores ajustados (g varia entre 6 y 10), n'_i es el número de observaciones en el i-ésimo grupo. O_i es la suma de las y's en el i-ésimo grupo y $\overline{p}_i$ es el promedio de las proporciones estimadas $\hat{p}_i$ del evento que está siendo considerando en el i-ésimo grupo.
Si C es mayor que χ^2_α con g-2 grados de libertad entonces se concluye que el modelo logístico no es adecuado.

5.2.3 Estadisticas Influenciales para regresión logística

Existen varios tipos de residuales que permiten cotejar si una observación es influencial o no para la regression logística.

a) Residuales de Pearson: Están definidos por

$$r_i = \frac{y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}}$$

donde, si los valores de la variable de respuesta están agrupadas, y_i representa el número de veces que y=1 entre las m_i repeticiones de X_i . Si los datos no están agrupados m_i=1 para todo i.

El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así un residual de Pearson en valor absoluto mayor que 2 indica un dato atípico.

b) Residuales de Devianza: Están dados por

$D_i = -\sqrt{2\,|\log(1-\hat{p}_i)|}$ si y_i=0 y por $D_i = \sqrt{2\,|\log(\hat{p}_i)|}$ si y_i=1.

Si el residual de devianza es mayor que 2 en valor absoluto entonces la observación correspondiente es atípico. Estos son los residuales dados por R.

Ejemplo 4. Aplicar modelos de regresión logística a los datos del ejemplo 2 y basados en la medidas de bondad de ajuste seleccionar el mejor modelos entre ellos.

```
> # Haciendo la regresion logistica simple con la predictora pesomama
> logis1<-glm(bajopeso~pesomama,data=pesobebe,family=binomial)
> summary(logis1)

Call:
glm(formula = bajopeso ~ pesomama, family = binomial, data = pesobebe)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
pesomama    -0.01406    0.00617  -2.279   0.0227 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 228.69  on 187  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 4

> #Haciendo la regresion logistica multiple usando todas las variables predictoras
> logis2<-glm(bajopeso~.,data=pesobebe,family=binomial)
> summary(logis2)

Call:
glm(formula = bajopeso ~ ., family = binomial, data = pesobebe)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8832  -0.8178  -0.5574   1.0288   2.1451

Coefficients:
              Estimate   Std. Error      z value Pr(>|z|)
(Intercept)  -0.078975    1.276254  -0.062  0.95066
edad         -0.035845    0.036472  -0.983  0.32569
pesomama     -0.012387    0.006614  -1.873  0.06111 .
raza          0.453424    0.215294   2.106  0.03520 *
fuma          0.937275    0.398458   2.352  0.01866 *
prematuros    0.542087    0.346168   1.566  0.11736
hipertensio   1.830720    0.694135   2.637  0.00835 **
uteroirrit    0.721965    0.463174   1.559  0.11906
chequeos      0.063461    0.169765   0.374  0.70854
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.19  on 180  degrees of freedom
AIC: 222.19

Number of Fisher Scoring iterations: 4

> #Haciendo otra vez la regresion logistica incluyendo solo las variables mas significativas
> logis3<-glm(bajopeso~pesomama+raza+fuma+hipertensio,data=pesobebe,family=binomial)
> summary(logis3)

Call:
glm(formula = bajopeso ~ pesomama + raza + fuma + hipertensio,
    family = binomial, data = pesobebe)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7988  -0.8865  -0.5847   1.0997   2.2503

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.357536   1.010584  -0.354  0.72350
pesomama     -0.015354   0.006523  -2.354  0.01858 *
raza          0.489555   0.207324   2.361  0.01821 *
fuma          1.080020   0.383735   2.814  0.00489 **
hipertensio   1.744272   0.687563   2.537  0.01118 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 211.04  on 184  degrees of freedom
AIC: 221.04.
Number of Fisher Scoring iterations: 4
```

Observando el valor de la devianza residual y del AIC el tercer modelo seria el mejor modelo. Notar que la Devianza Residual=211.04 el cual habría que compararlo con una Ji-Cuadrado con 184 grados de libertad para un nivel de significación dado. Usando un nivel de significación del 5%, la Ji-Cuadrado da 216.64 que es mayor que la Devianza Residual. En consecuencia, los datos parecen ajustarse a un modelo logístico. El pseudo-R^2 da .1007.

```
pihat=logis3$fit
pihatc=cut(pihat,br=c(0,quantile(pihat,p=seq(.1,.9,.1)),1))
table(pihatc)
pihatc=cut(pihat,br=c(0,quantile(pihat,p=seq(.1,.9,.1)),1),labels=F)
table(pihatc)
E=matrix(0,nrow=10,ncol=2)
O=matrix(0,nrow=10,ncol=2)
```

```
for(j in 1:10){
 E[j,2]=sum(pihat[pihatc==j])
E[j,1]=sum((1-pihat)[pihatc==j])
O[j,2]=sum(pesobebe$bajopeso[pihatc==j])
 O[j,1]=sum((1-pesobebe$bajopeso)[pihatc==j]) }
>sum((O-E)^2/E)
[1] 5.744907
> 1-pchisq(sum((O-E)^2/E),8)
[1] 0.6757812
```

Viendo el "p-value" de la prueba de χ^2 , se concluye que hay suficiente evidencia estadistica para aceptar que el tercer modelo satisface el modelo logistico.

De acuerdo a los residuales de Pearson las siguientes observaciones pueden ser influenciales

```
> y=pesobebe$bajopeso
>pihat=logis3$fit
>rp=(y-pihat)/sqrt(pihat*(1-pihat))
> rp[abs(rp)>2]
      13       132       147       152       155       170       183
-2.010543  2.539516  3.402709  2.048327  2.700368  2.178067  2.345670
```

De acuerdo a los residuales de devianza las siguientes observaciones pueden ser influenciales

```
>r1=sqrt(2*abs(log(pihat[y==1])))
>r2=-sqrt(2*abs(log(1-pihat[y==0])))
>rd=c(r2,r1)
>rd[abs(rd)>2]
    132      147      155
2.004045 2.250326 2.056837
```

Ejemplo 4. Aplicar regresión logistica a los datos de la tabla del ejemplo 3.

```
> horas
[1] 0 1 2 3 4 5 6 7
> # número de estudiantes que aprueban el examen  por hora de studio
>est.aprob=c( 1, 5, 25, 40, 45, 30, 17,  9)
> # número de estudiantes que fracas an el examen por hora de studio
>est.frac=c(19,45,55,80,45,10,3,1)
> logis4=glm(cbind(est.aprob,est.frac)~horas,family=binomial)
> summary(logis4)

Call:
glm(formula = cbind(est.aprob, est.frac) ~ horas, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.8166  -0.6576  -0.1846   0.4510   1.7415

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.59099    0.29779  -8.701  < 2e-16 ***
horas        0.68489    0.08474   8.082 6.35e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.0406  on 7  degrees of freedom
Residual deviance:  5.4156  on 6  degrees of freedom
AIC: 38.111

Number of Fisher Scoring iterations: 4
```

Comparando la Devianza residual que es 5.41 con el valor de la Ji-Cuadrado al 5% de significacion y 6 grados de libertad que resulta 12.59, se concluye de que hay evidencia de que los datos ajustan bien a un modelo logístico. El pseudo-R^2 resulta ser bastante bueno .9417.

5.2.4 Uso de la regresión logística en Clasificación:

Para efectos de clasificación la manera más facil de discriminar es considerar que si p>0.5 entonces la obervación pertenece a la clase que uno está interesado. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Métodos alternos son:

a) Plotear el porcentaje de observaciones que están en la clase de interés y que han sido correctamente clasificadas (**Sensitividad**) versus distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (**especifidad**) versus los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtiene intersectando las dos curvas.
b) Usar la curva ROC (receiver operating characteristic curva). En este caso se grafica la sensitividad versus (1-especifidad)100%, y se coge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto (0,100).

Ahora aplicaremos la regresión logística como un clasificador a los datos del ejemplo anterior. En lo que sigue vamos a considerar los resultados del segundo modelo obtenido en el ejemplo 3.
Prediciendo las clases con el segundo modelo usando el método mas simple es decir comparando el valor ajustado por la regresión logística con p=0.5 y asignando la observación a la clase 1 se obtienen 52 de las 189 observaciones mal clasificadas lo cual representa una tasa de mala clasificación del 27.51%
Haciendo la clasificación con el método más complicado calculando la sensitividad y especificidad se obtiene la siguiente tabla

Sensitividad	Especifidad	P	(1-especifidad)%
100.00	15.38	0.10	84.62
88.14	36.15	0.20	63.85
79.66	53.85	0.25	46.15

67.80	67.69	0.30	32.31
55.93	77.69	0.35	22.31
47.46	83.08	0.40	16.92
35.59	89.23	0.50	10.77
18.64	96.92	0.60	3.08
13.56	98.46	0.70	1.54
1.69	99.23	0.80	0.77
0.00	100.00	0.90	0.00

Notar que para p=.30 la curva está mas cerca a la esquina superior izquierda. La tasa de mala clasificación optima es= 0.3227513

Las gráficas de los dos métodos aparecen en las figuras 5.5 y 5.6 respectivamente y en ambos caso el p-óptimo a usarse es p=0.3

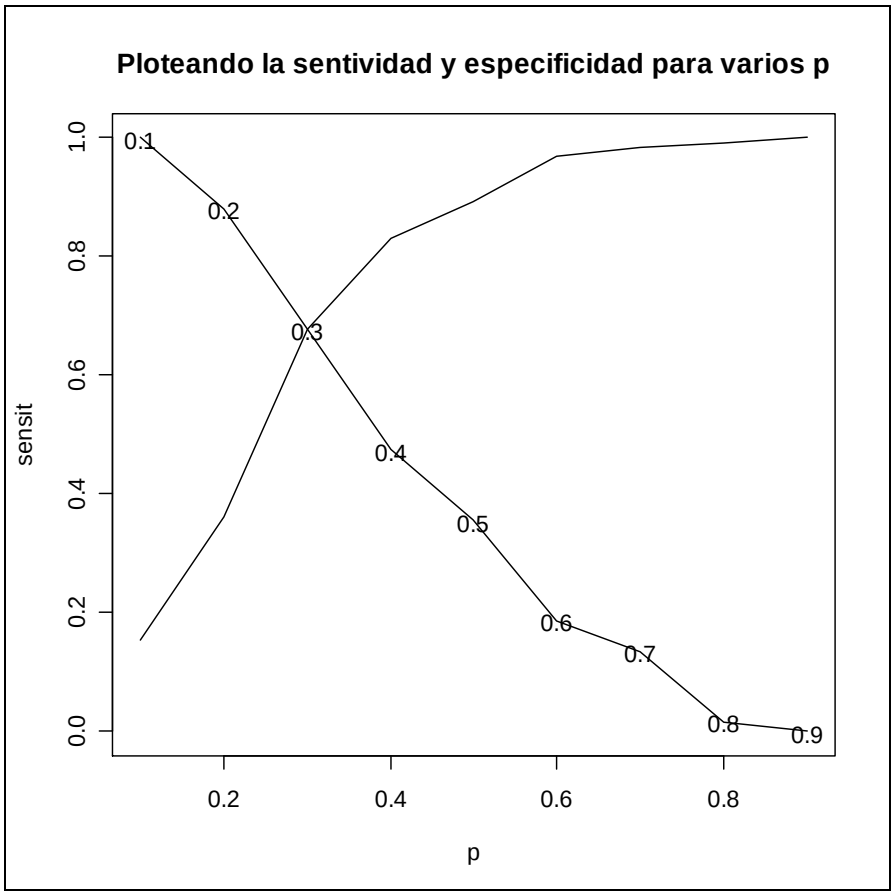

Figura 5.5. Determinación del p óptimo usando el método a

La regresión logística se puede extender al caso donde hay más de dos clases y recibe el nombre de regresión logística politómica. Este tipo de regresión es estudiada mas detalladamente en un curso de clasificación. También existe una relación entre regresión logistica y redes neuronales.
El commando **multinom** de la libreria nnet de R lleva a cabo regression logística politónmica.

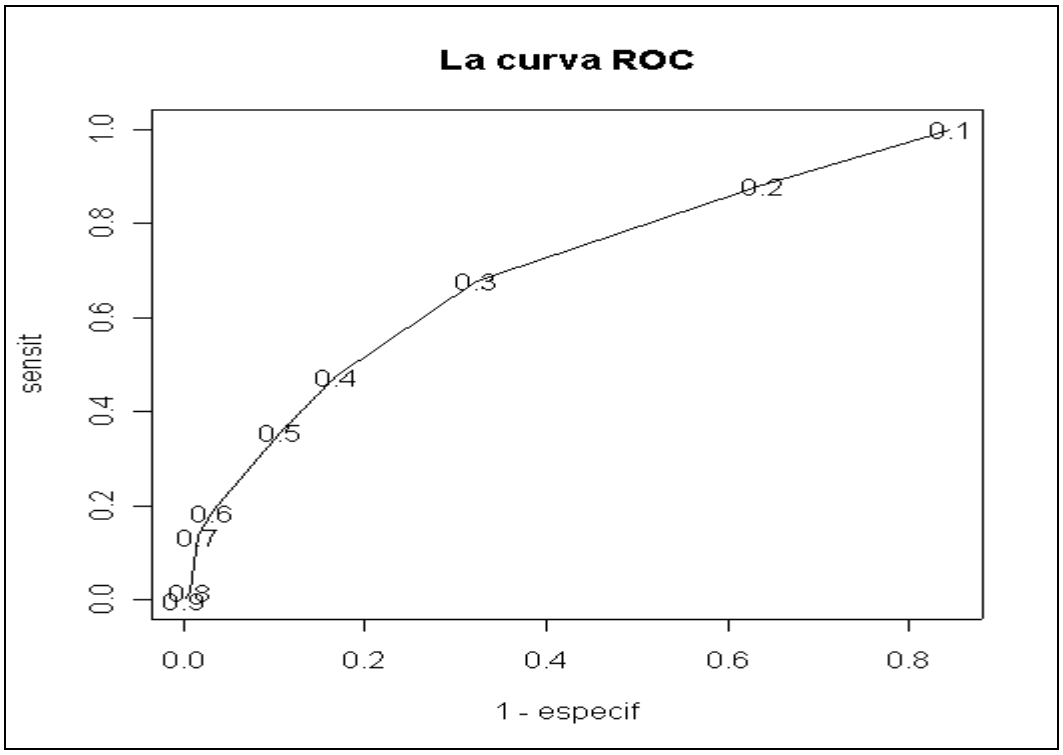

Figura 5.6. Determinación del p óptimo usando la curva ROC

Ejercicios

1. Comparando lineas de regresión. Considerar el conjunto de datos bajopeso disponible en la página de internet del texto y tomar al peso del bebé como Y y a peso de la mamá como X. Comparar las pendientes y los interceptos de la línea de regresión en los tres grupos de raza de la madre.

2. Considerar el conjunto de datos **heartc** disponible en math.uprm.edu/~edgar/datosclass.html en el cual se toman 13 mediciones a 297 pacientes para clasificarlos en propensos o no propensos a sufrir ataque cardíaco. Las clases están en la última columna y están codificadas como 1 y 2.

a) Usar los criterios de la Devianza y del AIC para determinar un modelo de logistico óptimo
b) Determinar la bondad de ajuste del modelo
c) Identificar posibles valores influenciales
d) Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

3**.** Considerar el conjunto de datos **breastw** disponible en math.uprm.edu/~edgar/datosclass.html en el cual se toman 9 mediciones a 699 mujeres para clasificarlas en propensas o no propensas a tener cáncer al seno. Las clases están en la última columna y están codificadas como 1 y 2.

a) Usar los criterios de la Devianza y del AIC para determinar un modelo de logistico óptimo
b) Determinar la bondad de ajuste del modelo
c) Identificar posibles valores influenciales
d) Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

4. Regresión logística con datos agrupados. En una Universidad se registra el número de estudiantes que pasaron con A de acuerdo a las veces que habian tomado de antemano un curso de estadistica. Los resultados se muestran en la siguiente tabla

Veces que habia tomado el curso anteriormente	Número de estudiantes	Estudiantes pasando con A
0	300	30
1	150	25
2	80	20
3	35	5
4	20	3
5	5	1

a) Construir una regresión logística para predecir la probabilidad de que un estudiante obtenga A en la clase de acuerdo a las veces que la ha tomado antes.
b) Probar si la variable predictora: número de veces que el estudiante ha tomado antes el curso e s significativa o no?
c) Determinar la bondad de ajuste del modelo.