https://doi.org/10.22491/2357-9730.98944

Modelagem estatística: perguntas que você sempre quis fazer, mas nunca teve coragem

Statistical modeling: questions you have always wanted to ask but never had the courage to

Vanessa Bielefeldt Leotti[1,2], Rogério Boff Borges[1], Aline Castello Branco Mancuso[1], Stela Maris de Jezus Castro[1,2], Vânia Naomi Hirakata[1], Suzi Alves Camey[1,2]

Clin Biomed Res. 2019;39(4):356-363

1 Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação (GPPG), Hospital de Clínicas de Porto Alegre (HCPA). Porto Alegre, RS, Brasil.

2 Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, RS, Brasil.

Autor correspondente:
Vanessa Bielefeldt Leotti
l-bioestatistica@hcpa.edu.br
Hospital de Clínicas de Porto Alegre (HCPA)
Rua Ramiro Barcelos, 2350.
90035-007, Porto Alegre, RS, Brasil.

RESUMO

Dando continuidade aos artigos da série "Perguntas que você sempre quis fazer, mas nunca teve coragem", que tem como objetivo responder e sugerir referências para o melhor entendimento das principais dúvidas dos pesquisadores do Hospital de Clínicas de Porto Alegre sobre estatística, este quarto artigo se propõe a responder às principais dúvidas levantadas sobre modelagem estatística. São discutidas questões referentes à classificação de variáveis em independentes e dependentes, diferenças entre correlação, associação e regressão, os principais tipos de regressão e quais etapas são necessárias na construção de modelos. Os conceitos são abordados numa linguagem acessível ao público leigo e diversas referências são sugeridas para os curiosos em relação ao tema.

Palavras-chave: *Modelo estatístico; regressão; variável dependente; variável independente; correlação; associação*

ABSTRACT

Continuing the series of articles "Questions you have always wanted to ask but never had the courage to," which aims to answer the most common questions of researchers at Hospital de Clínicas de Porto Alegre regarding statistics and to suggest references for a better understanding, this forth article addresses the topic of statistical modeling. Questions about classification of variables as dependent or independent, differences between correlation, association and regression, types of regression and steps for statistical modeling are discussed. The concepts are explained in plain language for lay readers and several references are suggested for those curious about the topic.

Keywords: *Statistical model; regression; dependent variable; independent variable; correlation; Association*

O QUE SÃO MEDIDAS DEPENDENTES E INDEPENDENTES?

Em termos estatísticos, a palavra *medida* é utilizada em referência aos resultados de uma variável aleatória observada (ou mensurada) na amostra de um determinado estudo. A diferença entre ser dependente ou independente está na relação da variável (medida) em questão com as demais.

Uma medida ou variável dependente é um fenômeno (quantitativo ou qualitativo) que aparece, desaparece, aumenta ou diminui conforme as mudanças ocorridas em ao menos uma outra variável. As variáveis aleatórias dependentes normalmente são os próprios desfechos dos estudos, pois geralmente se quer conhecer o que ocorre com o comportamento desta variável quando outras são alteradas. Já as variáveis que influenciam, afetam ou determinam uma variável dependente são denominadas de independentes. Tais variáveis podem, por vezes, ser controladas pelo pesquisador. Por exemplo, no estudo de Cadavid et al.[1], a variável dependente é o desfecho 'melhora confirmada

ISSN 2357-9730

da deficiência' e uma das variáveis independentes é o grupo ao qual o paciente é alocado (tratamento com opicinumab ou placebo).

Neste contexto, variável resposta ou desfecho são sinônimos para a variável dependente. Já para as variáveis independentes, os sinônimos são variáveis explicativas, variáveis preditoras, fatores ou covariáveis. O site da Khanacademy.org[2] apresenta algumas videoaulas que ilustram estes conceitos.

QUAL A DIFERENÇA ENTRE CORRELAÇÃO E ASSOCIAÇÃO?

Em estatística, muitas vezes se faz necessário avaliar se duas variáveis aleatórias estão relacionadas e, se sim, qual o grau dessa relação. No entanto, a técnica estatística empregada para analisar tal relação dependerá do tipo de variável em estudo. Quando as duas variáveis em questão são quantitativas, a técnica utilizada para avaliar a relação entre elas se chama Análise de Correlação. Quando as duas variáveis são categóricas, costuma-se utilizar as chamadas Medidas de Associação. É importante salientar que, em ambas as técnicas, não é possível estabelecer uma relação de causa e efeito entre as variáveis.

No caso das variáveis quantitativas, costuma-se utilizar o coeficiente de *Correlação*, que tem como objetivo quantificar o grau da relação entre duas variáveis. A vantagem desse coeficiente é a de ser um número puro, independe da unidade medida nas variáveis. Existem diferentes métodos para se calcular o coeficiente de correlação, que dependerá das características das variáveis envolvidas, dentre eles os mais conhecidos são: o coeficiente de correlação de momento-produto de Pearson e o coeficiente de correlação de Spearman. Além de estimar a magnitude da correlação entre duas variáveis, é também comum testar a significância desta estimativa, isto é, se o coeficiente de correlação estimado com a amostra pode ser considerado significativamente diferente de zero.

O coeficiente de correlação de Pearson é uma medida que avalia o grau (ou intensidade) da relação linear entre duas variáveis quantitativas contínuas, que varia de − 1 a + 1 e costuma ser representada pela letra **r**. Quando o coeficiente é significativamente diferente de zero, pode-se dizer que há evidências de correlação. Um coeficiente igual a + 1 indica uma relação linear direta ou positiva perfeita (Figura 1), e um coeficiente igual a − 1 indica uma relação linear inversa ou negativa perfeita (Figura 2). Quando não é possível perceber a relação entre as variáveis ele assume o valor zero, indicando a ausência de relação linear entre as duas variáveis avaliadas (Figuras 3a e 3b). Observa-se, ainda, que um coeficiente nulo não significa que as variáveis não estão relacionadas de alguma outra forma, como é o caso da Figura 3a, onde existe associação entre x e y, porém não na forma de uma reta. A Tabela 1 apresenta uma classificação da magnitude de correlação, sugerida por Hopkins[3].

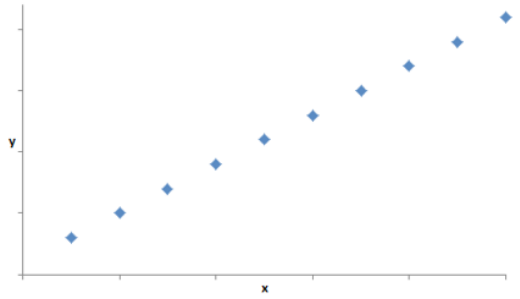

Figura 1: Correlação linear positiva perfeita (r = 1).

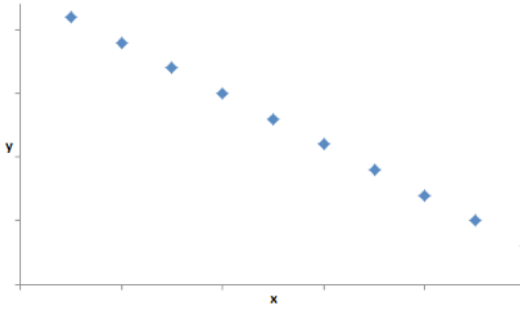

Figura 2: Correlação linear negativa perfeita (r = − 1).

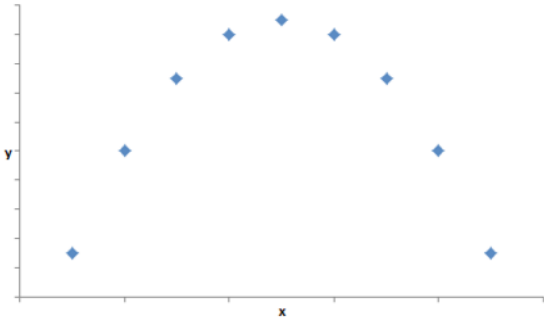

Figura 3a: Ausência de correlação linear (r ≈ 0).

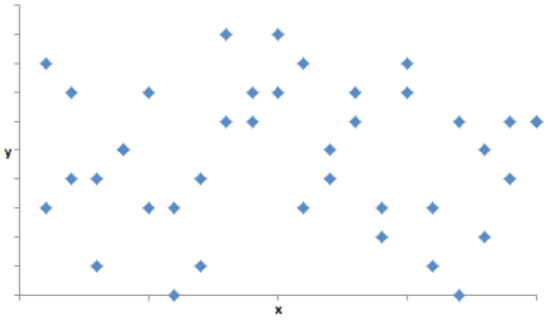

Figura 3b: Ausência de correlação linear (r ≈ 0).

Tabela 1: Classificação da magnitude de correlação (em valor absoluto).

Coeficiente de correlação	Classificação
0 - 0,1	muito fraca (trivial)
0,1 - 0,3	fraca
0,3 - 0,5	moderada
0,5 - 0,7	forte
0,7 - 0,9	muito forte
0,9 - 1,0	perfeita ou quase perfeita

Callegari-Jacques[4] exemplifica o conceito de correlação: Um professor deseja saber se existe correlação (relação) entre o tempo dedicado ao estudo e o desempenho dos alunos em determinada disciplina. A variável tempo (x) é medida em horas de estudo e a variável desempenho (y) é a nota obtida na prova por cada aluno. A Figura 4 apresenta os dados de tempo e desempenho plotados em um diagrama de dispersão. Aparentemente, alunos com mais tempo de dedicação aos estudos tendem a ter um melhor desempenho na prova e alunos com menos tempo de dedicação tendem a ter pior desempenho. O coeficiente de correlação de Pearson para estes dados é igual a 0,58, indicando uma relação linear positiva moderada entre o tempo (em horas) de dedicação aos estudos e o desempenho na prova da disciplina.

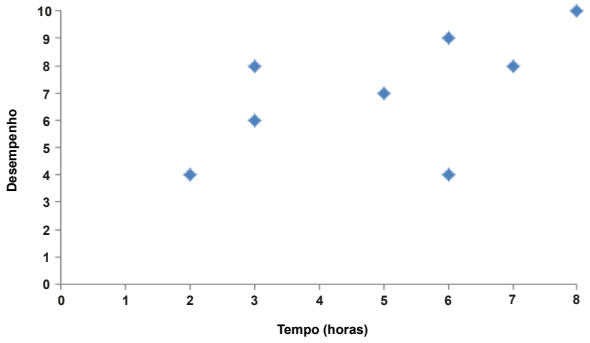

Figura 4: Diagrama de dispersão correspondente ao tempo dedicado ao estudo e o desempenho alcançado na prova.

Já nos casos em que se deseja avaliar a existência de uma relação entre duas variáveis categóricas (qualitativas), costuma-se utilizar o termo **Associação**. Geralmente, a avaliação da existência, ou não, de associação entre duas variáveis categóricas é feita através de testes de associação, onde a hipótese nula indica a ausência de associação entre duas variáveis e a hipótese alternativa indica a presença. Os testes mais conhecidos para este contexto são o Teste Exato de Fisher e o Teste Qui-Quadrado de Associação de Pearson. Zar[5] traz como exemplo dados coletados de presença e ausência de doença em uma planta (Doença na planta) e dados de presença e ausência de determinada espécie de inseto (Inseto), conforme a Tabela 2. A pergunta do pesquisador é: as variáveis 'Presença da espécie de inseto' e 'Presença de doença na planta' estão associadas?

Tabela 2: Tabela de contingência das variáveis 'presença da espécie de inseto' e 'presença de doença na planta'.

	Doença na planta		
Inseto	**Presente**	**Ausente**	**Total**
Presente	60	40	100
Ausente	0	40	40
Total	60	80	140

Para responder a pergunta do pesquisador, testa-se a hipótese nula de que as variáveis 'Presença da espécie de inseto' e 'Presença de doença na planta' não estão associadas. Neste caso, o teste de associação rejeita a hipótese nula, com valor $p < 0,0001$, evidenciando a presença de associação entre as duas variáveis.

Mas além do teste de hipóteses, também podem ser utilizadas medidas de associação tais como o coeficiente de Contingência (*Pearson coefficient of mean square contingency*), os coeficiente de Cramér Φ_1 (Phi de Cramér) ou Φ_2 (um coeficiente de contingência que varia de − 1 a + 1), o coeficiente de associação de Yule, entre outros[5]. No exemplo anterior, a partir do resultado do teste de associação, concluímos que as variáveis 'Presença da espécie de inseto' e 'Presença de doença na planta' estão associadas, no entanto não sabemos o grau desta associação. A medida de associação Φ_2, que tem uma interpretação semelhante ao coeficiente de correlação de Pearson, expressa não somente a força da associação entre as duas variáveis, mas também a direção desta associação. Para os dados da Tabela 2, Φ_2 é igual a 0,55, denotando uma associação moderada positiva entre a 'Presença de doença na planta' e a 'Presença da espécie de inseto'. Pode-se entender esta associação observando na Tabela 2 que, do total de 60 plantas doentes, 100% delas tem a presença de insetos.

QUAL A DIFERENÇA ENTRE CORRELAÇÃO E REGRESSÃO?

Em suma, correlação e regressão são duas técnicas que estudam o comportamento conjunto entre duas ou mais variáveis. Podem ser consideradas complementares, mas não são equivalentes.

Ao contrário da correlação, a análise de regressão é utilizada nas situações em que há razões para supor uma relação de dependência entre duas ou mais

variáveis. Pode ter como objetivo ajustar a melhor função matemática que relacione as variáveis e/ou predizer o valor da variável dependente com base nos valores das variáveis independentes (preditores). Existem diversos modelos de regressão, os principais podem ser vistos na quinta pergunta deste artigo.

Mais especificamente, a correlação linear de Pearson pode ser utilizada como uma etapa prévia a regressão linear, analisando e quantificando uma possível linearidade entre a variável dependente e a variável preditora, ambas quantitativas. Dada a existência da relação e dos demais pressupostos do modelo, a regressão linear ajusta a melhor reta que representa essa relação, permitindo a previsão da variável dependente com base na(s) variável(is) preditora(s). Zar[5] contrapõe estas duas técnicas e Callegari-Jacques[4] descreve estes e demais conceitos.

QUAL A DIFERENÇA ENTRE ANÁLISE MULTIVARIADA E MULTIVARIÁVEL?

É comum os termos *multivariável* e *multivariada* serem confundidos ou até mesmo utilizados como sinônimos. No entanto, não são equivalentes, pois representam análises diferentes.

Uma análise *multivariável* pode ser entendida como a extensão de uma análise que considera apenas um preditor, para determinar as influências de vários preditores sobre um único resultado. Uma regressão simples, por exemplo, que por definição tem apenas uma variável preditora (variável independente), pode ser estendida para regressão múltipla ou multivariável ao incluir várias variáveis preditoras, porém apenas uma variável dependente. Por exemplo, vários fatores podem ser associados ao desenvolvimento de angina, como tabagismo, obesidade, sedentarismo, diabetes, hipertensão, entre outros. A análise multivariável nos permite determinar a contribuição conjunta de cada um desses fatores para o desenvolvimento da doença[6,7].

Já o termo análise *multivariada* se refere a técnicas estatísticas que consideram simultaneamente múltiplas medidas dependentes, para o mesmo indivíduo, onde um resultado (desfecho) é representado por duas ou mais variáveis dependentes. Por exemplo, para avaliar o sono de universitários Taylor et al.[8] considera o tempo total de sono, um índice de eficiência do sono e outro de qualidade do sono, conjuntamente[6,9].

Cabe salientar que tanto a denominação *análise multivariada* quanto a *análise multivariável* correspondem a um grande número de métodos e técnicas estatísticas que utilizam várias variáveis. No contexto multivariável pode ser citado os diversos modelos de regressão (linear, logística, de Poisson, de Cox etc.) e de análise de variâncias (Anova), entre outros. Dentre as técnicas multivariadas, as mais conhecidas são análise de componentes principais, análise fatorial, análise de agrupamento, análise multivariada de variâncias (Manova), entre outras.

QUAIS OS PRINCIPAIS TIPOS DE REGRESSÃO E SUAS DIFERENÇAS?

Aqui, serão abordados os modelos de regressão multivariáveis. Esta família de modelos estatísticos é útil quando há múltiplos preditores para serem relacionados com um único desfecho (variável dependente). Basicamente, é o tipo do desfecho que determina o tipo de regressão.

Para *desfechos contínuos* (por exemplo, custo que o paciente teve com um tratamento), e se for razoável fazer a suposição de que o mesmo tem distribuição normal, pode-se utilizar a *regressão linear*. Caso a suposição de normalidade não seja razoável, uma possibilidade é utilizar algum modelo dentro da classe dos *modelos lineares generalizados* (sigla GLM, do inglês *Generalized Linear Models*)[10]. Nesta classe, pode-se trabalhar com outras distribuições como, por exemplo, a distribuição Gama, que modela desfechos contínuos com assimetria.

Para *desfechos dicotômicos*, ou seja, quando se pode classificar os respondentes em um de dois grupos (por exemplo, doente ou não-doente), pode-se trabalhar com a *regressão logística*, com a *regressão de Poisson com variância robusta*, entre outros. Ambos pertencem à classe dos GLM e sua diferença será explicada no próximo tópico deste artigo.

Para *desfechos categóricos*, com mais de duas categorias (por exemplo, local de moradia classificado em capital, região metropolitana ou interior), pode-se trabalhar com a *regressão logística politômica*[11]. Caso exista uma ordenação nas categorias (por exemplo, graus de escolaridade), o desfecho é chamado de ordinal e então pode-se optar pela *regressão logística ordinal*.

Para *desfechos do tipo tempo até um evento*, em que nem todos os sujeitos são monitorados até o evento ocorrer, utiliza-se os modelos de sobrevida, ou sobrevivência. Um desses modelos é a *regressão de Cox*[11]. Por exemplo, considerando pacientes transplantados, pode-se ter interesse de modelar o tempo até a rejeição, definindo-se um tempo máximo de 5 anos de acompanhamento após transplante. Pacientes que não apresentaram rejeição até 5 anos são ditos censurados. Assim, os modelos de sobrevida lidam conjuntamente com indivíduos que apresentaram o evento e indivíduos censurados.

Para *desfechos do tipo contagem* (por exemplo, número de reinternações em um específico período de tempo), pode-se aplicar a *regressão de Poisson*, que é um tipo de GLM. Este modelo também pode

ser utilizado para modelar taxas (por exemplo, taxa de mortalidade infantil por mil habitantes), pois taxas usualmente tem seu numerador formado por uma contagem[11].

Todos estes modelos compartilham similaridades, como por exemplo, a possibilidade de lidar com preditores contínuos ou categóricos. O impacto de cada preditor no desfecho é medido através de parâmetros chamados *coeficientes de regressão*[10]. Além de Dupont[11] e Vittinghoff et al.[10], outra referência recomendada sobre regressão em epidemiologia é Suárez Pérez[12].

QUANDO UTILIZAR REGRESSÃO LOGÍSTICA OU REGRESSÃO DE POISSON COM VARIÂNCIAS ROBUSTAS?

A regressão logística e a de Poisson com variâncias robustas são utilizadas para estimar razões em estudos cujo desfecho de interesse é dicotômico. No entanto, cada uma delas nos fornece estimativas de medidas diferentes: a regressão logística para razão de chances (*odds ratio*) e a regressão de Poisson robusta para razão de prevalências ou de incidências, conforme o delineamento do estudo (respectivamente: transversal ou longitudinal)[13,14]. Mais detalhes sobre a diferença entre as estimativas podem ser obtidas em Castro et al.[15].

Do ponto de vista metodológico, os estudos de caso-controle são preferíveis quando desfechos são raros[16-18] e, quando os desfechos são mais comuns, estudos transversais são mais frequentemente utilizados.

Embora seja possível estimar a razão de chances em qualquer delineamento (caso-controle, longitudinal ou transversal), não podemos interpretá-la como uma razão de prevalências ou de incidências que são as medidas apropriadas para esses delineamentos[17]. Além disso, quando o desfecho não é raro (até 10%) ou a magnitude do efeito for elevada, a razão de chances superestima o risco relativo, distorcendo, portanto, a magnitude do efeito que se deseja estimar[19-21]. Outro problema é que a razão de chances é uma medida de efeito de difícil interpretação[22-24].

Portanto, ao se analisar estudos transversais ou longitudinais, a análise recomendada é a regressão de Poisson com variâncias robustas, mas no caso de estudos de caso-controle, o uso da regressão logística é recomendado.

EXISTE UM MÉTODO OU ETAPAS PARA REALIZAR UMA REGRESSÃO?

Primeiramente é necessário planejar a coleta de dados em relação a escolha das informações a serem obtidas. O pesquisador precisa planejar quais as variáveis que respondem o objetivo do estudo. Maiores informações sobre planejamento de um estudo podem ser obtidas em Hulley et al.[17] e em Hirakata et al.[25].

Após a coleta ser realizada, é fundamental realizar a análise descritiva dos dados. Essa etapa é necessária e crucial para qualquer análise, uma vez que com ela é possível detectar possíveis erros de digitação e dados discrepantes, conforme discutido no primeiro artigo desta série, Mancuso et al.[26].

Finalmente, na etapa de construção do modelo de regressão multivariável, uma das maiores questões dos pesquisadores é: "Quais variáveis devem ser incluídas em um modelo multivariável?". A resposta dessa questão depende, principalmente, do objetivo do estudo. Se o objetivo for predição de novos valores, métodos automáticos de seleção podem ser apropriados para a construção do modelo[16]. No entanto, em estudos com o objetivo de avaliar a relação de um desfecho com um conjunto de variáveis preditoras, o ideal é selecionar variáveis que apresentam uma plausibilidade biológica/clínica para explicar o desfecho. Em todos os casos, busca-se um modelo parcimonioso, isto é, um modelo que explique bem o fenômeno de interesse com o mínimo de variáveis. Independentemente da estratégia utilizada para a seleção de variáveis, sua elaboração está sempre sujeita a erros[16,27].

Nesta seção será descrita a técnica de modelagem hierarquizada proposta por Victora et al.[28], os métodos automáticos de seleção de variáveis (melhor subgrupo, *forward* e *backward*) e o método de seleção proposital proposto por Hosmer e Lemeshow[29].

Victora et al.[28] propuseram uma técnica de *modelagem hierarquizada* para avaliar a relação das variáveis independentes com o desfecho. Essa proposta exige que o pesquisador construa previamente um modelo conceitual dos possíveis fatores de exposição e confusão com base no conhecimento já existente, o que os autores chamam de marco conceitual. Essa estrutura é mantida em todas as etapas do processo iterativo, possibilitando a seleção daquelas mais fortemente relacionadas ao desfecho[28,30,31].

Neste conceito, as variáveis são organizadas de forma hierárquica, de acordo com sua proximidade causal com o desfecho, do nível mais distal até o nível mais proximal, conforme ilustra o Quadro 1. Pressupõe-se que os níveis superiores (distais) da hierarquia influenciam nos níveis inferiores (mais proximais), que, por consequência, afetam, diretamente ou indiretamente, o desfecho de interesse. Assim, o modelo é ajustado sequencialmente, incluindo, em cada etapa, todas as variáveis do nível e avaliando seus efeitos. As variáveis selecionadas em um determinado nível permanecem nos modelos subsequentes, mesmo que, com a inclusão das

variáveis do próximo nível, elas venham a perder a significância.

Quadro 1: Modelo conceitual hipotético.

Nível hierárquico	**Exemplo utilizado por Victora et al.**[28]
1. Variáveis distais	Renda
2. Variáveis intermediárias	Saneamento
3. Variáveis proximais	Desnutrição
Desfecho	**Óbito por diarreia**

O Quadro 1 reproduz o exemplo utilizado por Victora et al.[28]. Nessa análise o primeiro modelo é construído somente com as variáveis de renda (Nível 1). Este modelo fornecerá o efeito de cada variável relacionada à renda sobre o desfecho, porém sem o ajuste das variáveis relacionadas ao saneamento e à desnutrição. Na segunda etapa são adicionadas as variáveis do próximo nível, no caso, variáveis de saneamento (Nível 2). O efeito obtido dessas variáveis está ajustado para as variáveis de renda, no entanto, ainda sem o ajuste das variáveis de desnutrição. Neste segundo modelo, o efeito restante das variáveis de renda, representa o efeito "independente" das variáveis de saneamento. Por fim, é incorporado ao modelo as variáveis relacionadas à desnutrição (Nível 3). O resultado do efeito dessas variáveis proximais está ajustado para as variáveis de renda e saneamento. Qualquer efeito residual que constar nas variáveis dos níveis anteriores, representa a magnitude do efeito "independente" dos respectivos níveis inferiores. A principal vantagem da análise hierárquica é que os critérios de seleção são previamente estabelecidos e há uma base teórica para essa decisão.

Caso o pesquisador não tenha nenhuma referência sobre quais variáveis incluir no modelo, a seleção pode ser realizada através de métodos automáticos, que utilizam critérios estatísticos para se chegar ao modelo multivariável final. Cabe salientar que esses procedimentos automáticos não garantem uma plausibilidade biológica ou um bom ajuste do modelo final[32].

Existem diversos métodos automáticos, mas todos exigem que haja alguma métrica de comparação dos possíveis modelos, que dependerá do tipo de modelo em questão. Usualmente se utiliza medidas que penalizam modelos com um número maior de variáveis, deste modo é necessário um ajuste satisfatório para justificar a adição ou a não remoção de variáveis. Entre essas medidas as mais comuns são o AIC e o BIC[27,33,34].

O método automático *melhor subgrupo* testa todos os possíveis subconjuntos de variáveis explicativas, ajustando modelos com todas as combinações possíveis de variáveis independentes. Ao final é escolhido o modelo com a melhor medida de ajuste definida pelo pesquisador

Outro método automático é o *forward*. Nesse procedimento, o modelo inicial possui somente o intercepto e então se testa a inclusão de novas variáveis. Se o ajuste dado ao modelo com a inclusão dessa variável for satisfatório, ela permanece no modelo e não sai mais (mesmo que sua contribuição venha a não ser significativa nas etapas seguintes), caso contrário essa variável é descartada e não volta mais ao processo iterativo.

Já no procedimento de eliminação *backward*, o ponto de partida é o modelo saturado (incluindo todos os possíveis preditores) e então se elimina as variáveis não significativas. Nessa metodologia, a variável que sai do modelo não tem a oportunidade de voltar a ser testada novamente.

O procedimento *both stepwise* é uma mistura do *forward* e o *backward*. Uma variável corre o risco de ser incluída em um passo e ser removida novamente nos próximos passos. Nesse procedimento, em cada etapa, todas as variáveis são verificadas, podendo ser adicionadas ou removidas do modelo.

Maiores informações sobre esses métodos automáticos podem ser obtidas em Jr.[32], Miller[34], Hocking[35], Draper[36], Fritz[37] e Hosmer[29].

Hosmer e Lemeshow[29] propõem um método de seleção *proposital* que é semelhante ao método *backward*, porém no modelo inicial são incluídas somente as variáveis independentes que obtiverem valor de p menor do que 0,10, 0,20 ou 0,25 na análise univariável. Na etapa seguinte, as variáveis não significativas e que não provoquem uma grande alteração (mais dos que 20% de mudança) nas estimativas dos coeficientes são removidas do modelo, uma de cada vez. Após verificar todas as variáveis, testa-se o efeito de cada variável que não foi incluída no modelo inicial, essa etapa é fundamental para verificar se a variável que sozinha não é significativa tem importância na presença de outras variáveis[29,38].

Após definir as variáveis do modelo multivariável, é necessário verificar seu ajuste e se seus pressupostos estão sendo atendidos. Além disso, é recomendado avaliar a influência de possíveis dados discrepantes sobre as estimativas[39,40]. O ajuste do modelo de regressão linear pode avaliado pelo coeficiente de determinação (R^2), ele indica o percentual de variação da variável dependente que pode ser explicado pelas variáveis independentes. Seus principais pressupostos são que os resíduos são independentes e identicamente distribuídos seguindo uma distribuição normal, maiores detalhes podem ser obtidos em Kutner et al.[41]. Hosmer e Lemeshow[28] mostram maneiras de avaliar o ajuste do modelo de regressão logística, cujos conceitos podem ser aplicados ao modelo de regressão Poisson

com variância robusta. Entre outras medidas, pode ser utilizado o teste de Hosmer-Lemeshow para avaliar o ajuste desses modelos. Este teste avalia se a frequência de eventos esperada pelo modelo corresponde a frequência de eventos observada nos dados[42,43]. Para o modelo de regressão de Cox tradicional é essencial avaliar a proporcionalidade do *hazard* ao longo do tempo observado[44].

REFERÊNCIAS

1. Cadavid D, Mellion M, Hupperts R, Edwards KR, Calabresi PA, Drulović J, et al. Safety and efficacy of opicinumab in patients with relapsing multiple sclerosis (SYNERGY): a randomised, placebo-controlled, phase 2 trial. *Lancet Neurol*. 2019;18(9):845-56.
2. Pré-álgebra: equações, expressões e inequações [Internet]. Mountain View: Khan Academy; c2020 [citado 2019 Ago 13]. Disponível em: https://pt.khanacademy.org/math/pre-algebra/pre-algebra-equations-expressions
3. Hopkins WG. New view of statistics: effect magnitudes. 2006 ago 7 [citado 2019 Out 8]. In: SPORTSCIENCE [Internet]. Disponível em: https://www.sportsci.org/resource/stats/effectmag.html.
4. Callegari-Jacques SM. Bioestatística: princípios e aplicações. Porto Alegre: Artmed; 2003.
5. Zar JH. Biostatistical Analysis. Upper Saddle River: Prentice Hall; 1999.
6. Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health*. 2013;103(1):39-40.
7. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med*. 2003;138(8):644-50.
8. Taylor DJ, Bramoweth AD, Grieser EA, Tatum JI, Roane BM. Epidemiology of insomnia in college students: relationship with mental health, quality of life, and substance use difficulties. *Behav Ther*. 2013;44(3):339-48.
9. Hair Jr. JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Análise multivariada de dados. Porto Alegre: Bookman; 2005.
10. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Nova York: Springer; 2012.
11. Dupont WD. Statistical modeling for biomedical researchers: a simple introduction to the analysis of complex data. 2a ed. Cambridge: Cambridge University Press; 2009.
12. Suárez Pérez EL, Pérez CM, Rivera R, Martínez MN, organizadores. Applications of regression models in epidemiology. Hoboken: Wiley; 2017.
13. Hirakata VN. Estudos transversais e longitudinais com desfechos binários: qual a melhor medida de efeito a ser utilizada? *Rev HCPA* [Internet]. 2009 [citado 2020 Jan 12];29(2):174-6. Disponível em: http://www.seer.ufrgs.br/index.php/hcpa/article/view/9737.
14. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol*. 2003;3:21.
15. Castro SMJ, Mancuso ACB, Leotti VB, Hirakata VN, Camey SA. Bioestatística e epidemiologia: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res*. 2019;39(3):258-65.
16. Rothman K, Greenland S, Lash T. Epidemiologia moderna. 3a ed. Porto Alegre: Artmed; 2011.
17. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Delineando a pesquisa clínica: uma abordagem epidemiológica. Porto Alegre: Artmed; 2003.
18. Breslow NE, Day NE. Statistical methods in cancer research volume I : the analysis of case-control studies. Lyon: IARC Scientific Publication; 1980.
19. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761-8.
20. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ*. 1998;317(7168):1318.
21. Nurminen M. To use or not to use the odds ratio in epidemiologic analyses. *Eur J Epidemiol*. 1995;11(4):365-71.
22. Lee J. Odds ratio or relative risk for cross-sectional data? *Int J Epidemiol*. 1994;23(1):201-3.
23. Axelson O, Fredriksson M, Ekberg K. Use of prevalence ratio v the prevalence odds ratio in view of confounding in cross sectional studies. *Occup Environ Med*. 1995;52(7):494.
24. Sackett D, Deeks JJ, Altman DG. Down with odds ratios! *BMJ Evid Based Med*. 1996;1:164-6.
25. Hirakata VN, Mancuso ACB, Castro SMJ. Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res* [Internet]. 2019 [citado 2019 Set 27];39(2):181-5. Disponível em: https://seer.ufrgs.br/hcpa/article/view/93649
26. Mancuso ACB, Castro SMJ, Guimarães LSP, Leotti VB, Hirakata VN, Camey SA. Estatística descritiva: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res* [Internet]. 2018 [citado 2019 Set 27];38(4). Disponível em: https://seer.ufrgs.br/hcpa/article/view/89242.
27. Akaike H. A new look at the statistical model identification. In: Parzen E, Tanabe K, Kitagawa G, organizadores. Selected papers of Hirotugu Akaike [Internet]. Nova York: Springer; 1974 [citado 2019 Out 8]. p. 215-22. Disponível em: http://link.springer.com/10.1007/978-1-4612-1694-0_16.
28. Victora CG, Huttly SR, Fuchs SC, Olinto MT. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. *Int J Epidemiol*. 1997;26(1):224-7.
29. Hosmer Jr. DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken: Wiley; 2013.

30. Fuchs SC, Victora CG, Fachel J. Modelo hierarquizado: uma proposta de modelagem aplicada à investigação de fatores de risco para diarreia grave. *Rev Saude Publica*. 1996;30(2):168-78.

31. Olinto MTA, Victora CG, Barros FC, Tomasi E. Determinantes da desnutrição infantil em uma população de baixa renda: um modelo de análise hierarquizado. *Cad Saude Publica*. 1993;9 (supl. 1):14-27.

32. Harrell Jr. FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Nova York: Springer; 2015.

33. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-4.

34. Miller A. Subset selection in regression. 2a ed. Boca Raton: CRC; 2002.

35. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32(1):1-50.

36. Draper NR, Smith H. Applied regression analysis. Hoboken: Wiley; 1998.

37. Fritz M, Berger PD. Can you relate in multiple ways? Multiple linear regression and stepwise regression. In: Improving the user experience through practical data analytics [Internet]. Elsevier; 2015 [citado 2019 Out 8]. p. 239-69. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/B9780128006351000100

38. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code Biol Med*. 2008;3:17.

39. Turkan S, Meral C, Oniz T. Outlier detection by regression diagnostics based on robust parameter estimates. *Hacet J Math Stat*. 2012;41(1):147-55.

40. Rahman K, Sathik M, Kaliyaperumal SK. Multiple linear regression models in outlier detection. *Int J Res Comput Sci*. 2012;2(2):23-8.

41. Neter J, Kutner MH, Wasserman W, Nachtsheim CJ. Applied linear regression models. Nova York: McGraw-Hill Education; 2003.

42. Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Theory Methods*. 1980;9(10):1043-69.

43. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115(1):92-106.

44. Carvalho MS, Andreozzi VL, Codeço CT, Campos DP, Barbosa MTS, Shimakura SE. Análise de sobrevivência: teoria e aplicações em saúde. Rio de Janeiro: Editora Fiocruz; 2011.

Recebido: 12 dez, 2019
Aceito: 13 dez, 2019