O QUE É E COMO SUPERAR A MULTICOLINARIEDADE? UM GUIA PARA CIÊNCIA POLÍTICA*

WHAT IS AND HOW TO OVERCOME MULTICOLLINEARITY? A GUIDE FOR POLITICAL SCIENCE

Dalson Figueiredo Filho†
Lucas Silva‡
Amanda Domingos§

Resumo: Este artigo explica como detectar e superar problemas de multicolinariedade. Em particular, apresentamos quatro procedimentos para lidar com altos níveis de correlação entre variáveis em um modelo de regressão linear: (1) verificar a codificação e a transformação das variáveis; (2) aumentar o tamanho da amostra; (3) utilizar alguma técnica de redução de dados e (4) consultar a literatura específica sobre o tema. Nosso público alvo são estudantes de graduação e pós-graduação em Ciência Política em fases iniciais de treinamento. O desenho de pesquisa utiliza simulação para demonstrar como a multicolinariedade afeta a eficiência dos coeficientes estimados. Defendemos que importante progresso pode ocorrer em nossa disciplina se os pesquisadores checarem seus dados utilizando o *checklist* apresentado neste artigo.

Palavras-chave: Multicolinariedade. Regressão linear. Métodos quantitativos.

Abstract: This paper explains how to detect and overcome multicollinearity problems. In particular, we describe four procedures to handle high levels of correlation among variables in a linear regression model: (1) to check variables coding and transformation; (2) to increase sample size; (3) to employ some data reduction techniques and (4) to check specific literature on the subject. Our target audience are both undergraduate and graduate Political Science students in early training stages. The research design uses simulation to show how multicollinearity affects coefficients efficiency. We argue that significant progress can occur in our discipline if scholars check their data using the checklist presented in this article.

Keywords: Multicollinearity. Linear regression. Quantitative methods.

∗ Este trabalho se beneficiou dos comentários de Ricardo Borges, Nicole Janz e Richard Ball. A pesquisa contou com suporte financeiro do CNPQ e do *Teaching Integrity in Empirical Research workshop* (TIER – Haverford College). Agradecemos também ao apoio da PROPESQ e da PROACAD (UFPE). Eventuais limitações são monopólio dos autores.

† Professor do Departamento de Ciência Política da Universidade Federal de Pernambuco (UFPE).
‡ Aluno do Departamento de Ciência Política da Universidade Federal de Pernambuco (UFPE).
§ Aluna do Departamento de Ciência Política da Universidade Federal de Pernambuco (UFPE).

1 Introdução

A regressão linear de mínimos quadrados ordinários é a ferramenta mais utilizada na pesquisa empírica em Ciência Política (KRUGER; LEWIS-BECK, 2008). Desde que os seus pressupostos sejam devidamente respeitados, as estimativas serão eficientes e não viesadas, o que os econometristas definem como BLUE (*Best Linear Unbiased Estimator*). Não viesado, já que não existe tendência sistemática em sobre-estimar ou subestimar o verdadeiro valor do parâmetro populacional. E eficiente, já que o coeficiente apresenta a menor variância possível (KENNEDY, 2005).

E o que acontece quando os pressupostos são violados? Diferentes procedimentos podem ser empregados com o objetivo de garantir estimativas confiáveis dos parâmetros populacionais (KELLSTEDT; WHITTEN, 2013; WOOLDRIDGE, 2009; ACHEN, 2002). Neste trabalho nós discutimos um problema em particular: multicolinariedade (FARRAR; LAUBER, 1967). Isso porque altos níveis de correlação entre variáveis independentes produzem efeitos adversos sobre a consistência dos coeficientes. Portanto, dado que os modelos de regressão formam as engrenagens básicas de nossa disciplina, é importante que os cientistas políticos compreendam qual é o significado substantivo e as consequências práticas da multicolinariedade (HAIR et al., 2009).

Este artigo apresenta quatro procedimentos para lidar com altos níveis de correlação entre as variáveis independentes em um modelo de regressão linear: (1) verificar a codificação e a transformação das variáveis; (2) aumentar o tamanho da amostra; (3) utilizar alguma técnica de redução de dados (análise fatorial ou análise de componentes principais); e (4) consultar a literatura específica sobre o tema. Metodologicamente, o desenho de pesquisa utiliza simulação básica para demonstrar como a multicolinariedade afeta a eficiência dos coeficientes estimados. Além disso, adotamos o protocolo TIER 2.0 com o objetivo de aumentar a transparência e garantir a replicabilidade dos resultados (KING, 1995; PARANHOS et al., 2014; JANZ, 2015).

O restante do artigo está organizado da seguinte forma: a primeira seção define o que é multicolinariedade. Depois são apresentadas as suas principais consequências sobre a eficiência das estimativas. A terceira seção descreve como identificar a multicolinariedade. A quarta parte apresenta quatro procedimentos para lidar com altos níveis de correlação entre as variáveis independentes. A quinta e última seção utiliza uma simulação para ilustrar o que acontece com os coeficientes quando a correlação entre as variáveis é excessivamente alta.

2 O que é multicolinariedade?[1]

A ausência de colinariedade perfeita é um pressuposto chave para todos os modelos de regressão (KENNEDY, 2005; HAIR et al., 2009). Matematicamente, é impossível calcular o erro padrão quando a correlação entre as variáveis independentes é 1 ou -1. Por essa razão, quando falamos em multicolinariedade estamos nos referindo a altos níveis de correlação, ao invés de correlação exata entre X e Z. Diferente da autocorrelação e da heterocedasticidade, que são

[1] Ver Angrist e Pischke (2010), Gujarati e Porter (2009), Kellstedt e Whitten (2013), Long (1997), Wooldridge (2009), Achen (2002), Agresti e Finlay (2009), Beck (2010), Greene (2012) e Stock e Watson (2011). Ver também POLS... (2016).

problemas estatísticos, a multicolinariedade é um problema dos dados (LONG, 1997). Dessa forma, é possível existir multicolinariedade mesmo quando todos os pressupostos do modelo linear de mínimos quadrados são respeitados (ACHEN, 2002).

Ainda que seja extremamente improvável observar correlações perfeitas na prática, é comum a existência de algum nível de associação entre as variáveis explicativas. A Tabela 1 apresenta um exemplo de colinearidade.

Tabela 1. Colinearidade com três variáveis independentes

	Y	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
Y	1			
X_1	0,720	1		
X_2	0,670	0,750	1	
X_3	0,600	0,740	0,960	1

Fonte: Elaboração dos autores (2016).

A correlação entre X_2 e X_3 (0,960) é mais forte do que correlação entre X_1 e Y (0,720), X_2 e Y (0,670) e X_3 e Y (0,600). Ou seja, tem-se mais variância compartilhada entre as variáveis independentes do que entre as variáveis explicativas e a variável dependente. Para os propósitos deste artigo, definimos multicolinariedade como altos níveis de correlação entre duas ou mais variáveis independentes em um modelo de regressão. E quanto maior a correlação, maiores são os eventuais problemas.

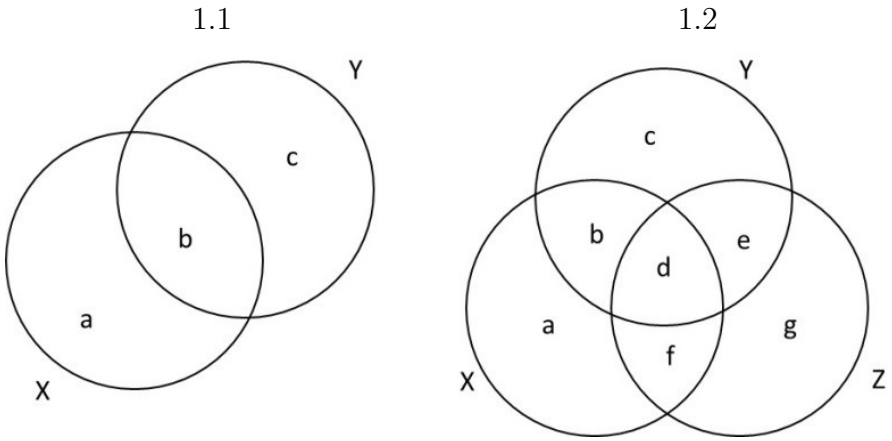
3 Consequências?

O principal efeito de altos níveis de correlação entre as variáveis explicativas é a ineficiência das estimativas. A multicolinariedade tende a sobrestimar a magnitude dos erros padrões dos coeficientes de regressão, prejudicando a confiabilidade dos testes de significância (p-valor e intervalos de confiança). Dessa forma, o p-valor tenderá a ser maior e os intervalos de confiança serão menos precisos. Isso porque, na presença de variáveis muito colineares, tem-se menos informação para calcular o efeito individual de cada variável independente sobre a variável dependente. Ao fim, menos informação gera uma maior variação, o que implica em uma menor precisão. Em geral, os coeficientes permanecem não viesados, mas perdem a propriedade da eficiência (menor variância possível).

Além disso, altos níveis de correlação entre as variáveis independentes podem produzir um modelo em que a maior parte dos coeficientes não são significativos, mas apresenta alto coeficiente de determinação (r^2), o que não faz sentido do ponto de vista substantivo.[2] Por fim, outro problema gerado por variáveis independentes colineares é a instabilidade dos coeficientes. Tanto a inclusão e/ou exclusão de um único caso e/ou o acréscimo de uma nova variável pode mudar dramaticamente a magnitude e, de forma mais preocupante, a direção dos coeficientes. A Figura 1 ilustra como esse problema afeta a eficiência das estimativas.

[2] Para os interessados em saber mais sobre as limitações do coeficiente de determinação, ver o artigo de King (1991), que foi traduzido nesta edição pela Revista Conexão Política. Ver também Figueiredo Filho, Silva e Rocha (2010).

Figura 1. Esquematização da multicolinariedade



Fonte: Elaboração dos autores (2016).

A Figura 1.1 apresenta duas variáveis. A interseção entre X e Y está ilustrada pela letra b e representa a correlação entre as variáveis. A letra c representa a variação de Y, que independe da variação em X. Já a Figura 1.2 tem duas variáveis independentes (X e Z) e a mesma variável dependente (Y). A área $d+f$ representa a correlação entre as variáveis independentes. Se apenas a variável X for utilizada para entender/explicar/predizer Y, tem-se informação referente à área $b+d$. Se apenas a variável Z for utilizada para entender/explicar/predizer Y, tem-se informação referente a área $d+e$. Mas o que acontece se forem utilizadas as variáveis X e Z ao mesmo tempo? A regressão linear de mínimos quadrados ordinários utiliza apenas a variância única entre cada variável independente e a variável dependente. Ou seja, toda a informação da área d seria perdida (área comum entre X e Z). Portanto, quanto maior for a correlação entre as variáveis independentes, menos informação estará disponível para calcular as estimativas dos coeficientes (KENNEDY, 2009). Logo, menor a eficiência.

4 Como detectar?

A forma mais simples de detectar a multicolinariedade é estimar uma matriz de correlação entre as variáveis independentes. Quanto maior a magnitude dos coeficientes, maiores os eventuais problemas. A literatura indica 0,9, independente do sinal, como parâmetro. Outra possibilidade é tratar cada variável independente, como se ela fosse uma variável dependente e estimar um modelo explicativo a partir das demais variáveis independentes. Quanto maior o coeficiente de determinação, mais graves são os problemas de multicolinariedade. Tecnicamente, a literatura indica duas medidas sínteses para diagnosticar problemas de colinariedade: (1) Tolerância e (2) Fator de Inflação da Variância (*Variance Inflation Factor – VIF*).

A tolerância é a quantidade de variabilidade de uma variável independente que não é explicada pelas demais variáveis independentes. Ela é calculada a partir de $1 - r^2$. Por exemplo, se o modelo explica 30% da variável independente, então a tolerância de X_1 é de $0,70(1–0,3)$. Quanto maior a tolerância, menor nível de colinariedade.

Por sua vez, o VIF é calculado como o inverso da tolerância. Por exemplo, se a tolerância

é de 0, 7, o VIF será de 1, 43(1/0, 7). Dessa forma, quanto maior o VIF, mais sérios os problemas de correlação entre as variáveis independentes. Uma propriedade interessante do VIF é que a sua raiz quadrada informa o aumento esperado na magnitude do erro padrão. Por exemplo, um VIF de nove indica que o erro padrão triplicou de tamanho, enquanto um VIF de quatro sugere que o erro padrão dobrou. E quanto maior o erro padrão, maiores serão os intervalos de confiança e mais difícil será de observar a significância estatística das estimativas. Como regra geral, sugerimos os seguintes parâmetros para interpretar o Fator de Inflação da Variância (VIF):

Gráfico 1: VIF

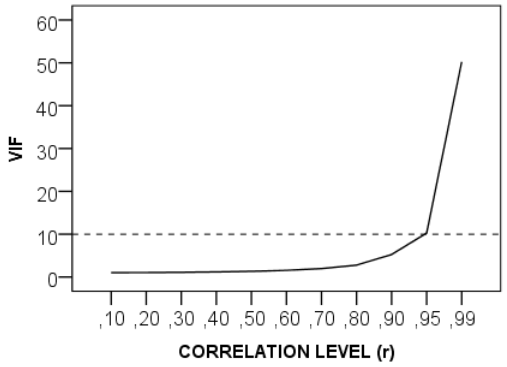


Tabela 2. Correlação x VIF

VIF	Significado
Até 1	Ausência de multicolinariedade
Entre 1 e 10	Multicolinariedade aceitável
> 10	Multicolinariedade problemática

5 Como superar?[3]

Este artigo apresenta quatro procedimentos para superar os problemas de multicolinariedade: (1) verificar a codificação e transformação das variáveis; (2) aumentar o tamanho da amostra; (3) utilizar alguma técnica de redução de dados; e (4) consultar a literatura específica sobre o tema.

A primeira recomendação é analisar as codificações e transformações realizadas nas variáveis. Isso porque, muitas vezes, os pesquisadores optam por recodificar uma variável e um simples deslize de atenção pode produzir efeitos perversos sobre a qualidade do modelo, principalmente em amostras pequenas. Por exemplo, uma recodificação mal renomeada pode criar problemas de colinariedade, já que a mesma variável vai ser duplamente incluída. Erros de digitação também podem produzir variáveis altamente correlacionadas ao acaso devido à presença de casos extremos. Por esse motivo, é importante checar a distribuição de cada variável por observações atípicas. Após a checagem dos dados, se os problemas persistirem, a próxima recomendação é aumentar o tamanho da amostra.

Kennedy (2005) e Achen (2002) sugerem que os problemas de multicolinariedade são especialmente recorrentes em amostras pequenas (micronumerosidade). Dessa forma, sugerimos elaborar um desenho de pesquisa que maximize a quantidade de observações (KING; KEOHANE; VERBA, 1994). Por exemplo, se a unidade de análise é o ente federativo (N = 27), uma forma

[3] Uma das opções é não fazer nada e reportar os coeficientes. Kennedy (2005) sugere duas regras que justificam a inércia diante da multicolinariedade: a) se o coeficiente de determinação (r2) do modelo for maior do que os coeficientes de determinação regredidos para cada variável independente e b) se a estatística t for maior do que dois para todas as variáveis explicativas, independente do sinal. Além disso, a multicolinariedade também não apresenta problemas para modelos puramente preditivos em que o foco está na capacidade conjunta das variáveis e não no efeito individual de cada regressor.

de aumentar a amostra é mudar a unidade amostral básica e coletar os dados por município ($N > 5.000$). Similarmente, se a amostra tem informações para os países da América Latina, é possível incluir mais casos e dessa forma minimizar esse problema. Uma opção adicional é manter a unidade de análise constante e aumentar a quantidade de períodos disponíveis, formando assim um painel.[4] Todavia, a adição de novas observações ou é muito caro, muito demorado ou o pesquisador já possui dados para a população.

Nossa terceira sugestão é utilizar alguma técnica de redução de dados (FIGUEIREDO FILHO; SILVA JUNIOR, 2010; FIGUEIREDO FILHO et al., 2014). Essas técnicas são especialmente adequadas para lidar com variáveis independentes, fortemente correlacionadas (HAIR et al., 2009). Dessa forma, é possível reduzir a dimensionalidade dos dados e criar um índice que carrega a informação das variáveis originais. Esse novo indicador pode ser utilizado como variável dependente ou independente em novos modelos explicativos (TABACHINICK; FIDELL, 2007). Uma desvantagem desse procedimento é a impossibilidade de observar o efeito individual de cada variável explicativa. Aqui vale a máxima: ninguém pode ter tudo.

Nossa última recomendação é consultar a literatura específica sobre o fenômeno de interesse e identificar as variáveis mais teoricamente relevantes. Muitos modelos incluem variáveis por comodidade ou simplesmente para "ver o que acontece". É importante que o pesquisador apenas inclua variáveis que são relevantes para explicar o seu fenômeno de interesse. Desaconselhamos fortemente a exclusão arbitrária de variáveis colineares. Isso porque a exclusão de uma variável teoricamente importante pode gerar problemas de especificação, que são mais graves do que os gerados por variáveis altamente correlacionadas. Um modelo mal especificado produz estimativas viesadas, um modelo com multicolonariedade não. Em síntese, os pesquisadores apenas devem excluir variáveis da análise quando existem razões substantivas. Caso contrário, o remédio pode ser pior do que a doença.

6 Simulando para entender

Para fixar a interpretação, optamos por simular o seguinte modelo:

$$y = 0,5 + 4X_1 + 2X_2 + \epsilon$$

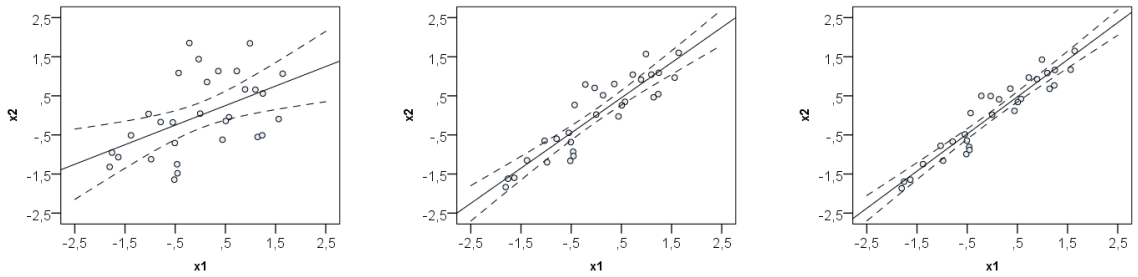
A variável dependente (y) vai ser calculada a partir dos seguintes parâmetros: constante igual a $0,5$, β_1 igual a 4 e β_2 igual a 2. O erro ϵ tem média zero e distribuição normal. X_1 e X_2 apresentam correlação de $0,5$, $0,9$ e $0,95$, respectivamente, conforme ilustra a Figura 2.

Para todos os casos, utilizamos uma amostra com 30 observações. Quanto maior a correlação entre as variáveis independentes, maior o tamanho do erro padrão dos coeficientes. Similarmente, os intervalos de confiança da estimativa aumentam à medida com que a colinariedade entre as variáveis cresce. Em particular, quando a correlação atinge $0,95$ com um VIF de $10,256$, β_2 deixa de ser significativo, já que o intervalo de confiança passa pelo zero. Em termos

[4] No entanto, a análise de dados longitudinais requer outras especificidades que estão fora do escopo deste trabalho. Para os interessados no assunto, ver Beck e Katz (1995), Hedeker e Gibbons (2006) e Fitzmaurice, Laird e Ware (2001). Ver também Torres-Reyna (2016).

Figura 2. Mesma amostra, diferentes correlações

(a) $r = 0,5$ **(b)** $r = 0,9$ **(c)** $r = 0,95$



(a)	(b)	(c)
$\alpha = 0,626(0,224)$	$\alpha = 0,626(0,224)$	$\alpha = 0,626(0,224)$
$\beta_1 = 4,308(0,263)$	$\beta_1 = 4,807(0,523)$	$\beta_1 = 5,136(0,729)$
$(3,768; 4,847)$	$(3,735; 5,879)$	$(3.639; 6.632)$
$\beta_2 = 1,612(0,263)$	$\beta_2 = 1,229(0,523)$	$\beta_2 = 0,924(0,729)$
$(1,072; 2,152)$	$(0,157; 2,301)$	$(-0,573; 2,421)$
Tolerância$= 0,750$	*Tolerância*$= 0,190$	*Tolerância*$= 0,097$
$VIF = 1,333$	$VIF = 5,263$	$VIF = 10,256$

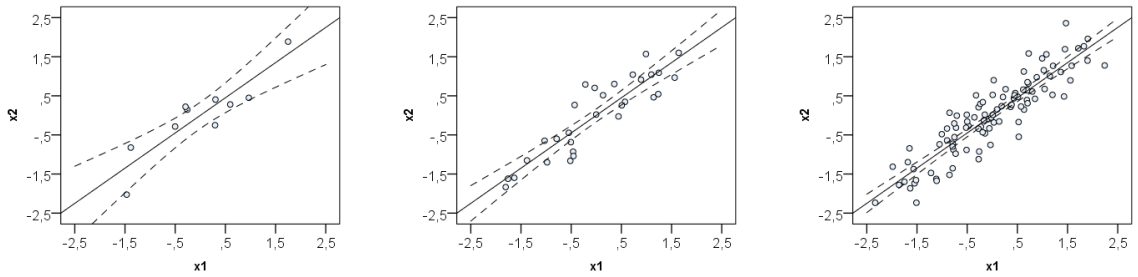
Fonte: elaborado pelos autores (2016).

substantivos, o pesquisador seria levado a cometer o erro do tipo 2: não rejeitar a hipótese nula quando ela deveria ser rejeitada. Nesse caso, β_2 é diferente de zero, mas o pesquisador concluiria que não existe relação entre $X2$ e Y.

Por fim, utilizamos outra simulação mantendo a correlação entre as variáveis constantes ($r = 0,9$) e variando o número de casos (10, 30 e 100). Todos os demais parâmetros são os mesmos. A Figura 3 ilustra essas informações.

Figura 3. Mesma amostra, diferentes correlações

(a) $n = 10$ **(b)** $n = 30$ **(c)** $n = 100$



(a)	(b)	(c)
$\alpha = 0,073(0,511)$	$\alpha = 0,626(0,224)$	$\alpha = 0,546(0,089)$
$\beta_1 = 4,203(1,235)$	$\beta_1 = 4,807(0,523)$	$\beta_1 = 3.617(0,205)$
$(1,282; 7,124)$	$(3,735; 5,879)$	$(3,209; 4,024)$
$\beta_2 = 1.781(1,235)$	$\beta_2 = 1,229(0,523)$	$\beta_2 = 2,305(0,205)$
$(-1,140; 4,702)$	$(0,157; 2,301)$	$(1,898; 2,713)$
Tolerância$= 0,190$	*Tolerância*$= 0,190$	*Tolerância*$= 0,190$
$VIF = 5,263$	$VIF = 5,263$	$VIF = 5,263$

Fonte: elaborado pelos autores (2016).

O tamanho da amostra não afeta o valor do VIF, que permanece constante em $5,263$. No entanto, o efeito da multicolinariedade sobre a eficiência das estimativas é menor quando o número de observações aumenta. Por exemplo, com uma amostra de 10 não seria possível rejeitar a hipótese nula de que β_2 é igual a zero, já que o intervalo de confiança passa pelo zero. Por outro lado, com 30 observações, a multicolinariedade não altera a interpretação dos

testes de significância. Em particular, com uma amostra de 100 casos, os intervalos de confiança variam pouco, ou seja, têm-se estimativas com menor variabilidade, ou seja, mais eficientes. É nesse sentido que a coleta de mais observações é um "santo remédio" para resolver problemas de multicolinariedade.

7 Conclusão

Este artigo apresentou uma introdução sobre como detectar e superar problemas de multicolinariedade. O foco repousou sobre a compreensão intuitiva dos conceitos, já que nosso público alvo são estudantes de graduação e pós-graduação em fases iniciais de treinamento. Nossa principal motivação é a escassez de material pedagógico, especialmente voltado para Ciência Política. Metodologicamente, reproduzimos as principais recomendações da literatura e utilizamos simulação para demonstrar o efeito de altos níveis de correlação entre as variáveis explicativas sobre a eficiência das estimativas.

Além disso, apresentamos quatro procedimentos que podem ajudar a resolver os problemas de multicolinariedade: (1) verificar a codificação e transformação das variáveis; (2) aumentar o tamanho da amostra; (3) utilizar alguma técnica de redução de dados; e (4) consultar a literatura específica sobre o tema. Todas as rotinas computacionais foram devidamente reportadas com o objetivo de aumentar a transparência e garantir a replicabilidade dos resultados. Com este artigo esperamos atingir dois objetivos complementares: a) incentivar a produção de trabalhos na área de metodologia política e b) aprimorar a qualidade dos resultados empíricos reportados pela Ciência Política nacional. Além disso, defendemos que importante progresso pode ocorrer em nossa disciplina se os pesquisadores checarem seus dados, utilizando o *checklist* apresentado neste artigo.

Referências

ACHEN, Christopher. Advice for students taking a first political science graduate coursein statistical methods. *The Political Methodologist*, v. 10, n. 2 p. 10-12, 2002.

AGRESTI, Alan; FINLAY, Barbara. *Statistical methods for the social sciences*: with SPSS from A to Z: a brief step-by-step manual. Pearson, 2009.

ANGRIST, Joshua; PISCHKE, Jörn-Steffen. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, v. 24, n. 2, p. 3-30, 2010.

BECK, Nathaniel; KATZ, Jonathan N. What to do (and not to do) with time-series cross-section data. *American political science review*, v. 89, n. 3, p. 634-647, 1995.

BECK, Nathaniel. Making regression and related output more helpful to users. *The Political Methodologist*, v. 18, n. 1, p. 4-9, 2010.

FARRAR, Donald; GLAUBER, Robert. Multicollinearity in regression analysis: the problem revisited. *The Reviewof Economic and Statistics*, v. 49, v, 1, p. 92-107, 1967.

FIGUEIREDO FILHO, D. B; SILVA JÚNIOR, José Alexandre da. Visão além do alcance: uma introdução à análise fatorial. *Opinião Pública*, v. 16, n. 1, p. 160-185, 2010.

GREENE, William. *Econometric analysis*. 7th ed. Upper Saddle River: Prentice Hall, 2012.

GUJARATI, Damodar; PORTER, Dawn. *Basic econometrics*. Mc Graw-Hill International Edition, 2009.

HAIR, Joseph et al. *Análise multivariada de dados*. Bookman, 2009.

JANZ, Nicole. Bringing the gold standard into the classroom: replication in University teaching. *International Studies Perspectives*, 2015.

KELLSTEDT, Paul; WHITTEN, Guy. *The fundamentals of political science research*. Cambridge University Press, 2013.

KENNEDY, Peter. *A guide to econometrics*. MIT press, 2005.

____. *A guide to econometrics*. Wiley-Blackwell, 2009.

KING, Gary. Replication, replication. *PS: Political Science & Politics*, v. 28, n. 3, p. 444-452, 1995.

KING, Gary; KEOHANE, Robert; VERBA, Sidney. *Designing social inquiry*: scientific inference in qualitative research. Princeton University Press, 1994.

KRUEGER, James; LEWIS-BECK, Michael. Is ols dead? *The Political Methodologist*, v. 15, n. 2. p. 2-4, 2008.

LONG, Scott. Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences Number 7*. Sage Publications, Thousand Oaks, 1997.

PARANHOS, R. et al. A importância da replicabilidade na Ciência Política: O caso do SIGOBR. *Revista Política Hoje*, v. 22, v. 2, p. 213-229, 2014.

POLS 509: the linear model - lecture 7 - violations of the OLS assumptions. Disponível em: <https://www.youtube.com/watch?v=kBlUStQcnJU>. Acesso em: 12 out. 16.

STOCK, J. H.; WATSON, M. W. Dynamic factor models. *Oxford Handbook of Economic Forecasting*, 1, p. 35-59, 2011.

TABACHNICK, B. G.; FIDELL, L. S. *Using multivariate statistics*. 5. ed. Needham Height: Allyn & Bacon, 2007

TORRES-REYNA, Oscar. *Panel data analysis fixed and random effects using stata*. v. 4.2. Dec. 2007. Disponível em: <https://www.princeton.edu/~otorres/Panel101.pdf>. Acesso em: 13 dez. 2016.

WOOLDRIDGE, J. M. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, v. 104, n. 3, p. 112-114, 2009.