Modelos con respuesta Binaria: La Regresión Logística

José Luis Vicente Villardón

Departamento de Estadística
Universidad de Salamanca

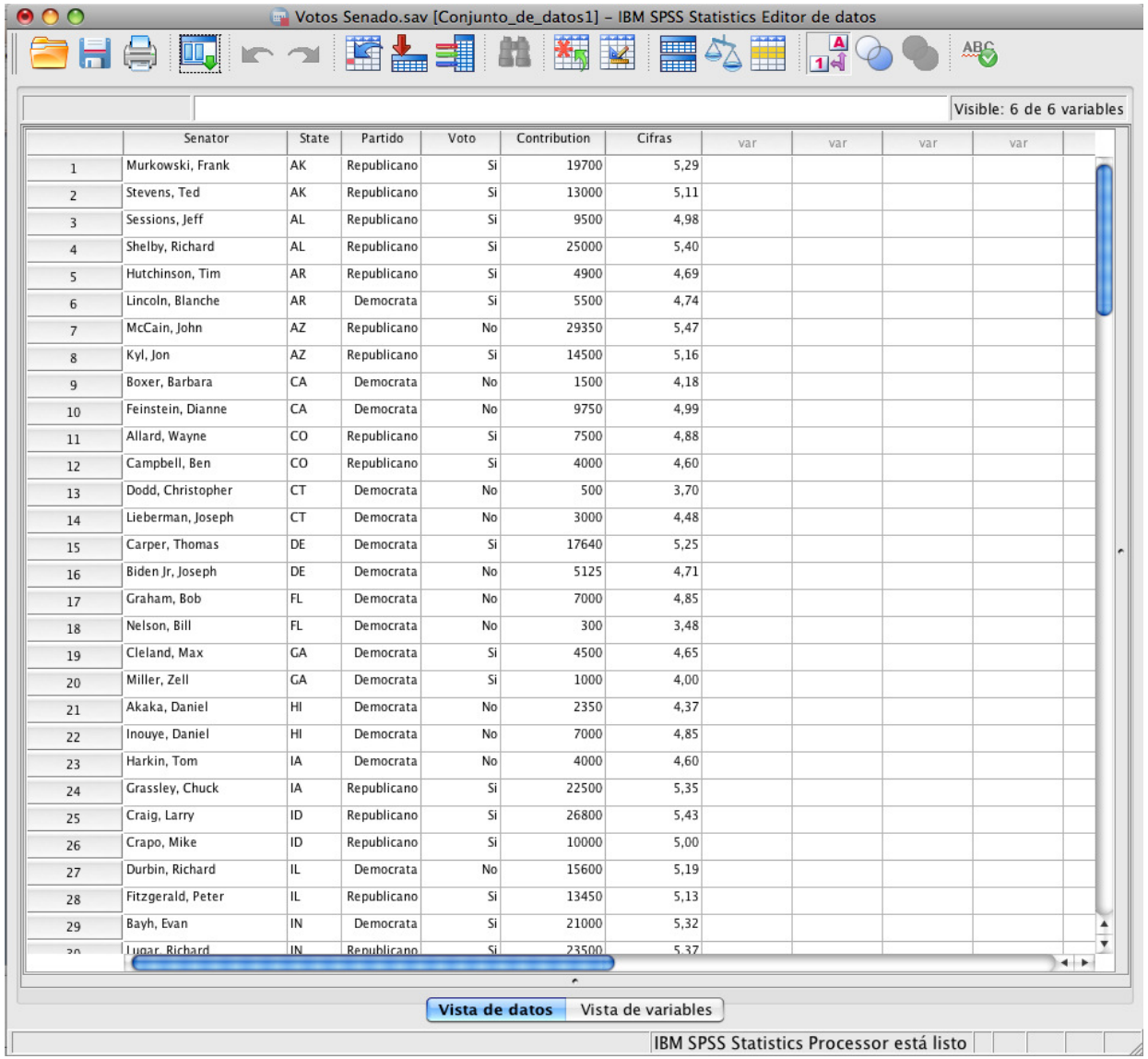
Análisis Multivariante
May 9, 2018

Ejemplo Inicial

At the end of February 2002 the U.S. Senate considered comprehensive energy legislation. Senators John McCain and John Kerry proposed raising the Corporate Average Fuel Economy (CAFE) standard for cars and trucks. On March 13, 2002 the United States Senate voted on the Levin amendment (No. 2997), charging the National Highway Traffic Safety Administration with the development of a new standard and effectively shelving the McCain/Kerry proposal. The dataset consists of information about each of the 100 U.S. Senators regarding their vote on the Levin amendment. A senator's vote (Vote) is the response variable. Provided explanatory variables include the state represented, political party affiliation (Party), and the lifetime total amount of contributions received from auto manufacturers (Amount).

- **Objetivo General** : Análisis del Voto
- **Objetivo Específico** : Calcular la probabilidad de voto positivo en función de variables relacionadas con el votante o clasificarlo en voto positivo ó negativo.

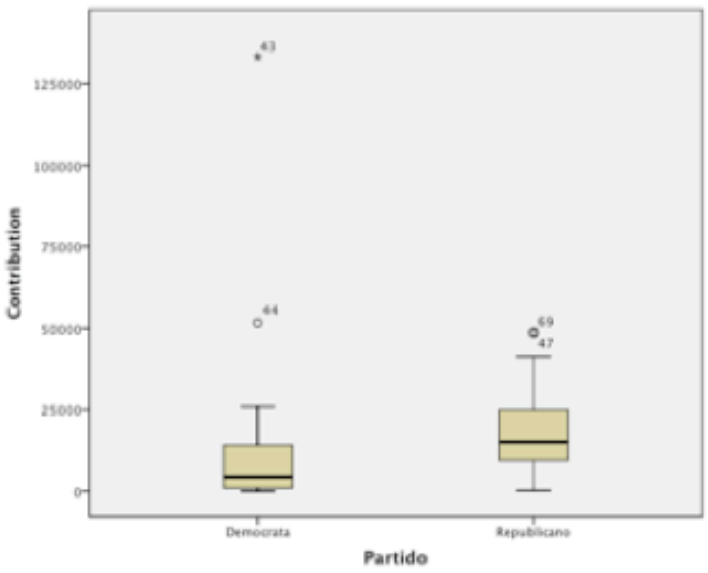
Datos

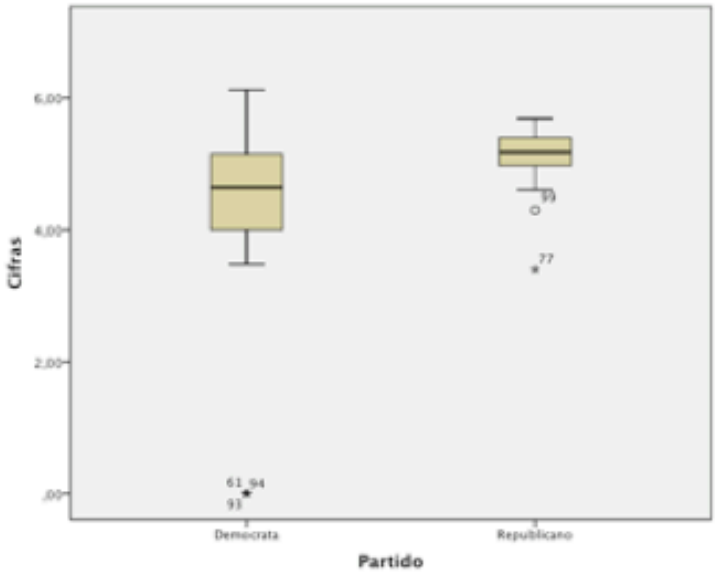


	Senator	State	Partido	Voto	Contribution	Cifras	var	var	var	var
1	Murkowski, Frank	AK	Republicano	Si	19700	5,29				
2	Stevens, Ted	AK	Republicano	Si	13000	5,11				
3	Sessions, Jeff	AL	Republicano	Si	9500	4,98				
4	Shelby, Richard	AL	Republicano	Si	25000	5,40				
5	Hutchinson, Tim	AR	Republicano	Si	4900	4,69				
6	Lincoln, Blanche	AR	Democrata	Si	5500	4,74				
7	McCain, John	AZ	Republicano	No	29350	5,47				
8	Kyl, Jon	AZ	Republicano	Si	14500	5,16				
9	Boxer, Barbara	CA	Democrata	No	1500	4,18				
10	Feinstein, Dianne	CA	Democrata	No	9750	4,99				
11	Allard, Wayne	CO	Republicano	Si	7500	4,88				
12	Campbell, Ben	CO	Republicano	Si	4000	4,60				
13	Dodd, Christopher	CT	Democrata	No	500	3,70				
14	Lieberman, Joseph	CT	Democrata	No	3000	4,48				
15	Carper, Thomas	DE	Democrata	Si	17640	5,25				
16	Biden Jr, Joseph	DE	Democrata	No	5125	4,71				
17	Graham, Bob	FL	Democrata	No	7000	4,85				
18	Nelson, Bill	FL	Democrata	No	300	3,48				
19	Cleland, Max	GA	Democrata	Si	4500	4,65				
20	Miller, Zell	GA	Democrata	Si	1000	4,00				
21	Akaka, Daniel	HI	Democrata	No	2350	4,37				
22	Inouye, Daniel	HI	Democrata	No	7000	4,85				
23	Harkin, Tom	IA	Democrata	No	4000	4,60				
24	Grassley, Chuck	IA	Republicano	Si	22500	5,35				
25	Craig, Larry	ID	Republicano	Si	26800	5,43				
26	Crapo, Mike	ID	Republicano	Si	10000	5,00				
27	Durbin, Richard	IL	Democrata	No	15600	5,19				
28	Fitzgerald, Peter	IL	Republicano	Si	13450	5,13				
29	Bayh, Evan	IN	Democrata	Si	21000	5,32				
30	Lugar, Richard	IN	Republicano	Si	23500	5,37				

Descripción inicial

La cantidad de dinero se ha transformado como $\log_{10}(10x + 1)$ para simetrizar la variable ya que hay un senador (o dos) que reciben cantidades elevadas comparados con los demás.





Modelo de predicción

- En este caso la variable respuesta (o variable dependiente) es el voto.
- Codificaremos numéricamente la variable como un 0 para el voto negativo y un 1 para el positivo.
- Podemos entender estos valores como la probabilidad observada de voto positivo.
- En general esto es así para cualquier variable cualitativa que toma solamente dos categorías.
- Trataremos de construir un modelo que nos ponga la probabilidad de voto observada en función del partido y de la cantidad de dinero recibida por el senador.
- Para un senador hipotético de un determinado partido que recibe una cantidad de dinero concreta es posible estimar la probabilidad de voto positivo y clasificarlo como *votante a favor* si esta es mayor de 0.5.

Diagrama de dispersión

Es obvio que, en este caso, un modelo regresión lineal tradicional no es valido ya que nos proporcionaría probabilidades observadas mayores de 1 y menores de cero (negativas) que no recogen la naturaleza de nuestra variable respuesta.

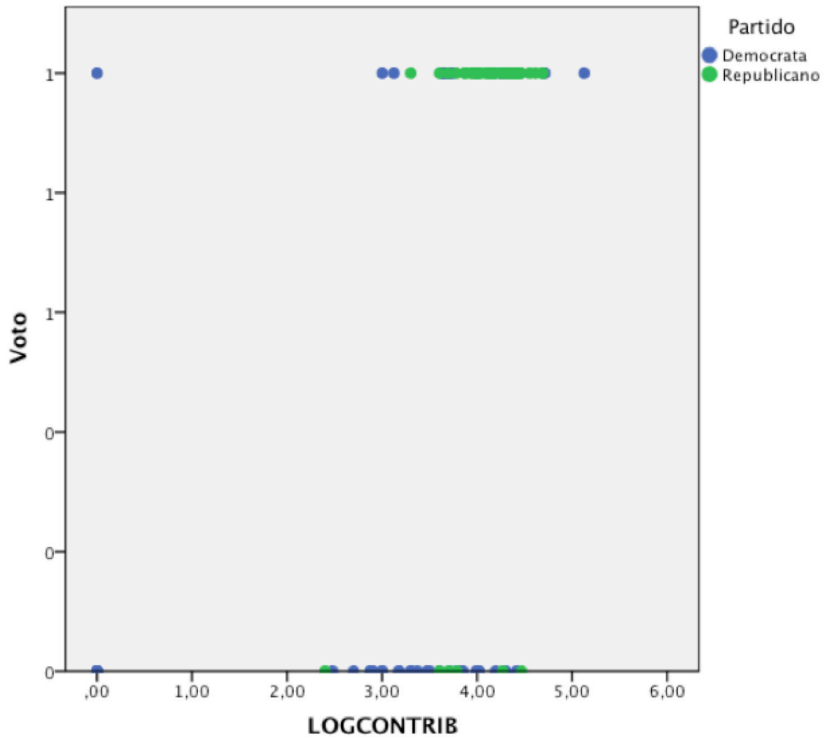


Tabla de frecuencias

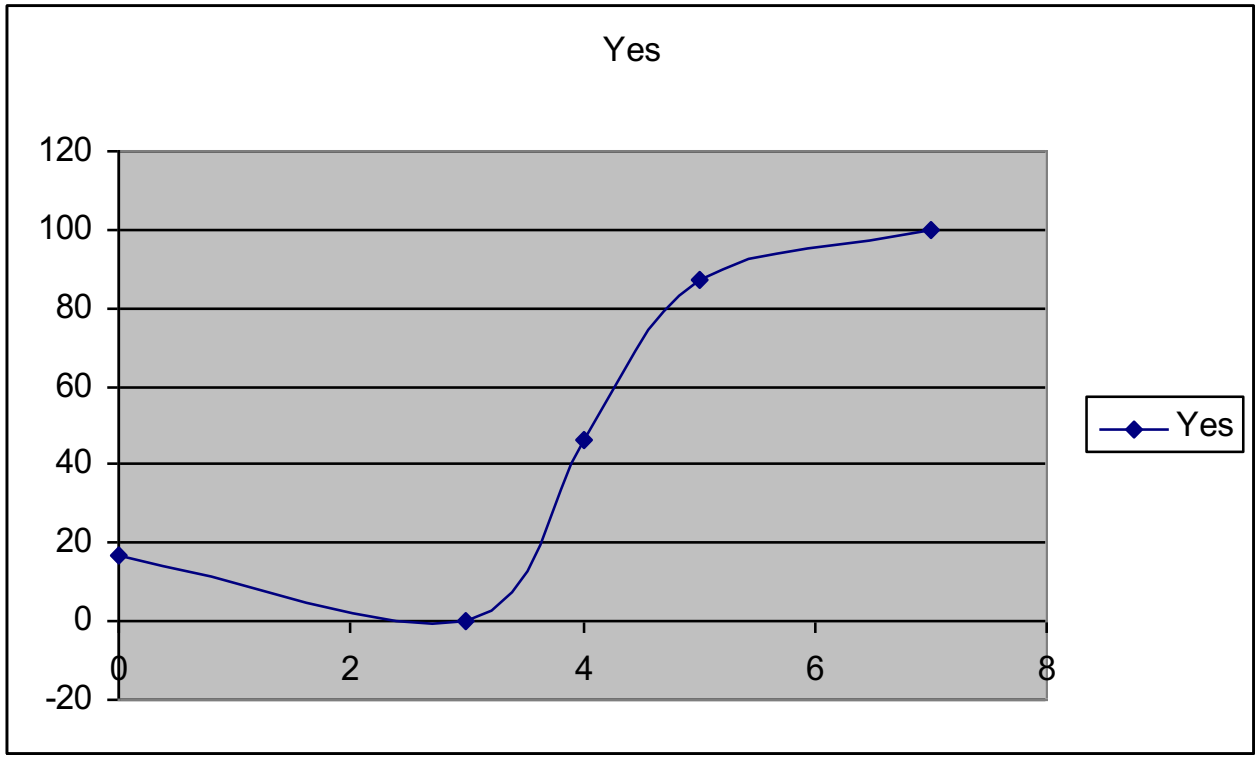
Agrupemos las cantidades numéricas (en cifras) y veamos como se comportan los porcentajes de voto positivo para cada cantidad.

ContribInterv * Vote Crosstabulation

			Vote		
			NO	YES	Total
ContribInterv	0	Count	5	1	6
		% within ContribInterv	83,3%	16,7%	100,0%
	3	Count	6	0	6
		% within ContribInterv	100,0%	,0%	100,0%
	4	Count	21	18	39
		% within ContribInterv	53,8%	46,2%	100,0%
	5	Count	6	42	48
		% within ContribInterv	12,5%	87,5%	100,0%
	7	Count	0	1	1
		% within ContribInterv	,0%	100,0%	100,0%
Total		Count	38	62	100
		% within ContribInterv	38,0%	62,0%	100,0%

Representación de los porcentajes

De forma general, los porcentajes de voto positivo van creciendo con el número de cifras de la cantidad recibida por el Senador.



La curva tiene forma sigmoide.

Curva Logística

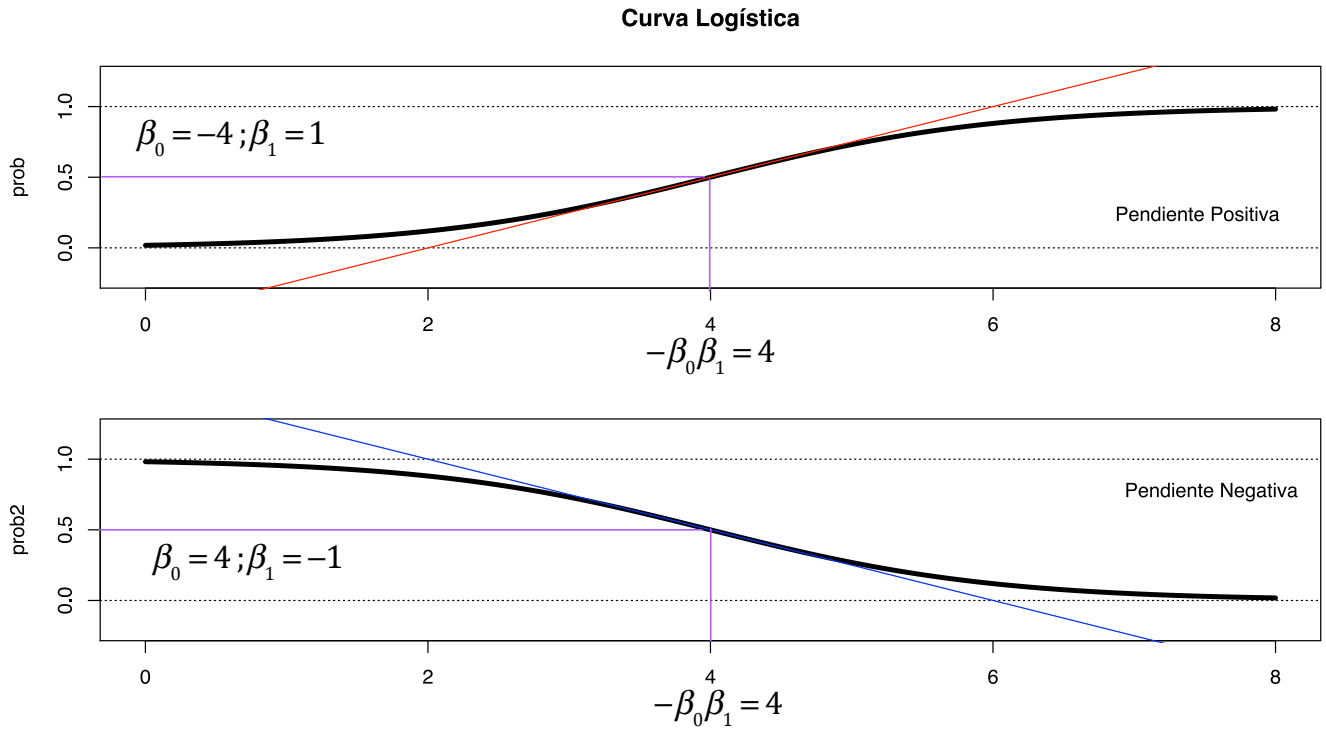
Sa entonces Y la variable respuesta (probabilidad de voto positivo) que toma valores 0 y 1. Si llamamos p_i a la probabilidad esperada de que un senador vote positivamente cuando recibe una cantidad de dinero $X = x_i$, la curva logística tiene la forma deseada y podemos escribirla como:

$$p_i = P(Y = 1/X = x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Hay otras alternativas a la curva logística, pero esta es la más utilizada en la práctica.

Interpretación de los parámetros de la curva

La curva logística simple depende de dos parámetros β_0 y β_1. $\frac{1}{4}\beta_1$ es la pendiente de la recta tangente a la curva en el punto para el que la probabilidad esperada es 0.5



El valor de X para el que se alcanza la probabilidad 0,5 es $-\beta_0\beta_1$.

Logística Múltiple

Normalmente disponemos de un conjunto de varios predictores $X = (X_1, \ldots, X_p)$. Para cada individuo (i) tenemos un vector de observaciones $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. El modelo podemos escribirlo ahora como:

$$p_i = P(Y = 1/\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}$$

Para incluir el término independiente el vector de observaciones lo tomamos como: $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$. Para simplificar la notación llamaremos $M_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ al término lineal en la ecuación, o bien $M_i = \mathbf{x'}_i\beta$ donde $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ es el vector de parámetros. El modelo de regresión logística es un modelo lineal en escala *logit*.

$$\text{logit}(p_i) = \log(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = M_i$$

Estimación de los parámetros: Máxima Verosimilitud

Para cada individuo tenemos un valor observado y_i que procede de una variable Y_i que tiene una distribución de Bernouilli de parámetro p_i $(i = 1, \ldots, n)$, es decir

$$P(Y_i = y_i/X = x_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

con

$$p_i = P\,(Y = 1/X = x_i) = \frac{e^{M_i}}{1 + e^{M_i}} = \frac{1}{1 + e^{-M_i}}$$

La función de probabilidad conjunta de la muestra completa (función de verosimilitud) es

$$L(\beta/\mathbf{x}_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)} = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i}(1 - p_i)$$

ya que suponemos que las respuesta de los individuos son mutuamente independientes.

Estimación de los parámetros: Máxima Verosimilitud

Teniendo en cuenta que $\frac{p_i}{1-p_i} = e^{\mathbf{x'}_i\beta}$ y $1 - p_i = \frac{1}{1+e^{\mathbf{x'}_i\beta}}$, podemos escribir

$$L(\beta/\mathbf{x}_i) = \prod_{i=1}^{n} \left(e^{\mathbf{x'}_i\beta}\right)^{y_i} \left(\frac{1}{1+e^{\mathbf{x'}_i\beta}}\right)$$

Buscamos los valores de los parámetros que hacen máxima la verosimilitud, estos valores hacen máximo también el logaritmo de la misma, que puede escribirse como

$$\log L(\beta/\mathbf{x}_i) = \sum_{i=1}^{n} y_i \mathbf{x'}_i\beta - \sum_{i=1}^{n} \log(1+e^{\mathbf{x'}_i\beta})$$

Derivamos la función con respecto a los parámetros, igualamos las derivadas a cero y resolvemos el sistema resultante.

Estimación de los parámetros: Máxima Verosimilitud

La derivada de esta función con respecto a β puede escribirse como:

$$\frac{\partial \log L(\beta/\mathbf{x}_i)}{\partial \beta} = \sum_{i=1}^{n} y_i \mathbf{x}'_i - \sum_{i=1}^{n} \left(\frac{e^{\mathbf{x}'_i \beta}}{1 + e^{\mathbf{x}'_i \beta}} \right) \mathbf{x}_i$$

Igualando el vector de derivadas parciales a cero para maximizar el logaritmo de la verosimilitud, se tiene

$$\sum_{i=1}^{n} \left(\frac{e^{\mathbf{x}'_i \beta}}{1 + e^{\mathbf{x}'_i \beta}} \right) \mathbf{x}_i = \sum_{i=1}^{n} y_i \mathbf{x}'_i$$

Tenemos entonces un sistema de ecuaciones no lineales que debemos resolver mediante métodos numéricos. La solución $\hat{\beta}$ serán los valores de los estimadores.

Estimación de los parámetros: Máxima Verosimilitud

Como se trata del estimador máximo-verosímil $\hat{\beta}$ de β , su matriz de covarianzas asintótica puede obtenerse de la matriz de información

$$I(\beta) = -E\left[\frac{\partial^2 \log L(\beta/\mathbf{x}_i)}{\partial\beta\partial\beta'}\right]$$

evaluada en $\beta = \hat{\beta}$. Calculando las derivadas y haciendo las operaciones oportunas, la matriz de varianzas-covarianzas de los estimadores es.

$$Cov(\hat{\beta}) = \left[\sum_{i=1}^{n}\left(\frac{e^{\mathbf{x}'_i\beta}}{(1+e^{\mathbf{x}'_i\beta})^2}\right)\mathbf{x}_i\mathbf{x}_j\right]^{-1}$$

$$= \left[\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)\mathbf{x}_i\mathbf{x}_j\right]^{-1} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$$

donde $\mathbf{V} = Diag\left\{\hat{p}_i(1-\hat{p}_i)\right\}$ contiene las varianzas estimadas de la observaciones.

Estimación de los parámetros: Máxima Verosimilitud

Las raíces cuadradas de los elementos de la diagonal de $Cov(\hat{\beta})$ son los errores estándar asintóticos estimados de los coeficientes del modelo que pueden usarse para calcular intervalos de confianza y contrastes para muestras grandes.
La forma más común en Estadística es resolver el sistema usando el método de Newton-Raphson, obteniendo el algoritmo que describimos a continuación:

Estimación de los parámetros: Newton-Raphson

- Comenzamos con estimadores iniciales $\hat{\beta}_0$, por ejemplo $\hat{\beta}_0 = \mathbf{0}$. Iniciamos el contador de iteraciones en $l = 0$
- Aumentamos el contador de iteraciones en una unidad $l = l + 1$
- Calculamos nuevos estimadores

$$\hat{\beta}_{l+1} = \hat{\beta}_l + (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \hat{\mathbf{p}}_l)$$

donde $\hat{\mathbf{p}}_l = (\hat{p}_{l1}, \ldots, \hat{p}_{ln})'$ con $\hat{p}_{li} = \frac{1}{1+e^{-\mathbf{x}'_i\hat{\beta}_l}}$ y $\mathbf{V}_l = Diag\left\{\hat{p}_{li}(1 - \hat{p}_{li})\right\}$.

- Las iteraciones se repiten hasta que la diferencia entre los estimadores en dos pasos sea menor que un índice de tolerancia ϵ ($\left|\hat{\beta}_{l+1} - \hat{\beta}_l\right| < \varepsilon$) hasta que la diferencia en las vero¡similitudes de dos iteraciones sucesivas esté por debajo del índice de tolerancia o hasta que se alcancen un número prefijado de iteraciones

Estimación de los parámetros: Newton-Raphson

En el proceso de estimación con Newton-Raphson podemos encontrarnos algunos problemas

- La matriz de información es singular. Por ejemplo, cuando se producen predicciones perfectas.
- Máximos locales.
- Necesitamos mayor número de individuos que de variables.
- Es conveniente estandarizar los predictores si tenemos escalas muy distintas.
- Si tenemos un predictor categórico es posible que el modelo sea inestable si alguna de las categorías tiene pocos casos.

Bondad del ajuste

Para este modelo utilizaremos un contraste de bondad de ajuste que resulta de comparar el modelo ajustado con el modelo saturado o modelo completo que explica completamente las observaciones. Si llamamos L_c a la verosimilitud del modelo en estudio (current model) y L_f a la verosimilitud del modelo completo o "full model", la diferencia mide la proximidad del modelo ajustado con el modelo perfecto, es decir, cuanto menor sea la diferencia, mejor se ajusta el modelo a los datos. El contraste de razón de verosimilitudes para contrastar el ajuste del modelo tiene como estadístico de contraste

$$D = -2 \log \left(\frac{L_c}{L_f} \right)$$

al que denominamos Deviance y que sigue asintóticamente una distribución chi-cuadrado con $n - p - 1$ grados de libertad.

Comparación de modelos

El estadístico *Deviance* se utiliza para comparar modelos de forma jerárquica.
Supongamos que tenemos dos modelos M_1 y M_2 para el mismo conjunto de datos, y que el modelo M_1 está incluido en el modelo M_2, es decir, M_1 tiene menos parámetros que M_2.
Supongamos que hemos obtenidos valores de la deviance D_1 con grados ν_1 de libertad y D_2 con ν_2 grados de libertad respectivamente $(\nu_1 > \nu_2)$.
La diferencia entre las dos $D = D_1 - D_2$ sigue asintóticamente una distribución chi-cuadrado con $(\nu_1 - \nu_2)$ grados de libertad y sirve para contrastar si el modelo más grande es significativamente mejor que el otro.

Análisis de la Varianza

Construiremos un contraste análogo al del análisis de la varianza en el modelo lineal, es decir, un test para comparar el modelo con sólo la constante y el modelo con todas las variables.
Supongamos que

$$M_1 = \beta_0$$

y

$$M_2 = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

.
Llamando D_{β_0} a la Deviance para el modelo con sólo la constante y $D_{\beta_0,\beta_1,\ldots,\beta_p}$ a la del modelo con el conjunto completo de variables, la diferencia $D_{\beta_0} - D_{\beta_0,\beta_1,\ldots,\beta_p}$ sigue una distribución chi-cuadrado con p grados de libertad.

Test de Wald

Sirve para contrastar la igualdad acero de los parámetros del modelo.
Se divide el valor del estimador entre su error estándar y se compara con el percentil de una normal estándar. El error estándar de la estimación se obtiene de la matriz de covarianzas asintótica que se obtuvo en el proceso de estimación.

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \rightarrow N(0,1)$$

Algunos programas elevan ese valor al cuadrado y lo comparan con el percentil de una Ji-Cuadrado, con un grado de libertad.

$$W^2 = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2 \rightarrow \chi^2_1$$

Pseudo R^2

Es posible calcular análogos del coeficiente de determinación (R^2) en una regresión lineal. Llamando L_{null} a la Deviance del modelo con sólo la constante, tenemos

- Efron

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{p}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- McFadden

$$R^2 = 1 - \frac{L_c}{L_{null}}$$

- Cox-Snell

$$R^2 = 1 - \left[\frac{L_{null}}{L_c}\right]^{2/n}$$

- Nagelkerke

$$R^2 = \frac{1 - \left[\frac{L_{null}}{L_c}\right]^{2/n}}{1 - (L_{null})^{2/n}}$$

Test de Hosmer y Lemeshow

Contraste de bondad de ajuste:

- Para el grupo de individuos analizados las probabilidades ajustadas se dividen en g grupos y en cada uno se calculan las frecuencias observadas y esperadas que se comparan con un test de bondad de ajuste ji-cuadrado.

$$\chi^2_{HL} = \sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(O_{kl} - E_{kl})^2}{E_{kl}} \equiv \chi^2_{(g-1)}$$

- La frecuencia esperada se calcula a partir de la probabilidad media de los individuos que están en cada intervalo.
- El modelo se ajusta bien a los datos cuando no es estadísticamente significativo. (Las frecuencias observadas y esperadas son próximas)

Salida Típica: Voto en función del número de cifras

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	23,720	1	,000
	Bloque	23,720	1	,000
	Modelo	23,720	1	,000

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Cifras	1,468	,453	10,500	1	,001	4,341	1,786	10,550
	Constante	-6,496	2,208	8,654	1	,003	,002		

a. Variable(s) introducida(s) en el paso 1: Cifras.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13,391	8	,099

Tabla de clasificación[a]

			Pronosticado		
			Voto		
	Observado		No	Si	Porcentaje correcto
Paso 1	Voto	No	18	20	47,4
		Si	4	58	93,5
	Porcentaje global				76,0

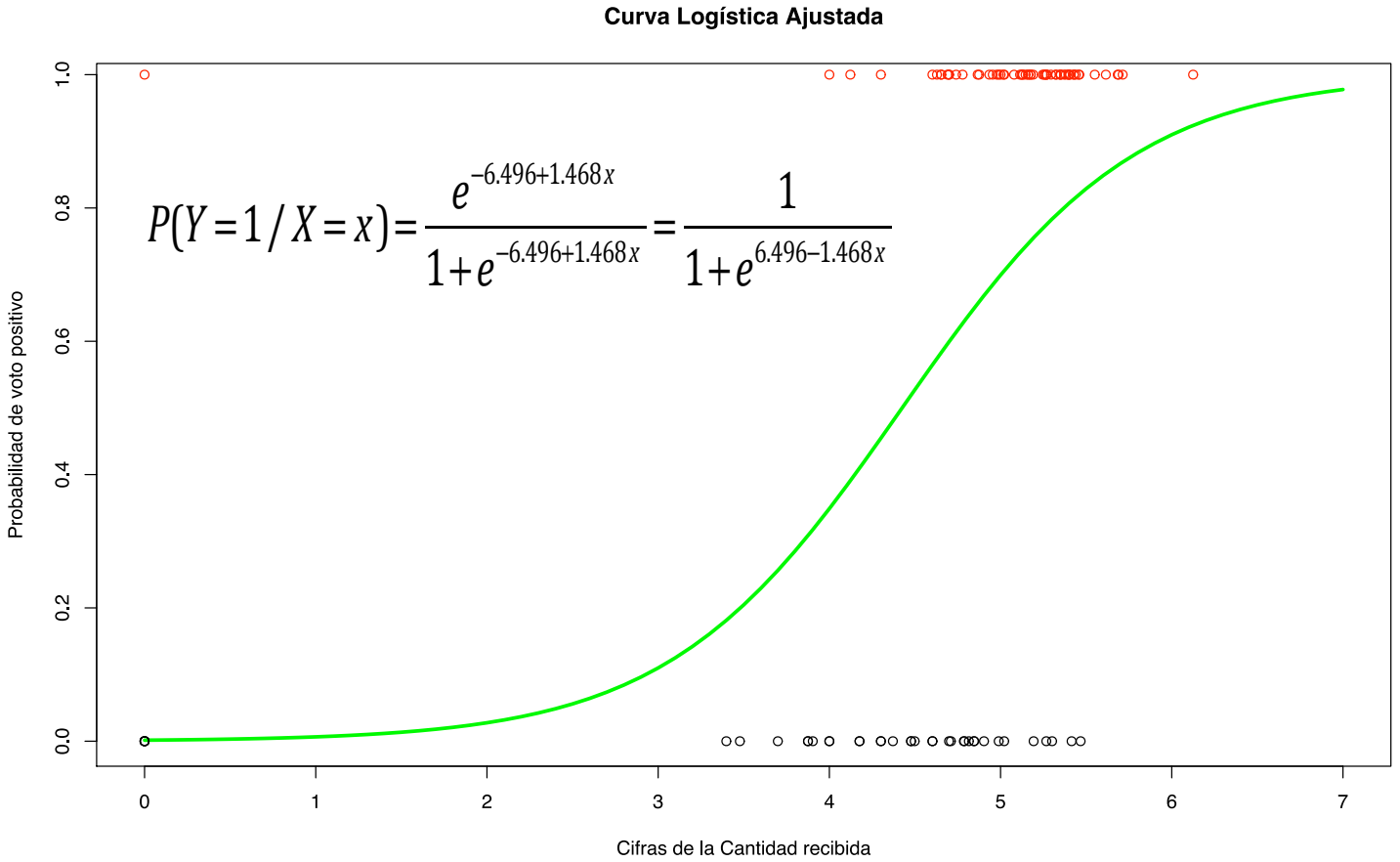
a. El valor de corte es ,500

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	109,093[a]	,211	,287

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Curva Ajustada



Interpretación

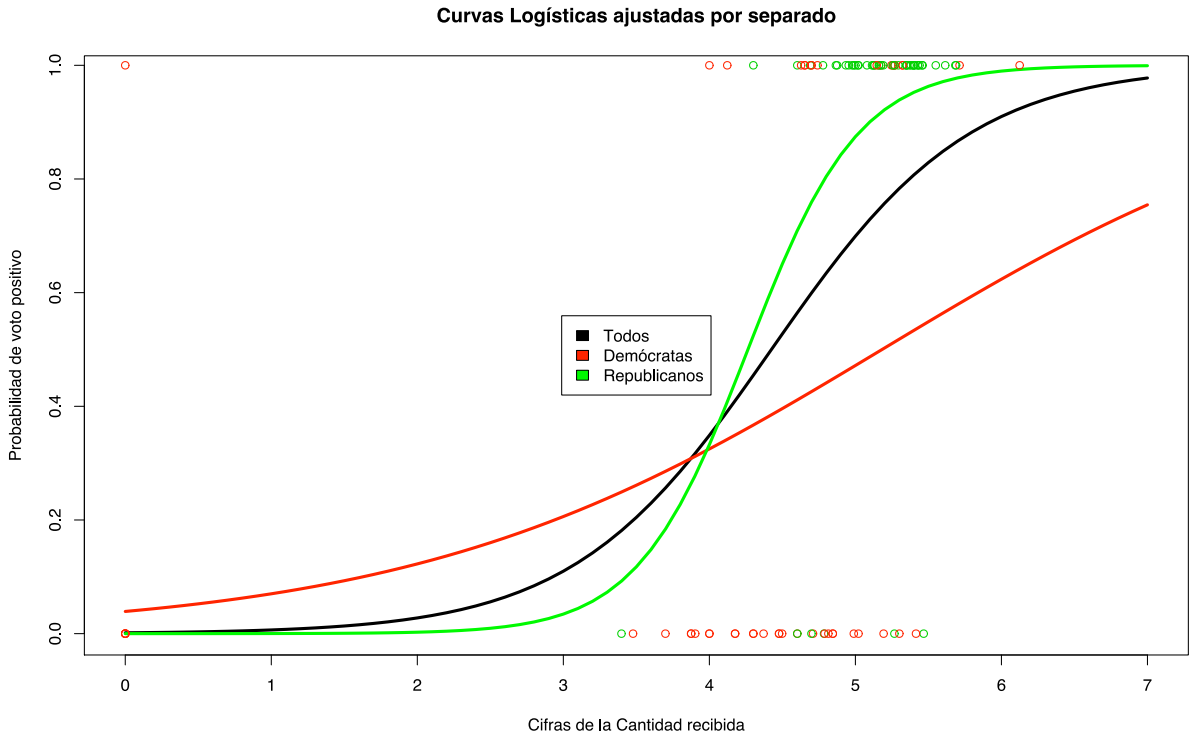
- La cantidad de dinero recibida en la campaña tiene un efecto significativo sobre el voto (Prruebas Omnibus-Modelo, chi-cuadrado=23.720, p-valor<0.0005)
- El incremento de una una unidad en el número de cifras produce un incremento de 1.468 unidades en el odds de la probabilidad de voto positivo frente a negativo.
- El test de Wald para el número de cifras proporciona un contraste estadísticamente significativo.
- El modelo se ajusta bien a los datos (Test de Hosmer y Lemeshow, chi-cuadrado=13.391, p-valor = 0.099) Recuérdese que es un test de bondad de juste.
- El pseudo R^2 de nagelkerke es 0.287. La variabilidad explicada no es muy alta, sibien en este tipo de modelos suele ser aceptable.
- Se necesita un número de 4.425 cifras (2661 dólares) para que sea más probable el voto positivo.

Discriminante Logístico

El modelo de Regresión Logística puede utilizarse como un modelo discriminante. La regla de clasificación es muy sencilla, clasificamos al grupo cuya probabilidad de pertenencia estimada por el modelo sea más alta. Eso es equivalente a clasificar en un grupo si la probabilidad de pertenencia es mayor de 0.5.
Si los grupos no están balanceados es posible utilizar un punto de corte distinto. Se pueden probar varios y elegir aquel que tenga una probabilidad más alta de clasificación correcta.
La tabla de confusión es una tabla de frecuencias que cruza las observaciones con las predicciones de la respuesta.
El 76% de los senadores queda correctamente clasificado en el modelo sin interacción.

Introducción de variables cualitativas

Supongamos que deseamos estudiar si la probabilidad de voto en función de la contribución y que tenemos dos partidos distintos, Republicano y Demócrata.



El ajuste conjunto no representa a ninguno de los dos y los ajustes separados pierden precisión.

Introducción de variables cualitativas

Trataremos de incluir ambas variables (Contribución y Partido) en el mismo modelo.
Definimos una nueva variable D que toma el valor 1 si el senador pertenece al partido republicano y el 0 si pertenece al demócrata, y ajustamos el modelo

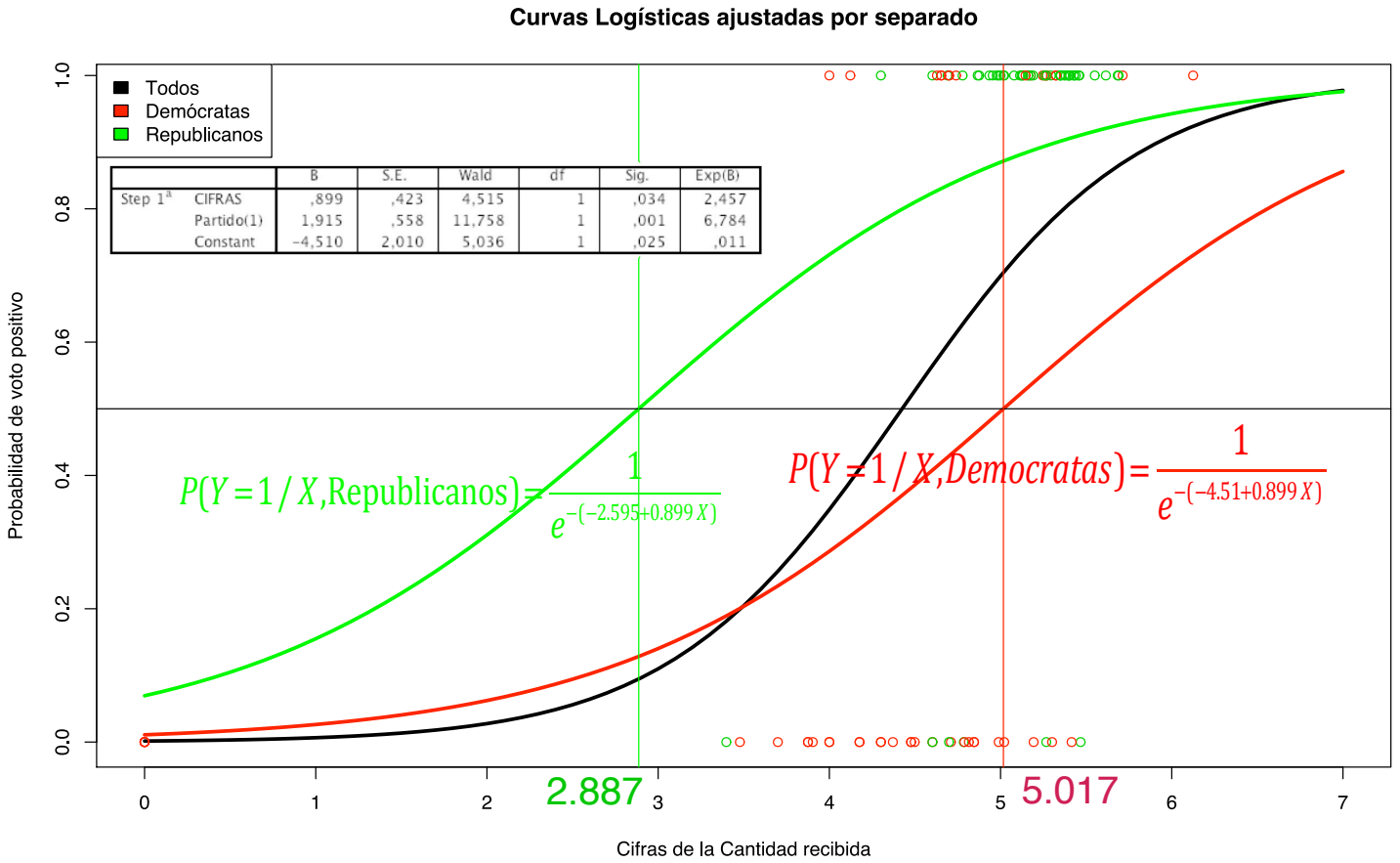
$$P(Y = 1/(X, D)) = \frac{1}{e^{-(\beta_0+\beta_1\ X+\delta\ D)}}$$

donde X es el número de cifras.

- Para un senador "Demócrata" $P(Y = 1/X) = \frac{1}{e^{-(\beta_0+\beta_1\ X)}}$
- Para un senador "Republicano" $P(Y = 1/X) = \frac{1}{e^{-((\beta_0+\delta\ )+\beta_1\ X)}}$

Entonces δ mide la "diferencia" entre un senador republicano y uno demócrata. Suponemos que la pendiente es la misma en ambos grupos. Los efectos de la contribución y del partido son aditivos.

Para los datos de los que disponemos



		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CIFRAS	,899	,423	4,515	1	,034	2,457
	Partido(1)	1,915	,558	11,758	1	,001	6,784
	Constant	-4,510	2,010	5,036	1	,025	,011

Introducción de variables cualitativas

Para los datos de los que disponemos

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	37,045	2	,000
	Bloque	37,045	2	,000
	Modelo	37,045	2	,000

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	93,816[a]	,312	,426

a. La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación[a]

			Pronosticado		
			Voto		
	Observado		No	Si	Porcentaje correcto
Paso 1	Voto	No	27	10	73,0
		Si	9	53	85,5
	Porcentaje global				80,8

a. El valor de corte es ,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Partido(1)	-1,915	,558	11,758	1	,001	,147	,049	,440
	Cifras	,899	,423	4,515	1	,034	2,457	1,072	5,630
	Constante	-2,595	2,171	1,429	1	,232	,075		

a. Variable(s) introducida(s) en el paso 1: Partido, Cifras.

Sólo el partido

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Partido(1)	2,459	,524	22,003	1	,000	11,693
	Constant	-,490	,291	2,823	1	,093	,613

a. Variable(s) entered on step 1: Partido.

Un valor interesante es el 11.693, que se conoce como *odds ratio*.
Si llamamos p_R a la probabilidad de voto positivo en los republicanos, el *odds* o *ventaja* del voto positivo frente al negativo, para los republicanos es, $odds_R = \frac{p_R}{1-p_R}$. Para los Demócratas sería $odds_D = \frac{p_D}{1-p_D}$.
El cociente de estas dos cantidades se conoce como *odds ratio*.

$$OR_{R/D} = \frac{odds_R}{odds_D}$$

y puede considerarse como una medida de la asociación o correlación entre las variables. En este caso, la ventaja del voto positivo frente al negativo es 11.693 veces mayor en los republicanos.

Interacción

En muchos casos los efectos de las dos variables no son aditivos sino que también hay un efecto combinado de ambas que se denomina *interacción*.
En este caso las diferencias que se producen en el voto a los distintos partidos depende de la cantidad monetaria, y las diferencias que producen las cantidades monetarias dependen del partido.
En la práctica se traduce en qué las curvas no tienen la misma pendiente.
El modelos que ajustamos ahora es

$$P(Y = 1/(X, D)) = \frac{1}{e^{-(\beta_0 + \beta_1 X + \delta D + \gamma DX)}}$$

Donde la nueva variable que hemos incluido es producto de lass dois que teníamos en el modelo anterior.

Interacción

Incluyendo ambas variables (Contribución y Partido) y la interacción en el mismo modelo, tenemos

- Para un senador "Demócrata"

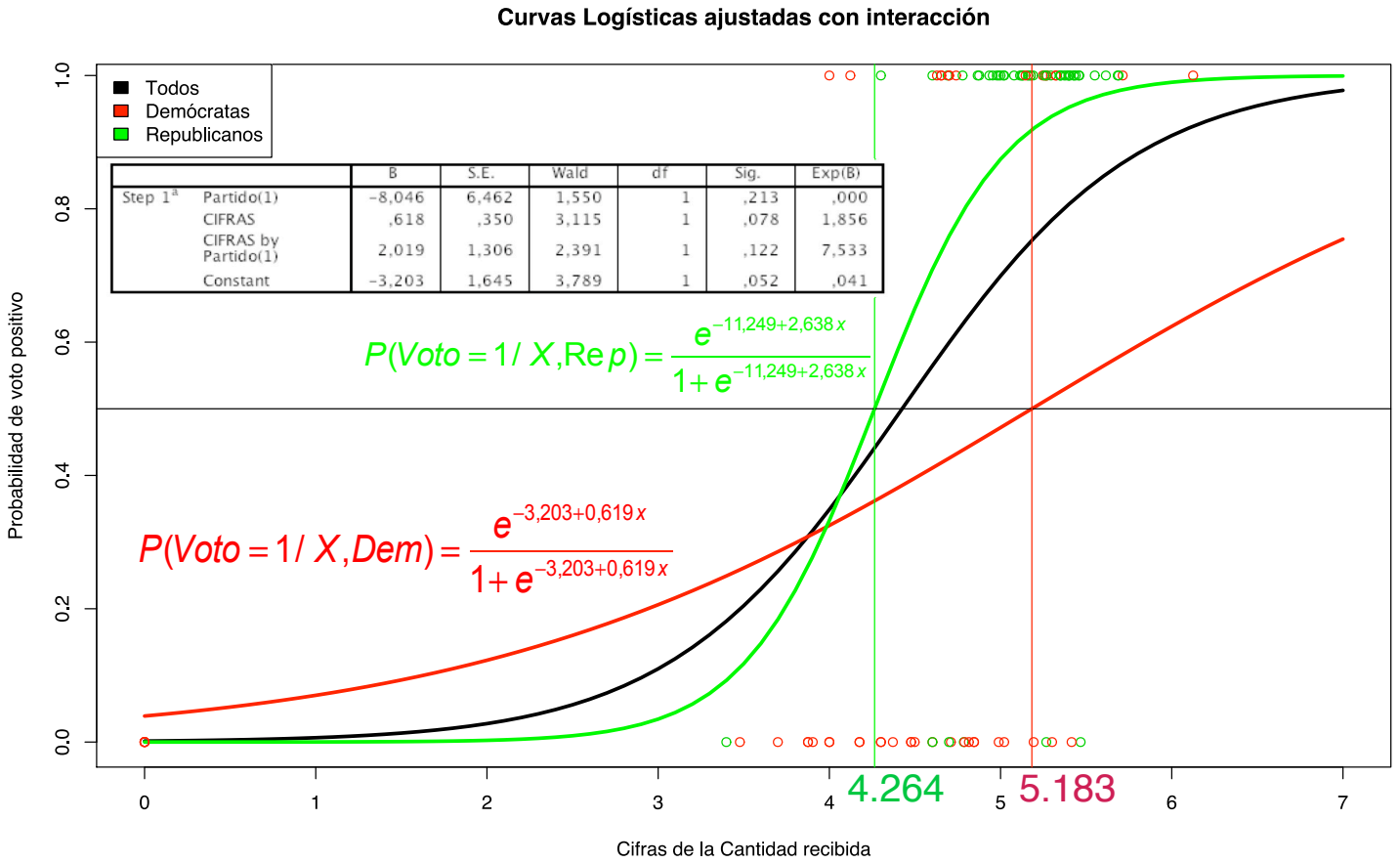
$$P(Y = 1/(X, Democratas)) = \frac{1}{e^{-(\beta_0+\beta_1\ X)}}$$

- Para un senador "Republicano"

$$P(Y = 1/(X, \mathrm{Republicanos})) = \frac{1}{e^{-((\beta_0+\delta)+(\beta_1+\gamma\ )\ X)}}$$

Entonces δ ya no mide la "diferencia" entre un senador republicano y uno demócrata ya que esta depende de la cantidad de dinero recibida. La pendiente no es la misma en ambos grupos. Los efectos de la contribución y del partido son multiplicativos y γ es el parámetro que mide la interacción.

Para los datos de los que disponemos



		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Partido(1)	-8,046	6,462	1,550	1	,213	,000
	CIFRAS	,618	,350	3,115	1	,078	1,856
	CIFRAS by Partido(1)	2,019	1,306	2,391	1	,122	7,533
	Constant	-3,203	1,645	3,789	1	,052	,041

Separación

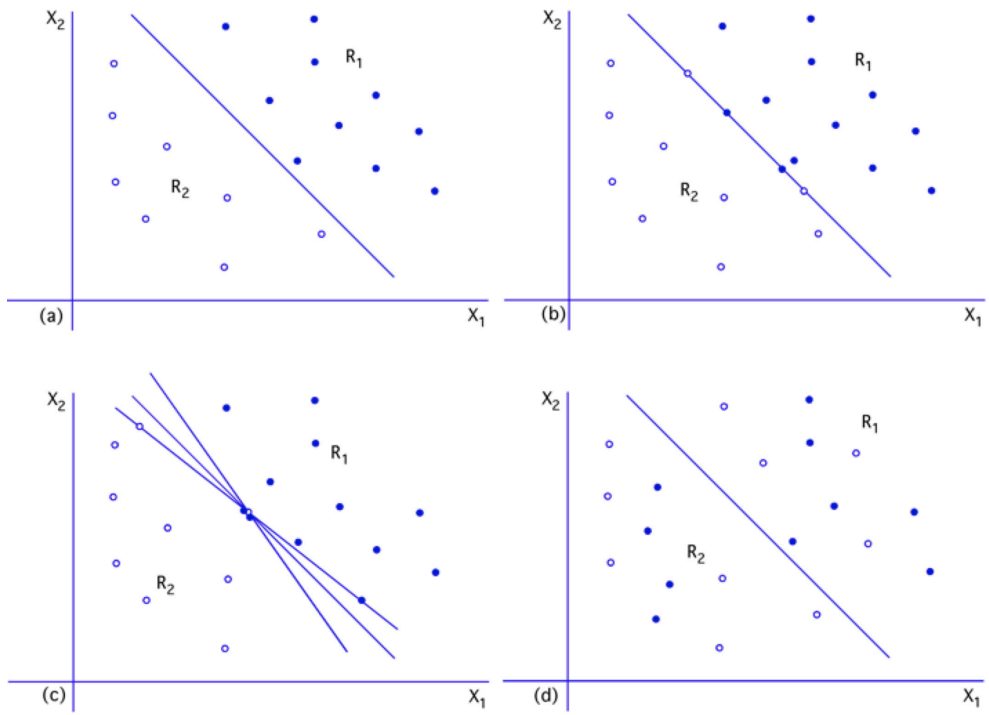


Figure: Posibles configuraciones de puntos para el caso de dos variables y dos grupos (puntos huecos y rellenos). Las regiones R1 y R2 definen las reglas de clasificación. (a) Separación completa. (b) Separación cuasicompleta con hiperplano de separación único. (c) Separación cuasicompleta con hiperplano de separación no único, hay tres puntos sobre la intersección de las líneas. (d) Solapamiento.

Separación

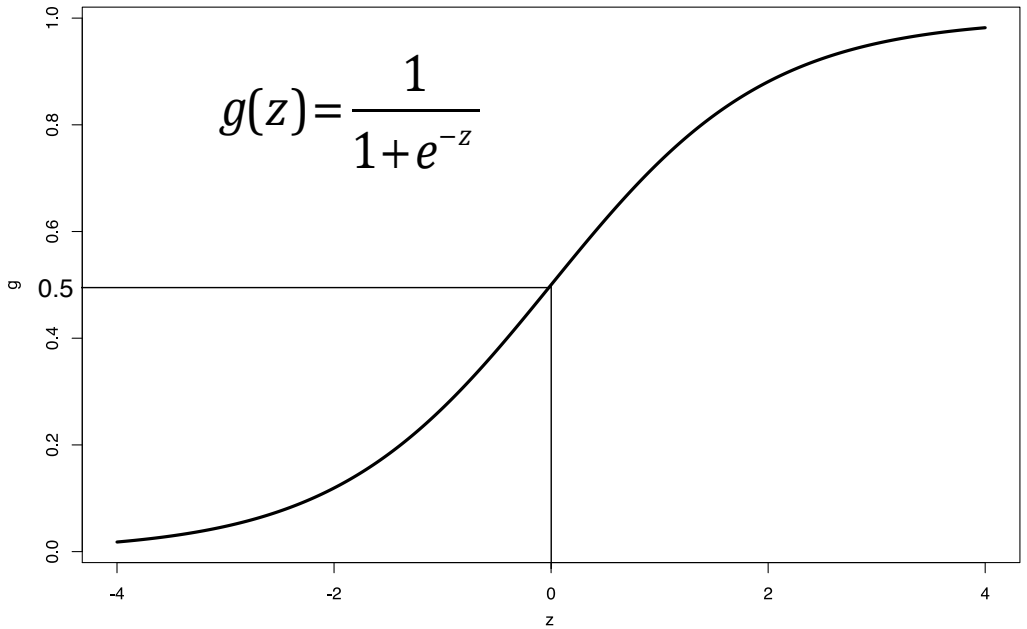
- Cuando hay una separación completa el método de máxima verosimilitud no converge y los estimadores tienden a infinito.
- Cuando hay una separación cuasi completa los estimadores son inestables.
- El problema puede solucionarse añadiendo una penalización sobre la magnitud de los estimadores en la función de verosimilitud de forma que no puedan crecer indefinidamente. (Firth, Ridge o LASSO)

Modelo de Regresión Logística en Machine Learning

Hipótesis: En machine learling la hipótesis es la función a justar, es decir, la función logística

Queremos $0 \leq h_\beta(x) \leq 1$

Ahora : $h_\beta(x) = g(\beta' x)$ con $g(z) = \frac{1}{1+e^{-z}}$ (función logística o sigmoide)



Nuestra hipótesis se interpreta ahora como la probabilidad de respuesta positiva $h_\beta(x) = P(y = 1/x; \beta)$.

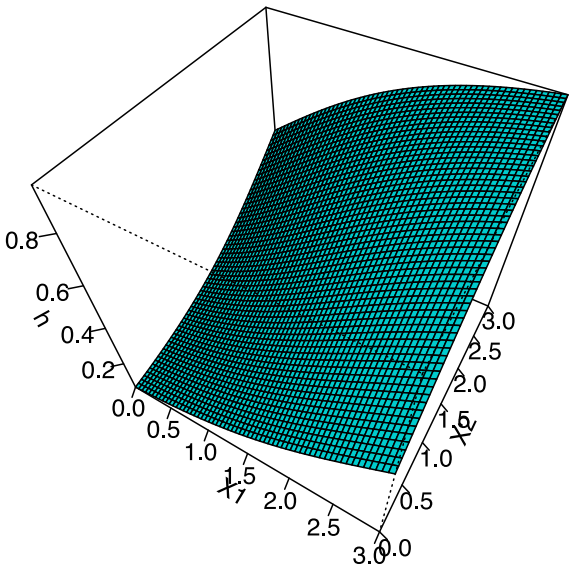
Modelo de Regresión Logística - Hipótesis:

$$h_\beta(x) = \frac{1}{1 + e^{-\beta' x}}$$

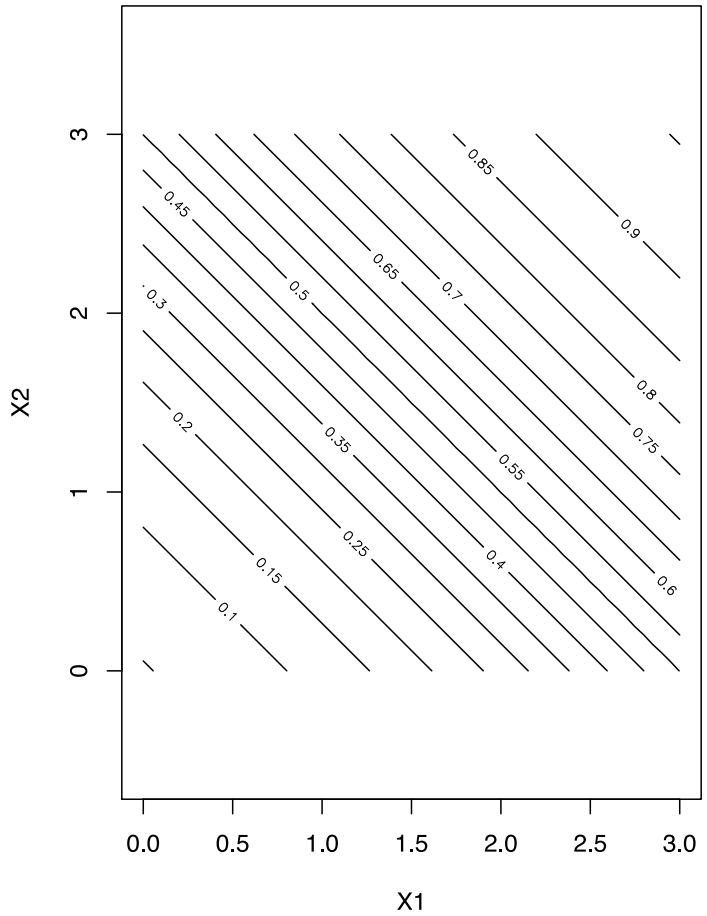
Predecimos respuesta positiva cuando $h_\beta(x) \geq 0.5$, que es equivalente a $\beta' x > 0$.
Si tenemos una sola variable, $h_\beta(x) = g(\beta_0 + \beta_1 x)$.
Despejando de $\beta_0 + \beta_1 x > 0$ obtenemos que el valor de x a partir del cual se predice presencia es: $x > \frac{-\beta_0}{\beta_1}$

Modelo de regresión Logística - Hipótesis con dos variables:

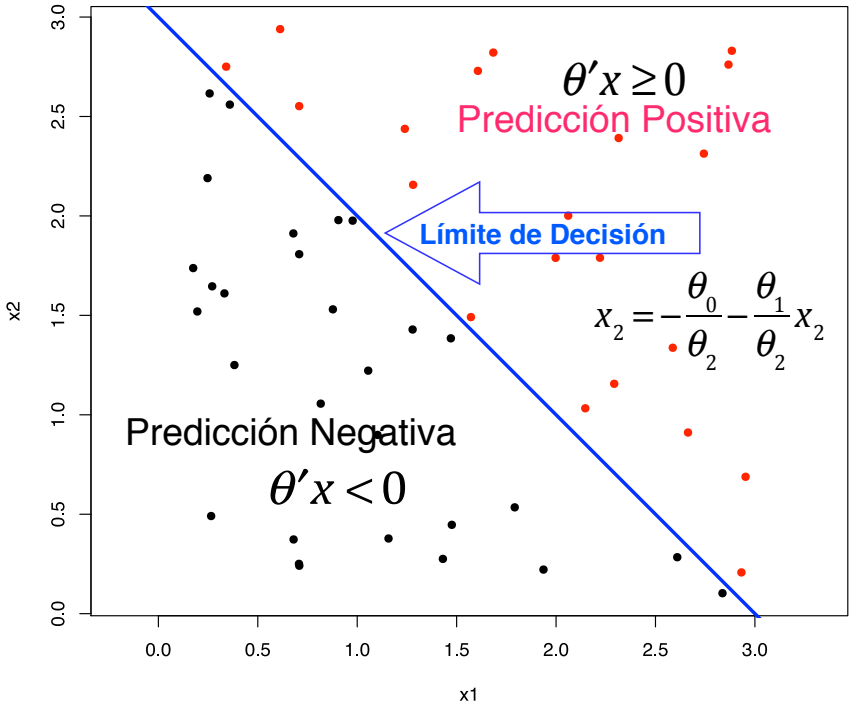
Función Logistica 3D



Curvas de Nivel de la Función Logistica

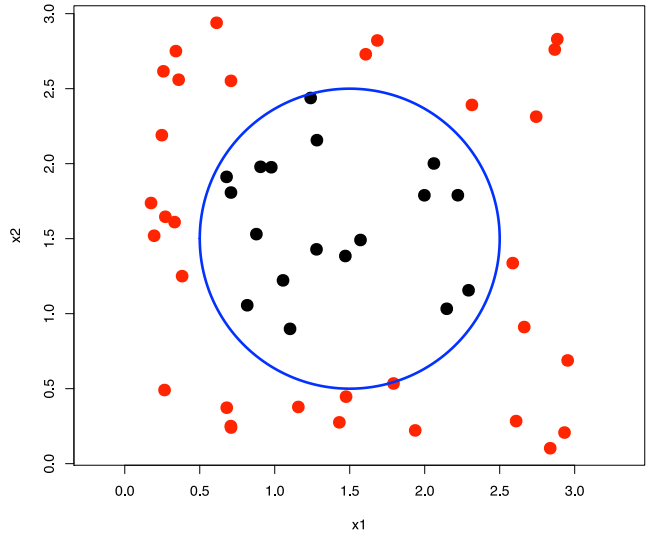


Modelo de regresión Logística - Límites de decisión:



Con dos variables el límite de decisión es una línea recta, con tres un plano y con más es un hiperplano.

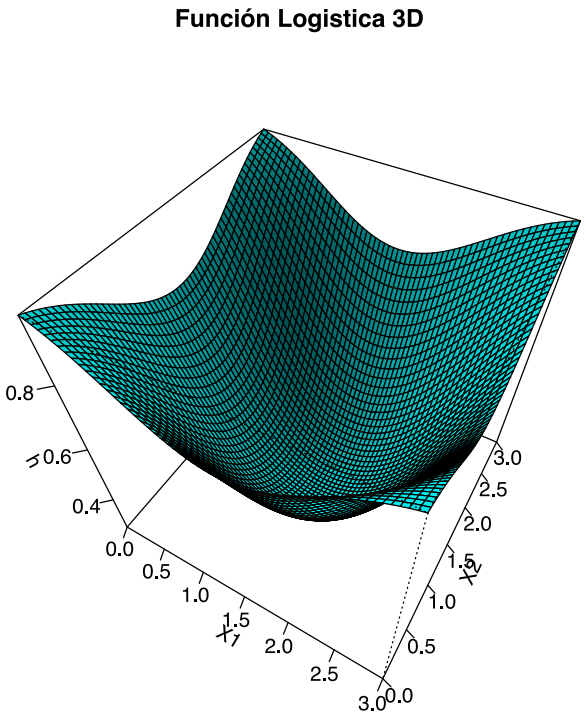
Límites de decisión no lineales: Podemos encontrar que los límites de decisión no son líneas rectas

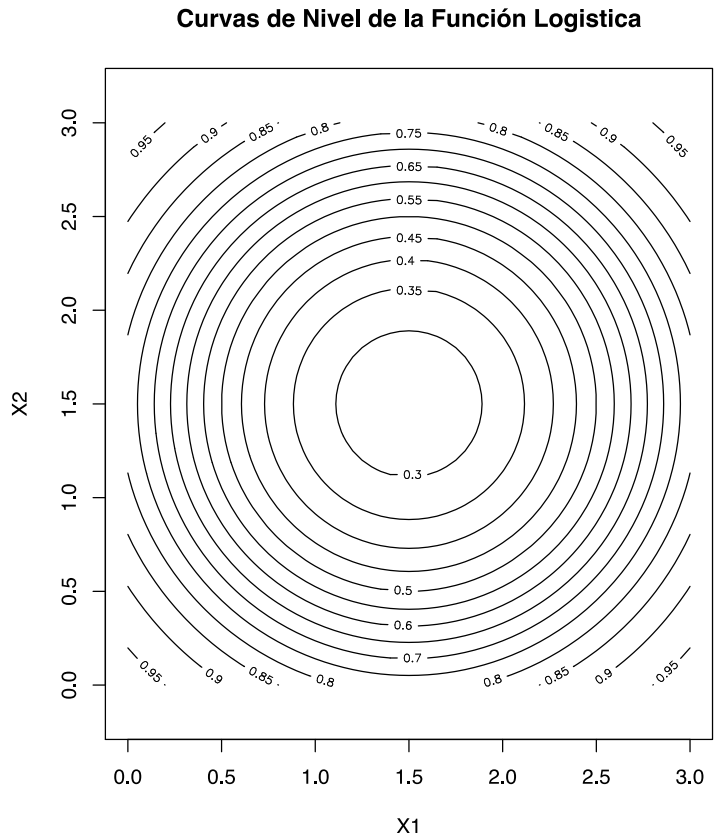


El problema se soluciona añadiendo términos cuadráticos y productos a la hipótesis

$$h_\beta(x) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2)$$

Límites de decisión no lineales:





$$h_\beta(x) = g(3.5 - 3x_1 - 3x_2 + x_1^2 + x_2^2)$$

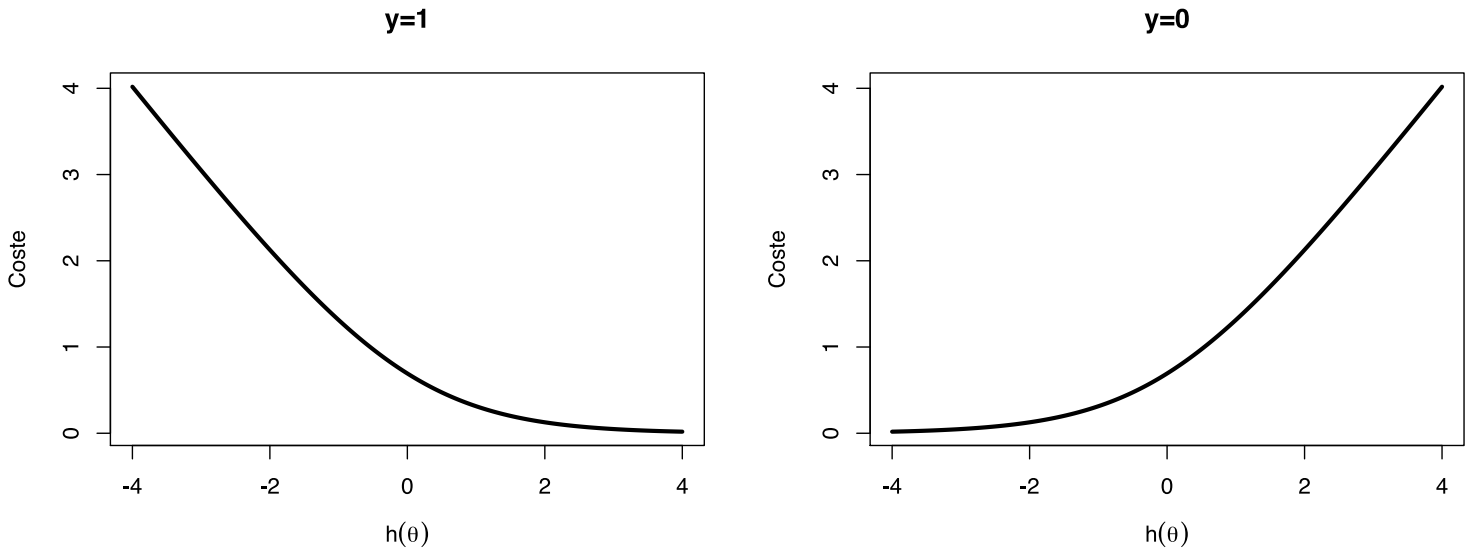
Función de coste:

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n} \text{Coste}(h_\beta(x_i), y_i)$$

- $\text{Coste}(h_\beta(x), y) = -\log(h_\beta(x))$ Si "$y = 1$"
- $\text{Coste}(h_\beta(x), y) = -\log(1 - h_\beta(x))$ Si "$y = 0$"

Obsérvese que la función de coste es la misma que la que utilizábamos en el método de máxima verosimilitud.

Función de coste:



Entonces

- $\text{Coste}(h_\beta(x), y) = 0$ Si $h_\beta(x) = y$
- $\text{Coste}(h_\beta(x), y) \to \infty$ Si $y = 0$ y $h_\beta(x) \to 1$
- $\text{Coste}(h_\beta(x), y) \to \infty$ Si $y = 1$ y $h_\beta(x) \to 0$

Función de coste:
Simplificando:

$$\text{Coste}(h_\beta(x), y) = -y \log(h_\beta(x)) - (1 - y) \log(1 - h_\beta(x))$$

Sumando para todos los ejemplos

$$J(\beta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(h_\beta(x_i)) - (1 - y_i) \log(1 - h_\beta(x_i))]$$

En forma matricial

$$h = g(X\beta)$$

$$J(\beta) = -\frac{1}{n}(y' \log(h) + (1 - y') \log(1 - h))$$

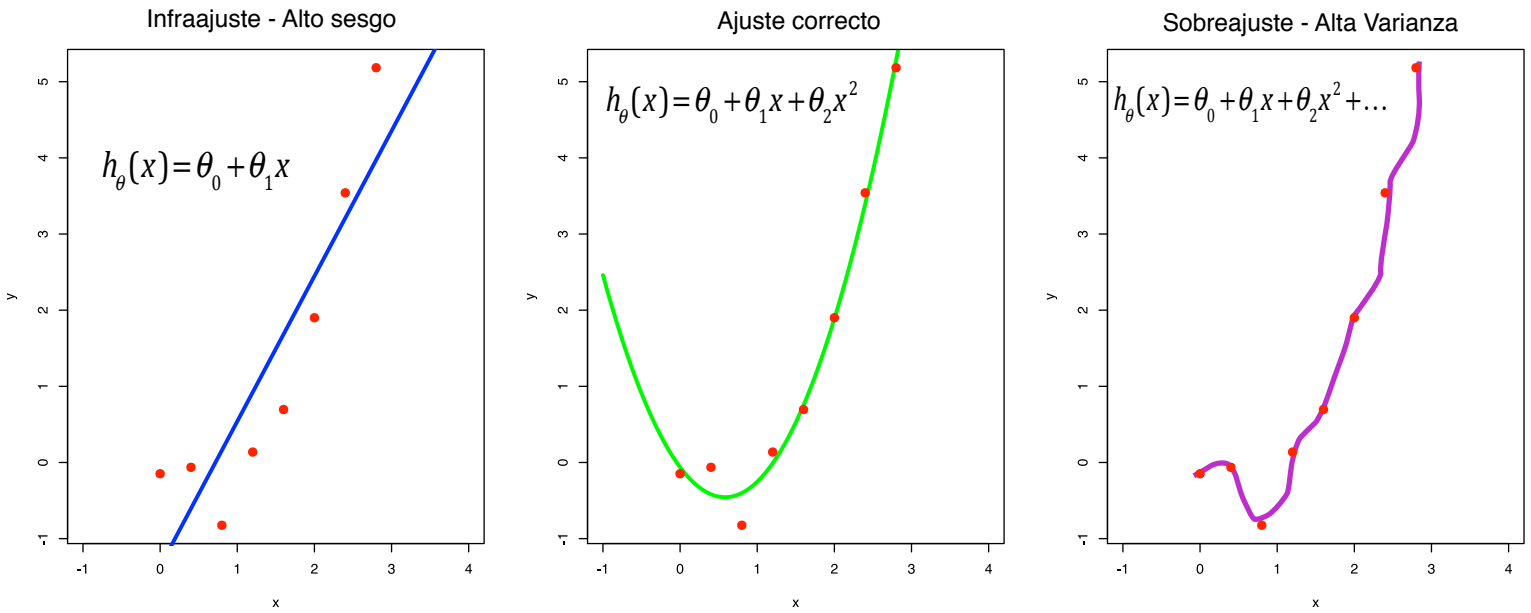
Gradiente: Repetir hasta la convergencia :

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\beta_0, \beta_1, \ldots, \beta_p)$$

que resulta ser similar al de la regresión múltiple,

$$\beta_j := \beta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} (h_\beta(x_i) - y_i) x_{j\,i}$$

Sobreajuste - OverFitting

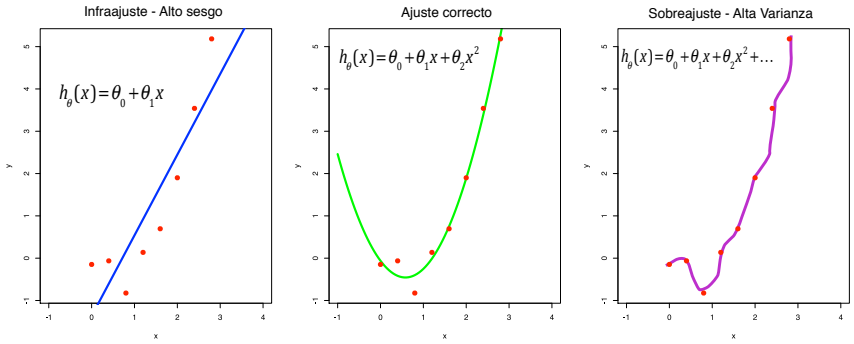


Si tenemos demasiados atributos, la hipótesis aprendida puede ajustarse bien al conjunto usado para su entrenamiento ($J(\beta) \approx 0$), pero puede que falle cuando se aplique a nuevos ejemplos

Tratamiento del Sobreajuste

- Reducir el número de atributos.
  - Seleccionar manualmente los atributos a eliminar
  - Algoritmos automáticos de selección de modelos.
- Regularización.
  - Mantener todos los atributos, pero reducir la magnitud de los parámetros θ_j
  - Funciona bien cuando tenemos muchos atributos, cada uno de los cuales contribuye un poco a predecir y..

Tratamiento del Sobreajuste - Regularización



Una posible solución del problema de sobreajuste sería penalizar los parámetros de los términos de grado mayor que 2 ($\beta_3, \beta_4, \ldots$) para que sean muy pequeños. Podríamos tomar la función de coste:

$$J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(h_\beta(x_i) - y_i)^2 + 1000\beta_3 + 1000\beta_4 + \ldots$$

La única forma de hacer la función de coste pequeña es que los parámetros sean pequeños.

Sobreajuste - OverFitting Tener valores más pequeños para los parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$, significa :

- Hipótesis más simples.
- El modelo está menos afectado por el sobreajuste..

Penalizaremos la función de coste para no permitir valores muy grandes en los parámetros.

Sobreajuste - OverFitting

De forma general añadimos la penalización a la función de coste, Por ejemplo, en regresión sería:

$$J(\beta) = \frac{1}{2n}\left[\sum_{i=1}^{n}(h_\beta(x_i) - y_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right]$$

Donde λ es el parámetro de regularización y determina cuanto añaden los parámetros al coste . Normalmente β_0 se deja sin penalizar. Si el valor de λ es demasiado grande, puede ocurrir que los reduzcamos demasiado y tengamos "infraajuste". Si es demasiado pequeño, puede que no resolvamos el problema. En otro contexto esto se conoce como "Regresión Ridge". Otra penalización popular es la conocida como LASSO que consiste en :

$$\min_{\beta} J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(h_\beta(x_i) - y_i)^2 \qquad \textit{sujeto} \sum_{j=1}^{p} \|\beta_j\| \leq t$$

Regularización - Regresión Logística

Coste:

$$J(\beta) = -\left[\frac{1}{n}\sum_{i=1}^{n} y_i \log(h_\beta(x_i)) + (1-y_i)\log(1-h_\beta(x_i))\right]$$

$$+\frac{\lambda}{2n}\sum_{j=1}^{p}\beta_j^2$$

Repetir hasta la convergencia :

$$\beta_0 := \beta_0 - \alpha\frac{1}{n}\sum_{i=1}^{n}(h_\beta(x_i) - y_i)x_{0i}$$

$$\beta_j := \beta_j - \alpha\frac{1}{n}\left[\sum_{i=1}^{n}(h_\beta(x_i) - y_i)x_{ji} + \lambda\beta_j\right] (j = 1, 2, \ldots, p)$$

Evaluación de la hipótesis

Supongamos que hemos entrenado nuestro modelo y no hemos obtenido una bondad de ajuste adecuada. ¿Qué podemos hacer?

- Obtener una muestra mayor.
- Intentar con conjuntos más pequeños de atributos.
- Añadir atributos adicionales.
- Añadir atributos polinómicos.
- Aumentar o disminuir λ.

Evaluación del error Una alta bondad de ajuste no siempre significa que el modelo que hemos ajustado sea adecuado para los datos. Si tenemos problemas de sobreajuste, puede que el modelo no se generalice bien a otro conjunto de datos. El modelo debería probarse en un conjunto de datos distinto al que se utilizó para ajustarlo.
Dividiremos nuestro conjunto de datos en dos partes:

- Conjunto de Entrenamiento: Un conjunto de datos para entrenar (ajustar) el modelo. Normalmente el de 70% los datos.
- Conjunto de prueba: Un conjunto de datos para probar el modelo. Normalmente el 30% de los datos.

El conjunto de prueba nos sirve para estimar el error de generalización.

Selección del modelo

Supongamos, por ejemplo, que queremos seleccionar el grado del polinomio que queremos ajustar. Para estimar el error de generalización necesitaríamos un nuevo conjunto de datos. Dividiremos ahora nuestro conjunto de datos en tres partes:

- Conjunto de Entrenamiento: Un conjunto de datos para entrenar (ajustar) el modelo. Normalmente el de 60% los datos. $J_{entr}(\beta)$
- Conjunto de Validación: Un conjunto de datos para explorar parámetros adicionales. Normalmente el de 20% los datos. $J_{valid}(\beta)$
- Conjunto de prueba: Un conjunto de datos para probar el modelo y estimar el error de generalización. Normalmente el 20% de los datos. $J_{gen}(\theta)$

Error para clases no balanceadas.

Supongamos, por ejemplo, que en un problema de clasificación de cánceres tenemos un 1% de error de clasificación en el conjunto de prueba (99% de diagnósticos correctos). Parece un buen resultado pero resulta que solamente el 0.5% de los pacientes tienen cáncer.

Si usamos la regla, "Clasificar todos como sanos", ignorando x. El porcentaje de diagnósticos correctos es mejor.

Necesitamos medidas nuevas de la bondad del ajuste.

Construimos la tabla de confusión

Obs./ Pred.	**1**	**0**	**Total**
1	Verdaderos Positivos	Falsos Negativos	Enfermos
0	Falsos Positivos	Verdaderos Negativos	Sanos
Total	Positivos	Negativos	Total

$$\text{Precision} = \frac{\text{Verdaderos} \quad \text{positivos}}{\text{Positivos}} = \frac{\text{Verdaderos} \quad \text{positivos}}{\text{Verd.} \quad \text{Pos.} + \text{Falsos} \quad \text{Pos.}}$$

(Valor Predictivo Positivo)

$$\text{Recuerdo} = \frac{\text{Verdaderos} \quad \text{positivos}}{\text{Enfermos}} = \frac{\text{Verdaderos} \quad \text{positivos}}{\text{Verd.} \quad \text{Pos.} + \text{Falsos} \quad \text{Neg.}}$$

(Sensibilidad)

Error para clases no balanceadas.

En algunos casos es conveniente modificar la regla de decisión cambiando el 0.5 por otro valor.

Pueden evaluarse distintos puntos de corte con una curva ROC (Sensibilidad frente a 1-Especificidad). En el contexto de Machine Learning se representan las dos cantidades anteriores.

Un compromiso entre precisión y recuerdo se puede obtener en el punto con mayor puntuación F_1 con

$$F_1 = 2\frac{PR}{P+R}$$