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, 2400/2° 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

10 Regressão linear simples e múltipla

Michele Strelow Moreira
Marina Petter Rodrigues
Charles Francisco Ferreira
Otto Henrique Nienov

Como normalmente existem muitas possíveis variáveis explicativas (ou independentes) em um estudo, fica difícil analisá-las de uma só vez. Por esta razão, é comum, primeiro, buscar identificar fatores associados a um determinado desfecho (resposta) realizando análises univariadas, seguidas por uma análise multivariada. Mas, nem sempre é possível ou necessário realizá-la.

A análise univariada avalia isoladamente a relação entre cada possível variável independente e a variável desfecho, sem levar em conta as demais. A análise bivariada inclui métodos de análise de duas variáveis, podendo ser ou não estabelecida uma relação de causa e efeito entre elas. São exemplos de métodos de análise bivariada, o teste de Qui-quadrado e os coeficientes de correlação (e.g. Correlação de Pearson, Correlação de Spearman).

A análise multivariada refere-se a modelos de regressão múltipla que buscam explicar uma variável desfecho com base em um conjunto de variáveis independentes. Em modelos estatísticos, denomina-se a variável dependente (desfecho) como aquela em que se tem interesse em analisar.

Análise de regressão

Em um estudo, quais os fatores que mais importam? O que podemos ignorar? Como esses fatores interagem uns com os outros? E, o mais importante, quão certos estamos sobre esses fatores? A análise de regressão ajuda a responder essas questões, visto que o coeficiente de cada variável preditora descreve a contribuição relativa de cada variável ao desfecho.

A análise de regressão é uma forma de prever algum resultado, relacionada à variável de resposta ou dependente (principal fator que você está tentando entender ou prever), a partir de uma ou mais variáveis preditoras, explicatórias ou independentes (fatores que você suspeita terem algum impacto no desfecho). Desta forma, conseguimos avaliar o impacto de cada variável independente sobre o desfecho.

Você pode estar se perguntado se, por exemplo, uma correlação não poderia fazer isso. A correlação pode ser uma ferramenta bastante útil, mas ela não nos informa sobre o poder preditivo das variáveis. Na análise de regressão, ajustamos um modelo preditivo aos nossos dados e, então, usamos esse modelo para prever valores da variável dependente a partir de uma ou mais variáveis independentes.

Quando queremos avaliar a influência de um conjunto de fatores sobre doenças ou outras características de interesse, a análise de regressão é o método estatístico utilizado, estabelecendo uma equação que simula os relacionamentos entre a variável dependente e os fatores que se deseja investigar (variáveis independentes).

Ainda podemos classificar a regressão conforme a quantidade de fatores preditores que se deseja investigar (variáveis independentes). Quando existe apenas uma variável independente, ou seja, uma única variável preditora, a definimos como sendo simples. Por outro lado, em regressões múltiplas, mais de uma variável independente é utilizada para prever o desfecho.

Construindo modelos de regressão

O modelo de regressão começa quando um pesquisador deseja descrever matematicamente a relação entre algumas variáveis preditoras (independentes) e a variável desfecho (dependente). Em um projeto de pesquisa, normalmente muitas variáveis são julgadas importantes para responder uma pergunta. Contudo, no modelo de regressão, incluímos apenas algumas delas. O pesquisador deve tentar eliminar as variáveis que não estão relacionadas e incluir apenas aquelas que ele acredite terem algum relacionamento verdadeiro, sendo possível considerar muitos modelos diferentes ao longo do percurso.

Mas, como determinar as variáveis que farão parte do modelo de regressão? Geralmente se busca o modelo mais parcimonioso que explique os dados. No entanto, para a obtenção deste "modelo mais parcimonioso" final, algumas etapas devem ser percorridas. Deixamos aqui alguns tópicos que podem auxiliar você na elaboração do modelo final de regressão:

1) Tamanho de amostra para regressão: existe uma divergência sobre o tamanho da amostra ideal para o modelo de regressão, a depender do número de variáveis preditoras. Alguns autores indicam o tamanho amostral de 10 indivíduos para cada previsor (variável independente) no modelo, outros apontam 15 indivíduos por variável independente. Assim, se você tivesse cinco variáveis independentes no modelo, seria necessário 50 ou 75 casos, respectivamente. O tamanho da amostra necessário irá depender do tamanho de efeito que estamos tentando detectar.

2) Número de variáveis incorporadas no modelo: um modelo pouco especificado tende a produzir estimativas não reais, assim como um modelo muito especificado pode produzir estimativas menos precisas. Se um excesso de variáveis independentes forem incluídas, maiores se tornam os erros e mais dependente será o modelo dos dados observados. Portanto, um modelo com as variáveis independentes mais pertinentes terá menor risco de viés e apresentará as estimativas mais precisas. Minimizando o número de variáveis independentes, o modelo matemático final terá maior probabilidade de ser generalizável.

3) Avaliação da multicolinearidade: variáveis independentes altamente correlacionadas entre si são denominadas multicolineares. Muitos tipos de modelos de regressão pressupõem que a multicolinearidade não deve estar presente no conjunto de dados, por causar problemas na classificação de variáveis com base em sua importância ou dificultar o trabalho na seleção da variável independente mais importante. A multicolinearidade indica que duas variáveis independentes (fatores preditores) analisados possuem explicação para o mesmo segmento do desfecho analisado.

4) Análise univariada: na prática, existe um grande número de variáveis explicativas (independentes). Logo, a análise univariada pode servir como critério de seleção das variáveis que entrarão em um modelo de regressão. Como se trata de uma etapa inicial e não-definitiva da análise de dados, podemos ser menos rigorosos e adotar níveis de significância maiores que o usual (por exemplo, $p < 0,150$, $p < 0,200$ ou $p < 0,250$) para não correr o risco de desprezar variáveis importantes. Denominamos estas de variáveis tendenciosas, ou seja, que apresentam um p-valor limítrofe para associação. Considere, também, aspectos como a análise dos resíduos do teste de Qui-quadrado, para verificar a indicação dessa associação; a direção (positiva ou inversa) e a intensidade (fraca ou forte) dos coeficientes de correlação; o valor do índice F da Análise de Variância (ANOVA), onde um valor elevado significa que há alguma diferença entre os grupos capaz de ser expressa adequadamente por meio de um modelo de regressão.

5) Aspectos encontrados na literatura: pesquise o que outros autores fizeram e incorpore essas descobertas na construção do seu modelo. Antes de iniciar a análise de regressão, desenvolva uma ideia de quais são as variáveis importantes, juntamente com suas relações, sinais de coeficiente e magnitudes de efeito. Com base nos resultados de outras pessoas, fica mais fácil coletar os dados corretos e especificar o melhor modelo de regressão sem a necessidade de usar mineração de dados. Em seguida, verifique se o modelo obtido se alinha com a teoria e faça ajustes. Por exemplo, com base na teoria, você pode incluir uma variável preditora no modelo mesmo que seu valor-p não seja significativo.

6) Variáveis de controle: insira no modelo variáveis que você julga importantes, inclusive para controlar variáveis do próprio modelo, mesmo estas não sendo significativas.

7) Teste modelos diferentes: após selecionadas as variáveis candidatas para o modelo, a última etapa é modelar juntas todas essas variáveis e testar modelos diferentes. Dessa forma, as variáveis mais associadas ao desfecho de interesse

são selecionadas para o modelo final. Esta etapa consiste em avaliar simultaneamente, um modelo de cada vez, o efeito das variáveis selecionadas sobre a resposta. Modelos mais simples geralmente produzem previsões mais precisas. Existem diferentes formas de abordagem, conforme apresentado no quadro 1.

Quadro 1. Formas de abordagem do modelo de regressão.

<table>
<tr><th>Abordagem</th><th colspan="2">Descrição</th></tr>
<tr><td>Método ENTER (forma manual)</td><td colspan="2">É o método mais indicado para se testar hipóteses, uma vez que é o pesquisador que determina a relevância das variáveis no modelo. Pode-se iniciar com a variável mais relevante e avalia-se a inclusão, uma a uma, das seguintes, ou, se a amostra é suficientemente grande (pelo menos 10 casos com o desfecho menos frequente por variável), pode-se começar a modelagem com todas as variáveis selecionadas, retirando-se, uma a uma, a menos relevante.</td></tr>
<tr><td rowspan="2">Método STEPWISE (forma automática)</td><td rowspan="2">É indicado quando não existem pesquisas de base para auxiliar na escolha das variáveis do modelo. No entanto, pode ser influenciado por variações aleatórias dos dados. As variáveis são selecionadas automaticamente, sem o controle do pesquisador, tanto para inclusão quanto para exclusão, de uma maneira sequencial, baseada apenas em critérios estatísticos. Há duas modelagens:</td><td>FORWARD: as variáveis são adicionadas uma a uma ao modelo, sendo a primeira variável candidata aquela mais significativa e, assim sucessivamente.</td></tr>
<tr><td>BACKWARD: todas as variáveis selecionadas pelo pesquisador são incluídas, sendo retirada, uma a uma, a variável menos significativa. Esta abordagem permite a comparação de modelos que talvez não fossem gerados pelo método manual.</td></tr>
</table>

É comum que variáveis estatisticamente significativas na análise univariada percam a importância na multivariada. Isso porque, quando analisamos o fator isoladamente, não estamos levando em conta outras características que podem estar relacionadas a este fator (viés). Embora seja menos comum, é possível também que uma variável não-significativa na univariada passe a ser significativa na multivariada, principalmente quando houver interação entre os fatores. A existência de interação significa que o efeito de um fator sobre a resposta depende de outro fator.

Regressão linear

A regressão linear estima os coeficientes da equação linear, com o objetivo prever o comportamento de uma variável dependente (quantitativa) em função de uma (simples) ou mais (múltipla) variáveis independentes (quantitativas ou qualitativas binárias). Assumimos a relação entre as variáveis por meio de uma reta (Figura 1) e, assim, utilizamos o resultado da função dessa reta para estimar valores, quando conhecemos as variáveis que afetam o seu comportamento.

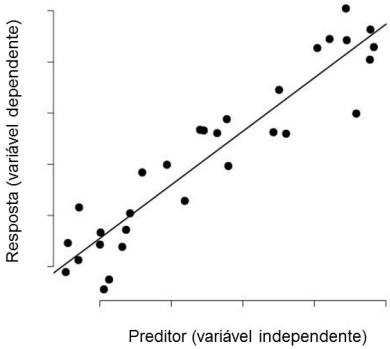

Figura 1. Reta da regressão linear.

No eixo Y da representação gráfica, temos a variável dependente (desfecho) que estamos tentando descobrir e, no eixo X da representação gráfica, as variáveis independentes (preditoras) que exercem influência sobre a variável dependente. A variável Y deve ter distribuição normal ou aproximadamente normal. Os parâmetros de distribuição β_0 e β_1 são denominados de coeficientes da regressão: β_0 é o intercepto em Y da equação de regressão (é o valor de Y quando X é igual a 0) e β_1 refere-se à inclinação da reta de regressão (indica a mudança na média de Y quando X é acrescido de uma unidade). O erro (ε) representa as influências não controladas, ou seja, influências que a variável dependente possui além da exercida pelas variáveis no modelo. Conforme mencionado, o modelo para essa relação entre as variáveis é linear, sendo representado pela equação:

$$Y = \beta_0 + \beta_1 * X_1 + ... + \varepsilon$$

Onde:

Y = variável dependente da equação de regressão linear

β_0 = intercepto em Y da equação de regressão linear
β_1 = inclinação da reta de regressão linear
ε = erro da equação de regressão linear.

Tanto na regressão linear simples quanto na múltipla, as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. Essa validação é feita através da análise dos resíduos, um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. Esses resíduos representam o erro que está presente no modelo, correspondendo a diferença entre o valor previsto e o observado, ou seja, a distância do ponto até a reta que foi prevista no modelo. Se o modelo se ajusta bem aos dados da amostra, todos os resíduos devem ser pequenos (se o modelo aderir perfeitamente aos dados, todos os pontos estarão sobre a linha de regressão e todos os resíduos serão iguais a zero). Se o modelo não tiver uma boa aderência aos dados da amostra, os resíduos serão grandes. Além disso, se qualquer caso destacar-se por ter um grande resíduo, ele poderá ser um valor atípico.

Os pré-requisitos para a realização dessa análise de resíduos são:

1) Normalidade dos resíduos: o erro (diferença entre a variável dependente e a estimação feita pelo modelo) deve ter distribuição normal. A distribuição normal dos resíduos é essencial para que os resultados do ajuste do modelo de regressão sejam confiáveis. Como visto no capítulo 7, podemos verificar essa suposição por meio de uma inspeção visual (histograma de resíduos padronizados e gráficos de probabilidade normal comparando a distribuição de resíduos padronizados com uma distribuição normal, também chamado de P-P plot) e de testes estatísticos (Shapiro-Wilk e Kolmogorov-Smirnov) para os resíduos.

2) Homocedasticidade: a variância do erro experimental (ε) para observações distintas deve ser constante, ou seja, deve ser homocedástico. Isso pode ser visto em um gráfico de resíduos do tipo *scatterplot* (Figura 2). Neste gráfico, é apresentada a relação dos valores preditos no eixo X e

dos valores residuais no eixo Y, onde a distribuição destes resíduos não deve exibir nenhum padrão óbvio. Desta forma, se os pontos estão aleatoriamente distribuídos, sem nenhum comportamento ou tendência, temos indícios de que a variância dos resíduos é homoscedástica. Já a presença de um "funil", é um indicativo da presença de heterocedasticidade. Além disso, os testes de Breusch-Pagan e de Goldfeld-Quandt também podem ser utilizados para testar se os resíduos são homoscedásticos. Baseado no teste multiplicador de Lagrange, o teste de Breusch-Pagan é bastante utilizado para testar a H_0 de que as variâncias dos erros são iguais versus a H_A de que as variâncias dos erros são uma função multiplicativa de uma ou mais variáveis, sendo que esta(s) variável(eis) pode(m) pertencer ou não ao modelo em questão. É indicado para grandes amostras e quando a suposição de normalidade nos erros é assumida. O teste de Goldfeld-Quandt também é utilizado para testar a homoscedasticidade dos resíduos, mas limita-se por exigir que a amostra seja relativamente grande. Se não houver homocedasticidade, será necessário transformar os dados, como por exemplo, logaritmo.

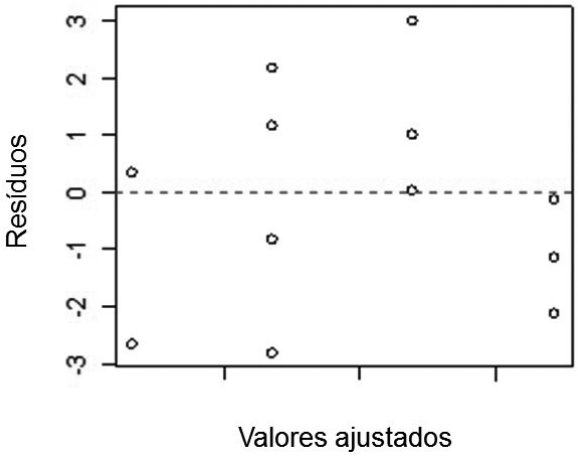

Figura 2. Gráfico de *scatterplot* dos resíduos *versus* valores ajustados.

3) Multicolinearidade: o valor do erro para uma observação deve ser independente dos valores das variáveis do modelo e do erro das outras observações, ou seja, não devemos ter uma relação entre as variáveis independentes. Esse diagnóstico é importante quando utilizarmos duas ou mais variáveis independentes no modelo de regressão múltipla. Desta forma, se não houver nenhum relacionamento entre elas, dizemos que são ortogonais. Na prática, é muito difícil que as variáveis de entrada sejam ortogonais e, felizmente, a falta de ortogonalidade não é grave. Mas, se as variáveis forem muito correlacionadas, as inferências baseadas no modelo de regressão podem ser errôneas ou pouco confiáveis. Um dos diagnósticos para multicolinearidade pode ser baseado na estatística K, uma análise de estrutura em função dos autovalores, onde um $K < 10$ indica que não há problemas de multicolinearidade; enquanto que $10 < K < 100$ indica uma multicolinearidade moderada e; para $K > 100$, temos uma multicolinearidade severa. Outro diagnóstico pode ser realizado pelo Fator de Inflação de Variância (VIF), que deve ser menor que 10. O VIF quantifica a extensão da correlação entre um preditor e os outros preditores em um modelo. Se maiores que 10, há indicação de multicolinearidade, ou seja, é difícil avaliar com precisão a contribuição dos preditores para um modelo. Ou conforme o índice de Tolerância (1/VIF), que deve ser maior que 0,2.

4) Autocorrelação: os resíduos devem ser independentes, isto é, não correlacionados. Para verificar esta suposição, utilizamos o teste de Durbin-Watson, que testa a suposição de independência dos erros. No entanto, diferentemente dos testes de hipóteses vistos anteriormente, devemos analisar os valores da estatística D, apresentados na tabela 1. Seguindo a linha do respectivo tamanho amostral (N) e a coluna do número de variáveis independentes consideradas (X), temos, para cada nível de significância (α): se D é maior que D-U, aceita-se H_0 (resíduos não correlacionados); se D menor que D-L, rejeita-se H_0 (resíduos correlacionados) ou; se D estiver entre D-L e D-U, o teste é inconclusivo.

Tabela 1. Valores da estatística D do testes de Durbin-Watson.

Observações		Número de variáveis independentes (X)							
		1		2		3		4	
N	α	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U
15	0,050	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97
	0,010	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70
20	0,050	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83
	0,010	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57
25	0,050	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77
	0,010	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52
30	0,050	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74
	0,010	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51
40	0,050	1,44	1,54	1,39	1,60	1,34	1,66	1,39	1,72
	0,010	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52
50	0,050	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72
	0,010	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54
60	0,050	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73
	0,010	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56
80	0,050	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74
	0,010	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60
100	0,050	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76
	0,010	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63

5) Pontos influentes e valores atípicos (também denominados *outliers)*: o valor atípico é um caso que difere substancialmente da maioria dos dados, que pode ser vista em gráficos de dispersão. Além de procurar valores atípicos olhando para os erros do modelo, também é possível buscar certos casos que influenciam os parâmetros do modelo. Um ponto influente é uma observação que pode influenciar em qualquer parte da análise de regressão, como a previsão de uma resposta ou o resultado de testes de hipóteses. Os valores extremos

têm potencial para serem influentes, podendo introduzir tendenciosidade no modelo, pois eles irão afetar os valores dos coeficientes de regressão estimados. Mas, temos que investigar para avaliar o quanto eles são influentes. Se um valor extremo for influente, ele interfere sobre a função de regressão ajustada (a inclusão ou não do ponto modifica substancialmente os valores ajustados). Por outro lado, uma dada observação pode ser considerada um valor atípico e não ser um ponto influente. Da mesma forma, podemos ter pontos que influenciam na análise de regressão, mas não são valores atípicos. Esse tipo de análise pode ajudar a determinar se o modelo de regressão é estável por toda a amostra ou se ele pode estar sendo influenciado somente por poucos casos.

Verificados os pré-requisitos da análise de resíduos, prosseguimos com a análise da regressão linear. Para verificar se a regressão é significativa, devemos inicialmente observar o resultado da ANOVA. Se o *p*-valor > 0,050, não temos evidência para dizer que o modelo de regressão linear é importante para explicar a variável desfecho, ou seja, as variáveis independentes não exercem influência na variável dependente. Por outro lado, se o p-valor ≤ 0,050, podemos dizer que ao menos uma das variáveis do modelo é importante para explicar a variável desfecho, ou seja, pelo menos uma variável independente exerce influência na variável dependente.

Para cada parâmetro estimado é realizado um teste de significância (teste t de Student), onde a H_0 é de que o respectivo coeficiente é igual a zero e, a H_A é de que o coeficiente é diferente de zero, sendo esta última significativa quando p-valor ≤ 0,050.

Constatada a significância do modelo, também verificamos o coeficiente de determinação (R^2). O R^2 é um indicador para a análise do ajuste do modelo adotado, indicando a proporção da variabilidade de Y que pode ser explicada pela variabilidade das variáveis X. O R^2 pode variar de 0 a 1: quanto mais próximo de zero, indica que o modelo não explica a variabilidade dos dados de desfecho ao redor de sua média; quanto mais próximo de um, indica que o modelo explica toda a variabilidade dos dados de desfecho ao redor de sua média.

Regressão linear simples

Para melhor compreender a análise de regressão, começaremos construindo uma regressão linear simples, onde há apenas uma variável independente. Logo, a equação resume-se a:

$$Y = \beta_0 + \beta_1 * X_1 + ... + \varepsilon$$

No "Banco de dados 2.sav" (disponível em https:// bit.ly/ bancosdedados), vamos verificar se a estatura depende da idade, supondo que a estatura tem distribuição normal. Lembre-se: é importante conhecer o banco de dados. Tome algum tempo para examiná-lo e conhecer as variáveis (Mínimo? Máximo? Média? Mediana? Intervalo interquartílico? Teste de normalidade?).

No menu "Analisar", "Regressão", clique em "Linear...". Na janela "Regressão linear", selecione a variável "Estatura" e a insira em "Dependente" e, a variável "Idade", em "Independente(s)". Em "Estatísticas", selecione as opções "Estimativas" e "Intervalo de confiança" em "Coeficiente de regressão"; "Ajuste do modelo", "Descritivos" e "Diagnóstico de colinearidade" (para avaliar a multicolinearidade) e; em "Residuais", "Durbin-Watson" (para verificar a autocorrelação) e "Diagnóstico por caso" (para analisar os valores atípicos além de três desvios padrão). Clique em "Continuar". Em "Diagramas", defina os valores preditos padronizados (ZPRED) no "X" e os resíduos padronizados (ZRESID) no "Y". ZPRED são formas padronizadas dos valores previstos pelo modelo e ZRESID são as diferenças padronizadas entre os dados observados e os valores que o modelo prevê. Marque as opções de "Histograma" e "Diagrama de probabilidade normal" em "Diagramas residuais padronizados". Clique em "Continuar".

É necessário salvar os resíduos e os valores previstos da análise de regressão. Para isso, clique em "Salvar" e selecione "Não padronizados" em "Valores previstos", "Não padronizado" em "Residuais" e, no "Intervalos de previsão", selecione "Média". Os valores preditos não padronizados correspondem ao valor que o modelo prediz para a variável dependente. Os residuais não padronizados indicam a diferença entre um valor observado e o valor predito pelo modelo. Os intervalos de predição para média indicam limites inferior e superior (duas variáveis) para o intervalo de predição da resposta média predita. Clique em "Continuar". Por fim, clique em "Ok" ou "Colar".

A análise de regressão apresenta uma série de quadros no arquivo de saída. No primeiro (*Descriptive Statistics*), temos os valores de média (*Mean*) e desvio padrão (*Std. Deviation*) e tamanho da amostra (N). Em seguida (*Correlations*), o coeficiente de correlação de Pearson. É importante verificarmos se a relação entre a variável dependente e as variáveis independentes é de fato linear. Conforme o gráfico (Figura 3), é visível a relação linear, sendo negativa fraca (r = -0,133, *p* = 0,002).

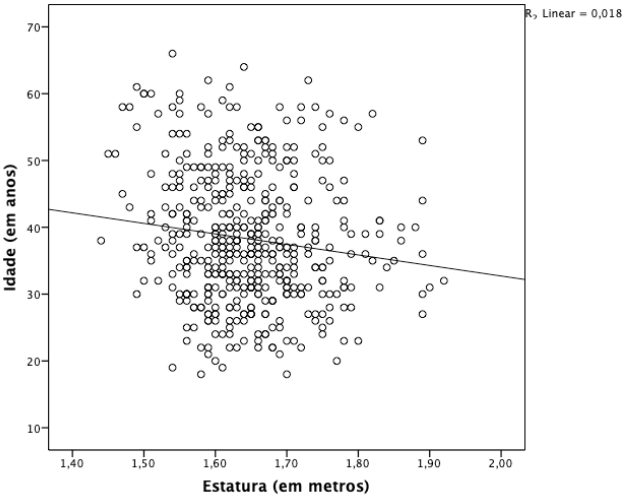

Figura 3. Correlação entre idade (em anos) e estatura (em metros).

Antes de validar os pré-requisitos da análise dos resíduos, vamos verificar a análise da regressão. O primeiro quadro que devemos analisar é o da ANOVA, que mostra o índice F. Esta análise é bastante familiar, pois foi vista no capítulo 7. Neste exemplo, temos como a H_0 de que a idade não explica a variação da estatura (R^2 = 0), enquanto que na H_A a idade explica a variação da estatura. O valor do índice F (8,459) é significativo, portanto, podemos assumir que o modelo explica uma quantidade significativa da variação da estatura em função da variabilidade da idade. Com *p*-valor igual a 0,004, o teste global mostra que há regressão entre as variáveis.

Cada variável é multiplicada por uma constante (β), que corresponde ao valor B para a constante (intercepto). O teste t verifica se essa constante é (H_0) ou não (H_A) igual a zero. No segundo

quadro a ser analisado (*Coefficients*), a variável "Idade" possui β = -0,001, sendo este significativamente diferente de zero pelo valor do índice t (p = 0,004). Assim, para cada unidade a mais na idade, a estatura média é menor em 0,001 metros.

Por fim, verificamos a qualidade do modelo (*Model Summary*), através do coeficiente de correlação (R), que é idêntico ao coeficiente de correlação de Pearson; do coeficiente de determinação (R^2) e; do coeficiente de determinação ajustado (R^2adj), que penaliza por cada variável colocada no modelo, baixando o coeficiente de determinação quando o modelo é construído com mais de uma variável. Estes servem como um parâmetro para comparar modelos de regressão.

Neste exemplo, temos que a correlação entre as variáveis é de 0,133 e a proporção da variabilidade de estatura que pode ser explicada pela variabilidade da idade é de 1,8% (R^2). A regressão linear simples mostrou que a idade prevê a estatura, [$F(1,467)$ = 8,459, p = 0,004, R^2 = 0,018]. A estatura, em metros, corresponde a 1,690 - 0,001 x (idade), sendo a idade medida em anos.

Para validar a regressão, devemos realizar a análise de resíduos. Lembre-se que a variável dependente é a estatura, pois é ela que depende da idade, e não o contrário. Os gráficos podem ajudar na validação das suposições de normalidade, linearidade e igualdade das variâncias. Além disso, são úteis para detectar valores discrepantes, observações atípicas e casos influentes.

O primeiro passo é a verificação do gráfico de *scatterplot* dos resíduos versus valores preditos (Figura 4). Conforme o gráfico, é visível que os pontos estão aleatoriamente distribuídos, sem nenhum comportamento ou tendência (sem comportamento cônico) e, assim, temos indícios de que a variância dos resíduos é homoscedástica.

A colinearidade (ou multicolinearidade) é a situação indesejável em que uma variável independente é uma função linear de outras variáveis independentes. Para avaliar a multicolineariedade, o quadro "*Coeficients*" apresenta o índice de Tolerância e o VIF, ambos quantificados em 1, ou seja, não há problemas de multicolinearidade. Para verificar a autocorrelação,

o quadro "*Model Summary*" traz a estatística D de Durbin-Watson, que foi de 2,038. Assim, para N > 100 e X igual a 1 (modelo com uma única variável independente), temos D maior que D-U, ou seja, podemos afirmar que não há autocorrelação.

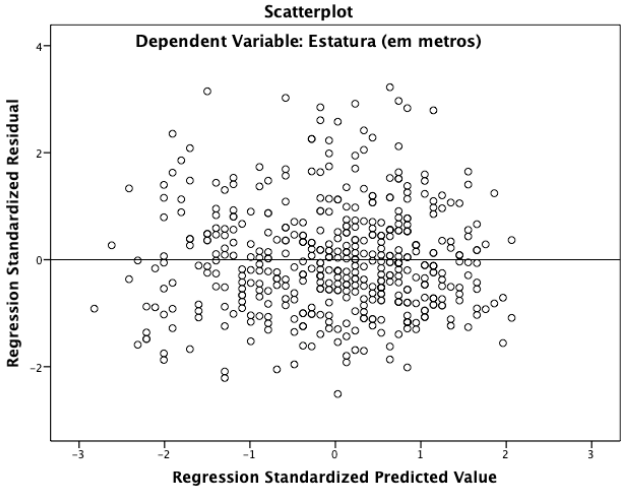

Figura 4. Gráfico de *scatterplot* dos resíduos *versus* valores preditos.

No quadro "*Casewise Diagnostics*", temos indicado três casos com valores atípicos (acima de três desvios-padrões). Neste caso, devemos realizar a análise de regressão com e sem os valores atípicos, para verificar a influência destes no modelo. Por fim, verificamos a normalidade dos resíduos. Pode-se utilizar um histograma ou um gráfico P-P *plot* normal (Figura 5). No gráfico P-P *plot*, é possível observar que os pontos geralmente seguem a reta normal (diagonal) sem desvios fortes. Isso indica que os resíduos aparentemente estão normalmente distribuídos, mas o ideal é confirmar com o teste de normalidade de Shapiro-Wilk.

Para solicitar o teste de normalidade dos resíduos, no menu "Analisar", "Estatística descritivas", clique em "Explorar..." e utilize a variável de resíduos criada na análise de regressão. O quadro "*Test of Normality*" apresenta o teste de normalidade com p-valor inferior a 0,001. Desta forma, rejeitamos a hipótese de normalidade dos resíduos, ou seja, o teste de regressão não é válido pois não cumpriu o pré-requisito de normalidade dos resíduos.

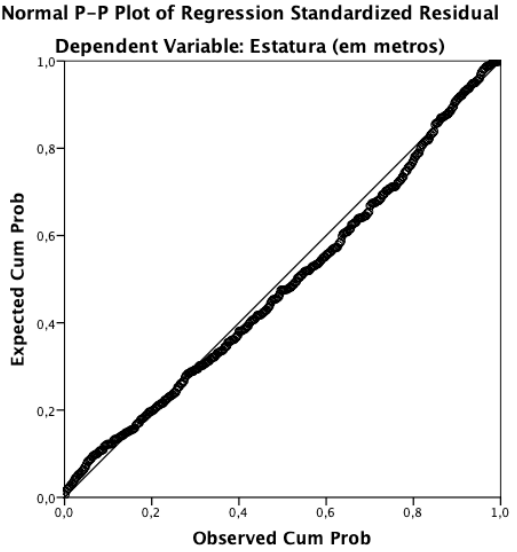

Figura 5. Gráfico P-P *plot* para verificar a normalidade dos resíduos.

Regressão linear múltipla

Existem muitos testes estatísticos que avaliam o grau de relação entre duas variáveis, como os coeficientes de correlação, por exemplo, mas estes não são capazes de explorar relações multivariadas, como mencionado. Comumente temos um desfecho que não depende apenas de uma variável, mas de várias. Nesses casos podemos utilizar a regressão linear múltipla, ou seja, com mais de uma variável independente, adotando a seguinte equação: $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 ... + \beta_k * X_k + \varepsilon$.

Assim, Y é a variável de resultado (desfecho), β_1 é o coeficiente do primeiro previsor (variável independente X1), β_2 é o coeficiente do segundo previsor (variável independente X_2), β_k é o coeficiente do k-ésimo previsor (variável independente X_k) e ε é a diferença entre o valor previsto e o observado de Y. Como na regressão linear simples, na múltipla também devemos validar as suposições do modelo ajustado para que os resultados sejam confiáveis, através da análise de resíduos.

Para exemplificar, no "Banco de dados 2.sav", vamos verificar se a estatura depende da idade e do sexo, supondo que a estatura tem distribuição normal. No menu "Analisar", "Regressão", clique

em "Linear...". Na janela "Regressão linear", selecione a variável "Estatura" e a insira em "Dependente" e, as variáveis "Idade" e "Sexo", como "Independente(s)". Em "Estatísticas", selecione as opções "Estimativas" e "Intervalo de confiança" em "Coeficiente de regressão"; "Ajuste do modelo", "Descritivos" e "Diagnóstico de colinearidade" (para avaliar a multicolinearidade) e; em "Residuais", "Durbin-Watson" (para verificar a autocorrelação) e "Diagnóstico por caso" (para analisar os valores atípicos além de três desvios padrão). Clique em "Continuar".

Em "Diagramas", defina o valor previsto (ZPRED) no "X" e o resíduo ajustado (ZRESID) no "Y". Marque as opções de "Histograma" e "Diagrama de probabilidade normal" em "Diagramas residuais padronizados". Clique em "Continuar". É necessário salvar os resíduos e os valores previstos da análise de regressão. Para isso, clique em "Salvar" e selecione "Não padronizados" em "Valores previstos", "Não padronizado" em "Residuais" e, no "Intervalos de previsão", selecione "Média". Clique em "Continuar". Por fim, clique em "Ok" ou "Colar".

No exemplo, note que a variável "Sexo" não é quantitativa, mas sim, qualitativa. Nestes casos, é comum a codificação indicadora, na qual as categorias são representadas por "1" ou "2", conforme rótulos calibrados para esta variável, por exemplo. Para a análise de regressão, deve-se tratá-la como as demais variáveis independentes.

Da mesma forma, no arquivo de saída, a análise de regressão apresenta vários quadros. No primeiro (*Descriptive Statistics*), temos os valores de média (*Mean*) e desvio padrão (*Std. Deviation*) e tamanho da amostra (N). Em seguida (*Correlations*), o coeficiente de correlação de Pearson. É importante verificarmos se a relação entre a variável dependente e as variáveis independentes é de fato linear. Conforme o gráfico (Figura 6), é visível a relação linear negativa fraca com a idade ($r = -0,133$, $p = 0,002$) e negativa forte com o sexo ($r = -0,628$, $p < 0,001$).

Seguindo a interpretação vista no exemplo anterior, primeiramente devemos analisar o teste global (ANOVA) e o teste t para os coeficientes. Com p-valor menor que 0,001, para ambos, o teste global mostra que há regressão entre as variáveis e, pelo teste t, os coeficientes são significativamente diferentes de zero.

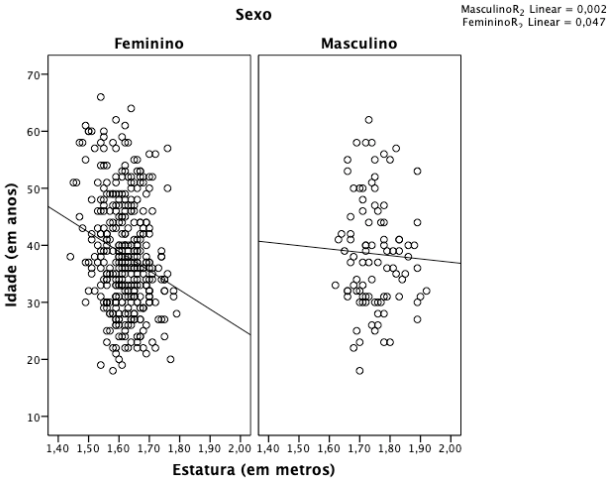

Figura 6. Correlação entre sexo, idade (em anos) e estatura (em metros).

Conforme a equação, 1,927 é a estatura constante, independentemente da idade e do sexo. O coeficiente angular da idade é negativo, indicando dependência inversa, sendo de -0,001 a variação na estatura em relação à unidade de variação na idade. Já o coeficiente para a variável sexo representa diferenças entre médias para cada grupo, isto é, -0,131 representa o efeito linear do grupo não omitido, que recebeu o código um (sexo masculino).

Em seguida, analisamos o coeficiente de determinação ajustado. Temos que a correlação entre as variáveis é de 0,643 e a proporção da variabilidade de estatura que pode ser explicada pela idade e pelo sexo é de 41,1% (R^2_{adj}). Com isso, concluímos que foi utilizada a regressão linear múltipla para verificar se idade e sexo são capazes de prever a estatura. A análise resultou em um modelo estatisticamente significativo [$F(2,466) = 164,519$; $p < 0,001$; $R^2 = 0,411$]. A idade ($\beta = -0,140$; $t = -3,947$; $p < 0,001$) e o sexo ($\beta = -0,629$; $t = -17,745$; $p < 0,001$) são preditores da estatura.

No entanto, lembre-se que é necessário validarmos os resultados pela análise dos resíduos. De acordo com o VIF e o índice de Tolerância apresentados no quadro "*Coefficients*[a]", não há presença de multicolinearidade. A estatística D de Durbin-Watson está apresentada em "*Model Summary*" e foi igual a 2,051. Isto é,

para N > 100 e X igual a 2, temos D maior que D-U, logo, podemos afirmar que não há autocorrelação. Analisando o *scatterplot* (Figura 7), sugere-se a homogeneidade de variâncias, pois não há um parâmetro de distribuição das variáveis. Verificada a normalidade dos resíduos pelo teste de Shapiro-Wilk (p = 0,197), podemos afirmar que o modelo de regressão é válido para esta análise.

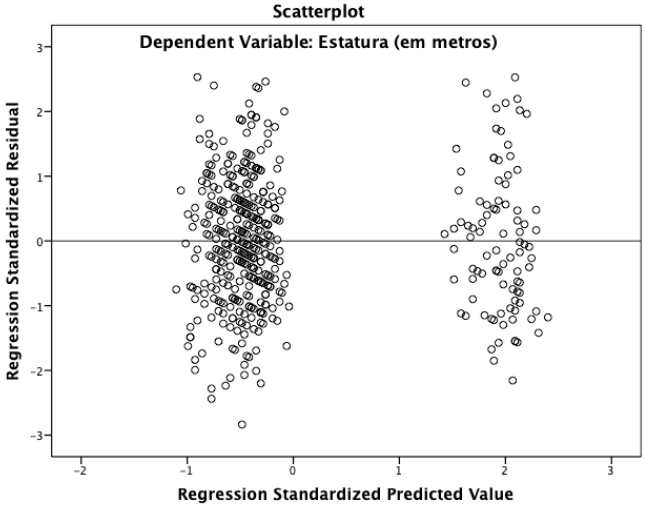

Figura 7. Gráfico de *scatterplot* dos resíduos *versus* valores preditos.

Você pode agrupar variáveis independentes em blocos e especificar métodos de entrada diferentes para diferentes subconjuntos de variáveis. Quando pretendemos testar diferentes modelos, de forma manual, seleciona-se, na janela "Regressão linear", o "Método inserir" (*Enter*) e, inclui-se uma variável por vez em cada "Bloco", utilizando as opções "Anterior" e "Próximo" para definir a ordem de entrada das variáveis no modelo (conforme Quadro 1, página 201). No arquivo de saída, obtém-se os resultados para cada modelo enumerado.

Também pode-se construir os modelos de forma automática (*stepwise*), utilizando os modelos "Avançar" (*Forward*) (as variáveis são inseridas sequencialmente no modelo, sendo a primeira variável aquela com a maior correlação positiva ou negativa com a variável dependente e, assim, sucessivamente),

"Retroceder" (*Backward*)" (todas as variáveis são inseridas na equação e, em seguida, todas as variáveis selecionadas pelo pesquisador são incluídas, sendo retirada, uma a uma, a variável menos significativa), "Remover" (todas as variáveis em um bloco são removidas em um único passo) e "Por etapa" (em cada passo, a variável independente fora da equação que tiver o menor índice F será inserida, se esse índice for suficientemente pequeno e, as variáveis que já estiverem na equação da regressão serão removidas se o índice F for suficientemente grande) no "Método".

Referências

Callegari-Jacques SM. Bioestatística: princípios e aplicações. Porto Alegre: ArtMed, 2011. 255p.

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

Harrell , FE. General Aspects of Fitting Regression Models. In: Regression Modeling Strategies. Springer Series in Statistics. Cham: Springer International Publishing; 2015:13-44.

Heumann C, Schomaker M, Shalabh. Linear Regression. In: Introduction to Statistics and Data Analysis. Cham: Springer International Publishing; 2016:249-295.

Montgomery DC. Fitting Regression Models. In: Design and Analysis of Experiments Eighth Edition. Vol 2. ; 2012:449-475.

Exercícios sugeridos

1. No "Banco de dados 2.sav", verifique se a pressão arterial sistólica depende da massa corporal, supondo distribuição normal da variável dependente. Realize a análise estatística, interprete os resultados e verifique os pré-requisitos da análise de resíduos.

2. No "Banco de dados 2.sav", verifique se a pressão arterial sistólica depende da circunferência da cintura e do sexo, supondo distribuição normal da variável dependente. Realize a análise estatística, interprete os resultados e verifique os pré-requisitos da análise de resíduos.