Primeira avaliação (A1)

Disciplina: Modelagem Estatística
Instrutor: Luiz Max Carvalho
Monitores: Eduardo Adame & Ezequiel Braga

15 de abril de 2024

- 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.
- 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. All about that interaction, baby!

O cantor e compositor Crazy Frog está trabalhando em seu novo álbum, mas está encontrando dificuldades para escrever uma música que seja realmente original e cativante. Para solucionar esse problema, ele contratou uma equipe de estatísticos de primeira linha para ajudá-lo a tomar decisões baseadas em modelos. Entretanto, Crazy Frog faltou às aulas de Modelagem Estatística durante a graduação e precisa de ajuda para interpretar alguns dos diferentes modelos apresentados. Sua tarefa é auxiliar o Frog neste desafio.

Considere o modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

onde $X_1, X_2, \ldots, X_n$ são constantes fixas e ε_i são variáveis aleatórias independentes e identicamente distribuídas com distribuição Normal$(0, \sigma^2)$.

Considere agora a seguinte reparametrização:

$$Y_i = \alpha_0 + \alpha_1 \cdot (X_i - \bar{X}_n) + \varepsilon_i,$$

onde $\bar{X}_n = n^{-1} \sum_{i=1} X_i$. Sejam $\widehat{\beta_0}$ e $\widehat{\beta_1}$ os estimadores de máxima verossimilhança (EMVs) para os coeficientes do modelo original e $\widehat{\alpha_0}$ e $\widehat{\alpha_1}$ os EMVs para os coeficientes do modelo reparametrizado.

a) (5 pontos) Mostre que

   (a) $\widehat{\alpha_0} = n^{-1} \sum_{i=1} Y_i$ e que $\widehat{\beta_0} \neq \widehat{\alpha_0}$;

   (b) $\widehat{\alpha_1} = \widehat{\beta_1}$. O que isso significa em termos de interpretação desses coeficientes?

b) (10 pontos) Crazy indagou sobre a possibilidade de incluir mais uma das covariáveis que ele consolidou. Portanto, vamos supor agora que temos duas covariáveis contínuas, X_1 e X_2, e que o modelo agora é

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + \varepsilon_i.$$

   (a) Descreva em palavras a interpretação de cada coeficiente;

   (b) Para $j = 1, 2$, defina $m_j = n^{-1} \sum_{i=1}^{n} X_{j,i}$. Considere a transformação $X^c_{j,i} = X_{j,i} - m_j$ e escreva o modelo

$$Y_i = \beta^\star_0 + \beta^\star_1 X^c_{1,i} + \beta^\star_2 X^c_{2,i} + \beta^\star_3 X^c_{1,i} X^c_{2,i} + \varepsilon_i.$$

   A interpretação dos coeficientes mudou? Se sim, como?

c) (15 pontos) Uma análise interessante seria mapear os coeficientes do segundo modelo nos coeficientes do primeiro modelo. É possível mostrar que

$$\beta_1 = \beta^\star_1 - \beta^\star_3 m_2.$$

Encontre as expressões para β_0, β_2 e β_3.

Conceitos trabalhados: regressão linear, estimação, interação. **Nível de dificuldade**: fácil.

Resolução:

Para responder o item (a), sabemos que em um problema de regressão linear simples,

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x},$$
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Seja $X_i^c = X_i - \bar{X}_n$. Assim, é fácil notar que $\bar{x}_n^c = 0$ e, usando a fórmula anterior, segue que $\widehat{\alpha_0} = \bar{y}$ e $\widehat{\beta_1} = \widehat{\alpha_1}$. Isso significa que uma mudança de 1 em X_i ou X_i^c provoca o mesmo efeito em Y, em média.

Para o modelo do item (b), observe que β_0 representa o efeito sobre Y_i marginalmente, isto é, quando as outras covariáveis são nulas. Para os outros coeficientes, é importante considerar a seguinte reescrita do modelo:

$$Y_i = \beta_0 + (\beta_1 + \beta_3 X_{2,i})X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

Fica mais fácil perceber que β_2 representa o efeito de $X_{1,i}$ sobre Y_i quando $X_{2,i} = 0$. Analogamente, β_2 mede o efeito de $X_{2,i}$ sobre Y_i quando $X_{1,i} = 0$. Por fim, β_3 mede o efeito da interação de $X_{1,i}X_{2,i}$ sobre Y_i. Ao centralizar as covariáveis em torno da média, observe agora que a função dos coeficientes é a mesma, mas representam efeitos diferentes: $\beta_1^\star$ representa o efeito de $X_{1,i}^c$ sobre Y_i quando $X_{2,i}^c = 0$, ou seja, quando $X_{2,i} = m_j$. Porém, observe também que é equivalente dizer que $\beta_1^\star$ representa o efeito de $X_{1,i}^c$ sobre Y_i quando $X_{2,i}^c = \bar{X}_{2,i}$, já que $\bar{X}_{2,i} = 0$. A mesma ideia segue para os demais coeficientes.

Para responder o último item, vale reescrever o modelo da seguinte forma:

$$\begin{aligned} Y_i &= \beta_0^\star + \beta_1^\star(X_{1,i} - m_1) + \beta_2^\star(X_{2,i} - m_2) + \beta_3^\star(X_{1,i} - m_1)(X_{2,i} - m_2) + \varepsilon_i, \\ &= (\beta_0^\star - \beta_1^\star m_1 - \beta_2^\star m_2 + \beta_3^\star m_1 m_2) + (\beta_1^\star - \beta_3^\star m_2)X_{1,i} + \\ &+ (\beta_2^\star - \beta_3^\star m_1)X_{2,i} + \beta_3^\star X_{1,i}X_{2,i} \end{aligned}$$

Isto resulta no seguinte mapeamento:

$$\begin{aligned} \beta_0 &= \beta_0^\star - \beta_1^\star m_1 - \beta_2^\star m_2 + \beta_3^\star m_1 m_2, \\ \beta_1 &= \beta_1^\star - \beta_3^\star m_2, \\ \beta_2 &= \beta_2^\star - \beta_3^\star m_1, \\ \beta_3 &= \beta_3^\star. \end{aligned}$$

■

Comentário: Nesta questão bem simples, vimos o alguns resultados básicos de regressão linear, além de cobrar a interpretação dos coeficientes no caso centrado, que foi discutido em aula. Além disso, descobrimos como mapear os coeficientes entre um modelo onde as covariáveis estão centradas e um em que não estão. Descobrimos assim que o coeficiente do termo de interação não muda.

2. *Logito ergo sum*

Palmirinha tem recebido muitas reclamações sobre clientes que sentem dor de barriga depois de frequentar sua barraca de pamonha. Ela suspeita que o que está acontecendo é que a sua pamonha é muito gostosa, e por isso as pessoas comem demais e acabam passando mal. Assim, ela quer estudar o fenômeno para depois confeccionar uma placa com os dizeres "Não é aconselhável comer mais que $x^{\star\star}$ gramas de pamonha por dia". Nesta questão, vamos ajudar Palmirinha a determinar essa quantidade.

Suponha que Palmirinha dispõe de um conjunto de dados que consiste em n observações $Y_i \in \{0, 1\}$, $i = 1, \ldots, n$, onde $Y_i = 1$ se o indivíduo teve dor de barriga e $Y_i = 0$ caso contrário. Ela dispõe ainda da quantidade em gramas de pamonha que o i-ésimo indivíduo consumiu, X_i.

Ela escolhe começar sua modelagem com

$$Y_i \mid X_i \sim \text{Bernoulli}(\theta(X_i)), \tag{1}$$

mas ainda está em dúvida sobre como especificar a média condicional $\theta(X_i)$.

Ela considera dois modelos lineares generalizados:

$$\begin{aligned} \text{Logit}: \quad & \theta(X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}, \\ \text{Probit}: \quad & \theta(X_i) = \Phi(\beta_0 + \beta_1 X_i), \end{aligned}$$

onde $\Phi(x) = (2\pi)^{-1} \int_{-\infty}^{x} \exp(-t^2/2)\, dt$ é a função de distribuição acumulada (f.d.a. ou c.d.f.) de uma normal padrão.

Palmirinha ajustou os dois modelos a $n = 100$ observações e produziu a Figura 1. Já o output do R (resumido) ficou assim:

```
> summary(fit.logit)
Call:
glm(formula = Ys ~ doses, family = binomial("logit"))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.044785 0.841447 -4.807 1.53e-06 ***
doses       0.017837  0.003259  5.474 4.40e-08 ***

> summary(fit.probit)
Call:
glm(formula = Ys ~ doses, family = binomial("probit"))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.216496 0.408629 -5.424 5.82e-08 ***
doses       0.009833  0.001551  6.338 2.32e-10 ***
```

a) (10 pontos) Mostre que o modelo em (1) pertence à família exponencial canônica.

b) (20 pontos) Considere $\theta_{\text{logit}}(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$ e $\theta_{\text{probit}}(x) = \Phi(\beta_0+\beta_1 x)$. Vamos encará-los como curvas de probabilidades.

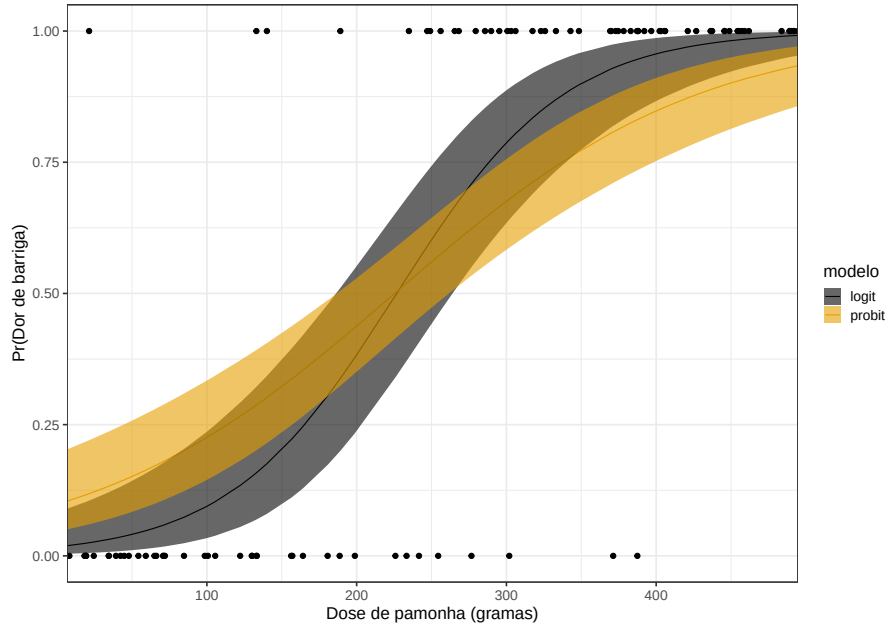

Figura 1: Dor de barriga *vs* quantidade de pamonha consumida. Pontos pretos mostram os dados coletados. Linhas coloridas mostram a predição no espaço da probabilidade e as bandas mostram o intervalo de predição de 95%.

i. Exiba $\theta'(x) := \frac{\partial}{\partial x}\theta(x)$ para cada um dos modelos acima.
Dica: Note que no nosso caso, o inverso da função de ligação g^{-1} pode ser interpretado como a c.d.f. de uma distribuição contínua e simétrica. Aplique a regra da cadeia.

ii. Uma quantidade importante é $x^\star$ tal que $\theta(x^\star) = 1/2$. Como os dois modelos serão ajustados no mesmo conjunto de dados, esperamos que $\theta'_{\text{logit}}(x^\star) \approx \theta'_{\text{probit}}(x^\star)$. Sob essa premissa, podemos computar

$$\frac{\hat{\beta}_1^{\text{ logit}}}{\hat{\beta}_1^{\text{ probit}}} \approx a.$$

Determine o valor (aproximado) de a e veja se ele está de acordo com os resultados empíricos obtidos por Palmirinha.

c) (10 pontos) Suponha que Palmirinha quer computar $x^{\star\star}$ de modo que $\theta(x^{\star\star}) = 0,8$, isto é, ela quer determinar a dose que faz com que 80% daqueles que comem $x^{\star\star}$ gramas de pamonha ou mais terem dor de barriga. Mostre a Palmirinha como computar essa quantidade sob cada um dos modelos.

Conceitos trabalhados: modelos lineares generalizados. **Nível de dificuldade**: médio.
Resolução: Como visto em aula, uma variável aleatória $\boldsymbol{X}$ está na família exponencial canônica se sua p.d.f. é dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left[\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right].$$

Para a Bernoulli do modelo em (1), temos

$$\begin{aligned} f_{Y_i|X_i}(y;\theta(X_i)) &= \theta(X_i)^y(1-\theta(X_i))^{1-y}, \\ &= \exp\left[y\log\theta(X_i) + (1-y)\log(1-\theta(X_i))\right], \\ &= \exp\left[y\log\frac{\theta(X_i)}{1-\theta(X_i)} + \log(1-\theta(X_i))\right]. \end{aligned}$$

Logo, segue que $h(y) = 1$, $\eta = \log\frac{\theta(X_i)}{1-\theta(X_i)}$ e $A(\eta) = \log(1+\exp(\eta))$.
Para o item (b), observe que usando as propriedades da função logit, temos

$$\begin{aligned} \theta'(x) &= \frac{\partial}{\partial x}(1+\exp(-(\beta_0+\beta_1 x))^{-1}, \\ &= (1+\exp(-\beta_0-\beta_1 x))^{-2}\exp(-\beta_0-\beta_1 x)\beta_1, \\ &= \theta(x)(1-\theta(x))\beta_1. \end{aligned}$$

Analogamente, para o modelo probit, $\theta'(x) = \phi(\beta_0+\beta_1 x)\beta_1$, onde ϕ é a p.d.f. da normal padrão. Para computar o valor de a, basta usar a premissa de que $\theta'_{\text{logit}}(x^\star) \approx \theta'_{\text{probit}}(x^\star)$ e que $\theta(x^\star) = 1/2$, ou seja, $\beta_0 + \beta_1 x^\star = \Phi^{-1}(1/2) = 0$. Daí, segue que $\theta'_{\text{probit}}(x^\star) = \phi(0)\beta_1^{\text{probit}} = 1/\sqrt{2\pi}\beta_1^{\text{probit}}$ e $\theta'_{\text{logit}}(x') = 1/4\beta_1^{\text{logit}}$. Logo,

$$1/4\beta_1^{\text{logit}} \approx 1/\sqrt{2\pi}\beta_1^{\text{probit}},$$

isto é, $a = 2\sqrt{\frac{2}{\pi}} \approx 1.6$. Se olharmos no *output*, temos $0.017837/0.009833 \approx 1.81$, o que não está longe do valor teórico. Para computar $x^{\star\star}$, basta calcular

$$\begin{aligned} \theta(x^{\star\star}) &= \frac{1}{1+\exp(\beta_0+\beta_1 x^{\star\star})} = 0.8, \\ \theta(x^{\star\star}) &= \Phi(\beta_0+\beta_1 x^{\star\star}) = 0.8. \end{aligned}$$

Isso nos dá $x^{\star\star}_{\text{logit}} = \frac{-\log(4)-\beta_0}{\beta_1}$ e $x^{\star\star}_{\text{probit}} = \frac{\Phi^{-1}(0.8)-\beta_0}{\beta_1}$.
■

Comentário: Nesta questão vimos como um GLM para uma variável resposta binária pode ser construído com diferentes funções de ligação. Além disso, nos aproveitamos do fato de que as curvas são parecidas em um particular ponto para obter uma razão aproximada entre os coeficientes "angulares"dos dois modelos. Então da próxima vez que você vir um coeficiente estimado sob um modelo *logit* e se perguntar "Qual seria a estimativa sob um modelo *probit*?", você já sabe como converter aproximadamente. Vimos como utilizar o GLM ajustado para responder a perguntas sobre que dose de *pamonha da Palmirinha* ™ causa dor de barriga em 80% da amostra?

3. Your AIC ain't that big...

A comparação de modelos é uma das principais preocupações da Estatística aplicada. Para um modelo paramétrico $\mathcal{M}$ qualquer, o *Akaike Information Criterion* (AIC) é dado por

$$\text{AIC}_{\mathcal{M}} = 2k - 2\log(\hat{L}),$$

onde k é número de parâmetros de $\mathcal{M}$ e $\hat{L} = f(\boldsymbol{z}; \hat{\theta})$ é a verossimilhança avaliada no estimador de máxima verossimilhança dos parâmetros θ, $\hat{\theta}$.

Tome $\boldsymbol{X}$ uma matriz real $n \times P$ e $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}^T \in \mathbb{R}^n$ um vetor contendo os valores da variável dependente. Considere o modelo

$$E[Y_i] =: \mu_i(\boldsymbol{\beta}) = \boldsymbol{X}_i\boldsymbol{\beta},$$

onde $\boldsymbol{\beta} \in \mathbb{R}^P$ é o vetor de coeficientes e parâmetro de interesse. Para completar a especificação do modelo, vamos assumir que os erros em torno da média condicional (preditor linear) são normalmente distribuídos com variância comum:

$$\begin{aligned} Y_i &= \mu_i(\boldsymbol{\beta}) + \varepsilon_i \\ \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2), \end{aligned}$$

com $\sigma^2 \in \mathbb{R}_+$ desconhecida.

a) (10 pontos) Exiba a expressão do AIC para um modelo linear com P coeficientes, que vamos chamar de modelo $\mathcal{M}_1$.

b) (10 pontos) Suponha agora que temos um conjunto de covariáveis que inclui as P do item anterior e mais Q covariáveis. Chamemos este modelo de $\mathcal{M}_2$. Mostre que

$$\text{AIC}_{\mathcal{M}_2} < \text{AIC}_{\mathcal{M}_1} \iff \frac{\text{SSE}_2}{\text{SSE}_1} < \exp\left(-2Q/n\right),$$

onde SSE_i é a soma de erros quadráticos do modelo $\mathcal{M}_i$.

c) (10 pontos) Uma quantidade importante na avaliação de modelos lineares é o coeficiente de determinação, R^2. Mostre que

$$\text{AIC}_{\mathcal{M}_2} < \text{AIC}_{\mathcal{M}_1} \iff R_2^2 - R_1^2 > \left(1 - \exp\left(-2Q/n\right)\right) \frac{\text{SSE}_1}{\text{SSE}_0},$$

onde SSE_0 é a soma dos erros quadráticos de um modelo só com intercepto.

Conceitos trabalhados: regressão múltipla, coeficiente de determinação, máxima verossimilhança.
Nível de dificuldade: médio .
Resolução: Para o item (a), basta observar que $\boldsymbol{Y} \sim \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$. Assim, segue que

$$L_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Como σ^2 também é um parâmetro, temos $P+1$ parâmetros no modelo. Além disso, sabemos que o MLE de σ é dado por $\widehat{\sigma} = n^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$. Logo, substituindo na log-verossimilhança, temos

$$\begin{aligned}\text{AIC}_{\mathcal{M}_1} &= 2(P+1) + n\log(2\pi\widehat{\sigma}^2) + \frac{1}{\widehat{\sigma}^2}n\widehat{\sigma}^2,\\ &= 2(P+1) + n[\log(2\pi\widehat{\sigma}^2)+1].\end{aligned}$$

Para o item (b), tome $\widehat{\sigma}_i^2$ o MLE de σ^2 do modelo $\mathcal{M}_i$. Assim, usando o resultado anterior, temos

$$\begin{aligned}\text{AIC}_{\mathcal{M}_2} - \text{AIC}_{\mathcal{M}_1} &= 2(P+Q+1) + n[\log(2\pi\widehat{\sigma}_2^2)+1] - 2(P+1) - n[\log(2\pi\widehat{\sigma}_1^2)+1],\\ &= 2Q + n\log\frac{\widehat{\sigma}_2^2}{\widehat{\sigma}_1^2}.\end{aligned}$$

Observando que $\widehat{\sigma}_i^2 = n^{-1}\text{SSE}_i$, a fórmula acima equivale a $2Q + n\log\frac{\text{SSE}_2}{\text{SSE}_1}$. Logo,

$$\begin{aligned}\text{AIC}_{\mathcal{M}_2} < \text{AIC}_{\mathcal{M}_1} &\iff 2Q + n\log\frac{\text{SSE}_2}{\text{SSE}_1} < 0,\\ &\iff -\frac{2Q}{n} > \log\frac{\text{SSE}_2}{\text{SSE}_1},\\ &\iff \exp\left[-\frac{2Q}{n}\right] > \frac{\text{SSE}_2}{\text{SSE}_1},\end{aligned}$$

onde a última equivalência decorre do fato de que a função exp é monotônica não decrescente.

Para o último item, basta notar que $R_i^2 = 1 - \frac{\text{SSE}_i}{\text{SSE}_0}$. Desse modo, $R_2^2 - R_1^2 = \frac{\text{SSE}_1 - \text{SSE}_2}{\text{SSE}_0}$. Logo,

$$\begin{aligned}R_2^2 - R_1^2 > \left(1-\exp\left(-2Q/n\right)\right)\frac{\text{SSE}_1}{\text{SSE}_0} &\iff \frac{\text{SSE}_1 - \text{SSE}_2}{\text{SSE}_0} > \left(1-\exp\left(-2Q/n\right)\right)\frac{\text{SSE}_1}{\text{SSE}_0},\\ &\iff 1 - \frac{\text{SSE}_2}{\text{SSE}_1} > \left(1-\exp\left(-2Q/n\right)\right),\\ &\iff \frac{\text{SSE}_2}{\text{SSE}_1} < \left(\exp\left(-2Q/n\right)\right).\end{aligned}$$

■

Comentário: Nesta questão estamos olhando indicadores de bondade do ajuste (*goodness-of-fit*) no modelo de regressão linear múltipla. O AIC foi pensando como uma medida da capacidade preditiva do modelo fora da amostra, e procura penalizar a verossimilhança máxima atingida pelo tamanho do modelo em questão. Já o coeficiente de determinação procura medir a quantidade de variância presente na variável resposta (dependente) que é explicada pelo modelo comparado com um modelo que só tem o intercepto (i.e. um modelo i.i.d.). Nós estabelecemos uma equivalência entre a diferença nos coeficientes de determinação e o AIC de um modelo ser menor que o outro – o que levaria à sua preferência.