PRINTS DURANTE A AULA 20/08/2024

Obs	Dummy Y	X_1	X_2	...	X_k
M	1 sim				⋮
J	0 não				
P	1 sim				
A	1 sim				
⋮	0 não				
	0 não				
	1 sim				
	⋮				
n	1 sim				

↳ dicotômica

↳ reg. logística binária!

Y:
1: evento (sim)
0: não evento (não)

P: probabilidade do evento.

1 - P: probabilidade do não evento.

$$\text{chance (odds)} = \frac{P}{1-P}$$

evento

não evento

$P = 0{,}80 \rightarrow \text{chance} = \frac{4}{1} = \boxed{4}$

$P = 0{,}25 \rightarrow \text{chance} = \frac{0{,}25}{0{,}75} = \boxed{\frac{1}{3}}$

$P = 0{,}50 \rightarrow \text{chance} = \frac{1}{1} = \boxed{1}$

MQO (OLS):

$$\hat{Y}_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Logística Binária: (Hosmer & Lemeshow Communications in Statistics, 1980)

$$\ln(chance_i) = \underbrace{\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}}_{Z: \text{ logito}}$$

$$\ln\left(\frac{p}{1-p}\right) = Z$$

$$\frac{p}{1-p} = e^{Z}$$

$$p \cdot (1 + e^{Z}) = e^{Z}$$

$$p = \frac{e^{Z}}{1+e^{Z}} = \frac{1}{1+e^{-Z}}$$

Prob. evento

$$p_i = \frac{1}{1 + e^{-(\alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki})}}$$

Origem: Juros Compostos (mat. Financeira).

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n \cong 2{,}71828$$

(~ começo séc 18)

Número Euler

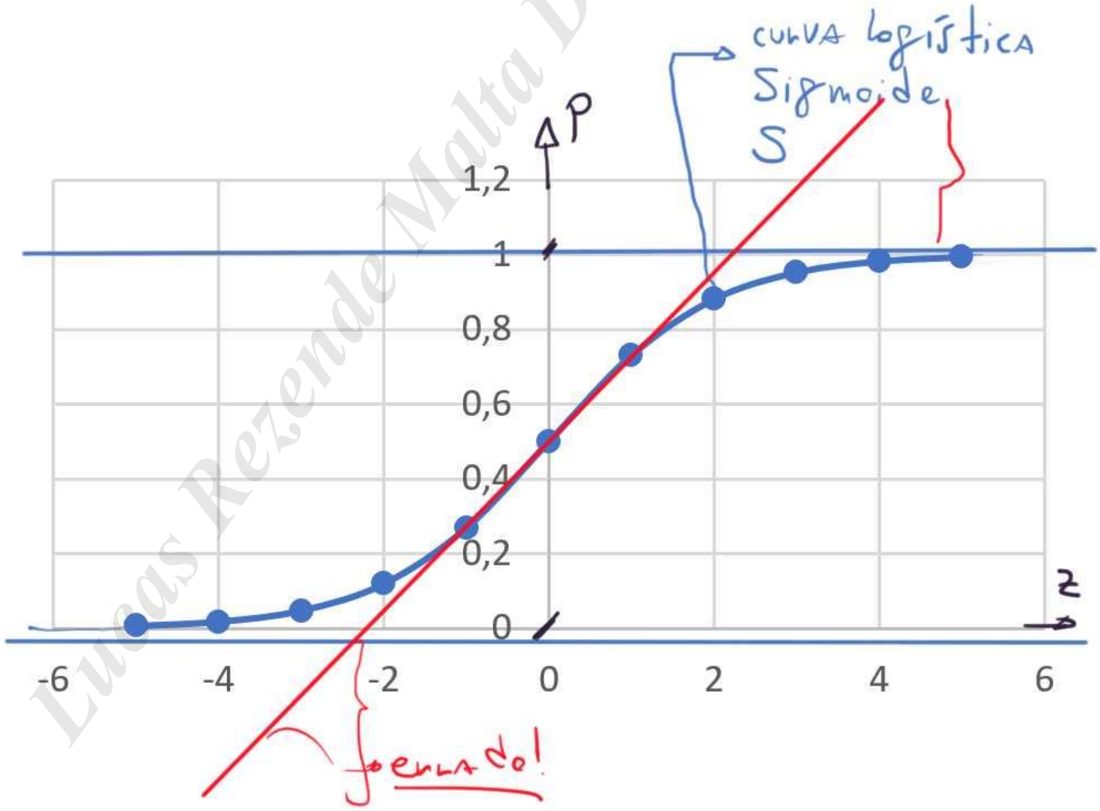
número Napier (~ final séc. 16) } "o descobridor de logaritmos".

$$\ln = \log_e$$

$$\underset{\substack{\text{evento}\\ y=1}}{P} = \frac{e^{z}}{1+e^{z}} = \boxed{\frac{1}{1+e^{-z}}}$$

$$\underset{\substack{\text{não evento}\\ y=0}}{P} = 1 - \frac{e^{z}}{1+e^{z}} = \boxed{\frac{1}{1+e^{z}}}$$

Soma = 1.



dist

atrasado
Y=1

sem

$$P_{\underset{y=1}{atrasado=sim_i}} = \frac{1}{1 + e^{-(\alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i)}}$$

$$p(Y_i) = p_i^{Y_i} \cdot (1 - p_i)^{1 - Y_i}$$

Distribuição Bernoulli.

$$p(1) = p^1 \cdot (1 - p)^0 = p$$

$$p(0) = p^0 \cdot (1 - p)^1 = 1 - p.$$

α	-26,16
β_1	0,19
β_2	2,36

Somatória LL_i	-50,46638

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               atrasado   No. Observations:                  100
Model:                            GLM   Df Residuals:                       97
Model Family:                Binomial   Df Model:                            2
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -50.466
Date:                Tue, 20 Aug 2024   Deviance:                       100.93
Time:                        21:34:51   Pearson chi2:                     86.7
No. Iterations:                     7   Pseudo R-squ. (CS):             0.2913
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -26.1665      8.442     -3.100      0.002     -42.712      -9.621
dist           0.1904      0.076      2.493      0.013       0.041       0.340
sem            2.3629      0.795      2.972      0.003       0.804       3.921
==============================================================================
```

Wald

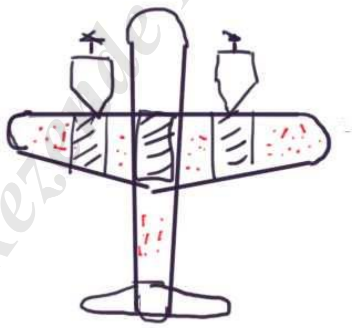
α

β_1

β_2

$$P_{atrasado_i}(Y=1) = \frac{1}{1 + e^{-(-26,16 + 0,19 \cdot dist_i + 2,36 \cdot sem_i)}}$$

Z de WAld: (1902-1950)



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               atrasado   No. Observations:                  100
Model:                          Logit   Df Residuals:                       97
Method:                           MLE   Df Model:                            2
Date:                Tue, 20 Aug 2024   Pseudo R-squ.:                  0.2544
Time:                        22:17:05   Log-Likelihood:                -50.466
converged:                       True   LL-Null:                       -67.686
Covariance Type:            nonrobust   LLR p-value:                 3.324e-08
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -26.1665      8.442     -3.100      0.002     -42.713      -9.620
dist           0.1904      0.076      2.493      0.013       0.041       0.340
sem            2.3629      0.795      2.972      0.003       0.804       3.921
==============================================================================
```

maximum likelihood est.

(Análogo p-value F) χ^2

α, β_1, β_2

$$\chi^2 = -2 \cdot (LL_0 - LL_m)$$

↑ melhor

$$\text{pseudo } R^2 = \frac{-2LL_0 - (-2LL_m)}{-2LL_0}$$

McFadden

↓ melhor

$$\boxed{AIC = -2LL_m + 2 \cdot (\underbrace{k+1}_{\beta's}) \quad BIC = -2 \cdot LL_m + (\underbrace{k+1}_{\beta's}) \cdot \ln(n)}$$

cutoff

Se phat $\geq$ cutoff $\Rightarrow$ Evento

Se phat $<$ cutoff $\Rightarrow$ Não Evento

Matriz de Confusão: (Classificação Cruzada)

cutoff = 0,5

True (Real)

classified (predito)

	1	0
1	46	16
0	13	25
	59	41

Sensitividade (Recall) $= \frac{46}{59}$ | Especificidade $= \frac{25}{41}$ | Acurácia (EGM) $= \frac{46+25}{100}$

LogLik (LLf)
pseudo R^2
AIC
BIC

} Indicador performance do modelo (independem do cutoff).

Sensitividade
Especificidade
Acurácia

} Dependem do cutoff.