2
Regressão Logística

Este capítulo foi dedicado à Regressão Logística, pois se trata de um método base para o entendimento do modelo proposto. Porém antes da apresentação dos conceitos daquela, faz-se necessário uma introdução aos Modelos Lineares Generalizados para um bom entendimento deste capítulo.

2.1
Revisão de Modelos Lineares Generalizados (MLG)

Os Modelos Lineares Generalizados foram propostos em (24), com a finalidade de permitir a modelagem, não apenas utilizando os modelos lineares clássicos, os quais assumem, dentre outras coisas, que a variável dependente (Y_i) segue uma distribuição Normal (ou Gaussiana). Assim os MLG´s admitem que Y_i possa seguir outras distribuições pertencentes à família exponencial. No mesmo trabalho é introduzido o conceito de *Deviance*, que é uma medida utilizada para comparar os modelos.

2.1.1
Componentes de um MLG

Assim como nos modelos lineares clássicos, o objetivo dos modelos lineares generalizados é descrever a relação entre y_i, que são as realizações da variável aleatória Y_i, e outras variáveis chamadas regressores (também conhecidas como variáveis explicativas, preditores ou covariáveis). A realização de uma variável explicativa, X_i, será representada por x_i e, é descrita por meio de um conjunto de parâmetros representado por $\beta_1, ..., \beta_p$, que ponderam a combinação linear dos valores de X_i, bem como a um erro aleatório (ou perturbação) ϵ_i, consegue descrever o comportamento da variável dependente através da seguinte expressão

$$y_i = \sum_{j=1}^{p} x_{ji}\beta_j + \epsilon_i, \quad i = 1, ..., n \tag{2-1}$$

ou ainda $y_i = \mathbb{E}(Y_i|\mathbf{x}_i) + \epsilon_i$, onde a principal suposição sob o erro no caso de modelos lineares é que o mesmo segue a distribuição Normal com média zero e variância constante e, por conseqüência, a distribuição da variável dependente condicional as variáveis explicativas será Normal com média $\mathbb{E}(Y_i|\mathbf{x}_i)$ e variância constante.

Descrevendo tal modelo na forma matricial teremos

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_1 \begin{pmatrix} x_{11}=1 \\ x_{12}=1 \\ \vdots \\ x_{1n}=1 \end{pmatrix} + \beta_2 \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} + ... + \beta_p \begin{pmatrix} x_{p1} \\ x_{p2} \\ \vdots \\ x_{pn} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{21} & ... & x_{p1} \\ 1 & x_{22} & ... & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & ... & x_{pn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

que pode ser expressa simplesmente por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

em que $\mathbf{y}$ e $\boldsymbol{\epsilon}$ são vetores $n \times 1$, $\mathbf{X}$ uma matriz $n \times p$ e $\boldsymbol{\beta}$ um vetor $p \times 1$.

Um elemento do vetor $\mathbf{y}$ é dado pela expressão

$$y_i = \mathbf{x}_i\beta + \epsilon_i, \quad i = 1, ..., n$$

que corresponde a forma matricial de (2-1), onde $\mathbf{x}_i = (1, x_{1i}, x_{2i}, ..., x_{pi})'$.

A fim de identificar os componentes de um MLG iremos assumir, neste primeiro momento, que $Y_1, ..., Y_n$ são variáveis independentes e normalmente distribuídas, as quais, também por suposição, são independentes, com distribuição, não necessariamente, mas usualmente, Normal e têm variância constante (σ^2_ϵ), um *Ruído Branco*.

Tal variância também é um parâmetro desconhecido e, desta maneira, além dos p parâmetros representados pelos β´s, teremos σ^2_ϵ totalizando $p+1$ parâmetros.

Porém diferentemente do caso linear Gaussiano aqui $y_i = \pi(\mathbf{x}_i) + \epsilon_i$, onde ϵ_i assume apenas dois valores dependendo daquele assumido por y_i. Se $y_i = 1$ então $\epsilon_i = 1 - \pi(\mathbf{x}_i)$ com probabilidade $\pi(\mathbf{x}_i)$. Caso $y_i = 0$, $\epsilon_i = -\pi(\mathbf{x}_i)$ com

probabilidade $1 - \pi(\mathbf{x}_i)$.

Com isso reescrevemos o modelo como

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mu_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n \tag{2-2}$$

Expressão que, na forma matricial será tal que

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

Após a estimação dos β´s podemos encontrar os valores de $\hat{\mu_1}, ..., \hat{\mu_n}$ escrevendo assim o modelo estimado da seguinte maneira

$$\hat{\mu_i} = \sum_{j=1}^{p} x_{ji}\hat{\beta}_j, \quad i = 1, ..., n$$

O modelo pode ser dividido em três partes:

– Componente aleatória: componente da variável aleatória Y_i, $i = 1, ...n$, admitindo que a mesma tenha distribuição pertencente à *família exponencial*;

– Preditor linear: representado por η e denominado por

$$\eta_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n \tag{2-3}$$

– Função de ligação: função $g(\cdot)$ monotônica diferenciável que liga o preditor linear ao valor esperado de Y_i, ou seja, $g(\mu_i) = \eta_i$, $i = 1, ...n$

2.1.2 Ligações Canônicas

No modelo linear clássico, a função que ligava o preditor linear ao valor esperado de Y_i era a identidade, pois sendo aquela uma variável aleatória com distribuição normal, a média e o preditor linear são idênticos. Em se tratando de variáveis dependentes que tenham uma distribuição pertencente à família exponencial, porém diferente da Normal, temos disponíveis outras funções de ligação clássicas como, por exemplo, para o caso de uma distribuição binomial, em que $\mu \in (0, 1)$

1. Logito: $g(\mu_i) = \log\left(\frac{\mu}{1-\mu}\right)$
2. Probito: $g(\mu_i) = \Phi^{-1}(\mu)$
   onde $\Phi(\cdot)$ é uma função de distribuição acumulada Normal padrão
3. Complemento Log-log: $g(\mu_i) = \log[-\log(1-\mu)]$

Temos também o caso clássico para contagens, que seguem uma distribuição de Poisson, cuja função de ligação é a logarítmica, $g(\mu_i) = \log(\mu)$. Além das distribuições citadas anteriormente, também fazem parte da família exponencial a distribuição gamma e a binomial negativa.

Utilizaremos algumas dessas funções na próxima seção, onde abordaremos a *Regressão Logística*, a qual em seu caso particular mais simples tem uma variável dependente dicotômica e possui uma distribuição binomial.

2.2 Dados binários (Regressão Logística)

Como mencionado, o caso mais simples de uma Regressão Logística ocorre quando a variável aleatória Y_i assume apenas dois valores, 0 ou 1. O primeiro é a ocorrência de um determinado evento *fracasso* e o segundo *sucesso*. Para isso, teremos que definir a probabilidade de interesse, ou *probabilidade de sucesso*, $\mathbb{P}(Y_i = 1) = \pi_i$ e a *probabilidade de fracasso* $\mathbb{P}(Y_i = 0) = 1 - \pi_i$.

Para investigar a relação entre a probabilidade de sucesso π_i e o vetor de covariáveis $\mathbf{X} = (X_1, X_2, ..., X_p)'$ escrevemos o modelo

$$\mathbb{E}(Y_i|\mathbf{x}_i) = \pi_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

Entretanto, como $0 < \pi < 1$, dificilmente a igualdade acima é verdadeira. Desta maneira usaremos uma transformação $g(\pi)$ para, corretamente, poder escrever o modelo. Nosso próximo passo é escolher a transformação, que será chamada função de ligação e assim formalizar a relação como segue

$$g(\pi_i) = \eta_i$$

$$g(\pi_i) = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

Optamos pela logito (ou função logística) por ser a ligação canônica.

$$\log\left[\frac{\mathbb{P}(Y_i=1|\mathbf{x}_i)}{\mathbb{P}(Y_i=0|\mathbf{x}_i)}\right]=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{j=1}^{p}x_{ji}\beta_j,\quad i=1,...,n \tag{2-4}$$

A probabilidade π_i pode ser escrita em função do preditor linear conforme:

$$\frac{\pi_i}{1-\pi_i}=e^{\sum_{j=1}^{p}x_{ji}\beta_j},\quad i=1,...,n \tag{2-5}$$

$$\pi_i=\frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{2-6}$$

2.2.1 Especificação do modelo de Regressão Logística

Antes de descrever os métodos de especificação do modelo de Regressão Logística, será deduzida a expressão da função de *Máxima Verossimilhança* e introduzido o conceito de *Deviance* (ou *Função Desvio*).

Se olharmos apenas para o caso em que π é um escalar temos a função de *Máxima Verossimilhança* para $y_1,...,y_n$, realizações da distribuição *Bernoulli* (π), dada por:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi^{y_i}(1-\pi)^{1-y_i}=\pi^{\sum_{i=1}^{n}y_i}(1-\pi)^{n-\sum_{i=1}^{n}y_i},\quad 0\leq\pi\leq 1.$$

A expressão do $\log L(\boldsymbol{\beta})$, também chamada *log-verossimilhança* é

$$l(\boldsymbol{\beta})=\sum_{i=1}^{n}y_i\log(\pi)+\left(n-\sum_{i=1}^{n}y_i\right)\log(1-\pi),\ 0\leq\pi\leq 1.$$

Na Regressão Logística π é função de outras covariáveis, $x_1,...,x_n$, assim temos:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi(\mathbf{x}_i)^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}=\pi(\mathbf{x}_i)^{\sum_{i=1}^{n}y_i}[1-\pi(\mathbf{x}_i)]^{n-\sum_{i=1}^{n}y_i} \tag{2-7}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i)] + \left(n - \sum_{i=1}^{n} y_i\right) \log[1 - \pi(\mathbf{x}_i))] \tag{2-8}$$

$$\begin{aligned} l(\boldsymbol{\beta}) = & \quad \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i)] + \left(n - \sum_{i=1}^{n} y_i\right) \log[1 - \pi(\mathbf{x}_i)] \\ = & \quad \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i)] + n \log[1 - \pi(\mathbf{x}_i)] - \sum_{i=1}^{n} y_i \log[1 - \pi(\mathbf{x}_i)] \end{aligned}$$

$$l(\boldsymbol{\beta}) = \left\{ \sum_{i=1}^{n} y_i \log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] + n \log[1 - \pi(\mathbf{x}_i)] \right\}. \tag{2-9}$$

Podemos notar em (2-9) o aparecimento da função logito, $\log \left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} \right]$, que é a função de ligação entre o preditor linear e o valor esperado de Y_i.

Da qual sabemos que

$$\log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] = \beta' \mathbf{x}_i$$

e de forma análoga

$$\pi(\mathbf{x}_i) = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}$$

logo

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

então

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \beta' \mathbf{x}_i - \log(1 + e^{\beta' \mathbf{x}_i}) \right]. \tag{2-10}$$

A *função Desvio*, também conhecida como *Deviance* (ver (19)) é basicamente a distância entre a log-verossimilhança do modelo contendo um parâmetro para cada uma das n observações (modelo saturado) e o modelo ajustado para p parâmetros, medindo assim a qualidade do ajuste. Se o seu valor for pequeno, indica que o ajuste do modelo com p parâmetros é próximo daquele com n e, segundo o princípio da parcimônia, escolhe-se o primeiro.

No caso Binomial a *Deviance* toma a forma

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - \hat{\pi}_i}{m_i - y_i}\right)\right], \qquad (2\text{-}11)$$

onde $0 \leq y_i \leq m_i$ e, no caso, $m_i = 1, \forall t$.

Segundo o apresentado em (15), quando tal expressão é computada para regressão linear simples é equivalente a soma dos quadrados dos resíduos. Porém, se tratando de uma Regressão Logística em que $y = 0$ ou 1, tal medida não pode ser utilizada como sinalizadora de um bom ajuste. Desenvolvendo a equação acima temos

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right)\right] \\ &= -2\sum_{i=1}^{n}\left[y_i \log(y_i) + (1 - y_i)\log(1 - y_i) - y_i \log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) - \log(1 - \hat{\pi}_i)\right] \end{aligned}$$

como y assume apenas os valores 0 e 1 temos que

$$y_i \log(y_i) = (1 - y_i)\log(1 - y_i) = 0.$$

Além disso, $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}'\mathbf{x}_i$, e desta maneira

$$\begin{aligned} D &= -2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - 2\sum_{i=1}^{n}\log(1 - \hat{\pi}_i) \\ &= -2\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\pi}} - 2\sum_{i=1}^{n}\log(1 - \hat{\pi}_i). \end{aligned}$$

A *Deviance* não pode ser utilizada como medida da qualidade do ajuste quando m_i é um número pequeno, geralmente para $m_i \leq 5$. Neste caso D passa a ter uma distribuição condicional degenerada, dado os valores de $\hat{\boldsymbol{\beta}}$ (ver (19)). A frente serão apresentadas as medidas da qualidade do ajuste para regressão logística. A função desvio poderá ser usada em testes de hipótese de nulidade dos parâmetros através da estatística F.

A seleção dos regressores pode ser feita utilizando-se uma metodologia proposta em (15), a qual é uma variante do método *Stepwise*, ou através dos critérios de informação, AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*).

Tais critérios penalizam a função de log-verossimilhança pela inclusão de

novas variáveis, respeitando o princípio da parcimônia. Escolhe-se o modelo que minimiza o AIC ou BIC, que estão descritos em (2-12) e (2-13)

$$AIC = -2\frac{l(\boldsymbol{\beta})}{n} + 2\frac{p}{n} \tag{2-12}$$

$$BIC = -2\frac{l(\boldsymbol{\beta})}{n} + p\frac{\log(n)}{n} \tag{2-13}$$

onde p é o número de parâmetros, n a quantidade de observações e $l(\boldsymbol{\beta})$ é o log da função de verossimilhança.

Já a metodologia proposta em (15) considerando um modelo com p variáveis explicativas segue alguns passos mostrados a seguir

1. Ajusta-se o modelo nulo somente com intercepto e $(p-1)$ modelos cada um contendo o intercepto mais uma das variáveis explicativas $\mathbf{x}_i$. Confronta-se cada um desses $(p-1)$ modelos, com o modelo nulo através da estatística de *Razão de Verossimilhança* dada por

$$\xi_{RV}^{(0)} = 2\ln\left[\frac{L(\hat{\boldsymbol{\beta}};\mathbf{y})}{L(\boldsymbol{\beta}^0;\mathbf{y})}\right] = 2\left[l(\hat{\boldsymbol{\beta}};\mathbf{y}) - l(\boldsymbol{\beta}^0;\mathbf{y})\right] \xrightarrow{a} \chi^2_{(p)}, \tag{2-14}$$

onde $l(\hat{\boldsymbol{\beta}};\mathbf{y})$ é a log verossimilhança do modelo com intercepto mais uma das variáveis explicativas e $l(\boldsymbol{\beta}^0;\mathbf{y})$ é a log verossimilhança do modelo apenas com intercepto. Tendo conhecimento do parâmetro de dispersão, no contexto de MLG denotado por ϕ, podemos expressar a razão de verossimilhança através da diferença entre as funções desvio e assim

$$\xi_{RV}^{(0)} = \phi\left[D(\mathbf{y};\hat{\boldsymbol{\mu}}^0) - D(\mathbf{y};\hat{\boldsymbol{\mu}})\right] \xrightarrow{a} \chi^2_{(p)}, \tag{2-15}$$

em que $\hat{\boldsymbol{\mu}}^0 = \mathbf{g}^{-1}(\hat{\boldsymbol{\eta}}^0)$, $\hat{\boldsymbol{\eta}}^0 = \mathbf{X}\boldsymbol{\beta}^0$. De maneira análoga, a estatística F, a seguir, pode ser utilizada como alternativa de teste das hipóteses, apresentando ainda a vantagem de não depender do parâmetro de dispersão, ϕ,

$$F = \frac{\left[D(\mathbf{y};\hat{\boldsymbol{\mu}}^0) - D(\mathbf{y};\hat{\boldsymbol{\mu}})\right]/q}{D(\mathbf{y};\hat{\boldsymbol{\mu}})/(N-p)} \xrightarrow{a} F_{q,(N-p)}. \tag{2-16}$$

Escolhe-se o modelo com o menor *p-valor*, ou seja, sendo $p_{e_i}^{(0)} = P\left(\chi^2_{(\nu)} > \xi_{RV}^{(0)}\right)$ o *p-valor* associado a $\mathbf{x}_i$, $\forall t$, onde $\nu = 1$ se $\mathbf{x}_i$ é contínua e $\nu = k-1$ se $\mathbf{x}_i$ é categórica com $k-1$ níveis. Assim, o modelo escolhido

é aquele que apresenta $p_{e_1} = \min\left[p_{e_i}^{(0)}\right]$, e a variável escolhida é denominada $\mathbf{x}_{e_1}$. Além disso, é determinada uma probabilidade de entrada, P_E, a partir da qual verifica-se a significância da variável escolhida, onde a seqüência da modelagem se dá caso $p_{e_1} < P_E$, quando se prossegue para o passo seguinte, e caso o contrário aconteça, o modelo especificado é escolhido. Geralmente $0,15 < P_E < 0,25$ (ver (15)).

2. Ajustam-se agora $(p-2)$ modelos incluindo mais uma variável explicativa, das que restaram em relação ao modelo que foi selecionado no passo anterior. Cada um desses modelos é avaliado em relação ao modelo do passo 1 como se segue

$$\xi_{RV_i}^{(1)} = 2\left[l_{e_1e_i}(\hat{\boldsymbol{\beta}};\mathbf{y}) - l_{e_1}(\hat{\boldsymbol{\beta}};\mathbf{y})\right], \quad i = 2,3,...,p.$$

A escolha da variável $\mathbf{x}_{e_2}$ se dá para $p_{e_2} = \min\left[p_{e_i}^{(1)}\right]$, onde $p_{e_i}^{(1)} = P\left(\chi^2_{(\nu)} > \xi_{RV_i}^{(1)}\right)$. Se $p_{e_2} < P_E$ siga para o passo 3, caso contrário pare.

3. Com o modelo formado pelo intercepto mais $\mathbf{x}_{e_1}$ e $\mathbf{x}_{e_2}$ deve-se testar se ao incluir esta última, a variável $\mathbf{x}_{e_1}$ deixa de ser significativa. Desta forma, calcula-se

$$\xi_{RV_i}^{(2)} = 2\left[l_{e_1e_2}(\hat{\boldsymbol{\beta}};\mathbf{y}) - l_{e_i}(\hat{\boldsymbol{\beta}};\mathbf{y})\right], \quad i = 1,2$$

$$p_{e_i}^{(2)} = P\left(\chi^2_{(\nu)} > \xi_{RV_i}^{(2)}\right).$$

Nesta situação, observa-se a variável que possui o maior *p-valor*, a fim de testar se ela irá ser retirada ou não do modelo. A variável escolhida é denominada $\mathbf{x}_{r_2}$. Além disso bem como a probabilidade de entrada é determinada uma probabilidade da variável ser retirada, P_R, onde $0,15 < P_R < 0,25$. Se $p_{r_2} > P_R$ então a variável é retirada, caso contrário a variável permanece e deve-se verificar a entrada de outra variável no modelo escolhido. Ainda neste passo ajusta-se $(p-3)$ modelos (supondo que $\mathbf{x}_{e_1}$ e $\mathbf{x}_{e_2}$ tenham permanecido) e calcula-se

$$\xi_{RV_i}^{(2^*)} = 2\left[l_{e_1e_2e_i}(\hat{\boldsymbol{\beta}};\mathbf{y}) - l_{e_1e_2}(\hat{\boldsymbol{\beta}};\mathbf{y})\right], \quad i = 3,4,...,p,$$

$$p_{e_i}^{(2^*)} = P\left(\chi^2_{(\nu)} > \xi_{RV_i}^{(2^*)}\right),$$

$$p_{e_3} = \min\left[p_{e_i}^{(2^*)}\right].$$

Se $p_{e_3^*} < P_E$ siga para o próximo passo, caso contrário pare.

Os próximos passos são semelhantes ao passo 3 até que sejam esgotadas as possibilidades de entrada e retirada de variáveis. Com o modelo, após a escolha das variáveis principais (ou efeitos principais), é verificada a significância de cada coeficiente (β) individualmente, através de testes de *Wald*[1], onde $H_0 : \beta = 0$, e aqueles que não forem estatisticamente significativos, ou seja, quando a hipótese nula não é rejeitada, determina-se a retirada de sua respectiva covariável do modelo.

Feito isto, os passos seguintes consistem na inclusão das interações de primeira ordem seguindo-se o mesmo procedimento de entrada e retirada feito anteriormente sem se esquecer de não eliminar os efeitos principais. Se for necessário incluir as interações de segunda e terceira ordem segue-se o mesmo padrão.

2.2.2
Estimação do modelo de Regressão Logística

A estimação dos parâmetros de uma Regressão Logística é feita por Máxima Verossimilhança, utilizando-se o método iterativo de *Newton-Raphson.*

Os cálculos das derivadas de $l(\boldsymbol{\beta})$ (log da verossimilhança) bem como o algoritmo do método são apresentados a seguir.

Derivando (2-16) em relação ao parâmetro β (*Função Escore*) teremos:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial \beta} &= \sum_{i=1}^{n} \mathbf{x}_i \left[y_i - \frac{e^{\beta' \mathbf{x}_i}}{(1+e^{\beta' \mathbf{x}_i})}\right] \\ &= \sum_{i=1}^{n} \mathbf{x}_i \left[y_i - \pi(\mathbf{x}_i)\right].\end{aligned}$$

Na forma matricial,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\pi}). \tag{2-17}$$

O algoritmo também requer a Hessiana, obtida por

[1] A estatística de teste é dada por: $W = \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}$

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{n}\left[\frac{(1+e^{\beta'\mathbf{x}_i})e^{\beta'\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i' - e^{\beta'\mathbf{x}_i}\mathbf{x}_i' e^{\beta'\mathbf{x}_i}\mathbf{x}_i'}{(1+e^{\beta'\mathbf{x_i}})^2}\right] \\
&= -\sum_{i=1}^{n}\left[\frac{(1+e^{\beta'\mathbf{x}_i})e^{\beta'\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i' - (e^{\beta'\mathbf{x}_i})^2\mathbf{x}_i\mathbf{x}_i'}{(1+e^{\beta'\mathbf{x_i}})^2}\right] \\
&= -\sum_{i=1}^{n}\left[\frac{e^{\beta'\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i'}{1+e^{\beta'\mathbf{x}_i}} - \left(\frac{e^{\beta'\mathbf{x}_i}}{1+e^{\beta'\mathbf{x}_i}}\right)^2\mathbf{x}_i\mathbf{x}_i'\right] \\
&= -\sum_{i=1}^{n}\left[\pi(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i' - \pi(\mathbf{x}_i)^2\mathbf{x}_i\mathbf{x}_i'\right] \\
&= -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)].
\end{aligned}$$

E matricialmente representada por

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta\partial\beta'} = -\mathbf{X}'\mathbf{W}\mathbf{X}. \tag{2-18}$$

A *Matriz de Informação de Fisher* para β é conhecida pela expressão $I(\beta) = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta\partial\beta'}\right]$.

Tendo esses elementos falta apenas determinar um valor inicial para β, chamaremos de $\beta^{(m)}$, e assim dar início ao algoritmo de Newton-Raphson, que a fim de obter a estimativa de Máxima Verossimilhança do parâmetro em questão, β, expande-se a Função Escore, $U(\beta)$ em torno do valor inicial, um $\beta^{(m)}$ qualquer, de maneira que

$$U(\beta) \cong U(\beta^{(m)}) + \frac{\partial}{\partial\beta'}U(\beta^{(m)})(\beta^{(m+1)} - \beta^{(m)}), \quad m = 0, 1, ...$$

Iterativamente obtem-se

$$\beta^{(m+1)} = \beta^{(m)} + \left[-\frac{\partial}{\partial\beta'}U(\beta^{(m)})\right]^{-1}U(\beta^{(m)}), \quad m = 0, 1, ...$$

A matriz $-\frac{\partial}{\partial\beta'}U(\beta^{(m)})$ deve ser positiva definida, e como não se pode garantir tal hipótese, substitui-se a mesma pelo seu valor esperado $E\left[-\frac{\partial}{\partial\beta'}U(\beta^{(m)})\right]^{-1} = I^{-1}(\beta^{(m)})$ e assim, continuando o processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + I^{-1}(\beta^{(m)})U(\beta^{(m)}), \quad m = 0, 1, ...$$

$$\beta^{(m+1)} = \beta^{(m)} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right)^{-1} \frac{\partial l(\beta)}{\partial\beta}, \quad m = 0, 1, ... \tag{2-19}$$

Trabalhando mais uma vez com a forma matricial em que $\mathbf{y}$ é o vetor $(n \times 1)$ de valores y_i, $\mathbf{X}$ a matriz $(n \times (p+1))$ de valores x_i, $\boldsymbol{\pi}$ o vetor $(n \times 1)$ das probabilidades ajustadas com o *i-ésimo* elemento igual a $\pi(x_i; \beta^{(m)})$ e $\mathbf{W}$ a matriz diagonal $(n \times n)$ de pesos com o *i-ésimo* elemento igual a $\pi(x_i; \beta^{(m)})(1 - \pi(x_i; \beta^{(m)}))$ tem-se

$$\begin{aligned} \beta^{(m+1)} = & \quad \beta^{(m)} + (\mathbf{X'WX})^{-1}\mathbf{X'}(\mathbf{y} - \boldsymbol{\pi}) \\ = & \quad (\mathbf{X'WX})^{-1}\mathbf{X'W}[\mathbf{X}\beta^{(m)} + \mathbf{W^{-1}}(\mathbf{y} - \boldsymbol{\pi})] \end{aligned}$$

$$\beta^{(m+1)} = (\mathbf{X'WX})^{-1}\mathbf{X'Wz} \tag{2-20}$$

onde $\mathbf{z} = \mathbf{X}\beta^{(m)} + \mathbf{W^{-1}}(\mathbf{y} - \boldsymbol{\pi})$.

Considerando $\mathbf{z}$ como se fosse o vetor de observações de uma variável dependente qualquer, também chamada de variável dependente ajustada, a equação $(\mathbf{X'WX})^{-1}\mathbf{X'Wz}$ seria o cálculo do estimador de *Mínimos Quadrados Ponderados* (ver (33)). A implementação do método pode ser feita considerando $z_i = \beta'\mathbf{x}_i + \frac{y_i - \pi(\mathbf{x}_i)}{\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]}$, $w_i = \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]$ e, conseqüentemente, $z_i = \beta'\mathbf{x}_i + \frac{y_i - \pi(\mathbf{x}_i)}{w_i}$, usando regressão linear ponderada para explicar z_i por $\mathbf{x}_i$. Este procedimento é conhecido como *Iteratively Reweighted Least Squares (IRLS)* e matricialmente é descrito por

$$\mathbf{z} = \mathbf{X}\beta^{(m)} + \mathbf{W^{-1}}(\mathbf{y} - \boldsymbol{\pi})$$

$$\beta^{(m+1)} \longleftarrow \arg\min_{\beta}(\mathbf{z} - \mathbf{X}\beta)^{-1}\mathbf{W}(\mathbf{z} - \mathbf{X}\beta), \quad m = 0, 1, ...$$

O valor inicial usual é $\beta = 0$.

Fazemos a regressão de z_0 em relação as covariáveis $x_1, ..., x_p$ com peso W_0 e assim achamos a estimativa de $\hat{\boldsymbol{\beta}}^{(1)}$, com isto alimento o modelo e faço o cálculo para z_1, acho $\hat{\boldsymbol{\beta}}^{(2)}$ e assim sucessivamente. O procedimento converge em um número finito de passos, podendo falhar apenas se uma ou mais componentes de $\hat{\boldsymbol{\beta}}$ forem infinitas, o que implica em algumas probabilidades ajustadas serem iguais a zero ou um. Caso isso ocorra, poderemos identificar as convergências anormais através da análise de *Deviance*, dado que as probabil-

idades ajustadas serão alteradas; outra forma é verificar a mudança no $\hat{\boldsymbol{\beta}}$ ou no preditor linear, $\hat{\boldsymbol{\eta}}$.

2.2.3 Avaliação do ajuste (Medidas de aderência)

Avaliamos os modelos através de alguns métodos, tais como: Tabela de Classificação (ou Tabela de Previsão ou ainda Matriz de Confusão), Área abaixo da Curva ROC (*Receiver Operating Characteristic*), χ^2 de Pearson e o Teste de Hosmer-Lemeshow. Os dois últimos não serão abordados neste trabalho, mas encontram-se vastamente explicados em (15), onde podem ser vistos de forma bastante aplicada.

– Tabela de Classificação: Para esta análise deve-se estipular um ponto de corte, c^*, geralmente usa-se o valor 0,5. Este será comparado aos valores estimados do modelo de regressão logística, $\hat{\pi}(\mathbf{x}_i)$, e desta forma obtém-se os valores de $\hat{y}_i$ da seguinte maneira:

$$\text{se } \hat{\pi}(\mathbf{x}_i) > c^* \text{ então } \hat{y}_i = 1$$
$$\text{se } \hat{\pi}(\mathbf{x}_i) \leq c^* \text{ então } \hat{y}_i = 0.$$

Feito isto, se monta uma tabela de contingência cruzando com os valores observados de y_i com os valores encontrados no procedimento anterior, $\hat{y}_i$ e verifica-se quantas classificações foram feitas corretamente.

Na Tabela 2.1 tiramos algumas medidas relevantes para avaliar o ajuste:

– Taxa de acerto total: $\left(\frac{A+D}{A+B+C+D}\right) \times 100\%$

– Taxa de acertos para 0: $\left(\frac{A}{A+B}\right) \times 100\%$ (*Especificidade*)

– Taxa de acertos para 1: $\left(\frac{D}{C+D}\right) \times 100\%$ (*Sensitividade*)

Tabela 2.1: Tabela de Classificação

	Predito (ŷ)		
Observado (y)	0	1	
0	A	B	A + B
1	C	D	C + D
	A + C	B + D	A + B + C + D

A taxa de acertos para 1, ou seja, a probabilidade de estimar o sucesso dado que o valor real observado é realmente 1, também é chamada de

Sensitividade. Da mesma forma, a taxa de acertos para 0 é conhecida como *Especificidade*.

– Área abaixo da Curva ROC: Diferentemente da Tabela de Classificação na qual a Especificidade e a Sensitividade provêm de um único ponto de corte, neste método, pode-se variar o valor do ponto de corte utilizando o maior número possível de opções, a fim de recalcular as duas medidas citadas. Após, é feito um gráfico da Sensibilidade contra (1 - Especificidade). A curva que se forma neste gráfico é chamada de Curva ROC.

  Pelo fato de se esperar que a Sensitividade e a Especificidade sejam complementares, a área abaixo da curva ROC que indica se o modelo discriminou corretamente os fracassos (zeros) e sucessos (uns) deve ser igual a 1.

  Na literatura encontra-se uma regra que descreve a área abaixo da ROC e a qualidade do ajuste ligada a ela (ver (15)) como descrito na Tabela 2.2.

Área abaixo da ROC	Discriminação
$= 0.5$	Sem discriminação
$0.7 \leq ROC < 0.8$	Aceitável
$0.8 \leq ROC < 0.9$	Excelente
≥ 0.9	Excepcional

Tabela 2.2: Qualidade do ajuste - ROC