Programa del curso

Cómputo Estadístico

- En este curso se discute la generalización de los modelos de regresión incluyendo los modelos de regresión lineal, logística y de Poisson, y se revisan las herramientas clásicas y metodologías modernas para evaluar y seleccionar los modelos. Se discuten también herramientas computacionales que facilitan la estimación de los parámetros de interés y su aplicación en la imputación de datos. Se ofrece también una introducción al análisis de datos temporales y sus aplicaciones.
- Los objetivos generales son proporcionar las bases teóricas que sustentan a las principales aplicaciones de los modelos estadísticos, con un enfoque moderno, haciendo uso de algoritmos computacionales intensivos.

Contenido del curso

I. Modelos estadísticos
- Modelos lineales generalizados
  - Modelos de regresión logística y Poisson
  - Modelos lineales generalizados, el caso general
- Modelos lineales de los análisis de varianza y covarianza
- Modelos log-lineales

II. Análisis de datos temporales
- Procesos estocásticos
- Autocovarianza y autocorrelación. Series de tiempo estacionarias
- Procesos autorregresivos
- Procesos de promedios móviles
- Procesos ARMA
- Análisis espectral de series de tiempo

Contenido del curso

III. Métodos de estimación computacionalmente intensivos

- Algoritmo EM
- Algoritmos MCMC
- Bootstrap

IV. Evaluación y selección de modelos en análisis de regresión

- Criterios para evaluar y seleccionar el modelo adecuado
- Estimación del error de predicción: Validación cruzada
- Métodos de selección de modelos: Métodos Stepwise, AIC, BIC
- Métodos de selección de variables: regularización, Ridge, LARS, LASSO

V. Métodos de imputación de datos

- Métodos basados en regresión y análisis de covarianza.
- Métodos basados en el algoritmo EM
- Imputación Bayesiana
- Métodos basados en técnicas de Machine Learning

Aprendizaje y evaluación

- Actividades de aprendizaje
  - Clases
  - Sesiones de ayudantías
  - Laboratorios de cómputo
  - Individuales: tareas, estudio
- Criterios y procedimientos de evaluación y acreditación
  - Exámenes parciales
  - Examen final
  - Evaluación de las tareas y actividades en clase.
- Asistente del curso: José Benito Hernández Chaudary

Bibliografía

1. Wakefield, J. (2013). Bayesian and frequentist regression methods. Springer.
2. Myers, R.H., Montgomery, D.C. & Vining, G.G. (2001). Generalized linear models with applications in engineering and the sciences. Wiley.
3. Rizzo, M. (2008) Statistical computing with R. Chapman & Hall.
4. Hastie, T., Tibshirani, R. & Friedman, J. (2009). The elements of statistical learning. Springer.
5. Rawlings, J.O., Pantula, S.G. & Dickey, D.A. (1998). Applied regression analysis: a research tool, 2nd Ed. Springer.
6. Little, R.J.A. & Rubin, D.B. (2002). Statistical analysis for missing data, 2nd Ed. Wiley.
7. Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. Wiley.
8. Wood, S. (2006). Generalized additive models. An introduction with R. Chapman & Hall.
9. Faraway, J. (2005). Linear models with R. Chapman & Hall.
10. Fuller, W. (1996). Introduction to statistical time series 2nd Ed. Wiley.
11. Box, G., Jenkins, G., Reinsel, G. (2008). Time series analysis, forecasting and control, 4th Ed. Wiley.

Regresión Logística

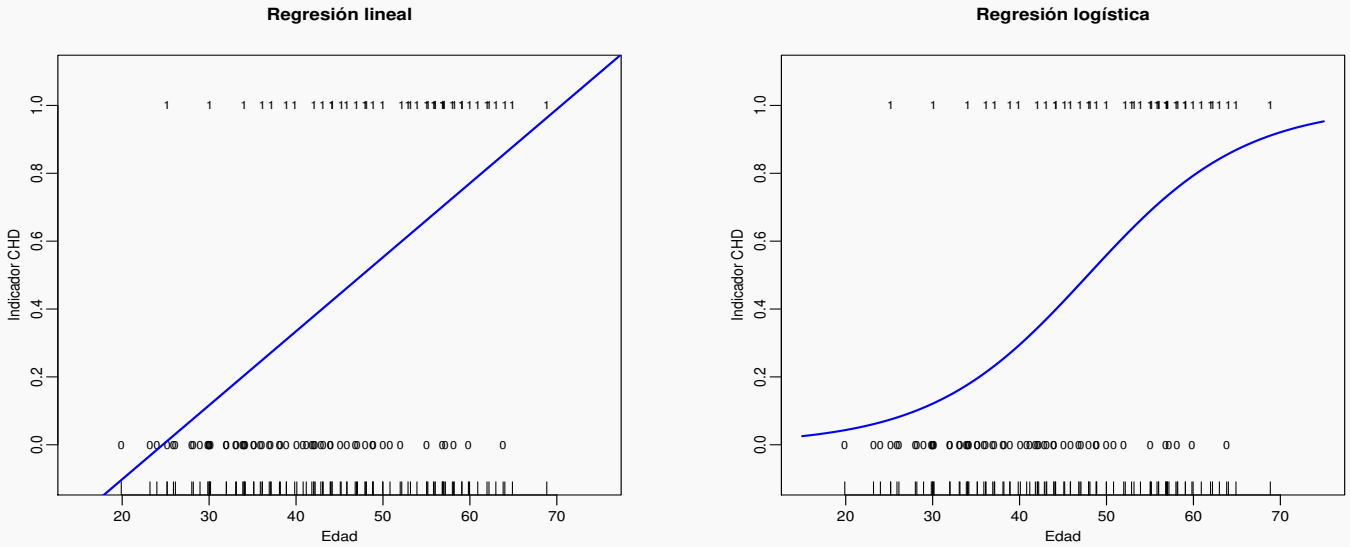
Modelos con respuesta binaria

- Los datos siguientes son edades (en años) e indicadores de presencia o ausencia de daño significativo en la coronaria de 100 individuos seleccionados para participar en el estudio.

edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD
20	0	34	0	41	0	48	1	57	0
23	0	34	0	42	0	48	1	57	1
24	0	34	1	42	0	49	0	57	1
25	0	34	0	42	0	49	0	57	1
25	1	34	0	42	1	49	1	57	1
26	0	35	0	43	0	50	0	58	0
26	0	35	0	43	0	50	1	58	1
28	0	36	0	43	1	51	0	58	1
28	0	36	1	44	0	52	0	59	1
29	0	36	0	44	0	52	1	59	1
30	0	37	0	44	1	53	1	60	0
30	0	37	1	44	1	53	1	60	1
30	0	37	0	45	0	54	1	61	1
30	0	38	0	45	1	55	0	62	1
30	0	38	0	46	0	55	1	62	1
30	1	39	0	46	1	55	1	63	1
32	0	39	1	47	0	56	1	64	0
32	0	40	0	47	0	56	1	64	1
33	0	40	1	47	1	56	1	65	1
33	0	41	0	48	0	57	0	65	1

Modelos

- Deseamos establecer una relación entre la edad de una persona y su propensión a padecer un problema en la coronaria. Las siguientes gráficas muestran dos posibles soluciones:



- La cuestión aquí no es ¿cuál ajusta mejor? sino, ¿cuál es la más adecuada?.

Modelos

- En regresión lineal modelamos el comportamiento medio de una variable de interés (variable de respuesta) como función de covariables

$$\mathrm{E}(y) = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k$$

- en regresión logística también se modela la media como función de covariables

$$\mathrm{E}(y) = h(\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k) = h(x^T \beta)$$

  o, equivalentemente $g(\ \mathrm{E}(y)\ ) = x^T \beta$

- Por ser y una variable Bernoulli, entonces tenemos que sus dos posibles valores los toma con probabilidades

$$P(y = 1) = p \qquad \text{y} \qquad P(y = 0) = 1 - p$$

Función logit

- Una forma muy usada (aparte de que es sensata) de modelar la dependencia de p sobre covariables es mediante la función logit:

$$\text{logit}(p) \;=\; \log\left(\frac{p}{1-p}\right) \;=\; x^T\beta$$

- o, equivalentemente (despejando para p), mediante la función logística (esta es la que se muestra a la derecha en hojas anteriores):

$$p = E(y|x) = \frac{1}{1+\exp(-x^T\beta)}$$

Este es el llamado Modelo de Regresión Logística.

Problemas básicos en Estadística

- En forma muy simplificada, muchos problemas en Estadística giran alrededor de responder las siguientes preguntas:
  - ¿Cómo modelo un fenómeno? (i.e. ¿Cómo parametrizo su comportamiento?)
  - ¿Cómo estimo, a partir de datos, los parámetros del modelo?
  - ¿Cómo valido ese modelo?
  - ¿Cómo uso ese modelo? (i.e. cómo me permite explicar un fenómeno, cómo descubre uno relaciones, cómo hago inferencias, cómo hago predicciones, etc.)

Modelo

- Supongamos que tenemos observaciones independientes $(x_1^T, y_1), (x_2^T, y_2), \cdots, (x_n^T, y_n)$, sobre n individuos, donde y_i es una variable binaria $(0/1)$ con 1 indicando la presencia de cierta característica de interés en el individuo i, y x_i es el correspondiente vector de covariables (o atributos).
- El modelo de regresión logística postula la relación entre la probabilidad de ocurrencia de un evento y los niveles de covariables (o variables predictoras):

$$p_i = P(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-x_i^T \beta)}$$

- El método estándar para estimar los parámetros del modelo es **Máxima Verosimilitud**.

Estimación

- La verosimilitud de los datos independientes (x_1^T, y_1), $\cdots, (x_n^T, y_n)$ es

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \qquad \text{con} \quad p_i = \frac{1}{1+\exp(-x_i^T\beta)}$$

Estimamos β como aquel valor que maximiza la logverosimilitud $l(\beta) = \log L(\beta)$

$$l(\beta) = \sum_{i=1}^{n} [\, y_i \log(p_i) + (1-y_i)\log(1-p_i) \,]$$

Derivando la logverosimilitud:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[y_i \frac{1}{p_i}\frac{\partial p_i}{\partial \beta} - (1-y_i)\frac{1}{1-p_i}\frac{\partial p_i}{\partial \beta} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right] \frac{\partial p_i}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i - p_i}{p_i(1-p_i)} \right] \frac{\partial p_i}{\partial \beta}$$

Estimación

- por otro lado,

$$\frac{\partial p_i}{\partial \beta} = -\left(1 + e^{-x_i^T \beta}\right)^{-2} e^{-x_i^T \beta}(-x_i)$$

$$= \frac{1}{1 + e^{-x_i^T \beta}} \frac{e^{-x_i^T \beta}}{1 + e^{-x_i^T \beta}} x_i = p_i(1 - p_i)x_i$$

entonces

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i - p_i}{p_i(1 - p_i)} \right] p_i(1 - p_i)x_i = \sum_{i=1}^{n} (y_i - p_i)\, x_i$$

- Así, las ecuaciones de verosimilitud son:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - p_i)\, x_i = 0, \qquad \text{donde las } x_i\text{'s son } \; p \times 1$$

en general, este es un sistema de p ecuaciones nolineales en p incógnitas. El método de Newton es un procedimiento iterativo que puede ser útil para resolver este tipo de sistemas.

Método de Newton 1

- Las ecuaciones de verosimilitud son:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - p_i)\, x_i = 0$$

- Para el caso de una sola covariable, se ven como:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - p_i) \begin{bmatrix} 1 \\ z_i \end{bmatrix} = \begin{bmatrix} \sum (y_i - p_i) \\ \sum (y_i - p_i) z_i \end{bmatrix}$$

$$\equiv \begin{bmatrix} h_1(\beta) \\ h_2(\beta) \end{bmatrix} \equiv H(\beta) = 0, \quad \text{(note que } H : \mathbb{R}^2 \to \mathbb{R}^2\text{)}$$

- Para resolver $H(\beta) = 0$, hacemos una aproximación de primer orden para H alrededor de algún valor inicial razonable β_0:

$$H(\beta) \approx H(\beta_0) + \frac{\partial H(\beta_0)}{\partial \beta}(\beta - \beta_0)$$

Método de Newton 2

- en la última expresión, tenemos

$$\frac{\partial H(\beta_0)}{\partial \beta} = \begin{bmatrix} \frac{\partial h_1(\beta_0)}{\partial \beta^T} \\ \frac{\partial h_2(\beta_0)}{\partial \beta^T} \end{bmatrix} \quad \text{es una matriz } 2 \times 2$$

  entonces, en vez de resolver $H(\beta) = 0$, resolvemos el problema más fácil $H(\beta_0) + \frac{\partial H(\beta_0)}{\partial \beta}(\beta - \beta_0) = 0$.
- Así, despejando para β, obtenemos

$$\beta_1 = \beta_0 - \left[\frac{\partial H(\beta_0)}{\partial \beta} \right]^{-1} H(\beta_0)$$

  esta expresión la iteramos hasta convergencia

$$\beta_{k+1} = \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k)$$

Método de Newton 3

- Las diferentes partes que intervienen en la expresión anterior son:

$$H(\beta) = \begin{bmatrix} h_1(\beta) \\ h_2(\beta) \end{bmatrix} = \begin{bmatrix} \sum(y_i - p_i) \\ \sum(y_i - p_i)z_i \end{bmatrix}$$

$$\frac{\partial H(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial h_1(\beta)}{\partial \beta^T} \\ \frac{\partial h_2(\beta)}{\partial \beta^T} \end{bmatrix} = -\begin{bmatrix} \sum \frac{\partial p_i}{\partial \beta} \\ \sum \frac{\partial p_i}{\partial \beta} z_i \end{bmatrix} = -\begin{bmatrix} \sum p_i(1-p_i)x_i^T \\ \sum p_i(1-p_i)z_i x_i^T \end{bmatrix}$$

$$\frac{\partial H(\beta)}{\partial \beta} = -\sum_{i=1}^{n} p_i(1-p_i)\begin{bmatrix} 1 \\ z_i \end{bmatrix} x_i^T = -\sum_{i=1}^{n} p_i(1-p_i)x_i x_i^T$$

Método de Newton 4

- Reescribimos en notación matricial. Definamos X y W de tamaños $n \times p$ y $n \times n$ respectivamente, y y p, $n \times 1$:

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad W = \begin{bmatrix} p_1(1-p_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n(1-p_n) \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{y} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

con esto, es fácil ver que

$$\beta_{k+1} = \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k) = \beta_k + (X^T W X)^{-1} X^T (y - p)$$

- Muchos procedimientos de optimización son de esta forma:

$$\text{nuevo} = \text{antiguo} + \text{tamaño de paso} \times \text{dirección}$$

Concavidad de la logverosimilitud

- Note que la segunda derivada de la logverosimilitud, $l(\beta)$ es:

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \frac{\partial}{\partial \beta} H(\beta) = -X^T W X$$

- Esta matriz es negativa definida para todo β, (salvo casos extremos de dependencias lineales exactas en las columnas de X), por lo tanto la logverosimilitud es estrictamente cóncava.
- En general, los problemas convexos de optimización son "fáciles". Casi cualquier algoritmo sensato va a lograr convergencia al óptimo.
- Por supuesto, en la práctica podemos tener problemas numéricos de "overflow" y siempre es bueno, arrancar los algoritmos iterativos desde diferentes puntos iniciales.
- En resumen: En general, gracias a su propiedad de convexidad, el método de Newton nos lleva al máximo de la función de verosimilitud.

IRWLS

- En la literatura de Modelos Lineales Generalizados encontramos que la estimación se hace comunmente mediante "Mínimos Cuadrados Ponderados Iterativamente" (Iteratively Reweighted Least Squares). Este método es precisamente el anterior que acabamos de ver.
- Si definimos el vector de "observaciones de trabajo"

$$\tilde{y} = X\beta_k + W^{-1}(y - p)$$

entonces, el Método de Newton es

$$\begin{aligned}\beta_{k+1} &= \beta_k + (X^T W X)^{-1} X^T (y - p) \\ &= (X^T W X)^{-1} X^T W X \beta_k + (X^T W X)^{-1} X^T W W^{-1} (y - p) \\ &= (X^T W X)^{-1} X^T W (X\beta_k + W^{-1}(y - p)) \\ &= (X^T W X)^{-1} X^T W \tilde{y}\end{aligned}$$

- Esto es, $\beta_{k+1} = (X^T W X)^{-1} X^T W \tilde{y}$, y de aquí es de donde le viene el nombre de "mínimos cuadrados ponderados".

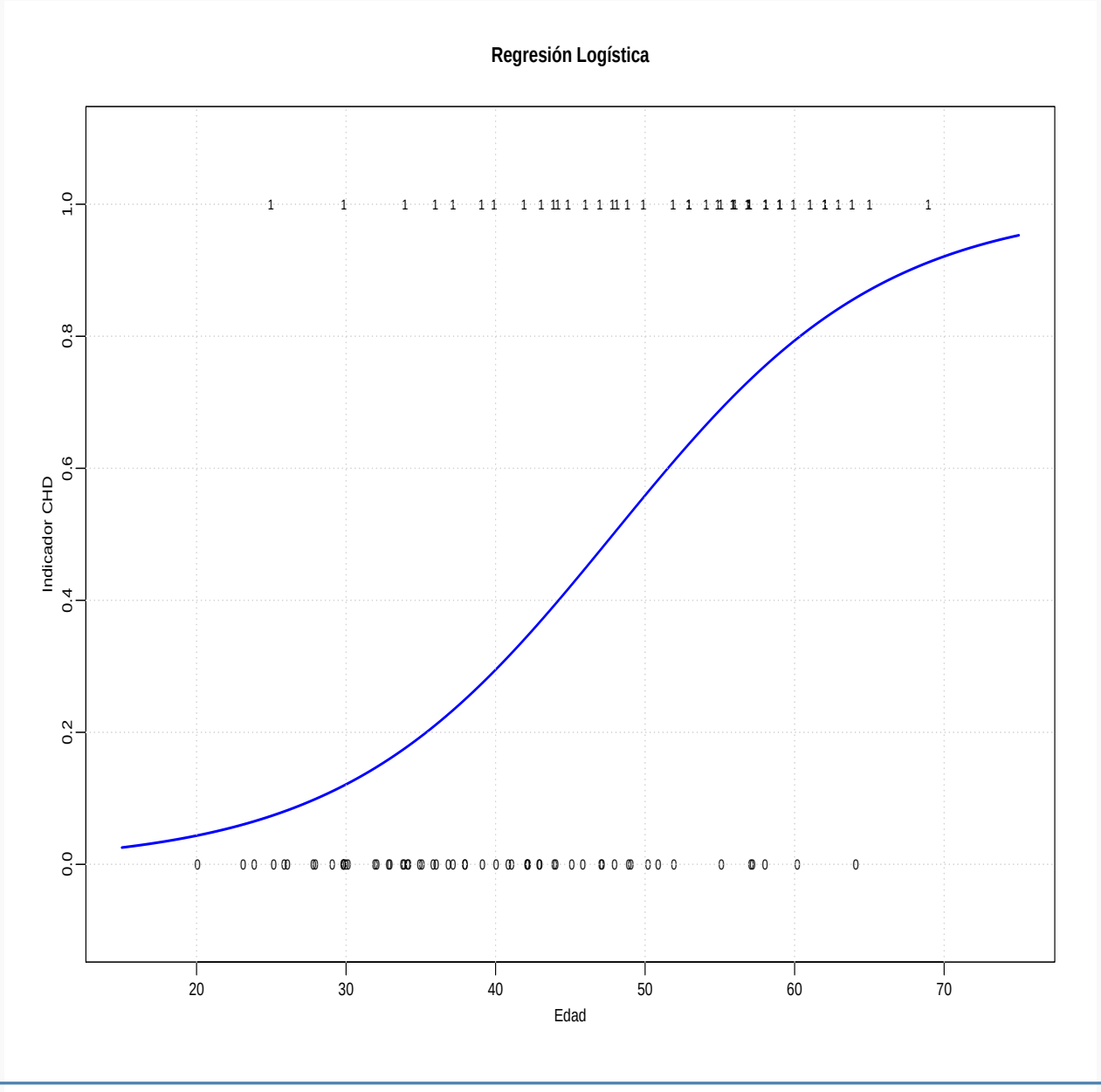
Código en R

```
# Hosmer, D.W. & Lemeshow, S.(1989) Applied logistic regression. Wiley
# Edad y Coronaria (daño significativo en coronaria)

edad <- c(
 20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32, 32, 33, 33,
 34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39, 40, 40, 41,
 41, 42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48,
 48, 48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57,
 57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64, 65, 69)

coro <- c(
 0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,1,0,
 0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,0,1,1,0,0,1,0,1,0,0,1,1,1,1,0,1,1,1,1,1,0,
 0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0,1,1,1)

# Gráfica de los datos

edaj <- jitter(edad)    # solo con fines de graficación

plot(edaj, coro, xlab="Edad", ylab="Indicador CHD", ylim=c(-.1,1.1),
 mgp=c(1.5,.5,0), cex.axis=.8, cex.lab=.8, cex.main=1, xlim=c(15,75), cex=.7,
 main="Regresión lineal", pch=ifelse(coro==1,"1","0"))
rug(edaj)
out <- lm(coro ~ edad)
abline(out,lwd=2,col="blue")

plot(edaj, coro, xlab="Edad", ylab="Indicador CHD", ylim=c(-.1,1.1),
 mgp=c(1.5,.5,0), cex.axis=.8, cex.lab=.8, cex.main=1, xlim=c(15,75), cex=.7,
 main="Regresión Logística", pch=ifelse(coro==1,"1","0"))
rug(edaj)
```

Código en R

```
# Resolviendo ecuaciones de verosimilitud
y       <- coro
n       <- length(y)
X       <- cbind(rep(1,n),edad)
b       <- c(-10,.2) # valores iniciales
# Las 4 líneas anteriores son específicas para los datos de coronaria
tolm    <- 1e-6      # tolerancia (norma minima de delta)
iterm   <- 100       # numero maximo de iteraciones
tolera  <- 1         # inicializar tolera
itera   <- 0         # inicializar itera
histo   <- b         # inicializar historial de iteraciones
while( (tolera>tolm)&(itera<iterm) ){
 p      <- 1/( 1+exp( -as.vector(X%*%b) ) )
 W      <- p*(1-p)
 delta  <- as.vector( solve(t(X*W)%*%X, t(X)%*%(y-p)) )
 b      <- b + delta
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
# histo -10.000000 0.20000000
# b      -1.497206 0.03767488
# b      -4.380358 0.09253679
# b      -5.221685 0.10918224
# b      -5.308597 0.11090419
# b      -5.309453 0.11092114
# b      -5.309453 0.11092114
# Agregamos curva logística a la gráfica original
xx   <- seq(15,75,length=200)
X    <- cbind(rep(1,n),xx)
p    <- 1/( 1+exp( -as.vector(X%*%b) ) )
lines(xx,p,lwd=2,col="blue")
grid()
```

Modelo ajustado a datos de Coronaria



Interpretación de los coeficientes del modelo

- En regresión lineal, por ejemplo, $E(y \mid z_1) = \beta_0 + \beta_1 z_1$, los coeficientes del modelo tienen una interpretación directa. Si los atributos, z_1, de un individuo se incrementan en una unidad, esto es, de z_1 pasa a $z_1 + 1$, entonces el impacto en la media esta dado por β_1, pues

$$E(y \mid z_1+1) - E(y \mid z_1) \;=\; (\beta_0 + \beta_1[z_1+1]) - (\beta_0 + \beta_1 z_1) \;=\; \beta_1$$

- En regresión logística sucede algo semejante, pero el impacto no es en la media sino en la diferencia de logits.

$$\text{logit}(p) = \text{logit}[\; P(y = 1 \mid z_1)\;] = \beta_0 + \beta_1 z_1$$

entonces

$$\text{logit}[\; P(y = 1 \mid z_1 + 1)\;] - \text{logit}[\; P(y = 1 \mid z_1)\;] \;=\; \beta_1$$

y resulta que esta diferencia de logits tiene una interpretación importante.

Tasa de momios

- Tasa de momios

$$\text{logit}[\, P(y = 1 \mid z_1 + 1)\,] - \text{logit}[\, P(y = 1 \mid z_1)\,]$$
$$= \log \frac{\text{Momios de } y = 1 \text{ con } z_1 + 1}{\text{Momios de } y = 1 \text{ con } z_1} = \beta_1$$

- Los momios de que ocurra un evento A se definen como $P(A)/[\,1 - P(A)\,]$. La tasa de momios ("odds ratio") compara los momios de un evento bajo dos escenarios (el base z_1 versus cuando la variable aumenta $z_1 + 1$).
- La tasa de momios, para propósitos prácticos, se interpreta como la comparación de dos probabilidades; por ejemplo si la tasa de momios es igual a 2 entonces decimos que aumentar z_1 en una unidad incrementa al doble la probabilidad de ocurrencia del evento con respecto al escenario base (Nota: Por supuesto, esta no es, formalmente, la interpretación correcta).

Tasa de momios

- La relación entre la tasa de momios y los coeficientes del modelo es:

$$\frac{\text{Momios de } y = 1 \text{ con } z_1 + 1}{\text{Momios de } y = 1 \text{ con } z_1} = e^{\beta_1}$$

- Por ejemplo, en el estudio sobre enfermedades coronarias tenemos $\widehat{\beta} = 0.1109$. Si comparamos dos individuos con 10 años de diferencia tenemos

$$\frac{\text{Momios de } y = 1 \text{ con } z_1 + 10}{\text{Momios de } y = 1 \text{ con } z_1} = e^{10\beta_1} = e^{1.109} = 3.03$$

- Si dos personas tienen 10 años de diferencia, entonces el riesgo de tener problemas en la coronaria para la persona mayor es 3 veces más alto que el riesgo que tiene la persona menor.

Tasa de momios

- La interpretación anterior supone que los logits se expresan linealmente en términos de las covariables **en toda la escala de las covariables**; esto pudiera no ser realista pues la comparación de los riesgos entre dos personas de 20 y 30 años podría no ser tan drástica como la comparación de los riesgos entre personas de 50 y 60 años. Esto lleva a un problema práctico el cual se podría enfrentar, por ejemplo, con transformaciones de la covariable, digamos log(edad).
- Las tasas de momios tienen una propiedad importante en aplicaciones: Son invariantes con respecto al tipo de estudio (prospectivo o retrospectivo, i.e. estudios de cohortes prospectivos o estudios casos-control) (o sobresimplificando: "caros y lentos" versus "baratos y rápidos"). Esta propiedad de invarianza se traduce en un mayor atractivo de los modelos de regresión logística.

Nota acerca de la tasa de momios 1

- Los datos de la siguiente tabla provienen de uno de los primeros estudios sobre la asociación entre cáncer de pulmón y fumar.

	Cáncer	Controles
Fuma	688	650
No Fuma	21	59
	709	709

- El estudio fue efectuado en 20 hospitales en Inglaterra; los controles fueron pacientes (sin cáncer) seleccionados del mismo sexo, mismos hospitales y aproximadamente de la misma edad que los pacientes con cáncer. La cantidad que es de interés es el **Riesgo Relativo**

$$RR = \frac{P(Can \mid Fum)}{P(Can \mid NoFum)}$$

sin embargo, para este estudio, estas cantidades no son estimables ¿por qué?.

Nota acerca de la tasa de momios 2

- Los **momios** de la ocurrencia de un evento A se definen como

$$\omega = \frac{P(A)}{1 - P(A)}$$

Así, si $A \equiv Can \mid Fum$, los momios de cáncer dado que la persona fuma, se definen como

$$\omega_1 = \frac{P(Can \mid Fum)}{1 - P(Can \mid Fum)} \equiv \frac{P(C \mid F)}{1 - P(C \mid F)}$$

- Para comparar estos momios contra los momios de cáncer dado que la persona no fuma

$$\omega_2 = \frac{P(Can \mid NoFum)}{1 - P(Can \mid NoFum)} \equiv \frac{P(C \mid NF)}{1 - P(C \mid NF)}$$

usamos la **tasa de momios**

$$\theta = \frac{\omega_1}{\omega_2} = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]}$$

Nota acerca de la tasa de momios 3

- La expresión anterior para la tasa de momios aparentemente tiene el mismo problema de no estimabilidad de las probabilidades que la conforman; sin embargo, tenemos la siguiente relación:

$$\theta = \frac{\omega_1}{\omega_2} = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]} = \frac{P(F \mid C)[1 - P(F \mid NC)]}{P(F \mid NC)[1 - P(F \mid C)]}$$

  las cuales **si** pueden ser estimadas del estudio retrospectivo.

$$\widehat{\theta} = \frac{[688/709]\ [59/709]}{[650/709]\ [21/709]} = \frac{688 \times 59}{650 \times 21} = 2.97$$

- De aquí que **los momios de cáncer en fumadores son 3 veces más altos que los momios de cáncer en no fumadores**.
- En general, en una tabla 2×2, los momios se calculan como:

$$\widehat{\theta} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}}$$

Nota acerca de la tasa de momios 4

- Nota técnica:

$$\theta = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]} = \frac{\frac{P(C,F)}{P(F)} \times \frac{P(NF) - P(C,NF)}{P(NF)}}{\frac{P(C,NF)}{P(NF)} \times \frac{P(F) - P(C,F)}{P(F)}}$$

$$= \frac{P(F,C)P(NF,NC)}{P(NF,C)P(F,NC)} = \frac{\frac{P(F,C)}{P(C)} \times \frac{P(NF,NC)}{P(NC)}}{\frac{P(NF,C)}{P(C)} \times \frac{P(F,NC)}{P(NC)}}$$

$$= \frac{P(F \mid C)}{P(NF \mid C)} \times \frac{P(NF \mid NC)}{P(F \mid NC)} = \frac{P(F \mid C)[1 - P(F \mid NC)]}{P(F \mid NC)[1 - P(F \mid C)]}$$

la cual es la relación que queríamos demostrar. Para obtener lo anterior, se usó la relación:

$$P(A) = P(A,B) + P(A,NB)$$

Tarea 1 (1)

1 La siguiente tabla muestra los resultados parciales de dos encuestas que forman parte de un estudio para evaluar el desempeño del Primer Ministro del Canadá. Se tomó una muestra aleatoria de 1600 ciudadanos canadienses mayores de edad y en los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después de la primera encuesta, sólo 880 aprueban su desempeño.

Primera encuesta	Segunda encuesta $Y = 1$, Aprueba	$Y = 0$, Desaprueba	Total
$x = 1$, Aprueba	794	150	944
$x = 0$, Desaprueba	86	570	656
Total	880	720	1600

Tarea 1 (2)

1 $\cdots$ Cont.

a Considere el modelo de regresión logística

$$\log \frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} = \beta_0 + \beta_1 x_i$$

Escriba la logverosimilitud correspondiente. Muestre explícitamente (i.e. maximizando la logverosimilitud), que el estimador máximo verosimilitud para β_1 es el logaritmo de la tasa de momios de la tabla dada (En general, en regresión logística los estimadores de máxima verosimilitud no tienen una forma explícita, sin embargo, en el presente caso si).

b Sea p_1 la proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial y sea p_2 la proporción correspondiente seis meses después. Considere la hipótesis $H_0 : p_1 = p_2$, ¿Cómo puede hacerse esta prueba?.

Tarea 1 (3)

2 Suponga $(x_1, y_1), \ldots, (x_n, y_n)$ observaciones independientes de variables aleatorias definidas como sigue:

$$Y_i \sim \text{Bernoulli}(p), \quad i = 1, \cdots, n$$
$$X_i \mid \{Y_i = 1\} \sim N(\mu_1, \sigma^2)$$
$$X_i \mid \{Y_i = 0\} \sim N(\mu_0, \sigma^2)$$

Usando el Teorema de Bayes, muestre que $\text{P}(Y_i = 1|X_i)$ satisface el modelo de regresión logística, esto es

$$\text{logit}(\ \text{P}(Y_i = 1|X_i)\ ) = \alpha + \beta X_i$$

con $\beta = (\mu_1 - \mu_0)/\sigma^2$.

Entregar: Miércoles 19 de agosto.

Errores estándar

- Una forma aproximada de calcular la varianza del estimador de máxima verosimilitud es mediante

$$V(\widehat{\beta}) \doteq \left[-\frac{\partial^2 l(\widehat{\beta})}{\partial \beta \partial \beta^T} \right]^{-1}$$

  esto es, la varianza (asintótica) de $\widehat{\beta}$ se aproxima por el inverso de la Matriz observada de Información.
- Interpretación heurística: La logverosimilitud es globalmente cóncava (para los modelos lineales generalizados), por lo tanto su segunda derivada es negativa, además, la segunda derivada mide el grado de curvatura de la logverosimilitud. Mientras más grande sea la segunda derivada, más "picuda" es la logverosimilitud y mejor definido está su máximo. Entonces, si tomamos el recíproco del negativo de la curvatura tenemos una medida de que tan mal está nuestro estimador, (i.e. que tanta varianza tiene), a mayor curvatura menor varianza.

Errores estándar en regresión logística

- Para el caso de regresión logística vimos que

$$\frac{\partial l(\beta)}{\partial \beta} = H(\beta) = \sum_{i=1}^{n}(y_i - p_i)x_i$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \frac{\partial H(\beta)}{\partial \beta^T} = -\sum_{i=1}^{n} p_i(1-p_i)x_i x_i^T = -X^T W X$$

- así que, la varianza se puede aproximar por:

$$V(\widehat{\beta}) = \left[X^T W X \right]^{-1}$$

- Los errores estándar de los estimadores se obtienen sacando raíz cuadrada a los elementos diagonales de V.

Código en R

```
# Cálculo de errores estándar
y  <- coro
n  <- length(y)
X  <- cbind(rep(1,n),edad)
# Las líneas anteriores son específicas para los datos CHD

p  <- 1/( 1+exp( -as.vector(X%*%b) ) )
W  <- p*(1-p)
V  <- solve( t(X*W)%*%X )
es <- sqrt( diag(V) )
# estimadores y sus desviaciones estándar
#              b         es
#      -5.3094534 1.13365464
# edad  0.1109211 0.02405984

# Usando glm
out <- glm(y ~ edad, family=binomial)
summary(out)

# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -1.9718  -0.8456  -0.4576   0.8253   2.2859

# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
# edad         0.11092    0.02406   4.610 4.02e-06 ***

#     Null deviance: 136.66  on 99  degrees of freedom
# Residual deviance: 107.35  on 98  degrees of freedom
# AIC: 111.35
# Number of Fisher Scoring iterations: 4
```

Usos de los modelos

- Los modelos de regresión logística contribuyen a:
  - Identificar factores de riesgo.
  - Evaluar el riego (probabilidad de ocurrencia del evento respuesta) para individuos específicos.
  - Clasificar / discriminar a grupos de individuos como alto riesgo o bajo riesgo.
- La formulación y estimación del modelo

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-x^T \beta}}$$

nos da una fórmula que nos permite evaluar el riesgo de un individuo específico (i.e. para un vector, x, de atributos específicos).

Problemas de clasificación

- Clasificamos a un individuo como de alto riesgo si la probabilidad de ocurrencia es mayor o igual a cierto umbral; entonces, diremos que un individuo con atributos x es de alto riesgo si

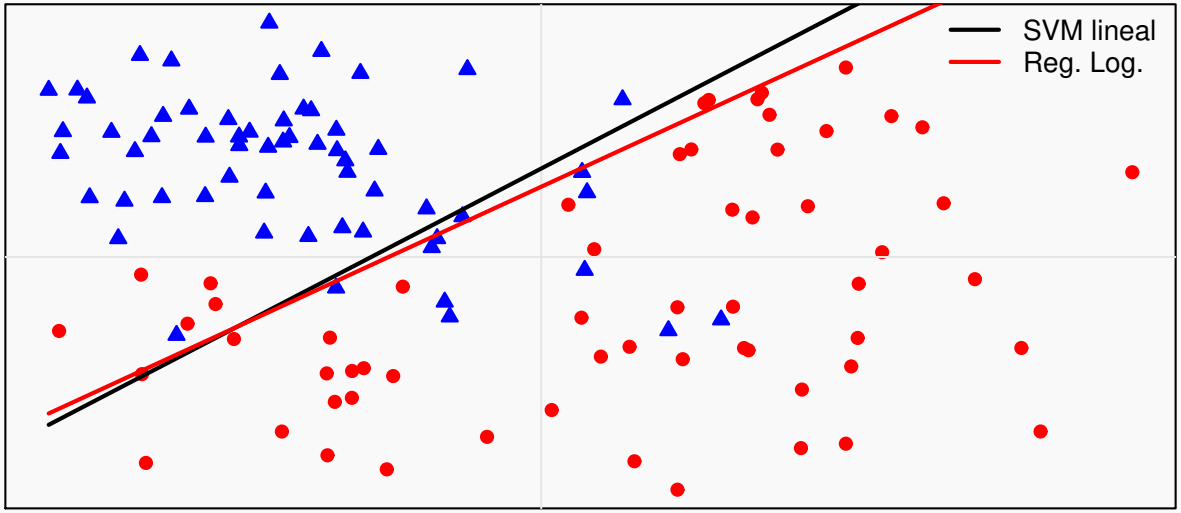
$$\frac{1}{1+e^{-x^T\beta}} > p$$

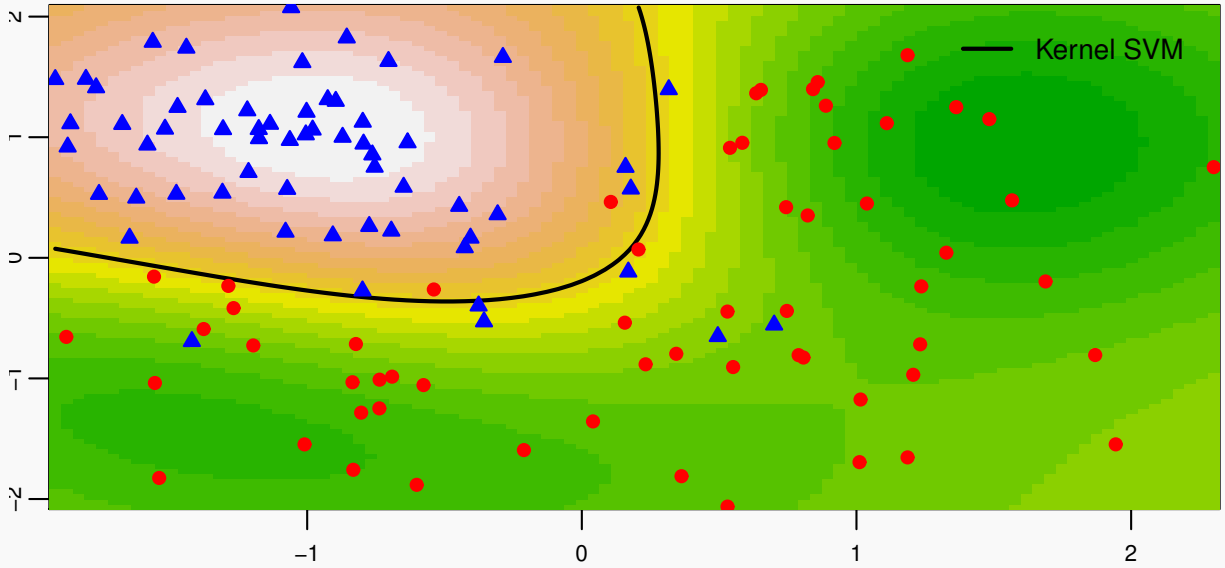
esto es, si

$$x^T\beta > \log\frac{p}{1-p}$$

- En el caso particular de $p = 0.5$ tenemos que los individuos de alto riesgo satisfacen $x^T\beta > 0$.
- La gráfica siguiente muestra el comportamiento de regresión logística, como función discriminante, comparada con el clasificador SVM. Regresión logística, en su forma estándar, no es tan poderosa como los SVM's, sin embargo, son muy útiles por sus usos en la identificación de factores de riesgo, lo cual, no es inmediato con los SVM's kernelizados.

Regresión logística vs SVM's





Ejemplo de un problema de clasificación

- Los datos consisten de digitalizaciones de dígitos manuscritos. Cada dígito es escaneado y estandarizado a un cuadro de 16×16 pixeles; de modo que la x correspondiente es un vector con 256 atributos (niveles de gris de cada pixel). En cada caso se cuenta con y, el valor verdadero del dígito escaneado.
- Queremos tener una regla discriminante entre, digamos, los dígitos 8 y 3, usando las ideas expuestas antes.

Dígitos

Código en R

```
# Datos para ajuste (o entrenamiento)
datos <- scan( "C:\\...\\digitrain.txt" )
dato  <- matrix(datos,ncol=257,byrow=T)            # 7291 x 257
base  <- 8
vs    <- 3
dat   <- dato[ (dato[,1]==base)|(dato[,1]==vs) ,]  # 1200 x 257
y     <- ifelse(dat[,1]==base,0,1)
dd    <- as.data.frame(dat[,-1])
colnames(dd) <- paste("X",1:(dim(dd)[2]),sep="")

# Graficamos una muestra de los digitos
set.seed(64646)
n     <- dim(dd)[1]
M     <- 15
aa    <- as.matrix(dd[sample(1:n,size=(M^2)),])
par(mfrow=c(M,M),mar=c(0, 0, 0, 0))           # grafica de M^2 digitos
for(i in 1:(M^2)){
 bb  <- matrix(aa[i,],ncol=16,byrow=T)
 bb  <- bb[16:1,]
 image(-t(bb), cex.axis=.7, col=terrain.colors(20),mgp=c(1.5,.5,0),
       xlab="",ylab="",xaxt="n",yaxt="n")}

# Ajustamos un modelo de regresión logística
# (no se alcanza convergencia con los defaults...)
# (pero no importa, seguir adelante, ... por qué?)

out   <- glm(y ~ ., family=binomial, data=dd)

# Leemos los datos que usaremos para probar el modelo de clasificación
datp <- scan( "C:\\...\\digitest.txt" )
```

Código en R

```
dap   <- matrix(datp,ncol=257,byrow=T)          # 2007 x 257
dp    <- dap[ (dap[,1]==base)|(dap[,1]==vs) ,]  # 332 x 257
yp    <- ifelse(dp[,1]==base,0,1)
d     <- as.data.frame(dp[,-1])
colnames(d) <- paste("X",1:(dim(d)[2]),sep="")

# Evaluamos la bondad de predicción con los datos de prueba
pre   <- predict(out, newdata=d, type="response")
ypre  <- ifelse(round(pre)==0,base,vs)
yobs  <- ifelse(yp==0,base,vs)
pp    <- table(ypre,yobs)

#        yobs
# ypre   3   8
#    3 152   9
#    8  14 157
# 100*sum(diag(pp))/sum(pp)
# 93.07229   # porcentaje de aciertos con los datos de prueba

# Graficamos una muestra
nt   <- 100
selg <- sample(1:(dim(d)[1]),size=nt)
par(mfrow=c(sqrt(nt),sqrt(nt)),mar=c(0, 0, 1, 0))
for(i in 1:nt){
 bb  <- matrix(as.numeric(d[selg[i],]),ncol=16,byrow=T)# grafica de digitos de prueba
 bb  <- bb[16:1,]               # se resaltan los digitos mal clasif.
 image(-t(bb), cex.axis=.7, col=terrain.colors(20),mgp=c(1.5,.5,0),
       xlab="",ylab="",xaxt="n",yaxt="n")
 mtext(ypre[selg[i]], side = 3,
       col = ifelse(ypre[selg[i]]==yobs[selg[i]],"blue","black"),
       cex= ifelse(ypre[selg[i]]==yobs[selg[i]],1,1.5)) }
```

Dígitos

Ejemplo: Peso al nacer

- Consideremos un estudio sobre pesos de recién nacidos. Se tienen 188 registros de nacimientos de los cuales 58 fueron bebés de peso bajo (< 2.5 kgm).

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	10	1	29	130	1	0	0	0	1	2	1021
2	11	1	34	187	2	1	0	1	0	0	1135
3	13	1	25	105	3	0	1	1	0	0	1330
4	15	1	25	85	3	0	0	0	1	0	1474
5	16	1	27	150	3	0	0	0	0	0	1588
...											
185	223	0	35	170	1	0	1	0	0	1	4174
186	224	0	19	120	1	1	0	0	0	0	4238
187	225	0	24	116	1	0	0	0	0	1	4593
188	226	0	45	123	1	0	0	0	0	1	4990

- Dicotomizamos la variable `bwt` haciendo `lbw=1` si `bwt` $<$ 2500 (la columna `bwt` es el peso, en gramos, al nacer).
- Consideraremos solo las columnas `age`, `lwt`, `race`, `ftv` como variables predictoras de `lbw`.

Ejemplo: Peso al nacer

variable	descripción
low	indicadora 1 si `bwt` < 2500
age	edad
lwt	peso de la madre en el último periodo menstrual
race	blanca $= 1$, negra $= 2$, otra $= 3$
smoke	si $= 1$
ptl	nacimientos prematuros (no$= 0$, uno$= 1$, dos o más$= 2$)
ht	hipertensión (si $= 1$)
ui	irritabilidad uterina (si $= 1$)
ftv	frecuencia de visitas médicas en primer trimestre
bwt	peso al nacer (gramos)

Ejemplo tomado de

- Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013). Applied logistic regression. Wiley

Ejemplo: Usando `glm()`

```
lowbwt    <- read.table( "c:\\...\\lowbwt.dat", header=T )
names(lowbwt) <- c(
 "id","low","age","lwt","race","smoke","ptl","ht","ui","ftv","bwt")
attach(lowbwt)

lbw <- ifelse(bwt<2500,1,0)
r1  <- ifelse(race==2,1,0)
r2  <- ifelse(race==3,1,0)
out <- glm(lbw ~ age+lwt+r1+r2+ftv, family=binomial)
summary(out)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.395792   1.079401   1.293   0.1960
age         -0.028703   0.034132  -0.841   0.4004
lwt         -0.014224   0.006558  -2.169   0.0301 *
r1           0.992963   0.498179   1.993   0.0462 *
r2           0.384988   0.365125   1.054   0.2917
ftv         -0.034574   0.167569  -0.206   0.8365

    Null deviance: 232.33  on 187  degrees of freedom
Residual deviance: 220.38  on 182  degrees of freedom
AIC: 232.38
Number of Fisher Scoring iterations: 4
```

Ejemplo: Optimización directa

```
# Resolviendo ecuaciones de verosimilitud a "pie"
y      <- lbw
n      <- length(y)
X      <- cbind(rep(1,n),age,lwt,r1,r2,ftv)
b      <- c(1,0,0,1,.3,0)          # valores iniciales
tolm   <- 1e-6      # tolerancia (norma minima de delta)
iterm  <- 100       # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones
while( (tolera>tolm)&(itera<iterm) ){
 p      <- 1/( 1+exp( -as.vector(X%*%b) ) )
 W      <- p*(1-p)
 delta  <- as.vector( solve(t(X*W)%*%X, t(X)%*%(y-p)) )
 b      <- b + delta                  # al final, b = estimadores Max.V.
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
b     1.306589 -0.030773 -0.015373 -0.720189 0.037751 -0.032447
b     0.816949 -0.023867 -0.010344  2.701179 0.496114 -0.027881
b     1.020153 -0.027075 -0.011502  0.276587 0.412367 -0.041692
b     1.362326 -0.028317 -0.014030  1.057008 0.387026 -0.033292
b     1.395792 -0.028703 -0.014224  0.992963 0.384988 -0.034574
```

Ejemplo: Errores estándar

```
# Cálculo de errores estándar

y       <- lbw
n       <- length(y)
X       <- cbind(rep(1,n),age,lwt,r1,r2,ftv)
p       <- 1/( 1+exp( -as.vector(X%*%b) ) )
W       <- p*(1-p)
V       <- solve( t(X*W)%*%X )
es      <- sqrt( diag(V) )

cbind(b,es)

#              b          es
#      1.39579193 1.079401170
# age -0.02870331 0.034131750
# lwt -0.01422444 0.006558137
# r1   0.99296316 0.498178969
# r2   0.38498843 0.365125203
# ftv -0.03457411 0.167568833

# (Igual que con glm)
```

Prueba global de ajuste: Cociente de Verosimilitudes

- Una prueba global de ajuste

$$H_0 : \text{Los factores no afectan} \quad \text{vs} \quad H_1 : \text{Si afectan}$$

- La hipótesis nula es equivalente a que

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

- El estadístico Cociente de Verosimilitudes es

$$\Lambda = \frac{\text{Verosimilitud optimizada bajo } H_0}{\text{Verosimilitud optimizada bajo el modelo completo}}$$

- Bajo H_0: $G = -2 \log \Lambda \sim \chi^2_{gl}$ (asintóticamente), donde gl es la diferencia entre el número de parámetros del modelo completo y el número de parámetros del modelo reducido, (en este caso $gl = 6 - 1 = 5$).
- Rechazamos la hipótesis nula si $G = -2 \log \Lambda > \chi^2_{5,\alpha}$.
  Observe que

$$G = -2 \log \Lambda = -2 \text{ logvero. bajo nula} - 2 \text{ logvero.}$$

Ejemplo: Prueba global

```
# Prueba global
# -2loglik del modelo (residual deviance)
aa   <- -2*(sum(y*log(p) + (1-y)*log(1-p)))                  # 220.3819
# -2loglik del modelo nulo (null deviance)
bb   <- -2*(sum(y)*log(mean(y))+(n-sum(y))*log(1-mean(y))) # 232.3318
# -2 log(Cociente de verosimilitudes)
G    <- bb-aa                  # 11.94995
pval <- 1-pchisq(G,df=5)       # 0.035 => rech Ho
```

- De aquí que, con un nivel de significancia del 5%, podemos decir que las variables ayudan a explicar el peso bajo de los bebés.
- Si queremos la significancia de, por ejemplo, la variable `ftv`, podemos usar la prueba de Wald, en la estimación la dividimos entre su error estándar y comparamos contra un cuantil de la normal estándar.

```
z  <- b[6]/es[6]        # -0.2063278
2*(1-pnorm(abs(z)))     #  0.8365349
```

Ejemplo: Pruebas individuales

- En la prueba de Wald anterior, el p-valor es grande y, por lo tanto, el factor `ftv` no es significativo (esta prueba también la da `glm()` directamente).
- Una segunda forma, es la prueba de cociente de verosimilitudes: Ajustando dos modelos, uno con todas las variables y el otro con todas las variables menos la que se quiere evaluar. Calculamos el estadístico G y lo comparamos contra una χ^2_1.
- Nota: El AIC (Akaike Information Criteria), para un modelo dado, se define como

$$AIC = -2\, l(\widehat{\beta}) + 2\, p$$

es un balance entre "ajuste y complejidad". En el ejemplo, tenemos

$$AIC = 220.3819 + 2 * 6 = 232.3819$$

para comparación de modelos, valores menores del AIC son preferibles.

Ejemplo: Comparación de modelos

- Viendo la salida de `glm()`, aparentemente solo las variables `lwt` y raza son importantes (al menos en forma individual).
- Consideremos un modelo reducido con solo estas variables y lo contrastamos con el modelo completo.
- Usando `glm()`, el valor de -2 veces la logverosimilitud del modelo bajo consideración se encuentra bajo el nombre `deviance`, así que basta con tomar la diferencia entre ellas (la del completo y el reducido) y compararla contra una ji-cuadrada de 2 grados de libertad

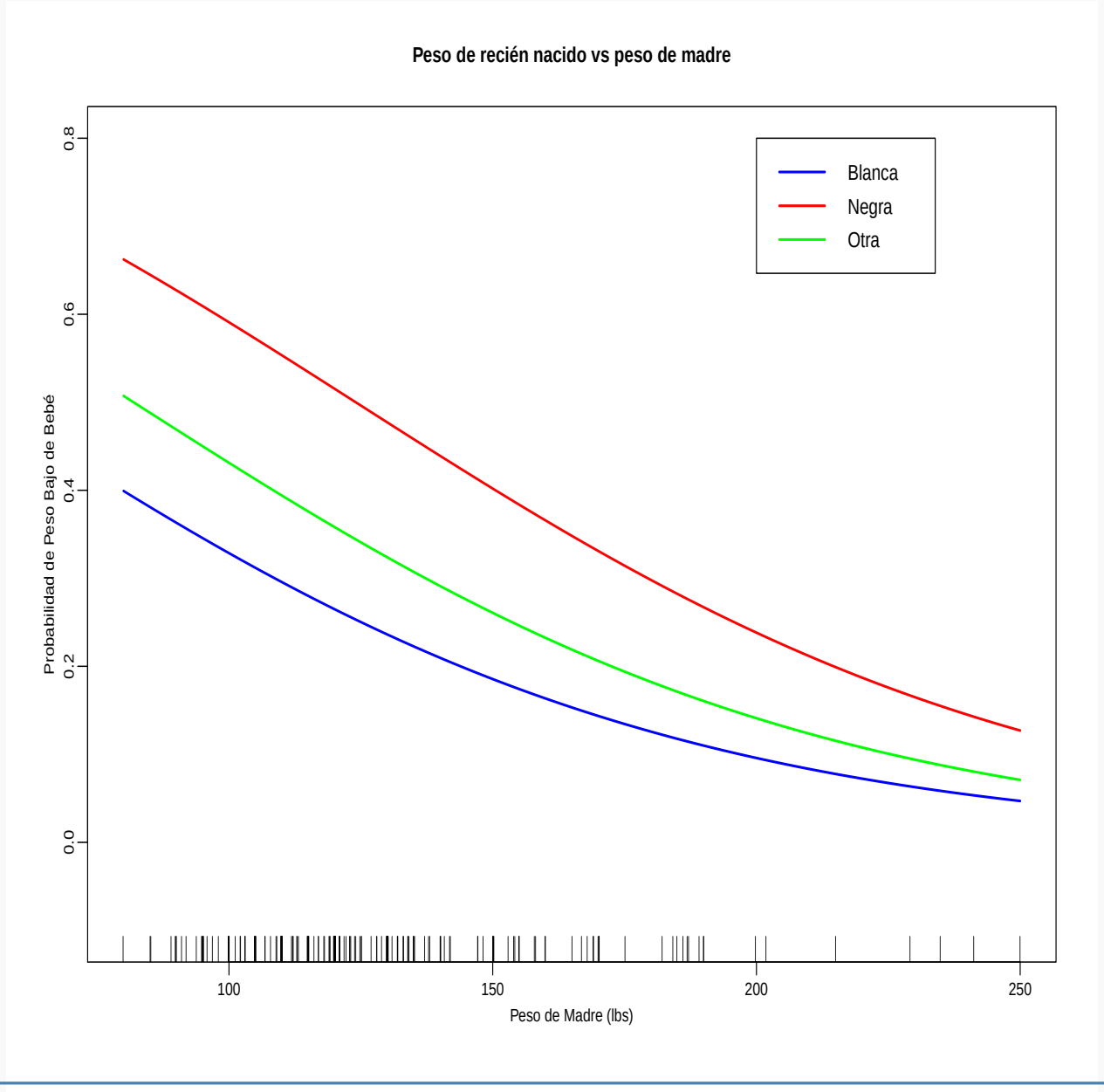
```
# modelo completo vs reducido
mcomp <- glm(lbw ~ age+lwt+r1+r2+ftv, family=binomial)
mred  <- glm(lbw ~ lwt+r1+r2, family=binomial)
G     <- mred$deviance - mcomp$deviance   # 0.8516
pval  <- 1-pchisq(G,df=2)                 # 0.6532467
# no rechazamos el modelo nulo (i.e. el modelo reducido)
```

Modelo reducido

- El p-valor es grande, luego no rechazamos la hipótesis de que el modelo reducido es tan bueno como el modelo completo.
- Esto es, para explicar el peso bajo de bebés, un modelo basado en `lwt` y raza es tan bueno como uno que incluye las demás variables; por supuesto, un modelo parsimonioso es preferible a uno más complejo.

```
mred  <- glm(lbw ~ lwt+r1+r2, family=binomial)
summary(mred)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.815520   0.847607   0.962   0.3360
lwt         -0.015301   0.006461  -2.368   0.0179 *
r1           1.082278   0.488243   2.217   0.0266 *
r2           0.437738   0.359270   1.218   0.2231

    Null deviance: 232.33  on 187  degrees of freedom
Residual deviance: 221.23  on 184  degrees of freedom
AIC: 229.23

Number of Fisher Scoring iterations: 4
```

Peso al nacer

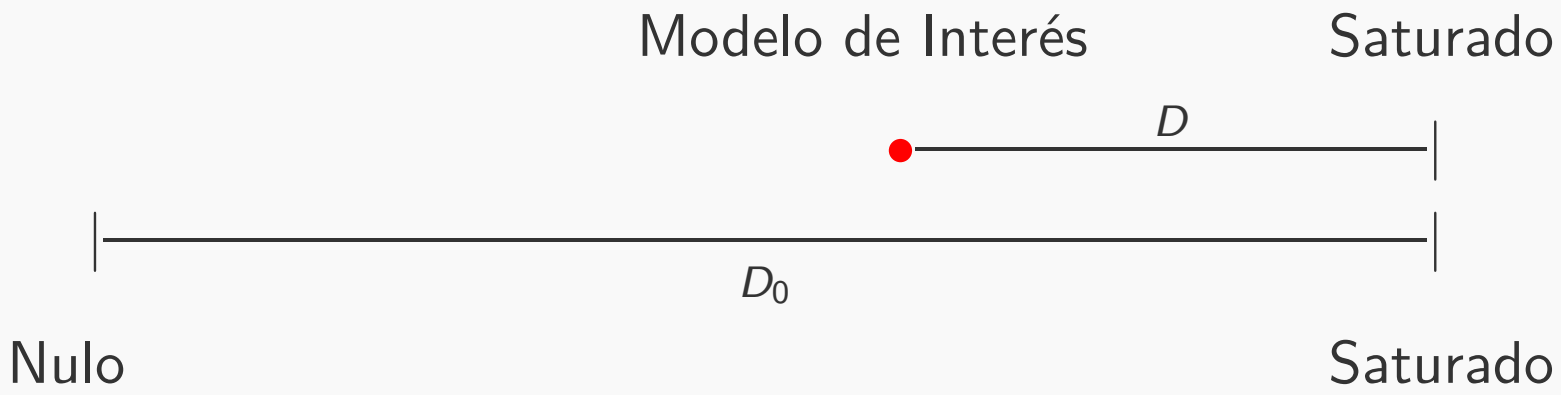


Gráfica

```
# Gráfica de peso al nacer vs peso de la madre

br <- mred$coeff
xx <- seq(80,250,length=200)
X  <- cbind(rep(1,200),xx)
p  <- 1/( 1+exp( -as.vector(X%*%br[1:2]) ) )
p1 <- 1/( 1+exp( -as.vector(X%*%(br[1:2]+c(br[3],0))) ) )
p2 <- 1/( 1+exp( -as.vector(X%*%(br[1:2]+c(br[4],0))) ) )

plot(xx,p,xlab="Peso de Madre (lbs)",
 ylab="Probabilidad de Peso Bajo de Bebé",
 ylim=c(-.1,.8), mgp=c(1.5,.5,0),cex.axis=.8,cex.lab=.8,
 cex.main=1,xlim=c(80,250),cex=.7,lwd=2,col="blue",type="l",
 main="Peso de recién nacido vs peso de madre")
lines(xx,p1,lwd=2,col="red")
lines(xx,p2,lwd=2,col="green")
rug(jitter(lwt))
legend(200,.8,legend=c("Blanca","Negra","Otra"),lwd=2,
       col=c("blue","red","green"))
```

Modelo saturado, modelo nulo

- Los modelos nulo y saturado son dos modelos extremos. En el nulo, el predictor lineal, $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, es de la forma $\eta_i = \beta_0$ y, básicamente, lo que suponemos es que $y_1, \cdots, y_n$ son i.i.d. $f(y; \theta)$, donde θ es un sólo parámetro. Por otro lado, en el modelo saturado tenemos un número maximal de parámetros (puede haber hasta el máximo número: n), que, por supuesto, será el modelo que mejor ajuste a los datos. El modelo de interés es un modelo intermedio:



Devianza

- La devianza nos ayuda a medir "distancias" entre modelos. Supongamos que el parámetro de dispersión, ϕ, satisface $a(\phi) = \phi/\omega$, donde ω es conocido. La devianza de un modelo particular se define como

$$D = -2\left[l(\widehat{\beta}) - l(\widehat{\beta}_{\text{sat}}) \right] \phi$$

donde $l(\widehat{\beta})$ es la logverosimilitud del modelo bajo consideración, evaluada en el máximo verosímil y $l(\widehat{\beta}_{\text{sat}})$ es la logverosimilitud del modelo saturado, evaluada en el estimador máximo verosímil del parámetro de ese modelo. Es claro que esta "distancia" está basada en el valor del estadístico cociente de verosimilitudes para la comparación del modelo de interés contra el saturado.

Curva ROC

- Cuando usamos un modelo de regresión logísica para clasificación, tenemos que definir el umbral, p, a partir del cual declaramos un "positivo".
- Las curvas ROC grafican las tasas TPR vs FPR para diferentes umbrales p.

$$TPR = \text{True Positive Rate} = \frac{TP}{P} = \text{"sensitividad"}$$

$$FPR = \text{False Positive Rate} = \frac{FP}{N} = 1 - \text{"especificidad"}$$

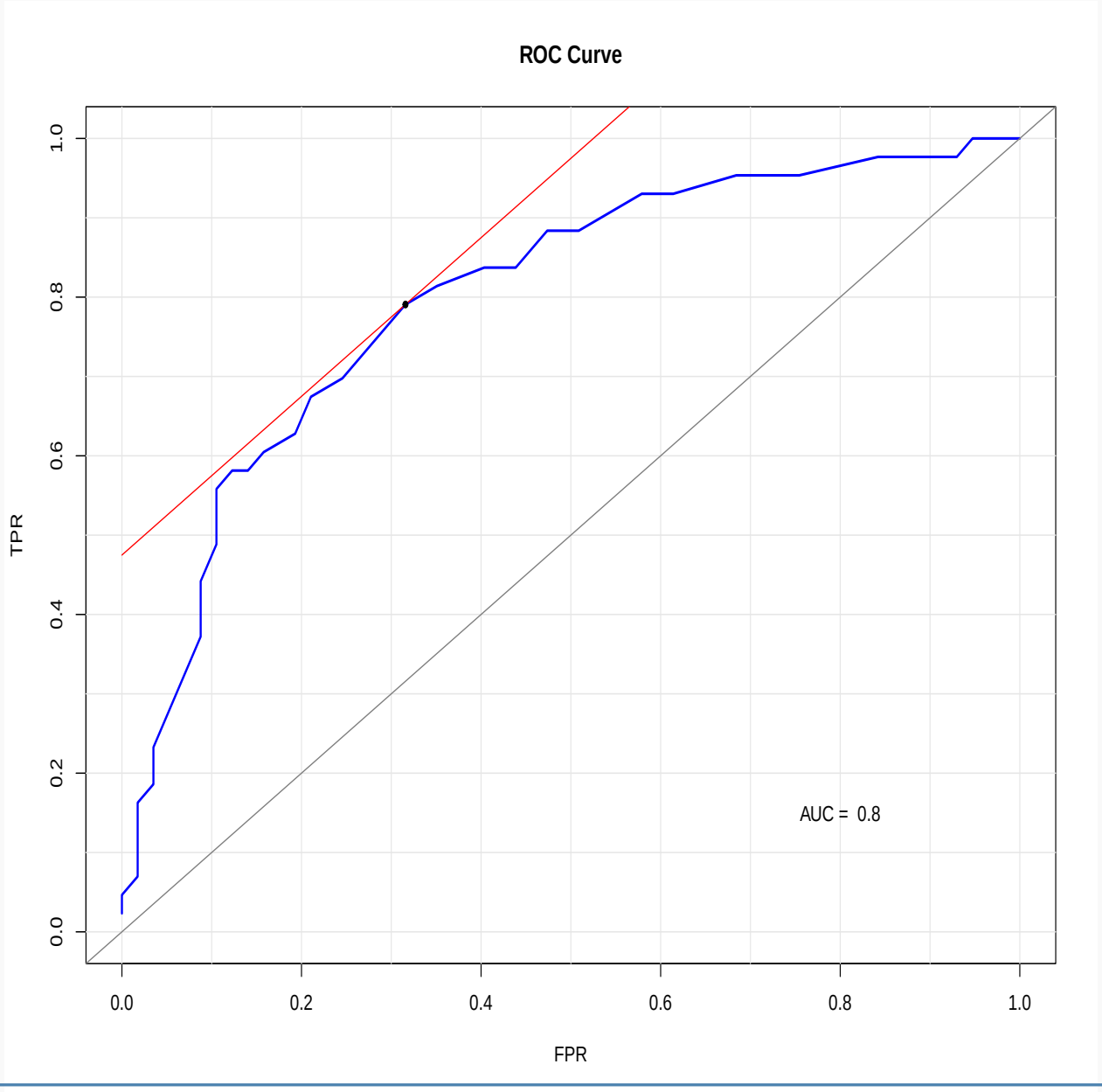
	Observados	
	1	0
Decisión = 1	TP	FP
Decisión = 0	FN	TN
	P	N

Curva ROC

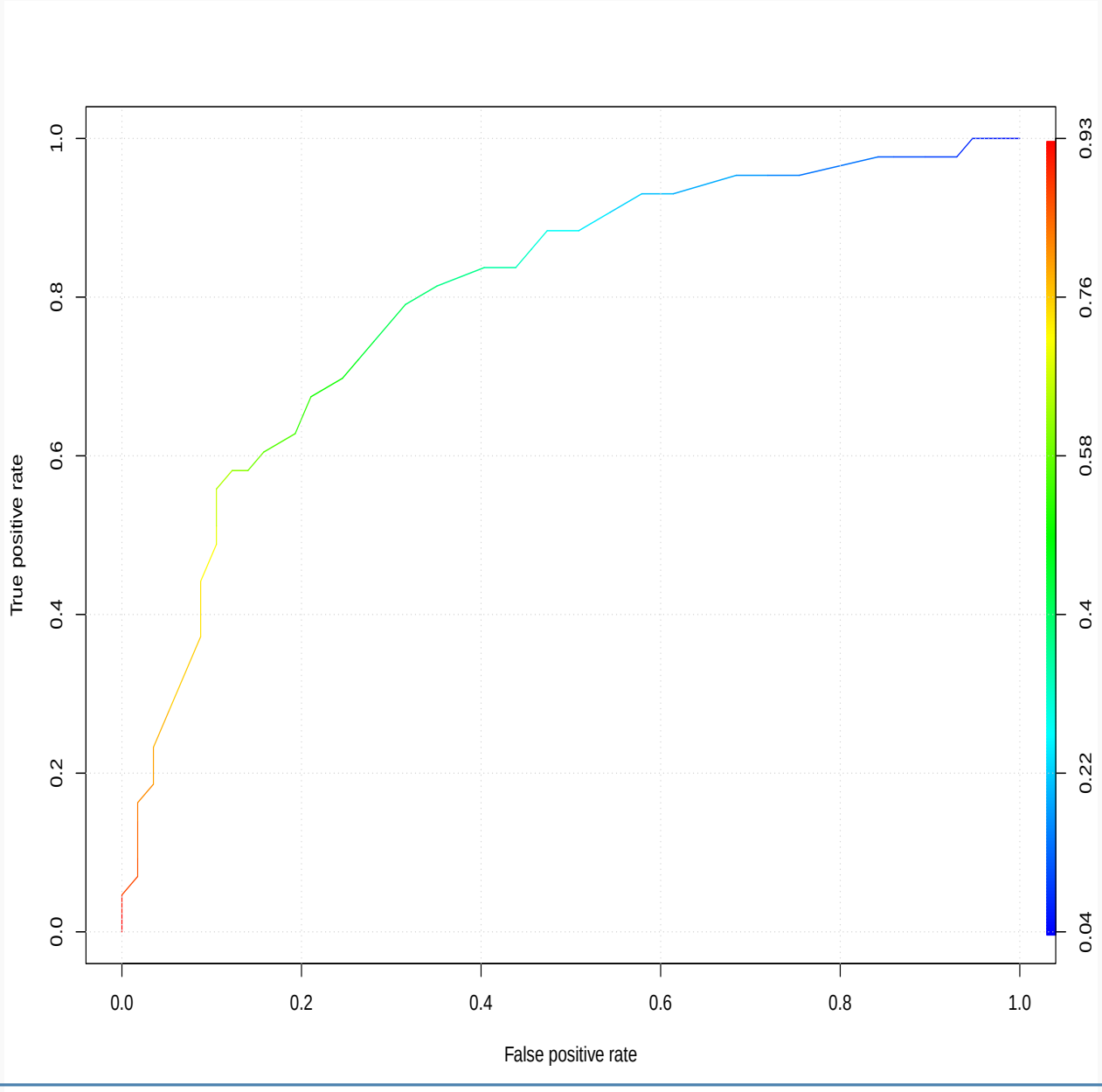
- La gráfica de *TPR* vs *FPR* puede interpretarse como una gráfica de "poder" vs "error tipo I".
- Idealmente, una regla de decisión estaría en el punto $(0, 1)$
- El área bajo la curva, AUC, puede verse, es la probabilidad de que un individuo de los positivos, tomado al azar, tenga un riesgo estimado mayor que un individuo de los negativos, tomado al azar.
- El estadístico J de Youden, es una medida que, con un sólo número, trata de capturar el desempeño de una prueba de diagnóstico. Es la máxima distancia vertical, entre la diagonal y la curva ROC, o equivalentemente

$$J = \text{sensitividad} - (1 - \text{especificidad})$$

ROC para datos de coronaria



ROC con librería ROCR



Curvas ROC en R

```
# Vectores de datos coro y edad definidos antes
n  <- length(coro)
plot(0,0, xlim=c(0,1), ylim=c(0,1), xlab="FPR", ylab="TPR",
     type="n", main="ROC Curve")
out    <- glm(coro ~ edad, family=binomial)
prob   <- out$fitted.values  # sólo se necesita la salida del glm
M      <- 101  # así que puede ser con cualquier número de predictoras
umbral <- seq(min(prob), max(prob), length=M)
ROC    <- matrix(0,M,2)
for(i in 1:M){
  yg     <- ifelse(prob >= umbral[i], 1, 0)
  yg     <- factor(yg,levels=c(0,1))
  aa     <- table(coro,yg)
  TN     <- aa[1,1]
  FP     <- aa[1,2]
  FN     <- aa[2,1]
  TP     <- aa[2,2]
  ROC[i,1] <- FP/(FP+TN)
  ROC[i,2] <- TP/(FN+TP) }

abline(h=seq(0,1,by=.1),v=seq(0,1,by=.1),col=gray(.9))
lines( ROC[,1], ROC[,2], type="l", xlab="FPR", ylab="TPR",
       lwd=2, col="blue")
```

Curvas ROC en R

```
# Cálculo del área bajo la curva
auc <- 0
for(i in 1:100){
  auc <- auc + (ROC[i,1]-ROC[i+1,1])*(ROC[i,2]+ROC[i+1,2])/2 }
# Área bajo la curva: auc = 0.7998776

# Cálculo de la probabilidad umbral
aux <- ROC[,2]-ROC[,1]
aa  <- which(aux==max(aux))              # tomar 40
umbral[ aa[2] ]   #  0.3823832

# Agregar la línea correspondiente al umbral óptimo
aau <- ROC[aa[2],2]-ROC[aa[2],1]
segments(0,aau,1,1+aau,col="red")
points(ROC[aa[2],1],ROC[aa[2],2],pch=20)
text(x=0.8, y=0.15, labels=paste("AUC = ",round(auc,3)))
abline(a=0,b=1,col=gray(.5))

ROC[aa[2],1]  # 0.315789 False Positive Rate
ROC[aa[2],2]  # 0.7906977 True Positive Rate
```

Curvas ROC en R

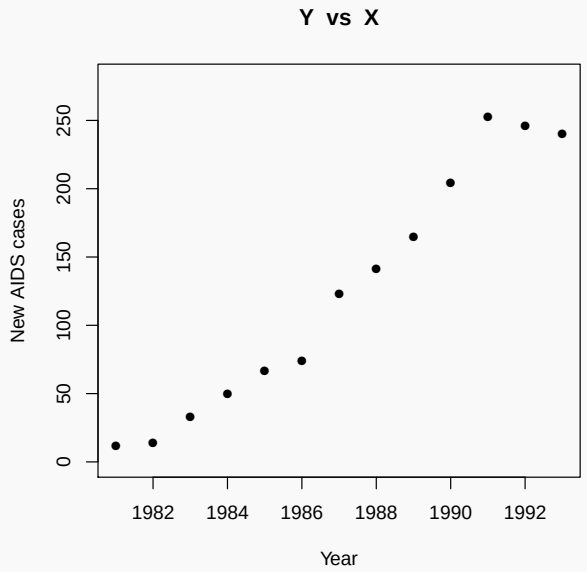
```
library(ROCR)

# Con datos de coronaria
n    <- length(coro)
xx   <- edad
X    <- cbind(rep(1,n),xx)
p    <- 1/( 1+exp( -as.vector(X%*%b) ) )
df   <- data.frame(predictions=p,labels=coro)
pred <- prediction(df$predictions, df$labels)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE)
grid()
```

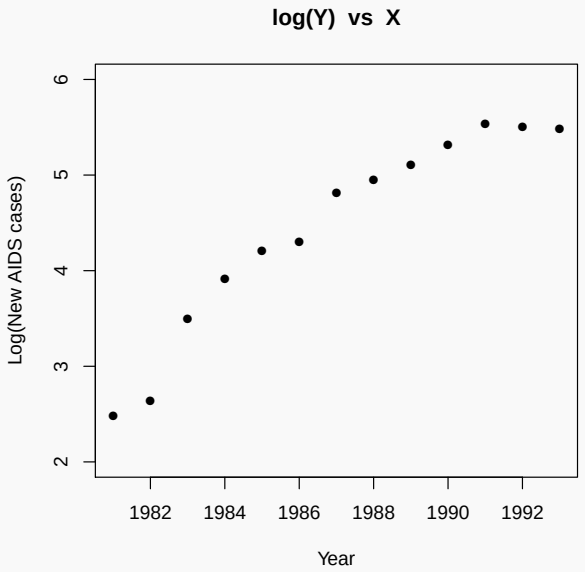
Regresión Poisson

Modelos con respuesta Poisson

- El siguiente conjunto de datos muestra el número de casos nuevos de SIDA en Bélgica (en los 80's). (Wood, S.N. (2006) Generalized Additive Models) .

```
y  <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
n  <- length(y);  x  <- 1:n
plot(x+1980,y,xlab="Year",ylab="New AIDS cases",
  ylim=c(0,280),pch=16, main="Y  vs  X")
plot(x+1980,log(y),xlab="Year",ylab="Log(New AIDS cases)",
  ylim=c(2,6), main="log(Y)  vs  X",pch=16)
```





Modelos con respuesta Poisson

- Para cada año, postulamos que el número de casos se comporta como una variable Poisson. La media, como se ve de las gráficas, depende del tiempo.
- Un par de posibles modelos, sugeridos de las gráficas anteriores
  - Liga identidad con término cuadrático para la dependencia temporal

$$y_i \sim \text{Poisson}(\lambda_i), \qquad \lambda_i = E(y_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$$

  - Liga log con término cuadrático para la dependencia temporal

$$y_i \sim \text{Poisson}(\lambda_i), \qquad \log(\lambda_i) = \log(\text{E}(y_i)) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$$

- Ambas opciones son válidas. La segunda es la llamada "liga canónica", con propiedades directas de suficiencia (veremos esto).

logverosimilitud para regresión Poisson

- Tenemos datos independientes $(y_1, x_1), \ldots, (y_n, x_n)$, con $x_i = (1, t_i, t_i^2)^T$.
- La verosimilitud es

$$L(\beta) = \prod_{i=1}^{n} e^{-\lambda_i} \lambda^{y_i} / y_i!, \qquad \text{con} \quad \log(\lambda_i) = x_i^T \beta$$

- La logverosimilitud es

$$l(\beta) = \sum_{i=1}^{n} \{y_i \log(\lambda_i) - \lambda_i\}$$

donde $\lambda_i = \exp(x_i^T \beta)$

- Igual que vimos con regresión logística, la estimación es iterativa

$$\beta_{k+1} = \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k)$$

Método de Newton

- Las expresiones a iterar, partiendo de valores iniciales β_0

$$\beta_{k+1} = \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k)$$

donde

$$H(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \lambda_i) x_i = X^T (y - \lambda)$$

y

$$\frac{\partial H(\beta_k)}{\partial \beta} = -\sum_{i=1}^{n} \lambda_i x_i x_i^T = -X^T \Lambda X$$

Código en R para regresión Poisson 1

```
y  <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
n  <- length(y)
x  <- 1:n
X  <- cbind( rep(1,n), x, x^2 )  # modelo cuadratico
b  <- c(2,1,0)      # valores iniciales de parámetros

tolm   <- 1e-6      # tolerancia (norma mínima de delta)
iterm  <- 1000      # número máximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 eta    <- as.vector(X%*%b)
 lamb   <- exp(eta)              # liga canónica log(lamb) = eta
 z      <- eta + (y-lamb)/lamb
 aa     <- as.vector(solve(t(X*lamb)%*%X, t(X*lamb)%*%z))
 delta  <- aa-b
 b      <- aa
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
```

Código en R para regresión Poisson 2

```
histo 2.0000000  1.00000000  0.0000000000
b     1.1211293  0.97953719  0.0008592407
... (14 iteraciones)
b     1.9014586  0.55600327 -0.0213462716

lsat   <- sum( y*log(y) - y - lfactorial(y) )  # logv(saturado)
eta    <- as.vector(X%*%b)
lamb   <- exp(eta)
errstd <- sqrt(diag(solve(t(X*lamb)%*%X))) # 0.186877 0.045780 0.002659
lmax   <- sum( y*log(lamb) - y - lfactorial(y) )   # logv(mod interés)
lmax   <- sum( y*log(lamb) - lamb - lfactorial(y) ) # logv(mod interés)
lambN  <- mean(y)
lnull  <- sum( y*log(lambN) - y - lfactorial(y) )  # logv(nulo)
NullD  <- -2*( lnull - lsat )   # 872.2058   con 13-1=12 gl
ResD   <- -2*( lmax - lsat )    # 9.240248   con 13-3=10 gl
AIC    <- -2*lmax + 2*3         # 96.92358

# Usando glm
m1 <- glm(y~x+I(x^2),poisson)
par(mfrow=c(2,2),mar=c(4,4,2,2))
summary(m1)
```

Código en R para regresión Poisson 3

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45903  -0.64491   0.08927   0.67117   1.54596

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.901459   0.186877  10.175  < 2e-16 ***
t            0.556003   0.045780  12.145  < 2e-16 ***
I(t^2)      -0.021346   0.002659  -8.029 9.82e-16 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 872.2058  on 12  degrees of freedom
Residual deviance:   9.2402  on 10  degrees of freedom
AIC: 96.924
Number of Fisher Scoring iterations: 4

# Comparar cuadrático vs lineal
m0 <- summary( glm(y~x,poisson) )
m1 <- summary( glm(y~x+I(x^2),poisson) )

# Estadístico de prueba
ji <- m0$deviance - m1$deviance
gl <- m0$df.residual - m1$df.residual

1-pchisq(ji,gl)   # = 0 .... se rechaza fuertemente el modelo lineal
```

Tarea 2

1 La siguiente tabla muestra conteos de células T_4 por mm^3 en muestras de sangre de 20 pacientes (en remisión) con enfermedad de Hodgkin, así como conteos en 20 pacientes en remisión de otras enfermedades. Una cuestión de interés es si existen diferencias en las distribuciones de conteos en ambos grupos.

H	396	568	1212	171	554	1104	257	435	295	397
No-H	375	375	752	208	151	116	736	192	315	1252
H	288	1004	431	795	1621	1378	902	958	1283	2415
No-H	675	700	440	771	688	426	410	979	377	503

a. Haga una comparación gráfica exploratoria de estos datos.
b. Ajuste un modelo de Poisson apropiado.
c. Usando la normalidad asintótica de los estimadores de máxima verosimilitud, dé un intervalo del 90% de confianza para la diferencia en medias. ¿Hay evidencia de diferencias en los dos grupos en cuanto a las medias de los conteos?.

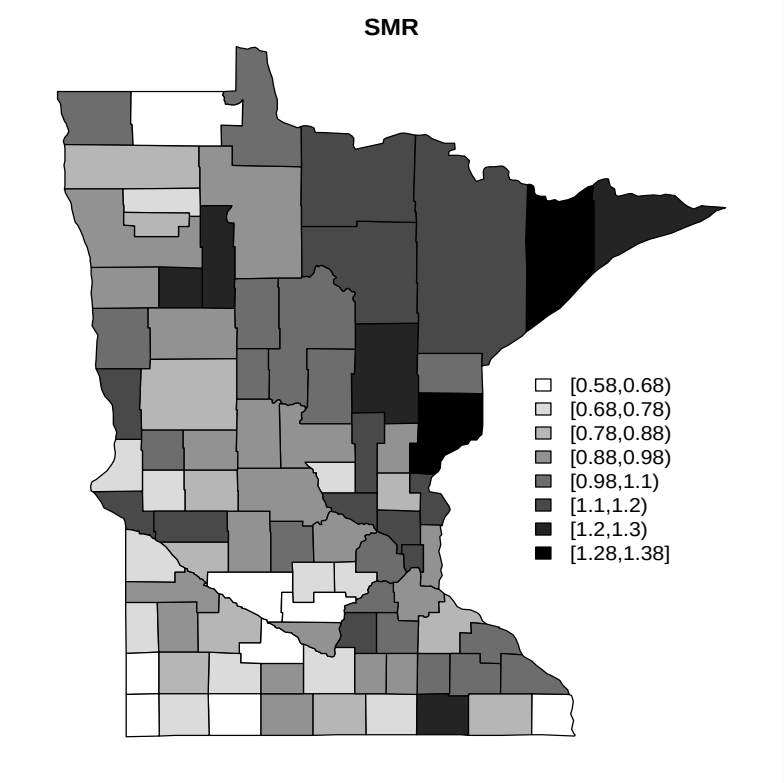
(Problema adaptado de Wakefield (2013), p.300)

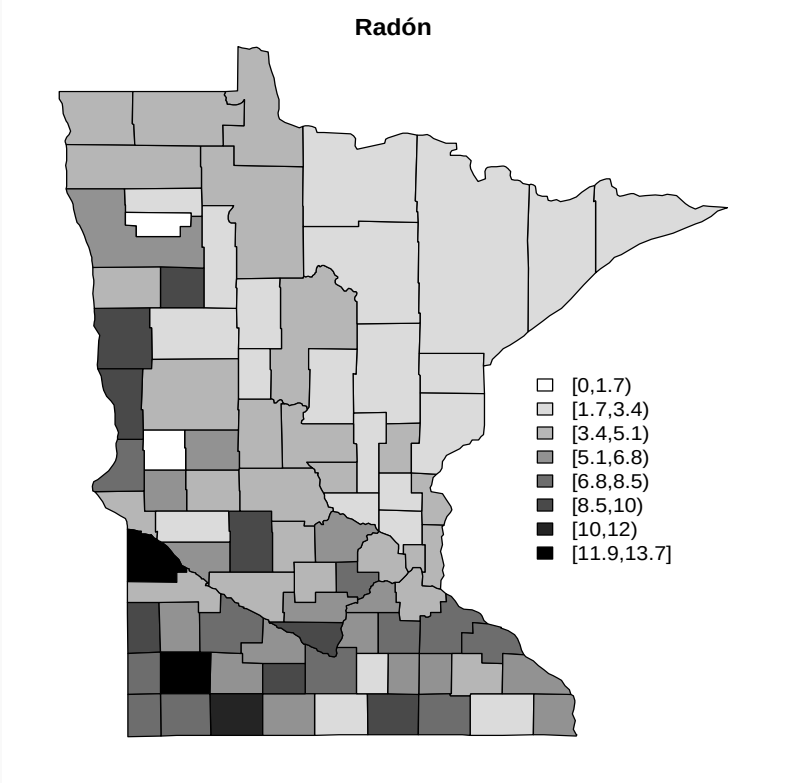
Ejemplo Poisson: Incidencia de cáncer y niveles de radón

- (Tomado de Wakefield, 2013) Interés en la asociación entre incidencia de cancer pulmonar y niveles de radón residencial por condado de Minnesota (1998-2002) (El radón ocurre en forma natural por descomposición del uranio disuelto en suelos, rocas y agua).
- Y_i conteos de casos de cáncer, x_i niveles de radón a nivel condado, $i = 1, \ldots, 87$
- Edad y sexo están fuertemente asociadas con cáncer. Una forma de controlar por estos factores, es calculando "conteos esperados" $E_i = \sum_j^J N_{ij} q_j$, donde q_j es una probabilidad ("de referencia") de caáncer en el estrato j, de modo que E_i es el número esperado de casos. De modo que cocientes $SMR_i = Y_i/E_i$ (tasa de morbilidad estandarizada) mayores a 1 nos indican incidencias mayores de lo esperado.

Ejemplo Poisson: Incidencia de cáncer y niveles de radón

- Se desea establecer relaciones entre las SMR_i's y las x_is.
- Los siguientes mapas muestran estas variables.

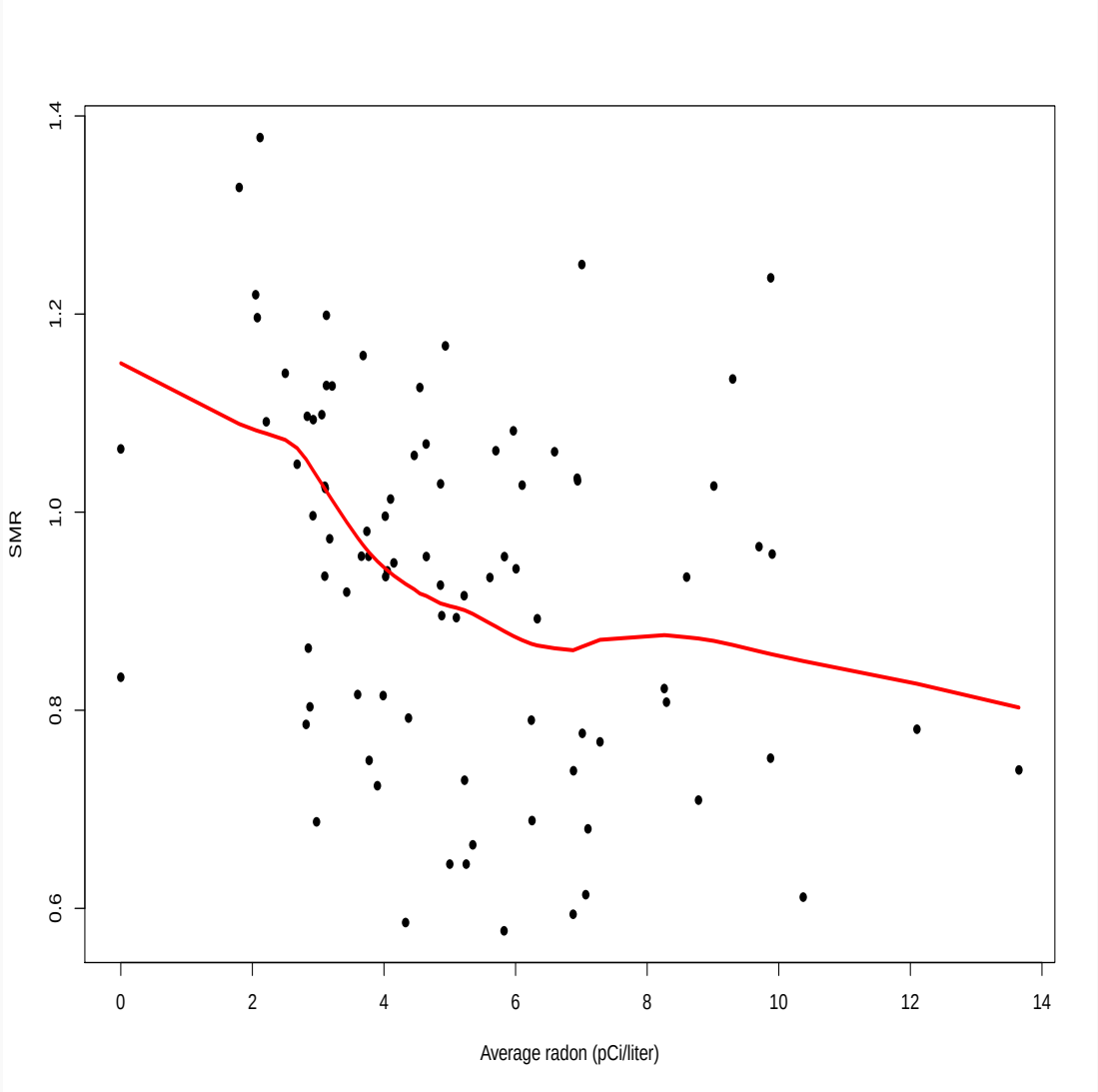




Mapas (Wakefield, 2013)

```
library(maps)
lung  <-read.table("MNlung.txt", header=TRUE, sep="\t")
radon <-read.table("MNradon.txt", header=TRUE)
MNmap <- function(data,ncol=5,figmain="",digits=5,type="e",lower=NULL, upper=NULL,udig=2){
  if (is.null(lower)) lower <- min(data)
  if (is.null(upper)) upper <- max(data)
  if (type=="q"){p <- seq(0,1,length=ncol+1)
  br <- round(quantile(data,probs=p),2)}
  if (type=="e"){br <- round(seq(lower,upper,length=ncol+1),udig)}
  shading <- gray((ncol-1):0/(ncol-1))
  data.grp <- findInterval(data,vec=br,rightmost.closed=T,all.inside=T)
  data.shad <- shading[data.grp]
  map("county", "minnesota", fill=TRUE, col=data.shad)
  leg.txt<-paste("[",br[ncol],",",br[ncol+1],"]",sep="")
  for(i in (ncol-1):1){leg.txt<-append(leg.txt,paste("[",format(br[i],digits=2),
            ",", format(br[i+1],digits=2),")",sep=""),)}
  leg.txt<-rev(leg.txt)
  legend(-91.9,46.7, legend=leg.txt,fill=shading,bty="n",ncol=1,text.width=1)
  title(main=figmain,cex=1.5);   invisible() }
Obs  <- apply(cbind(lung[,3], lung[,5]), 1, sum)
Exp  <- apply(cbind(lung[,4], lung[,6]), 1, sum)
SMR  <- Obs/Exp
par(mar=c(1,1,1,1)+.1)  # bottom/left/top/right
MNmap(SMR, ncol=8, type="e", figmain="SMR", lower=min(SMR), upper=max(SMR))
rad.avg <- rep(0, length(lung$X))
for(i in 1:length(lung$X)) { rad.avg[i] <- mean(radon[radon$county==i,2]) }
rad.avg[26] <- 0
rad.avg[63] <- 0
MNmap(rad.avg,ncol=8,type="e",figmain="Radón",lower=min(rad.avg),upper=max(rad.avg),udig=1)
plot(Obs/Exp~rad.avg,xlab="Average radon (pCi/liter)",ylab="SMR", pch=16)
lines(lowess(Obs/Exp~rad.avg), lwd=3, col="red")
```

Exploración de la relación *SMR* vs *x*



Se supone que Radón es un cancerígeno!

Relación *SMR* vs *x*

- Un posible modelo es

$$\log E\left(\frac{Y_i}{E_i}\middle| x_i\right) = \beta_0 + \beta_1 x_i$$

el cual es un modelo de regresión Poisson con "offset", i.e. una variable con coeficiente igual a 1, e.g. muy útil para conteos por unidad de área, volumen, tiempo, etc.

$$\log E\left(Y_i|x_i\right) = \beta_0 + \beta_1 x_i + \log E_i$$

```
x        <- rad.avg
x[26]    <- NA
x[63]    <- NA
poismod  <- glm(Obs~offset(log(Exp))+rad.avg,family="poisson")

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.164144   0.026301   6.241 4.35e-10 ***
rad.avg     -0.034922   0.005302  -6.586 4.51e-11 ***

    Null deviance: 290.43  on 86  degrees of freedom
Residual deviance: 245.87  on 85  degrees of freedom
AIC: 785.77
Number of Fisher Scoring iterations: 4
```