Universidade do Estado do Rio de Janeiro
Instituto Politécnico
Departamento de Modelagem Computacional

# Disciplina de Modelos Lineares 2012-2

## Professora Ariane Ferreira

# Regressão Logística

O modelo de regressão logístico é semelhante ao modelo de regressão linear. No entanto, no modelo logístico a variável resposta $Y_i$ é binária. Uma variável binária assume dois valores, como por exemplo, $Y_i = 0$ e $Y_i = 1$, denominados "fracasso" e "sucesso", respectivamente. Neste caso, "sucesso" é o evento de interesse.

No modelo linear temos

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Assumindo que $E(\varepsilon_i) = 0$, obtemos que

$$E(Y_i) = \beta_0 + \beta_1 x_i. \quad (4.1)$$

A variável resposta $Y$ tem distribuição Bernoulli $(1, \pi)$, com probabilidade de sucesso $P(Y_i = 1) = \pi_i$ e de fracasso $P(Y_i = 0) = 1 - \pi_i$. Desta forma

$$E(Y_i) = \pi_i. \quad (4.2)$$

Igualando (4.2) e (4.1), temos

$$E(Y_i) = \pi_i = \beta_0 + \beta_1 x_i.$$

Essa igualdade viola as suposições do modelo linear. De fato,

**i)** Os erros não são normais, pois:

- $y_i = 1 \quad \Rightarrow \quad \varepsilon_i = 1 - \beta_0 - \beta_1 x_1$
- $y_i = 0 \quad \Rightarrow \quad \varepsilon_i = 0 - \beta_0 - \beta_1 x_1$

Assim não faz sentido assumirmos a normalidade dos erros.

**ii)** Não homogeneidade da variância.

Temos que $\mathrm{Var}(Y_1) = \pi_i(1 - \pi_i) = (\beta_0 + \beta_1 x_1)(1 - \beta_0 - \beta_1 x_1)$ então a variância de $Y_i$ depende de $x_i$, e consequentemente, não é constante.

**iii)** Restrição para a resposta média $E(Y_i)$. Como a resposta média é obtida em probabilidades temos que $0 \leq \beta_0 + \beta_1 x_1 \leq 1$. Entretanto, esta restrição é inapropiada para resposta em um modelo linear, que assume valores no intervalo $(-\infty, \infty)$. Uma forma de resolver esse problema é utilizar o modelo logístico.

Muitas funções foram propostas para a análise de variáveis com respostas dicotômicas. Dentre elas a mais simples é a que dá origem ao modelo logístico. Do ponto de vista estatístico este modelo é bastante flexível e de fácil interpretação.

# Regressão logistica Simples

# Modelo Estatístico

Um modelo de regressão logística simples é usado para o caso de regressão com uma variável explicativa.

Suponha uma amostra de $n$observações independentes da terna $(x_i, m_i, y_i), i = 1, 2, \ldots, n$, sendo que:

- $x_i$ é o valor da variável explicativa;
- $m_i$ é a quantidade de itens verificados na amostra (número de ensaios);
- $y_i$ número de ocorrência de um evento (exemplo: quantidade de peças não conforme) em $m_i$ ensaios; e
- $n$ é o tamanho da amostra.

Com isso, assumimos que a variável resposta tem distribuição de probabilidade binomial $(Y_i \sim B(m_i, \pi_i))$, tal que

$$P[Y_i = y_i] = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}.$$

Para adequarmos a resposta média ao modelo linear usamos a função de ligação

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}, \; i = 1, \ldots, n,$$

que pode ser escrita como

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i.$$

As figuras a seguir ilustram a forma do modelo logístico para $\beta_1$ positivo e negativo.

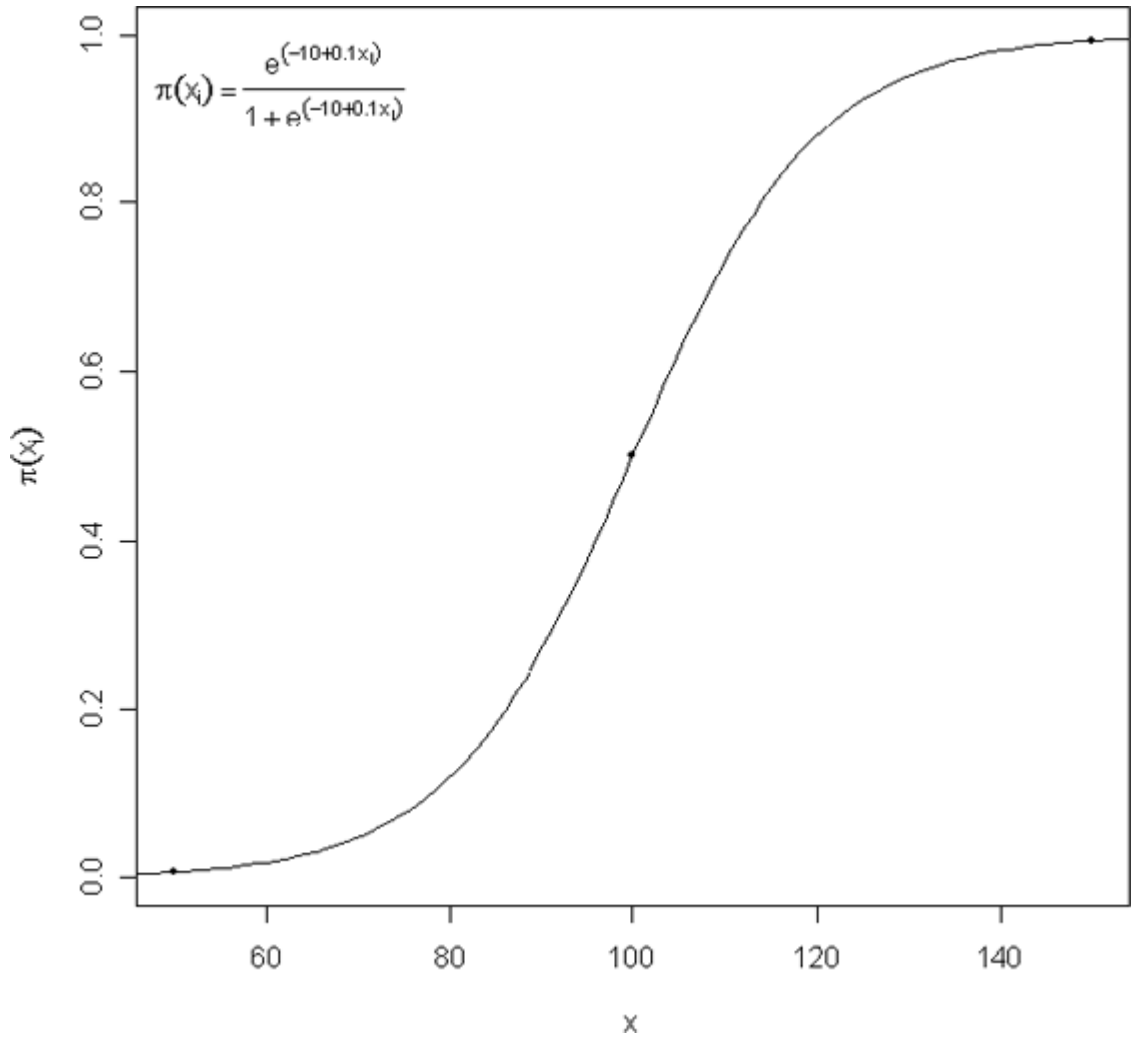


**Figura 4.1.1.1:** Modelo logístico com $\beta_1$positivo.

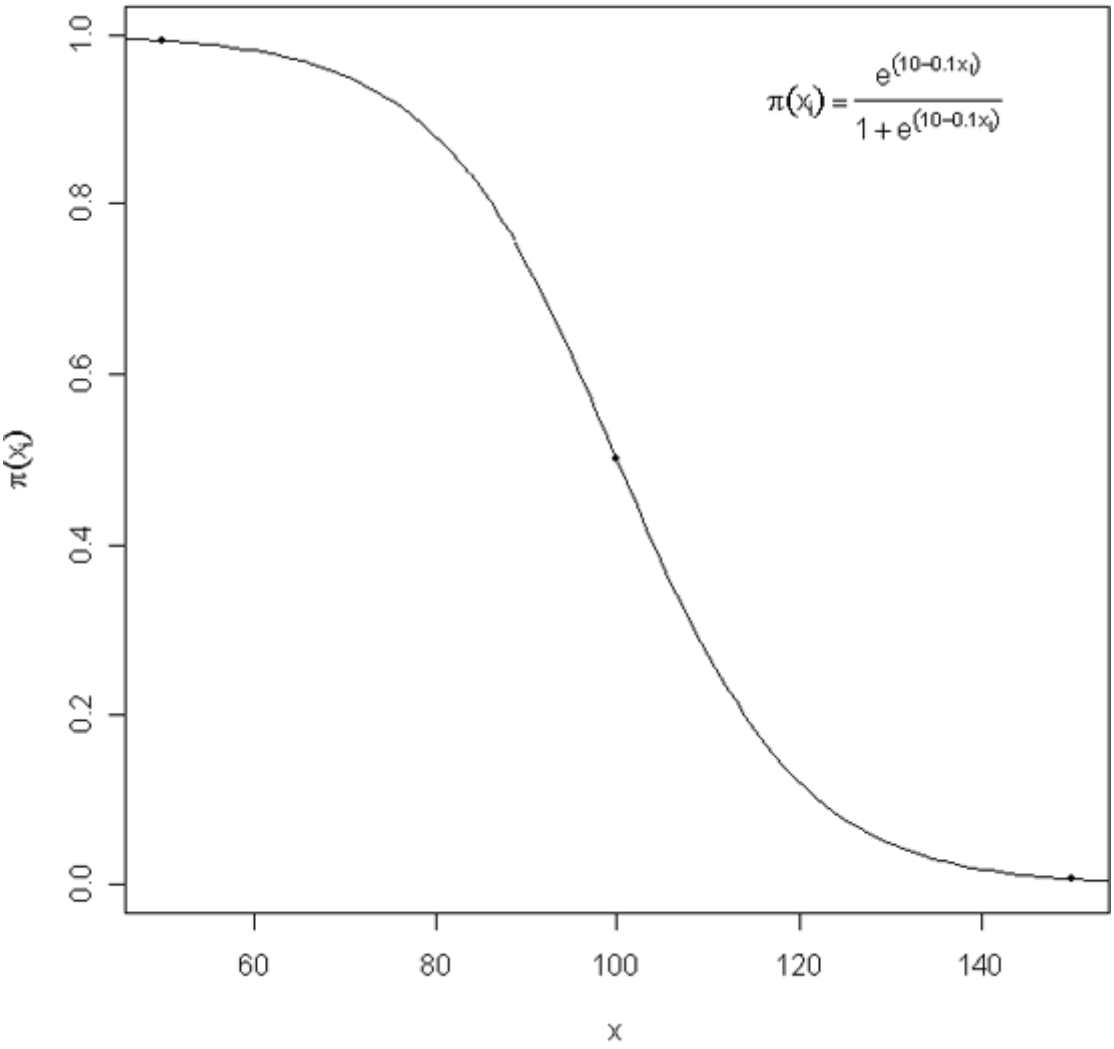


**Figura 4.1.1.2:** Modelo logístico com $\beta_1$negativo.

Neste caso, utilizamos o método da máxima verossimilhança para estimarmos os parâmetros $(\beta_0, \beta_1)$. De forma genérica, o método de máxima verossimilhança nos fornece valores para os parâmetros desconhecidos que maximizam a probabilidade de se obter determinado conjunto de dados.

Assumindo que $(x_0, m_0, y_0), \ldots, (x_n, m_n, y_n)$ são independentes, a função de verossimilhança é da seguinte forma

$$P[Y = y_1, \ldots, y_n | \beta_0, \beta_1] = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

$$= \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i} (1-\pi_i)^{-y_i}$$

$$= \prod_{i=1}^{n} \binom{m_i}{y_i} \frac{\pi_i^{y_i}(1-\pi_i)^{m_i}}{(1-\pi_i)^{y_i}}$$
$$= \prod_{i=1}^{n} \binom{m_i}{y_i} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)^{m_i}$$

Ignorando o termo constante $\binom{m_i}{y_i}$, que não depende de $x_i$, e tomando o logaritmo $(\ln)$ em ambos os lados da expressão anterior, temos

$$Ł\ (\beta_0, \beta_1 | (x_i; m_i; y_i)) = \ln \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} + \ln\ (1-\pi_i)^{m_i}$$
$$Ł\ (\beta_0, \beta_1 | (x_i; m_i; y_i)) = y_i \ln \left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ln(1-\pi_i) \qquad (4.1.1.1)$$

Detalhando $\ln \left(\frac{\pi_i}{1-\pi_i}\right)$, e considerando que,

$\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$, temos

$$\ln \left(\frac{\pi_i}{1-\pi_i}\right) = \ln \left(\frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}\right)$$
$$= \ln \left(\frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}}\right)$$

Assim a expressão (4.1.1.1), pode ser reescrita como:

$$Ł\ (\beta_0, \beta_1 | (x_i; m_i; y_i)) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ln(1-\pi_i)\right]$$
$$= \sum_{i=1}^{n} \left[y_i\ (\beta_0+\beta_1 x_i) + m_i \ln \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right]$$
$$= \sum_{i=1}^{n} \left[y_i\ (\beta_0+\beta_1 x_i) + m_i \ln \left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)\right]$$
$$= \sum_{i=1}^{n} \left[y_i\ (\beta_0+\beta_1 x_i) + m_i\ (\ln 1 - \ln(1+e^{\beta_0+\beta_1 x_i}))\right]$$

$$= \sum_{i=1}^{n} y_i \ (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} m_i \ \ln(1 + e^{\beta_0 + \beta_1 x_i})$$

Portanto,

$$\mathrm{L}\ (\beta_0, \beta_1 | (x_i; m_i; y_i)) = \sum_{i=1}^{n} y_i \ (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} m_i \ \ln(1 + e^{\beta_0 + \beta_1 x_i}) \qquad (4.1.1.2)$$

Para simplificar a notação faremos $\mathrm{L}\ (\beta_0, \beta_1 | (x_i; m_i; y_i)) = \mathrm{L}\ (\beta_0, \beta_1)$.

# Estimação dos Parâmetros do modelo

Para ajustar um modelo de regressão devemos estimar os parâmetros $\beta_0$e $\beta_1$do modelo. Os estimadores de máxima verossimilhança para os parâmetros $\beta_0$e $\beta_1$são os valores de $\widehat{\beta}_0$e $\widehat{\beta}_1$que maximizam o logaritmo da função de verossimilhança. A função de verossimilhança tem máximo, pois $0<P[Y_i = y_i \mid x_i]<1$,pois a função logaritmo é estritamente crescente.

Para maximizar a função de verossimilhança basta derivarmos em relação aos parâmetros do modelo, da seguinte forma

$$\frac{\partial}{\partial \beta_0} \mathrm{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$\frac{\partial}{\partial \beta_1} \mathrm{L}\ (\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \ x_i - \sum_{i=1}^{n} m_i \, x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Igualando estas derivadas a zero e substituindo os parâmetros ($\beta_0, \beta_1$) pelos estimadores $(\widehat{\beta}_0, \widehat{\beta}_1)$,

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}} = 0$$

$$\sum_{i=1}^{n} y_i \ x_i - \sum_{i=1}^{n} m_i x_i \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}} = 0$$

Porém estas equações são não-lineares nos parâmetros e para resolvê-las é preciso recorrer a métodos numéricos interativos, como Newton-Raphson (Gourieroux e Monfort, 1995). Este método é definido expandindo-se a função **U**($\boldsymbol{\beta}$) em torno do ponto inicial $\boldsymbol{\beta}^{(0)}$, tal que

$$U(\boldsymbol{\beta}) \approx U(\boldsymbol{\beta}^{(0)}) + U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}), \qquad (4.1.2.1)$$

sendo que **U**($\boldsymbol{\beta}$) são as derivadas de primeira or$U'(\boldsymbol{\beta})$do logaritmo da função de verossimilhança em relação aos parâmetros do modelo e são as derivadas de ordem 2 do logaritmo da função de verossimilhança.

Se repetirmos o processo (4.1.2.1) chegaremos ao processo iterativo

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [-U'(\boldsymbol{\beta}^{(m)})]^{-1}U'(\boldsymbol{\beta}^{(m)}),$$

sendo que $m = 0, 1, \ldots$

Como a matriz $-U'(\boldsymbol{\beta})$pode não ser positiva definida, e portanto não invertível, ela é substituída pela matriz de informação de Fisher. Assim

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\text{-I}(\boldsymbol{\beta}^{(m)})^{-1}]U'(\boldsymbol{\beta}^{(m)}), \quad m = 0, 1, \ldots \quad (4.1.2.2)$$

A matriz de informação de Fisher, para o modelo logístico com uma variável, tem a seguinte forma:

$$I(\widehat{\beta}) = -\begin{bmatrix} \frac{\partial^2}{\partial\beta_0^2}\ \ln\ \mathrm{L}(\beta_0,\beta_1) & \frac{\partial^2}{\partial\beta_0\beta_1}\ \ln\ \mathrm{L}(\beta_0,\beta_1) \\ \frac{\partial^2}{\partial\beta_0\beta_1}\ \ln\ \mathrm{L}(\beta_0,\beta_1) & \frac{\partial^2}{\partial\beta_1^2}\ \ln\ \mathrm{L}(\beta_0,\beta_1) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \\ \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i^2 \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \end{bmatrix} \quad (4.1.2.3)$$

Após obter as estimativas dos parâmetros do modelo é possível calcular as probabilidades estimadas

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}} \quad (4.1.2.4)$$

# Estimação dos Parâmetros do modelo

Para ajustar um modelo de regressão devemos estimar os parâmetros $\beta_0$e $\beta_1$do modelo. Os estimadores de máxima verossimilhança para os parâmetros $\beta_0$e $\beta_1$são os valores de $\widehat{\beta}_0$e $\widehat{\beta}_1$que maximizam o logaritmo da função de verossimilhança. A função de verossimilhança tem máximo, pois $0<P[Y_i = y_i \mid x_i]<1$,pois a função logaritmo é estritamente crescente.

Para maximizar a função de verossimilhança basta derivarmos em relação aos parâmetros do modelo, da seguinte forma

$$\frac{\partial}{\partial \beta_0} \mathrm{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$\frac{\partial}{\partial \beta_1} \mathrm{L}\ (\beta_0, \beta_1) = \sum_{i=1}^{n} y_i\ x_i - \sum_{i=1}^{n} m_i\, x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Igualando estas derivadas a zero e substituindo os parâmetros ($\beta_0, \beta_1$) pelos estimadores ($\widehat{\beta}_0, \widehat{\beta}_1$),

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}} = 0$$

$$\sum_{i=1}^{n} y_i\ x_i - \sum_{i=1}^{n} m_i x_i \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}} = 0$$

Porém estas equações são não-lineares nos parâmetros e para resolvê-las é preciso recorrer a métodos numéricos interativos, como Newton-Raphson (Gourieroux e Monfort, 1995). Este método é definido expandindo-se a função **U**($\boldsymbol{\beta}$) em torno do ponto inicial $\boldsymbol{\beta}^{(0)}$, tal que

$$U(\boldsymbol{\beta}) \approx U(\boldsymbol{\beta}^{(0)}) + U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}), \qquad (4.1.2.1)$$

sendo que **U**($\boldsymbol{\beta}$) são as derivadas de primeira ordem do logaritmo da função de verossimilhança em relação aos parâmetros do modelo e $U'(\boldsymbol{\beta})$são as derivadas de ordem 2 do logaritmo da função de verossimilhança.

Se repetirmos o processo (4.1.2.1) chegaremos ao processo iterativo

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [-U'(\boldsymbol{\beta}^{(m)})]^{-1} U'(\boldsymbol{\beta}^{(m)}),$$

sendo que $m = 0, 1, \ldots$

Como a matriz $-U'(\boldsymbol{\beta})$pode não ser positiva definida, e portanto não invertível, ela é substituída pela matriz de informação de Fisher. Assim

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathrm{I}(\boldsymbol{\beta}^{(m)})^{-1}] U'(\boldsymbol{\beta}^{(m)}), \quad m = 0, 1, \ldots \quad (4.1.2.2)$$

A matriz de informação de Fisher, para o modelo logístico com uma variável, tem a seguinte forma:

$$I(\widehat{\beta}) = -\begin{bmatrix} \frac{\partial^2}{\partial \beta_0^2} \ln \mathrm{L}(\beta_0, \beta_1) & \frac{\partial^2}{\partial \beta_0 \beta_1} \ln \mathrm{L}(\beta_0, \beta_1) \\ \frac{\partial^2}{\partial \beta_0 \beta_1} \ln \mathrm{L}(\beta_0, \beta_1) & \frac{\partial^2}{\partial \beta_1^2} \ln \mathrm{L}(\beta_0, \beta_1) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \\ \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i^2 \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \end{bmatrix} \quad (4.1.2.3)$$

Após obter as estimativas dos parâmetros do modelo é possível calcular as probabilidades estimadas

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}} \quad (4.1.2.4)$$

# Estimação dos Parâmetros do modelo

Para ajustar um modelo de regressão devemos estimar os parâmetros $\beta_0$e $\beta_1$do modelo. Os estimadores de máxima verossimilhança para os parâmetros $\beta_0$e $\beta_1$são os valores de $\widehat{\beta}_0$e $\widehat{\beta}_1$que maximizam o logaritmo da função de verossimilhança. A função de verossimilhança tem máximo, pois $0<P[Y_i = y_i \mid x_i]<1$, pois a função logaritmo é estritamente crescente.

Para maximizar a função de verossimilhança basta derivarmos em relação aos parâmetros do modelo, da seguinte forma

$$\frac{\partial}{\partial \beta_0} \mathrm{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

$$\frac{\partial}{\partial \beta_1} \mathrm{L}\ (\beta_0, \beta_1) = \sum_{i=1}^{n} y_i\ x_i - \sum_{i=1}^{n} m_i\, x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

Igualando estas derivadas a zero e substituindo os parâmetros $(\beta_0, \beta_1)$ pelos estimadores $(\widehat{\beta}_0, \widehat{\beta}_1)$,

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}} = 0$$

$$\sum_{i=1}^{n} y_i\ x_i - \sum_{i=1}^{n} m_i x_i \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}} = 0$$

Porém estas equações são não-lineares nos parâmetros e para resolvê-las é preciso recorrer a métodos numéricos interativos, como Newton-Raphson (Gourieroux e Monfort, 1995). Este método é definido expandindo-se a função **U**($\boldsymbol{\beta}$) em torno do ponto inicial $\boldsymbol{\beta}^{(0)}$, tal que

$$U(\boldsymbol{\beta}) \approx U(\boldsymbol{\beta}^{(0)}) + U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}), \quad (4.1.2.1)$$

sendo que **U**($\boldsymbol{\beta}$) são as derivadas de primeira ordem $U'(\boldsymbol{\beta})$ do logaritmo da função de verossimilhança em relação aos parâmetros do modelo e são as derivadas de ordem 2 do logaritmo da função de verossimilhança.

Se repetirmos o processo (4.1.2.1) chegaremos ao processo iterativo

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [-U'(\boldsymbol{\beta}^{(m)})]^{-1} U'(\boldsymbol{\beta}^{(m)}),$$

sendo que $m = 0, 1, \ldots$

Como a matriz $-U'(\boldsymbol{\beta})$ pode não ser positiva definida, e portanto não invertível, ela é substituída pela matriz de informação de Fisher. Assim

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathrm{I}(\boldsymbol{\beta}^{(m)})^{-1}] U'(\boldsymbol{\beta}^{(m)}), \quad m = 0, 1, \ldots \quad (4.1.2.2)$$

A matriz de informação de Fisher, para o modelo logístico com uma variável, tem a seguinte forma:

$$I(\widehat{\beta}) = -\begin{bmatrix} \frac{\partial^2}{\partial \beta_0^2} \ln \mathrm{L}(\beta_0, \beta_1) & \frac{\partial^2}{\partial \beta_0 \beta_1} \ln \mathrm{L}(\beta_0, \beta_1) \\ \frac{\partial^2}{\partial \beta_0 \beta_1} \ln \mathrm{L}(\beta_0, \beta_1) & \frac{\partial^2}{\partial \beta_1^2} \ln \mathrm{L}(\beta_0, \beta_1) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \\ \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{n} m_i x_i^2 \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \end{bmatrix} \quad (4.1.2.3)$$

Após obter as estimativas dos parâmetros do modelo é possível calcular as probabilidades estimadas

$$\widehat{\pi}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}} \quad (4.1.2.4)$$

# Interpretação dos parâmetros do modelo

A interpretação dos parâmetros de um modelo de regressão logística é obtida comparando a probabilidade de sucesso com a probabilidade de fracasso, usando a função *odds ratio* - OR (razão de chances). Essa função é obtida a partir da função *odds.*

$$g(x) = \frac{\pi(x)}{[1-\pi(x)]} = \frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}} = \frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}} = e^{\beta_0+\beta_1 x_i}.$$

Assim, ao tomarmos dois valores distintos da variável explicativa, $x_j$e $x_{j+1}$, obtemos

$$OR = \frac{g(x_{j+1})}{g(x_j)} = \frac{e^{\beta_0+\beta_1\ x_{j+1}}}{e^{\beta_0+\beta_1\ x_j}}. \quad (4.1.2.1.1)$$

Temos ainda que:

$$\ln(OR) = \ln\left[\frac{g(x_{j+1})}{g(x_{j+1})}\right] = \ln\left[g(x_{j+1})\right] - \ln\left[g(x_j)\right]$$
$$= \beta_0 + \beta_1 x_{j+1} - \beta_0 - \beta_1 x_j = \beta_1(x_{j+1} - x_j).$$

Fazendo $x_{j+1} - x_j = 1$unidade, então

$$\ln(OR) = \ln(e^{\beta_1}) = \beta_1.$$

Assim, temos o quão provável o resultado ocorrerá entre os indivíduos $x_{j+1}$em relação aos indivíduos $x_j$, fazendo, portanto, algumas análises:

$$\beta_j{>}0 \ \Rightarrow \ OR{>}1 \Rightarrow \ \pi_i(x_{j+1}) \ > \ \pi_i(x_j)$$
$$\beta_j{<}0 \ \Rightarrow \ OR{<}1 \Rightarrow \ \pi_i(x_{j+1}) \ < \ \pi_i(x_j)$$

Veja "variáveis independentes categóricas" quando a variável explicativa é categórica.

# Estimativa dos desvios padrão

No modelo de regressão logístico o desvio padrão dos estimadores é obtido a partir da matriz de informação de Fisher. Podemos ainda obter a matriz de informação de Fisher $I(\widehat{\beta})$para o modelo logístico a partir dos dados, da seguinte forma, $I(\widehat{\beta}) = X^{'} V X$,sendo que

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \ V = \text{diag}[m_1\hat{\pi}_1(1-\hat{\pi}_1), \ldots, m_n\hat{\pi}_n(1-\hat{\pi}_n)],$$

$m_i$é o número de repetições para cada elemento da amostra, $i = 1, \ldots, n.$

As variâncias e covariâncias dos estimadores $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ são obtidos, invertendo a matriz de informação de Fisher, isto é, calculando $\widehat{\Sigma} = I^{-1}(\widehat{\beta})$.

O $j$-ésimo elemento da diagonal principal da matriz $\widehat{\Sigma}$ é a variância do estimador $\widehat{\beta}_j$, denominada $\widehat{\sigma}^2(\widehat{\beta}_j)$. Os demais elementos da matriz $\widehat{\Sigma}$ são as covariâncias entre $(\widehat{\beta}_j; \widehat{\beta}_u)$, $j \neq u$.

Desta forma o desvio padrão é definido como:

$$\widehat{DP}(\widehat{\beta}_j) = \sqrt{\widehat{\sigma}^2(\widehat{\beta}_j)}.$$

# Inferência em um modelo logístico simples

Após estimar os coeficientes, temos interesse em assegurar a significância das variáveis no modelo. Isto geralmente envolve formulação e teste de uma hipótese estatística para determinar se a variável independente no modelo é significativamente relacionada com a variável resposta. Para isso, temos os testes de hipóteses. Os testes de hipóteses mais utilizados são os testes da Razão da Verossimilhança, Wald e Escore. A seguir, temos a abordagem de cada um deles.

# Teste de Wald

O teste de Wald é obtido por comparação entre a estimativa de máxima verossimilhança do parâmetro ($\widehat{\beta_1}$) e a estimativa de seu erro padrão. A razão resultante, sob a hipótese $\mathrm{H}_0 : \beta_1 = 0$, tem distribuição normal padrão.

A estatística do teste Wald para a regressão logística é $W_j = \dfrac{\widehat{\beta}_1}{\widehat{DP}(\widehat{\beta}_1)}$.

O p-valor é definido como $P(|Z| > |W_j|)$, sendo que Z denota a variável aleatória da distribuição normal padrão.

Hauck e Donner (1977) examinaram o desempenho do teste de Wald e discobriram que ele se comporta de maneira estranha, em determinadas situações; frequentemente não rejeitando a hipótese nula quando o coeficiente é significativo. Eles recomendam a utilização do teste da razão de verossimilhança para testar se realmente o coeficiente não é significativo quando o teste de Wald não rejeita a hipótese nula.

# Teste da Razão de Verossimilhança

Na regressão linear o interesse está no valor da SQR. Um valor alto da SQR sugere que a variável independente é importante, caso contrário, a variável independente não é útil na predição da variável resposta.

Na regressão logística a ideia é a mesma: comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. A comparação dos observados com os valores preditos é baseado no log da verossimilhança. Para entender melhor essa comparação, é útil pensar em um valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto observações.

A comparação dos observados com os valores preditos usando a função de verossimilhança é baseada na seguinte expressão:

$$D = -2ln\left[\frac{(verossimilhanca\ do\ modelo\ ajustado)}{(verossimilhanca\ do\ modelo\ saturado)}\right].$$

Com o propósito de assegurar a significância de uma variável independente, comparamos o valor da D com e sem a variável na equação. A mudança em D devido a inclusão da variável no modelo é obtida da seguinte maneira:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável).$$

Podemos então escrever a estatística G como:

$$G = -2ln\left[\frac{(verossimilhanca\ sem\ a\ variável)}{(verossimilhanca\ com\ a\ variável)}\right].$$

ou ainda:

$$G = -2ln(L_s) + 2ln(L_c),$$

em que $L_s$é a verossimilhança do modelo sem a covariável e $L_c$é a verossimilhança do modelo com a covariável.

Queremos testar:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Sob a hipótese nula, a estatística G tem distribuição chi-quadrado com 1 grau de liberdade.

**Exemplo 4.1.3.2.1**

Vamos considerar o Exemplo 4.1.2.1 para verificar se a variável "horas de treinamento" é significativa para explicar o erro na montagem, através do teste da razão de verossimilhança (TRV).

O valor do log da verossimilhança do modelo apenas com o intercepto ( $L_s$) é -1064,183 e do modelo com a variável ( $L_c$) é -1035,089.

Assim, o valor da estatística teste é:

$$G=-2(-1064,183)-(-2(-1035,089))= 58,188.$$

O p-valor $P(\chi_1^2 > G = 58,188) < 0,0001$.

Rejeitamos a hipótese nula. Assim, pelo TRV, temos que a variável horas de treinamento é significativa para o modelo.

# Teste Score

A estatística teste para o Teste Score é:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{(\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2)^{1/2}},$$

em que $\bar{y} = \hat{\pi}$(proporção de sucessos na amostra).

No Teste Score também temos o interesse em testar:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

O p-valor é definido como $P(|Z| > |ST|)$, sendo que Z denota a variável aleatória da distribuição normal padrão.

## Intervalo de Confiança para os parâmetros

A base da construção das estimativas do intervalo de confiança para os parâmetros é a mesma teoria estatística que usamos para os testes de significância do modelo. Em particular, um intervalo de confiança para a inclinação e intercepto são baseados em seus respectivos testes de Wald. O intervalo de confiança de $100(1-\alpha)$% para o parâmetro $\beta_1$é:

$$IC(\beta_1, 1-\alpha) = [\hat{\beta}_1 - z_{1-\alpha/2}DP(\hat{\beta}_1); \quad \hat{\beta}_1 + z_{1-\alpha/2}DP(\hat{\beta}_1)].$$

E para o intercepto:

$$IC(\beta_0, 1-\alpha) = [\hat{\beta}_0 - z_{1-\alpha/2} DP(\hat{\beta}_0); \quad \hat{\beta}_0 + z_{1-\alpha/2} DP(\hat{\beta}_0)],$$

em que $z_{1-\alpha/2}$é o ponto da normal padrão correspondente a $100(1-\alpha/2)\%$.

## Intervalo de Confiança para Logito

A logito é a parte linear do modelo de regressão logística. O estimador para logito é:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

O estimador da variância do estimador da logito requer a obtenção da variância da soma. No caso é:

$$\hat{Var}[\hat{g}(x)] = \hat{Var}(\hat{\beta}_0) + x^2 \hat{Var}(\hat{\beta}_1) + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1). \qquad (4.1.4.2.1)$$

O intervalo de confiança para a logito é:

$$IC(g(x), 1-\alpha) = [\hat{g}(x) - z_{1-\alpha/2} DP(\hat{g}(x)); \quad \hat{g}(x) + z_{1-\alpha/2} DP(\hat{g}(x))],$$

em que $DP(\hat{g}(x))$é a raiz quadrada de 4.1.4.2.1 e $z_{1-\alpha/2}$é o ponto da normal padrão.

## Intervalo de Confiança para os valores ajustados

O estimador do logito e seu intervalo de confiança fornece o estimador dos valores ajustados. O intervalo de confiança dos valores ajustados é dado por:

$$IC(\pi, 1-\alpha) = \left[ \frac{e^{\hat{g}(x) - z_{1-\alpha/2} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) - z_{1-\alpha/2} DP(\hat{g}(x))}}; \quad \frac{e^{\hat{g}(x) + z_{1-\alpha/2} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) + z_{1-\alpha/2} DP(\hat{g}(x))}} \right]. \qquad (4.1.4.2.2)$$

## Intervalo de Confiança para a Odds Ratio

Sejam os limites do intervalo de confiança para $\beta_1$:

$$\beta_I = \hat{\beta}_1 - z_{1-\alpha/2} DP(\hat{\beta}_1) \text{e } \beta_S = \hat{\beta}_1 + z_{1-\alpha/2} DP(\hat{\beta}_1).$$

O intervalo de confiança para a Odds Ratio é:

$$IC(Odds\ Ratio, 1-\alpha) = [e^{\beta_I}; \quad e^{\beta_S}]. \qquad (4.1.4.2.3)$$

## Bibliografia:

Neter ,J.; Wasserman, William; Kutner, M.H., Applied linear statistical models;
Draper,N.R.; Smith,H., Applied Regression Analysis.
Montgomery and Peck, Introduction to Linear Regression Analysis;
Seber, G.A.F., Linear Regression Analysis.
Myers and Montgomery, Generalized Linear Models.