Avaliação substitiva (AS)

Disciplina: Modelagem Estatística
Instrutor: Luiz Max Carvalho
Monitor: Ezequiel Braga

08 de Julho de 2025

- O tempo para realização da prova é de 3 horas;
- Leia a prova toda com calma antes de começar a responder;
- Responda todas as questões sucintamente;
- Marque a resposta final claramente com um quadrado, círculo ou figura geométrica de sua preferência;
- Você tem direito a trazer **uma folha de "cola"** tamanho A4 frente e verso, que deverá ser entregue junto com as respostas da prova.

1. Compute me, baby!

Considere uma variável aleatória Y com função de densidade

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

onde $a : \mathbb{R}_{>0} \to \mathbb{R}_{>0}, b : \mathbb{R} \to \mathbb{R}_{>0}, c : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$, para $\theta \in \mathbb{R}$ e $\phi \in \mathbb{R}_{>0}$. Distribuições com a densidade acima pertencem ao que chamamos de família de dispersão exponencial.

a) (10 pontos) Mostre que $\mathbb{E}[Y] = b'(\theta)$ e que $\mathrm{Var}(Y) = a(\phi)b''(\theta)$.

b) (10 pontos) Considere uma amostra aleatória $V_1, V_2, \ldots, V_m$ tal que $V_i \sim \mathrm{Bin}(n_i, p_i)$, para $i = 1, 2, \ldots, m$. Defina $Y_i = V_i/n_i$.

   i) Encontre a distribuição de Y_i e mostre que ela pertence a família de dispersão exponencial.

   ii) Considere observações de uma covariável $x_1, x_2, \ldots, x_m$ e assuma que $g(\mathbb{E}[Y_i \mid x_i]) = \eta_i$, para $\eta_i := \beta_0 + \beta_1 x_i$. Encontre a função de ligação canônica, g.

Conceitos trabalhados: família exponencial, função de ligação, modelos de regressão. **Nível de dificuldade**: fácil.
Resolução: Defina $l(\xi) = \log f(y;\xi)$, para $\xi = (\theta, \phi)$. Logo, $\frac{\partial l(\xi)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$ e $\frac{\partial^2 l(\xi)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$. Por outro lado, perceba que $\mathbb{E}\left[\frac{\partial l(\xi)}{\partial \theta}\right] = 0$ e que $-\mathbb{E}\left[\frac{\partial^2 l(\xi)}{\partial \theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial l(\xi)}{\partial \theta}\right)^2\right]$, o que conclui o item a. Para o item b, temos

$$\begin{aligned}\Pr(Y_i = y_i) = \Pr(V_i = n_i y_i) &= \exp\left\{\log\binom{n_i}{n_i y_i} + n_i y_i \log p_i + (n_i - n_i y_i)\log(1-p_i)\right\}, \\ &= \exp\left\{\log\binom{n_i}{n_i y_i} + n_i y_i \log\frac{p_i}{1-p_i} + n_i \log(1-p_i)\right\},\end{aligned}$$

ou seja, $\theta_i = \log\frac{p_i}{1-p_i}$, $a(\phi_i) = \frac{1}{n_i}$, $b(\theta_i) = -\log(1-p_i) = \log(1+\exp(\theta_i))$ e $c(y,\phi) = \log\binom{n_i}{n_i y_i}$. Usando o que vimos no item a, $\mathbb{E}[Y_i] = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ e, como a g canônica satisfaz $g(\mathbb{E}[Y_i \mid X_i]) = \theta_i$, temos $g(t) = \log\frac{t}{1-t}$. ■
Comentário: Nesta questão trabalhamos mais uma vez os fundamentos da família exponencial de distribuições, em particular a classe das famílias de dispersão. Além disso, vimos que uma transformação útil de uma v.a. binomial ainda continua na família exponencial de dispersão, o que possibilita a montagem de modelos de regressão bem similares à regressão logística.

2. Propensity to be awesome.

Em muitos estudos epidemiológicos, os pesquisadores estão interessados em entender a relação entre uma variável resposta binária $Y_i \in \{0, 1\}$ (por exemplo,

presença ou ausência de uma doença) e um conjunto de variáveis explicativas X_i (como idade, exposição a um fator de risco, características genéticas, etc).

Quando o evento de interesse (isto é, $Y_i = 1$) é **raro** — o que significa que $\pi_i = \mathbb{P}(Y_i = 1 \mid X_i)$ costuma ser pequeno — torna-se ineficiente ou mesmo inviável coletar uma amostra aleatória grande o suficiente para estudá-lo com boa precisão.

Para lidar com isso, utiliza-se com frequência um **desenho caso-controle**, em que: todos ou muitos dos **casos** (indivíduos com $Y_i = 1$) são incluídos; apenas uma **amostra dos controles** (indivíduos com $Y_i = 0$) é coletada.

Esse desenho gera uma **amostra enviesada** da população, pois a probabilidade de inclusão no estudo depende da variável resposta. Mais precisamente, define-se $Z_i = 1$ se o indivíduo i foi incluído na amostra; $\rho_1 = \Pr(Z_i = 1 \mid Y_i = 1)$ como a fração de amostragem dos casos; e $\rho_0 = \Pr(Z_i = 1 \mid Y_i = 0)$ como a fração de amostragem dos controles.

Embora a amostragem dependa do desfecho Y_i, nosso interesse continua sendo estimar a relação entre X_i e Y_i, frequentemente usando um modelo de regressão logística. Para isso, tome uma amostra aleatória $Y_1, \ldots, Y_n$, com $\pi_i = \Pr(Y_i = 1 \mid X_i) = \frac{\exp[\alpha + X_i^\top \boldsymbol{\beta}]}{1+\exp[\alpha + X_i^\top \boldsymbol{\beta}]}$.

a) (10 pontos) Mostre que se $X_i = \begin{bmatrix} X_{1i} \\ X_{1i} + c \end{bmatrix}$, para uma constante conhecida $c \in \mathbb{R}$, o modelo é não-identificável. Forneça uma reparametrização que resolva este problema.

b) (10 pontos) Mostre que $\exp(\beta_j)$ pode ser escrito como uma razão de chances (*odds-ratio*) e forneça uma interpretação dessa quantidade quando os π_i's são pequenos.

c) (20 pontos) Mostre que

$$\Pr(Y_i = 1 \mid X_i, Z_i = 1) = \frac{\exp(\alpha^* + X_i^\top \boldsymbol{\beta})}{1 + \exp(\alpha^* + X_i^\top \boldsymbol{\beta})},$$

onde $\alpha^* = \alpha + \log(\rho_1/\rho_0)$. O que esse resultado significa, na prática, em termos de *odds-ratio*?

Conceitos trabalhados: Modelos lineares generalizados; viés de amostragem; reparametrização. **Nível de dificuldade**: médio.
Resolução: Seja $\eta_i := \alpha + X_i^\top \boldsymbol{\beta}$. Temos

$$\begin{aligned} \eta_i &= \alpha + X_{1i}\beta_1 + (X_{1i} + c)\beta_2, \\ &= \alpha + c\beta_2 + (\beta_1 + \beta_2)X_{1i}. \end{aligned}$$

Tome $\alpha = \tilde{\alpha} + cd$, $\beta_1 = \tilde{\beta}_1 + d$ e $\beta_2 = \tilde{\beta}_2 - d$, para $d \in \mathbb{R}$. Perceba que η_i é o mesmo $\forall d \in \mathbb{R}$, isto é, a verossimilhança é invariante. Para resolver isto, podemos definir $\alpha' = \alpha + c\beta_2$ e $\beta_1' = \beta_1 + \beta_2$. Para o item b, note que

$$\frac{\Pr(Y_i = 1 \mid X_{ij} = x_{ij})}{\Pr(Y_i = 0 \mid X_{ij} = x_{ij})} = \exp\left[\alpha + x_{ij}\beta_j + \sum_{k \neq j} x_{ik}\beta_k\right].$$

Logo,

$$\frac{\frac{\Pr(Y_i=1|X_{ij}=x_{ij}+1)}{\Pr(Y_i=0|X_{ij}=x_{ij}+1)}}{\frac{\Pr(Y_i=1|X_{ij}=x_{ij})}{\Pr(Y_i=0|X_{ij}=x_{ij})}} = \exp\left[\beta_j\right].$$

Quando π_i é pequeno para todo i, $1-\pi_i \approx 1$, o que implica que $\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_{i'}}{1-\pi_{i'}}} \approx \frac{\pi_i}{\pi_{i'}}$, ou seja, o risco relativo. Para o último item, perceba que

$$\begin{aligned}
\Pr(Y_i = 1 \mid X_i, Z_i = 1) &= \frac{\rho_1\pi_1}{\rho_1\pi_1 + \rho_0\pi_0}, \\
&= \frac{\rho_1 \frac{\exp[\alpha + X_i^\top \boldsymbol{\beta}]}{1+\exp[\alpha + X_i^\top \boldsymbol{\beta}]}}{\rho_1 \frac{\exp[\alpha + X_i^\top \boldsymbol{\beta}]}{1+\exp[\alpha + X_i^\top \boldsymbol{\beta}]} + \rho_0 \frac{1}{1+\exp[\alpha + X_i^\top \boldsymbol{\beta}]}}, \\
&= \frac{\exp[\alpha + \log(\rho_1/\rho_0) + X_i^\top \boldsymbol{\beta}]}{\exp[\alpha + \log(\rho_1/\rho_0) + X_i^\top \boldsymbol{\beta}] + 1},
\end{aligned}$$

o que mostra que a razão de chances (*odds-ratio*) não muda em comparação com o modelo inicial. ■

Comentário: Nesta questão trabalhamos um conceito fundamental na aplicação de modelos de regressão, que é levar em conta a amostragem. Vimos também uma maneira de reparametrizar o modelo reescrevendo o preditor linear. Por fim, descobrimos que um dos principais estimandos do modelo, a razão de chances, não muda em relação ao modelo inicial, facilitando a interpretação sob diferentes parametrizações.

3. Making it rain

Neste problema analisaremos dados que fazem parte de um estudo longitudinal sobre renda nos EUA, o *Panel Study of Income Dynamics*, iniciado em 1968. O subconjunto analisado consiste em 42 chefes de família que tinham entre 25 e 39 anos em 1968. As variáveis incluídas são:

- `lincm`, log da renda nominal anual;
- `age`, a idade em 1968;
- `cyear`, codificado como -10 em 1968, 0 em 1978 e 10 em 1988;
- `educ`, anos de escolaridade em 1968;
- `sex`, M = masculino, F = feminino.

Abaixo está um trecho da saída da função de ajuste de um modelo linear misto (LMM) com a função `lme` em R, onde a resposta é o logaritmo da renda, `lincm`:

$$\texttt{lincm}_{ij} = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{cyear}_{ij} + \beta_3 \texttt{educ}_i + \beta_4 \texttt{sex}_i + b_i + \varepsilon_{ij},$$

onde b_i representa os efeitos aleatórios, para $i = 1, \ldots, 42$ e $j = 1968, 1978, 1988$.

```
Linear mixed-effects model fit by REML
 Data: psid2
     AIC       BIC    logLik
  320.0178 339.5883 -153.0089
Random effects:
 Formula: ~1 | fid
       (Intercept) Residual
StdDev: 0.0419192 0.7505293

Fixed effects: lincm ~ age + cyear + educ + factor(sex)
                Value  Std.Error   t-value
(Intercept)  7.386823 0.6317104 11.693370
age         -0.020930 0.0152069 -1.376316
cyear        0.084163 0.0081889 10.277583
educ         0.116343 0.0275823  4.218021
factor(sex)M 1.311661 0.1422471  9.221007

Correlation:
             (Intr)   age  cyear  educ
age          -0.831
cyear         0.000  0.000
educ         -0.685  0.201 0.000
factor(sex)M 0.003 -0.217 0.000 0.041

Standardized Within-Group Residuals:
      Min         Q1      Med       Q3      Max
-3.4411400 -0.4003431 0.1070887 0.5602338 1.6037724

Number of Observations: 126
Number of Groups: 42
```

a) (10 pontos) Formule o modelo na forma matricial e explique o significado de cada parte. Declare cuidadosamente as suposições do modelo.

b) (5 pontos) Determine um intervalo de confiança aproximado de 95% para o coeficiente da variável `cyear`.

c) (10 pontos) Descreva como os efeitos aleatórios b_i podem ser estimados.

d) (15 pontos) Use os valores da saída do R para calcular a matriz de covariância estimada para a resposta $(Y_{i,1968}, Y_{i,1978}, Y_{i,1988})^T$.

Conceitos trabalhados: modelos lineares mistos; intervalos aproximados; componentes de variância.
Nível de dificuldade: .
Resolução: Considere o vetor de respostas $\boldsymbol{Y}_i^\top = (Y_{i,1968}, Y_{i,1978}, Y_{i,1988})^T$. Então, podemos escrever

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \begin{bmatrix}1\\1\\1\end{bmatrix} b_i + \boldsymbol{\varepsilon}_i,$$

para $\boldsymbol{X}_i = \begin{bmatrix} 1 & \texttt{age}_i & \texttt{cyear}_{i,1968} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i,1978} & \texttt{educ}_i & \texttt{sex}_i \\ 1 & \texttt{age}_i & \texttt{cyear}_{i,1988} & \texttt{educ}_i & \texttt{sex}_i \end{bmatrix}$, $b_i \sim \mathrm{N}(0, \tau^2)$ e $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(0, \sigma^2 I_3)$, onde $b_i \perp \boldsymbol{\varepsilon}_i$. Para o próximo item, sabemos que $\hat{\beta}_k \pm 1.96\,\mathrm{se}(\hat{\beta}_k)$ é intervalo (Wald) aproximado de 95% para β_k. Logo, observando o *output*, temos o intervalo $(0.068, 0.100)$. Para estimar b_i, precisamos notar que

$$\begin{bmatrix} b_i \\ \boldsymbol{Y}_i \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ \boldsymbol{X}_i\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tau^2 & \tau^2 \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \\ \tau^2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} & \tau^2 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \sigma^2 I_3 \end{bmatrix} \right)$$

e, usando as propriedades de multivariada normal,

$$\mathbb{E}[b_i \mid \boldsymbol{Y_i} = \boldsymbol{y}_i] = \tau^2 \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \left(\tau^2 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \sigma^2 I_3 \right)^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

Consequentemente, podemos obter estimadores de τ, σ e $\boldsymbol{\beta}$ e estimar b_i. Além disso, a matriz de covariância de $\boldsymbol{Y}_i$ é dada por

$$\Sigma_{\boldsymbol{Y}_i} = \tau^2 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \sigma^2 I_3,$$

ou seja, basta substituir τ e σ por seus estimadores obtidos no *output* do R, 0.0419192 e 0.7505293, respectivamente. ■

Comentário: Neste exercício trabalhamos as habilidades de extrair do *output* de um programa estatístico as quantidades necessárias para fazer inferências sobre modelos lineares multinível.