Bioestatística quantitativa aplicada

Edison Capp
Otto Henrique Nienov
Organizadores

Caroline Darski
Charles Francisco Ferreira
Cristiana Palma Kuhl
Fernanda Dapper Machado
Fernanda Vargas Ferreira
Hellen Meiry Grosskopf Werka
Johanna Ovalle Diaz
Marina Petter Rodrigues
Michele Strelow Moreira
Nadine de Souza Ziegler
Paula Barros Terraciano
Pedro Henrique Comerlato
Sinara Santos

Universidade Federal do Rio Grande do Sul
Faculdade de Medicina
Programa de Pós-Graduação em Ciências da Saúde:
Ginecologia e Obstetrícia

Bioestatística Quantitativa Aplicada

Porto Alegre 2020
UFRGS

U58b Universidade Federal do Rio Grande do Sul. Faculdade de Medicina. Programa de Pós-Graduação em Ciências da Saúde: Ginecologia e Obstetrícia
Bioestatística quantitativa aplicada/ Universidade Federal do Rio Grande do Sul; organizadores: Edison Capp e Otto Henrique Nienov – Porto Alegre: UFRGS, 2020.

260p.
ISBN: 978-65-86232-43-1
E-Book: 978-65-86232-44-8

1. Epidemiologia e Bioestatística 2. Estatística 3. SPSS I. Capp, Edison, org. II. Nienov, Otto Henrique, org. III Título.

NLM: WA950

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
(Bibliotecária Shirlei Galarça Salort – CRB10/1929)

Endereço:
PPG em Ciências da Saúde: Ginecologia e Obstetrícia
FAMED – UFRGS
Rua Ramiro Barcellos, 2400/2º andar
CEP 900035-003 – Porto Alegre – RS
Telefone: +55 51 3308 5607
E-mail: ppggo@ufrgs.br

Editoração e diagramação: Edison Capp
Capa: Edison Capp, imagens: www.freepik.com/starline

Edison Capp
Otto Henrique Nienov
Organizadores

Caroline Darski
Charles Francisco Ferreira
Cristiana Palma Kuhl
Fernanda Dapper Machado
Fernanda Vargas Ferreira
Hellen Meiry Grosskopf Werka
Johanna Ovalle Diaz
Marina Petter Rodrigues
Michele Strelow Moreira
Nadine de Souza Ziegler
Paula Barros Terraciano
Pedro Henrique Comerlato
Sinara Santos

11 Modelos lineares generalizados

Sinara Santos
Nadine Ziegler
Pedro Henrique Comerlato
Edison Capp
Otto Henrique Nienov

Os modelos lineares generalizados correspondem a um conjunto de análises de regressão que são uma generalização dos modelos lineares gerais. Enquanto que os modelos lineares gerais são usados para modelos de regressão e variância com uma distribuição normal e uma função de ligação identidade, os modelos lineares generalizados permitem a utilização de variáveis com distribuição assimétrica.

Para definir qual modelo linear generalizado tem de ser aplicado, deve-se considerar o tipo de variável dependente (Quadro 1) e o tipo de delineamento do estudo (Quadro 2).

Quadro 1. Análise de regressão de acordo com o tipo de variável dependente.

Tipo de variável	**Tipo de regressão**
Qualitativa nominal	Regressão logística binária
	Regressão de Poisson
	Regressão logística politômica ou multinominal
Qualitativa ordinal	Regressão logística ordinal

Regressão logística

A regressão logística é muito similar à regressão linear, pois ambas avaliam a relação entre uma variável dependente (variável desfecho ou de resposta) e uma ou mais variáveis preditoras (variável independente ou de exposição). O quadro 3 elenca as principais diferenças e semelhanças entre as regressões linear e logística.

Quadro 2. Análise de regressão de acordo com o delineamento do estudo.

<table>
<tr><th>Tipo de estudo</th><th>Tipo de regressão</th></tr>
<tr><td>Caso-controle</td><td>Regressão logística</td></tr>
<tr><td>ECR</td><td rowspan="3">Regressão de Poisson</td></tr>
<tr><td>Transversal</td></tr>
<tr><td rowspan="2">Coorte</td></tr>
<tr><td>Regressão de Cox</td></tr>
</table>

Em relação as semelhanças entre a regressão linear e a logística, existe uma boa razão para que não possamos aplicar a regressão linear diretamente a uma situação onde a variável de saída é dicotômica: uma das hipóteses da regressão linear é que o relacionamento entre as variáveis seja linear. Portanto, para o modelo de regressão linear ser válido, os dados observados devem ter um relacionamento linear. Quando a variável de saída é dicotômica, essa hipótese é normalmente violada.

Quadro 3. Comparação entre regressão linear e regressão logística.

<table>
<tr><th>Características</th><th>Regressão linear</th><th>Regressão logística</th></tr>
<tr><td>Variável dependente</td><td>Analisa respostas quantitativas</td><td>Analisa desfechos categóricos</td></tr>
<tr><td>Variáveis independentes</td><td colspan="2">Podem ser categóricas (também nomeadas como fatores) e/ou contínuas (nomeadas como covariáveis)</td></tr>
<tr><td>Interpretação dos resultados</td><td>Obtém-se uma associação em relação a uma função linear</td><td>Associação é quantificada por meio de uma razão da probabilidade de ocorrência</td></tr>
<tr><td>Reta da regressão</td><td>Reta</td><td>Curva</td></tr>
</table>

Regressão logística binária

A regressão logística binária determina a probabilidade de ocorrência das categorias de uma variável dicotômica, ou seja, prevê qual de duas categorias é mais provável, dado um conjunto de informações. Descomplicando, isso quer dizer que é possível prever a qual de duas categorias uma pessoa é provável de pertencer dado certas informações. A figura 1 descreve algumas considerações pertinentes para a análise de regressão logística binária, relacionadas a variável dependente e aos preditores do modelo.

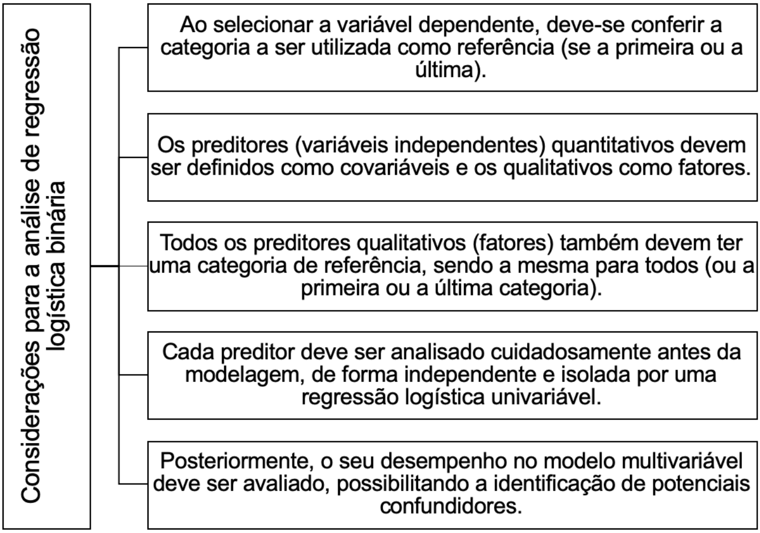

Figura 1. Considerações referentes a variável dependente e aos preditores (variáveis independentes) para a análise de regressão logística binária.

Na sua forma simples, quando existe um único previsor X, a equação da regressão logística, a partir da qual a probabilidade da variável Y é prevista, define-se em:

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \varepsilon)}}$$

Onde P(Y) é a probabilidade de Y ocorrer, e é a base dos logaritmos naturais e os demais coeficientes da equação formam uma combinaçao linear muito semelhante à regresão linear simples. A parte da equação entre parênteses é idêntica à regressão linear, pois existe uma constante (β_0), uma variável previsora (X) e um coeficiente (ou peso) agregado ao previsor (β_1).

Da mesma forma que para a regressão linear, é possível estender essa equação para incluir diversas variáveis previsoras. Quando isso ocorre, a equação toma a seguinte forma:

$$P(Y)= \frac{1}{1 + e -(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_n X_n + \varepsilon)}$$

A regressão logística, ao contrário da regressão linear, não pressupõe normalidade. Mas, da mesma forma, deve-se investigar a multicolinearidade entre as variáveis independentes e a presença de valores atípicos. Também se recomenda um tamanho amostral mínimo para a análise de regressão, dependendo do número de variáveis preditoras. Alguns autores indicam 10 casos para cada previsor no modelo, outros 15 casos por previsor.

Para exemplificar, faremos uma regressão logística binária univariável utilizando o "Banco de dados 7.sav" (disponível em https://bit.ly/bancosdedados), supondo um delineamento de estudo do tipo caso-controle, em que o critério de seleção foi a ocorrência ou não de polineuropatia periférica (PNP) (onde, "0" = "Não neuropata" e "1" = "Neuropata") após a cirurgia bariátrica. Pretende-se determinar se a idade (variável quantitativa) e o sexo (variável qualitativa) (onde, "1" = "Masculino" e "2" = "Feminino") possuem associação com o desfecho PNP. Lembre-se: é importante conhecer o banco de dados. Tome algum tempo para examiná-lo e conhecer as variáveis.

Atenção! Deve-se ter cuidado com as tabelas de contingência contendo células vazias. Incluir estas variáveis no modelo causará resultados numéricos indesejados e a melhor estratégia para resolver este problema é o agrupamento desta variável em um conjunto menor de categorias.

Como visto no capítulo 10, para definirmos um modelo de regressão, podemos utilizar como critério de seleção as variáveis que se mostraram estatisticamente significativas ou com p-valor limítrofe para associação na análise univariada. No exemplo, houve uma diferença significativa entre as medianas de idade dos sujeitos com e sem PNP (50 anos versus 34 anos, respectivamente; $p < 0,001$) pelo teste U de Mann-Whitney e, uma indicação de associação de PNP com o sexo masculino (61,8%; $p < 0,001$) pelo teste de correção de continuidade de Yates. Essas associações nos motivam a avaliar o impacto da idade e do sexo na PNP.

Primeiro, precisamos verificar a colinearidade e a presença de valores atípicos. Essa verificação é feita da mesma forma que na regressão linear. No menu "Analisar", "Regressão", clique em "Linear...". Na janela "Regressão linear", selecione a variável "Polineuropatia periférica" na tela da esquerda e a insira em "Dependente" e, as variáveis "Idade" e "Sexo" como "Independente(s)". Clique em "Estatísticas" e selecione as opções "Diagnóstico de colinearidade" e "Diagnóstico por caso". Clique em "Continuar". Por fim, clique em "Ok" ou "Colar".

No arquivo de saída, verifique os parâmetros de avaliação de colinearidade, índice de Tolerância e VIF. Lembre-se que o VIF deve ser menor que 10 e o índice de Tolerância maior que 0,2. Neste caso, temos um VIF de 1,177 e um índice de Tolerância de 0,850, ou seja, não há colinearidade entre as variáveis. Na análise, também, foi identificado um *outlier* (caso 47). Nestes casos, o recomendado é realizar a análise de regressão em um modelo sem e o outro com o caso, para verificar a influência deste sobre o resultado.

Verificados a colinearidade e valores atípicos, podemos prosseguir para a regressão logística binária univariável, avaliando primeiro a idade (variável quantitativa). No menu "Analisar", "Modelos lineares generalizados", clique em "Modelos lineares generalizados...". Na janela "Modelos lineares generalizados", precisaremos definir alguns parâmetros. Na aba "Tipo de Modelo", selecione a opção "Logística binária". Na aba "Resposta", selecione e insira em "Variável dependente" a variável "Polineuropatia periférica". Na mesma aba, em "Tipo de variável dependente (apenas distribuição binomial)", marque

"Binário". Clique em "Categoria de referência..." e selecione a opção "Primeiro (menor valor)", onde "0" = "Não neuropata". Clique em "Continuar". Na aba "Preditores", selecione na tela da esquerda e insira em "Covariáveis" a variável "Idade" (variável quantitativa). Na aba "Modelo", selecione em "Fatores e covariáveis" a variável "Idade" e mova-a para "Modelo" marcando como "Tipo: Efeito Principal". Na aba "Estatísticas", marque as opções "Razão de probabilidade" em "Estatísticas de quiquadrado", "Probabilidade do perfil" em "Tipo de intervalo de confiança" e "Incluir estimativas de parâmetro exponencial" em "Imprimir". Por fim, clique em "Ok" ou "Colar".

Nos resultados da análise, no arquivo de saída, os primeiros quadros mostram as informações do modelo, as estatísticas descritivas da variável dependente e da variável preditora. O primeiro quadro (*Model Information*) mostra as informações gerais do modelo. O segundo quadro (Case Processing Summary) apresenta o número de casos incluídos, excluídos e total. Em seguida (*Categorical Variable Information*), temos a proporção absoluta e relativa para as categorias da variável dependente (PNP), onde 27,5% (n = 55) da amostra é neuropata. No quarto quadro (*Continuous Variable Information*), temos o valor absoluto (N), média (*Mean*), desvio padrão (*Std. Deviation*) e os valores mínimo (*Minimum*) e máximo (*Maximum*) para a covariável (Idade).

Após, encontramos no quadro "*Goodness of Fit*" diferentes medidas para avaliar a qualidade de ajuste do modelo, úteis para comparar modelos concorrentes. Em todas, grandes valores indicam falta de ajuste. O resultado do "*Omnibus Test*" nos indica se o modelo de regressão foi significativo. Neste caso, com $p < 0{,}001$, temos evidência que o modelo da regressão é estatisticamente significativo.

No quadro "*Tests of Model Effects*", temos o teste de Qui-quadrado de *Likelihood Ratio* para a significância dos preditores. Esse verifica a existência de diferenças entre o modelo com o preditor e o modelo nulo (intercepto). Se significativo (neste caso tem um $p < 0{,}001$), o teste informa que o coeficiente do preditor difere de zero. Se isto ocorrer, pode-se dizer que o preditor está contribuindo de modo significativo para o modelo.

Constatada a significância do preditor, o quadro "*Parameter Estimates*" apresenta os parâmetros estimados. Encontramos o coeficiente B (que corresponde a constante β), o erro padrão (*Std. Error*), o limite inferior (95% *Profile Likelihood Confidence Interval: Lower*) e superior (95% *Profile Likelihood Confidence Interval: Upper*) do intervalo de confiança para B. Também é apresentado o teste Qui-quadrado de Wald, que testa as mesmas hipóteses do teste Qui-quadrado de *Likelihood Ratio*. No entanto, utilizamos os resultados do teste Qui-quadrado de *Likelihood Ratio* pois é mais poderoso que o teste de Wald. Além disso, a estatística de Wald deve ser vista com cautela porque quando o coeficiente de regressão (β) é grande, o erro padrão tende a ficar inflacionado, resultando em uma estatística de Wald subestimada. A inflação do erro padrão aumenta a probabilidade de que o previsor seja rejeitado quando na realidade ele contribui para o modelo (erro do Tipo II).

Nas últimas colunas, encontra-se o OR, expresso pelo valor de Exp(B) e seu intervalo de confiança. O mais importante para a interpretação da regressão logística é o valor da Exp(B), que é um indicador da mudança nas probabilidades resultantes da mudança de uma unidade no previsor. O OR (Exp(B)) indica a presença de um fator de risco ou proteção (Quadro 4) e é interpretado da mesma forma vista no capítulo 9.

Quadro 4. Interpretação dos valores de OR.

Valor de OR	Interpretação
Exp(B) > 1	Indica que à medida que o previsor aumenta, aumentam as chances de ocorrência do desfecho.
Exp(B) = 1	O efeito é nulo.
Exp(B) < 1	Indica que à medida que o previsor aumenta, as chances de o desfecho ocorrer diminuem.

Conclui-se que a idade é um preditor de PNP. Usando a primeira categoria da variável PNP como referência, PNP igual à zero (não neuropata), podemos afirmar que a chance de ter PNP aumenta em 41,5% (IC95%: 28,6%-60,3%) para cada ano de incremento na idade ($p < 0,001$).

Em seguida, vamos avaliar o sexo (variável qualitativa dicotômica). No menu "Analisar", "Modelos lineares generalizados", clique em "Modelos lineares generalizados...". Na aba "Tipo de modelo", selecione a opção "Logística binária". Na aba "Resposta", selecione na tela da esquerda e insira em "Variável dependente" a variável "Polineuropatia periférica". Na mesma aba, em "Tipo de variável dependente (apenas distribuição binominal)", marque "Binário". Clique em "Categoria de referência..." e selecione a opção "Primeiro (menor valor)", onde "0" = "Não neuropata". Clique em "Continuar". Na aba "Preditores", selecione na tela da esquerda e insira em "Fatores" a variável "Sexo" (variável dicotômica). Como se trata de uma variável categórica, também precisamos informar a categoria de referência. Clique em "Opções", localizado abaixo da tela de "Fatores" e selecione como "Crescente" (assume que a categoria de referência é a última em todos os fatores). A opção "Decrescente" assume que a categoria de referência é a primeira em todos os fatores e, "Usar ordem de dados" assume que a categoria de referência é a que tem maior probabilidade de ocorrência do desfecho. Clique em "Continuar". Na aba "Modelo", selecione em "Fatores e covariáveis" a variável "Sexo" e mova-a para "Modelo" marcando como "Tipo: Efeito Principal". Na aba "Estatísticas", marque as opções "Razão de probabilidade" em "Estatísticas de quiquadrado", "Probabilidade do perfil" em "Tipo de intervalo de confiança" e "Incluir estimativas de parâmetro exponencial" em "Imprimir". Por fim, clique em "Ok" ou "Colar".

Os resultados dos primeiros quadros são idênticos aos obtidos para a variável "Idade", porém com ajustes para um preditor categórico. O quadro "*Omnibus Test*" nos indica que o modelo de regressão foi significativo (p < 0,001). A principal diferença está nos parâmetros estimados da tabela "*Parameter Estimates*", onde teremos um parâmetro para cada categoria da variável. Os resultados estão em comparação ao sexo feminino, cuja categoria é "2". Por isso, em "[SEXO=2]" o coeficiente está zerado, assim como o OR (Exp(B)) é um ("1"), pois a categoria está sendo comparada com ela mesma.

Assim, conclui-se que o sexo é um preditor significativo para a ocorrência de PNP (p < 0,001). A probabilidade de ter PNP é 11,4 (IC95%: 5,6-24,4) vezes maior para o sexo masculino em comparação ao sexo feminino.

Considerando o mesmo exemplo, será que um modelo multivariado (com as variáveis "Sexo", "Idade" e "Massa corporal") melhor prediz a ocorrência de PNP? Como já mencionado, podemos utilizar como critério de seleção as variáveis que se mostraram estatisticamente significativas ou com p-valor limítrofe para associação na análise univariada. Para a massa corporal, obteve-se um p-valor limítrofe para associação ($p = 0{,}062$) pelo teste U de Mann-Whitney.

Modelos com múltiplas variáveis analisam a relação entre o desfecho e as variáveis preditoras controlando a influência de outras variáveis. No menu "Analisar", "Modelos lineares generalizados", clique em "Modelos lineares generalizados...". Na aba "Tipo de modelo", selecione "Logística binária". Na aba "Resposta", selecione na tela da esquerda e insira em "Variável dependente" a variável "Polineuropatia periférica". Na mesma aba, em "Tipo de variável dependente (apenas distribuição binominal)", marque "Binário". Clique em "Categoria de referência..." e selecione a opção "Primeiro (menor valor)", onde "0" = "Não neuropata". Clique em "Continuar". Na aba "Preditores", selecione em "Fatores" a variável "Sexo" e, em "Covariáveis", as variáveis "Idade" e "Massa corporal". Como a variável "Sexo" é uma variável categórica, precisamos informar a categoria de referência. Clique em "Opções", localizado abaixo da tela de "Fatores" e selecione como "Crescente" (assume que a categoria de referência é a última em todos os fatores). Na aba "Modelo", selecione em "Fatores e covariáveis" as variáveis "Sexo", "Idade" e "Massa corporal" e as insira em "Modelo", marcando "Tipo: Efeitos principais". Na aba "Estatísticas", marque as opções "Razão de probabilidade" em "Estatísticas de quiquadrado", "Probabilidade do perfil" em "Tipo de intervalo de confiança" e "Incluir estimativas de parâmetro exponencial" em "Imprimir". Por fim, clique em "Ok" ou "Colar".

Novamente, no arquivo de saída, temos os mesmos quadros vistos na análise univariável, com a descrição do modelo, das variáveis e do ajuste do modelo. Assim como na análise univariável, o quadro "*Omnibus Test*" nos indica que o modelo de regressão foi significativo ($p < 0{,}001$). Entretanto, na análise multivariável, a tabela "*Tests of Model Effects*" ganha maior importância, pois no modelo multivariável ela não se equivale à tabela "*Omnibus Test*".

A significância de um modelo multivariado na tabela "*Omnibus Test*" denota a significância de pelo menos um dos preditores (análise global do modelo), para tanto, deve-se checar se todas as variáveis independentes devem permanecer no modelo. Para o exemplo em questão, os preditores "Sexo" e "Idade" possuem coeficiente significativamente diferente de zero, mas o "Massa corporal" não. Neste momento, deve-se ponderar se este será o modelo final ou se algum preditor deve sair ou entrar antes da análise dos coeficientes, pois os mesmos são influenciados pelo conjunto de variáveis independentes.

Constatada a significância dos preditores, a tabela "*Parameter Estimates*" apresenta os parâmetros estimados. Concluindo, pode-se afirmar que, controlando por sexo e massa corporal, as chances de ter PNP aumentam em 38,0% (IC95%: 25,6%-56,1%, p < 0,001) para cada aumento de um ano de idade. Controlando por sexo e idade, a massa corporal não prediz a ocorrência de PNP (p = 0,727). E, controlando-se o efeito da idade e da massa corporal, as chances de PNP em homens aumentam em 7,8 (IC95%: 2,5-26,2, p < 0,001) vezes em relação as mulheres.

Regressão de Poisson

O modelo de regressão de Poisson é adequado quando os dados das variáveis dependentes são contáveis, ou seja, quando se descreve o número de vezes que um evento ocorre em um espaço de observação finito. Tem por característica a análise de dados contados na forma de proporções ou razões de contagem, ou seja, leva em consideração o total de pessoas com uma determinada doença.

Esta análise de regressão é utilizada em estudos com delineamento transversal (razão de prevalências) ou longitudinal (razão de incidências), inclusive em ensaios clínicos randomizados, cujos desfechos (variáveis dependentes) são variáveis binárias ou dicotômicas do tipo óbito/não óbito ou doença/não doença, por exemplo. As variáveis independentes (fatores em estudo) podem ser categóricas ou contínuas. É imprescindível que, ao se utilizá-la para avaliar desfechos dicotômicos, seja aplicado um ajuste robusto nas variâncias.

Para exemplificar, no “Banco de dados 7.sav” (disponível em https://bit.ly/bancosdedados), vamos avaliar a magnitude da associação (através da razão de prevalências) entre o sexo e a presença de PNP, supondo um estudo de delineamento transversal. No menu “Analisar”, “Modelos lineares generalizados”, clique em “Modelos lineares generalizados...”. Como o teste está localizado no mesmo local onde foi realizada a regressão logística, clique em “Redefinir”. Na aba “Tipo de modelo”, defina a distribuição da variável dependente como “Linear de log de Poisson”. Na aba “Resposta”, selecione a variável dependente “Polineuropatia periférica” na tela da esquerda e a insira em “Variável dependente”. Na aba “Preditores”, selecione a(s) variável(is) independente(s), da mesma forma vista na regressão logística: se variável categórica, insira em “Fatores”; se contínua, mova para “Covariáveis”. Como a variável “Sexo” é uma variável dicotômica, insira em “Fatores” e defina, em “Opções”, a categoria de referência como “Crescente”. Na aba “Modelo”, selecione a variável “Sexo” em “Fatores e Covariáveis” e a insira em “Modelo” marcando a opção “Tipo: Efeitos principais”. Na aba “Estimativa”, marque “Estimador robusto” em “Matriz de covariância” e, na aba “Estatísticas”, selecione a opção “Incluir estimativas de parâmetro exponencial”. Clique em “Ok” ou “Colar”.

Nos resultados da análise, os primeiros quadros mostram as informações utilizadas na análise: a variável dependente, a distribuição e a função de ligação, a matriz de correlação, o número total de sujeitos e suas discriminações pelas variáveis independentes (fatores). O primeiro quadro que temos que verificar é o “*Omnibus Test*”, que nos indica se o modelo de regressão foi significativo. Neste caso, com $p < 0,001$, temos evidência de que o modelo é estatisticamente significativo. Na tabela “*Test of Model Effects*”, temos os resultados dos efeitos principais e da interação entre as variáveis, caso solicitada no modelo. Pelo teste de Qui-quadrado de Wald, neste exemplo, o efeito do sexo foi significativo ($p < 0,001$), ou seja, há uma diferença nas prevalências de PNP entre homens e mulheres.

Em seguida, avalia-se o valor da razão de prevalências no quadro “*Parameter Estimates*”. Neste quadro, tem-se os resultados dos coeficientes calculados pelo modelo para cada uma das variáveis, juntamente com erros-padrão, a exponencial do coeficiente (razão de prevalências) e intervalos de confiança para cada variável. Como foi indicado a categoria de referência como sendo "Crescente", a última categoria ("2" = "Feminino") foi determinada como a categoria de referência. Assim, temos que, com uma razão de prevalências de 4,608, nos homens a chance de apresentar PNP é 4,6 (IC95%: 3,0-7,2; $p < 0,001$) vezes maior em relação às mulheres.

Análise de sobrevida

Em estudos que se caracterizam por tempo de seguimento (longitudinais), além da variável de desfecho, como ocorrência de um evento, também precisamos levar em consideração o tempo médio para o evento ocorrer. Neste sentido, a análise de sobrevida possui como variável dependente o tempo de ocorrência de determinado evento. Por exemplo, a taxa de sobrevida entre pacientes submetidos a certo medicamento ou procedimento. O estado do paciente (ter ou não o evento) é uma variável dicotômica (binária) que define a situação individual de sua participação no seguimento. Este recebe um código zero ("0") se o evento não ocorrer e um ("1") se o evento ocorrer.

Uma vantagem dessa análise é permitir a utilização da informação de todos os participantes até o momento em que desenvolvem o evento ou são censurados. Ou seja, se o paciente sem o evento desistir ou abandonar o estudo no meio do seguimento, registra-se o tempo de permanência do sujeito no estudo e coloca-se o código zero ("0") (referente a não ocorrência do evento). Esses pacientes com código zero, pois não desenvolveram o evento, são chamados de censurados. As informações a partir de sujeitos censurados contribuem positivamente para a estimação do modelo.

Regressão de Cox

A regressão de Cox permite modelar, de forma uni ou multivariada, as variáveis que influenciam no tempo de ocorrência de um determinado evento. O Modelo de Cox é dito como "semi-paramétrico", pois assume que as taxas de falhas são proporcionais, ou seja, o risco de falha das variáveis é constante ao longo do tempo.

A regressão é muito utilizada para verificar inicialmente se a variável independente possui uma relação com a variável desfecho e avaliar o efeito das variáveis explicativas. Para realizar a regressão de Cox univariável são necessárias três variáveis: tempo, estado do paciente e variável independente. Já para a multivariável, utilizam-se duas ou mais variáveis independentes ou explicativas (categóricas e contínuas).

A interpretação dos coeficientes obtidos dá-se pela razão de taxas de falha ou RR. A interpretação do RR é similar à do OR da regressão logística. Quando o evento medido é óbito, a função de risco hazard ratio (HR), também chamada taxa de risco (de falha), representa a taxa instantânea de morte de um indivíduo no intervalo de tempo t a t+1, sabendo que sobreviveu até o momento t. Valores altos do HR indicam grande potencial de que o evento ocorra.

Para exemplificar, no "Banco de dados 7.sav" (disponível em https://bit.ly/bancosdedados), iremos medir o tempo na ocorrência da PNP (o evento é ter ou não PNP). Inicialmente, realizaremos uma análise de regressão de Cox univariável com a variável independente "Hipertensão" (variável categórica).

No menu "Analisar", "Sobrevivência", clique em "Regressão de Cox...". Na janela "Regressão de Cox", selecione na tela da esquerda a variável "Tempo pós-bariátrica" e a insira em "Tempo", com o auxílio do botão da seta. Em seguida, na mesma tela, selecione a variável "Polineuropatia periférica" e a mova para "Status". Clique em "Definir evento" e indique o código da categoria referente a ocorrência do evento em "Valor único" ("1" = "Neuropata"). Clique em "Continuar". Ainda na mesma janela, insira o fator ou a covariável de interesse em "Covariáveis" (neste caso, "Hipertensão"). Nos modelos multivariados coloca-se mais de uma variável nesse quadro. Como temos uma variável independente categórica, clique em "Categórico...". Selecione para a direita somente a variável "Hipertensão" e indique se a menor (ou a maior) categoria é a de referência. Neste caso, nossa referência será a primeira ("0" = "Não hipertenso"). Não é possível classificar uma categoria intermediária como referência. Não esqueça de clicar em "Alterar" se mudar de categoria. Clique em "Continuar".

Em seguida, clique em "Opções" e selecione "CI para exp(B)". Clique em "Continuar". Por fim, clique em "Ok" ou "Colar".

Quando trabalhamos com variáveis categóricas, indicamos uma das suas categorias para ser a referência. A categoria "Não hipertenso" da variável "Hipertensão" é a categoria de referência, neste exemplo. O SPSS mostra em um quadro (*Categorical Variable Codings*) no arquivo de saída, as variáveis com suas categorias de referência. O valor zero ("0") na coluna "1" mostra qual categoria é a de referência. O quadro "*Variables in the Equation*" possui as informações necessárias para podermos interpretar a relação entre as variáveis estudadas. A primeira informação encontrada é do parâmetro β (B) e seu erro-padrão (SE). A exponencial do parâmetro B (Exp(B)) e seu intervalo de confiança (IC95%) são encontrados nas últimas colunas. A Exp(B) foi de 2,780, ou seja, indivíduos hipertensos apresentam taxa de risco para PNP 2,78 (IC95%: 1,63-4,73, $p < 0,001$) vezes maior do que indivíduos não hipertensos.

Em seguida, vamos realizar uma análise de regressão de Cox univariável com a variável independente "Massa corporal" (variável quantitativa). No menu "Analisar", "Sobrevivência", clique em "Regressão de Cox...". Primeiro, na janela "Regressão de Cox", vamos "Redefinir" as opções de teste. Após, selecione na tela da esquerda a variável "Tempo pós-bariátrica" e a insira em "Tempo", com o auxílio do botão da seta. Em seguida, na mesma tela, selecione a variável "Polineuropatia periférica" e a mova para "Status". Clique em "Definir evento" e indique o código da categoria referente a ocorrência do evento em "Valor único" ("1" = "Neuropata"). Clique em "Continuar". Ainda na mesma janela, insira o fator ou a covariável de interesse em "Covariáveis" (neste caso, "Massa corporal"). Em seguida, clique em "Opções" e selecione "CI para exp(B)". Clique em "Continuar". Por fim, clique em "Ok" ou "Colar".

Quando temos uma variável contínua como variável independente, a sua interpretação é relacionada ao acréscimo de uma unidade. Neste caso, a massa corporal obteve uma Exp(B) igual a 1,078. Assim, interpretamos que para cada aumento de 1 kg na massa corporal, o risco de PNP aumenta em 7,8% (IC95%: 5,9-9,7, $p < 0,001$). As interpretações para modelo multivariável seguem da mesma forma.

Referências

Barton, B, Peat, J. Medical Statistics: A guide to SPSS. Data analysis and Critical Appraisal. Second edition. John Wiley & Sons, Ltda. 2014.

Field A. Descobrindo a estatística usando o SPSS. Tradução: Lorí Viali. 2. ed. Porto Alegre: Artmed, 2009. 684 p.

Exercícios sugeridos

1. No "Banco de dados 7.sav", supondo um estudo caso-controle, em que o critério de seleção foi a ocorrência ou não de polineuropatia periférica ("0" = "Não neuropata" e "1" = "Neuropata") após a cirurgia bariátrica, verifique se o tipo de cirurgia realizada ("1" = Bypass gástrico, RYGB e "2" = gastrectomia vertical, SG) possui associação com o desfecho. Realize a análise estatística e interprete os resultados.

2. No "Banco de dados 7.sav", supondo um estudo transversal, avalie a magnitude da associação (através da razão de prevalências) entre o sexo e a presença de hipertensão. Realize a análise estatística e interprete os resultados.