Advanced Econometrics

Mestrado Profissional em Economia

Insper Instituto de Ensino e Pesquisa

Prof. Hedibert Freitas Lopes

Lista de Exercícios 2

Data de entrega via blackboard: 23 de novembro de 2026 (até 19h30)
Cada grupo de trabalho pode conter até 4 estudantes

Dados

Os dados vêm de um estudo com uma amostra aleatória de $n = 40$ adultos residentes no condado de Alachua, Flórida (EUA), que investigou a relação entre índices de saúde mental e algumas variáveis explicativas. As variáveis são:

- y_i = `impair`: índice de *comprometimento mental*, que agrega diversas dimensões de sintomas psiquiátricos (incluindo aspectos de ansiedade e depressão). Valores maiores indicam maior gravidade. Na amostra varia de 17 a 41, com $\bar{y} = 27{,}30$ e $s_y = 5{,}46$.
- x_{1i} = `life`: índice de *eventos de vida*, medida composta do número e da gravidade de eventos marcantes vividos pelo indivíduo nos últimos três anos (por exemplo, morte na família, divórcio, perda de emprego, mudança de residência). Na amostra varia de 3 a 97, com $\bar{x}_1 = 44{,}42$ e $s_{x_1} = 22{,}62$.
- x_{2i} = `ses`: índice de *status socioeconômico* (SES), composto a partir de ocupação, renda e escolaridade, numa escala de 0 a 100. Na amostra varia de 3 a 97, com $\bar{x}_2 = 56{,}60$ e $s_{x_2} = 25{,}28$.

Fonte: Agresti, A. *et al.* (2026), *Foundations of Bayesian Statistics for Data Scientists*, Taylor & Francis, Seção 4.1.5 (arquivo `Mental.dat`); originalmente em Agresti & Finlay, *Statistical Methods for the Social Sciences* (Tabela 11.1; dados cedidos por Dr. Charles Holzer).

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
17	46	84	24	33	67	27	60	70	31	95	29
19	39	97	24	18	39	28	97	89	31	63	53
20	27	24	25	81	87	28	37	50	31	42	7
20	3	85	26	22	95	28	30	90	32	38	32
20	10	15	26	50	40	28	13	56	33	45	55
21	44	55	26	48	52	28	40	56	34	70	58
21	37	78	26	45	61	29	5	40	34	57	16
22	35	91	27	21	45	30	59	72	34	40	29
22	78	60	27	55	88	30	44	53	41	49	3
23	32	74	27	45	56	31	35	38	41	89	75

(A tabela deve ser lida por colunas de blocos; a ordem das linhas é irrelevante para a análise.)

Para facilitar, os dados em formato `R` (na ordem do arquivo original):

```
impair <- c(17,19,20,20,20,21,21,22,22,23,24,24,25,26,26,26,26,27,27,27,
          27,28,28,28,28,28,29,30,30,31,31,31,31,32,33,34,34,34,41,41)
life <- c(46,39,27, 3,10,44,37,35,78,32,33,18,81,22,50,48,45,21,55,45,
          60,97,37,30,13,40, 5,59,44,35,95,63,42,38,45,70,57,40,49,89)
ses <- c(84,97,24,85,15,55,78,91,60,74,67,39,87,95,40,52,61,45,88,56,
          70,89,50,90,56,56,40,72,53,38,29,53, 7,32,55,58,16,29, 3,75)
Mental <- data.frame(impair, life, ses)
n <- nrow(Mental) # 40
# Alternativamente:
# Mental <- read.table("http://bayes4ds.rwth-aachen.de/data/Mental.dat", header=TRUE)
```

Modelos

Para facilitar a elicitação das prioris, trabalhe com as variáveis explicativas *padronizadas*, $z_{ji} = (x_{ji} - \bar{x}_j)/s_{x_j}$, $j = 1, 2$. Considere os seis modelos abaixo, em que β_0 é o nível médio de y e cada β_j é o efeito (em pontos do índice de comprometimento) de um desvio-padrão a mais em x_j:

	Erros Gaussianos	Erros Student-t
Só `life`	$M_1:\ y_i = \beta_0 + \beta_1 z_{1i} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$	$M_4:$ idem, com $\varepsilon_i \sim t_\nu(0, \sigma^2)$
Só `ses`	$M_2:\ y_i = \beta_0 + \beta_2 z_{2i} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$	$M_5:$ idem, com $\varepsilon_i \sim t_\nu(0, \sigma^2)$
`life` e `ses`	$M_3:\ y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$	$M_6:$ idem, com $\varepsilon_i \sim t_\nu(0, \sigma^2)$

Nos modelos M_4–M_6 o grau de liberdade ν é *desconhecido* e deve ser aprendido a partir dos dados. Lembre que a t_ν admite a representação como mistura de escala de normais: $\varepsilon_i \mid \omega_i \sim N(0, \sigma^2/\omega_i)$, com $\omega_i \sim \text{Gama}(\nu/2, \nu/2)$.

Prioris (informativas, mas não muito concentradas). Use, em todos os modelos, as mesmas prioris para os parâmetros comuns:

$$\beta_0 \sim N(25,\ 5^2), \qquad \beta_j \sim N(0,\ 3^2)\ (j = 1, 2), \qquad \sigma^2 \sim \text{GI}(3,\ 50), \qquad \nu \sim \text{Gama}(2,\ 0{,}1),$$

todas independentes. A priori de β_0 diz que o nível médio do índice está quase certamente entre 15 e 35; a de β_j, que efeitos de 1 desvio-padrão de magnitude superior a 6 pontos são pouco plausíveis; a GI$(3, 50)$ tem média $E(\sigma^2) = 25$ (isto é, $\sigma \approx 5$); e a Gama$(2; 0{,}1)$ (Juárez & Steel, 2010) tem média 20 e dá peso razoável tanto a caudas pesadas ($\nu < 10$) quanto a caudas próximas da normal ($\nu > 30$). Vocês podem propor outras prioris, desde que justifiquem e as mantenham comuns aos seis modelos.

Exercícios

1. Análise exploratória e correlações.

(a) Reporte média, mediana, desvio-padrão, mínimo e máximo de y, x_1 e x_2, e faça os histogramas das três variáveis.

(b) Faça os diagramas de dispersão $y \times x_1$, $y \times x_2$ e $x_1 \times x_2$ (por exemplo, `pairs(Mental)`).

(c) Calcule as correlações de Pearson $r(y, x_1)$, $r(y, x_2)$ e $r(x_1, x_2)$. Para cada uma, construa um intervalo de confiança de 95% (via transformação z de Fisher) e teste $H_0 : \rho = 0$. Compare também com a correlação de Spearman.

(d) Interprete sinais e magnitudes. A correlação entre as duas variáveis explicativas é forte o suficiente para gerar preocupação com multicolinearidade? O que você espera, a partir das correlações, sobre os coeficientes do modelo com ambas as variáveis quando comparados aos dos modelos com uma variável só?

2. Comparação de modelos via fator de Bayes.

(a) Escreva a verossimilhança dos modelos Gaussianos (M_1–M_3) e dos modelos Student-t (M_4–M_6). Para estes últimos, escreva também a verossimilhança aumentada pelas variáveis latentes $\omega_1, \ldots, \omega_n$.

(b) Obtenha amostras da distribuição a posteriori de cada um dos seis modelos via MCMC (por exemplo, Gibbs para M_1–M_3; Gibbs com os ω_i e um passo Metropolis–Hastings para ν em M_4–M_6; ou `Stan`/`brms`). Reporte média, desvio-padrão e intervalo de credibilidade de 95% de todos os parâmetros, e verifique a convergência das cadeias (traços, autocorrelações, tamanho efetivo).

(c) Para os modelos M_4–M_6, compare graficamente a priori e a posteriori de ν. Os dados trazem informação sobre ν? Há evidência de caudas pesadas?

(d) Estime a verossimilhança marginal (preditiva a priori)

$$p(y \mid M_k) = \int p(y \mid \theta_k, M_k)\, p(\theta_k \mid M_k)\, d\theta_k,$$

para $k = 1, \ldots, 6$, por pelo menos um método adequado (por exemplo, amostragem por importância com proposta t multivariada centrada na moda a posteriori, *bridge sampling* (pacote `bridgesampling`), método de Chib, ou aproximação de Laplace). Evite o estimador da média harmônica e comente por quê. *Dica:* trabalhe sempre na escala logarítmica.

(e) Monte uma tabela com $\log p(y \mid M_k)$, os log-fatores de Bayes $\log B_{k*} = \log p(y \mid M_k) - \log p(y \mid M_*)$ em relação ao melhor modelo M_*, e as probabilidades a posteriori dos modelos supondo $P(M_k) = 1/6$. Interprete usando a escala de Kass & Raftery (1995).

(f) Responda: (i) as duas variáveis explicativas são necessárias? (ii) erros Student-t são preferidos aos Gaussianos?

(g) *Sensibilidade à priori.* Repita o item (e) multiplicando por 3 os desvios-padrão a priori de β_0, β_1 e β_2. Os fatores de Bayes mudam? E as posterioris dos coeficientes? Relacione com o paradoxo de Lindley–Bartlett.

3. Comparação de modelos via validação cruzada *leave-one-out*. Para cada observação $i = 1, \ldots, 40$ e para cada modelo M_k, ajuste o modelo com as prioris do item 2 usando apenas os dados sem a observação i, $y_{(-i)}$, e obtenha a distribuição preditiva a posteriori de y_i:

$$p(y_i \mid y_{(-i)}, M_k) = \int p(y_i \mid \theta_k, M_k)\, p(\theta_k \mid y_{(-i)}, M_k)\, d\theta_k \approx \frac{1}{G} \sum_{g=1}^{G} p\big(y_i \mid \theta_k^{(g)}, M_k\big),$$

em que $\theta_k^{(1)}, \ldots, \theta_k^{(G)}$ são amostras de $p(\theta_k \mid y_{(-i)}, M_k)$. Isso requer $6 \times 40 = 240$ ajustes (pense em como reaproveitar código e paralelizar).

(a) Para cada modelo, calcule os três critérios abaixo, avaliados na observação deixada de fora:

- *Erro quadrático médio:* $\mathrm{MSE}_k = \frac{1}{n} \sum_i \big(y_i - \hat{y}_{i,k}\big)^2$, com $\hat{y}_{i,k} = E(y_i \mid y_{(-i)}, M_k)$;
- *Erro absoluto médio:* $\mathrm{MAE}_k = \frac{1}{n} \sum_i \big|y_i - \tilde{y}_{i,k}\big|$, com $\tilde{y}_{i,k}$ a mediana de $p(y_i \mid y_{(-i)}, M_k)$;
- *Escore logarítmico preditivo:* $\mathrm{LPD}_k = \sum_i \log p(y_i \mid y_{(-i)}, M_k)$ (quanto maior, melhor).

Por que faz sentido usar a média preditiva no MSE e a mediana preditiva no MAE?

(b) Apresente uma tabela com MSE, MAE e LPD dos seis modelos e o ordenamento dos modelos segundo cada critério. Os três critérios concordam entre si?

(c) Faça um gráfico das contribuições individuais $\log p(y_i \mid y_{(-i)}, M_6) - \log p(y_i \mid y_{(-i)}, M_3)$ contra i (ou contra y_i). Quais observações mais favorecem os erros Student-t? Elas aparecem como atípicas nos gráficos do item 1?

(d) Compare as conclusões da validação cruzada com as do fator de Bayes (item 2). Lembre que $\exp(\mathrm{LPD}_k)$ é o produto das ordenadas preditivas condicionais (CPO) e que $\exp(\mathrm{LPD}_k - \mathrm{LPD}_l)$ é o *pseudo-fator de Bayes* de Geisser & Eddy (1979). Por que os dois enfoques podem discordar? Qual deles é mais sensível à priori? Repita o cálculo do LPD com as prioris mais difusas do item 2(g) e comente.

(e) *(Para os mais energizados)* Em vez de reajustar o modelo 40 vezes, aproxime $p(y_i \mid y_{(-i)}, M_k)$ a partir de um único ajuste com todos os dados, usando a identidade $p(y_i \mid y_{(-i)}) = \left\{ E[\, 1/p(y_i \mid \theta) \mid y \,] \right\}^{-1}$ (Gelfand, Dey & Chang, 1992), ou o PSIS-LOO do pacote `loo` (Vehtari, Gelman & Gabry, 2017). Compare com os valores exatos do item (b) e comente as diferenças observação a observação.

Observações. Entregue um relatório com gráficos, tabelas de estimativas e comentários, acompanhado do código utilizado (`R`, `Python` ou outra linguagem de sua preferência). Fixe a semente do gerador de números aleatórios e reporte o número de iterações, o *burn-in* e o tamanho efetivo das cadeias.

Referências. Gelfand, Dey & Chang (1992), *Bayesian Statistics 4*, 147–167. Geisser & Eddy (1979), *JASA* 74, 153–160. Juárez & Steel (2010), *JBES* 28, 52–66. Kass & Raftery (1995), *JASA* 90, 773–795. Vehtari, Gelman & Gabry (2017), *Statistics and Computing* 27, 1413–1432.