PRINTS DURANTE A AULA 20/08/2024

P: probabilidade do evento.

$1-P$: probabilidade do NÃO evento.

$$\text{chance (odds)} = \frac{P}{1-P} \quad \begin{array}{l}\text{evento}\\ \text{NÃO evento}\end{array}$$

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

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

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

MQO (OLS):

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

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

$$\ln(\text{chance}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

$$\ln\left(\frac{p}{1-p}\right) = z \qquad z: \text{logito}$$

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

$$p \cdot (1 + e^z) = e^z$$

$$P = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \to \text{Prob. evento}$$

$$P_i = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K)}}$$

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 $\Big\}$ "o descobridor
(~ final séc 16) de logaritmos".

$$\ln = \log_e$$

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

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

$$\underline{Soma = 1.}$$

CURVA logística
Sigmoide
S

$\uparrow P$

1,2
1
0,8
0,6
0,4
0,2
0

-6   -4   -2   0   2   4   6

z

foerrado!

dist

atrasado
$y=1$

sem
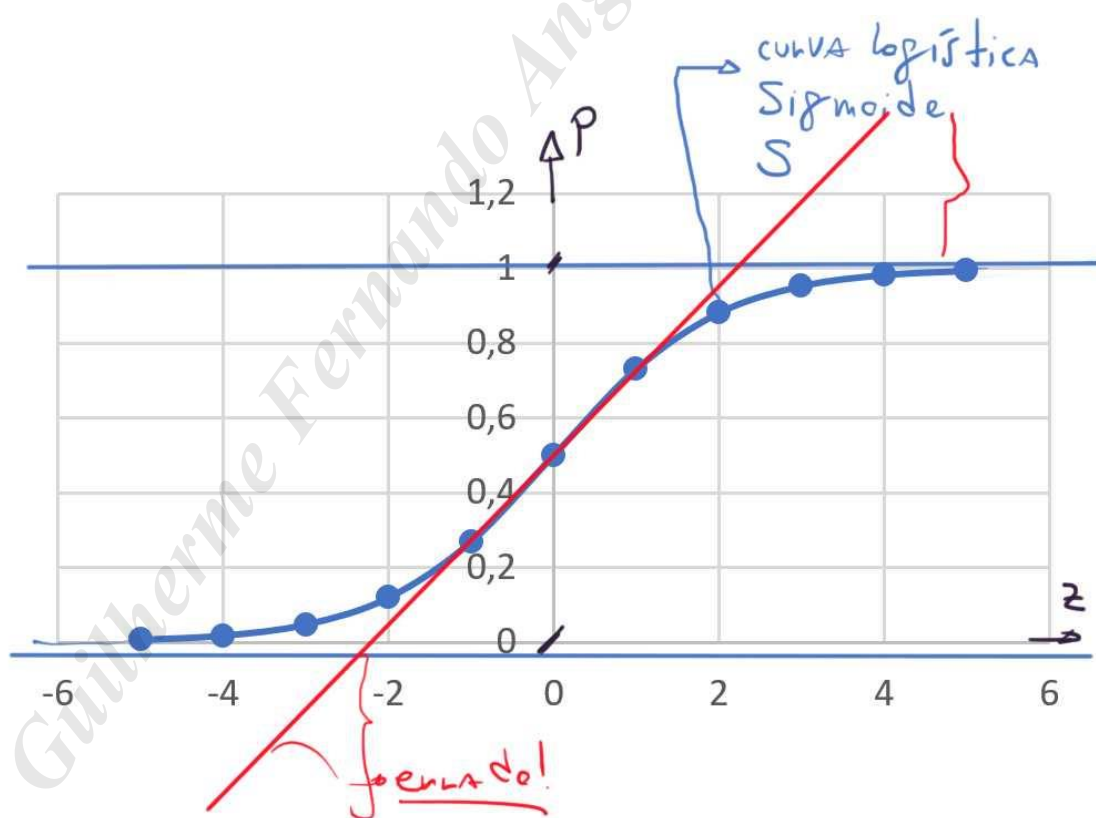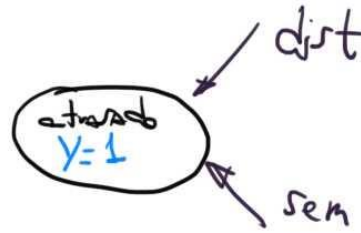
$$P_{\substack{atrasado = sim \\ y=1}_i} = \cfrac{1}{1 + e^{-(\alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i)}}$$

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

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 | Somatória LL_i | -50,46638

β_2 | 2,36

```
                 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
------------------------------------------------------------------------------
```
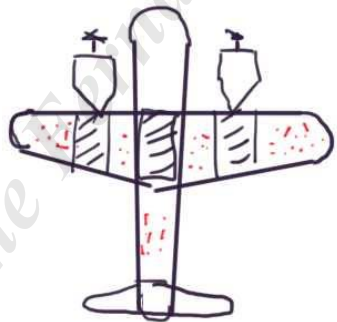
(handwritten annotations on table: α, β_1, β_2; WALD above z and P>|z|)

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

Z de WALd: (1902 - 1950)

```
                        Logit Regression Results
===============================================================================
Dep. Variable:              atrasado   No. Observations:                 100
Model:                         Logit   Df Residuals:                      97
Method:  maximum likelihood est.  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:  (Análogo p-value F) 3.324e-08  χ²
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept  α   -26.1665      8.442     -3.100      0.002     -42.713      -9.620
dist       β₁    0.1904      0.076      2.493      0.013       0.041       0.340
sem        β₂    2.3629      0.795      2.972      0.003       0.804       3.921
===============================================================================
```

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

↑ melhor

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

↓ melhor

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

cutoff.

$$\text{Se } \hat{p} \geq \text{cutoff} \Rightarrow \text{Evento}$$

$$\text{Se } \hat{p} < \text{cutoff} \Rightarrow \text{NÃO Evento}$$

Matriz de Confusão: $\begin{pmatrix} Classificação \\ Cruzada \end{pmatrix}$

cutoff = 0,5

True (Real)

	①	0	
classified (previstol) 1	㊺ 46	16	
0	13	㉕ 25	
	59	41	

$$\text{Sensitividade (Recall)} = \frac{46}{59} \quad \Big| \quad \text{Especificidade} = \frac{25}{41} \quad \Big| \quad \text{Acurácia (FGV)} = \frac{46+25}{100}$$

$$\left.\begin{array}{l} \text{LogLik (LLf)} \\ \text{pseudo } R^2 \\ \text{AIC} \\ \text{BIC} \end{array}\right\} \text{Indicador performance do Modelo. (independem de cutoff).}$$

$$\left.\begin{array}{l} \text{Sensitividade} \\ \text{Especificidade} \\ \text{Acurácia} \end{array}\right\} \text{Dependem do cutoff.}$$