

Estudos e aplicações da distribuição normal multivariada em modelos de regressão linear

Studies and applications of the multivariate normal distribution in linear regression models

RESUMO

Aline Beatriz Noveli
alinenoveli@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procopio, Paraná, Brasil.

Roberto Molina de Souza
rmolinasouza@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procopio, Paraná, Brasil.

O uso de análise multivariada tem grande aplicação nas análises de dados mais complexas encontrada recentemente na literatura. Um dos fatores pode ser atribuído ao uso da análise Bayesiana que dá flexibilidade à construção de novos modelos. Muitos problemas em que tratava duas variáveis possivelmente correlacionadas como independentes, hoje podem ser resolvidos sem grandes custos computacionais utilizando-se uma distribuição bivariada quando os pressupostos desta distribuição são garantidos. O objetivo deste trabalho é apresentar quatro modelos bivariados distintos para regressão linear múltipla, baseados na distribuição normal, sob o enfoque Bayesiano, e aplicá-los em um conjunto de dados reais. Como resultado, os modelos bivariados propostos se ajustaram melhor aos dados analisados quando comparados ao modelo que pressupõe independência das variáveis respostas.

PALAVRAS-CHAVE: Análise multivariada. Análise de regressão. Análise bayesiana.

ABSTRACT

The use of multivariate analysis has great application in the more complex data analysis found recently in the literature. One of the factors can be attributed to the use of Bayesian analysis, which gives flexibility to the construction of new models. Many problems in which two possibly correlated variables were treated as independent, today can be solved without great computational costs using a bivariate distribution when the assumptions of this distribution are guaranteed. The objective of this work is to present four different bivariate models for multiple linear regression, based on the normal distribution, under the Bayesian approach, and to apply them to a set of real data. As a result, the proposed bivariate models adjusted better to the data analyzed when compared to the model that assumes independence of the response variables.

KEYWORDS: Multivariate analysis. Regression analysis. Bayesian analysis.

Recebido: 19 ago. 2020.

Aprovado: 01 out. 2020.

Direito autoral: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.



INTRODUÇÃO

As técnicas de análises estatística multivariada vem ganhando um amplo espaço em diversos estudos e aplicações nas mais variadas áreas de conhecimento. Entre um dos casos particulares mais utilizados está a distribuição normal bivariada que toma por base duas variáveis aleatórias, ambas com comportamento normal. Este modelo permite a análise simultânea de duas variáveis aleatórias estando elas correlacionadas ou não.

Na área da saúde, por exemplo, pode-se ter o interesse em estudar os níveis de pressão arterial sistólica (PAS) e diastólica (PAD) de um indivíduo. Neste caso, estas variáveis poderiam ser estudadas de forma conjunta, uma vez que as duas foram oriundas do mesmo indivíduo (MATTEI et al., 2019). Na comparação de dois fármacos, em um estudo de bioequivalência, por exemplo, cada indivíduo fornece duas medidas, sendo uma que diz respeito a absorção do medicamento e outra a concentração máxima obtida (SOUZA et al., 2009 e SOUZA et al., 2016).

O objetivo geral deste trabalho é estudar a distribuição normal bivariada aplicada a modelos de regressão linear, bem como a aplicação deste modelo em um conjunto de dados reais. Os objetivos específicos são: estudar a aplicação da distribuição normal bivariada em modelos de regressão linear; estudar a inferência Bayesiana e seus aspectos computacionais (Software JAGS e R); obter um conjunto de dados reais para modelagem estatísticas e aplicação dos modelos e métodos de estimação estudados; e analisar os dados, comparar os modelos e concluir.

METODOLOGIA

Nesta seção serão descritos quatro modelos que levam em conta a possível dependência entre duas variáveis aleatórias. Os mesmos podem ser descritos em (SOUZA et al., 2009 e SOUZA et al., 2016).

Sejam Y_1 e Y_2 duas variáveis aleatórias possivelmente correlacionadas e com distribuição aproximadamente normal, ou seja, $Y_j \sim N(\mu_j; \sigma_j^2)$, $j=1,2$, em que μ denota a média e σ^2 a variância de uma distribuição normal.

Considerando uma amostra aleatória, independente e identicamente distribuída de n indivíduos, com $i=1,2,\dots,n$, sob o enfoque Bayesiano, alguns modelos conjuntos que capturem a possível dependência entre Y_1 e Y_2 serão apresentados abaixo:

Modelo 1:

$$Y_{1i} = \beta_{01} + \beta_{11}X_{1i} + \dots + \beta_{k1}X_{ki} + e_{1i} \quad (1)$$

$$Y_{2i} = \beta_{02} + \beta_{12}X_{1i} + \dots + \beta_{k2}X_{ki} + e_{2i}$$

em que β_{0j} é o intercepto do modelo referente a variável Y_j ($j = 1,2$); $\beta_{1j}, \dots, \beta_{kj}$ são os parâmetros que representam a influência de $X_1, \dots, X_k$, na variável aleatória Y_j . Finalmente, $e_i \sim N_2(0, \Sigma_e)$ sendo $\Sigma_e = \begin{pmatrix} \sigma_{1e}^2 & \sigma_{12e} \\ \sigma_{12e} & \sigma_{2e}^2 \end{pmatrix}$, em que $N_2(0, \Sigma_e)$ denota uma distribuição normal bivariada com vetor de médias $\mathbf{0}$ e matriz de variância / covariâncias Σ_e . Logo, no Modelo 1, a possível correlação entre Y_1 e Y_2 está sendo capturada nos resíduos, e pode ser calculada por:

$$\rho_e = \frac{\sigma_{12e}}{\sqrt{\sigma_{1e}^2 \sigma_{2e}^2}} \quad (2)$$

Modelo 2: Este modelo é análogo ao Modelo 1, porém com uma componente que considera os parâmetros regressores correlacionados, ou seja:

$$\beta_0 \sim N_2 \left(\begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix}, \Sigma_{\beta_0} \right) \quad (3)$$

$$\beta_1 \sim N_2 \left(\begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix}, \Sigma_{\beta_1} \right)$$

⋮

$$\beta_k \sim N_2 \left(\begin{pmatrix} \beta_{k1} \\ \beta_{k2} \end{pmatrix}, \Sigma_{\beta_k} \right)$$

Modelo 3: Uma vez que as variáveis Y_1 e Y_2 são possivelmente correlacionadas, uma alternativa seria utilizar efeitos aleatórios, estabelecendo a correlação a partir destes efeitos. Logo, o modelo pode ser escrito na forma:

$$\mu_{1i} = \beta_{01} + \beta_{11}X_1 + \dots + \beta_{k1}X_k + \alpha_{1i} \quad (4)$$

$$\mu_{2i} = \beta_{02} + \beta_{12}X_1 + \dots + \beta_{k2}X_k + \alpha_{2i}$$

em que $\alpha_i \sim N_2(0, \Sigma_\alpha)$ sendo $\Sigma_\alpha = \begin{pmatrix} \sigma_{1\alpha}^2 & \sigma_{12\alpha} \\ \sigma_{12\alpha} & \sigma_{2\alpha}^2 \end{pmatrix}$. Observe que o Modelo 3 não é igual ao Modelo 1, uma vez que no Modelo 1 a distribuição imposta para as variáveis aleatórias Y_1 e Y_2 é uma normal bivariada, enquanto para o Modelo 3, as distribuições impostas para as variáveis aleatórias Y_1 e Y_2 são distribuições normais univariadas.

Modelo 4: No Modelo 4, considera-se a estratégia proposta no modelo 3, combinada com a componente que correlaciona as variáveis regressoras já introduzidas no Modelo 2.

Os detalhes referentes a estimação dos parâmetros dos modelos propostos no contexto da inferência Bayesiana podem ser vistos em Achcar et al. (2019). A implementação dos modelos propostos será feita utilizando o Software livre JAGS (PLUMMER, 2003) a partir do pacote *R2jags* disponível no Software R Core Team (2020). O uso do Software livre JAGS requer apenas importação do conjunto de dados, especificação das distribuições imposta aos dados e distribuições a priori para os parâmetros, bem como os hiperparâmetros destas distribuições. Quando a escolha dos hiperparâmetros leva a distribuições a priori não informativas, as estimativas dos parâmetros se aproximam às estimativas obtidas com os estimadores de máxima verossimilhança, por exemplo.

APLICAÇÃO

Os dados utilizados para aplicação dos modelos propostos referem-se a um estudo publicado por Mattei et al. (2019). Trata-se de um estudo controlado randomizado, com seguimento de 12 meses, que foi realizado em uma clínica de saúde primária no estado do Paraná. Pacientes adultos com hipertensão foram alocados aleatoriamente para intervenção ($n = 47$) e grupos de cuidados habituais ($n = 47$). O gerenciamento de casos de enfermagem inclui consultas de

enfermagem, contato telefônico, visitas domiciliares, educação em saúde e encaminhamentos adequados. Resultados do paciente (sangue pressão, índice de massa corporal, entre outras) foram avaliados no início e no seguimento de 6 e 12 meses no grupo intervenção e no início e no seguimento de 12 meses no grupo controle.

O objetivo da pesquisa de Mattei et al. (2019) é observar a possível redução das pressões sistólica e diastólica, obtidas de uma mesma unidade amostral. Logo, estes dados podem ser analisados a partir dos modelos propostos neste trabalho em que a variação nas pressões sistólica e diastólica serão as variáveis dependentes, enquanto o grupo onde o indivíduo foi alocado será a variável independente.

ANÁLISES INICIAIS

Foram analisadas as variáveis dependentes segundo a variável grupo. Pode-se observar que houve redução tanto na pressão arterial sistólica quanto na diastólica para os pacientes em estudo que fizeram parte do grupo intervenção. Para análise dos dados no modelo proposto, será utilizada a variação tanto da PAS quanto da PAD, para cada indivíduo, dada pela pressão aferida no início do estudo, subtraída da pressão aferida no final do estudo e dividida pela pressão inicial. Portanto, se a variação for negativa, implica que houve redução na pressão arterial.

MODELAGEM

Como o Software JAGS utiliza métodos de simulação do tipo *Gibbs Sampling* (GELFAND; SMITH, 1990) e *Metropolis Hastings* (CHIB; GREENBERG, 1995) baseado nos métodos de Monte Carlo em Cadeia de Markov (MCMC) (HASTINGS, 1970) para a obtenção das quantidades aleatórias (ou estimativas dos parâmetros) de interesse, foram geradas 55000 amostras para cada parâmetro de interesse. Destas, as primeiras 5000 foram excluídas para eliminar um possível efeito dos valores iniciais. Além disso, para eliminar uma possível correlação na série de valores gerados, eles foram tomados de 50 em 50, resultando assim em uma amostra final de 1000 valores para cada parâmetro. Estas amostras foram validadas a partir do uso de gráficos garantindo assim a convergência do algoritmo para todos os modelos.

Nas Tabelas 1 a 5 são apresentadas as estimativas dos parâmetros, seu erro padrão e intervalos com 95% de credibilidade. Foram estimados também os parâmetros para o Modelo 0, que não contempla qualquer tipo de correlação.

Tabela 1 – Estimativas dos parâmetros, desvio-padrão e intervalo com 95% de credibilidade para o Modelo 0.

Variável	Parâmetro	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	β_0	0,0157 (0,0181)	(-0,0202; 0,0519)
		-0,0710 (0,0261)	(-0,1214; -0,0212)
PAD	β_1	0,0083 (0,0217)	(-0,0346; 0,0489)
		-0,0820 (0,0311)	(-0,1421; -0,0212)

Fonte: Autoria própria (2020).

Tabela 2 – Estimativas dos parâmetros, desvio-padrão e intervalo com 95% de credibilidade para o Modelo 1.

Variável	Parâmetro	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	β_0	0,0149 (0,0189)	(-0,0239; 0,0512)
		-0,0704 (0,0269)	(-0,1213; -0,0214)
PAD	β_1	0,0068 (0,0220)	(-0,0364; 0,0506)
		-0,0805 (0,0323)	(-0,1415; -0,0147)
	ρ_e	0,6928 (0,0604)	(0,5637; 0,7956)

Fonte: Autoria própria (2020).

Tabela 3 – Estimativas dos parâmetros, desvio-padrão e intervalo com 95% de credibilidade para o Modelo 2.

Variável	Parâmetro	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	β_0	0,0147 (0,0180)	(-0,0203; 0,0503)
		-0,0694 (0,0262)	(-0,1205; -0,0188)
PAD	β_1	0,0065 (0,0221)	(-0,0384; 0,0495)
		-0,0795 (0,0319)	(-0,1417; -0,0162)
	ρ_e	0,6899 (0,0615)	(0,5463; 0,7930)
	ρ_{β_1}	0,0425 (0,6514)	(-0,9692; 0,9893)

Fonte: Autoria própria (2020).

Tabela 4 – Estimativas dos parâmetros, desvio-padrão e intervalo com 95% de credibilidade para o Modelo 3.

Variável	Parâmetro	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	β_0	0,0158 (0,0217)	(-0,0247; 0,0599)
		-0,0710 (0,0311)	(-0,1341; -0,0100)
PAD	β_1	0,0070 (0,0253)	(-0,0420; 0,0552)
		-0,0797 (0,0357)	(-0,1494; -0,0096)
	ρ_α	0,9569 (0,0505)	(0,7885; 0,9996)

Fonte: Autoria própria (2020).

Tabela 5 – Estimativas dos parâmetros, desvio-padrão e intervalo com 95% de credibilidade para o Modelo 4.

Variável	Parâmetro	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	β_0	0,0146 (0,0225)	(-0,0290; 0,0599)
		-0,0690 (0,0322)	(-0,1347; -0,0052)
PAD	β_1	0,0063 (0,0261)	(-0,0455; 0,0580)
		-0,0789 (0,0361)	(-0,1505; 0,0098)
	ρ_α	0,9632 (0,0389)	(0,8538; 0,9991)
	ρ_{β_1}	0,0678 (0,6532)	(-0,9637; 0,9890)

Fonte: Autoria própria (2020).

Na Tabela 6 são apresentados os valores de *DIC* (*Deviance Information Criterion*). Estes são importantes para fazer a discriminação dos modelos, ou seja, quando vários modelos são comparados, entende-se que o melhor modelo ajustado aos dados seja aquele que apresenta o menor valor de *DIC*.

Tabela 6 – Valores para *DIC* para os modelos ajustados.

Modelo	DIC
0	-229,5
1	-277,3
2	-277,3
3	-294,4
4	-293,1

Fonte: Autoria própria (2020).

Assim, pode-se tirar as seguintes conclusões a partir das Tabelas 1 a 6:

- Todos os modelos introduzidos neste trabalho são superiores ao Modelo 0, ou seja, aquele que analisa as variáveis separadamente. Os valores de *DIC* para os Modelos 1 a 4 são menores quando comparados ao Modelo 0 e os coeficientes de correlação ρ_α e ρ_e são diferentes de 0, uma vez que este valor não está incluído nos respectivos intervalos de credibilidade;
- Comparando os modelos propostos aos pares, ou seja, Modelos 1 e 2 e Modelos 3 e 4 não se observa ganhos quando da inclusão da correlação entre a covariável grupo no modelo, uma vez que os valores de *DIC* são praticamente os mesmos. Isto também pode ser observado no intervalo de credibilidade de ρ_{β_1} que inclui o valor 0.
- Finalmente, comparando o Modelo 1 com o Modelo 3, observa-se menor valor de *DIC* para o Modelo 3, indicando que capturar a correlação entre as variáveis a partir do efeito aleatório trouxe melhor ajuste do modelo aos dados.

Assim, tomando os dois modelos mais bem ajustados (Modelo 1 e Modelo 3) pode-se fazer os testes de hipóteses de interesse, ou seja, se existem evidências do efeito de grupo na redução da PAS e PAD. O parâmetro responsável por esta informação é o β_1 . Logo, o teste de hipótese proposto é na forma:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \quad (5)$$

Portanto, se a hipótese nula $\beta_1 = 0$ for rejeitada, ao nível de 0,05 de significância, tem-se evidências de que houve redução da pressão arterial para o grupo intervenção.

Utilizando os intervalos de credibilidade é bastante simples realizar este teste de hipóteses, bastando apenas observar se o valor 0 está contido nos intervalos de 95% de credibilidade construídos. Tanto para o Modelo 1 quanto para o Modelo 3 e para a PAS e PAD os intervalos de confiança para β_1 não contém o valor 0. Isto implica que existem evidências que o grupo intervenção foi eficaz na redução da pressão arterial ao nível de 0,05 de significância.

Uma consideração importante, observando-se os intervalos para β_1 na Tabela 2 e 4, está na amplitude deles. Observe que os intervalos de credibilidade para o Modelo 3 tem amplitude menor que para o Modelo 1. Isto provavelmente ocorra, uma vez que o Modelo 3 apresenta melhor ajuste que o Modelo 1.

Quanto à capacidade preditora dos modelos, pode-se estimar a redução da PAS e PAD de acordo com o grupo ao qual o indivíduo faça parte, utilizando a expressão $\hat{y}_j = \hat{\beta}_{0j} + \hat{\beta}_{1j}X_1$, para $j = 1, 2$. Logo, assumindo o Modelo 3, uma vez que o mesmo foi o que melhor se ajustou aos dados, observa-se na Tabela 7 que um indivíduo submetido ao grupo intervenção tem, em média, uma redução da PAS de 5,56%, variando de 1,25% a 9,96% e, na PAD, uma redução média de 7,3%, variando de 2,33% a 12,11%.

Tabela 7 – Preditores estimados e intervalos com 95% de credibilidade.

Variável	Estimativa (Desvio Padrão)	Intervalo de Credibilidade
PAS	-0,0556(0,0223)	(-0,0996; - 0,0125)
PAD	-0,0733 (0,0254)	(-0,1211; 0,0233)

Fonte: Autoria própria (2020).

CONCLUSÕES

A partir da Revisão da Literatura pode-se notar que o uso da distribuição normal bivariada ou multivariada são bastante úteis nos casos em que variáveis obtidas de uma mesma unidade amostral podem estar correlacionadas. Na Metodologia foi apresentada a distribuição normal bivariada e, a partir de seus parâmetros de localização e escala, introduzidos 4 modelos que foram implementados com o uso da inferência Bayesiana.

Na aplicação, utilizando o indicador DIC para a comparação de modelos, ficou evidente que no problema apresentado, assumir modelos que levam em conta a correlação entre as variáveis observadas, torna a análise superior, no sentido da adequação dos modelos bivariados quando comparado ao modelo normal univariado. Dentre os 4 modelos bivariados propostos, destacam-se aqueles em que a correlação foi modelada a partir dos efeitos aleatórios.

Finalmente, a construção de modelos que represente o problema de maneira fidedigna, comparando-os com modelos mais simples, traz mais segurança na decisão. Sendo assim, os modelos aqui introduzidos e resultados apresentados podem ser importantes não somente para área da saúde, mas também para áreas em que necessitam do uso da estatística aplicada.

REFERÊNCIAS

ACHCAR, J. A.; BARROS, E. A. C.; SOUZA, R. M.; MARTINEZ, E. Z. **Uma introdução aos métodos bayesianos aplicados à análise de dados**. 1. ed. Timburi: Cia do E-book, 2019.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American statistician**, v. 49, n. 4, p. 327-335, 1995.
<http://web1.sph.emory.edu/users/hwu30/teaching/statcomp/papers/chibGreenbergMH.pdf>. Acesso em: 25 ago. 2020.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American statistical association**, v. 85, n. 410, p. 398-409, 1990.

HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, v. 57, n. 1, p. 97-109, 1970.
https://pdfs.semanticscholar.org/143d/2e02ab91ae6259576ac50b664b8647af8988.pdf?_ga=2.232079322.419226227.1598378356-542825385.1598378356. Acesso em: 25 ago. 2020.

MATTEI, A. T. S.; MANTOVANI, M. F.; MOREIRA, R. C.; ARTHUR, J. P.; SOUZA, R. M. Nursing case management for people with hypertension in primary health care: A randomized controlled trial. **Research in Nursing & Health**, v. 43, n. 1, p. 68-78, 2019.

PLUMMER, M. et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: **Proceedings of the 3rd international workshop on distributed statistical computing**. 2003. p. 1-10. <https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf>. Acesso em: 25 ago. 2020.

R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2020. <https://www.R-project.org/>. Acesso em: 25 ago. 2020.

SOUZA, R. M.; ACHCAR, J. A.; MARTINEZ, E. Z. Use of bayesian methods for multivariate bioequivalence measures. **Journal of Biopharmaceutical Statistics**, v. 19, n. 1, p. 42-66, 2009.

SOUZA, R. M.; ACHCAR, J. A.; MARTINEZ, E. Z.; MAZUCHELI, J. The use of asymmetric distributions in average bioequivalence. **Statistics in Medicine**, v. 35, n. 15, p. 2525-2542, 2016.