Primeira avaliação (A1)

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

16 de abril 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;
- A prova vale 100 pontos. A pontuação restante é contada como bônus;
- Você tem direito a trazer **uma folha de "cola"** tamanho A4 frente e verso (impressa ou escrita à mão), que deverá ser entregue junto com as respostas da prova.

1. Elementary, my dear Watson

Um dos principais usos do modelo linear é a inclusão de termos de interação, que permitem maior flexibilidade na modelagem do relacionamento das cováriaveis entre si e com a variável dependente. Nesta questão você vai usar seus conhecimentos sobre o modelo linear e suas habilidades de detetive para preencher as lacunas deixadas por uma analista que sumiu em circunstâncias suspeitas.

Palmirinha, a renomada estatística, estava estudando um conjunto de dados que incluía o peso Y ao nascer de n bebês, bem como o número X_1 de semanas que o bebê passou na barriga da mãe até o parto – "age" – e o sexo ("sex") X_2 do bebê, codificado como $X_2 = 0$ para sexo feminino (*female*) e $X_2 = 1$ para o sexo masculino (*male*). Sendo esperta, Palmirinha escolheu trabalhar com $\tilde{X}_1 = X_1 - m_1$, onde m_1 é a média dos X_1 na amostra. Ela estava considerando modelos **lineares** para a esperança condicional $\mathbb{E}[Y \mid \boldsymbol{X}] =: \mu(\boldsymbol{X}) = \mu(X_1, X_2)$. Em particular, Palmirinha suspeitava fortemente de que haveria uma **interação** entre X_1 e X_2.

Ela fez algumas descobertas que devem ter aborrecido as pessoas erradas, porque logo depois de terminar a sua análise, ela desapareceu. Seu escritório estava uma bagunça completa, e aqui analisaremos os alfarrábios restantes. Com a sua ajuda, os investigadores do World Statistics Covenant (WSC) serão capazes de completar a análise e, quem sabe, obter pistas sobre o paradeiro de Palmirinha.

a) (5 pontos) Escreva um modelo para Y que reflita a crença de Palmirinha sobre a presença de uma interação. Explique a interpretação de <u>todos</u> os parâmetros do modelo.

b) (10 pontos) No computador de Palmirinha, ainda aberto em uma janela do R, os investigadores viram o seguinte *output*:

```
Residuals:
   Min      1Q  Median     3Q    Max
-246.69 -138.11 -39.13 176.57 274.28

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2884.17      52.51  54.924  < 2e-16 ***
age_c           130.40      30.00   4.347 0.000313 ***
sexmale         163.16      74.25   2.198 0.039924 *
age_c:sexmale   -18.42      41.76  -0.441 0.663893

Residual standard error: 180.6 on 20 degrees of freedom
Multiple R-squared: 0.6435,  Adjusted R-squared: 0.59
F-statistic: 12.03 on 3 and 20 DF, p-value: 0.000101
```

Você acha que esses resultados justificam a suspeita da nossa heroína de que havia de fato uma interação significativa entre X_1 e X_2? O que uma interação estatisticamente significativa implica na prática, isto é, qual é a sua *interpretação*?

c) (20 pontos) No meio da bagunça os investigadores encontraram uma folha com a seguinte tabela, que infelizmente estava manchada de sangue nas caselas marcadas de (i) a (v). Ajude os investigadores completando a

Parâmetro	Descrição matemática	Valor
β_0	$\mu(age = m_1, sex = f)$	(i)
β_1	(ii)	130.40
β_2	$\mu(age = m_1, sex = m) - \mu(age = m_1, sex = f)$	(iii)
β_3	(iv)	(v)

Tabela 1: Evidência # 32: tabela com parâmetros, sua descrição matemática e estimativa pontual.

tabela com que deveria estar escrito em (i) – (v).

d) (5 pontos) Se Palmirinha tivesse utilizado X_1 e não $\tilde{X}_1$, o que aconteceria com β_0? E com β_1? Porquê?

e) (10 pontos) Nossa heroína também precisava calcular o efeito sobre a média de mudar uma unidade em X_1 mantendo X_2 fixa. Mostre aos investigadores como ela teria feito. Além disso, use o *output* que ela deixou para calcular esta quantidade para $X_2 = 1$.

Conceitos trabalhados: regressão linear; intervalos de confiança; interação; preditor linear; interpretação. **Nível de dificuldade**: médio.
Resolução: Um modelo que atende o desejo de Palmirinha é

$$\mathbb{E}[Y_i \mid \boldsymbol{X}] = \beta_0 + \beta_1 \tilde{X}_{1i} + \beta_2 X_{2i} + \beta_3 \tilde{X}_{1i} X_{2i}.$$

Neste modelo, β_0 representa o efeito esperado quando $\tilde{X}_{1i} = X_{2i} = 0$. Agora, para facilitar a interpretação, considere a seguinte reescrita:

$$\mathbb{E}[Y_i \mid \boldsymbol{X}] = \beta_0 + (\beta_1 + \beta_3 X_{2i})\tilde{X}_{1i} + \beta_2 X_{2i}.$$

A partir disso, é fácil ver que β_2 representa o efeito de X_{2i} quando $\tilde{X}_{1i}$ é 0; β_1, o efeito de mudar $\tilde{X}_{1i}$ em uma unidade quando X_{2i} é 0; e β_3, o efeito da interação $\tilde{X}_{1i} X_{2i}$. Note que $\tilde{X}_{1i} = 0$ significa que X_{1i} está exatamente na média m_1.

Ao observar o *output* do R, percebemos que não há efeito significativo da interação, baseado na magnitude do parâmetro e no teste de significância. Isso quer dizer que o efeito de considerar a idade do indivíduo quando é do sexo masculino é pequeno, ou seja, a diferença entre aumentar em uma unidade a idade de um indivíduo do sexo masculino e aumentar em uma unidade a idade de um indivíduo do sexo feminino é baixa, em média. Em geral, quando temos essa interação onde uma das variáveis é binária, o coeficiente da interação representa a diferença na média de uma das variáveis dependendo se a outra toma o valor 0 ou 1. Quando as variáveis são contínuas ou têm outras formas, as interpretações variam.

Para completar os valores da tabela, basta olhar o *output* fornecido, ou seja: $(i) = 2884.17, (iii) = 163.16$ e $(v) = -18.42$. Para as descrições, basta escrever as interpretações que colocamos no item a de maneira matemática, ou seja,

$$
\begin{aligned}
(ii) &= \mu(age = x+1, sex = f) - \mu(age = x, sex = f),\\
(iv) &= \mu(age = x+1, sex = m) - \mu(age = x, sex = m) -\\
&\quad - [\mu(age = x+1, sex = f) - \mu(age = x, sex = f)].
\end{aligned}
$$

Para responder d, vamos escrever o modelo em função de X_{1i}:

$$
\begin{aligned}
\mathbb{E}[Y_i \mid \boldsymbol{X}] &= \beta_0 + \beta_1(X_{1i} - m_1) + \beta_2 X_{2i} + \beta_3(X_{1i} - m_1)X_{2i},\\
&= \beta_0 - \beta_1 m_1 + \beta_1 X_{1i} + (\beta_2 - \beta_3 m_1)X_{2i} + \beta_3 X_{1i}X_{2i}.
\end{aligned}
$$

Logo, agora temos novos coeficientes $\tilde{\beta_0} = \beta_0 - \beta_1 m_1$ e $\tilde{\beta_2} = \beta_2 - \beta_3 m_1$, ou seja, o efeito de β_0 e β_2 mudam, mas β_1 e β_3 permanecem os mesmos.

Para analisar o efeito de mudar X_{1i} em uma unidade mantendo X_{2i} fixa, basta ver a diferença

$$
\mu(age = x_1 + 1, sex = x_2) - \mu(age = x_1, sex = x_2) = \beta_1 + \beta_3 x_2.
$$

Tomando $x_2 = 1$, o *output* nos dá $\beta_1 + \beta_3 = 111.98$. ■

Comentário: Neste *muder mystery* digno de Agatha Christie[1] nós usamos nossos conhecimentos de regressão linear para interpretar o *output* de um programa simples e extrair dele informações importantes. Manipulando algebricamente a expressão para a esperança condicional, é possível obter várias quantidades interessantes, em particular é possível escrever os coeficientes do modelo como operações com a média condicional $\mu(\cdot)$.

2. Exponential power

A família exponencial forma a base de muitos modelos e métodos importantes em Estatística. Uma vasta classe de modelos lineares pode ser construída a partir de membros da família exponencial, com aplicações em Epidemiologia, Finanças, Ecologia e muitas outras áreas.

a) (15 pontos) Considere uma amostra aleatória $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ com distribuição pertencente a família exponencial canônica, isto é,

$$
f(x_i; \eta) = h(x_i)\exp\{\eta T(x_i) - A(\eta)\},
$$

com η escalar. Mostre que

i) $T(\boldsymbol{X}) = \sum_{i=1}^{n} T(X_i)$ é suficiente para η;

ii) $\mathbb{E}_\eta[T(\boldsymbol{X})] = nA'(\eta)$;

iii) $\hat{\eta}_{\text{EMV}} = (A')^{-1}\left(\frac{1}{n}T(\boldsymbol{X})\right)$;

[1] Agatha Mary Clarissa Christie (1890–1976) foi uma escritora inglesa famosa por suas tramas de mistério, como “Assassinato no Expresso Oriente” (1934).

iv) $I_n(\eta) = nA''(\eta)$, onde $I_n(\eta)$ é a informação de Fisher de η para $\boldsymbol{X}$.

b) (5 pontos) Suponha que $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ são variáveis aleatórias Bernoulli com

$$\Pr(X_i = 1) = \frac{\exp(\alpha + \beta t_i)}{1 + \exp(\alpha + \beta t_i)}, \tag{1}$$

onde $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ são constantes conhecidas. Mostre que a distribuição conjunta para $\boldsymbol{X}$ forma uma distribuição exponencial com dois parâmetros e identifique as estatísticas suficientes T_1 e T_2.

c) (10 pontos) Considere a distribuição Pareto, na qual para $y > 1$ e $\theta > 0$ a densidade vale

$$f(y; \theta) = \theta y^{-\theta-1}.$$

A família em questão está na forma canônica? Exiba a variância da estatística suficiente e compute a informação de Fisher de θ.

Conceitos trabalhados: famĺia exponencial e suas propriedades. **Nível de dificuldade**: fácil.

Resolução: Para o primeiro item, note que

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \eta) = \exp\left\{\eta \sum_{i=1}^{n} T(x_i) - nA(\eta)\right\} \prod_{i=1}^{n} h(x_i).$$

Pelo teorema da fatorização, é evidente que $T(\boldsymbol{X})$ é suficiente. Agora, observe que

$$\int_{\mathcal{X}} f_X(x; \eta) dx = 1.$$

Diferenciando sob o sinal da integral,

$$\int_{\mathcal{X}} [T(x) - A'(\eta)]\, f_X(x; \eta) dx = 0, \tag{2}$$

ou seja,

$$\mathbb{E}_\eta[T(X)] = A'(\eta).$$

Como temos uma amostra iid, $\mathbb{E}_\eta[T(\boldsymbol{X})] = nA'(\eta)$. Para o EMV, seja $l_n(\eta) = \log f_{\boldsymbol{X}}(\boldsymbol{x}; \eta)$. Logo,

$$l'_n(\eta) = T(\boldsymbol{X}) - nA'(\eta),$$

isto é,

$$l'_n(\hat{\eta}_{\text{EMV}}) = 0 \Leftrightarrow T(\boldsymbol{X}) - nA'(\hat{\eta}_{\text{EMV}}) = 0 \Leftrightarrow \hat{\eta}_{\text{EMV}} = (A')^{-1}\left(\frac{1}{n} T(\boldsymbol{X})\right).$$

Por fim, observe que

$$I(\eta) = \mathbb{E}\left[(T(X) - A'(\eta))^2\right] = \mathrm{Var}(T(X))$$

e que, a partir de (2),

$$\int_{\mathcal{X}} [T(x) - A'(\eta)]^2 f_X(x;\eta)dx - A''(\eta) = 0,$$

ou seja, $I(\eta) = A''(\eta)$. Novamente, como $\boldsymbol{X}$ é iid, $I_n(\eta) = nA''(\eta)$.
Para o item b, note que

$$f_{X_i}(x_i \mid \theta, t_i) = \left[\frac{\exp(\alpha + \beta t_i)}{1 + \exp(\alpha + \beta t_i)}\right]^{x_i} \left[\frac{1}{1 + \exp(\alpha + \beta t_i)}\right]^{1-x_i},$$

para $\boldsymbol{\theta} = (\alpha, \beta)$. Logo,

$$f_{\boldsymbol{X}}(\boldsymbol{X}; \theta; \boldsymbol{t}) = \frac{\exp\{\alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i t_i\}}{\prod_{i=1}^n [1 + \exp(\alpha + \beta t_i)]}.$$

Definindo $T(\boldsymbol{X}) = (T_1, T_2)^\top = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i t_i\right)^\top$, temos

$$f_{\boldsymbol{X}}(\boldsymbol{X}; \theta; \boldsymbol{t}) = \frac{\exp\left\{\boldsymbol{\theta}^\top T(\boldsymbol{X})\right\}}{\prod_{i=1}^n [1 + \exp(\alpha + \beta t_i)]},$$

como queríamos.

Para o último item, perceba que a distribuição em questão não está na forma canônica. Para facilitar as contas, vamos reparametrizar:

$$f(y;\theta) = \exp\{\log(\theta) + (-\theta - 1)\log(y)\} = \exp\{\log(-\eta - 1) + \eta \log(y)\},$$

para $\eta = -\theta - 1$, ou seja, $A(\eta) = -\log(-\eta - 1)$. Assim, pelo item a, $T(\boldsymbol{Y}) = \sum_{i=1}^n \log(Y_i)$ é suficiente e $I_n(\eta) = \mathrm{Var}_\eta[T(\boldsymbol{Y})] = nA''(\eta) = \frac{n}{(\eta+1)^2}$, ou seja, $\mathrm{Var}_\theta[T(\boldsymbol{Y})] = \frac{n}{\theta^2}$. Para a informação de Fisher, vamos usar a seguinte propriedade: $\tilde{I}(\theta) = [\eta'(\theta)]^2 I(\eta)$. Logo, $I_n(\theta) = \frac{n}{\theta^2}$.

■

Comentário: Nesta questão nós sedimentamos os conhecimentos sobre a família exponencial relembrando alguns resultados clássicos e depois aplicando alguns deles à família Pareto. Note como certos resultados só valem para a família na forma canônica, mas é geralmente simples transformar as coisas para essa forma. *Sopa no mel*!

3. We shall prevail

Suponha que desejamos estimar a proporção $\theta \in (0,1)$ de indivíduos infectados com um determinado patógeno em uma população. Suponha ainda que dispomos de um teste laboratorial, que produz o resultados $R = \{-,+\}$ indicando se o indivíduo (Y_i) é livre (0) ou infectado (1).

a) (5 pontos) Se o teste fosse perfeito, mostre como poderíamos escrever a probabilidade de observar $y = \sum_{i=1}^{n} y_i$ testes positivos em n testes realizados, $\Pr(Y = y \mid \theta, n)$. Assuma independência condicional entre os testes dado θ.

b) (10 pontos) Infelizmente, o teste não é perfeito, acertando o diagnóstico com probabilidades fixas da seguinte forma

$$\Pr(\ R = + \mid Y_i = 0) =: u, \tag{3}$$
$$\Pr(R = - \mid Y_i = 1) =: v. \tag{4}$$

Assumindo $u + v > 1$, mostre que sob esse modelo de erro de observação,

$$\Pr(R = + \mid \theta, u, v) = \theta(1-v) + (1-\theta)u. \tag{5}$$

c) (5 pontos) Escreva a verossimilhança em termos de $Z = \sum_{i=1}^{n} \mathbb{I}(R_i = +)$ neste novo modelo.

d) (10 pontos) Encontre o estimador de máxima verossimilhança (EMV) para θ.

e) (10 pontos) Escolha e justifique uma distribuição *a priori* para θ – lembre-se que neste exercício u e v são fixos. Além disso, deduza a posteriori $p(\theta \mid z, n, u, v)$.

Conceitos trabalhados: modelos probabilísticos; erro de observação; inferência.
Nível de dificuldade: médio.
Resolução: Para responder o item a, note que se o teste fosse perfeito, $\Pr(Y_i = 1) = \theta$. Logo, pela independência condicional,

$$\Pr(Y = y \mid \theta, n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Agora, observe que

$$\begin{aligned}\Pr(\ R = + \mid \theta, u, v) &= \Pr(\ R = + \mid Y_i = 0, \theta, u, v)\Pr(Y_i = 0 \mid \theta, u, v) + \\ &+ \Pr(\ R = + \mid Y_i = 1, \theta, u, v)\Pr(Y_i = 1 \mid \theta, u, v), \\ &= u(1-\theta) + (1-v)\theta.\end{aligned}$$

Para o item c, temos a soma de variáveis Bernoulli, condicionalmente independentes, cuja probalidade é dada pelo item anterior, ou seja,

$$\Pr\left(Z = z \mid \theta, n, u, v\right) = \binom{n}{z} \left[u + \theta(1-(u+v))\right]^{z} \left[1 - u - \theta(1-(u+v))\right]^{n-z}.$$

Para encontrar o EMV, considere a log-verossimilhança:

$$l(\theta) \propto z \log\left[u + \theta(1-(u+v))\right] + (n-z)\log\left[1 - u - \theta(1-(u+v))\right].$$

Assim,

$$l'(\theta) = \frac{z[1-(u+v)]}{u+\theta[1-(u+v)]} - \frac{(z-n)[1-(u+v)]}{1-u-\theta[1-(u+v)]},$$

ou seja,

$$l'(\hat{\theta}) = 0 \Leftrightarrow \bar{z} - u - \hat{\theta}[1-(u+v)] = 0 \Leftrightarrow \hat{\theta} = \frac{\bar{z}-u}{1-(u+v)}.$$

Por fim, como θ é uma proporção, uma escolha comum para a priori é $\theta \sim \text{Beta}(a,b)$, o que nos leva a posteriori

$$p(\theta \mid z, n, u, v) = \frac{\left[u + \theta(1-(u+v))\right]^{z} \left[1 - u - \theta(1-(u+v))\right]^{n-z} \theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \left[u + t(1-(u+v))\right]^{z} \left[1 - u - t(1-(u+v))\right]^{n-z} t^{a-1}(1-t)^{b-1}dt}.$$

■

Comentário: Nesta questão, trabalhamos a modelagem de um problema por primeiros princípios, usando a lei da probabilidade total para derivar a distribuição marginal dos dados e depois realizando inferência tanto sobre o ponto de vista frequentista quanto bayesiano. A estimação da prevalência é um problema importante em epidemiologia, com uma longa história. Aqui nós fixamos v e u, mas na prática essas quantidades também são incertas (ver Gelman & Carpenter, 2020) e recebem distribuições *a priori*. Além disso, também é possível fazer mais de um teste ou só retestar os positivos – ver Bastos, Carvalho & Gomes (2021) para uma revisão.

Bibliografia

Bastos, L. S., Carvalho, L. M., and Gomes, M. F. (2021). Modelling misreported data. In *Building a Platform for Data-Driven Pandemic Prediction*, pages 113–140. Chapman and Hall/CRC.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1269–1283.