Revista de Sociologia e Política
ISSN: 0104-4478
ISSN: 1678-9873
Universidade Federal do Paraná

Fernandes, Antônio Alves Tôrres; Figueiredo, Dalson Britto;
Rocha, Enivaldo Carvalho da; Nascimento, Willber da Silva

Read this paper if you want to learn logistic regression

Revista de Sociologia e Política, vol. 28, no. 74, 2020, October-December
Universidade Federal do Paraná

DOI: 10.1590/1678-987320287406en

Available in: http://www.redalyc.org/articulo.oa?id=23865754006

Artigos Originais

Revista de Sociologia e Política

DOI 10.1590/1678-987320287406

Leia este artigo se você quiser aprender regressão logística

Antônio Alves Tôrres Fernandes[I], Dalson Britto Figueiredo Filho[I], Enivaldo Carvalho da Rocha[I], Willber da Silva Nascimento[I]

[I]Programa de Pós-Graduação em Ciência Política, Universidade Federal de Pernambuco, Recife, PE, Brasil.

RESUMO Introdução: E se a minha variável resposta for categórica binária? Este artigo apresenta uma introdução intuitiva à regressão logística, técnica estatística mais adequada para lidar com variáveis dependentes dicotômicas. **Materiais e Métodos:** estimamos o efeito dos escândalos de corrupção sobre a chance de reeleição de candidatos concorrentes a deputado federal no Brasil a partir dos dados de Castro e Nunes (2014). Em particular, mostramos a implementação computacional no R e explicamos a interpretação substantiva dos resultados. **Resultados:** disponibilizamos todos os materiais de replicação, permitindo que estudantes e profissionais utilizem os procedimentos discutidos aqui em suas atividades de estudo e pesquisa. **Discussão:** esperamos incentivar o uso da regressão logística e difundir a replicabilidade como ferramenta de ensino de análise de dados.

PALAVRAS-CHAVE: regressão; regressão logística; replicação; métodos quantitativos; transparência.

Recebido em 19 de Outubro de 2019. Aprovado em 7 de Maio de 2020. Aceito em 16 de Maio de 2020.

I. Introdução[1]

[1]Materiais de replicação disponíveis em: <https://osf.io/nv4ae/>. Este artigo também se beneficiou dos comentários do professor Jairo Nicolau e das sugestões recebidas pelos pareceristas anônimos da *Revista de Sociologia e Política.* Agradecemos ainda ao *Berkeley Initiative for Transparency in the Social Sciences* e ao *Teaching Integrity in Empirical Research.*

O modelo linear de mínimos quadrados ordinários (MQO) é uma das ferramentas mais utilizadas na Ciência Política (Kruger & Lewis-Beck, 2008). Desde que os seus pressupostos sejam respeitados, os coeficientes estimados a partir de uma amostra aleatória fornecem a Melhor Estimativa Linear Não Viesada (*best linear unbiased estimator)* dos parâmetros populacionais (Kennedy, 2005). Não viesada porque nem sobreestima nem subestima sistematicamente o valor do parâmetro, e melhor porque apresenta a menor variância dentre todas as possíveis estimativas (Lewis-Beck, 1980).

E quando os pressupostos forem violados? Nesse caso, devemos adotar técnicas mais adequadas à natureza dos dados. Por exemplo, imagine uma pesquisa que investiga o impacto da receita de campanha sobre a chance de um candidato ser eleito ou não. Como a variável dependente é binária, alguns pressupostos do modelo de mínimos quadrados são violados (homocedasticidade, linearidade e normalidade) e as estimativas podem ser inconsistentes. A regressão logística é a melhor ferramenta para lidar com variáveis dependentes dicotômicas, ou seja, quando o *y* apenas pode assumir duas categorias: eleito ou não eleito; adotou a política ou não adotou; votou no presidente Bolsonaro ou não. Lottes, DeMaris e Adler (1996) argumentam que, apesar da popularidade da regressão logística nas Ciências Sociais, ainda existe grande confusão a respeito de sua correta utilização. Pela nossa experiência pedagógica, essa dificuldade se explica pela escassez de material didático intuitivo. Além disso, muitos cursos de graduação e pós-graduação, assim como livros didáticos, encerram o conteúdo em regressão linear, o que reduz a disseminação de outras técnicas de análise de dados.

Para preencher essa lacuna, este artigo apresenta uma introdução à regressão logística. Nosso objetivo é facilitar a compreensão da aplicação prática dessa técnica. Em termos de audiência, escrevemos para estudantes em estágios iniciais de treinamento e professores que necessitam de materiais didáticos para

ministrar disciplinas de métodos quantitativos. Metodologicamente, reproduzimos os dados de Castro e Nunes (2014) sobre a relação entre envolvimento em escândalos de corrupção (Mensalão[2] e Sanguessugas[3]) e a chance de reeleição dos candidatos concorrentes ao cargo de deputado federal no Brasil em 2006. Todos os dados e *scripts* estão disponíveis no sítio eletrônico do *Open Science Framework* (OSF)[4].

Ao final, o leitor deve ser capaz de identificar quando a regressão logística deve ser utilizada, implementar computacionalmente o modelo e interpretar os resultados. Estamos cientes de que este trabalho não substitui a leitura detalhada das fontes primárias sobre o assunto e de materiais mais técnicos. Apesar disso, esperamos facilitar a compreensão da regressão logística e disseminar a replicabilidade como ferramenta de ensino de análise de dados.

O restante do trabalho está dividido da seguinte forma: a próxima seção explica os fundamentos da regressão logística. A terceira parte identifica os principais requisitos técnicos que devem ser satisfeitos para garantir que as estimativas do modelo sejam consistentes. A quarta seção descreve as principais estatísticas que devem ser observadas. Por fim, apresentamos algumas recomendações sobre como melhorar a qualidade do treinamento metodológico oferecido aos alunos da graduação e pós-graduação em Ciência Política no Brasil.

[2]Para uma breve retrospectiva do mensalão, ver O julgamento do Mensalão (2012).
[3]Para uma apresentação do escândalo dos sanguessugas, ver Entenda o Escândalo dos sanguessugas (2006).
[4]Veja o curso sobre regressão logística oferecido pelo *Coursera* (https://www.coursera.org/course/logisticregression). Sugerimos também o curso de análise de dados categóricos ofertado pelo Programa de Treinamento Intensivo em Metodologia Quantitativa da Universidade Federal de Minas Gerais (MQ – UFMG).

II. A lógica da regressão logística[5]

A utilização de variáveis dependentes categóricas binárias é comum na pesquisa empírica em Ciência Política. Por exemplo: votou ou não (Nicolau, 2007; Soares, 2000), venceu ou perdeu a disputa eleitoral (Speck & Mancuso, 2013; Peixoto, 2009), aderiu à política pública ou não (Furlong, 1998), democracia ou não democracia (Goldsmith, Chalup & Quinlan, 2008), iniciou guerra ou não (Henderson & Singer, 2000), recorreu ou não de uma decisão judicial (Epstein, Landes & Posner, 2013). Para todas essas situações a regressão logística é a técnica mais adequada para modelar a variação da variável dependente em função de um conjunto de variáveis independentes.

Na regressão logística a variável dependente tem apenas duas categorias[6]. Em geral, a ocorrência do evento de interesse é codificada como "1" e a ausência como "0". Lembrando que a codificação altera o sinal dos coeficientes e, portanto, sua interpretação substantiva. Para melhor entender o funcionamento da regressão logística é necessário compreender a lógica da análise de regressão de forma geral. Vejamos a notação clássica do modelo linear:

$$Y = \alpha + \beta X + \varepsilon \quad (1)$$

Y representa a variável dependente, ou seja, aquilo que queremos entender/explicar/predizer. X representa a variável independente. O intercepto, (α), representa o valor de Y quando X assume valor zero. O coeficiente de regressão, (β), representa a variação observada em Y associada ao aumento de uma unidade em X. O termo estocástico, (ε), representa o erro do modelo. Tecnicamente, é possível estimar se existe relação linear entre uma variável dependente (Y) e diferentes variáveis independentes. Além disso, o modelo permite observar a magnitude do efeito e testar a significância estatística dos coeficientes (p-valor e intervalos de confiança). A regressão logística pode ser interpretada como um caso particular de modelos lineares generalizados (MLG)[7] em que a variável dependente é dicotômica. A Figura 1 compara os modelos linear e logístico.

Como a variável dependente no modelo logístico assume apenas dois valores (0 ou 1), a probabilidade predita pelo modelo também deve se limitar ao

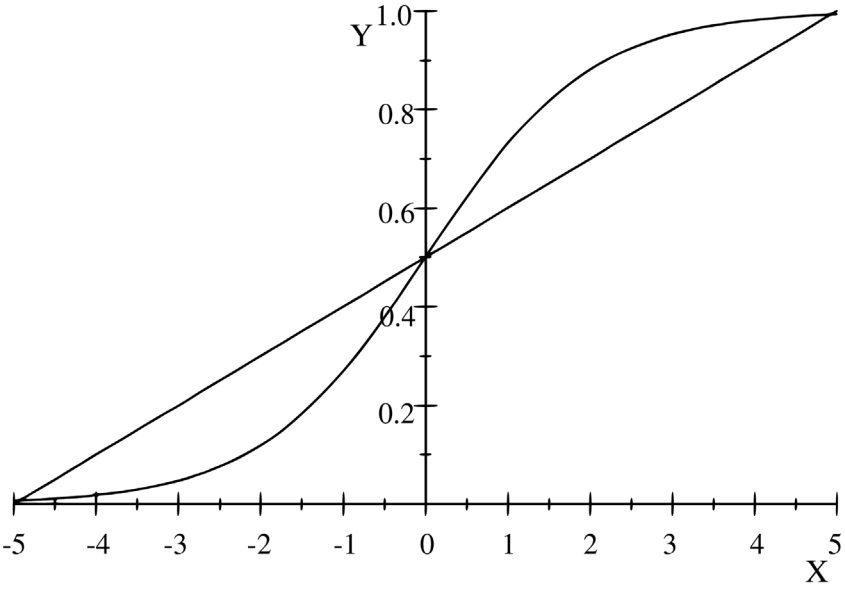
[5]Não discutiremos os fundamentos matemáticos da regressão logística. Para leitores interessados no assunto sugerimos ver Long (1997) e Pampel (2000).

[6]Existem extensões do modelo logístico que permitem modelar a variação de variáveis ordinais (regressão logística ordinal) e policotômicas (regressão logística multinomial).

[7]Nelder e Wedderburn (1972) demonstraram que é possível utilizar o mesmo algoritmo para estimar modelos da família da distribuição exponencial, tais como

Logístico, Probit, Poisson, Gama e Normal Inversa. Não se preocupe com as fórmulas desses modelos. O importante é compreender para que serve cada um deles, quando devem ser utilizados e como os coeficientes devem ser interpretados.

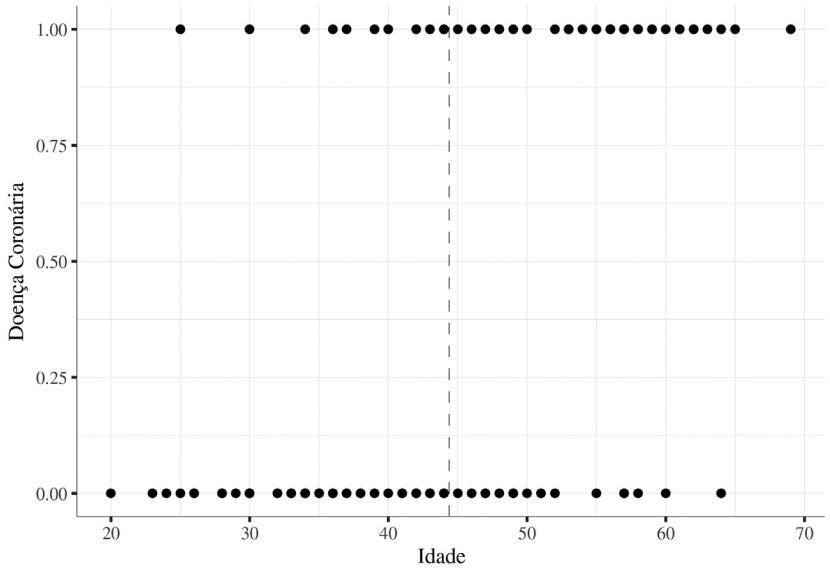
Figura 1 - Reta de regressão linear versus curva logística



Fonte: elaboração própria, com base em Hair, A. *et al.* (2019).

referido intervalo. Quando X (variável independente) assume valores mais baixos, a probabilidade se aproxima de zero. No outro oposto, na medida em que X aumenta, a probabilidade se aproxima de um. Para Kleibaum e Klein (2010), o fato de que a função logística varia entre 0 e 1 explica a popularidade desse modelo. Isso porque como a natureza binária da variável dependente viola alguns pressupostos do modelo linear (homocedasticidade[8], linearidade[9], normalidade), a utilização do modelo linear para analisar variáveis binárias pode gerar coeficientes ineficientes e viesados[10]. Para melhor compreender a relação entre os modelos linear e logístico reproduzimos os dados de Hosmer, Lemeshow e Sturdivant (2013) sobre a associação entre idade e doenças coronárias (Gráfico 1)[11].

Gráfico 1 - Idade x doença coronária



Fonte: elaboração própria, com base em Hosmer, Lemeshow e Sturdivant (2013).

[8]Hair *et al.* (2009) afirmam que homocedasticidade refere-se ao pressuposto de que a variável dependente exibe níveis iguais de variância em toda a gama de variável preditora (Hair *et al.*, 2009, p. 83).

[9]Para Hair *et al.* (2009), um pressuposto implícito de todas as técnicas de análise multivariada com base em medidas correlacionais de associação, incluindo regressão linear múltipla e regressão logística, é a linearidade (Hair *et al.*, 2009, p. 85).

[10]Um estimador é *Best Linear Unbiased Estimator* quando as seguintes propriedades são satisfeitas. Melhor significa eficiente, que produz a menor variância, linear refere-se ao tipo de relação esperada entre os parâmetros e não-viesamento diz respeito à distribuição amostral do estimador. Um estimador viesado é aquele que sistematicamente superestima ou subestima o valor do parâmetro populacional.

[11]Os dados estão disponíveis em: <http://www.ats.ucla.edu/stat/

stata/examples/alr2/alr2stata1.htm>.

A linha pontilhada vertical representa a média da idade: 44,38 anos. Os casos foram codificados como 1 (desenvolveu doença coronária) e 0 (não desenvolveu). A tendência é bastante clara: a medida em que a idade aumenta, cresce a quantidade de pessoas diagnosticadas com doenças coronárias. Uma forma intuitiva de observar esse padrão é examinar o quantitativo de casos tomando a média como parâmetro de comparação. Por exemplo, para as pessoas acima da média, existem mais casos de doentes, enquanto para as pessoas abaixo da média, a concentração maior é na categoria "não desenvolveu". Ou seja, o gráfico está informando que existe uma associação entre idade e doença coronária. É nesse sentido que a regressão logística informa a probabilidade da ocorrência do evento que foi codificado como 1, no caso, desenvolveu doença coronária. A Tabela 1 apresenta esses dados por grupo de idade.

Basta observar a última coluna para chegar à mesma conclusão apresentada pelo Gráfico 1: quanto maior a idade, maior é a chance de desenvolver doenças coronárias. Uma opção adicional para visualizar a relação entre essas variáveis é representar graficamente o percentual de doentes para cada grupo de idade (Gráfico 2).

Observamos uma correlação positiva entre idade (eixo X) e a probabilidade de desenvolver doenças cardíacas (eixo Y). A regressão logística vai informar a direção, a magnitude e o nível da significância estatística dessa relação. Em síntese, o pesquisador deve utilizar a regressão logística quando a variável dependente for categórica binária. Dado que muitas variáveis em Ciências Humanas são categóricas, os benefícios analíticos associados à correta aplicação e interpretação do modelo logístico são evidentes[12].

III. Planejando uma regressão logística

[12]A regressão logística também acomoda variáveis com mais de duas categorias. Quando não existe hierarquia entre as categorias, como por exemplo na distribuição do estado civil, devemos utilizar a regressão multinomial. Por sua vez, a regressão logística ordinal é ideal para modelar a distribuição de variáveis ordinais, ou seja, quando existe uma estrutura de intensidade entre as categorias.

A Tabela 2 descreve os cinco estágios que devem ser observados.

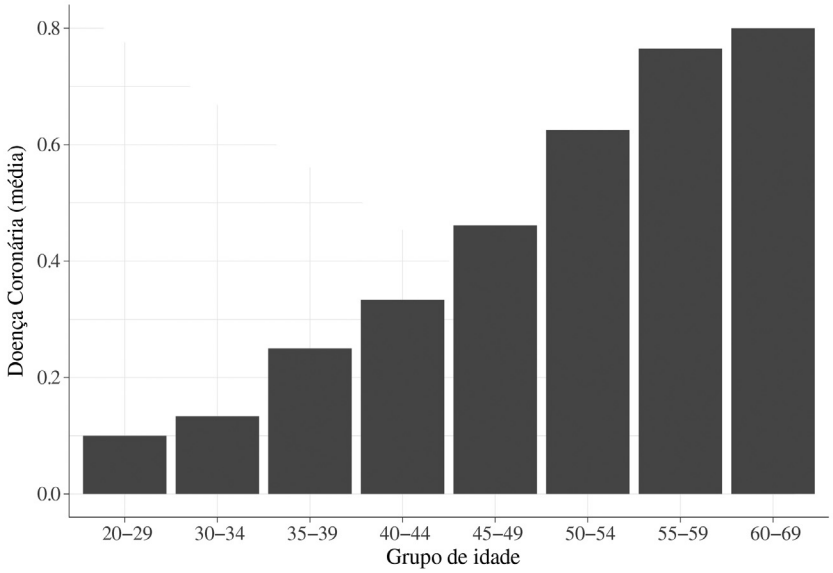
O primeiro passo é identificar uma questão de pesquisa em que a variável dependente seja originalmente dicotômica. Por exemplo, dada a popularidade da regressão logística na área de saúde, uma variável tipicamente utilizada é: viveu/morreu; doente/não doente; fumante/não fumante. Em geral, o pesquisador deve ser dissuadido de recodificar uma variável contínua ou discreta em uma variável categórica dicotômica. Para explicar, suponha que a variável de interesse é renda per capita. É errado recodificar a renda com o objetivo de produzir duas categorias: rico versus pobre. Tecnicamente, a recodificação de

Tabela 1 - Grupo de idade x doença coronária

		Doença		
Grupo Idade	**N**	**Sim**	**Não**	**Sim (%)**
20-29	10	1	9	0,1
30-34	15	2	13	0,13
35-39	12	3	9	0,25
40-44	15	5	10	0,33
45-49	13	6	7	0,46
50-54	8	5	3	0,63
55-59	17	13	4	0,76
60-69	10	8	2	0,8
Total	**100**	**43**	**57**	

Fonte: elaboração própria, com base em Hosmer, Lemeshow e Sturdivant (2013).

Gráfico 2 - Grupo de idade x doença coronária



Fonte: elaboração própria, com base em Hosmer, Lemeshow e Sturdivant (2013).

Tabela 2 - Planejamento de uma regressão logística em cinco estágios

Estágio	Descrição
1º	Identificar a variável dependente
2º	Observar os requisitos técnicos
3º	Estimar e ajustar o modelo
4º	Interpretar os resultados
5º	Validar os resultados

Fonte: elaboração própria a partir de Hair *et al.* (2009).

uma variável quantitativa em categórica implica em perda de informação e isso reduz a consistência das estimativas (Fernandes *et al.*, 2019)[13].

No segundo estágio, deve-se observar os requisitos técnicos. Apesar de ser mais flexível do que outras técnicas estatísticas, a regressão logística é sensível, por exemplo, a problemas de multicolinearidade (altos níveis de correlação entre as variáveis independentes)[14]. Existem diferentes procedimentos para minimizar esse problema. O mais simples é aumentar o número de observações (Kennedy, 2005). Uma saída adicional é utilizar alguma técnica de redução de dados para criar uma medida síntese a partir da variância das variáveis originais. Não devemos simplesmente excluir uma das variáveis independentes, sob pena de produzir erros de especificação no modelo. Na regressão logística o tamanho da amostra é fundamental (Hair *et al.*, 2009). Amostras pequenas tendem a produzir estimativas inconsistentes. Por outro lado, amostras excessivamente grandes aumentam o poder dos testes estatísticos de tal sorte que qualquer efeito tende a ser estatisticamente significativo, independentemente da magnitude. Hosmer e Lemeshow (2000) sugerem um *n* mínimo de 400 casos. Hair *et al.* (2009) sugerem uma razão de 10 casos para cada variável independente incluída no modelo. Pedhazur (1982) recomenda uma razão de 30 casos para cada parâmetro estimado.

[13]A categorização de variáveis tende a produzir estimativas viesadas e ineficientes (Taylor & Yu, 2002). Por esse motivo, enfatizamos o termo "originalmente dicotômicas" e recomendamos nunca reduzir o nível de mensuração de variáveis contínuas, discretas ou ordinais com o objetivo de aplicar modelos de regressão logística. Ainda na dúvida? Veja Fernandes *et al.* (2019).

[14]Quando a correlação é muito alta (alguns usam a regra de ouro de $r \geq 0{,}90$), o erro padrão dos coeficientes é grande, dificultando avaliar a importância relativa das variáveis explicativas. Para entender melhor os problemas que altos níveis de correlação entre as variáveis independentes podem gerar,

ver Figueiredo, Silva e Domingos (2015).

Outra eventual fonte de problema são os *outliers*. Os casos extremos produzem resultados desastrosos em análise de dados e no caso da regressão logística a presença de observações atípicas pode prejudicar o ajuste do modelo. Uma vez detectados os casos aberrantes o pesquisador deve decidir o que fazer com eles. Algumas vezes um caso extremo não passa de um erro de digitação, o que pode ser facilmente resolvido. Uma opção é excluir os *outliers* da estimação do modelo e mensurar o impacto de sua inclusão sobre os coeficientes. Outro procedimento comumente adotado é recodificar o caso, imputando-lhe um valor menos extremo, a média por exemplo. De toda forma, é importante descrever detalhadamente o que foi feito para lidar com eventuais observações extremas[15].

[15]Para uma introdução sobre como detectar *outliers*, ver Figueiredo Filho e Silva (2016), disponível em: <https://cienciapolitica.org.br/system/files/documentos/eventos/2017/04/outlier-que-pertuba-seu-sono-como-identificar-e-manejar.pdf>.

No terceiro estágio, o pesquisador deve estimar o modelo. Aqui dois procedimentos são essenciais: a) reportar o *software* e b) compartilhar os materiais de replicação, que incluem os dados originais, os dados tratados e os *scripts* computacionais[16]. Esses procedimentos aumentam a transparência e facilitam replicablidade dos resultados (King, 1995; Paranhos *et al.*, 2013; Janz, 2016; Figueiredo Filho *et al.*, 2019). Depois de estimar o modelo, o próximo passo é avaliar a qualidade do ajuste. Isso pode ser feito a partir da comparação do modelo nulo (apenas intercepto) com o modelo que incorpora as variáveis independentes. Uma diferença estatisticamente significativa entre os modelos indica que as variáveis explicativas ajudam a prever a ocorrência da variável dependente. A Figura 2 mostra a lógica subjacente à comparação de modelos quando lidamos com a regressão logística.

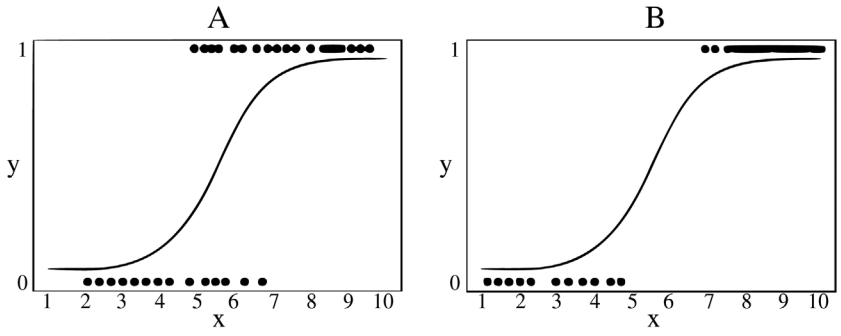
[16]O pesquisador pode disponibilizar os dados no *Dataverse* da Universidade de Harvard. O *Open Science Framework* também pode ser utilizado para disponibilização de dados em projetos mais amplos. No Brasil, sugerimos o Consórcio de Informações Sociais (CIS).

Comparativamente, o modelo B é mais bem ajustado do que o modelo A. Isso pode ser observado pela diferença na capacidade discriminatória. Enquanto o modelo A apresenta alta variabilidade, o modelo B é mais preciso. Para Tabachnick, Fidell e Ullman,

> [...] "a regressão logística, assim como análise de frequência múltipla, pode ser utilizada para ajustar e comparar modelos. O modelo mais simples (e pior ajustado) inclui apenas a constante e nenhum dos preditores. O modelo mais complexo (e melhor ajustado) inclui a constante, todos os preditores e, talvez, interações entre os preditores. Muitas vezes, no entanto, nem todos os preditores (e interações) estão relacionados com o resultado. O investigador usa testes de qualidade do ajuste para escolher o modelo que faz o melhor trabalho de previsão com o menor número de preditores" (Tabachnick, Fidell & Ullman, 2007, p.439).

O quarto estágio consiste na interpretação dos resultados. Infelizmente, muitos trabalhos se limitam a analisar a significância estatística das estimativas e não conferem atenção à magnitude dos coeficientes. Sugerimos que os pesquisadores interpretem os coeficientes e discutam substantivamente

Figura 2 - Comparando o ajuste dos modelos logísticos



Fonte: Hair *et al.* (2009).

como os resultados se relacionam com a hipótese de pesquisa. Diferente da regressão linear, em que os coeficientes são fáceis de interpretar, as estimativas produzidas no modelo logístico são menos intuitivas[17]. Isso porque a transformação *logit* informa o efeito da variável independente sobre a variação do logaritmo natural da chance da variável dependente. Por exemplo, ao se considerar um coeficiente de 0,6, espera-se um acréscimo de 0,6 unidades no *logit* de Y sempre que X aumenta uma unidade. A principal desvantagem dessa abordagem é a falta de inteligibilidade. Afirmar que a quantidade de *logit* aumentou em 0,6 unidades é pouco intuitivo e não ajuda a entender a relação entre as variáveis.

Uma segunda possibilidade é analisar o impacto das variáveis independentes sobre a chance (*odds)* de Y. Para isso o pesquisador deve obter o exponencial do próprio coeficiente. Em nosso exemplo, o exponencial de 0,6 = 1,82. Isso significa que a cada unidade adicional em X, espera-se um aumento de 1,82 na chance de ocorrência de Y, mantendo as demais variáveis constantes[18]. O Gráfico 3 ilustra a distribuição de uma função exponencial de uma simulação em que x varia entre -5 e 5.

Na regressão logística, o exponencial de um valor positivo (+) produz um coeficiente maior do que 1. Contrariamente, um coeficiente negativo (-) retorna um Exp (β) menor do que 1. Um coeficiente de valor zero produz um Exp (β) igual a 1, indicando que a variável independente não afeta a chance de ocorrência da variável dependente. Então, anote aí no seu caderno: quanto mais distante o coeficiente estiver de um, independente da direção, maior é o impacto de uma determinada variável independente sobre a chance da ocorrência do evento de interesse[19].

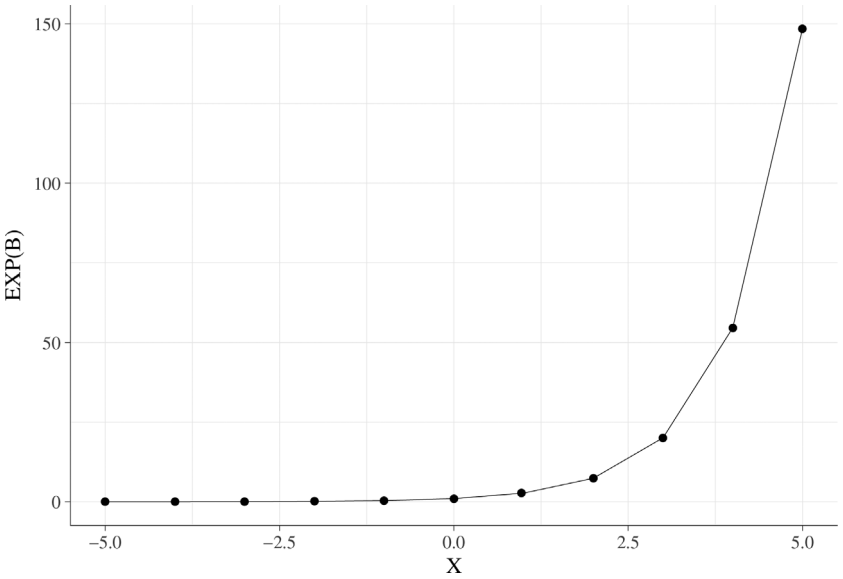
A terceira possibilidade é estimar o aumento percentual na chance de ocorrência de Y. Para tanto, deve-se subtrair uma unidade do coeficiente de regressão exponencializado e multiplicar o resultado por 100, no caso, (1,82-1 * 100). Temos então que o aumento de uma unidade em X está associado a um incremento de 82% na chance de ocorrência de Y (*ceteris paribus)*. A interpretação dos coeficientes da regressão logística pode ficar um pouco mais complicada quando a chance é menor do que 1, ou seja, quando o coeficiente (β) é negativo. Uma solução é inverter o coeficiente (1/valor do coeficiente) o que

[17]No modelo linear, o coeficiente de regressão é interpretado como a variação observada na variável dependente (Y) quando a variável independente (X) aumenta em uma unidade. Na regressão logística, o coeficiente indica a variação no logaritmo da chance da variável dependente ao se elevar a variável explicativa em uma unidade.

[18]Leitores pouco familiarizados com o conceito de chance devem consultar o Apêndice metodológico deste artigo antes de continuar a leitura. Para um tratamento mais detalhado, ver Hilbe (2009).

[19]Na hora de interpretar a significância estatística do intervalo de confiança do coeficiente de regressão da razão de chance devemos observar se o intervalo inclui o valor um (1). Em caso afirmativo, estamos diante de um resultado não significativo. Por exemplo, em um intervalo de confiança em que o coeficiente varia entre 0,8 e 1,6, não é possível rejeitar a hipótese nula.

Gráfico 3 - Função exponencial



Fonte: elaboração própria, com base em Hosmer, Lemeshow e Sturdivant (2013).

facilita a interpretação. Por exemplo, um coeficiente de 0,639 quando invertido indica que quando a variável independente diminui em uma unidade, espera-se um aumento médio de 1,56 na chance de ocorrência da variável dependente.

Por fim, o pesquisador deve validar os resultados observados com uma sub-amostra de sua base de dados original. Esse procedimento confere maior confiabilidade aos resultados de pesquisa, principalmente quando se trabalha com amostras pequenas. De acordo com Hair *et al.* (2009),

> "a abordagem mais comum para estabelecer a validade externa é que a avaliação da taxa geral de acerto seja utilizando uma amostra separada (amostra de *hold-out*) seja através de simulações (*bootstrapping*). Validade externa é ratificada quando a taxa de acertos da abordagem selecionada excede os padrões de comparação que representam a precisão preditiva esperada ao acaso" (Hair *et al.*, 2009, p. 330).

Infelizmente, esse procedimento raramente é utilizado pelos cientistas políticos. Desconfiamos que a reduzida utilização da validação se explica, parcialmente, pela falta de treinamento sobre as especificidades da regressão logística. A próxima seção apresenta um exemplo aplicado da regressão logística e explica como os resultados devem ser interpretados.

IV. Um exemplo aplicado

Para ilustrar a aplicação do modelo de regressão logística replicamos os dados de Castro e Nunes (2014) sobre corrupção e reeleição[20]. Todavia, como o nosso foco é puramente metodológico não iremos explorar o significado substantivo das conclusões reportadas pelos autores. Seguindo o planejamento da seção anterior, o primeiro passo é identificar a variável dependente, qual seja: assume valor "1" para os candidatos que foram reeleitos em 2006 e valor "0" caso contrário[21].

O segundo passo é verificar os requisitos técnicos para a estimação da regressão logística. Nessa etapa é importante observar a presença de eventuais *outliers*, existência de alta correlação entre as variáveis independentes e adequada quantidade de observações. Por limitação de espaço, reproduziremos apenas um dos modelos apresentados por Castro e Nunes (2014). Em particular, a amostra utilizada para estimar o modelo 5 da Tabela 6 possui um total de 217 observações e uma proporção de 19 casos para cada variável independente. Não encontramos casos destoantes e o nível de correlação entre as variáveis incluídas no modelo é aceitável. Dessa forma, podemos seguir para a próxima fase.

O terceiro estágio consiste na estimação do modelo[22]:

$$\text{logit}(Y) = \alpha + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5 + X_6\beta_6 + X_7\beta_7 + X_8\beta_8 + X_9\beta_9 + X_{10}\beta_{10} + X_{11}\beta_{11} + \varepsilon \quad (2)$$

O Quadro 1 sumariza como as variáveis foram mensuradas.

Testaremos três hipóteses:

H_1: estar envolvido em escândalo de corrupção reduz a probabilidade de reeleição;

H_2: quanto maior a despesa de campanha, maior a probabilidade de reeleição;

H_3: quanto maior a execução de emendas, maior a probabilidade de reeleição.

[20] Seguindo as melhores práticas científicas, os autores disponibilizaram os dados e *scripts* no seguinte endereço eletrônico: <http://thedata.harvard.edu/dvn/dv/felipenunes>.

[21] A principal vantagem de utilizar a codificação 0/1 é que a média da distribuição será igual à proporção de casos 1 na amostra. Em uma distribuição com 100 ocorrências, em que 25 casos foram codificados como 1, a média será 0,25, o que representa exatamente a proporção de eventos codificados como 1.

[22] Castro e Nunes (2014) estimaram o modelo de regressão a partir da função de ligação probit. A função logit é mais adequada para trabalhar com amostras pequenas (n < 20) uma vez que apresenta maior taxa de convergência. Em amostras grandes, por outro lado, não existem diferenças significativas entre essas funções de ligação. Para mais informações sobre o assunto, ver Freitas (2013).

Quadro 1 - Nível de mensuração das variáveis

Variáveis	Descrição
Sexo (Controle)	*Dummy:* Feminino (0); Masculino (1)
Idade (Controle)	Contínua: idade na eleição.
Escolaridade (Controle)	Categórica ordinal: Lê e escreve (0); Ensino Fundamental incompleto (1); Ensino Fundamental completo (2); Ensino Médio incompleto (3); Ensino Médio completo (4); Superior incompleto (5); Ensino Superior (6).
Pobreza (Controle)	Contínua: percentual de pessoas pobres no estado.
Ideologia (Controle)	Categórica: Esquerda (0); Centro (1); Direita (2).
Aumento Voto 2006 (Controle)	*Dummy*: Aumentou (1); Diminuiu (0).
Migrante (Controle)	*Dummy*: Migrou de partido (1); Não migrou (0).
Pork (Controle)	Contínua: taxa de sucesso de execução das emendas parlamentares.
Cadeiras por Estado (Controle)	Contínua: número de cadeiras de cada estado na Câmara dos Deputados.
Despesa (Controle)	Contínua: despesa de campanha
Escândalo (VI)	*Dummy:* Envolvido em escândalo (1); Não envolvido (0).
Reeleição (VD)	*Dummy:* Reeleito (1); Não Reeleito (0).

Fonte: elaboração própria a partir de Castro e Nunes (2014, p. 38-40).

V. Resultados

O primeiro passo é analisar a distribuição da variável dependente. A Tabela 3 sumariza essas informações.

Existem informações para 451 casos. Desse total, 60,53% dos deputados federais foram reconduzidos em 2006, o que significa 273 ocorrências[23]. Dizemos então que a probabilidade de reeleição é de 0,605. Por sua vez, a chance de ser reeleito pode ser calculada pela divisão entre as probabilidades (sim/não), no caso, 0,605/0,395 = 1,53. A Tabela 4 ilustra essas informações.

Ao considerar apenas os candidatos envolvidos em escândalos de corrupção, o percentual de reeleição foi de 17,86%, já que 10 dos 56 parlamentares

[23]O pesquisador deve se certificar de que nenhuma categoria tenha uma distribuição inferior a 5%. Isso porque enquadra o fenômeno como evento raro, sendo necessário aplicar correções específicas para lidar com essa situação. Para os leitores interessados ver King e Zeng (2001).

Tabela 3 - Distribuição de frequência da variável dependente (reeleito)

Reeleito	N	%
Sim	273	60,53
Não	178	39,47
Total	**451**	**100,0**

Fonte: elaboração própria.

Tabela 4 - Taxa comparativa de reeleição (envolvidos x não envolvidos) (%)

Envolvido em escândalo	Reeleito		Total
	Sim	Não	
Sim	10 (17,86)	46 (82,14)	**56 (100,0)**
Não	263 (66,58)	132 (33,42)	**395 (100,0)**
Total	**273 (60,53)**	**178 (39,47)**	**451 (100,0)**

Fonte: elaboração própria.

conseguiram um novo mandato[24]. Isso quer dizer que, para esse grupo, a probabilidade de reeleição é 0,179 e a chance de reeleição é de 0,22. Para os candidatos não envolvidos em escândalos de corrupção, a chance de ser reeleito é de 1,9. Fundamentalmente, em nosso exemplo de replicação, a regressão logística consiste na análise comparativa do percentual de reeleição entre candidatos envolvidos em escândalos de corrupção e os não envolvidos[25].

Em termos de ajuste geral do modelo, um dos principais testes utilizados é o de Hosmer e Lemeshow (2000). Esse teste é considerado mais robusto do que o teste de chi-quadrado comum, principalmente quando existem variáveis independentes contínuas ou quando o tamanho da amostra é pequeno (Garson, 2011). A Tabela 5 sumariza as informações de interesse (valor do teste, os graus de liberdade e a significância estatística) para o Teste de Hosmer e Lemeshow e a Tabela 6 apresenta as mesmas informações para o teste Omnibus dos coeficientes do modelo.

Um resultado não significativo ($p > 0,05$) sugere que o modelo estimado com as variáveis independentes é melhor do que o modelo nulo. O modelo estimado apresentou um chi-quadrado (χ^2) de 6,832 e p-valor de 0,555, sugerindo um ajuste adequado. Outra medida de ajuste comumente utilizada é o omnibus teste dos coeficientes (*Omnibus test of model coefficients*). É um teste de chi-quadrado comparando a variância do seu modelo com variáveis independentes e o modelo nulo (apenas o intercepto).

Diferente do teste de Hosmer e Lemeshow, um resultado significativo ($p < 0,05$) sugere um ajuste adequado. De acordo com os dados, o modelo apresentou um chi-quadrado de 56,356 (p-valor < 0,001), ou seja, o modelo ajustado é melhor do que o modelo nulo. Assim, devemos inferir que as variáveis independentes influenciam a variação da variável dependente[26]. Não encontramos esses testes nem no artigo de Castro e Nunes (2014), nem nos scripts computacionais. A Tabela 7 sumariza os coeficientes estimados do modelo de regressão logística na tentativa de reproduzir os resultados reportados na Tabela 6 de Castro e Nunes (2014).

Assim como na regressão linear, o primeiro passo é examinar os coeficientes estimados (β). Aqui o pesquisador deve observar o sinal das estimativas e comparar com a direção esperada em suas hipóteses de trabalho. X_{11} (Escândalo) tem um efeito negativo (-1,677) sobre a probabilidade de reeleição. Diferente do modelo linear, o coeficiente da regressão logística não tem uma interpretação direta.

Existem duas principais formas de interpretar os coeficientes: a) analisar a razão de chance e b) transformar razão de chance em percentual. Pelo primeiro critério, devemos concluir então que o envolvimento em escândalos de corrup-

[24]Esses achados divergem residualmente das informações reportadas pelas Tabelas 4 e 5 de Castro e Nunes (2014) que indica 9 reeleitos de um total de 50 parlamentares, o que equivale a 18%.

[25]E isso pode ser calculado a partir da razão de chance, que é calculada pela divisão entre as chances de reeleição de cada grupo, no caso, 1,9/0,22. Ou seja, candidatos não envolvidos em escândalos de corrupção tem cerca de 8 vezes mais chance de serem reeleitos quando comparados com os deputados citados nos esquemas do mensalão e/ou sanguessugas, assim como mensurado por Castro e Nunes (2014).

[26]Para Garson (2011), o teste omnibus pode ser interpretado como um teste para a capacidade conjunta de todos os preditores do modelo preverem a variável resposta (dependente). Um resultado significativo indica que o ajuste está adequado aos dados, sugerindo que pelo menos um dos preditores é significativamente relacionado com a variável resposta.

Tabela 5 - Teste de Hosmer e Lemeshow

χ^2	**gl**	**Sig**
6,832	8	0,555

Fonte: elaboração própria.

Tabela 6 - Teste Omnibus dos coeficientes do modelo

χ^2	**gl**	**Sig**
56,356	11	0,000

Fonte: elaboração própria.

Tabela 7 - Coeficientes do modelo de regressão logística*

	β	Erro Padrão	Z(Wald)	Sig.	Exp(β)	(exp(β)-1) x 100
(Intercepto)	0,552	1,568	0,352	0,725	1,737	73,734
Pobreza	1,171	1,419	0,825	0,409	3,224	222,386
Masculino	-0,005	0,560	-0,009	0,993	0,995	-0,484
Idade	-0,014	0,017	-0,830	0,406	0,986	-1,409
Escolaridade	-0,060	0,161	-0,370	0,712	0,942	-5,789
Ideologia	-0,125	0,224	-0,561	0,575	0,882	-11,782
Aumentou Votos	0,908	0,341	2,663	0,008	2,480	148,030
Migrante	0,078	0,382	0,205	0,838	1,081	8,136
Emendas Parlamentares	**-0,272**	**0,639**	**-0,425**	**0,671**	**0,762**	**-23,785**
Candidato/vagas	-0,005	0,009	-0,516	0,606	0,995	-0,469
Despesas de Campanha	**0,000**	**0,000**	**3,920**	**0,000**	**1,000**	**0,000**
Escândalo	**-1,677**	**0,528**	**-3,176**	**0,001**	**0,187**	**-81,299**

Fonte: elaboração própria.
Variável dependente: reeleito.
* Como em qualquer modelo de regressão, os coeficientes não padronizados de variáveis com escalas diferentes não podem ser diretamente comparados. O STATA tem o comando (listcoef, std help) que produz coeficientes padronizados na variável independente, dependente e em ambas. Menard (2004) apresenta seis diferentes formas de padronizar os coeficientes em regressão logística.

ção reduz a chance de ser reeleito. Em termos percentuais, estar envolvido em corrupção diminui em 81,2% a probabilidade de ser reeleito, tal como esperado teoricamente pela hipótese 1. Ao se considerar a despesa de campanha, o efeito foi nulo, com um Exp (β) = 1,000.

Assim como Castro e Nunes (2014), não encontramos efeitos significativos da variável emendas parlamentares sobre a chance de reeleição, levando em conta a magnitude do p-valor e o erro padrão duas vezes maior do que a própria estimativa do impacto[27].

Depois de analisar os coeficientes associados às variáveis de interesse, o próximo passo é avaliar a qualidade do ajuste do modelo. A Tabela 8 sumariza algumas medidas de ajuste tipicamente reportadas em modelos estimados por máxima verossimilhança[28].

É comum aparecer nas saídas dos diferentes pacotes estatísticos o número de iterações utilizadas pelo computador para estimar o modelo. Ao informar que o modelo convergiu após a 5 iteração, isso quer dizer que os coeficientes foram estimados via máxima verossimilhança. Em geral, quanto mais rápido o modelo convergir (menos iterações), melhor. Se o modelo não convergir, os coeficientes não são confiáveis. Um dos principais fatores que explicam a não conversão

[27]No original, "a alocação bem-sucedida de recursos particularistas *(pork)* não apresenta, diferentemente do que era esperado, associação positiva com reeleição. O resultado parece ser nulo e não relevante para explicar as chances de reeleição, em 2006, também quando variáveis socioeconômicas e institucionais são incluídas no modelo" (Castro & Nunes, 2014, p. 42).

[28]O método de máxima verossimilhança é um processo iterativo que procura ajustar o modelo através de várias repetições. No entanto, algumas vezes o modelo simplesmente não converge. Isso pode acontecer por vários motivos, desde problemas nos algoritmos utilizados para estimar a função de ligação até a distribuição fortemente assimétrica das variáveis independentes.

Tabela 8 - Medidas de ajuste do modelo*

-2log likelihood nulo	-2log likelihood	Cox e Snell R^2	Nagelkerke R^2	BIC
3.057.559	237.4225	0,229	0,308	301.891

Fonte: elaboração própria.
*A estatística - 2 log likelihood (-2LL) é uma medida de ajuste. Quanto menor, melhor é o ajuste. O pesquisador pode utilizá-la para comparar os ajustes de diferentes modelos (incluindo e retirando variáveis independentes, mas preservando a mesma variável dependente).

do modelo é a insuficiência de casos em relação ao número de variáveis independentes incluídas no modelo.

De acordo com Menard (2002), o *log likelihood* é uma medida de seleção de parâmetros no modelo de regressão logística. No entanto, a maior parte dos pacotes estatísticos reporta o *-2 log likelihood* (-2LL) e sua interpretação é a seguinte: quanto maior, pior é a capacidade explicativa/preditiva do modelo. Intuitivamente, ele pode ser interpretado como uma medida do erro ao tentar utilizar um determinado conjunto de variáveis independentes (modelo) para explicar a variação da variável dependente. O pesquisador pode solicitar a *iteration history* da estimação. O procedimento vai produzir o *-2 log likelihood* do modelo nulo e do modelo ajustado. A diferença entre elas é medida em termos de chi-quadrado. Como ele é uma medida de erro, quanto maior for o chi-quadrado, maior é a redução do erro do modelo ajustado (com as variáveis independentes) em relação ao modelo nulo.

A Tabela 8 apresenta o valor do -2LL para facilitar a comparação entre os modelos. No modelo nulo, o -2LL era de 3.057.559, e o modelo com as variáveis independentes foi de 237.4225. Nesse caso, observamos uma redução considerável. Isto significa que o modelo com variáveis independentes tem um ajuste superior ao modelo nulo. Do mesmo modo, o BIC (*Bayesian Information Criterion*) é mais uma medida baseada em máxima verossimilhança. Quanto menor, melhor. O modelo testado apresentou um BIC de 301,891, enquanto o do modelo nulo foi 3.066,105. Podemos extrapolar isso e comparar diversos modelos e não apenas o nulo.

Diferente do modelo linear, a regressão logística não tem uma medida síntese da variação na variável dependente explicada pelo modelo, tal como o coeficiente de determinação[29]. No entanto, algumas medidas foram desenvolvidas no sentido de orientar o pesquisador em relação ao poder explicativo/preditivo do modelo[30]. As mais comumente empregadas são o pseudo R^2 de Cox e Snell e o pseudo R^2 de Nagelkerke[31]. Para Menard (2002),

> "R_i^2 é uma redução proporcional em -2LL ou uma redução proporcional do valor absoluto do log-likelihood, onde a quantidade sendo minimizada para selecionar os parâmetros do modelo é tomada como uma medida da variação" (Menard, 2002, p. 25).

Para os propósitos deste artigo adotamos a seguinte interpretação: quanto mais próximo de zero, menor é a diferença entre o modelo nulo (sem nenhuma variável independente) e o modelo estimado. Quanto mais próximo de um, maior é a diferença entre o modelo nulo e o modelo proposto pelo pesquisador. No limite, um pseudo R^2 de zero indica que as variáveis independentes incluídas não ajudam a explicar a variação da variável dependente. Um pseudo R^2 de 1 sugere que as variáveis explicam/predizem perfeitamente a variação de Y. Lembrando que devemos ser menos exigentes com o modelo logístico do que com o modelo linear em termos de variância explicada pelo R^2.

Por fim, o pesquisador deve analisar a tabela de classificação (*classification table*). Essa saída é particularmente interessante pois fornece uma medida da capacidade preditiva do modelo. A Tabela 9 ilustra as informações de interesse.

A tabela de classificação é chamada, frequentemente, de tabela de confusão. Para Garson (2011),

> "apesar da classificação correta como medida do ajuste do modelo seja preferível às medidas de pseudo-R^2, eles têm algumas limitações severas para esta finalidade. Tabelas de Classificação não devem ser usadas exclusivamente como medidas de ajuste porque eles ignoram probabilidades preditas reais e em vez disso usam previsões dicotômicas baseadas em um ponto de corte (ex.: 0,50). Por exemplo, em regressão logística binária, prevendo uma dependente de 0 ou 1, a

[29] Existe um debate sobre as vantagens e limitações do r^2 como medida síntese para avaliar a qualidade do ajuste dos modelos de regressão linear. Salvo melhor juízo, King (1986) representa o primeiro alerta sistemático sobre o assunto na pesquisa empírica em Ciência Política. Figueiredo Filho, Silva Júnior e Rocha (2012) apresentam uma discussão pedagógica sobre o tema.

[30] Hair *et al.* (2009) afirmam que o ajuste do modelo logístico pode ser avaliado a partir de dois principais procedimentos: (1) os pseudo r2s, similarmente à regressão linear e (2) estimar a capacidade preditiva do modelo.

[31] Existem ainda o McFadden's pseudo R^2, McKelvey e Savoina pseudo R^2, McFadden pseudo R^2 ajustado, Cragg e Uhler pseudo R^2 e Efron pseudo R^2. Para o leitor interessado em aprofundar seus conhecimentos sobre o assunto ver Hagle e Mitchell (1992) e Menard (2000).

Tabela 9 - Tabela de classificação

		Predito		Total
		Não reeleito	Reeleito	
Real	**Não reeleito**	23,50	17,51	41,01
	Reeleito	10,60	48,39	58,99
	Total	34,10	65,90	100,00

Fonte: elaboração própria.

> Tabela Classificação não revela quão perto de 1 foram as previsões corretas, nem quão perto de 0 foram os erros. Um modelo no qual as predições, corretas ou não, estão muito próximas do ponto de corte 0,5, não tem um ajuste tão bom como um modelo onde os clusters das pontuações preditas estejam próximos de 1 ou 0. Além disso, porque a taxa de acerto pode variar consideravelmente para o mesmo modelo de logística em diferentes amostras, o uso da Tabela Classificação para comparar diferentes amostras não é recomendado" (Garson, 2011, p. 173).

Nossa matriz de classificação utiliza o padrão convencional de 50% para alocar os casos como 1 (se a probabilidade predita foi maior que 0,5) ou 0 (menor que 0,5). Podemos avaliar essa tabela utilizando três conceitos: acurácia, sensibilidade e especificidade. A acurácia do modelo é a proporção de casos verdadeiros positivos e verdadeiros negativos. De acordo com a Tabela 9, a acurácia do nosso modelo foi de 71,89% (23,50% + 48,29%). Contudo, nem sempre a acurácia do modelo é o mais importante. Em determinados casos, importa maximizar a taxa de verdadeiros positivos, ou verdadeiros negativos.

Passemos à sensibilidade. Ela diz respeito ao percentual de casos que tem a característica de interesse (foi reeleito) que foram corretamente preditos pelo modelo (verdadeiros positivos / (falsos positivos + verdadeiros positivos). No nosso exemplo, 48,39% dos candidatos reeleitos foram corretamente classificados de um total de 58,99% que realmente foram reeleitos. Isto nos dá uma sensibilidade de 82,03% (48,39%/58,99%). Já a especificidade do modelo diz respeito ao percentual de casos que não tem a característica de interesse (não foram reeleitos) que foram corretamente classificados pelo modelo, isto é (verdadeiros negativos / (falsos negativos + verdadeiros negativos)). Como pode ser observado, 23,50% dos candidatos não-reeleitos foram corretamente identificados em um total de 41,01% de não reeleitos. Isto nos dá uma especificidade de 57,30% (23,50%/41,01%). Existe um *tradeoff* entre sensibilidade e a especificidade. Ao aumentar uma, a outra diminui. Embora às vezes a sensibilidade do modelo seja mais importante (prever que terá uma doença, já que você poderá tratá-la), outras vezes o melhor seria aumentar a especificidade (impedir que políticos corruptos sejam reeleitos).

VI. Conclusão

Esperamos ajudar estudantes e professores a melhor compreenderem o funcionamento da regressão logística. A ausência de cursos de cálculo, álgebra linear e matricial e estatística avançada limita a capacidade de compreender técnicas mais avançadas de análise de dados. Por esse motivo, nossa abordagem se concentrou na exposição intuitiva dos resultados. Acreditamos também que entender a lógica intuitiva da regressão logística é o primeiro passo para melhor compreender os diferentes procedimentos existentes para lidar com dados categóricos. O avanço computacional permite que pesquisadores com menor treinamento específico em Matemática e Estatística possam se beneficiar das vantagens associadas às diferentes técnicas multivariadas. Dado que muitas

variáveis em Ciência Política são categóricas, os benefícios analíticos associados à correta aplicação e interpretação do modelo logístico são evidentes. Com esse artigo, esperamos difundir a utilização da regressão logística.

E como melhorar a qualidade do treinamento metodológico e técnico ofertado aos alunos da graduação e pós-graduação em Ciência Política no Brasil? Recomendamos o seguinte: (1) incorporar a replicação como ferramenta pedagógica em cursos de análise de dados; (2) incluir cursos obrigatórios de matemática, cálculo, probabilidade e estatística no componente curricular dos cursos de graduação e pós-graduação. Além disso, os alunos devem receber treinamento específico em alguma linguagem de programação; (3) realizar exercícios práticos envolvendo análise de dados com problemas tipicamente de Ciência Política. A ênfase em problemas abstratos reduz o interesse dos alunos pelo tema; (4) incentivar a participação dos discentes em cursos de verão como o MQ-UFMG e IPSA-USP; (5) promover cursos de epistemologia e filosofia da ciência. A definição dos métodos e técnicas de pesquisa dependem da visão epistemológica do que é o conhecimento científico e de como ele deve ser implementado; (6) difundir a leitura crítica de artigos que utilizam técnicas avançadas de análise de dados; (7) acompanhar a produção acadêmica de revistas com ênfase metodológica, como, por exemplo, a *Political Analysis* e a *Political Science Research and Methods;* (8) fomentar a publicação de artigos metodológicos em periódicos nacionais; (9) favorecer a criação de grupos de pesquisa e mesas redondas sobre metodologia e técnicas em análise de dados em congressos profissionais; e (10) financiar projetos de pesquisa especialmente voltados para aprofundar o *status* do conhecimento sobre a principal característica da ciência: o método.

Antônio Alves Tôrres Fernandes (antonio.alvestorres@ufpe.br) é mestrando em Ciência Política pela Universidade Federal de Pernambuco.

Dalson Britto Figueiredo Filho (dalson.figueiredofo@ufpe.br) é professor do Programa de Pós-graduação em Ciência Política da Universidade Federal de Pernambuco e autor do livro "Métodos quantitativos em Ciência Política", Editora InterSaberes.

Enivaldo Carvalho da Rocha (enivaldocrocha@gmail.com) é professor Titular aposentado do Programa de Pós-graduação em Ciência Política da Universidade Federal de Pernambuco.

Willber da Silva Nascimento (nascimentowillber@gmail.com) é doutor em Ciência Política pela UFPE e bolsita de pós-doutorado do PPGCP/UFPE/FACEPE.

Referências

Altman, D. (1991) Categorising continuous variables. *British Journal of Cancer*, 64(5), p. 975. DOI: 10.1038/bjc.1991.441

Bonney, G. (1987) Logistic regression for dependent binary observations. *Biometrics*, 43(4), pp. 951-973. DOI: 10.2307/2531548

Brant, R. (1996) Digesting logistic regression results. *The American Statistician,* 50(2), pp. 117-119. DOI: 10.2307/2684422

Castro, M.M.M. & Nunes, F. (2014) Candidatos corruptos são punidos?: accountability na eleição brasileira de 2006. *Opinião Pública*, 20(1), pp. 26-48. DOI: 10.1590/S0104-62762014000100002

Codato, A.; Cervi, E. & Perissinoto, R. (2013) Quem se elege prefeito no Brasil? Condicionantes do sucesso eleitoral em 2012. *Cadernos Adenauer*, 14(2), pp. 61-84.

Cohen, J. (1983) The Cost of Dichotomization. *Applied Psychological Measurement*, 7(3), pp. 249–253. DOI: 10.1177/014662168300700301

Cook, R. & Weisberg, S. (1997) Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association*, 92(438), pp. 490-499. DOI:10.1080/01621459.1997.10474002

R Core Team. (2019) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Disponível em: https://www.R-project.org/. Acesso em: 28 set. 2020.

DeMaris, A. (1995) A tutorial in logistic regression. *Journal of Marriage and the Family*, pp. 956-968. DOI: 10.2307/353415

Eno, D. & Terrell, G. (1999) Scatterplots for logistic regression. *Journal of Computational and Graphical Statistics*, 8(3), pp. 413-425. DOI: 10.1080/10618600.1999.10474822

Epstein, L.; Landes, W. & Posner, R. (2013) *The behavior of federal judges: a theoretical and empirical study of rational choice*. Cambridge: Harvard University Press.

Fernandes, A. *et al.* (2019) Why quantitative variables should not be recoded as categorical. *Journal of Applied Mathematics and Physics,* 7(7), pp. 1519-1530. DOI: 10.4236/jamp.2019.77103

Figueiredo Filho, D.; Silva, L. & Domingos, A. (2015) O Que é e como Superar a Multicolinearidade? Um Guia Para Ciência Política. *Conexão Política*, 4(2), pp. 95-104. DOI: 10.26694/rcp.issn.2317-3254.v4e2.2015.p95-104

Figueiredo Filho, D. & Silva Júnior, J. (2016) O outlier que pertuba o seu sono: Como identificar casos extremos? In: *10º Encontro da Associação Brasileira de Ciência Política*. Belo Horizonte. Disponível em: https://cienciapolitica.org.br/system/files/documentos/eventos/2017/04/outlier-que-pertuba-seu-sono-como-identificar-e-manejar.pdf. Acesso em: 13 out. 2020.

Figueiredo Filho, D. & Silva Júnior, J. (2010) Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). *Revista Política Hoje.* 18(1), pp. 115-146. Disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/view/3852>. Acesso em: 15 de maio 2020.

Figueiredo Filho, D. *et al.* (2011) O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). *Revista Política Hoje*, 20(1), pp. 44-99. Disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/view/3808/31622>. Acesso em: 15 de maio 2020.

Figueiredo Filho, D. *et al.* (2019) Seven Reasons Why: A User's Guide to Transparency and Reproducibility. *Brazilian Political Science Review*, 13(2), pp. e0001. DOI: 10.1590/1981-3821201900020001

Figueiredo Filho, D.; Silva Júnior, J. & Rocha, E. (2012) Classificando regimes políticos utilizando análise de conglomerados. *Opinião Pública*, 18(1), pp. 109-128. DOI: 10.1590/S0104-62762012000100006

Fox, J. (1991) *Regression diagnostics: An introduction* Vol. 79. Thousand Oaks, CA: Sage Publications.

Freitas, L. (2013) *Comparação das funções de ligação logit e probit em regressão binária considerando diferentes tamanhos amostrais*. Tese de Doutorado. Viçosa: Universidade Federal de Viçosa.

Furlong, E. (1998) A logistic regression model explaining recent state casino gaming adoptions. *Policy Studies Journal*, *26*(3), pp. 371-383, DOI: 10.1111/j.1541-0072.1998.tb01907.x

Garson, G.D. (2011) *Multiple regression: Overview*. Statnotes: Topics in Multivariate Analysis. Disponível em: https://faculty.chass.ncsu.edu/garson/PA765/statnote.htm. Acesso em: 13 set. 2020.

Garson, G.D. (2014) *Logistic Regression: Binary and Multinomial*. [s.l.]: Statistical Associates Publishing.

Goldsmith, B.; Chalup, S. & Quilan, M. (2008) Regime type and international conflict: towards a general model. *Journal of Peace Research*, 45(6), pp. 743-763. DOI: 10.1177/0022343308096154

Guthery, F. & Binghan, R. (2007) A primer on interpreting regression models. *The Journal of Wildlife Management*, 71(3), pp. 684-692. DOI: 10.2193/2006-285

Hagle, T. & Mitchell, G. (1992) Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36(3), pp. 762-784. DOI: 10.2307/2111590

Hair, J. *et al.* (2009) *Análise multivariada de dados*. Porto Alegre: Bookman Editora.

Henderson, E. & Singer, J. (2000) Civil war in the post-colonial world, 1946-92. *Journal of Peace Research*, 37(3), pp. 275-299. DOI: 10.1177/0022343300037003001

Hilbe, J. (2009) *Logistic regression models*. London: Chapman and Hall/CRC.

Hosmer Jr, D.; Lemeshow, S. & Sturdvanty, R. (2013) *Applied logistic regression* Vol. 398. New York: John Wiley & Sons.

Hosmer Jr, D. & Lemeshow, S. (2000) *Applied Logistic Regression*. New York: John Wiley & Sons.

Jaccard, J. & Jaccard, J. (2001) *Interaction effects in logistic regression*. Thousand Oaks: Sage Publications.

Janz, N. (2016) Bringing the gold standard into the classroom: replication in university teaching. *International Studies Perspectives*, 17(4), pp. 392-407. DOI: 10.1111/insp.12104

Kay, R. & Little, S. (1987) Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74(3), pp. 495-501. DOI: 10.2307/2336688

Kennedy, P. (2005) *A guide to econometrics*, Oxford: Maldon.

Keprt, A. & Snásel, V. (2004) Binary Factor Analysis with Help of Formal Concepts. In: *The Second International Conference on Concept Lattices and Their Applications (CLA)*. Ostrava, pp. 90-101. Disponível em: http://ceur-ws.org/Vol-110/paper10.pdf. Acesso em: 13 out. 2020.

King, G. (1986) How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 30(3), pp. 666-687. Disponível em: https://ssrn.com/ABSTRACT=084228. Acesso em: 28 set. 2020.

King, G. (1995) Replication, replication. *PS: Political Science & Politics*, 28(3), pp. 444-452. Disponível em: https://gking.harvard.edu/files/gking/files/replication.pdf. Acesso em: 13 out. 2020.

King, G. & Zeng, L. (2001) Logistic regression in rare events data. *Political analysis*, 9(2), pp. 137-163. DOI: 10.1093/oxfordjournals.pan.a004868

Kleinbaum, D. & Klein, M. (2010) *Logistic regression: A Self-Learning Text*. New York: Springer-Verlag. DOI: 10.1007/978-1-4419-1742-3

Krueger, J. & Lewis-Beck, M. (2008) Is ols dead? *The Political Methodologist*, 15(2), pp. 2-4, 2008.

Landwehr, J.; Pregibon, D. & Shoemaker, A. (1984) Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79(385), pp. 61-71. DOI: 10.1080/01621459.1984.10477062

Lewis-Beck, M. (1980) *Applied Regression.* Thousand Oaks: Sage Publications.

Long, J. (1997) Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, 7(s/n), Thousand Oaks: Sage Publications.

Lottes, I.; DeMaris, A. & Adler, M. (1996) Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology*, 24(3), pp. 284-298. DOI: 10.2307/1318743

Menard, S. (2000) Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), pp. 17-24. DOI: 10.1080/00031305.2000.10474502

Menard, S. (2002). *Applied logistic regression analysis*. Thousand Oaks: Sage Publications. DOI: 10.4135/9781412983433

Menard, S. (2004) Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, *58*(3), pp. 218-223. DOI: 10.1198/000313004X946

Nelder, J. & Wedderburn, R. (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A (General),* 135(3), pp. 370-384. DOI: 10.2307/2344614

Nicolau, J. (2000) An analysis of the 2002 presidential elections using logistic regression. *Brazilian Political Science Review*, 1(1), pp. 125-135.

O'Brien, S. M & Dunson, D. (2004) Bayesian multivariate logistic regression. *Biometrics*, 60(3), pp. 739-746. DOI: 10.1111/j.0006-341X.2004.00224.x

O'Connell, A. (2006) *Logistic regression models for ordinal response variables*. Thousand Oaks: Sage Publications. DOI: 10.4135/9781412984812

Pampel, F. (2000) *Logistic regression: A primer.* Thousand Oaks: Sage Publications. DOI: 10.4135/9781412984805

Paranhos, R.; Figueiredo Filho, D.; Rocha, E. & Carmo, E. (2013) A importância da replicabilidade na ciência política: o caso do SIGOBR. *Revista Política Hoje*, *22*(2), pp. 213-229.

Pardoe, I. & Cook, R. (2002) A graphical method for assessing the fit of a logistic regression model. *The American Statistician*, 56(4), pp. 263-272. DOI: 10.1198/000313002560

Pedhazur, E. (1982) *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston.

Peixoto, V. (2009) Financiamento de campanhas: o Brasil em perspectiva comparada. *Perspectivas: revista de ciências sociais*, 35(s/n), pp. 91-116.

Press, S. & Wilson, S. (197) Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), pp. 699-705. DOI: 10.1080/01621459.1978.10480080

Revelle, W. (2018) *Psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, USA. Disponível em: https://cran.r-project.org/web/packages/psych/. Acesso em: 28 set. 2020.

Ribeiro, E.; Carreirão, Y. & Borba, J. (2011) Sentimentos partidários e atitudes políticas entre os brasileiros. *Opinião Pública*, 17(2), pp. 333-368. DOI: 10.1590/S0104-62762011000200003

Roberts, G.; Rao, N. & Kumar, S. (1987) Logistic regression analysis of sample survey data. *Biometrika*, 74(1), pp. 1-12. DOI: 10.2307/2336016

Schwab, J. (2002) *Multinomial logistic regression: Basic relationships and complete problems*. Austin, Texas: University of Texas.

Soares, G. (2000) Em busca da racionalidade perdida: alguns determinantes do voto no Distrito Federal. *Revista Brasileira de Ciências Sociais*, 15(43), pp. 5-23. DOI: 10.1590/S0102-69092000000200001

Speck, B. & Mancuso, W. (2013) O que faz a diferença? Gastos de campanha, capital político, sexo e contexto municipal nas eleições para prefeito em 2012. *Cadernos Adenauer*, 14(2), pp. 109-126.

Stock, J. & Watson, M. (2015) *Introduction to Econometrics*. 3ª Edition. United Kingdom: Pearson.

Tabachnick, B.; Fidell, L. & Ullman, J. (2007) *Using multivariate statistics*. Boston, MA: Pearson.

Taylor, J. & Yu, M. (2002) Bias and Efficiency Loss Due to Categorizing an Explanatory Variable. *Journal of Multivariate Analysis*, 83(s/n), pp. 248–263. DOI: 10.1006/jmva.2001.2045

Wong, G. & Mason, W. (1985) The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80(391), pp. 513-524. DOI: 10.1080/01621459.1985.10478148

Artigos e jornais

O julgamento do Mensalão (2012) *Folha de São Paulo*. São Paulo. 12. Jun. Disponível em: <https://www1.folha.uol.com.br/especial/2012/ojulgamentodomensalao/>. Acessado em: 10 nov. 2020.

Entenda o Escândalo dos sanguessugas (2006) Estado de São Paulo. São Paulo. 11.dez. Disponível em: <https://politica.estadao.com.br/noticias/geral,entenda-o-escandalo-dos-sanguessugas,20061211p60113>. Acessado em: 10 nov. 2020.

Read this paper if you want to learn logistic regression

ABSTRACT: Introduction: What if my response variable is binary categorical? This paper provides an intuitive introduction to logistic regression, the most appropriate statistical technique to deal with dichotomous dependent variables. **Materials and Methods:** we estimate the effect of corruption scandals on the probability of reelection of candidates running for the Brazilian House of Representatives using data from Castro and Nunes (2014). Specifically, we show the computational implementation in R and we explain the substantive interpretation of the results. **Results:** we share replication materials which quickly enables students and professionals to use the procedures presented here for their studying and research activities. **Discussion:** we hope to facilitate the use of logistic regression and to spread replication as a data analysis teaching tool.

KEYWORDS: regression; logistic regression; replication; quantitative methods; transparency.

A produção desse manuscrito foi viabilizada através do patrocínio fornecido pelo Centro Universitário Internacional Uninter à *Revista de Sociologia e Política.*

Apêndice

Nesta seção apresentamos algumas informações que podem auxiliar o pesquisador na interpretação dos coeficientes da regressão logística. Em particular, examinamos a interpretação da razão de chance. Além disso, listamos algumas ferramentas de aprendizagem.

[32]Esta seção foi baseada em Schwab (2002).

• Entendendo a razão de chance (*odds ratio*)[32]

O termo razão de chance não é tão difundido na pesquisa aplicada em Ciência Política como média ou probabilidade. Em geral, como o pesquisador está comparando grupos/categorias, ele se interessa em analisar que grupo/categoria tem mais chance de ocorrer em relação ao outro grupo/categoria. Considere o seguinte exemplo: suponha que a probabilidade (p) de ocorrência um determinado evento é de 0,9. Dessa forma, ao se calcular o complementar, q = 1 - p, então 1 - 0,9 = 0,1. Chance é a divisão da probabilidade de ocorrência (p) pela probabilidade de não ocorrência (q). Então, 0,9/0,1 = 9. Afirma-se, então, que a chance de sucesso é 9 para 1. Por sua vez, a chance de fracasso dá-se por 0,1/0,9 = 0,11. Dizemos, então, que a chance de fracasso é de 1 para 9. Diferente da probabilidade que apenas pode assumir valores entre 0 e 1, a chance pode variar de 0 a infinito. Quando a probabilidade de ocorrência de um evento é maior do que a probabilidade de não ocorrência, a chance será maior do que 1. Quando a probabilidade de não ocorrência é maior, a chance será menor do que 1. Quando as probabilidades são iguais (ex. lançamento de uma moeda), a chance é igual a 1. Dado os propósitos pedagógicos deste artigo, é importante replicar os dados de Schawb (2002) para melhor entender esse conceito (Tabela 1A).

Tabela 1A - Frequência

Pena	**N**	**%**
Pena de morte	50	34
Prisão perpétua	97	66
Total	**147**	**100,0**

Fonte: Schwab (2002).

A Tabela 1A mostra que 34% dos presos foram condenados à pena de morte (n = 50/147). Isso quer dizer que a probabilidade de ocorrência desse evento é de 0,34. Por sua vez, a chance de ser condenado à pena capital é de 0,516 (50/97). Outra forma de dizer é que se tem aproximadamente a metade da chance de ser condenado à pena capital em relação a passar o resto da vida na prisão. Por fim, é possível inverter a interpretação e considerar que a prisão perpétua é cerca de duas vezes mais provável do que a pena de morte.

Até então não se tem nenhuma variável independente. O que o modelo logístico vai informar é o impacto de uma determinada variável sobre a chance de ocorrência de variável dependente. Por exemplo, considere a relação entre cor e tipo de sentença (Tabela 2A).

Tabela 2A - Tipo de pena por cor

Pena	**Negros**	**Não Negros**	**Total**
Pena de morte	28	22	50
Prisão perpétua	45	52	97
Total	**73**	**74**	**147**

Fonte: Schwab (2002).

É possível então calcular a chance para cada grupo específico: negros e não-negros. Para os negros tem-se 28/45 = 0,622. Para os não-negros tem-se 22/52 = 0,423. O impacto de ser negro pode ser representado pela divisão da chance do negro receber pena de morte (0,622) pela chance de um não negro receber a pena capital (0,423). 0,622/0,423 = 1,47. Para interpretar: a) negros tem 1,47 mais chance de receber a pena de morte do que não negros; b) ser negro aumenta 47% a chance de receber a pena capital (1,47-1*100).

Ferramentas de aprendizagem

http://www.icpsr.umich.edu/icpsrweb/sumprog/

No plano internacional, o *Summer Program in Quantitative Methods of Social Research* (ICPRS) é uma das principais iniciativas na difusão de métodos e técnicas de pesquisa.

http://www.fafich.ufmg.br/~mq/index.html

Curso intensivo de Metodologia Quantitativa em Ciências Humanas. É o curso mais tradicional no ensino de métodos e técnicas de pesquisa em Ciências Sociais no Brasil.

http://summerschool.ipsa.org/

Curso de verão organizado pela Associação Internacional de Ciência Política, Departamento de Ciência Política e o Instituto de Relações da Universidade de São Paulo (USP).

http://gking.harvard.edu/

Gary King disponibiliza artigos sobre metodologia, *softwares* específicos e bancos de dados para pesquisadores interessados em fazer replicações.

http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm

David Garson apresenta diferentes tópicos em estatística multivariada utilizando o Statistical Package for Social Sciences. Ao final de cada seção, tem-se uma bibliografia sugerida que pode ser utilizada como referência para ganhar mais profundidade no assunto.

http://www.statsoft.com/textbook/

Apresenta diferentes técnicas multivariadas utilizando o software Statistica.

http://www.ats.ucla.edu/stat/

Sítio eletrônico da Universidade da Califórnia (UCLA) especializado em técnicas multivariadas. Aqui o usuário encontra aplicações de diferentes softwares (SAS, SPSS, STATA, R, etc.), inclusive com vídeo aulas e tutoriais.

http://www.socr.ucla.edu/SOCR.html

Nesse endereço o leitor encontra jogos, aplicações, análises, entre outras ferramentas relacionadas ao ensino de Estatística e diferentes técnicas de pesquisa.

http://pan.oxfordjournals.org/

Political Analysis é um dos periódicos mais influentes da Ciência Política contemporânea e publica artigos na área de metodologia.

http://www.amstat.org/publications/jse/

Periódico especializado na divulgação de técnicas de ensino e aprendizagem de Estatística.

http://www.politicahoje.ufpe.br/index.php/politica

A Revista Política Hoje do Departamento de Ciência Política da UFPE publicou recentemente uma edição especial dedicada a Metodologia e Epistemologia em Ciência Política e Relações Internacionais.