Modelos Lineares Generalizados

Unidade II

Terezinha K. A. Ribeiro

terezinha.ribeiro@unb.br

Instituto de Ciências Exatas
Departamento de Estatística
Universidade de Brasília

Organização

- Modelo nulo e modelo saturado
- Função desvio
- Casos particulares
- Coeficiente de determinação
- Testes de hipóteses

Modelo nulo

Suponha que $Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{FE}\left(\mu_i, \phi\right)$ e $g\left(\mu_i\right) = \beta_1$, para $i = 1, 2, \ldots, n$.

Observe que o modelo suposto para os dados é um MLG que possui apenas o intercepto β_1. Este modelo é usualmente conhecido por modelo nulo. Neste caso, a matriz do modelo X se reduz a um vetor coluna, formado de 1's.

O modelo nulo atribui toda a variação da resposta ao componente aleatório dos MLGs.

Como fica a estimativa de máxima verossimilhança para β_1? E a estimativa de μ_i?

Neste caso, tem-se $\mu_i = g^{-1}\left(\beta_1\right)$ e $x_{i1} = 1$ para todo $i = 1, 2, \ldots, n$.

Modelo nulo

Já vimos que o escore para β_j sob o MLG com p parâmetros é expresso por

$$U_{\beta_j} = \phi \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij}.$$

Se o modelo contém apenas o intercepto β_1, então

$$U_{\beta_1} = \phi \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i},$$

em que μ_i, $V_i = V_i(\mu_i)$ e $d\mu_i/d\eta_i = [1/g'(\mu_i)]$ são constantes dependentes apenas de β_1 para todo i pois $\mu_i = g^{-1}(\beta_1)$.

Modelo nulo

Daí, a equação $U_{\beta_1} = 0$ implica que

$$\sum_{i=1}^{n}[y_i - g^{-1}(\widehat{\beta}_1)] = 0 \ \Leftrightarrow \ \widehat{\mu}_i = g^{-1}(\widehat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}, \forall i.$$

Note que

$$\widehat{\beta}_1 = g\left(\widehat{\mu}_i\right) = g(\overline{y}).$$

Portanto, os ajustes dos valores observados da resposta via o MLG apenas com o intercepto são diretamente a média amostral da resposta, isto é,

$$\widehat{y}_i = \widehat{\mu}_i = \overline{y}, \ \ \forall i.$$

Modelo saturado

Em contrapartida, considere que $Y_i \mid \boldsymbol{x}_i \stackrel{\text{ind}}{\sim} \text{FE}\,(\mu_i, \phi)$ e $g\,(\mu_i) = \beta_i$, para $i = 1, 2, \ldots, n$.

Observe que o modelo suposto para os dados é um MLG que possui n parâmetros, $\beta_1, \beta_2, \ldots, \beta_n$. Mais especificamente,

$$\mu_1 = g^{-1}(\beta_1), \;\; \mu_2 = g^{-1}(\beta_2), \quad \ldots, \quad \mu_n = g^{-1}(\beta_n).$$

Este modelo é usualmente conhecido por modelo saturado ou modelo completo. Neste caso, a matriz do modelo é a matriz identidade de ordem n, isto é, $X = I_n$.

O modelo saturado atribui toda a variação da resposta ao componente sistemático dos MLGs pois ajusta-se perfeitamente aos dados.

Modelo saturado

Como ficam os EMVs para os parâmetros $\beta_1, \beta_2, \ldots, \beta_n$? E a estimativa de μ_i?

Note que a relação das respostas médias com os β's é única e expressa por $g(\mu_i) = \beta_i$. Portanto, sob o modelo saturado podemos estimar diretamente as médias $\mu_1, \mu_2, \ldots, \mu_n$, e posteriormente, estabelecer as estimativas de $\beta_1, \beta_2, \ldots, \beta_n$.

Assim, o escore para μ_j, $j = 1, 2, \ldots, n$, sob o modelo saturado é

$$U_{\mu_j} = \phi \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V_i}.$$

Modelo saturado

O vetor escore para $\boldsymbol{\mu} = (\mu_1 \ \mu_2 \ \ldots \ \mu_n)^\top$ é

$$U_{\boldsymbol{\mu}} = \phi V^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

em que $V = \text{diag}\{V_1, V_2, \ldots, V_n\}$.

Daí, o EMV para o vetor de parâmetros $\boldsymbol{\mu}$ é obtido de

$$\phi V^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}) = \mathbf{0} \quad \Leftrightarrow \quad V^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}) = \mathbf{0} \quad \Leftrightarrow \quad \widetilde{\boldsymbol{\mu}} = \boldsymbol{y}.$$

em que $\widetilde{\boldsymbol{\mu}}$ denota o EMV de $\boldsymbol{\mu}$ sob o modelo saturado.

Logo, obtivemos que $\widetilde{\mu}_i = y_i, \forall i$, e consequentemente,

$$\widetilde{\beta}_i = g(\widetilde{\mu}_i) = g(y_i).$$

Perceba que devido a $\widehat{y}_i = \widetilde{\mu}_i = y_i, \forall i$, o modelo saturado possui o ajuste perfeito aos dados.

Função desvio

Na prática, queremos ter um bom ajuste de um modelo suposto aos dados que possua $p < n$ parâmetros.

Para avaliar o quão bom é o ajuste sob um modelo com $p < n$ parâmetros, pode-se comparar a distância entre o ajuste via o modelo saturado e o modelo com p parâmetros. Essa mensuração pode ser feita através da função desvio.

A função desvio é definida por

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} {d^*}^2(y_i; \widehat{\mu}_i) = 2\left\{\ell(\widetilde{\boldsymbol{\mu}}) - \ell(\widehat{\boldsymbol{\mu}})\right\},$$

em que $\ell(\boldsymbol{\mu})$ é o logaritmo da função de verossimilhança avaliado em $\boldsymbol{\mu}$.

Por definição $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \geq 0$.

Função desvio

A quantidade $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ é uma medida de qualidade de ajuste do modelo.

Note que para $p = n$ tem-se $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 0$.

Um valor pequeno para o desvio $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ indica que, para um menor número de parâmetros $p < n$, obtemos um ajuste aos dados tão bom quanto sob o modelo saturado.

Portanto, sob um modelo bem ajustado aos dados, espera-se um valor pequeno para o desvio.

A função desvio pode ser definida sob qualquer classe de modelos de regressão. Não é uma medida que é útil apenas para os MLGs.

Função desvio

Na família exponencial uniparamétrica, supondo ϕ conhecido,

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{n} \ell(\mu_i),$$

em que

$$\ell(\mu_i) = \phi[y_i\theta_i - b(\theta_i)] + c(y_i;\phi).$$

Segue que

$$\begin{aligned}
D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) &= 2\left\{\ell(\widetilde{\boldsymbol{\mu}}) - \ell(\widehat{\boldsymbol{\mu}})\right\} \\
&= 2\sum_{i=1}^{n}[\ell(\widetilde{\mu}_i) - \ell(\widehat{\mu}_i)] \\
&= 2\sum_{i=1}^{n}\left\{\phi[y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)] + c(y_i;\phi) - \phi[y_i\widehat{\theta}_i - b(\widehat{\theta}_i)] - c(y_i;\phi)\right\} \\
&= 2\sum_{i=1}^{n}\left\{\phi[y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)]\right\}.
\end{aligned}$$

Função desvio

Assim, a função desvio sob os MLGs pode ser expressa através de

$$D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = \phi D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}),$$

com

$$D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d^2(y_i;\widehat{\mu}_i), \quad {d^*}^2(y_i;\widehat{\mu}_i) = \phi d^2(y_i;\widehat{\mu}_i),$$

$$d^2(y_i;\widehat{\mu}_i) = 2\left\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)\right\}.$$

A quantidade $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})$ é chamado de desvio não escalonado.

Veremos mais à frente que os desvios individuais ${d^*}^2(y_i;\widehat{\mu}_i)$ serão utilizados para construir os resíduos conhecidos como componente do desvio.

Casos particulares - Distribuição normal

Sob o modelo normal, já vimos que $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2 = \mu_i^2/2$.

Assim, obtemos que

- as estimativas sob o modelo saturado são $\widetilde{\theta}_i = \widetilde{\mu}_i = y_i$, $b(\widetilde{\theta}_i) = y_i^2/2$;
- as estimativas sob o modelo com p parâmetros são $\widehat{\theta}_i = \widehat{\mu}_i$, $b(\widehat{\theta}_i) = \widehat{\mu}_i^2/2$.

Daí, segue que o componente do desvio não escalonado é

$$
\begin{aligned}
d^2(y_i; \widehat{\mu}_i) &= 2\left\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)\right\} \\
&= 2\left\{y_i(y_i - \widehat{\mu}_i) + \widehat{\mu}_i^2/2 - y_i^2/2\right\} \\
&= 2\left\{\frac{2y_i^2 - 2y_i\widehat{\mu}_i + \widehat{\mu}_i^2 - y_i^2}{2}\right\} \\
&= (y_i - \widehat{\mu}_i)^2.
\end{aligned}
$$

Casos particulares - Distribuição normal

Neste caso, $\phi = 1/\sigma^2$, então o componente do desvio escalonado fica dado por

$$d^{*2}(y_i; \widehat{\mu}_i) = \frac{(y_i - \widehat{\mu}_i)^2}{\sigma^2}.$$

Portanto, a função desvio (escalonada) é

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\sigma^2} = \frac{\mathrm{SQRes}}{\sigma^2},$$

em que $\mathrm{SQRes} = \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$ é a conhecida soma de quadrados dos resíduos sob o modelo de regressão normal.

Casos particulares - Distribuição normal

Sob a suposição $Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \sigma^2)$, mostra-se que

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \frac{\mathrm{SQRes}}{\sigma^2} \sim \chi^2_{n-p},$$

com χ^2_{n-p} denota a distribuição qui-quadrado com $(n-p)$ graus de liberdade.

Este resultado implica que $\mathrm{E}(D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})) = n - p$. Por esta razão, um valor do desvio $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ próximo de $(n-p)$ indica que o modelo está bem ajustado.

Casos particulares - Distribuição Poisson

Sob o modelo Poisson, $\theta_i = \log(\mu_i)$, $b(\theta_i) = e^{\theta_i} = \mu_i$, e $\phi = 1$.

Assim, obtemos que

- as estimativas sob o modelo saturado são $\widetilde{\theta}_i = \log(\widetilde{\mu}_i) = \log(y_i)$, $b(\widetilde{\theta}_i) = \widetilde{\mu}_i = y_i$ com $y_i > 0$;
- as estimativas sob o modelo com p parâmetros são $\widehat{\theta}_i = \log(\widehat{\mu}_i)$, $b(\widehat{\theta}_i) = \widehat{\mu}_i$.

Daí, segue que o componente do desvio escalonado é

$$
\begin{aligned}
{d^*}^2(y_i; \widehat{\mu}_i) &= d^2(y_i; \widehat{\mu}_i) \\
&= 2\left\{ y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i) \right\} \\
&= 2\left\{ y_i(\log(y_i) - \log(\widehat{\mu}_i)) + \widehat{\mu}_i - y_i \right\} \\
&= 2\left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (\widehat{\mu}_i - y_i) \right\}.
\end{aligned}
$$

Casos particulares - Distribuição Poisson

O que acontece quando $y_i = 0$? Vamos utilizar a definição de função desvio. Sabemos que

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\left\{\ell(\widetilde{\boldsymbol{\mu}}) - \ell(\widehat{\boldsymbol{\mu}})\right\} = 2\sum_{i=1}^{n}[\ell(\widetilde{\mu}_i) - \ell(\widehat{\mu}_i)].$$

Então, para a observação y_i,

$$d^{*2}(y_i; \widehat{\mu}_i) = 2[\ell(\widetilde{\mu}_i) - \ell(\widehat{\mu}_i)] = 2[\log(f(y_i; \widetilde{\mu}_i)) - \log(f(y_i; \widehat{\mu}_i))].$$

Quando $y_i = 0$, a função de probabilidade da distribuição Poisson é $f(0; \mu_i) = \mathrm{e}^{-\mu_i}$. Daí, segue que

$$f(0; \widetilde{\mu}_i) = \mathrm{e}^{-\widetilde{\mu}_i} = \mathrm{e}^{-0} = 1; \quad f(0; \widehat{\mu}_i) = \mathrm{e}^{-\widehat{\mu}_i}.$$

Casos particulares - Distribuição Poisson

Assim, para $y_i = 0$,

$$d^{*2}(0; \widehat{\mu}_i) = 2[\log(f(0; \widetilde{\mu}_i)) - \log(f(0; \widehat{\mu}_i))] = 2[\log(1) + \widehat{\mu}_i] = 2\widehat{\mu}_i.$$

Portanto, o componente do desvio escalonado é expresso por

$$d^{*2}(y_i; \widehat{\mu}_i) = 2\left\{y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (\widehat{\mu}_i - y_i)\right\}, \quad y_i \in \{1, 2, 3, \ldots\},$$

e para $y_i = 0$ é $d^{*2}(0; \widehat{\mu}_i) = 2\widehat{\mu}_i$.

Para $\mu_i \to \infty$, $\forall i$, $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ se aproxima da distribuição χ^2_{n-p}.

Casos particulares - Distribuição gama

Sob o modelo gama, $\theta_i = -1/\mu_i$, $b(\theta_i) = \log(\mu_i)$.

Assim, obtemos que

- as estimativas sob o modelo saturado são $\widetilde{\theta}_i = -1/\widetilde{\mu}_i = -1/y_i$, $b(\widetilde{\theta}_i) = \log(\widetilde{\mu}_i) = \log(y_i)$;
- as estimativas sob o modelo com p parâmetros são $\widehat{\theta}_i = -1/\widehat{\mu}_i$, $b(\widehat{\theta}_i) = \log(\widehat{\mu}_i)$.

Daí, segue que o componente do desvio não escalonado é

$$
\begin{aligned}
d^2(y_i;\widehat{\mu}_i) &= 2\left\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)\right\} \\
&= 2\left\{y_i\left(-\frac{1}{y_i} + \frac{1}{\widehat{\mu}_i}\right) + \log(\widehat{\mu}_i) - \log(y_i)\right\} \\
&= 2\left\{\frac{(y_i - \widehat{\mu}_i)}{\widehat{\mu}_i} + \log\left(\frac{\widehat{\mu}_i}{y_i}\right)\right\}.
\end{aligned}
$$

Casos particulares - Distribuição gama

Consequentemente, $d^{*2}(y_i; \widehat{\mu}_i) = \phi d^2(y_i; \widehat{\mu}_i)$.

Quando ϕ é grande, mostra-se que

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}.$$

Alguns trabalhos na literatura apontam que valores de ϕ a partir de 5 já se mostram grandes o suficiente para a aproximação ser razoável.

Casos particulares - Distribuição normal inversa

Sob o modelo normal inverso, $\theta_i = -1/2\mu_i^2$, $b(\theta_i) = -1/\mu_i$.

Assim, obtemos que

- as estimativas sob o modelo saturado são $\widetilde{\theta}_i = -1/2\widetilde{\mu}_i^2 = -1/2y_i^2$, $b(\widetilde{\theta}_i) = -1/\widetilde{\mu}_i = -1/y_i$;
- as estimativas sob o modelo com p parâmetros são $\widehat{\theta}_i = -1/2\widehat{\mu}_i^2$, $b(\widehat{\theta}_i) = -1/\widehat{\mu}_i$.

Daí, segue que o componente do desvio não escalonado é

$$\begin{aligned} d^2(y_i;\widehat{\mu}_i) &= 2\left\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)\right\} \\ &= 2\left\{y_i\left(-\frac{1}{2y_i^2} + \frac{1}{2\widehat{\mu}_i^2}\right) + \frac{1}{y_i} - \frac{1}{\widehat{\mu}_i}\right\}. \end{aligned}$$

$$
\begin{aligned}
d^2(y_i;\widehat{\mu}_i) &= -\frac{1}{y_i}+\frac{y_i}{\widehat{\mu}_i^2}+\frac{2}{y_i}-\frac{2}{\widehat{\mu}_i}\\
&= \frac{-\widehat{\mu}_i^2+y_i^2+2\widehat{\mu}_i^2-2y_i\widehat{\mu}_i}{y_i\widehat{\mu}_i^2}\\
&= \frac{(y_i-\widehat{\mu}_i)^2}{y_i\widehat{\mu}_i^2}.
\end{aligned}
$$

Portanto, o componente do desvio escalonado é

$$
{d^*}^2(y_i;\widehat{\mu}_i)=\phi d^2(y_i;\widehat{\mu}_i)=\phi\frac{(y_i-\widehat{\mu}_i)^2}{y_i\widehat{\mu}_i^2}.
$$

Analogamente, para ϕ é grande, mostra-se que

$$
D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}})\sim\chi^2_{n-p}.
$$

Casos particulares - Distribuição binomial

Suponha que $Y_i \sim \text{binomial}(n_i, \mu_i)$, $0 < \mu_i < 1$, $\phi_i = n_i$, $i = 1, 2, \ldots, k$.

Neste caso, $Y_i^* = Y_i/n_i \in \text{FE}(\mu_i, \phi_i)$ com $\mu_i = \mathrm{E}(Y_i^*)$. Daí, tem-se que $\theta_i = \log(\mu_i/(1-\mu_i))$, e $b(\theta_i) = -\log(1-\mu_i)$.

Perceba que

$$\widetilde{\mu}_i = y_i^* = \frac{y_i}{n_i}.$$

Assim,

- as estimativas sob o modelo saturado são

$$\widetilde{\theta}_i = \log\left(\frac{\widetilde{\mu}_i}{1-\widetilde{\mu}_i}\right) = \log\left(\frac{y_i^*}{1-y_i^*}\right) = \log\left(\frac{y_i/n_i}{1-y_i/n_i}\right),$$

$$b(\widetilde{\theta}_i) = -\log(1-\widetilde{\mu}_i) = -\log(1-y_i^*) = -\log\left(1-y_i/n_i\right);$$

Casos particulares - Distribuição binomial

- as estimativas sob o modelo com p parâmetros são

$$\widehat{\theta}_i = \log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right), \quad b(\widehat{\theta}_i) = -\log(1-\widehat{\mu}_i).$$

Daí, segue que o componente do desvio não escalonado é

$$\begin{aligned} d^2(y_i^*;\widehat{\mu}_i) &= 2\left\{y_i^*(\widetilde{\theta}_i - \widehat{\theta}_i) + b(\widehat{\theta}_i) - b(\widetilde{\theta}_i)\right\} \\ &= 2\left\{\frac{y_i}{n_i}\left[\log\left(\frac{y_i/n_i}{1-y_i/n_i}\right) - \log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right)\right]\right. \\ &\left. + \log(1-y_i/n_i) - \log(1-\widehat{\mu}_i)\right\} \\ &= 2\left\{\frac{y_i}{n_i}\left[\log(y_i/n_i) - \log(1-y_i/n_i) - \log(\widehat{\mu}_i) + \log(1-\widehat{\mu}_i)\right]\right. \\ &\left. + \log(1-y_i/n_i) - \log(1-\widehat{\mu}_i)\right\} \end{aligned}$$

Casos particulares - Distribuição binomial

$$
\begin{aligned}
d^2(y_i^*;\widehat{\mu}_i) &= 2\left\{\frac{y_i}{n_i}\left[\log\left(\frac{y_i}{n_i\widehat{\mu}_i}\right) - \log\left(\frac{1-y_i/n_i}{1-\widehat{\mu}_i}\right)\right] + \log\left(\frac{1-y_i/n_i}{1-\widehat{\mu}_i}\right)\right\} \\
&= 2\left\{\frac{y_i}{n_i}\log\left(\frac{y_i}{n_i\widehat{\mu}_i}\right) + \left(1-\frac{y_i}{n_i}\right)\log\left(\frac{1-y_i/n_i}{1-\widehat{\mu}_i}\right)\right\}.
\end{aligned}
$$

Daí, o componente do desvio escalonado fica dado por

$$
\begin{aligned}
d^{*2}(y_i^*;\widehat{\mu}_i) &= n_i d^2(y_i^*;\widehat{\mu}_i) \\
&= 2\left\{y_i\log\left(\frac{y_i}{n_i\widehat{\mu}_i}\right) + (n_i-y_i)\log\left(\frac{1-y_i/n_i}{1-\widehat{\mu}_i}\right)\right\},
\end{aligned}
$$

desde que $y_i \in \{1, 2, \ldots, n_i - 1\}$ ou $y_i^* \in \{1/n_i, 2/n_i, \ldots, (n_i - 1)/n_i\}$.

Casos particulares - Distribuição binomial

Quando $y_i^* = 0$ $(y_i = 0)$ e $y_i^* = 1$ $(y_i = n_i)$, a quantidade anterior não está bem definida.

Pela definição de função desvio, sabemos que para a observação y_i,

$$d^{*2}(y_i^*; \widehat{\mu}_i) = 2[\ell(\widetilde{\mu}_i) - \ell(\widehat{\mu}_i)] = 2[\log(f(y_i^*; \widetilde{\mu}_i)) - \log(f(y_i^*; \widehat{\mu}_i))].$$

Para $y_i^* = 0$ (ou $y_i = 0$), a função de probabilidade de Y_i^* fica dada por

$$f(0; \mu_i) = \binom{n_i}{n_i \times 0} \mu_i^{n_i \times 0} (1 - \mu_i)^{n_i - n_i \times 0} = (1 - \mu_i)^{n_i}.$$

Casos particulares - Distribuição binomial

Segue que

$$f(0;\widetilde{\mu}_i) = (1-\widetilde{\mu}_i)^{n_i} = (1-\widetilde{\mu}_i)^{n_i} = (1-0)^{n_i} = 1;$$

$$f(0;\widehat{\mu}_i) = (1-\widehat{\mu}_i)^{n_i};$$

Então, o componente do desvio escalonado para $y_i^* = 0$ (ou $y_i = 0$) é

$$\begin{aligned} {d^*}^2(0;\widehat{\mu}_i) &= 2[\log(f(0;\widetilde{\mu}_i)) - \log(f(0;\widehat{\mu}_i))] \\ &= 2[\log(1) - \log((1-\widehat{\mu}_i)^{n_i})] \\ &= -2n_i \log(1-\widehat{\mu}_i). \end{aligned}$$

Casos particulares - Distribuição binomial

Para $y_i^* = 1$ (ou $y_i = n_i$), a função de probabilidade de Y_i^* fica dada por

$$f(1;\mu_i) = \binom{n_i}{n_i \times 1} \mu_i^{n_i \times 1}(1-\mu_i)^{n_i - n_i \times 1} = \mu_i^{n_i}.$$

Segue que

$$f(1;\widetilde{\mu}_i) = \widetilde{\mu}_i^{n_i} = 1^{n_i} = 1;$$

$$f(1;\widehat{\mu}_i) = \widehat{\mu}_i^{n_i}.$$

O componente do desvio escalonado para $y_i^* = 1$ (ou $y_i = n_i$) é

$$\begin{aligned} {d^*}^2(1;\widehat{\mu}_i) &= 2[\log(f(1;\widetilde{\mu}_i)) - \log(f(1;\widehat{\mu}_i))] \\ &= 2[\log(1) - \log(\widehat{\mu}_i^{n_i})] \\ &= -2 n_i \log(\widehat{\mu}_i). \end{aligned}$$

Casos particulares - Distribuição binomial

Resumindo, o componente do desvio escalonado fica dado por

$$d^{*2}(y_i^*; \widehat{\mu}_i) = \begin{cases} 2\left\{ y_i \log\left(\frac{y_i}{n_i \widehat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{1 - y_i/n_i}{1 - \widehat{\mu}_i}\right)\right\}, & y_i^* \in \left\{\frac{1}{n_i}, \ldots, \frac{n_i - 1}{n_i}\right\} \\ -2n_i \log(1 - \widehat{\mu}_i), & y_i^* = 0, \\ -2n_i \log(\widehat{\mu}_i), & y_i^* = 1, \end{cases}$$

em que $0 < \widehat{\mu}_i < 1$.

Para n_i, $\forall i$ grande, mostra-se que

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \sim \chi^2_{k-p}.$$

Coeficiente de determinação

Sob o modelo de regressão linear normal, define-se o coeficiente de determinação ou coeficiente de explicação do modelo de regressão ajustado através de

$$R^2 = 1 - \frac{\text{SQRes}}{\text{SQT}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Por definição $0 \leq R^2 \leq 1$ é a proporção da variação da resposta Y que é explicada pela regressão linear normal ajustada.

Esta medida possui algumas problemas:

- R^2 tende a ser maior quando o tamanho amostral é pequeno;
- R^2 sempre aumenta com inclusão de novas covariáveis.

Coeficiente de determinação

Para corrigir os problemas citados, utiliza-se o coeficiente de determinação ajustado definido por

$$R_a^2 = 1 - \frac{\mathrm{SQRes}/(n-p)}{\mathrm{SQT}/(n-1)}.$$

Diferentemente de R^2, a medida R_a^2 não é uma proporção.

Mostra-se que

$$R_a^2 = R^2 - \frac{p-1}{n-p}(1-R^2).$$

Assim, $R_a^2 \leq R^2$.

Também, mostra-se que $R_a^2 < 0$ quando $R^2 < (p-1)/(n-1)$.

Coeficiente de determinação

Uma extensão natural do coeficiente de determinação para os MLGs é

$$R^2 = 1 - \frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{D(\boldsymbol{y};\overline{\boldsymbol{y}})},$$

em que $D(\boldsymbol{y};\overline{\boldsymbol{y}})$ denota o desvio do modelo apenas com o intercepto.

Segundo Paula (2023), o coeficiente de determinação para os MLGs (exceto para o caso normal) raramente é maior do que 0,40. Assim, este valor é usado como referencial de um ótimo ajuste aos dados.

Existem diversos pseudo-R^2 na literatura. Pode-se definir, por exemplo,

$$R_p^2 = r_{\eta,g}^2,$$

em que $r_{\eta,g}$ denota a correlação linear entre o preditor ajustado $\widehat{\eta}$ e a resposta transformada $g(\boldsymbol{y})$.

Coeficiente de determinação

Uma extensão natural do coeficiente de determinação para os MLGs é

$$R^2 = 1 - \frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{D(\boldsymbol{y}; \overline{\boldsymbol{y}})},$$

em que $D(\boldsymbol{y}; \overline{\boldsymbol{y}})$ denota o desvio do modelo apenas com o intercepto.

Segundo Paula (2023), o coeficiente de determinação para os MLGs (exceto para o caso normal) raramente é maior do que 0,40. Assim, este valor é usado como referencial de um ótimo ajuste aos dados.

Existem diversos pseudo-R^2 na literatura. Pode-se definir, por exemplo,

$$R_p^2 = r_{\eta,g}^2,$$

em que $r_{\eta,g}$ denota a correlação linear entre o preditor ajustado $\widehat{\boldsymbol{\eta}}$ e a resposta transformada $g(\boldsymbol{y})$.

Testes de hipóteses - Razão de verossimilhanças

Considere o particionamento $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top \ \boldsymbol{\beta}_2^\top)^\top \in \mathbb{R}^p$ com $\boldsymbol{\beta}_1 \in \mathbb{R}^q$ e $\boldsymbol{\beta}_2 \in \mathbb{R}^{p-q}$.

Inicialmente considere que o parâmetro de precisão ϕ seja um valor conhecido.

Queremos testar H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ contra H_1: $\boldsymbol{\beta}_1 \neq \mathbf{0}$. Para tanto, considere a estatística

$$\lambda(\boldsymbol{y}) = \frac{L(\widehat{\boldsymbol{\mu}}^0)}{L(\widehat{\boldsymbol{\mu}})},$$

em que $\widehat{\boldsymbol{\mu}}^0$ e $\widehat{\boldsymbol{\mu}}$ são os estimadores de máxima verossimilhança para $\boldsymbol{\mu}$ sob H_0 (restrito a H_0) e sob o modelo irrestrito, respectivamente.

Testes de hipóteses - Razão de verossimilhanças

A estatística $\lambda(\boldsymbol{y})$ é chamada de razão de verossimilhanças, e satisfaz $0 \leq \lambda(\boldsymbol{y}) \leq 1$. Rejeita-se H_0 para valores baixos da estatística $\lambda(\boldsymbol{y})$.

O teste da razão de verossimilhanças (TRV) rejeita H_0 ao nível de significância $\alpha \cdot 100\%$ se $\lambda(\boldsymbol{y}) \leq \lambda_0$ em que λ_0 é obtido de

$$\mathrm{P}_{H_0}(\lambda(\boldsymbol{Y}) \leq \lambda_0) = \alpha.$$

Note que precisamos conhecer a distribuição de $\lambda(\boldsymbol{Y})$ sob H_0.

Geralmente não conhecemos a distribuição exata de $\lambda(\boldsymbol{Y})$ sob H_0. Por esta razão, utiliza a distribuição assintótica de uma transformação de $\lambda(\boldsymbol{Y})$.

Testes de hipóteses - Razão de verossimilhanças

Sob H_0 e para n grande, mostra-se que

$$\xi_{RV} = -2\log[\lambda(\boldsymbol{Y})] \sim \chi^2_q.$$

No nosso caso, temos

$$\begin{aligned}
\xi_{RV} &= -2\log\left[\frac{L(\widehat{\boldsymbol{\mu}}^0)}{L(\widehat{\boldsymbol{\mu}})}\right] \\
&= -2\{\log[L(\widehat{\boldsymbol{\mu}}^0)] - \log[L(\widehat{\boldsymbol{\mu}})]\} \\
&= 2\{\ell(\widehat{\boldsymbol{\mu}}) - \ell(\widehat{\boldsymbol{\mu}}^0)\} \\
&= 2\{\ell(\widehat{\boldsymbol{\mu}}) - \ell(\widetilde{\boldsymbol{\mu}}) + \ell(\widetilde{\boldsymbol{\mu}}) - \ell(\widehat{\boldsymbol{\mu}}^0)\} \\
&= D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) \\
&= \phi[D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})].
\end{aligned}$$

Note que ϕ deve ser conhecido para que ξ_{RV} seja uma estatística.

Testes de hipóteses - Razão de verossimilhanças

Como $\lambda(\boldsymbol{y}) \leq 1$, tem-se que $\xi_{RV} \geq 0$. Assim, temos que

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) - D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \geq 0 \Rightarrow D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) \geq D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}).$$

O desvio do modelo restrito sempre será maior ou igual que o desvio do modelo irrestrito.

Deve-se rejeitar H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ para valores altos de ξ_{RV}. Assim, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $\xi_{RV} \geq \chi^2_{q,\alpha}$ em que $\chi^2_{q,\alpha}$ é obtido de

$$\mathrm{P}(\chi^2_q \geq \chi^2_{q,\alpha}) = \alpha.$$

Testes de hipóteses - Razão de verossimilhanças

Para o caso em que o parâmetro ϕ é desconhecido, considere

$$
\begin{aligned}
F_0 &= \frac{\xi_{RV}/q}{D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)} \\
&= \frac{\phi[D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})]/q}{\phi D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)} \\
&= \frac{[D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})]/q}{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)}.
\end{aligned}
$$

Observe que

- F_0 não depende de ϕ ;
- o numerador e o denominador de F_0 são variáveis qui-quadrado (aproximadamente) divididas pelos seus graus de liberdade;
- mostra-se independência entre numerador e denominador.

Sob H_0: $\beta_1 = \mathbf{0}$, para n grande ou para n qualquer e ϕ grande,

$$F_0 \sim F_{q,n-p},$$

em que $F_{q,n-p}$ denota a distribuição F de Snedecor com q graus de liberdade no numerador e $(n-p)$ graus de liberdade no denominador.

Deve-se rejeitar H_0 para valores grandes de F_0. Logo, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $F_0 \geq F_{q,n-p,\alpha}$ em que $F_{q,n-p,\alpha}$ é obtido de

$$\mathrm{P}(F_{q,n-p} \geq F_{q,n-p,\alpha}) = \alpha.$$

Testes de hipóteses - Razão de verossimilhanças

Agora, considere as hipóteses H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ contra H_1: $\boldsymbol{\beta} \neq \boldsymbol{\beta}^0$, com $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ vetor de constantes preestabelecido.

Para ϕ conhecido, a estatística do TRV é

$$\xi_{RV} = \phi[D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})],$$

em que $\widehat{\boldsymbol{\mu}}^0 = g^{-1}(\widehat{\boldsymbol{\eta}}_0)$ com $\widehat{\boldsymbol{\eta}}_0 = X\boldsymbol{\beta}^0$.

Observe que, neste caso, $\widehat{\boldsymbol{\mu}}^0$ é uma constante que depende de $\boldsymbol{\beta}^0$ estabelecido previamente em H_0. Em outras palavras, aqui não é necessário estimar $\boldsymbol{\mu}$ sob H_0.

Além do TRV, podemos testar estas hipóteses através dos testes de Wald e escore.

Para testar H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, a estatística do teste de Wald é definida por

$$\xi_W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top [\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}})}]^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0),$$

em que $\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}})}$ denota a matriz de covariâncias assintótica de $\widehat{\boldsymbol{\beta}}$ estimada (avaliada em $\widehat{\boldsymbol{\beta}}$).

Sob H_0 e para n grande, mostra-se que

$$\xi_W \sim \chi^2_p.$$

Perceba que se H_0 é verdadeira, a estatística ξ_W deve assumir valor próximo a zero, caso contrário, assumirá valor discrepante positivo.

Assim, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $\xi_W \geq \chi^2_{p,\alpha}$ em que $\chi^2_{p,\alpha}$ é obtido de

$$\mathsf{P}(\chi^2_p \geq \chi^2_{p,\alpha}) = \alpha.$$

Em particular, se $p = 1$, a estatística de Wald é

$$\xi_W = \frac{(\widehat{\beta} - \beta^0)^2}{\widehat{\mathrm{Var}(\widehat{\beta})}}.$$

Daí, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $\xi_{SR} \geq \chi^2_{1,\alpha}$.

Neste caso, a estatística ξ_W equivale ao quadrado da estatística:

$$T_0 = \frac{\widehat{\beta} - \beta^0}{\widehat{\mathrm{EP}(\widehat{\beta})}},$$

em que $\widehat{\mathrm{EP}(\widehat{\beta})}$ denota o erro-padrão de $\widehat{\beta}$ estimado.

Sob H_0 e para n grande, $T_0 \sim \mathrm{N}(0, 1)$. Daí, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $T_0 \geq z_{\alpha/2}$ ou se $T_0 \leq -z_{\alpha/2}$, em que $z_{\alpha/2}$ é obtido de

$$\mathrm{P}(Z \geq z_{\alpha/2}) = \alpha/2, \quad Z \sim \mathrm{N}(0, 1)$$

Testes de hipóteses - Wald

A matriz de covariâncias assintótica de $\widehat{\beta}$ sob os MLG é

$$\mathrm{Var}(\widehat{\beta}) = \phi^{-1}(X^\top W X)^{-1}.$$

Assim, se ϕ é conhecido, a estatística de Wald fica expressa por

$$\xi_W = \phi(\widehat{\beta} - \beta^0)^\top (X^\top \widehat{W} X)(\widehat{\beta} - \beta^0).$$

Note que valores grandes da estatística ξ_W indicam evidências contra H_0. Logo, rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $\xi_W \geq \chi^2_{p,\alpha}$.

A estatística Wald é umas das mais utilizadas na prática pela sua forma simples de ser calculada.

Testes de hipóteses - Escore

Para testar H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, a estatística do teste escore ou teste de Rao se baseia em $U_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, e é definida por

$$\xi_{SR} = U_{\boldsymbol{\beta}}(\boldsymbol{\beta}^0)^\top \widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}})} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}^0),$$

em que

- $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}^0)$ denota o vetor escore de $\boldsymbol{\beta}$ avaliado em $\boldsymbol{\beta}^0$;
- $\widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}})}$ denota a matriz de covariâncias assintótica de $\widehat{\boldsymbol{\beta}}$ estimada sob H_0 (avaliada em $\boldsymbol{\beta}^0$).

Sob H_0 e para n grande, mostra-se que

$$\xi_{SR} \sim \chi^2_p.$$

Se H_0 é verdadeira, o valor de ξ_{SR} deve ser próximo a zero, enquanto que se H_0 é falsa, ξ_{SR} deve assumir um valor grande.

Vimos que para os MLGs,

$$U_{\beta} = \phi X^{\top} W^{1/2} V^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}), \quad \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \phi^{-1} (X^{\top} W X)^{-1}.$$

Daí,

$$U_{\beta}(\boldsymbol{\beta}^0) = \phi X^{\top} \widehat{W}_0^{1/2} \widehat{V}_0^{-1/2} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0), \quad \widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}})} = \phi^{-1} (X^{\top} \widehat{W}_0 X)^{-1}.$$

Para os MLGs, se ϕ é conhecido, a estatística de escore fica dada por

$$\begin{aligned} \xi_{SR} &= U_{\beta}(\boldsymbol{\beta}^0)^{\top} \widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}})} U_{\beta}(\boldsymbol{\beta}^0) \\ &= \phi^{-1} U_{\beta}(\boldsymbol{\beta}^0)^{\top} (X^{\top} \widehat{W}_0 X)^{-1} U_{\beta}(\boldsymbol{\beta}^0). \end{aligned}$$

A estatística ξ_{SR} é mais conveniente em situações em que a hipótese alternativa é bem mais complexa do que a hipótese nula. Portanto, esta pode uma boa alternativa quando estimar as quantidades necessárias sob H_0 é relativamente mais simples.

Rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $\xi_{SR} \geq \chi^2_{p,\alpha}$ em que $\chi^2_{p,\alpha}$ é obtido de

$$\mathrm{P}(\chi^2_p \geq \chi^2_{p,\alpha}) = \alpha.$$

Vale ressaltar que os três testes, em geral, são válidos para tamanhos amostrais grandes.

Testes de hipóteses - Caso normal

Vamos obter as expressões das três estatísticas de teste para o caso normal com função de ligação identidade e σ^2 conhecido.

A função desvio não escalonada para o caso normal com ligação identidade é

$$D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})^2.$$

Para a estatística do TRV,

$$\begin{aligned}
\xi_{RV} &= \phi[D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^0)^2 - \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})^2\right] \\
&= \frac{1}{\sigma^2}\left[(\boldsymbol{y} - X\boldsymbol{\beta}^0)^\top(\boldsymbol{y} - X\boldsymbol{\beta}^0) - (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})\right].
\end{aligned}$$

Testes de hipóteses - Caso normal

Ainda, observe que

$$
\begin{aligned}
(\boldsymbol{y} - X\beta^0)^\top(\boldsymbol{y} - X\beta^0) &= (\boldsymbol{y} - X\widehat{\beta} + X\widehat{\beta} - X\beta^0)^\top(\boldsymbol{y} - X\widehat{\beta} + X\widehat{\beta} - X\beta^0) \\
&= (\boldsymbol{y} - X\widehat{\beta})^\top(\boldsymbol{y} - X\widehat{\beta}) + (\boldsymbol{y} - X\widehat{\beta})^\top(X\widehat{\beta} - X\beta^0) \\
&+ (X\widehat{\beta} - X\beta^0)^\top(\boldsymbol{y} - X\widehat{\beta}) + (X\widehat{\beta} - X\beta^0)^\top(X\widehat{\beta} - X\beta^0) \\
&= (\boldsymbol{y} - X\widehat{\beta})^\top(\boldsymbol{y} - X\widehat{\beta}) + 2(X\widehat{\beta} - X\beta^0)^\top(\boldsymbol{y} - X\widehat{\beta}) \\
&+ (\widehat{\beta} - \beta^0)^\top X^\top X(\widehat{\beta} - \beta^0).
\end{aligned}
$$

Testes de hipóteses - Caso normal

Sob o modelo normal com ligação identidade, vimos que o vetor escore para β é

$$U_{\boldsymbol{\beta}} = \frac{1}{\sigma^2} X^\top (\boldsymbol{y} - \boldsymbol{\mu}).$$

Segue que

$$U_{\boldsymbol{\beta}} = \boldsymbol{0} \Leftrightarrow X^\top (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{0},$$

com $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$.

Então,

$$X^\top (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = X^\top \boldsymbol{y} - X^\top X\widehat{\boldsymbol{\beta}} = \boldsymbol{0} \Leftrightarrow X^\top X\widehat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.$$

Desde que $r(X) = r(X^\top X) = p$ (X é de posto completo), obtemos

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Testes de hipóteses - Caso normal

Assim,

$$\begin{aligned} 2(X\widehat{\beta} - X\beta^0)^\top(\mathbf{y} - X\widehat{\beta}) &= 2(\widehat{\beta} - \beta^0)^\top X^\top(\mathbf{y} - X(X^\top X)^{-1}X^\top\mathbf{y}) \\ &= 2(\widehat{\beta} - \beta^0)^\top(X^\top\mathbf{y} - X^\top X(X^\top X)^{-1}X^\top\mathbf{y}) \\ &= 2(\widehat{\beta} - \beta^0)^\top(X^\top\mathbf{y} - X^\top\mathbf{y}) \\ &= 0. \end{aligned}$$

Segue que,

$$(\mathbf{y} - X\beta^0)^\top(\mathbf{y} - X\beta^0) = (\mathbf{y} - X\widehat{\beta})^\top(\mathbf{y} - X\widehat{\beta}) + (\widehat{\beta} - \beta^0)^\top X^\top X(\widehat{\beta} - \beta^0).$$

Portanto, a estatística do TRV fica expressa por

$$\xi_{RV} = \frac{1}{\sigma^2}(\widehat{\beta} - \beta^0)^\top X^\top X(\widehat{\beta} - \beta^0).$$

Testes de hipóteses - Caso normal

A estatística de Wald é

$$\xi_W = \phi(\widehat{\beta} - \beta^0)^\top (X^\top \widehat{W} X)(\widehat{\beta} - \beta^0).$$

No modelo normal com ligação identidade, sabemos que $W = I_n$, então $\widehat{W} = I_n$. Logo, obtemos que

$$\xi_W = \frac{1}{\sigma^2}(\widehat{\beta} - \beta^0)^\top (X^\top X)(\widehat{\beta} - \beta^0).$$

Note que $\xi_W = \xi_{RV}$.

Por fim, a estatística do teste escore é

$$\xi_{SR} = \phi^{-1} U_\beta(\beta^0)^\top (X^\top \widehat{W}_0 X)^{-1} U_\beta(\beta^0).$$

Mas,

$$U_\beta(\beta^0) = \frac{1}{\sigma^2} X^\top (\boldsymbol{y} - X\beta^0).$$

Testes de hipóteses - Caso normal

Também, $\widehat{W}_0 = I_n$. Então, segue que

$$\begin{aligned}
\xi_{SR} &= \sigma^2 \left[\frac{1}{\sigma^2} X^\top (\boldsymbol{y} - X\boldsymbol{\beta}^0)\right]^\top (X^\top \widehat{W}_0 X)^{-1} \frac{1}{\sigma^2} X^\top (\boldsymbol{y} - X\boldsymbol{\beta}^0) \\
&= \frac{1}{\sigma^2} (\boldsymbol{y} - X\boldsymbol{\beta}^0)^\top X (X^\top X)^{-1} X^\top (\boldsymbol{y} - X\boldsymbol{\beta}^0) \\
&= \frac{1}{\sigma^2} (X^\top \boldsymbol{y} - X^\top X\boldsymbol{\beta}^0)^\top (X^\top X)^{-1} (X^\top \boldsymbol{y} - X^\top X\boldsymbol{\beta}^0).
\end{aligned}$$

Mas, $X^\top \boldsymbol{y} = X^\top X\widehat{\boldsymbol{\beta}}$. Logo,

$$\begin{aligned}
\xi_{SR} &= \frac{1}{\sigma^2} (X^\top X\widehat{\boldsymbol{\beta}} - X^\top X\boldsymbol{\beta}^0)^\top (X^\top X)^{-1} (X^\top X\widehat{\boldsymbol{\beta}} - X^\top X\boldsymbol{\beta}^0) \\
&= \frac{1}{\sigma^2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top (X^\top X)^\top (X^\top X)^{-1} (X^\top X)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \\
&= \frac{1}{\sigma^2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top (X^\top X)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0).
\end{aligned}$$

Testes de hipóteses

Quando o parâmetro de precisão ϕ é desconhecido, pode-se utilizar a estatística

$$F_0 = \frac{[D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})]/p}{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})/(n-p)}.$$

Sob H_0 e para ϕ grande, $F_0 \sim F_{p,n-p}$.

Também, se no denominador de F_0 tivermos uma estimativa consistente para ϕ^{-1}, o resultado será válido quando n cresce.

Rejeita-se H_0 ao nível de significância $\alpha \cdot 100\%$ se $F_0 \geq F_{p,n-p}$.

Testes de hipóteses - Modelos encaixados

Considere $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top \ \boldsymbol{\beta}_2^\top)^\top \in \mathbb{R}^p$ com $\boldsymbol{\beta}_1 \in \mathbb{R}^q$ e $\boldsymbol{\beta}_2 \in \mathbb{R}^{p-q}$.

Esta partição implica em $X = [X_1 \ X_2]$ com X_1 matriz $n \times q$ e X_2 matriz $n \times (p-q)$.

Suponha que o parâmetro de precisão ϕ seja um valor conhecido.

Queremos testar H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$ contra H_1: $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$. Aqui $\boldsymbol{\beta}_1^0$ pode ser diferente do vetor nulo.

A estatística do TRV fica dada por

$$\xi_{RV} = \phi[D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})],$$

em que $\widehat{\boldsymbol{\mu}}^0$ denota o EMV restrito a H_0. Em outras palavras, fixa-se $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$ e estima-se $\boldsymbol{\beta}_2$.

Sob H_0 e n grande, $\xi_{RV} \sim \chi^2_q$. Daí, rejeita-se H_0 se $\xi_{RV} \geq \chi^2_{q,\alpha}$.

Testes de hipóteses - Modelos encaixados

Para ϕ conhecido, a estatística do teste de Wald é dada por

$$\xi_W = (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)^\top [\widehat{\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1)}]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0),$$

em que $\widehat{\boldsymbol{\beta}}_1$ sai do vetor $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^\top \; \widehat{\boldsymbol{\beta}}_2^\top)^\top$.

Usando resultados de álgebra matricial, mostra-se que

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) = \phi^{-1} [X_1^\top W^{1/2} M_2 W^{1/2} X_1]^{-1},$$

com $M_2 = I_n - H_2$ com $H_2 = W^{1/2} X_2 (X_2^\top W X_2)^{-1} X_2^\top W^{1/2}$.

Esta última expressão facilita a implementação da estatística de Wald.

Analogamente, sob H_0 e n grande, $\xi_W \sim \chi_q^2$. Rejeita-se H_0 se $\xi_W \geq \chi_{q,\alpha}^2$.

Testes de hipóteses - Modelos encaixados

Para ϕ conhecido, a estatística do teste escore é dada por

$$\xi_{SR} = U_{\boldsymbol{\beta}_1}(\widehat{\boldsymbol{\beta}}^0)^\top \widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}}_1)} U_{\boldsymbol{\beta}_1}(\widehat{\boldsymbol{\beta}}^0),$$

em que $\widehat{\boldsymbol{\beta}}^0 = (\boldsymbol{\beta}_1^0,\ \widehat{\boldsymbol{\beta}}_2^0)$, e

$$U_{\boldsymbol{\beta}_1}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} = \phi X_1^\top W^{1/2} V^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}).$$

Pode-se mostrar que

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) = \phi^{-1}(R^\top W R)^{-1}, \quad R = X_1 - X_2 C,$$

$$C = (X_2^\top W X_2)^{-1} X_2^\top W X_1.$$

Testes de hipóteses - Modelos encaixados

O vetor escore para β pode ser reescrito na forma

$$U_{\boldsymbol{\beta}} = \phi^{1/2} X^\top W^{1/2} \boldsymbol{r}_p,$$

em que

$$\boldsymbol{r}_p = \phi^{1/2} V^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu})$$

é conhecido como resíduo de Pearson (veremos mais à frente).

Daí, a estatística do teste escore pode ser expressa através de

$$\xi_{SR} = \widehat{\boldsymbol{r}}_{p0}^\top \widehat{W}_0^{1/2} X_1 (\widehat{R}_0^\top \widehat{W}_0 \widehat{R}_0)^{-1} X_1^\top \widehat{W}_0^{1/2} \widehat{\boldsymbol{r}}_{p0},$$

em que as quantidades $\widehat{\boldsymbol{r}}_{p0}$, $\widehat{W}_0$ e $\widehat{R}_0$ são avaliadas em $\widehat{\boldsymbol{\beta}}^0$. Nesse caso,

$$\widehat{\boldsymbol{r}}_{p0} = \phi^{1/2} \widehat{V}_0^{-1/2} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0).$$

Testes de hipóteses - Modelos encaixados

Veja aplicação deste resultado para o modelo de análise de variância com um fator em Paula (2023), páginas 38-39.

Sob H_0 e para n grande, $\xi_{SR} \sim \chi^2_q$.

Portanto, rejeita-se H_0 se $\xi_{SR} \geq \chi^2_{q,\alpha}$.

A estatística do teste de Wald ainda pode ser reescrita dependendo da matriz R definida no teste escore:

$$\begin{aligned}\xi_W &= (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)^\top [\widehat{\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1)}]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0) \\ &= \phi(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)^\top [\widehat{R}^\top \widehat{W} \widehat{R}]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0).\end{aligned}$$

E se o parâmetro ϕ for desconhecido?

Para testar H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$ contra H_1: $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$ quando o parâmetro de precisão ϕ é desconhecido, as estatísticas ficam definidas como apresentando a seguir.

A estatística do teste Wald é

$$\begin{aligned}
\xi_W &= (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)^\top [\widehat{\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1)}]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0) \\
&= \widehat{\phi} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)^\top [\widehat{R}^\top \widehat{W} \widehat{R}]^{-1} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0).
\end{aligned}$$

Sob H_0 e para n grande, $\xi_W \sim \chi^2_q$.

Testes de hipóteses - Modelos encaixados

A estatística do teste escore é

$$\begin{aligned}
\xi_{SR} &= U_{\beta_1}(\widehat{\boldsymbol{\theta}}^0)^\top \widehat{\mathrm{Var}_0(\widehat{\boldsymbol{\beta}}_1)} U_{\beta_1}(\widehat{\boldsymbol{\theta}}^0) \\
&= \widehat{\boldsymbol{r}}_{p0}^\top \widehat{W}_0^{1/2} X_1 (\widehat{R}_0^\top \widehat{W}_0 \widehat{R}_0)^{-1} X_1^\top \widehat{W}_0^{1/2} \widehat{\boldsymbol{r}}_{p0},
\end{aligned}$$

com $\widehat{\boldsymbol{\theta}}^0 = (\widehat{\boldsymbol{\beta}}^0, \ \widehat{\phi}^0)^\top$ sendo o EMV para $\boldsymbol{\theta}$ sob H_0, e

$$\widehat{\boldsymbol{r}}_{p0} = \widehat{\phi}_0^{1/2} \widehat{V}_0^{-1/2} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0).$$

Sob H_0 e para n grande, $\xi_{SR} \sim \chi^2_q$.

Testes de hipóteses - Modelos encaixados

Para definir a estatística ξ_{RV}, considere o resultado a seguir.

Para os modelos normal, gama e normal inverso, vale a seguinte decomposição:

$$c(y;\phi) = d(\phi) + \phi a(y) + u(y),$$

em que $a(\cdot)$, $d(\cdot)$, e $u(\cdot)$ são funções diferenciáveis.

A estatística do TRV fica expressa na forma

$$\xi_{RV} = 2\{\widehat{\phi} t(\widehat{\boldsymbol{\mu}}) - \widehat{\phi}^0 t(\widehat{\boldsymbol{\mu}}^0)\} + 2n\{d(\widehat{\phi}) - d(\widehat{\phi}^0)\},$$

em que

$$t(\boldsymbol{\mu}) = \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i) + a(y_i)].$$

Sob H_0 e para n grande, $\xi_{RV} \sim \chi^2_q$.

Testes de hipóteses - Modelos encaixados

Para o modelo gama,

$$
\begin{aligned}
c(y;\phi) &= (\phi-1)\log(y) + \phi\log(\phi) - \log(\Gamma(\phi)) \\
&= \phi\log(y) - \log(y) + \phi\log(\phi) - \log(\Gamma(\phi)).
\end{aligned}
$$

Tome $a(y) = \log(y)$, $u(y) = -\log(y)$, e

$$d(\phi) = \phi\log(\phi) - \log(\Gamma(\phi)).$$

Também, temos que $\theta_i = -1/\mu_i$ e $b(\theta_i) = \log(\mu_i)$. Daí, obtemos que

$$
\begin{aligned}
t(\boldsymbol{\mu}) &= \sum_{i=1}^{n}[y_i\theta_i - b(\theta_i) + a(y_i)] \\
&= \sum_{i=1}^{n}\left[-\frac{y_i}{\mu_i} - \log(\mu_i) + \log(y_i)\right] \\
&= \sum_{i=1}^{n}\left[\log\left(\frac{y_i}{\mu_i}\right) - \frac{y_i}{\mu_i}\right].
\end{aligned}
$$