Regresión Logística

Regresión Logística
Accidente Challenger 28/01/1986

Regresión Logística

Accidente Challenger 28/01/1986

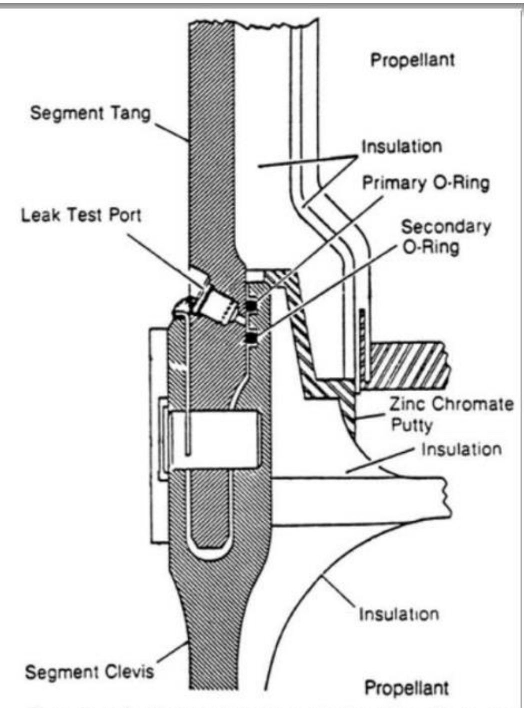
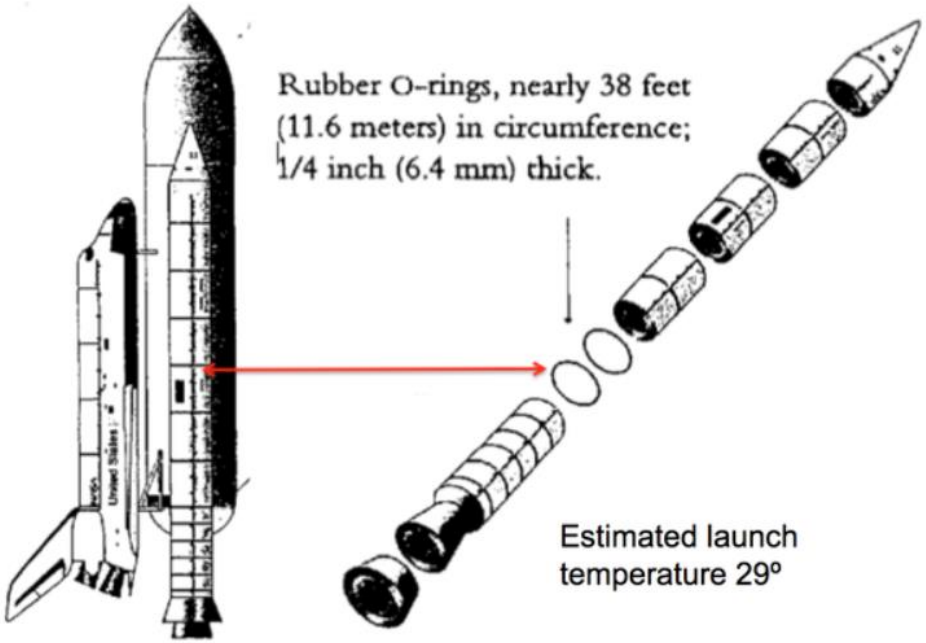


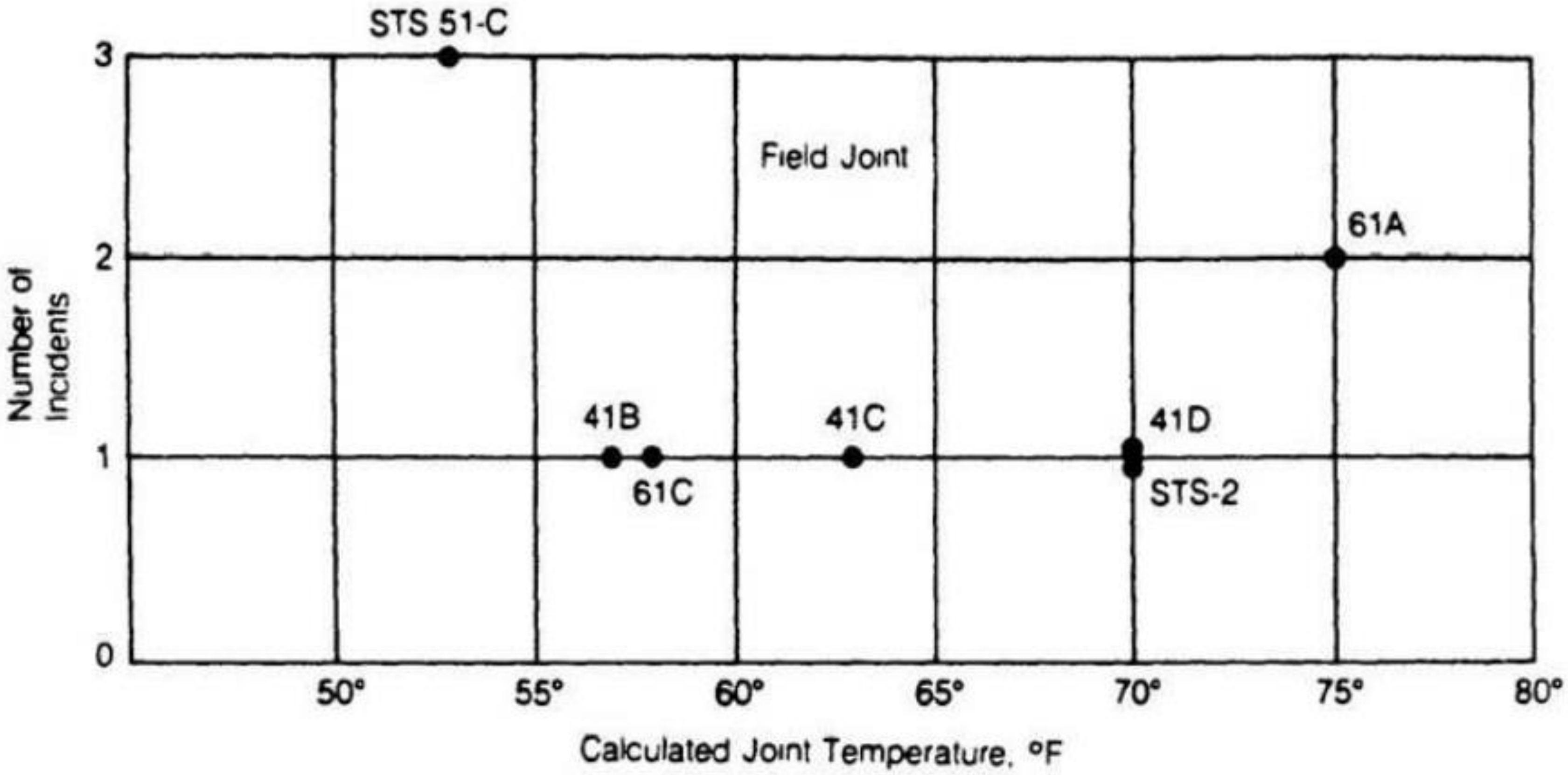
Figure 3. Solid Rocket Motor Cross Section: Tang, Clevis, and O-Rings.



https://es.wikipedia.org/wiki/Accidente_del_transbordador_espacial_Challenger

Regresión Logística

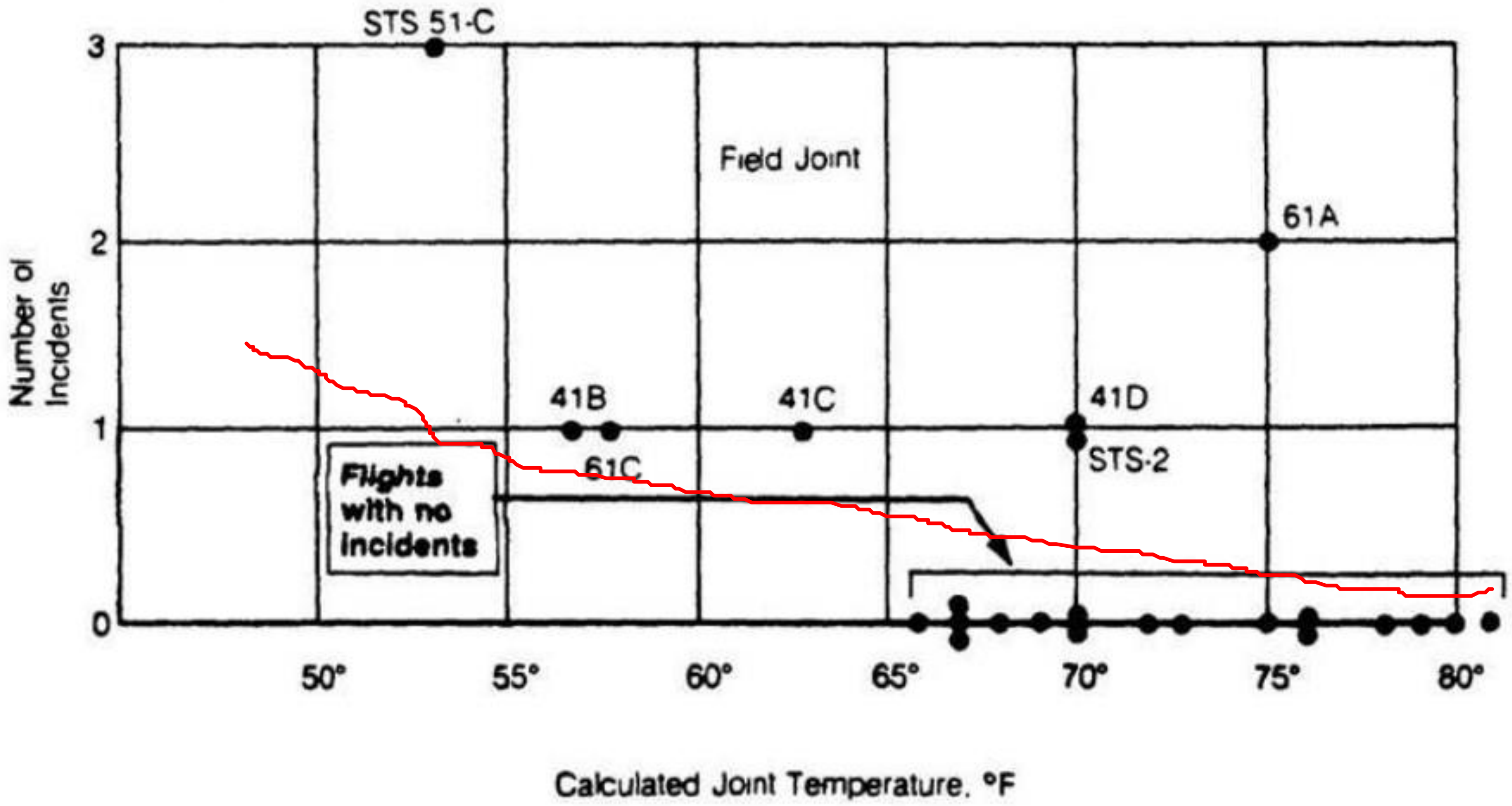
Accidente Challenger 28/01/1986



https://history.nasa.gov/rogersrep/v1ch6.htm#6.3

Regresión Logística

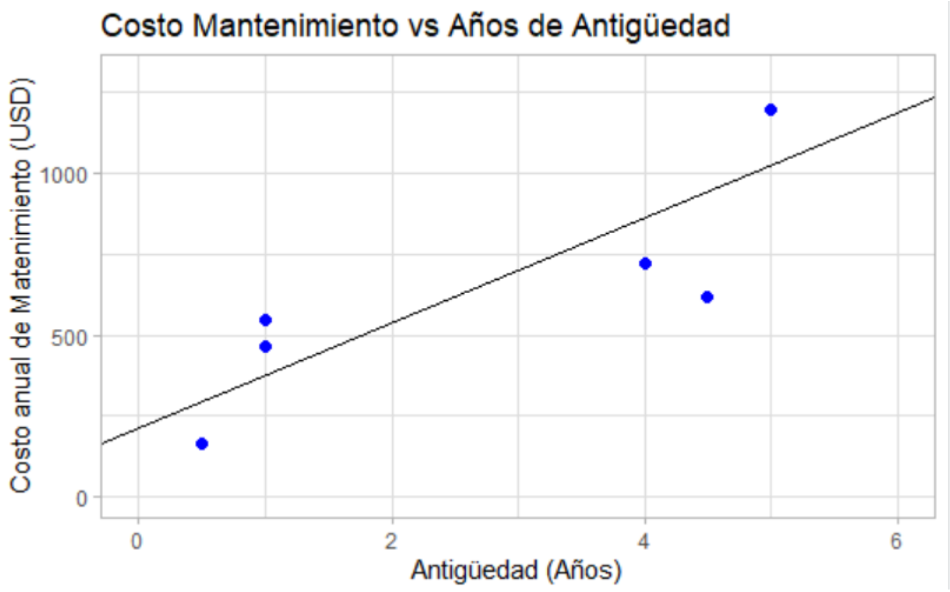
Accidente Challenger 28/01/1986

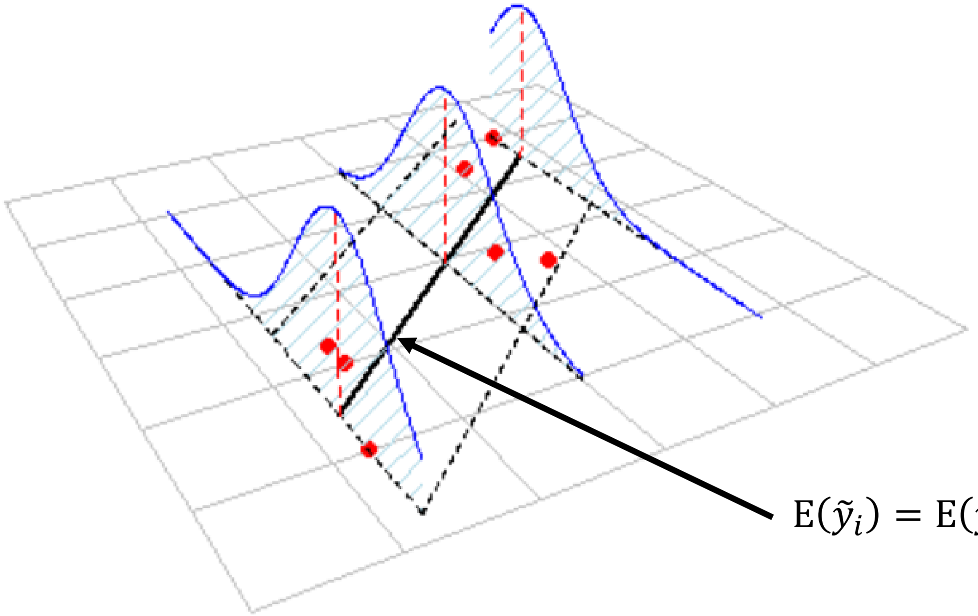


https://history.nasa.gov/rogersrep/v1ch6.htm#6.3

Regresión Logística

Interpretación del modelo de regresión lineal



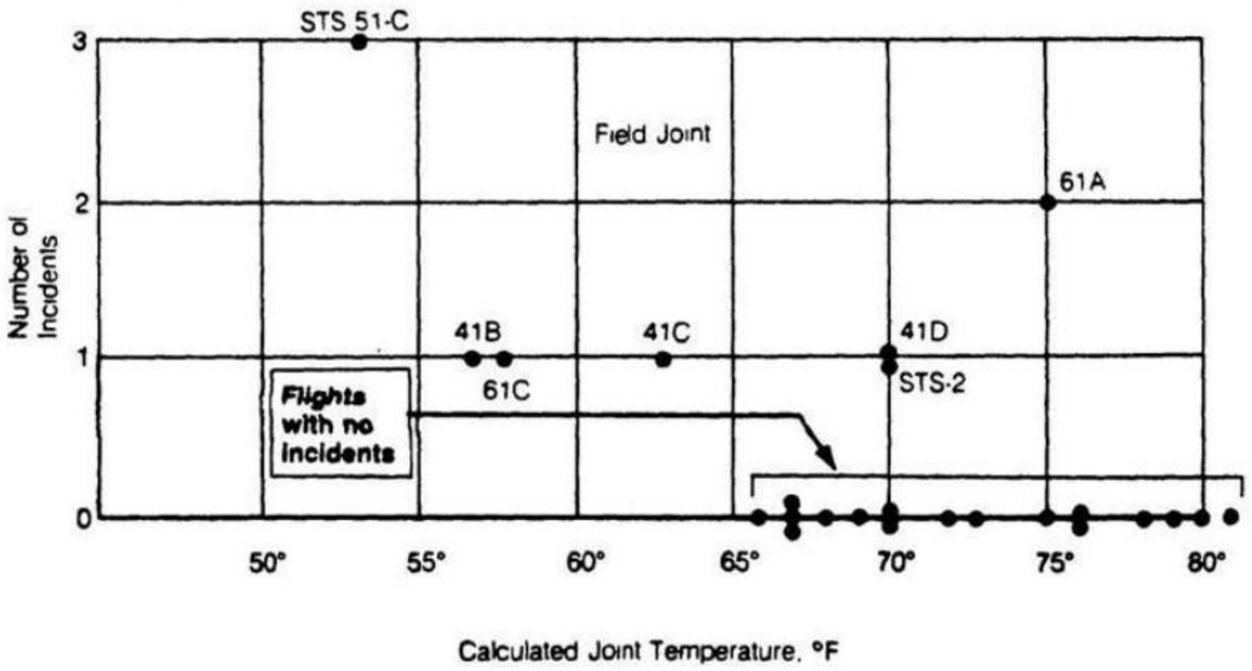


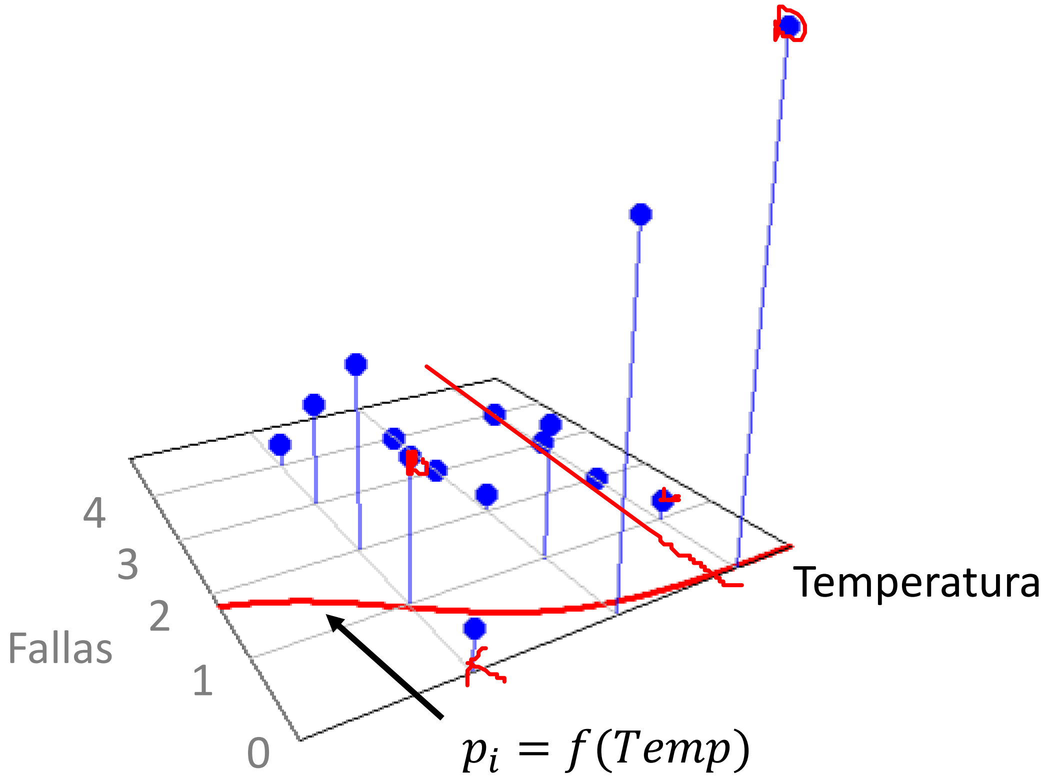
$$\mathrm{E}(\tilde{y}_i) = \mathrm{E}(\tilde{y}|x_i) = \beta_o + \beta_1 x_i = \mu_i$$

$$\tilde{y}_i \sim \mathrm{N}(\mu_i = \beta_o + \beta_1 x_i;\ \sigma^2)$$

Regresión Logística

Interpretación del modelo de regresión logística



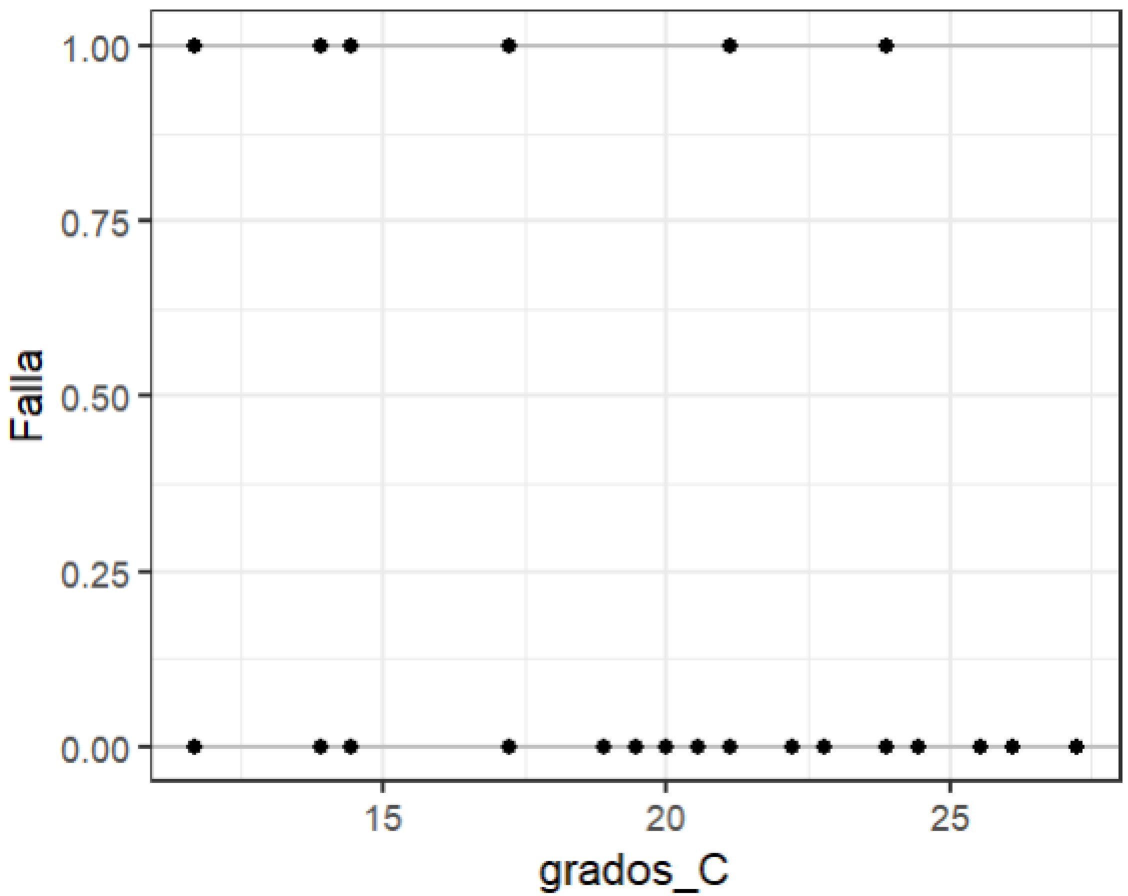


$$\tilde{y}_i \sim \mathrm{Bi}(n = 4; p_i = f(Temp))$$

Regresión Logística

Datos Challenger

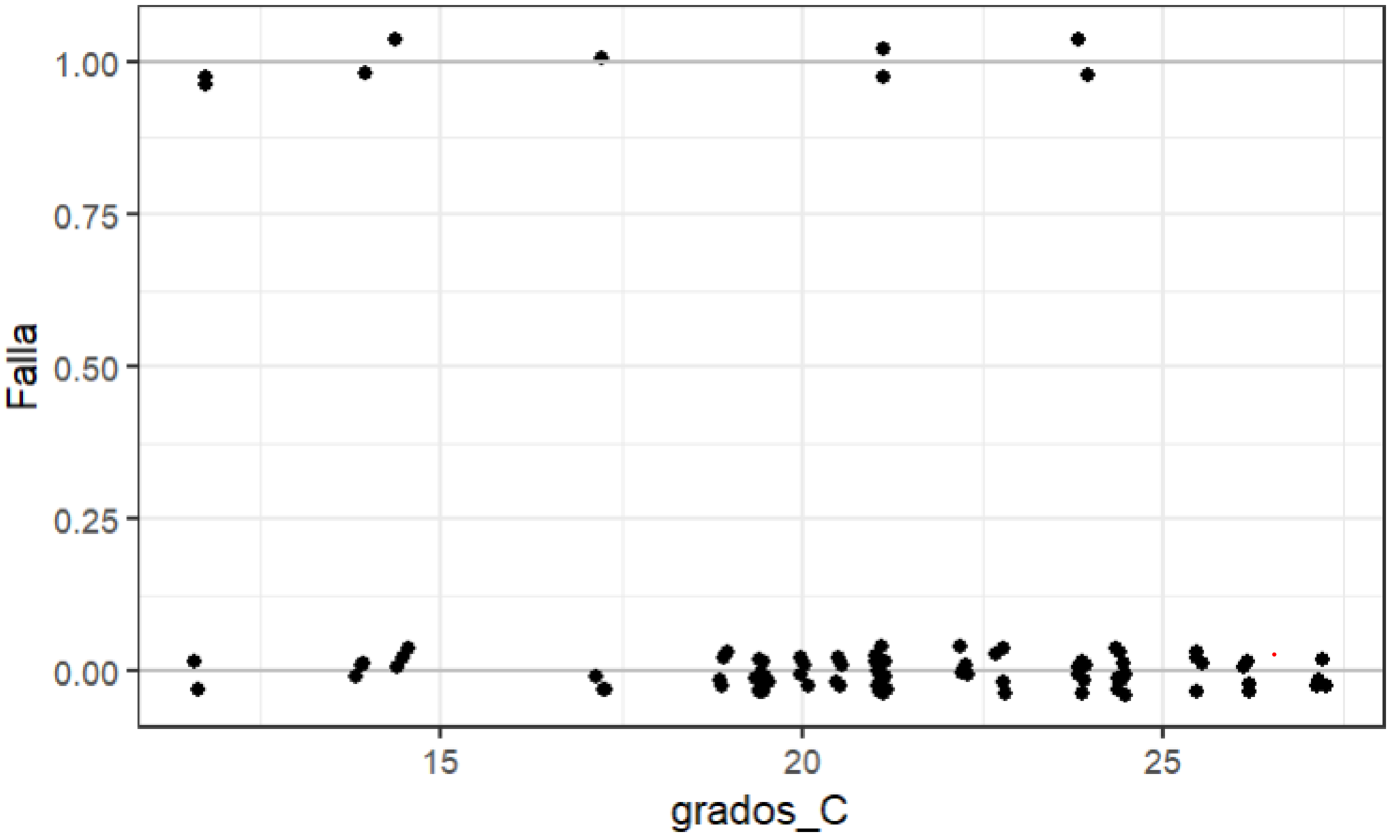
	F	grados_C	Falla
1	53	11.66667	1
2	53	11.66667	1
3	53	11.66667	0
4	53	11.66667	0
5	57	13.88889	1
6	57	13.88889	0
7	57	13.88889	0
8	57	13.88889	0
9	58	14.44444	1
10	58	14.44444	0
11	58	14.44444	0



Regresión Logística

Datos Challenger

	F	grados_C	Falla
1	53	11.66667	1
2	53	11.66667	1
3	53	11.66667	0
4	53	11.66667	0
5	57	13.88889	1
6	57	13.88889	0
7	57	13.88889	0
8	57	13.88889	0
9	58	14.44444	1
10	58	14.44444	0
11	58	14.44444	0

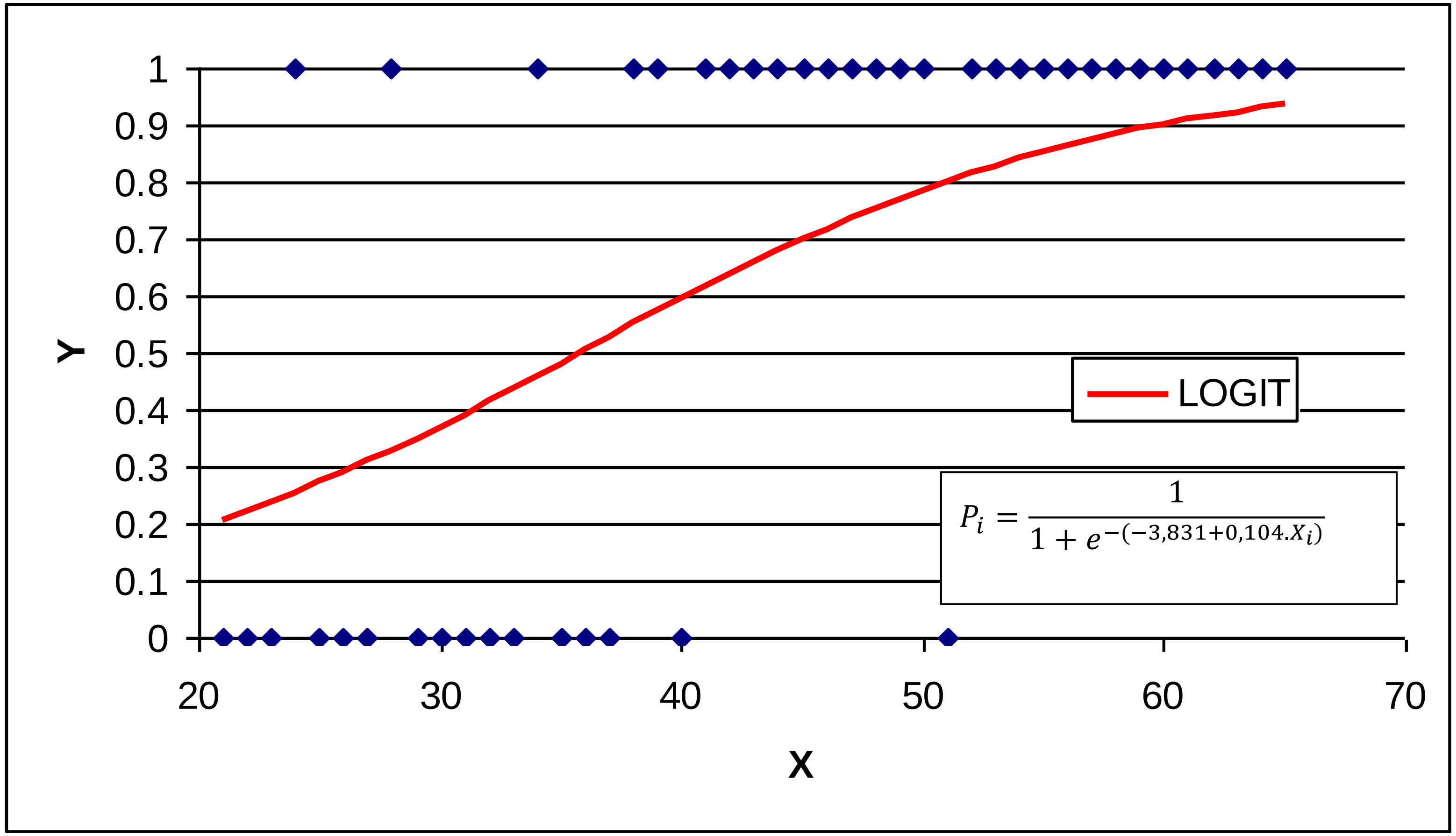


Regresión Logística

Modelo de Regresión Logística

$$\pi_i = \frac{e^{Z_i}}{1 + e^{Z_i}} = \frac{1}{1 + e^{-Z_i}}$$

$$Z_i = \beta_0 + \beta_1 X_i$$

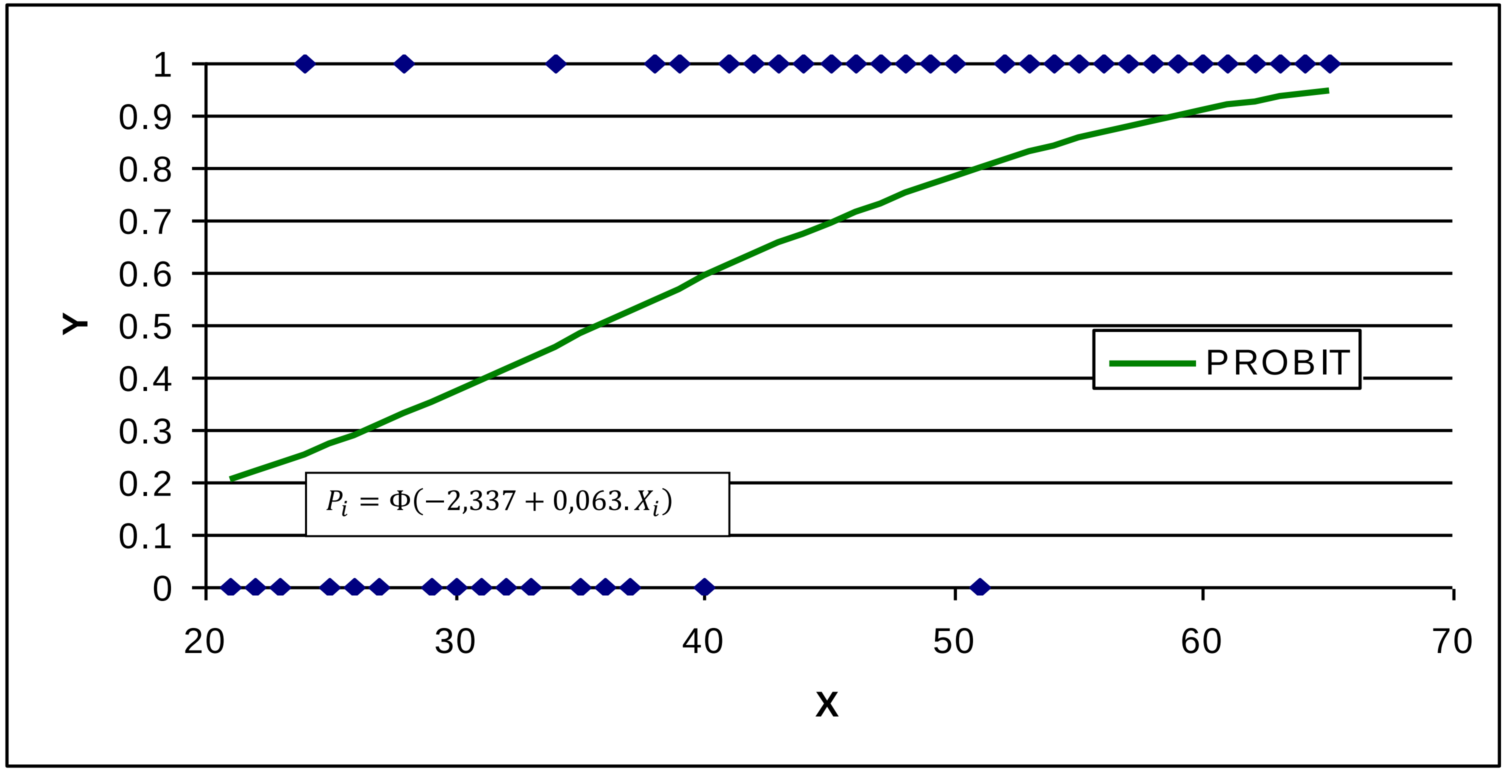


Regresión Logística

Modelo de Regresión Probit

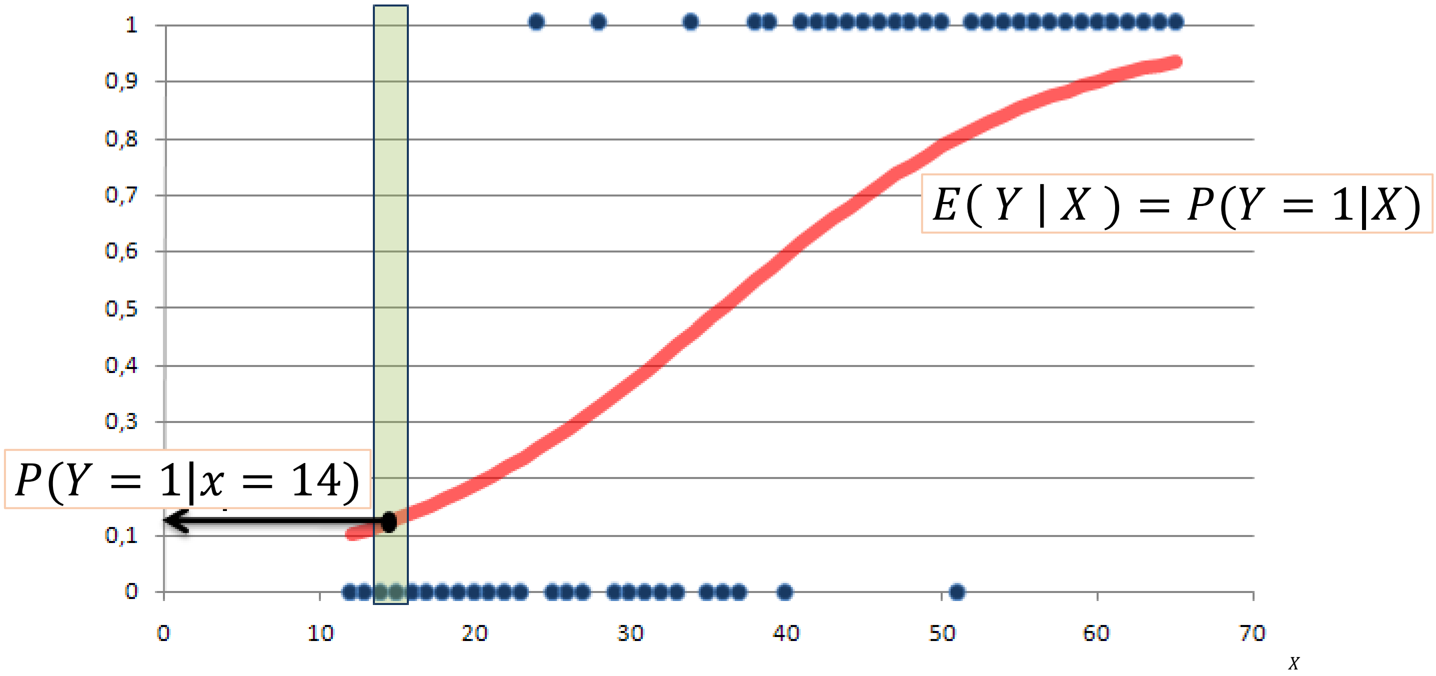
$$\pi_i = \Phi(\beta_0 + \beta_1 X_i)$$

$$= \int_{-\infty}^{\beta_0+\beta_1.X_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$



Regresión Logística

Interpretación de la ecuación de regresión



Estimación de parámetros

Regresión Logística
Estimación de parámetros

$$\mathcal{L} = P(Y_1 \cap Y_2 \cap Y_3 \ldots Y_n) = \prod_{i=1}^{n} {\pi_i}^{Y_i} (1 - \pi_i)^{1-Y_i}$$

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{k_i})}}$$

Distribución Bernoulli

Y	P(Y)
1	P
0	1-p
	1

$$P_{Bern}(Y \mid \pi) = \pi^Y (1 - \pi)^{1-Y}$$

Regresión Logística
Estimación de parámetros

$$\mathcal{L} = P(Y_1 \cap Y_2 \cap Y_3 \ldots Y_n) = \prod_{i=1}^{n} {\pi_i}^{Y_i} (1-\pi_i)^{1-Y_i}$$

$$Ln(\mathcal{L}) = \sum_{i=1}^{n} [Y_i \; Ln(\pi_i) + (1-Y_i) \; Ln(1-\pi_i)] \longrightarrow \text{Máximo}$$

$$\begin{cases} \dfrac{\partial Ln(\mathscr{L})}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i - \pi_i) = 0 \\ \dfrac{\partial Ln(\mathscr{L})}{\partial \beta_1} = \sum_{i=1}^{n} x_{1i}.(Y_i - \pi_i) = 0 \\ \ldots. \\ \dfrac{\partial Ln(\mathscr{L})}{\partial \beta_p} = \sum_{i=1}^{n} x_{ki}.(Y_i - \pi_i) = 0 \end{cases}$$

Regresión Logística

Estimación de parámetros

	F	grados_C	Falla
1	53	11.66667	1
2	53	11.66667	1
3	53	11.66667	0
4	53	11.66667	0
5	57	13.88889	1
6	57	13.88889	0
7	57	13.88889	0
8	57	13.88889	0

Verosimilitud Bernoulli

$$\mathcal{L} = P(Y_1 \cap Y_2 \cap Y_3 \ldots Y_8) = \pi_1\, \pi_2 (1-\pi_3)(1-\pi_4)\, \pi_5 (1-\pi_6)(1-\pi_7)(1-\pi_8)$$

Verosimilitud Binomial

$$\mathcal{L} = P(Y_1 \cap Y_2) = \binom{4}{2} {\pi_1}^2 (1-\pi_1)^2 \binom{4}{1} {\pi_2}^1 (1-\pi_2)^3$$

Bondad de Ajuste

Regresión Logística

Verosimilitud

0

$Ln(\mathcal{L}_{b_0})$

Logaritmo de la verosimilitud
del modelo nulo

Mínima Verosimilitud Posible

$Ln(\mathcal{L}_{b_0;b_1;\ldots;b_k})$

Logaritmo de la verosimilitud
del modelo estimado

$Ln(\mathcal{L}_{Mod\ Sat})$

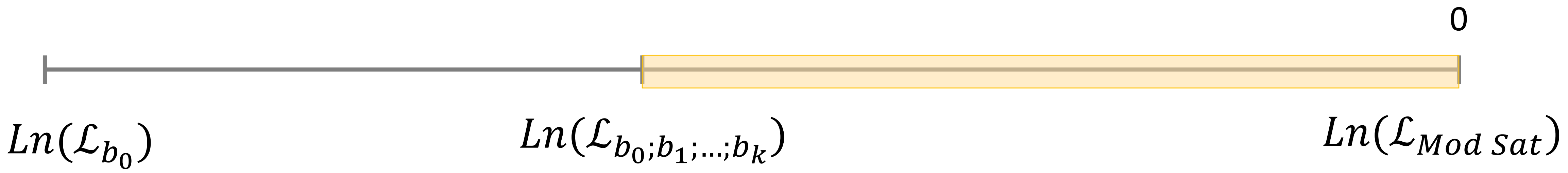
Logaritmo de la verosimilitud
del modelo saturado

Máxima Verosimilitud Posible

Regresión Logística

Devianza y R^2 McFadden

$$D = -2\left[Ln\left(\mathcal{L}_{b_0;b_1;\dots;b_k}\right) - Ln(\mathcal{L}_{Mod\ Sat})\right]$$
$$= -2\ Ln\left(\mathcal{L}_{b_0;b_1;\dots;b_k}\right)$$



$$R^2_{MF} = 1 - \frac{Ln\left(\mathcal{L}_{b_0;b_1;\dots;b_k}\right)}{Ln\left(\mathcal{L}_{b_0}\right)} = 1 - \frac{D}{D_0}$$

$$D_0 = -2Ln(\mathcal{L}_{b_0})$$

Regresión Logística

AIC y BIC

$$AIC = -2\,Ln\left(\mathcal{L}_{b_0;b_1;\ldots;b_k}\right) + 2\mathrm{k}$$

$$BIC = -2\,Ln\left(\mathcal{L}_{b_0;b_1;\ldots;b_k}\right) + 2\mathrm{k}\log(\mathrm{n})$$

Regresión Logística

Significancia de los coeficientes

$$\frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}} \sim \mathrm{N}(0;1)$$

$$\frac{\hat{\beta}_i^2}{Var(\hat{\beta}_i)} \sim \chi^2_{v=1}$$

Regresión Logística
Interpretación Salida R

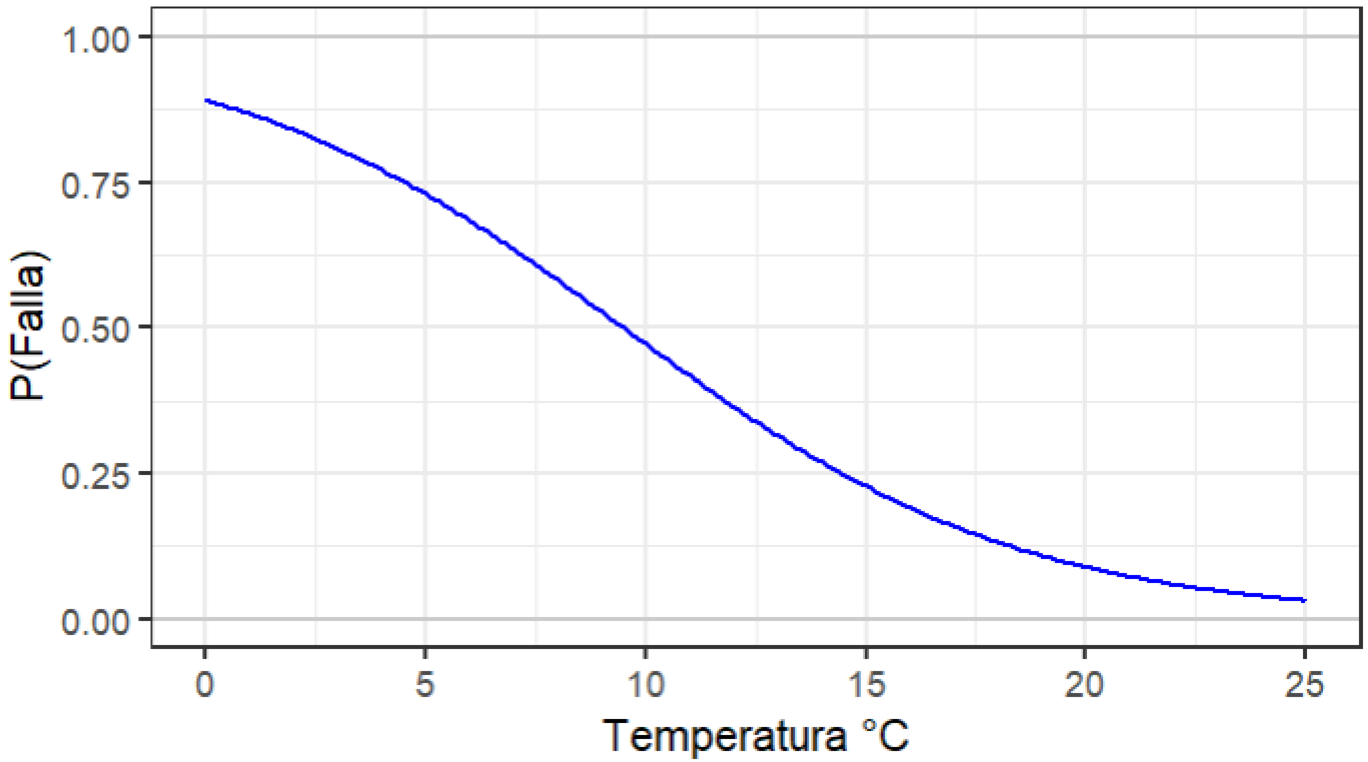
```
Call:
glm(formula = Falla ~ grados_C, family = binomial, data = challenger)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9813  -0.4578  -0.3837  -0.2678   2.5407

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.10334    1.67007   1.259   0.2079
grados_C    -0.22146    0.08924  -2.482   0.0131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58.932  on 91  degrees of freedom
Residual deviance: 52.469  on 90  degrees of freedom
AIC: 56.469

Number of Fisher Scoring iterations: 5
```

Regresión Logística

INTERPRETACION DE LOS PARAMETROS

Regresión Logística

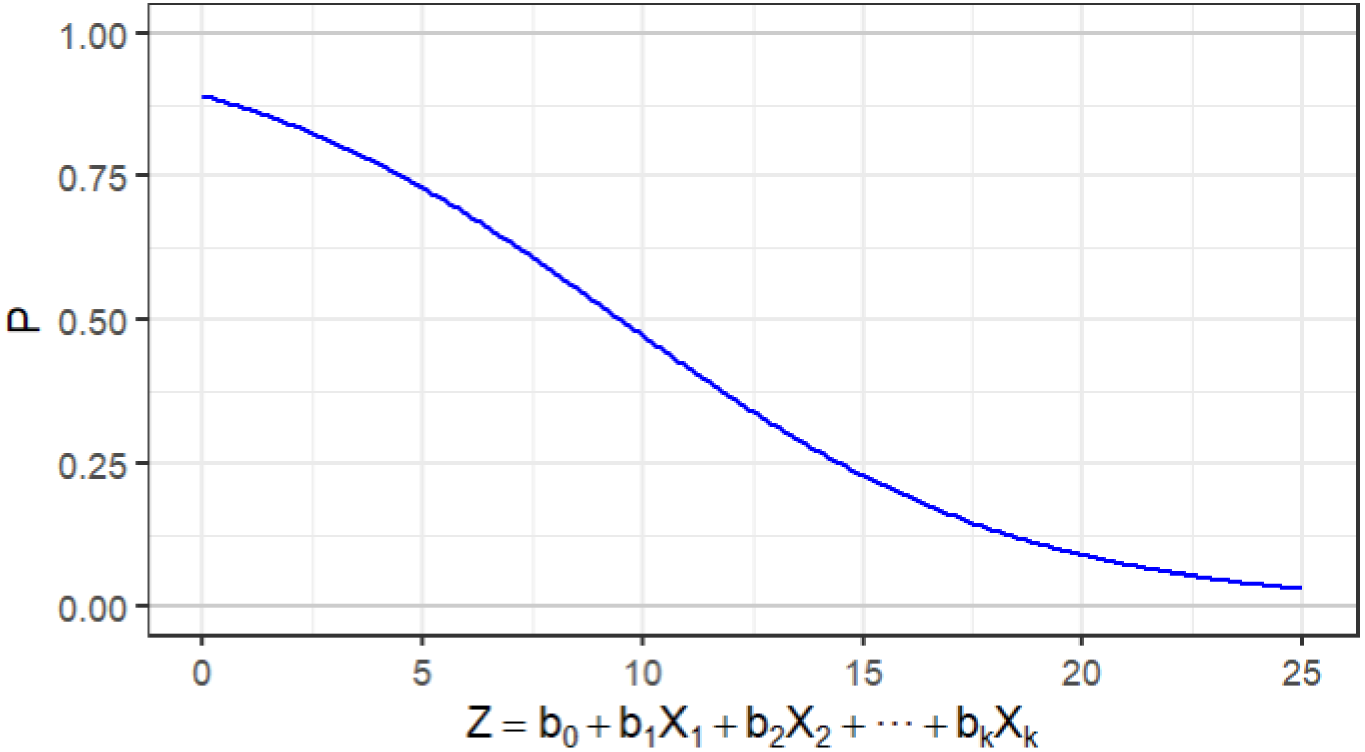
Interpretación de los parámetros



$$P(Falla|Temperatura) = \frac{1}{1 + e^{-(b_0 + b_1 Temperatura)}}$$

Regresión Logística

Interpretación de los parámetros



$$P(Y = 1|X_1; X_2; \ldots; X_k) = \frac{1}{1+e^{-(b_0+b_1X_1+\ldots+b_kX_k)}} = \frac{1}{1+e^{-Z}}$$

Regresión Logística

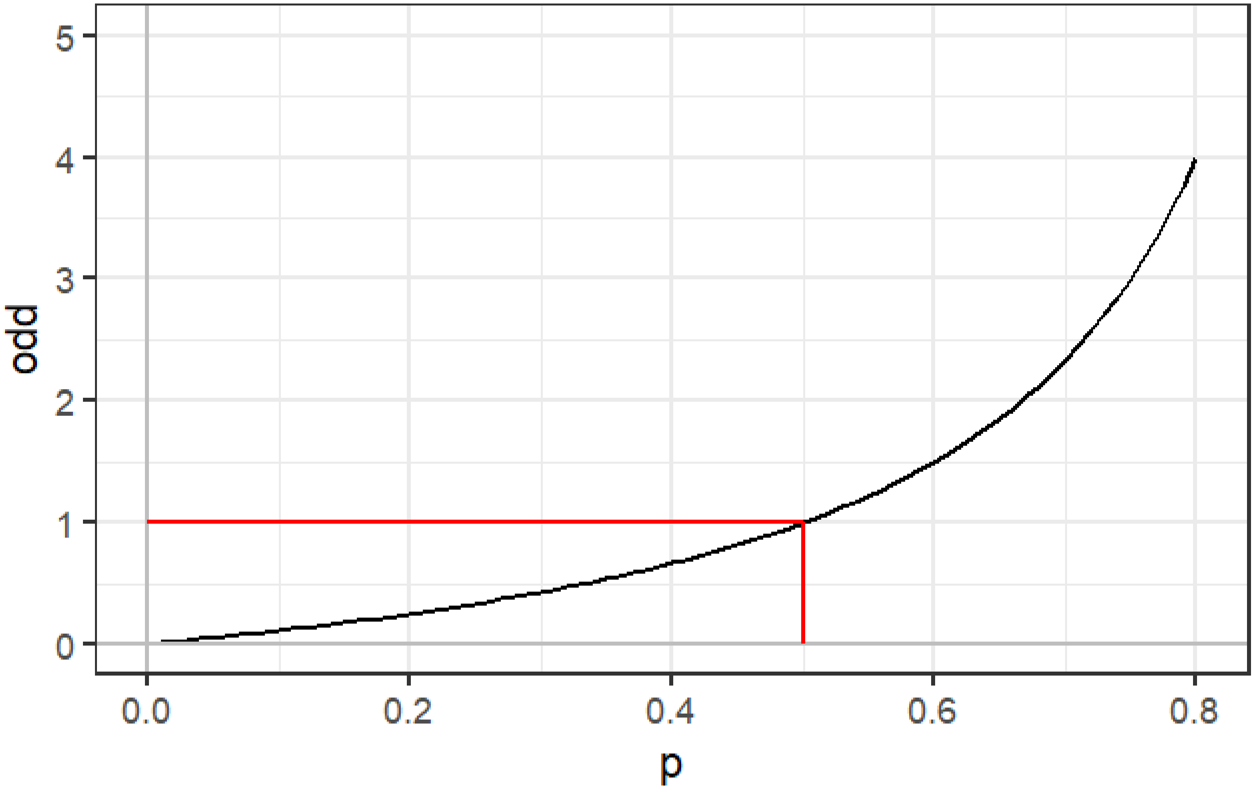
Función Logit

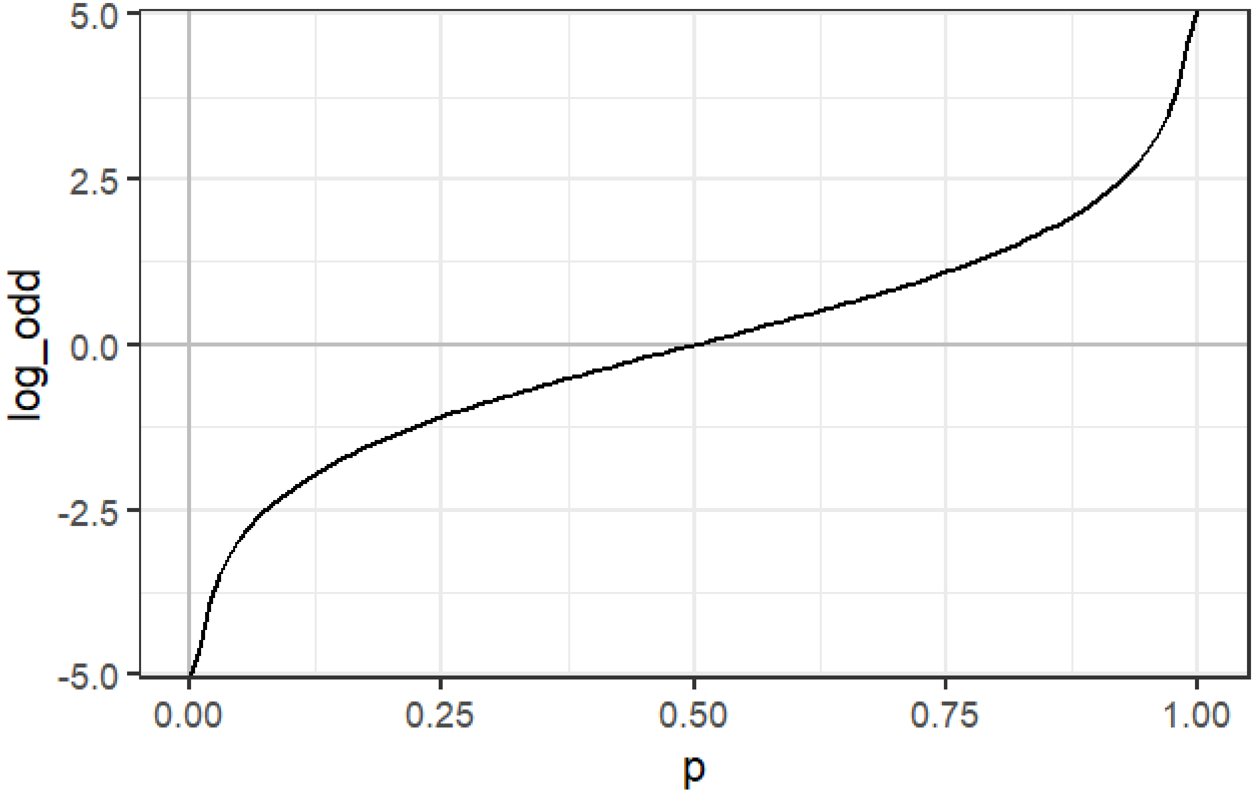
$$\pi = \frac{1}{1 + e^{-Z}} \longrightarrow ln\left(\frac{\pi}{1-\pi}\right) = Z = \beta_0 + \beta_1 X_1 + \ \dots + \beta_k X_k$$

$$ln\left(\frac{P(Y=1|Z)}{1-P(Y=1|Z)}\right) = ln\left(\underbrace{\boxed{\frac{P(Y=1|Z)}{P(Y=0|Z)}}}_{\text{ODD}}\right) = \beta_0 + \beta_1 X_1 + \ \dots + \beta_k X_k$$

Regresión Logística

ODD





Regresión Logística

ODD Ratio

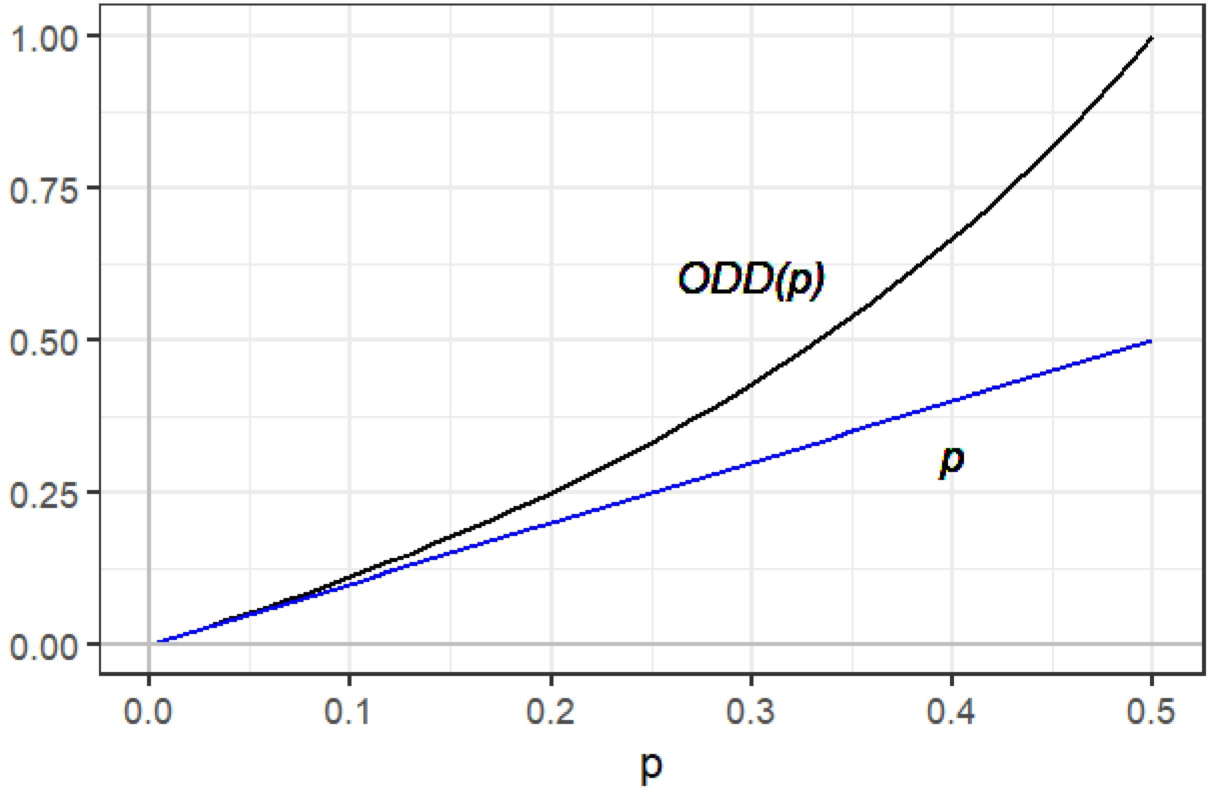
$$ln\left(ODD_{X_1}\right) = \beta_0 + \beta_1 X_1 + \ \dots + \beta_k X_k$$
$$ln\left(ODD_{X_1+1}\right) = \beta_0 + \beta_1 (X_1{+}1) + \ \dots + \beta_k X_k$$

$$ODD_{X_1} = e^{\beta_0+\beta_1 X_1+\ \dots\ +\beta_k X_k}$$
$$ODD_{X_1+1} = e^{\beta_0+\beta_1 (X_1+1)+\ \dots\ +\beta_k X_k}$$

$$OR = \frac{ODD_{X_1+1}}{ODD_{X_1}} = e^{\beta_1}$$

Regresión Logística

ODDs vs P

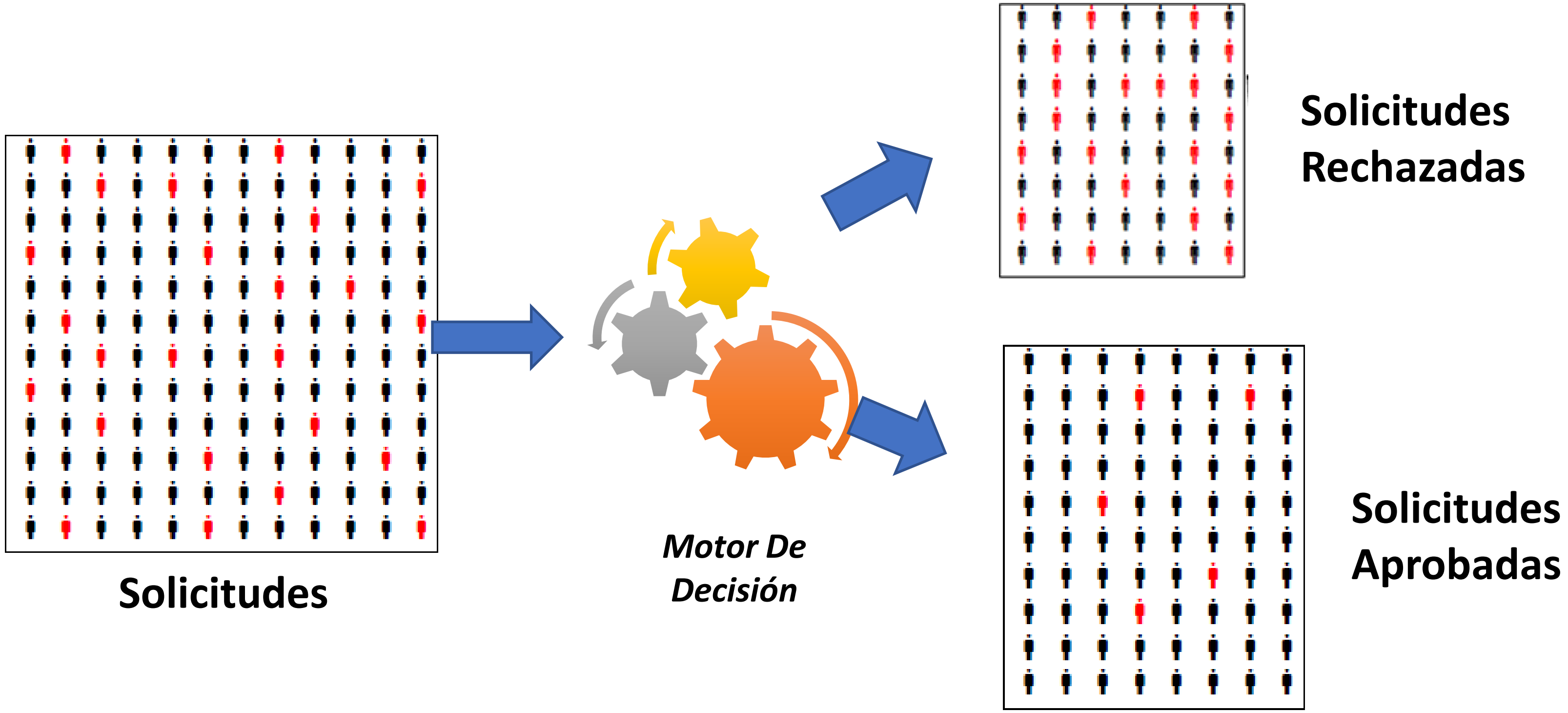


Regresión Logística

REGRESIÓN LOGÍSTICA VS CLASIFICACIÓN

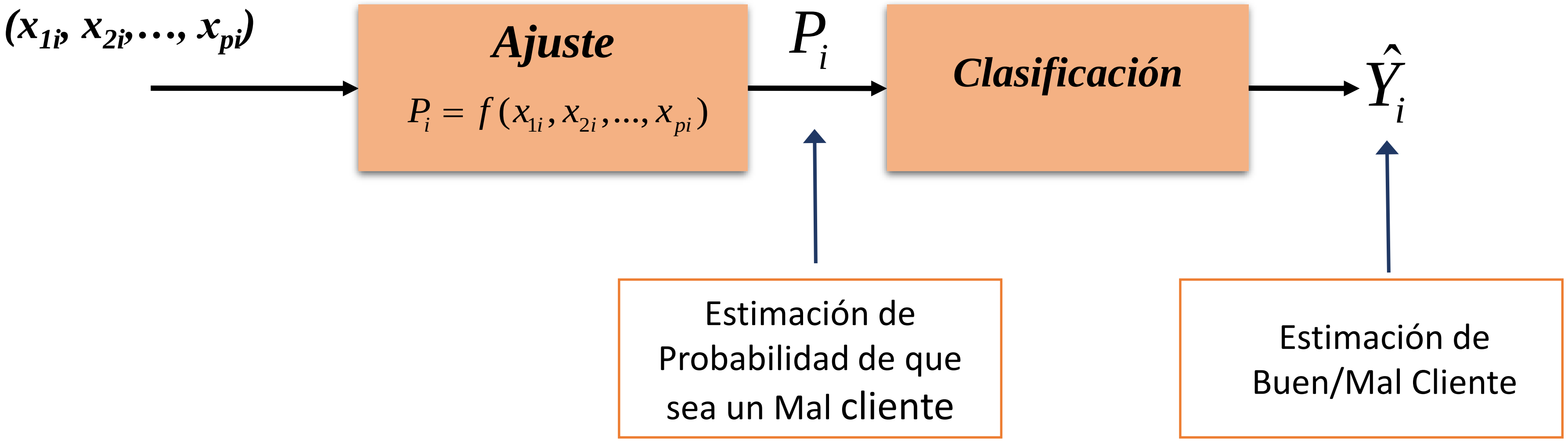
Regresión Logística

Regresión Logística y Clasificación



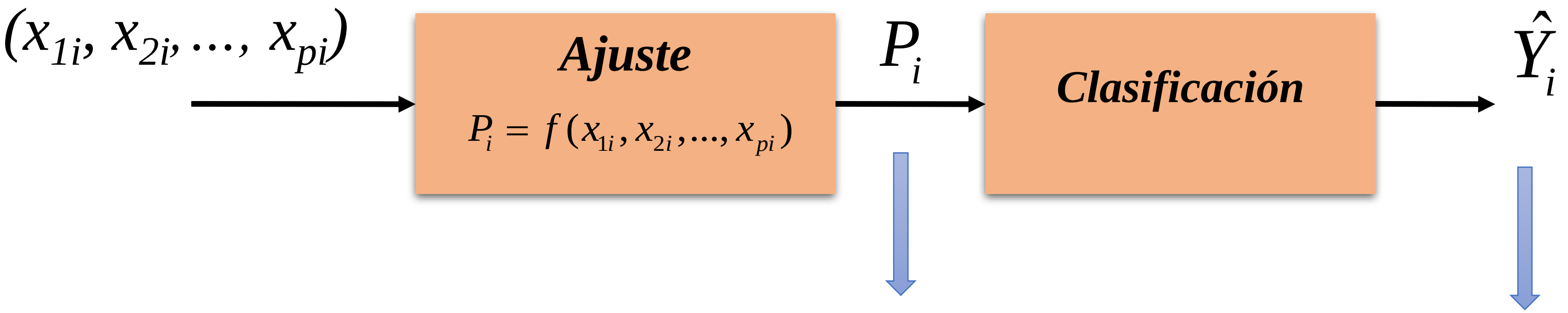
Regresión Logística

Regresión Logística y Clasificación



Regresión Logística

Regresión Logística y Clasificación

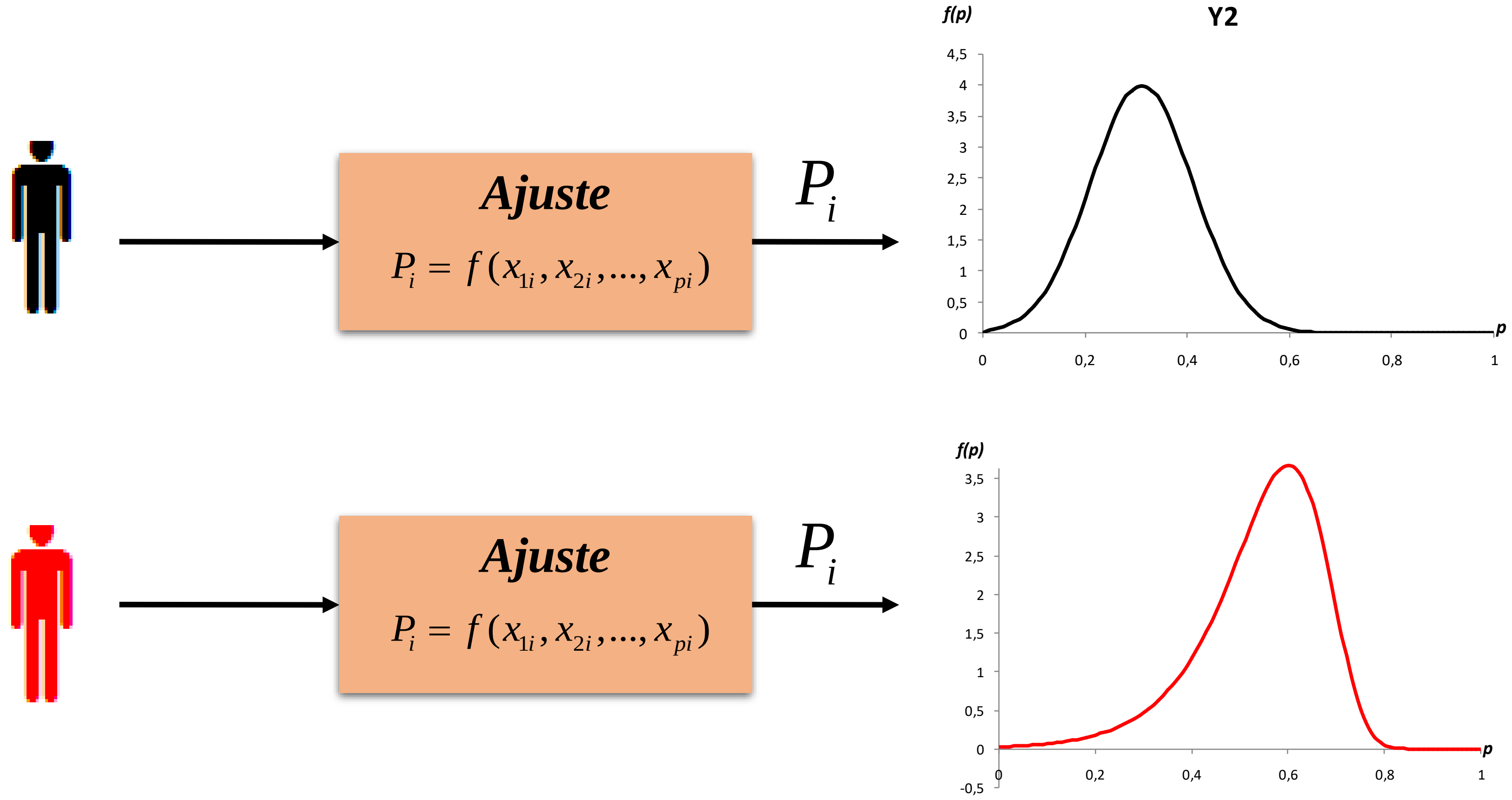


$$P_i = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \ldots + b_k X_k)}}$$

$$\hat{Y}_i = \begin{cases} 0 & P_i \leq K \\ 1 & P_i > K \end{cases}$$

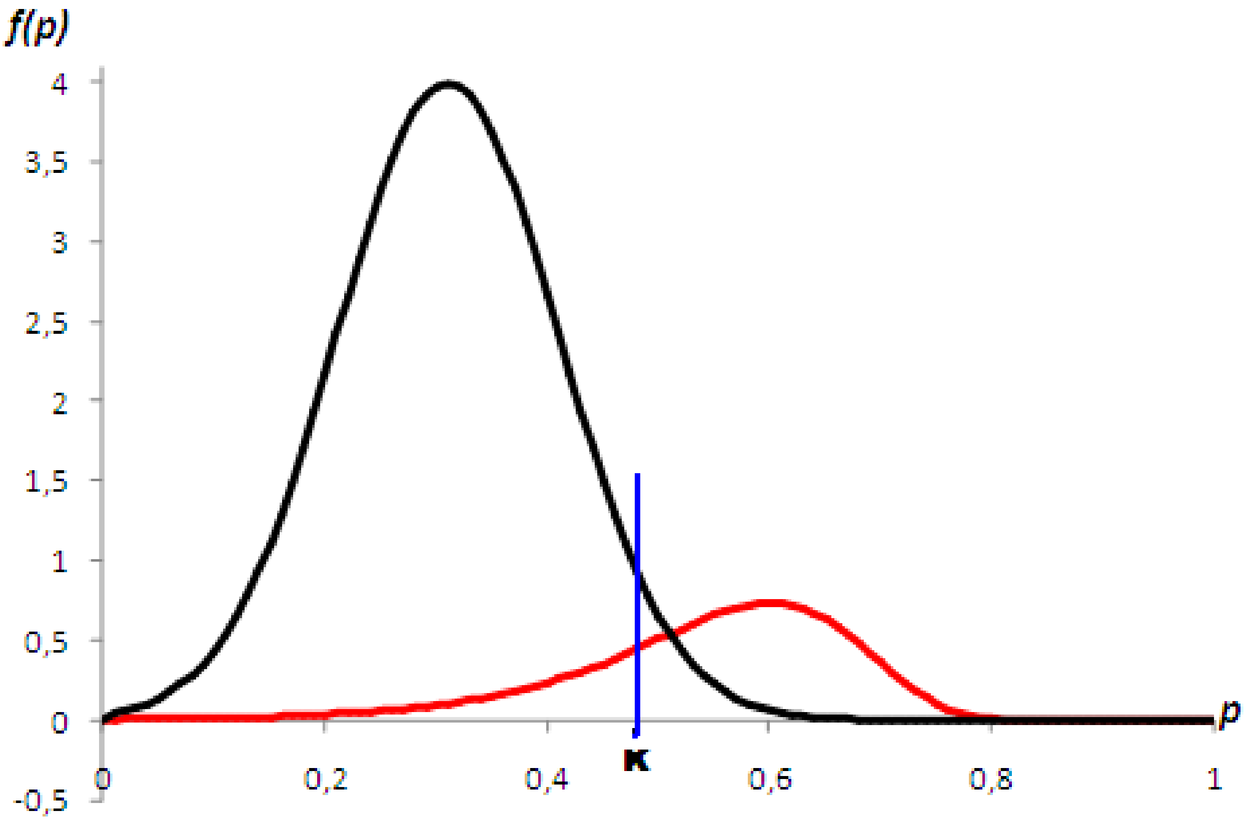
Regresión Logística

Regresión Logística y Clasificación



Regresión Logística

Regresión Logística y Clasificación



Regresión Logística

MATRIZ DE CONFUSIÓN

Regresión Logística

Matriz de confusión

		Valores Reales	
		$\mathbf{Y=1}$	$\mathbf{Y=0}$
Valores Predichos	$\hat{Y}=1$	TP	FP
	$\hat{Y}=0$	FN	TN

- **TP**: Verdaderos Positivos (True Positives)
- **FP**: Falsos Positivos (False Positives)
- **TN**: Verdaderos Negativos (True Negatives)
- **FN**: Falsos Negativos (False Negatives)

Regresión Logística

Matriz de confusión

		Valores Reales	
		$\mathbf{Y} = 1$	$\mathbf{Y} = 0$
Valores Predichos	$\hat{Y} = 1$	TP	FP
	$\hat{Y} = 0$	FN	TN

$$Exactitud\ (Accuracy) = \frac{TP + TN}{TP + FP + FN + TN}$$

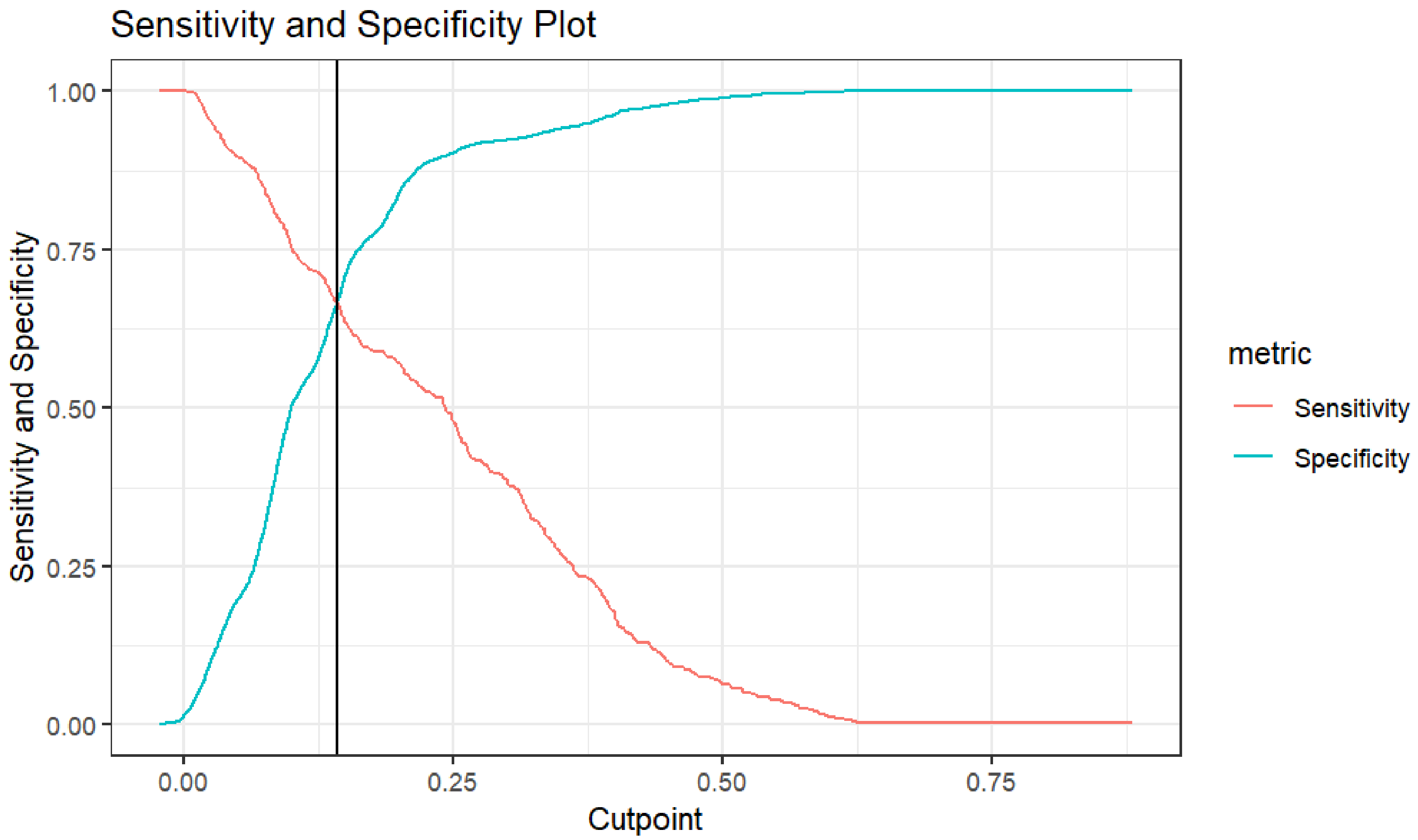
$$Sensibilidad\ (Sensitivity) = \frac{TP}{TP + FN}$$

$$Precisión\ (Precision) = \frac{TP}{TP + FP}$$

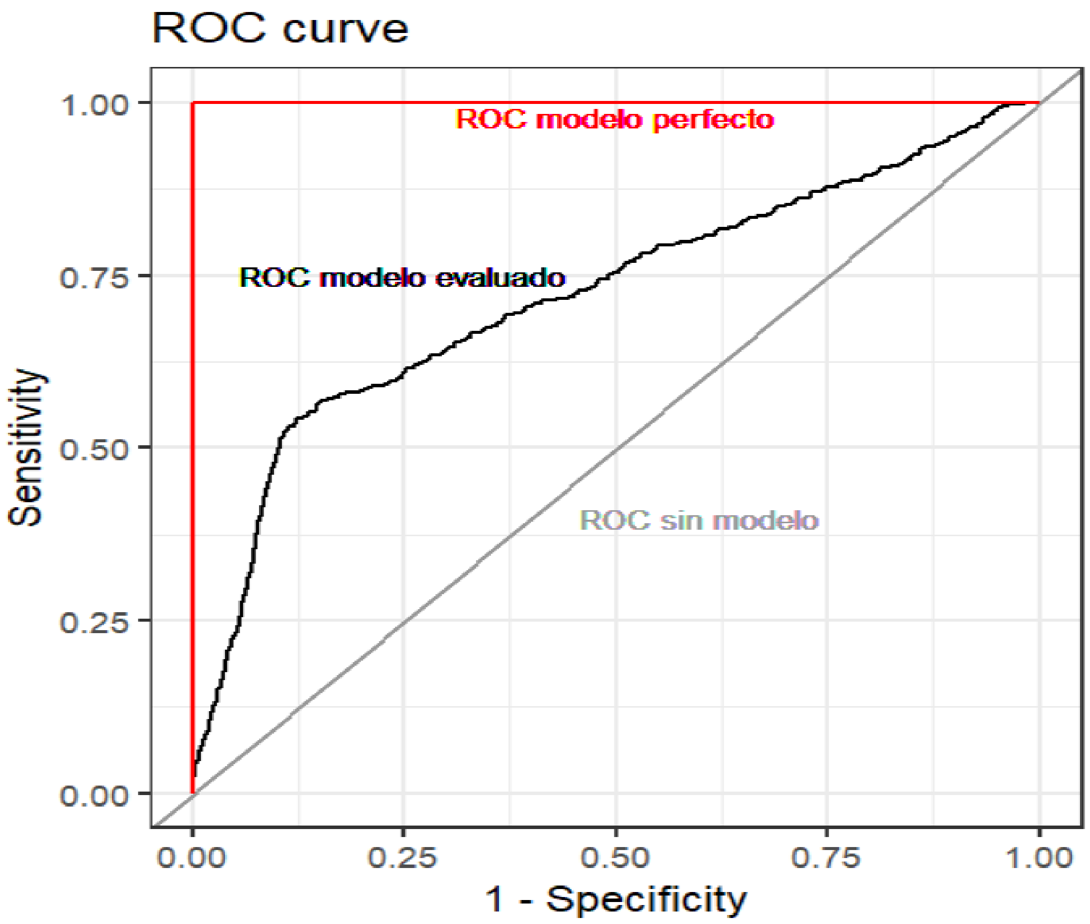
$$Especifidad\ (Specificity) = \frac{TN}{TN + FP}$$

Regresión Logística

Gráfico de Sensibilidad / Especificidad



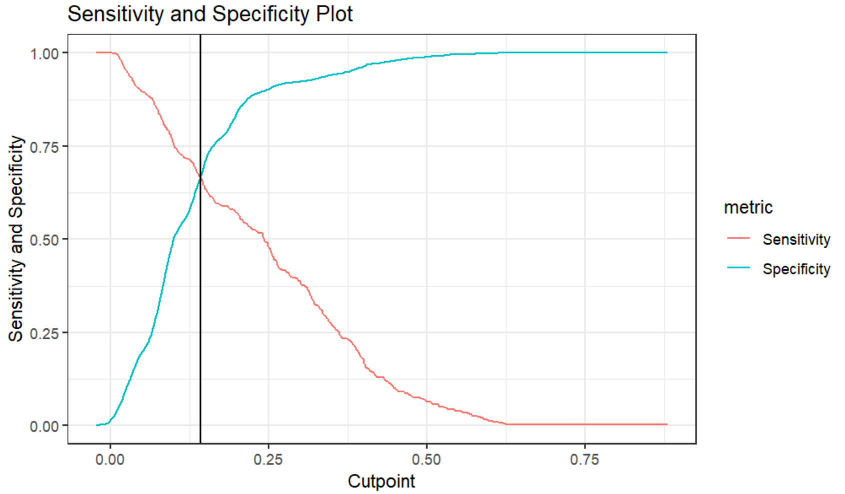
Regresión Logística
Curva ROC

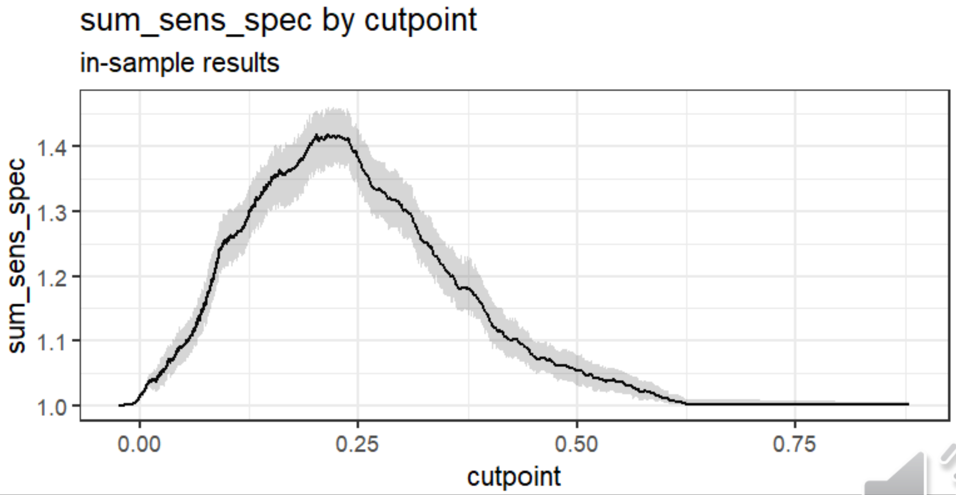


Regresión Logística

Selección de K

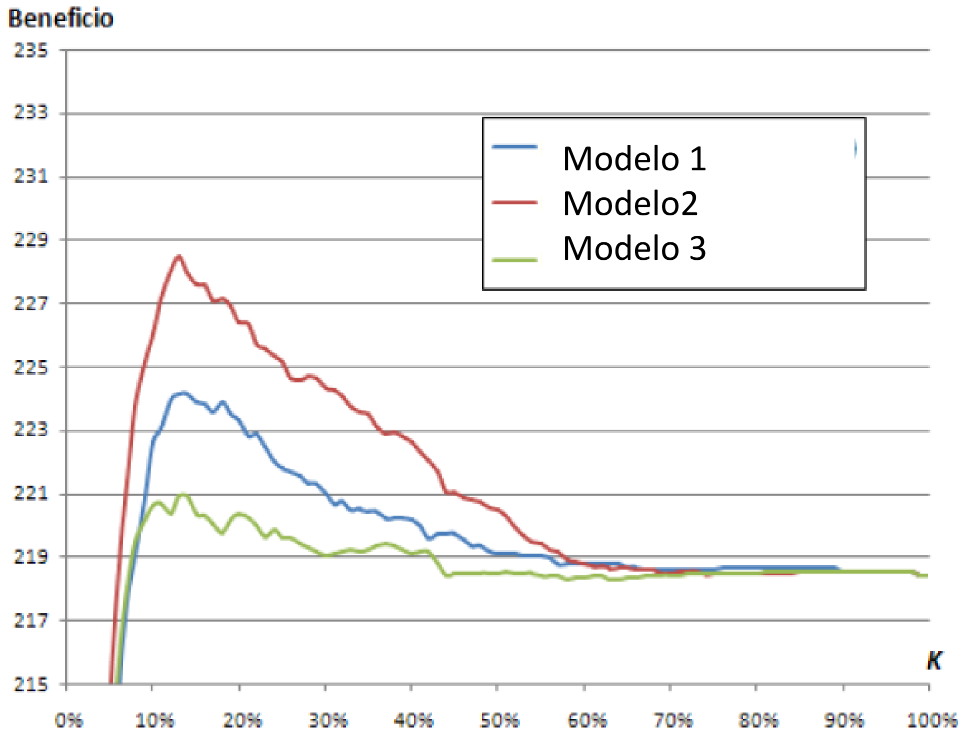
- $K = 0{,}50$
- $K = \hat{p} = \dfrac{TP+FN}{TP+FP+FN+TN}$
- $K =$ Valor donde se cruzan Sensitividad y Sensibilidad
- $K =$ Valor con máxima Sensitividad + Sensibilidad
- $K =$ Valor con máximo Beneficio Económico





Regresión Logística
Selección de K – Maximización de beneficio económico

		Comportamiento Pronosticado	
		Buen Comportamiento	Mal Comportamiento
Comportamiento Real	Buen Comportamiento	B_{VB}	C_{FM}
	Mal Comportamiento	C_{FB}	B_{VM}



B_{VB}: Beneficio obtenido por un Verdadero cliente con Buen Comportamiento

B_{VM}: Beneficio obtenido por un Verdadero cliente con Mal Comportamiento

C_{FM}: Costo incurrido por un Falso cliente con Mal Comportamiento

C_{FB}: Costo incurrido por un Falso cliente con Buen Comportamiento