O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de mínimos quadrados ordinários (MQO) *

Dalson Figueiredo Filho (UFPE)

Felipe Nunes (UCLA)

Enivaldo Rocha (UFPE)

Manoel Santos (UFMG)

Mariana Batista (UFPE)

José Alexandre Silva (UFG/UFPE)

[Artigo publicado na revista *Política Hoje*: vol 20, n. 1, 2011]†

*O título desse artigo foi diretamente inspirado no trabalho de Beck e Katz (1995): *What to do (and not to do) with Times-Series Cross-Section Data* publicado na *American Political Science Review*. Os autores agradecem a Natalia Leitão pelos comentários em versões preliminares e ao parecerista anônimo da revista "Política Hoje" por suas valiosas contribuições. Lembrando sempre que omissões remanescentes são integralmente creditas aos autores.

†Os autores registram que esse trabalho consiste na superação de um desafio nada trivial no contexto da produção científica. Como se trata de um trabalho coletivo, com efetiva participação de todos os envolvidos, caracteriza-se como uma realização nada comum. No contexto de comunicação e cooperação necessárias à sua conclusão, para além do produto final, fica o saldo positivo do aprendizado coletivo e a certeza de que vale a pena trabalhar em grupo. Para os cientistas políticos, esse resultado pode ser resumido como um jogo de *soma positiva*. Contudo, vencer os problemas de ação coletiva e a distância exigem, além de compromisso, apoio institucional. Nesse sentido, é importante registrar que essa atividade teve sucesso sobretudo em função do apoio recebido do Professor Enivaldo Rocha (UFPE) e Carlos Ranulfo Melo (UFMG) respectivos coordenadores do convênio PROCAD/CAPES celebrado entre os departamentos de Ciência Política da UFPE e da UFMG. Os recursos por eles disponibilizados e a atenção pedagógica direta permitiram não apenas a nossa mobilidade, como também a ação articulada entre professores e estudantes de ambas as instituições.

Resumo

Para que serve o modelo de regressão de mínimos quadrados ordinários? Como os cientistas sociais podem utilizar essa ferramenta em seus desenhos de pesquisa? Como evitar aplicações inadequadas dessa técnica? O principal objetivo desse artigo é apresentar a lógica do modelo de regressão linear de mínimos quadrados ordinários (MQO) a partir de uma abordagem intuitiva. Por meio de uma simulação de Monte Carlo demonstramos os procedimentos que devem ser seguidos para planejar, interpretar e avaliar o modelo de regressão linear. Esperamos com esse trabalho difundir, em geral, o uso da referida técnica na ciência social no Brasil, e em particular, na ciência política.

Abstract

What is ordinary least squares regression model good for? How could social scientists employ this technique in their research designs? How to avoid common problems with its application? The main objective of this paper is to present the underlying logic of the ordinary least square regression model based on an intuitive approach. Through a Monte Carlo simulation we demonstrate the main procedures to plan, interpret and evaluate such technique. We aim with this work to diffuse the use of the OLS in the Brazilian social sciences in general, and in political science in particular.

'A unidade de toda a ciência consiste apenas em seu método, não em seu material'
(Pearson, 1982: 16)

1 Introdução

A análise de regressão de mínimos quadrados ordinários (MQO) é o modelo estatístico mais usualmente empregado na ciência política contemporânea. Após analisarem 1.756 artigos publicados entre 1990 e 2005 em três importantes periódicos da área (*American Political Science Review, American Journal of Political Science e Journal of Politics*), Krueger e Lewis-Beck (2008) reportam que 30,8% das publicações utilizam a regressão linear de mínimos quadrados ordinários (*Ordinary Least Squares - OLS*). No Brasil a utilização pode ser considerada tímida, sobretudo se comparada a norte-americana. Para Soares (2005), existe uma "hostilidade em relação aos métodos quantitativos e à estatística [na ciência social brasileira]"(Soares, 2005: 27). Os trabalhos de Werneck Vianna et al (1988), Valle e Silva (1999) e Santos e Coutinho (2000) corroboram esse diagnóstico: a utilização de técnicas básicas de estatística descritiva e inferencial ainda é bastante limitada nas Ciências Sociais, isso independentemente do tipo de produção (artigos, dissertações ou teses). O resultado prático é o enfraquecimento generalizado do conhecimento "científico". Em especial, porque é sabido que a utilização dos métodos qualitativos não tem melhor sorte. Ou seja, segundo Soares (2005), a ausência de métodos quantitativos não significa a presença de métodos qualitativos. Regra geral, o padrão é o não método.

Diante desse quadro, o que pode ser feito? Partindo da hipótese de que a resistência é causada pelo não domínio das técnicas (e num sentido mais geral dos fundamentos básicos de estatística), algumas medidas podem ser pensadas. Primeiro, pode-se aumentar a oferta de cursos de metodologia quantitativa, quer seja nos currículos regulares quer seja via cur-

sos de especialização[1]. Segundo, parece bem-vindo examinar criticamente a literatura para identificar quais são as demandas mais latentes. Os dois caminhos têm o mesmo objetivo: assegurar que a preocupação com o método seja uma constante. O principal objetivo desse texto é contribuir com essa perspectiva através de uma introdução à análise de regressão linear de mínimos quadrados ordinários (MQO). Nossa principal meta é apresentar as principais características desse modelo de regressão, discutindo os pressupostos que devem ser obedecidos, assim como formas simples de compreender a sua aplicabilidade.

Para tanto, o artigo está dividido da seguinte forma. A primeira parte apresenta a estrutura básica do modelo de regressão. A meta é familiarizar o leitor com os componentes do modelo. A segunda seção discute alguns dos pressupostos que precisam ser satisfeitos, bem como as consequências de sua violação sobre a consistência das estimativas. A terceira parte ilustra a aplicação prática de uma análise de regressão, identificando os principais requisitos técnicos que devem ser satisfeitos pelo pesquisador. O objetivo é auxiliar a construção de um desenho de pesquisa que favoreça a utilização da referida técnica. A quarta parte apresenta a simulação dos dados utilizados nesse trabalho bem como os resultados, destacando as principais estatísticas de interesse e a sua interpretação substantiva. Nesta seção enfatiza-se a utilização de gráficos como ferramenta fundamental para a interpretação dos resultados do modelo. Na parte final do trabalho discutimos sumariamente alguns cuidados que os pesquisadores devem tomar durante a utilização do modelo de mínimos quadrados ordinários.

[1]Em particular, a escassez de cursos de métodos e técnicas, sejam eles quantitativos e qualitativos, acaba prejudicando a formação dos profissionais na área de ciências sociais, além de reduzir a qualidade técnica da produção acadêmica. Um dos principais esforços para minorar esse problema foi materializado através do curso de Metodologia Quantitativa (MQ) em Ciências Humanas realizado anualmente pelos departamentos de Sociologia e Ciência Política da Universidade Federal de Minas Gerais (UFMG). No plano internacional destaca-se o EMAS organizado pela Universidade de Salamanca, Espanha, o *Summer Program in Quantitative Methods of Social Research*, ICPSR, Michigan, EUA e o *Summer School in Methods and Techniques* organizado pelo European Consortium for Political Research. Tem-se, ainda, a *Essex Summer School in Social Sciences and Data Analysis*, Londres, Inglaterra. Entre 31 janeiro e 12 de fevereiro de 2011 a IPSA realizou um curso de verão na Universidade de São Paulo (USP) - *"Concepts, Methods, and Techniques in Political Science"*.

2 Entendendo o modelo de regressão de mínimos quadrados ordinários (MQO)[2]

O modelo regressão linear é uma poderosa ferramenta em análise de dados[3]. Hair et al (2009) afirmam que "a análise de regressão múltipla é uma técnica estatística que pode ser usada para analisar a relação entre uma única variável dependente e múltiplas variáveis independentes (preditoras)"(Hair et al, 2009: 176)[4]. Com a regressão é possível estimar o grau de associação entre Y, variável dependente e X_i, conjunto de variáveis independentes (explicativas). O objetivo é resumir a correlação entre X_i e Y em termos da direção (positiva ou negativa) e magnitude (fraca ou forte) dessa associação. Mais especificamente, é possível utilizar as variáveis independentes para predizer os valores da variável dependente. Em regressões multivariadas compostas de mais de uma variável independente é possível também identificar a contribuição de cada variável independente sobre a capacidade preditiva do modelo como um todo. Tecnicamente, dizer que o modelo é ajustado utilizando a forma funcional de mínimos quadrados ordinários significa que uma reta que minimiza a soma dos quadrados dos resíduos será utilizada para resumir a relação linear entre Y e X_i[5] (Krueger e Lewis-Beck, 2008). Pedagogicamente, é importante apresentar a notação do modelo de regressão linear:

$$Y = \alpha + \beta_1 X_1 + \epsilon$$

[2]Para os propósitos desse artigo minimizamos o grau de complexidade matemática dos conceitos apresentados. Para os leitores interessados em aprofundar seus conhecimentos sugerimos cobrir a bibliografia citada. Em particular, para uma introdução bastante didática à análise multivariada de dados ver Hair et al (2009). Para uma opção mais avançada ver Tabachnick e Fidell (2007). Em relação a conceitos elementares de estatística sugerimos Moore e McCabe (2009). Em Econometria sugerimos Wooldridge (2009), Kennedy (2009) e Gujarati (2000).

[3]Hair et al (2006) afirmam que "a principal razão para a popularidade de regressão tem sido a sua capacidade de prever e explicar variáveis métricas"(Hair et al 2006: 269).

[4]Similarmente, Pallant (2007) afirma que "regressão múltipla não é apenas uma técnica, mas uma família de técnicas que podem ser usadas para explorar a relação entre uma variável dependente contínua e um número de variáveis independentes ou preditoras"(Pallant, 2007: 146).

[5]Nas palavras de Hair et al (2009) "procedimento de estimação utilizado na regressão simples e múltipla em que os coeficientes de regressão são estimados de forma a minimizar a soma total dos quadrados dos resíduos"(Hair et al, 2009: 172).

Y representa a variável dependente, ou seja, aquilo que queremos explicar/entender/predizer. X_1, por sua vez, representa a variável independente, aquilo que o pesquisador acredita que pode ajudar a explicar/entender/predizer a variação de Y. O intercepto (α), também chamado de constante, representa o valor de Y quando X_1 assume valor zero. Dito de outra forma, na ausência de variáveis independentes, o intercepto (α) representa o valor da média esperada de Y. O coeficiente de regressão (β_1) representa a mudança observada em Y associada ao aumento de uma unidade em X_1. Finalmente, o termo estocástico (ϵ) representa o erro em explicar/entender/predizer Y a partir de X_1. Em particular, ϵ é a diferença entre os valores observados e os valores preditos de Y, ou seja, os resíduos do modelo. Os resíduos de um modelo de regressão são parte fundamental para que se avalie a capacidade do pesquisador em produzir um modelo (representação formal do mundo) que represente de forma acurada a realidade estudada (aqui representada pelos dados analisados). É essa abordagem teórica que nos permite afirmar (com bastante cautela) que quanto menor os resíduos encontrados, melhor é o ajuste do nosso modelo à realidade a ser explicada. Para os propósitos desse artigo julgamos importante ilustrar graficamente o funcionamento do modelo de regressão linear. A figura 1 abaixo apresenta as informações relevantes.

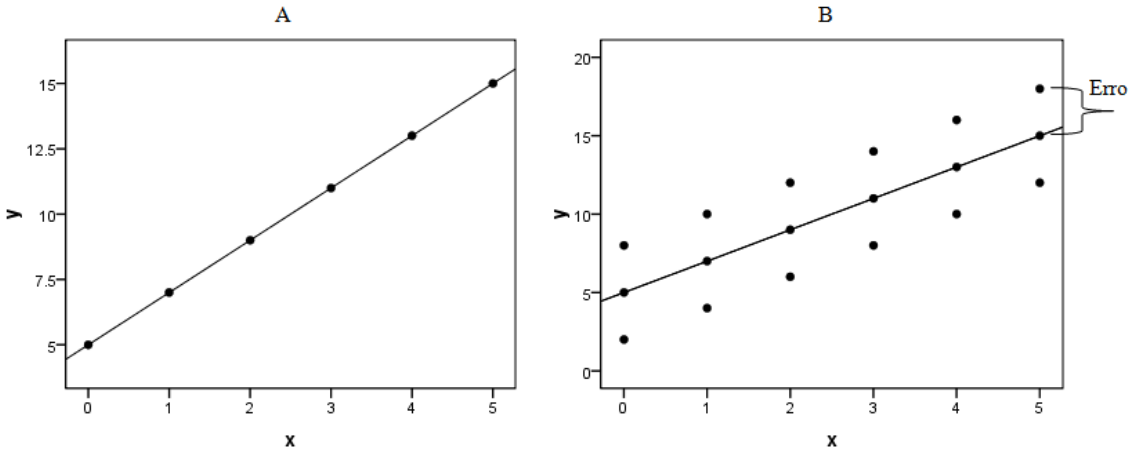

Figura 1: Funcionamento do modelo de regressão (MQO)

Na figura A, existe uma relação linear perfeita entre X (variável independente) e Y (o fenômeno que o pesquisador quer entender/explicar/predizer). Isso quer dizer que ao saber o valor de X, é possível determinar, exatamente, o valor de Y[6]. Na figura B, a relação entre as variáveis é inexata, ou seja, existe erro em predizer o valor de Y a partir dos valores de X. A forma funcional de mínimos quadrados é assim denominada porque minimiza os erros de estimação entre os valores observados e os valores preditos, ou seja, minimiza os resíduos. Dito de outra forma, o modelo de mínimos quadrados ordinários minimiza o erro em entender/explicar/predizer os valores de Y a partir dos valores de X. Essas estimações são eficientes desde que os pressupostos subjacentes à análise de regressão sejam devidamente respeitados. A próxima seção apresenta alguns desses pressupostos e as consequências de sua violação sobre a consistência das estimativas.

3 Os pressupostos do modelo de mínimos quadrados ordinários (MQO)

Diferentes autores apresentam pressupostos distintos que precisam ser satisfeitos para que a análise de regressão de mínimos quadrados ordinários possa ser adequadamente utilizada, produzindo o Melhor Estimador Linear Não-Viesado (MELNV)[7]. Por exemplo, Lewis-Beck

[6]Dificilmente o pesquisador irá observar uma relação perfeita entre suas variáveis de interesse. Nas palavras de Lewis-Beck (1980), "Um exemplo do mundo real com o qual todos estamos familiarizados é Y = 32 + 9/5X onde a temperatura em Fahrenheit (Y) em uma função linear exata de temperatura em Celsius (X). Em contraste, as relações entre as variáveis nas ciências sociais são quase sempre inexatas"(Lewis-Beck, 1980: 10).

[7]Um estimador é *Best Linear Unbiased Estimator* (*BLUE*) quando as seguintes propriedades são satisfeitas: (*Best*) Melhor significa eficiente, que produz a menor variância, (*Linear*) linear refere-se ao tipo de relação esperada entre as variáveis e (*Unbiased*) não-viesado diz respeito à distribuição amostral do estimador. Um estimador enviesado é aquele que sistematicamente sobreestima ou subestima o valor do parâmetro populacional. Para Kennedy (2009), "um estimador β^* é considerado um estimador não-viesado (ou não tendencioso) de β se a média de sua distribuição amostral é igual a β, isto é, se o valor médio de β^* em amostras repetidas é igual a β"(Kennedy, 2009: 15). Mais adiante, Kennedy (2009) afirma que "um estimador linear não-viesado, e que tem variância mínima entre todos os estimadores lineares não-viesados, é chamado de melhor estimador linear não-viesado (*Best Linear Unbiased Estimator - BLUE*)"(Kennedy, 2009: 16).

(1980) e Kennedy (2009) elencam os seguintes pressupostos: (1) a relação entre a variável dependente e as variáveis independentes deve ser linear; (2) as variáveis foram medidas adequadamente, ou seja, assume-se que não há erro sistemático de mensuração; (3) a expectativa da média do termo de erro é igual a zero; (4) homocedasticidade, ou seja, a variância do termo de erro é constante para os diferentes valores da variável independente; (5) ausência de autocorrelação, ou seja, os termos de erros são independentes entre si; (6) a variável independente não deve ser correlacionada com o termo de erro; (7) nenhuma variável teoricamente relevante para explicar Y foi deixada de fora do modelo e nenhuma variável irrelevante para explicar Y foi incluída no modelo; (8) as variáveis independentes não apresentam alta correlação, o chamado pressuposto da não multicolinearidade; (9) assume-se que o termo de erro tem uma distribuição normal e (10) há uma adequada proporção entre o número de casos e o número de parâmetros estimados.

Tecnicamente, a violação de cada pressuposto está associada a um determinado problema. Dessa forma, é importante entender, ainda que de maneira geral, qual é a função de cada um desses pressupostos. Para os propósitos desse artigo, elencamos dez pressupostos que precisam ser satisfeitos na utilização do modelo de regressão linear de mínimos quadrados para que as estimativas produzidas sejam consistentes.

O primeiro pressuposto que deve ser respeitado é a linearidade dos parâmetros, ou seja, deve-se esperar que a relação entre as variáveis independentes e a variável dependente possa ser representada por uma função linear[8]. Quanto mais a relação se distanciar de uma função linear, menor é a aplicabilidade da forma funcional de mínimos quadrados para ajustar o modelo. Em outras palavras cresce a diferença entre os parâmetros estimados e os observados.

[8]Note que o requisito da linearidade é nos parâmetros e não nas variáveis. Para Hair et al (2009), "um pressuposto implícito de todas as técnicas de análise multivariada com base em medidas correlacionais de associação, incluindo regressão múltipla, regressão logística, análise fatorial e modelagem de equações estruturais, é a linearidade. Porque correlações representam apenas a associação linear entre as variáveis, os efeitos não-lineares não estarão representados no valor de correlação. Esta omissão resulta em uma subestimação da força real da relação. É sempre prudente examinar todas as relações para identificar eventuais desvios da linearidade que podem afetar a correlação"(Hair et al, 2009: 85).

Em um modelo bivariado, uma forma simples de observar a relação entre X_1 e Y é através de um gráfico de dispersão. Na estimação do modelo, a linearidade implica que o aumento de uma unidade em X_1 gera o mesmo efeito sobre Y, independente do valor inicial de X_1 (Wooldridge, 2009). Em uma relação não linear mesmo que exista uma associação entre as variáveis explicativas incluídas no modelo e o fenômeno de interesse do pesquisador, não será possível detectar essa relação utilizando o método dos mínimos quadrados ordinários. Em uma frase: a violação desse pressuposto impede que a estimação por mínimos quadrados ordinários produza o melhor estimador linear não-viesado (MELNV)[9].

O segundo pressuposto diz respeito à mensuração das variáveis. Para Tabachnick e Fidell (2007), "a análise de regressão assume que as variáveis são medidas sem erro, uma clara impossibilidade em muitas pesquisas nas ciências sociais e comportamentais"(Tabachnick e Fidell, 2007: 122). Neste sentido, tem-se problemas de confiabilidade e validade dos indicadores utilizados. De acordo com Lewis-Beck (1980), a importância de incluir variáveis bem medidas no modelo é evidente: variáveis mal medidas produzirão estimativas inconsistentes. Em particular, se as variáveis independentes são medidas com erro, as estimativas (intercepto e coeficiente de regressão) serão viesadas. Além disso, os testes de significância e o intervalo de confiança serão afetados. Caso apenas a variável dependente seja medida com erro, ainda existe chance do estimador ser não-viesado, assumindo que a distribuição do erro é aleatória. No entanto, é comum observar ineficiência no erro padrão da estimativa, reduzindo a consistência dos testes de significância[10].

[9]Caso o pesquisador identifique que a relação entre as variáveis de interesse é não linear ele pode tomar algumas medidas. A mais comum é transformar as variáveis. Um procedimento alternativo consiste na criação de novas variáveis para modelar a relação não linear. É possível também utilizar modelos não lineares. Dada a complexidade de operacionalização e também interpretação desses últimos, recomendamos a transformação de variáveis como procedimento padrão para produzir linearidade. Operacionalmente, a transformação de variáveis pode ser facilmente conduzida na maior parte dos pacotes estatísticos, sendo necessário apenas que o pesquisador identifique qual é o remédio mais adequado para cada situação. Para os propósitos desse artigo reportamos nos anexos desse trabalho algumas das transformações mais usualmente empregadas.

[10]Kennedy (2009) sugere três principais remédios para superar problemas de erro de mensuração: a) modelos de regressão generalizados; b) variáveis instrumentais e c) modelo de equações estruturais.

O terceiro pressuposto refere-se ao termo aleatório de erro (ϵ). A importância do valor médio do termo de erro ser igual a zero dado X significa que os fatores não incluídos no modelo (que compõem o termo de erro) não afetam sistematicamente o valor médio de Y (os pontos positivos e negativos se anulam por serem equidistantes). A violação desse pressuposto compromete a consistência da estimativa do intercepto. Dessa forma, enquanto o coeficiente de regressão (*slope*) não é afetado, o pesquisador deve ter cuidado com a interpretação substantiva da constante. Para Kennedy (2009), "o erro pode ter uma média diferente de zero devido a presença de erros de mensuração sistematicamente positivos ou negativos no cálculo da variável dependente"(Kennedy, 2009: 109).

A homocedasticidade é o quarto pressuposto, ou seja, homogeneidade da variância é um pressuposto central do modelo de regressão de mínimos quadrados ordinários[11]. Mas o que é homocedasticidade afinal? Os resíduos, ou seja, a diferença entre os resultados observados e os resultados preditos pelo modelo devem variar uniformemente. Se a medida que o valor de Y aumenta, os erros de predição também aumentam, tem-se heterogeneidade na variância, ou seja, tem heterocedasticidade (variância diferente). Fundamentalmente, a violação desse pressuposto é preocupante na medida em que afeta a confiabilidade dos testes de significância e intervalos de confiança. Para Lewis-Beck (1980), "violar a suposição da homocedasticidade é mais grave. Isso porque mesmo que as estimativas dos mínimos quadrados continuem a ser não-viesados, os testes de significância e intervalos de confiança estariam errados"(Lewis-Beck, 1980: 28). Antes de reportar os resultados, o pesquisador deve analisar o ajuste do modelo, identificando eventuais problemas de heterocedasticidade[12]. Isso porque

[11]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. Homocedasticidade é desejável porque a variância da variável dependente a ser explicada na relação de dependência não deve ser concentrada em apenas uma gama limitada de valores independentes"(Hair et al, 2009: 83).

[12]Uma forma de identificar a presença de heterocedasticidade é analisar a dispersão dos erros. Quanto mais aleatória for a distribuição, maior é a confiança do pesquisador em ter ajustado um modelo homocedástico. A observação de qualquer outro tipo de padrão é um indício de heterocedasticidade. Outra alternativa é analisar a distribuição da variável dependente a partir das categorias de uma determinada variável independente categórica utilizando o gráfico de *Box-plot*. É possível ainda utilizar o teste de homogeneidade de variâncias de Levene. Uma vez detectada heterocedasticidade, o pesquisador pode seguir as seguintes diretrizes para tentar superar esse problema: a) aumentar o número de casos e b) transformar as variáveis.

modelos de mínimos quadrados ordinários com distribuição heterocedástica do erro perdem a propriedade de melhor estimativa dos parâmetros populacionais. Para Tabachick e Fidell (2007), a presença de erros de mensuração nas variáveis independentes é uma das causas de heterocedasticidade.

A quinta premissa é a da ausência de autocorrelação entre os casos, que se refere à situação em que o valor de uma observação medida em determinado período (t_1) não influencia o valor de uma observação medida em um momento posterior (t_2). Significa dizer que as observações são independentes, ou seja, que não existe correlação entre os termos de erro. Enquanto os valores dos coeficientes permanecem não-viesados, tem-se problemas na confiabilidade dos testes de significância e intervalos de confiança[13].

O sexto diz respeito à correlação entre as variáveis independentes e o termo de erro. Para Lewis-Beck (1980) é difícil satisfazer esse pressuposto em desenhos de pesquisa não experimentais. Como o pesquisador não pode manipular o valor da variável independente, é importante que todas as variáveis teoricamente importantes sejam incorporadas ao modelo explicativo. Se, por exemplo, uma variável X_1 está correlacionada com outra variável explicativa X_2, mas o pesquisador não incluir está última em seu modelo, as estimativas serão viesadas.

A sétima recomendação diz respeito à especificação adequada do modelo. Aqui deve-se observar dois procedimentos. Primeiro, todas as variáveis independentes teoricamente relevantes devem ser incluídas na equação de regressão. Segundo, nenhuma variável teoricamente irrelevante deve ser incluída no modelo já que isso produz ineficiência nos estimadores, aumentando o erro padrão da estimativa. Em conformidade com o pressuposto 2 (ausência de erros de mensuração), a correta especificação do modelo é central para produzir estimativas

[13]De acordo com Garson (2011), o pesquisador pode utilizar o teste de Durbin-Watson, d, para detectar a presença de autocorrelação em seus dados. A estatística d varia entre 0 e 4 de tal modo de quanto mais perto de 0 maior é a autocorrelação positiva e quanto mais perto de 4 maior é a autocorrelação negativa. Valores entre 1,5 e 2,5 sugerem independência das observações.

não-viesadas.

O oitavo pressuposto refere-se à multicolinearidade. Kennedy (2009) argumenta que "o estimador OLS na presença de multicolinearidade permanece não viesado e, de fato, ainda é o melhor estimador linear não viesado (BLUE) (...) na verdade, uma vez que todos os pressupostos da CLR (*Classical Linear Regression*) continuam a ser observados (estritamente falando, claro), o estimador MQO mantém todas as suas propriedades desejáveis"(Kennedy, 2009: 193). A maior dificuldade de modelos com problemas de multicolinearidade é o aumento da magnitude da variância dos parâmetros estimados. Isso porque a presença de altos níveis de correlação entre as variáveis independentes impossibilita estimar, com precisão, o efeito de cada variável sobre a variável dependente, no caso, o fenômeno que o pesquisador procura entender/explicar/predizer[14].

Por exemplo, suponha que o modelo explicativo utiliza duas variáveis altamente correlacionadas, X_1 e X_2, para explicar a variação de uma variável dependente qualquer, Y. A variação total é formada pela variação associada a X_1 mais a variação associada a X_2 mais a variação comum entre X_1 e X_2. O modelo de regressão de mínimos quadrados ordinários utiliza apenas a variação única de cada variável para estimar os coeficientes, ignorando a variância compartilhada. Eis a essência do problema: quanto maior a correlação entre as variáveis independentes, menos informação estará disponível para estimar os coeficientes associados às variáveis explicativas. Para Kennedy (2009), "qualquer estimativa baseada em pouca informação não pode ser realizada com muita confiança - ela terá uma alta

[14]Para Garson (2011), "multicolinearidade refere-se à correlação excessiva entre as variáveis preditoras. Quando a correlação é excessiva (alguns usam a regra de ouro de r $\geq$ 0,90), os erros padrão dos coeficientes de b e beta se tornam grandes, tornando difícil ou impossível avaliar a importância relativa das variáveis preditoras. Multicolinearidade é menos importante quando a finalidade da pesquisa é a predição já que os valores preditos da variável dependente permanecem estáveis, mas a multicolinearidade é um problema grave quando a finalidade da pesquisa inclui a modelagem causal"(Garson, 2011). Tecnicamente, o pesquisador pode analisar o *Variance Inflation Factor (VIF)* para verificar em que medida suas variáveis independentes apresentam problemas de multicolinariedade. Quanto maior, pior. A raiz quadrada do *VIF* de uma determinada variável independente informa ao pesquisador o aumento esperado no erro padrão do coeficiente da variável em comparação ao coeficiente esperado na ausência de multicolinearidade.

variância"(Kennedy, 2009: 194)[15]. A figura 2 abaixo ilustra esse argumento[16].

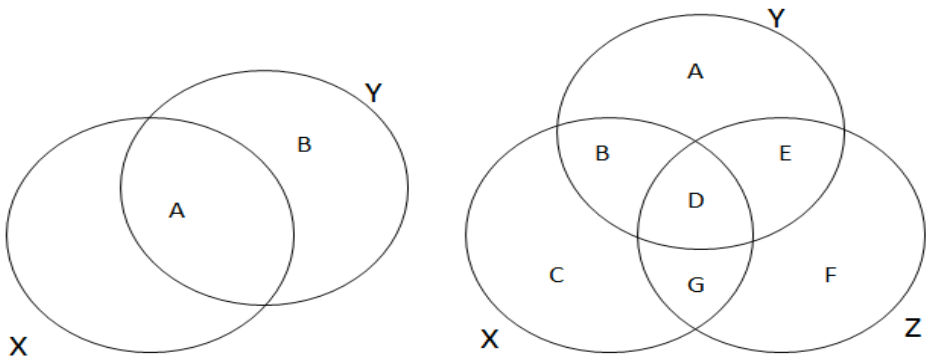

Figura 2: Multicolinearidade utilizando diagrama de Vein

No primeiro modelo (figura da esquerda) tem-se duas variáveis. A área comum entre X e Y está ilustrada pela letra A. Por sua vez, B representa a variação de Y que independe da variação em X, no caso, explicada pelo termo de erro. No segundo modelo tem-se duas variáveis independentes (X e Z) e a mesma variável dependente, Y. Como pode ser observado, existe uma correlação entre as variáveis independentes representada pela área D + G. 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

[15]Mas o que o pesquisador pode fazer para minimizar problemas de multicolineariedade? A literatura sugere várias medidas. Por exemplo, a incorporação de mais informação, ou seja, aumentar o número de observações. Além disso, o pesquisador pode se certificar de que não existem problemas de variáveis omitidas, averiguar se a forma funcional do modelo foi devidamente estipulada, identificar a presença de *outliers* e assegurar que as variáveis foram medidas de maneira adequada. Uma saída adicional é utilizar a análise de componentes principais ou a análise fatorial para criar uma medida síntese a partir da variância das variáveis originais. O pesquisador não deve excluir uma das variáveis independentes sob pena de produzir erros de especificação do modelo, a não ser que a correlação entre a variável excluída e as demais variáveis independentes seja zero. Tecnicamente, uma forma de detectar problemas de multicolineariedade é estimar uma correlação entre as variáveis independentes de tal modo que coeficientes próximos ou superiores a 0,9 indicam a presença de multicolinearidade. Outra regra importante é verificar se o R^2 da regressão com a variável dependente é menor do que o R^2 estimado a partir de uma variável independente pela outra. Em caso afirmativo, tem-se problemas de multicolinearidade. Seguindo os ensinamentos de Goldberger (1989), entendemos que muitas vezes os problemas de multicolinearidade estão associados a "micronumerosidade", ou seja, amostras pequenas. Como regra geral, portanto, sugerimos que o pesquisador evite, sempre que possível, a micronumerosidade e com ela toda a sorte de problemas associadas a estimações com N reduzido. As diferentes técnicas de *Fuzzy set/Qualitative Comparative Analysis (QCA)* são adequadas para trabalhar com amostras pequenas e intermediárias.

[16]Para uma introdução a utilização do digrama de Vein ver em especial Ip (2001) e Kennedy (2002).

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. Isso quer dizer que ao se estimar β_x apenas a área B seria utilizada, e ao se estimar β_z apenas a área E seria utilizada. Ou seja, toda a informação da área D seria perdida (área comum entre X e Z). Kennedy (2009) explica que essa informação não é utilizada "porque reflete a variação em Y que é determinada pela variação em ambos X e Z, as contribuições relativas dos quais não são conhecidas a priori"(Kennedy, 2009: 46). Portanto, quanto maior for a correlação entre as variáveis independentes (multicolinearidade), menos informação estará disponível para calcular as estimativas dos coeficientes. No limite, na existência de multicolinearidade perfeita as áreas B e E desaparecem, impossibilitando a estimação[17].

O nono pressuposto refere-se à distribuição do termo de erro. De acordo com as premissas do teorema de Gauss-Markov, o erro amostral deve seguir uma distribuição aproximadamente normal para que os estimadores de β_1, β_2 e σ (sigma) encontrados a partir do método de mínimos quadrados ordinários sejam não-viesados e eficientes.

Por fim, deve-se observar a proporção mínima entre o número de caso e de parâmetros. O número de casos deve exceder a quantidade de parâmetros estimados. Essa é uma condição matemática básica. Como o algoritmo computacional inverte a matriz para realizar os cálculos, caso o número de parâmetros a serem estimados supere a quantidade de observações, a estimação torna-se matematicamente impossível. O pesquisador deve maximizar o número de observações em sua análise dada as propriedades desejáveis de amostras grandes. Isso porque a partir do Teorema Central do Limite (*Central Limit Theorem*) sabe-se que a distribuição amostral de variáveis aleatórias converge para a distribuição normal quando o tamanho da amostra aumenta.

[17]Para Kennedy (2009), "além de criar altas variações nas estimativas dos coeficientes, a multicolinearidade está associada a problemas indesejáveis nos cálculos com base na matriz de dados que sejam instáveis, ou seja, nos quais pequenas variações na matriz de dados, tais como a adição ou supressão de uma única observação, pode levar a grandes mudanças nas estimativas dos parâmetros"(Kennedy, 2009: 198/199).

4 O planejamento de uma análise de regressão

O quadro abaixo sumariza o planejamento de um desenho de pesquisa em cinco estágios.

Estágio	**Procedimento**
1	Definir o problema de pesquisa, selecionar a variável dependente (VD) e identificar as variáveis independentes (VIs), ou seja, proceder a especificação do modelo. Aqui o pesquisador deve definir qual é a relação esperada entre VD e VIs.
2	Maximizar o número de observações no sentido de aumentar o poder estatístico (*statistical power*), a capacidade de generalização e reduzir toda sorte de problemas associados a estimação de parâmetros populacionais a partir de dados amostrais com N reduzido.
3	Verificar em que medida os dados disponíveis satisfazem os pressupostos da análise de regressão de mínimos quadrados ordinários (ver seção anterior). Como procedimento padrão, o pesquisador deve reportar as técnicas utilizadas para corrigir eventuais violações (transformações, recodificações, aumento de N, etc.).
4	Estimar o modelo
5	Interpretar os resultados

Metodologicamente, é importante apresentar, de forma clara e objetiva, qual é o problema de pesquisa que o pesquisador está interessado em investigar. Depois disso, deve-se observar o nível de mensuração da variável dependente. Isso porque a análise de regressão de mínimos quadrados ordinários requer que a variável dependente seja quantitativa, discreta ou contínua[18]. Por fim, o pesquisador deve identificar as variáveis independentes, especificando o modelo. Ele deve definir qual é a relação esperada entre a variável dependente (VD) e as variáveis independentes (VIs).

No que diz respeito ao segundo estágio, como regra geral, é importante garantir a maior quantidade possível de observações. Estimativas oriundas de amostras pequenas são instáveis, podem apresentar problemas com os graus de liberdade do modelo e apenas relações ex-

[18]O modelo requer variáveis discretas ou contínuas, mas alguns tipos dessas variáveis podem não ter o tratamento mais adequando com o modelo de Mínimos Quadrados Ordinários. Esse é o caso de variáveis censuradas e variáveis de contagem. Para esses casos modelos específicos (por exemplo, *Probit* ou *Tobit*) oferecem melhores resultados.

tremamente fortes serão detectadas. Por outro lado, quanto maior o tamanho da amostra, maior é chance de detectar a existência de uma relação entre as variáveis, independente de sua magnitude. Em relação ao tamanho da amostra, Hair et al (2009) sugerem que a razão entre o número de casos e o número de variáveis independentes nunca deve ser inferior a cinco, ou seja, para cada variável independente, o pesquisador deve ter, ao menos, cinco casos disponíveis. Tabachnick e Fidell (2007) sugerem utilizar $N \geq 50 + 80X$ (em que X representa o número de variáveis independentes incluídas na análise). Stevens (1996) recomenda uma proporção de 15 observações por variável para produzir estimativas confiáveis. Nossa recomendação é que o pesquisador utilize a maior proporção de observações por variável possível, e em casos em que precise trabalhar com o mínimo, o indicado é referenciar na literatura especializada e ser ortodoxo quanto aos pressupostos do modelo.

Em relação ao terceiro estágio, o pesquisador deve verificar em que medida os dados disponíveis satisfazem os pressupostos da análise de regressão de mínimos quadrados ordinários (ver seção anterior). Como procedimento padrão, o pesquisador deve reportar as técnicas utilizadas para corrigir eventuais violações (transformações, recodificações, aumento de N, etc). Essa fase é central para garantir a confiabilidade do trabalho, quer seja possibilitando a replicação, quer seja assegurando a avaliação crítica da consistência dos resultados. A transparência na coleta, no tratamento e na análise dos dados são características desejáveis de qualquer trabalho acadêmico. Nas palavras de King, Keohane e Verba (1994), "nossa primeira e mais importante diretriz para melhorar a qualidade dos dados é: registrar e relatar o processo pelo qual os dados são gerados. Sem essa informação não podemos determinar se mesmo utilizando procedimentos padrão na análise dos dados não estamos produzindo inferências viesadas"(King, Keohane e Verba, 1994: 23)[19].

[19]Em outro momento os autores afirmam que "Se o método e a lógica de observação dos pesquisadores e suas inferências são deixados implícitas ou obscuras, a comunidade acadêmica não tem como julgar a validade do que foi feito (...) não podemos aprender com os seus métodos ou replicar os seus resultados. Essa investigação não é um ato público. Ou, ainda, não faz uma boa leitura e portanto não é uma contribuição à ciência social"(King, Keohane e Verba, 1994: 08).

Após checar os pressupostos o próximo passo é estimar o modelo. Nessa fase é importante que as estatísticas de interesse sejam devidamente reportadas (erro padrão da estimativa, R^2, R^2 ajustado, teste F, níveis de significância, intervalos de confiança, etc). Como cada área do conhecimento tende a enfatizar determinadas formas de reportar os dados, é desejável que o pesquisador adote os padrões consolidados nos principais periódicos de seus respectivos ramos do conhecimento[20].

Por fim, depois de reportar as estatísticas de interesse o pesquisador deve interpretá-las. Não basta citar a magnitude dos coeficientes, é necessário discutir o tamanho do efeito a luz da teoria existente sobre o assunto. Similarmente, não basta mencionar o nível de significância de uma determinada relação, é necessário observar o peso explicativo dela a partir da literatura especializada sobre o tema. Em uma frase: é importante que o pesquisador deixe claro como as estatísticas estimadas se relacionam com a sua hipótese de pesquisa, discutindo os resultados empíricos de forma substantiva.

5 Simulando o uso do modelo linear de mínimos quadrados ordinários[21]

Para ilustrar como o modelo de regressão de mínimos quadrados ordinários pode ser utilizado em um desenho de pesquisa envolvendo temas relevantes para cientistas políticos, optamos por produzir uma simulação capaz de demonstrar não só as potencialidades da referida técnica, mas também suas limitações. Embora a utilização de simulações ainda seja limitada no Brasil, acreditamos que a técnica de simulação é a melhor alternativa metodológica para

[20]Em termos estritamente gráficos, replicamos aqui a sugestão de King (1995): o pesquisador deve evitar gráficos carregados (*too much information*), adotando tons de cinza e branco na elaboração de gráficos e tabelas.

[21]Todos os códigos apresentados aqui também estão disponíveis on-line. Além disso, os arquivos com os dados simulados também está disponível em `http://dvn.iq.harvard.edu/dvn/dv/felipenunes`. A partir da publicação deste texto, os comandos e os dados devem ser usados para fins acadêmicos.

demonstrar a aplicabilidade do modelo de regressão linear. Isso porque o pesquisador pode controlar os parâmetros que dão origem aos dados analisados. Dessa forma, é possível averiguar os resultados obtidos através do modelo linear, comparando-os com os valores utilizados para a criação dos dados. Em outras palavras, a simulação permite definir valores que representam a 'verdadeira' relação entre as variáveis, permitindo avaliar quão bem nosso modelo captura tal realidade.

A simulação utilizada neste texto tem dois propósitos. Primeiro, ela serve para ilustrar o mecanismo pelo qual um banco de dados é produzido. Com isso, pretendemos ressaltar que o modelo de regressão linear nada mais é do que uma estimação baseada no pressuposto de que há uma combinação linear de vetores (ou variáveis) presente nos dados. Tal combinação é ponderada pela multiplicação de coeficientes que expressam a relação linear entre cada vetor X e o vetor Y. Além disso, a especificação do mecanismo para produção dos dados também permite definir e controlar as distribuições pelas quais as variáveis serão construídas, garantindo assim que os pressupostos sejam devidamente satisfeitos. Nosso segundo objetivo com o uso de simulações é apresentar ao leitor as particularidades que um banco de dados 'real' pode conter, e que tendem a distorcer (para o bem e para mau) os resultados obtidos usando o método de mínimos quadrados. Para explorar exaustivamente essa dimensão, usaremos gráficos que representam visualmente os problemas discutidos nas seções anteriores. Embora nem sempre seja possível usar gráficos dessa natureza, é consenso na literatura aplicada que o simples uso de tais ferramentas poderia evitar a maioria dos erros em análise de dados (Tufte, 1990; Gelman, 2004)[22].

Por fim, toda a parte empírica do trabalho foi elaborada usando o programa R 2.13[23]. Além de permitir a simples elaboração de simulações, o R também contém pacotes gratuitos

[22]Como não se trata de um texto com pretensões explicativas, não seguiremos os padrões dos textos acadêmicos empíricos. Ao invés de focar na perguntar de pesquisa, nas motivações que nos levaram a ela, e nas hipóteses derivadas das implicações teóricas formuladas, nosso texto abordará as questões técnicas relevantes para o uso adequado da regressão linear de mínimos quadrados ordinários.

[23]O R e seus respectivos pacotes podem ser instalados gratuitamente a partir de `http://www.r-project.org/`

disponíveis na internet que possibilitam a produção de análises quantitativas de forma fácil e com excelente apresentação. Os autores recomendam fortemente a adoção de tal programa nos cursos de análise de dados espalhados no Brasil. Para além dos benefícios óbvios relativos aos custos de obtenção e manutenção do programa, o R é também uma poderosa ferramenta para análise de dados já que permite a interação completa do pesquisador com os resultados produzidos[24].

Para o nosso exemplo, o banco de dados apresenta 200 casos e quatro variáveis: x_1, x_2, x_3 e y. Todas elas foram construídas a partir de uma distribuição normal padronizada, ou seja, com média 0 e desvio padrão 1. As três primeiras (x_1, x_2, x_3) cumprirão o papel de variáveis independentes, tendo y como variável dependente, ou seja, a variável cuja variação se pretende entender/explicar/predizer. Além dessa configuração geral, as variáveis x_2 e x_3 foram construídas como funções de x_1. Em particular, x_2 é contínua e apresenta uma baixa correlação positiva com x_1 ($\rho = 0,3$). Por sua vez, x_3 é uma variável categórica dicotômica (*dummy*), assumindo valores 0 ou 1 fixados a partir da sua média. Ou seja, em x_3 valores acima da média recebem atributo 1, enquanto que os outros recebem 0, lembrando que x_3 também apresenta correlação de 0,25 com x_1. As três variáveis independentes combinadas formam y, sendo que tal combinação linear é ponderada pelos seguintes parâmetros: $\alpha = 1,\ \beta_1 = 4,\ \beta_2 = 0,5\ e\ \beta_3 = -2$.

O primeiro passo do pesquisador deve ser identificar as variáveis. Medidas de tendência central e de dispersão devem ser produzidas com o objetivo de conhecer as distribuições das variáveis, é o que a literatura denomina de analise exploratória dos dados (*Exploratory Data Analysis - EDA*). Em termos gráficos, sugerimos a utilização de histogramas e Box-plots. Visualizar os dados é uma tarefa que toma tempo e requer paciência, mas é um procedimento

[24]Além disso, a comunidade que utiliza R tem a possibilidade de atualizar as ferramentas já disponíveis e adicionar novas utilidades, contribuindo assim para o aprimoramento do programa e o contínuo acúmulo de conhecimento aplicado. Gostaríamos de lembrar que o professor Jackson Alves Aquino da Universidade Federal do Ceará (UFCE) vem trabalhando pioneiramente em um manual do R em português aplicado às ciências sociais.

fundamental para garantir que o pesquisador domine o banco de dados, explore a relação entre suas variáveis, e corrija eventuais erros de digitação, etc. Neste momento, também é recomendável que o pesquisador tenha atenção aos pontos que estejam destoando do restante da distribuição, os chamados *outliers*. Na tabela 1 abaixo apresentamos a estatística descritiva das variáveis.

Tabela 1: Estatísticas Descritivas para as Variáveis Usadas no Modelo de Regressão

Variável	N	Média	Mediana	Desvio-padrão	Q_1	Q_2	Min	Max	Omissos
y	200	0,86	0,72	6,23	-2,95	5,41	-18,91	14,97	0
x_1	200	0,06	0,12	0,92	-0,50	0,70	-2,30	2,18	0
x_2	200	-0,13	-0,24	0,98	-0,76	0,54	-3,02	2,80	0
x_3	200	0.53	1,00	0,50	0,00	1,00	0,00	1,00	0

O pesquisador deve sempre reportar o número de casos analisados, descrevendo a origem da amostra e indicando eventuais casos omissos (*missing cases*). No nosso exemplo, a amostra não tem casos omissos, totalizando, portanto 200 observações. Partindo do pressuposto de que uma boa explicação é precedida por uma boa descrição, o primeiro passo é caracterizar cada uma das variáveis de forma que o leitor entenda: (1) como cada medida foi construída e (2) qual é o significado substantivo dos valores observados. Isso é importante para que a interpretação dos coeficientes de regressão seja feita de forma adequada, já que eles são expressos em termos das unidades de medida das variáveis utilizadas. Tecnicamente, a análise gráfica é o procedimento mais eficiente para visualizar a distribuição das variáveis, como demonstram as figuras 3 e 4 abaixo.

O gráfico de Box-plot ilustra as distribuições de y, x_1 e x_2. Dois procedimentos devem ser observados. Primeiro, deve-se usar a mesma escala para comparar as distribuições. O mais adequado é fixar a escala da variável que tem a maior distribuição e plotar as demais seguindo tal medida. Segundo, o pesquisador deve observar a presença de pontos destoantes (*outliers*). No nosso exemplo, os box-plot revelam a presença de dois *outliers* em y e um

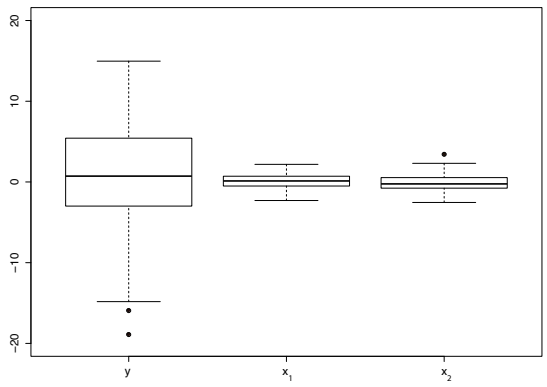

Figura 3: Box-plot com a distribuição das variáveis y, x_1 e x_2

ponto destoante em x_2. É comum, por exemplo, que um caso (*outlier*) altere radicalmente o ajuste da reta de regressão, 'enviesando' as estimativas. Nesses casos, é recomendável que transformações do tipo *Box-cox* sejam implementadas para tornar as distribuições mais bem comportadas. A figura abaixo ilustra a distribuição das variáveis a partir de histogramas.

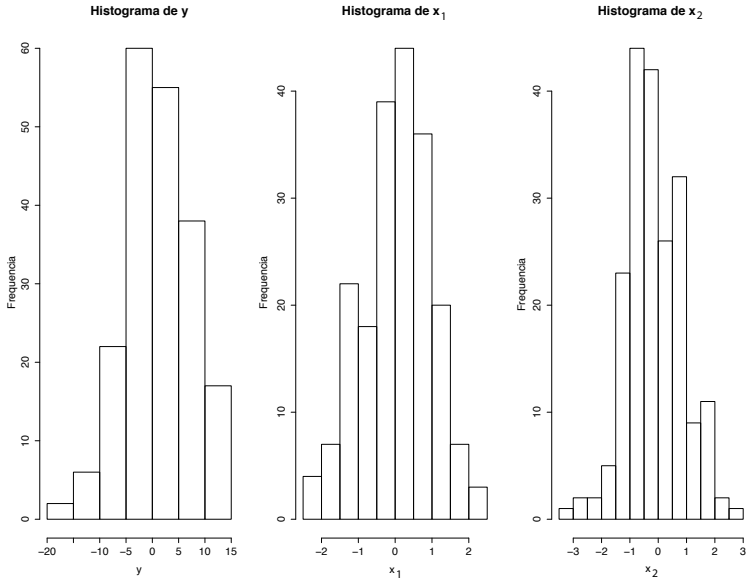

Figura 4: Histogramas com a distribuição das variáveis y, x_1 e x_2

Como os histogramas ilustram a distribuição real dos dados, eles são de fundamental importância para que eventuais correções e cuidados sejam devidamente tomados. Para distribuições assimétricas, por exemplo, é possível utilizar transformações logarítmicas. Outra

recomendação é conduzir uma análise sistemática sobre os casos destoantes da amostra com o objetivo de desvendar os mecanismos causais que explicam a sua posição de *outlier* (Geddes, 1994).

Após o reconhecimento dos dados, o próximo passo é definir o modelo que melhor descreve a relação entre as variáveis tendo em vista suas respectivas distribuições formas. Como este artigo trata especificamente da forma funcional de mínimos quadrados ordinários (MQO), nosso foco de agora em diante volta-se inteiramente para ele[25]. Em nosso exemplo a variável dependente y poderia ser o 'número de votos recebidos por determinado candidato a deputado federal'. Quanto maior o seu valor, maiores as chances que o candidato vença a eleição. As variáveis independentes incluídas no modelo são x_1, x_2 e x_3, sendo que elas poderiam representar, por exemplo, 'volume de dinheiro gasto na campanha', a 'popularidade do candidato dentro do partido', e 'se o candidato está concorrendo pela primeira vez (1) ou não (0)'. As duas primeiras foram criadas como variáveis contínuas e a última como categórica dicotômica (*dummy*).

O próximo passo é rodar o modelo de regressão e interpretar os seus resultados. Um ponto merece especial atenção neste momento: todas as variáveis independentes devem ser incluídas no modelo de regressão ao mesmo tempo[26]. Isso porque nosso objetivo é capturar as variações em y e x_1, controlando pela variação de x_2 e x_3. Esquematicamente, o modelo geral estimado

[25]Gostaríamos de ressaltar, no entanto, que há outras formas funcionais que podem ser usadas na estimação de um modelo de regressão. Os chamados modelos de mínimos quadrados generalizados são bons exemplos de técnicas que devem ser implementadas quando se trabalha com eventos raros, ou distribuições assimétricas. Com isso queremos chamar atenção para a importância do pesquisador e das decisões tomadas por ele. O apropriado uso de técnicas estatísticas é de responsabilidade do investigador e requer conhecimento dos limites e potenciais de cada ferramenta.

[26]Como dados observacionais não nos permitem observar a operação de tratamentos aleatórios, é impossível tirar conclusões causais usando o desenho de pesquisa discutido aqui. O melhor que nós podemos realizar é incluir as variáveis de controle relevantes no modelo de forma a capturar a relação explicitada acima. Quando a idéia do controle foi introduzida nos modelos de regressão múltipla o objetivo era exatamente emular um experimento usando dados observacionais. No entanto, pesquisas mais recentes demonstram a incapacidade da referida ferramenta e apontam novas estratégias como o uso de variáveis instrumentais, modelos de matching e de regressão em descontinuidade (ver Angrist e Pischke, 2009).

é o seguinte[27]:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \ + \beta_3 x_3 + \epsilon$$

sendo que α representa o ponto de origem da curva de regressão (constante), β_1 representa o coeficiente associado a variável x_1, β_2 representa o coeficiente associado a variável x_2, e β_3 representa o coeficiente associado a variável x_3.

6 A análise dos resultados

Tabela 2: Resumo das estatísticas do modelo final ajustado

r^2	r^2 ajustado	Erro padrão da estimativa
0,35	0,34	5,05

A primeira estatística de interesse é o coeficiente de determinação (r^2) que é uma medida de aderência dos dados em torno da reta de regressão e é usualmente interpretada como a proporção da variância na variável dependente explicada pela variação das variáveis independentes, ou seja, a qualidade do ajuste do modelo aos dados. O r^2 ajustado é uma medida similar ao r^2, mas que controla pelo número de observações e variáveis incluídas no modelo[28]. Quanto maior o tamanho da amostra, menor será a diferença entre essas duas estimativas. O r^2 ajustado sugere que 34% da variação na variável dependente pode ser explicada pela variação nas demais variáveis independentes. Por fim, o erro padrão residual (erro padrão da estimativa) é uma medida do grau com que a média da amostra se desvia da média das possíveis médias amostrais. Quanto menor, melhor já que ele representa a estimativa do efeito que o erro exerce sobre o ajuste geral modelo.

[27]Outras especificações serão testadas, como se pode ver na Tabela 4 abaixo. Mas esse é o modelo que nós acreditamos se ajustar melhor aos dados que nós temos.

[28]Para Hair et al (2009), o coeficiente de determinação ajustado é uma "versão modificada do coeficiente de determinação que leva em conta o número de variáveis independentes e o tamanho da amostra. Embora a adição de variáveis independentes irá sempre fazer com que o coeficiente de determinação suba, o coeficiente de determinação ajustado pode cair se as variáveis independentes acrescentadas têm pouco poder explicativo tornando o grau de liberdade demasiado pequeno. Esta estatística é bastante útil para a comparação entre equações com diferentes números de variáveis independentes, diferentes tamanhos de amostra, ou ambos"(Hair et al, 2009: 170).

Há grande controvérsia atualmente quanto à utilidade do coeficiente de determinação. O consenso na disciplina é de que não se pode avaliar a capacidade explicativa de um modelo de regressão a partir do r^2. O foco da análise é na magnitude dos coeficientes, e não na produção de um r^2 maior (King, 1986). No entanto, o tamanho do r^2 pode servir como um indicador para avaliar em que medida a relação entre as variáveis pode ser descrita por uma função linear. No nosso exemplo o r^2 não é alto o que poderia sugerir um ajuste inadequado entre o modelo estimado e os dados observados. Contudo, os dados foram simulados para atender a todos os pressupostos do método de MQO, sendo que as variáveis foram combinadas linearmente, e o mais importante, os coeficientes estimados correspondem aos valores estipulados na simulação. É esse tipo de exemplo que nos mostra a importância de focar na interpretação dos coeficientes e não na produção de r^2. A tabela 3 a seguir apresenta a análise de variância (ANOVA).

Tabela 3: Análise de variância (ANOVA) para modelo final

Modelo	Soma dos Quadrados	GL	Média dos Quadrados	F	P-valor
Regressão	7717,01	3	2527,34	35,39	0,000
Resíduos	5005,59	196	25,54		
Total	12722,60	199			

GL - Graus de Liberdade

A análise de variância (ANOVA) compara se o modelo estimado é melhor do que o modelo nulo (sem nenhuma variável independente). O teste avalia se algum dos coeficientes estimados (intercepto e coeficientes de regressão) é significativamente diferente de zero. Em termos técnicos, a estatística F é calculada a partir da divisão da média dos quadrados atribuída à regressão (2572,34) pela média dos quadrados dos resíduos (25,54). A probabilidade de que o resultado observado é proveniente de erro amostral pode ser examinada através do teste de significância (p-valor), assumindo que o modelo nulo é melhor do que o modelo estimado. No referido caso, a probabilidade de que o resultado observado esteja errado é muito pequena dado que p-valor é menor do que 0,000.

O próximo passo é analisar separadamente os coeficientes estimados. A forma tradicional de

interpretar resultados de regressão é através da leitura cuidadosa das estimativas. A tabela 4 abaixo sumariza essas informações.

Tabela 4: Regressão Linear Múltipla (MQO)

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo Final
Constante	0,65	0,73	1,63***	1,79*	1,87*
	(0,37)	(0,37)	(0,63)	(0,52)	(0,52)
x_1	3,81*	3,70*		3,99*	3,88*
	(0,39)	(0,41)		(0,40)	(0,40)
x_2		0,54	1,19**		0,54
		(0,38)	(0,44)		(0,37)
x_3			-1,16	-2,20**	-2,20**
			(0,87)	(0,73)	(0,72)
N	200	200	200	200	200
r^2 ajustado	0,31	0,31	0,03	0,32	0,34

Erros padrões entre parênteses

$***$ significância a $p < 0,05$

$**$ significância a $p < 0,01$

$*$ significância a $p < 0,001$

O primeiro passo é observar a correspondência entre o sinal dos coeficientes e a relação teoricamente esperada. Em outras palavras, em que medida os resultados oferecem evidências em favor das hipóteses de trabalho. Esse procedimento deve ser realizado a partir do modelo final, ou seja, aquele modelo que na opinião do pesquisador apresenta a especificação mais adequada (ver seção sobre os pressupostos). Nesse artigo nós testamos três hipóteses:

- **Hipótese 1**: x_1 exerce um efeito positivo sobre y
- **Hipótese 2**: x_2 exerce um efeito positivo sobre y
- **Hipótese 3**: x_3 exerce um efeito positivo sobre y

Conforme observado no modelo final, à exceção de x_3, todos os coeficientes apresentam a direção teoricamente esperada. Tais resultados sugerem a rejeição imediata da terceira hipótese. Enquanto a nossa teoria sugeria uma relação positiva entre x_3 e y, nossos resultados apontam na direção contraria. Nesse caso, o pesquisador deve ser capaz de revisar a

explicação apresentada, justificando o que poderia estar produzindo tal correlação inesperada. As motivações para a justificativa devem passar tanto por questionamentos teóricos, quanto metodológicos. O próximo passo é interpretar a magnitude dos coeficientes estimados.

O primeiro coeficiente a ser observado refere-se à constante do modelo (α) que representa o valor esperado da variável dependente quando todas as variáveis independentes assumem valor igual a zero. A constante do modelo final é de 1,87 ($p < 0,000$), ou seja, sobrestima o valor real do parâmetro ($\alpha = 1$). Em termos estritamente teóricos, os pesquisadores estão menos preocupados com a estimação consistente desse coeficiente. Isso porque em grande parte dos modelos estimados o valor da constante não tem uma interpretação substantiva[29]. Nesse caso, é recomendável suprimir esse valor da tabela apresentada. Quando a constante tiver uma interpretação inteligível, esse coeficiente deve ser devidamente reportado e suas implicações analisadas. No nosso exemplo quando x_1, x_2 e x_3 são iguais a zero espera-se que y seja, em média, igual a 1,87.

O próximo passo é analisar os coeficientes associados às variáveis independentes do modelo final. O β_1 (efeito de x_1 sobre y) assume valor 3,88, muito próximo do parâmetro especificado na simulação ($\beta_1 = 4$), o que quer dizer que o aumento de uma unidade em x_1 eleva em média em 3,88 o valor de y, mantendo tudo mais constante. Nossa conclusão é reforçada pela estimativa do erro associada ao ponto estimado. De acordo com os nossos resultados, a probabilidade de se estar errado ao rejeitar que β_1 é diferente de zero é muito pequena ($p < 0,000$). Logo, há fortes evidências sugerindo que o efeito positivo teoricamente esperado de x_1 sobre y pode ser corroborado.

[29]Por exemplo, ao se utilizar o Produto Interno Bruto (PIB) per capita (variável independente) para explicar o nível de democratização de um determinado país (variável dependente), o valor da constante expressa a média do grau de democratização quando o PIB per capita assume valor zero. Embora tecnicamente correta, essa interpretação não condiz com a realidade da distribuição do PIB. Portanto, essa leitura não nos auxilia na compreensão substantiva do modelo especificado. Um dos procedimentos mais comuns para evitar esse problema é centralizar as demais variáveis pela média.

Por sua vez, o β_2 (efeito de x_2 sobre y) assume valor 0,54, também muito próximo ao parâmetro da simulação ($\beta_2 = 0,5$). Nesse caso, mantendo as demais variáveis constantes, podemos dizer que o aumento de uma unidade em x_2 produz um efeito positivo de 0,54 em y. Embora o modelo linear de mínimos quadrados ordinários tenha sido capaz de estimar um coeficiente muito similar ao valor 'verdadeiro' da relação entre x_2 e y, e o resultado aponte para a direção esperada teoricamente, não é possível ter segurança quanto à rejeição da hipótese nula de que $\beta_2 = 0$. O que não significa que o efeito teoricamente esperado não exista, mas apenas que não fomos capazes de demonstrá-lo com segurança estatística usando o p-valor.

Resultados não significativos podem ser explicados por diversos motivos, exigindo, portanto, que o pesquisador justifique bem suas escolhas teórico-metodológicas e não suprima os resultados encontrados. Embora incomum em publicações científicas, a não rejeição da hipótese nula do coeficiente igual a zero também auxilia no debate disciplinar. Se um modelo estatístico é escolhido para descrever uma relação causal, é de se esperar apenas que os resultados convirjam para o valor verdadeiro do efeito quando a média dos efeitos de todos os trabalhos produzidos na área é calculada. Se os resultados inesperados não são reportados, o que se observa é um viés no nosso conhecimento sobre determinada relação causal. Esse problema é conhecido como viés de publicação. De acordo com Gerber, Green e Nickerson (2001), os artigos que não rejeitam suas hipóteses nulas tendem a não ser publicados. Nas palavras dos autores, 'esse fenômeno é conhecido como viés de publicação e representa uma tendência de pareceristas e pesquisadores em sobrestimar a importância da significância estatística nos achados da pesquisa'.

Por fim, o β_3 (efeito de x_3 sobre y) sugere a rejeição da nossa hipótese de trabalho. O coeficiente observado assume valor negativo, sendo que nossa hipótese esperava um efeito positivo. A interpretação do coeficiente indica que a cada unidade adicional de x_3 observa-se um efeito negativo de 2,20 em y ($\beta_3 = -2,20$). É possível concluir também que o efeito

estimado é diferente de zero, já que o resultado foi estatisticamente significativo ($p-valor < 0,05$), ou seja, nós temos 95% de confiabilidade que o coeficiente estimado é diferente de zero.

Mas qual é o principal problema em se utilizar os outros modelos especificados como referencia para avaliar as hipóteses de trabalho? Nos modelos 1, 2, 3 e 4 nós falhamos em atender a premissa da correta especificação do modelo linear de MQO. Dentre as principais consequências da violação desse pressuposto, listamos: (1) no modelo 1 há uma ligeira subestimação do efeito de x_1 sobre y, o que cresce no modelo 2 com a inclusão de uma variável irrelevante (x_2); (2) no terceiro modelo a omissão da principal variável independente (x_1) leva a incorreta conclusão de que x_2 exerce um efeito diferente de zero sobre y. Além disso, o pesquisador concluiria também incorretamente que o efeito de x_3 sobre y é igual a zero; e (3) o valor do efeito de x_1 é sobrestimado quando uma variável correlacionada é excluída do modelo[30]. Isso nos indica que parte da explicação atribuída a x_1 no modelo 1 devia-se a não inclusão de x_2 como variável de controle. Faz sentido observar esse resultado já x_1 e x_2 tem uma correlação de 0,3.

Para que a análise dos resultados fique completa é preciso analisar também a magnitude dos coeficientes, conferindo significado substantivo para cada um deles. Neste trabalho sugerimos o uso dos intervalos de confiança como a metodologia mais apropriada para tal prática. A estimativa de um intervalo de confiança nos permite identificar quais hipóteses nula não podem ser rejeitadas tendo como referência os dados e os modelos empregados. A vantagem do intervalo de confiança em relação ao uso do teste de significância (p-valor) para os pontos estimados, forma tradicional de testar hipóteses, é a possibilidade de se levar em conta o conjunto de alternativas que não podem ser rejeitadas. Ao invés de se considerar apenas a hipótese nula tradicional de que o coeficiente estimado não é diferente de zero, a inferência a

[30]Vale ressaltar que a omissão de uma variável independente não correlacionada com as demais variáveis independentes já incluídas no modelo não gera o problema aqui discutido. Embora a especificação ainda fique prejudicada, não haverá necessariamente um viés nos demais estimadores.

partir dos intervalos de confiança considera todas as outras hipóteses que se referem a valores dentro do intervalo de confiança. O gráfico 5 abaixo ilustra como intervalos de confiança podem ser utilizados para analisar as estimativas do resultado de regressão.

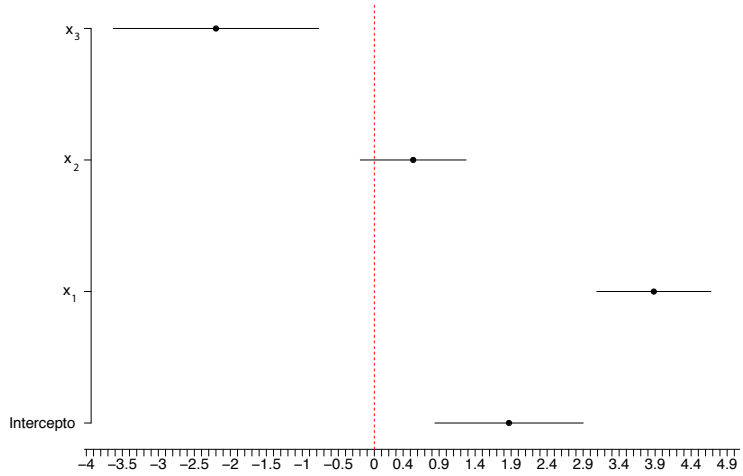

Figura 5: Pontos estimados e intervalos de confiança para variáveis analisadas

O eixo X representa a magnitude dos coeficientes e o eixo Y cada um dos coeficientes de interesse. Os segmentos na horizontal representam o intervalo de confiança de cada estimativa, ou seja, o grau de incerteza quanto aos valores que cada coeficiente pode assumir. O círculo preto por sua vez representa o valor médio para cada estimativa, nesse caso, $\alpha = 1,87,\ \beta_1 = 3,88,\ \beta_2 = 0,54,\ \beta_3 = -2,20$. Dois elementos precisam ser considerados para a correta interpretação desse gráfico. Primeiro, a distância entre as barras horizontais e a linha pontilhada vertical. Nesse caso, quanto maior for a distância, maior será a confiabilidade na rejeição das hipóteses nula de que os coeficientes β_1 e β_2 seriam iguais a zero. Segundo, é preciso prestar atenção à amplitude das barras horizontais. Isso porque elas nos apresentam uma medida da precisão dos coeficientes estimados. Quanto menores forem as barras, maior será a precisão da nossa estimativa, e portanto, maior a nossa confiança nas conclusões tiradas a partir dos mesmos.

No nosso exemplo, β_2 não foi estatisticamente significativo (ver tabela 3) porque um dos possíveis valores do seu intervalo de confiança é igual a zero. Isso significa que x_2 e y

podem apresentar uma correlação igual a zero. No entanto, o que a tabela 4 não mostra é que nós também não podemos rejeitar a hipótese nula de que β_2 seja igual 1,4 (a outra extremidade do intervalo de confiança). A vantagem dessa análise, portanto, é que saímos de uma interpretação pontual que pode rejeitar equivocadamente a importância substantiva de x_2 para uma análise mais detalhada, especificando o conjunto de hipóteses que devem ser consideradas, chamando atenção para a interpretação substantiva dos coeficientes e dos erros estimados. Nos casos de x_1 e x_3 fica visível a magnitude da distância entre as barras horizontais (intervalos de confiança) e a linha pontilhada vertical (representando o 0). Em consequência disso, nossa confiança na rejeição da hipótese nula é bem grande.

Quanto a precisão das nossas estimativas, é possível afirmar que embora o intervalo de confiança de β_2 cruze a linha referente ao 0, a maior parte de sua distribuição contém valores positivos, próximos ao valor 'verdadeiro' de β_2. β_1 também tem uma distribuição concentrada de seus valores, sendo todos positivos. β_3, ao contrário, embora distante de zero e totalmente distribuído na parte negativa do gráfico, pode ser caracterizado como uma estimativa imprecisa. isso porque β_3 pode assumir valores entre -4 e -0,7. O pesquisador não saberia afirmar se o efeito é negativo ou se praticamente não existe efeito de x_3 sobre y.

A figura 6 a seguir ilustra a distribuição dos valores estimados de y em função dos diferentes valores observados nas variáveis independentes. O gráfico A, por exemplo, é construído plotando os valores esperados de y para cada valor de x_1, fixando x_2 e x_3 em suas respectivas médias. O mesmo procedimento foi adotado para analisarx_2.

O gráfico A mostra os valores preditos de y quando x_1 varia de -2 a 2 (sua distribuição real - ver tabela 2), mantendo x_2 e x_3 fixados em suas respectivas médias. A partir dessa especificação é possível dizer que mantendo as outras variáveis constantes, um aumento de 1 uma unidade na variável x_1 produz uma elevação, em média, de 4 pontos em y (observe a área cinza no gráfico entre 0 e 1 na variável x_1, e entre 0,64 e 4,53 na variável y). O gráfico B, por sua vez, apresenta os valores esperados de y quando x_2 varia de -3 a 3, fixando x_1 e

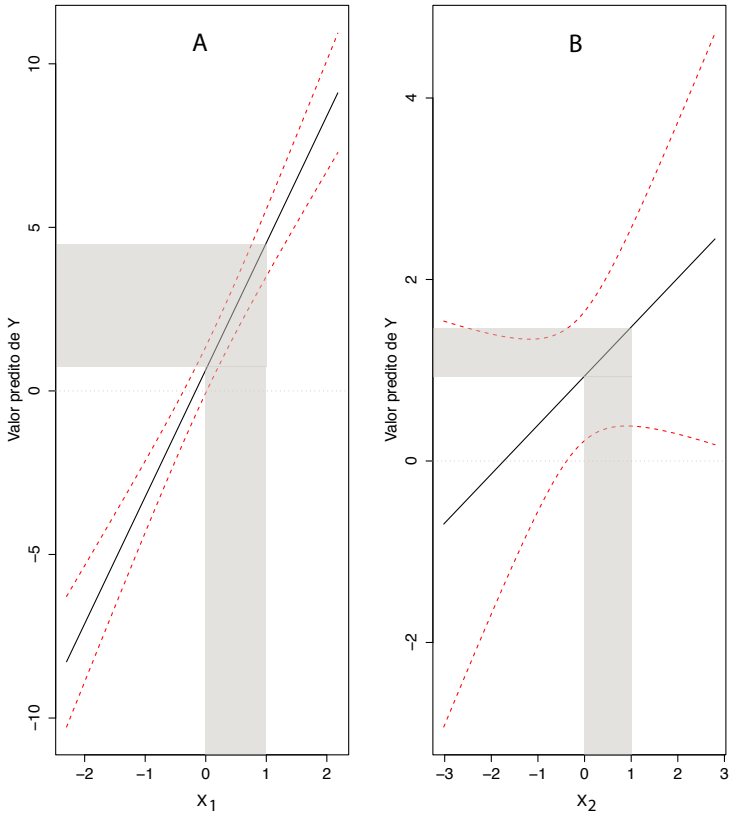

Figura 6: Valores preditos de y para valores fixos de x_1 e x_2

x_3 nas suas médias. Em termos práticos, aumentar x_2 de 0 a 1 produz um aumento em y de 0,94 a 1,48 (ver área cinza no gráfico B).

Imagine, por exemplo, que y representa o 'numero de votos recebidos por um determinado candidato a deputado federal' e x_1 representa o 'volume de dinheiro gasto na campanha'. A conclusão seria de que a cada R$ 1 adicional investido produz um aumento de 4 votos para o deputado. Digamos que os dados usados aqui fossem referentes às últimas eleições legislativas no Brasil. Um dos deputados menos votado no estado de São Paulo precisou de algo em torno de 20.000 votos para se eleger, isso significa que o candidato precisaria gastar cerca de R$ 5.000 para obter essa vaga. Esse resultado nos levaria a concluir que gasto em campanha é uma variável fundamental para se entender vitória eleitoral.

Mas poderia significar o contrário. Se estivéssemos estudando o impacto do 'número de mandatos de um deputado federal' sobre a 'quantidade de projetos de sua autoria aprovados (por legislatura)', e observássemos o mesmo efeito de x_1 sobre y (de 1 para 4), concluiríamos que a cada novo mandato o deputado teria um aumento de 4 projetos de sua autoria aprovados no Congresso. Levando-se em conta que os deputados aprovam em média mais de 50 projetos por mandato, seria preciso mais de 4 mandatos para que um deputado novato conseguisse se destacar no plenário da casa. O que, hipoteticamente, não expressaria um resultado substancialmente significante. Queremos com isso, chamar atenção para a importância de se conhecer as escalas das variáveis usadas e a variação real observada em cada uma delas.

Por fim, é preciso interpretar de forma substantiva o que significa observar um coeficiente $\beta_3 = -2,20$. Por se tratar de uma variável dicotômica (0 ou 1), o coeficiente de regressão informa o impacto esperado em y quando x_3 varia de uma categoria a outra. No nosso exemplo, y assume valor esperado de 2.02 quando $x_3 = 0$, mas cai para -0,18 quando $x_3 = 1$. O efeito negativo observado, além de estatisticamente significativo, também é substantivamente relevante.

Outra importante contribuição desse tipo de gráfico é a observação do intervalo de confiança para todos os valores de x_1 e x_2. Note que a incerteza quanto aos pontos estimados é muito menor para a primeira variável do que para a segunda. Observe ainda que o intervalo de confiança é menor nos locais da distribuição do eixo x_1 onde há maior frequência de casos. Isso porque existe uma correlação negativa entre a frequência de observações e o nível de incerteza das estimativas, ou seja, quanto maior o número de observações em um determinado valor da variável independente, menor o nível de incerteza, logo, mais precisa será a estimativa. Dado que as três variáveis são distribuídas normalmente é de se esperar que a incerteza expressa nos gráficos seja maior nas extremidades da distribuição. Essa observação reforça, portanto, a vantagem da análise dos intervalos de confiança em detrimento do teste de significância. Enquanto o primeiro fornece uma noção completa sobre nossas estimativas,

o segundo pode ser facilmente afetado aumentando o número de casos usados.

Por se tratar de uma simulação, os dados utilizados satisfazem todos os pressupostos do modelo de regressão de mínimos quadrados ordinários. No entanto, a realidade não nos é tão favorável quanto uma simulação computacional. O pesquisador deve avaliar em que medida os seus dados satisfazem esses pressupostos, tomando as medidas cabíveis em casos de violação. Para ilustrar como isso pode ser feito, a próxima seção discute algumas alternativas técnicas que auxiliam o pesquisador a superar eventuais problemas.

7 Os cuidados antes de se usar uma regressão linear

Uma tarefa fundamental para uma boa análise de dados é diagnosticar problemas em modelos de regressão. Esses diagnósticos informam em que medida os dados observados podem ser representados pelo modelo ajustado. No caso especifico da forma funcional de mínimos quadrados ordinários, a estrutura dos dados analisados precisa satisfazer uma serie de pressupostos para que os estimadores sejam consistentes. Quando tais pressupostos são violados, os estimadores de mínimos quadrados não fornecem, por exemplo, a melhor estimativa linear nao-viesada.

Nessa seção apresentamos os problemas mais comumente encontrados pelos pesquisadores ao ajustar seus modelos de regressão, sugerindo alternativas para como supera-los. Mais especificamente, trataremos dos seguintes obstáculos: (1) presença e influencia de *outliers*, (2) observação de resíduos com distribuição não normal, (3) erros com variância não constante (heterocedasticidade) e (4) multicolinearidade entre as variáveis independentes. Há diversas referencias tratando de cada um desses problemas de forma detalhada, dentre elas recomendamos Fox (2008)[31]. Ao contrário de outros pacotes estatísticos, o R nos possibilita

[31]Professor John Fox desenvolveu o principal pacote em R para produção simples de diagnósticos em regressão linear (*car*). Usaremos esse pacote nesta seção e disponibilizaremos os códigos no apêndice digital.

realizar diagnósticos de forma fácil e rápida, como será mostrado em seguida.

Comecemos nosso diagnóstico pelos valores fora da curva, os chamados *outliers*. Esses valores são assim denominados por apresentarem comportamento destoante do restante dos valores preditos. A estatística padrão para detectar *outliers* na regressão são os *'studentized residuals'* para o modelo (ver mais em Fox, 2004)[32]. Sugerimos o uso de um gráfico que compara se o modelo estimado consegue descrever bem a contribuição de todos os casos usando uma combinação linear (QQ plot padronizado). A lógica é comparar quão bem os pontos se distribuem sobre a reta continua. Quanto mais distantes os pontos, maior é a contribuição dos mesmos no enviesamento das estimativas. Tecnicamente, o gráfico mostra os quartis dos resíduos da regressão com os de uma distribuição padronizada ajustada. Esse gráfico plota resíduos padronizados contra seus respectivos quartis numa distribuição t com n-k-2 graus de liberdade, em que n representa o numero de casos e k a quantidade de parâmetros estimados. A figura 7 abaixo ilustra essas informações.

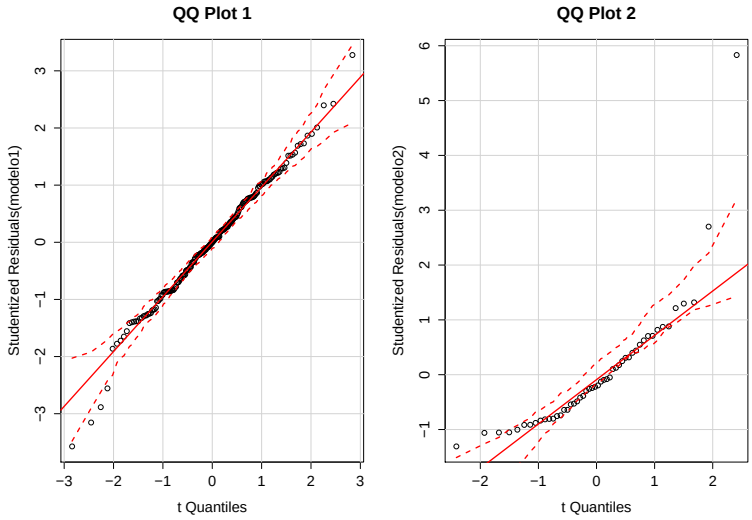

Figura 7: QQ-plots para resíduos padronizados numa distribuição t

O QQ-plot 1 apresenta uma distribuição sem *outliers*. Já no QQ-plot 2 observa-se a presença

[32]Recomendamos que leitores menos familizariados com noções básicas de estatística iniciem seus diagnósticos através de gráficos de *box-plot*, histogramas e de dispersão. Além disso, sugerimos a utilização de *QQ plots* das variáveis e testes de normalidade.

de pontos que estão literalmente fora da curva ajustada. Observe como os pontos no início e no final da distribuição do gráfico à direita se posicionam fora do intervalo de confiança (linha tracejada). No caso de se observar um gráfico como esse, talvez seja mais apropriado usar uma regressão robusta que corrija o efeito dos *outliers* (Fox, 2008).

Observações que estão relativamente distantes do centro da distribuição dos valores preditos, já considerando-se o padrão de correlação entre as variáveis independente, tem um grande potencial para influenciar os coeficientes da regressão de mínimos quadrados. Estes pontos são chamados de 'pontos de alavancagem' (*'high leverage'*). A forma mais simples de avaliar esse problema é plotando os gráficos de regressão parcial. A figura 8 abaixo ilustra esse procedimento.

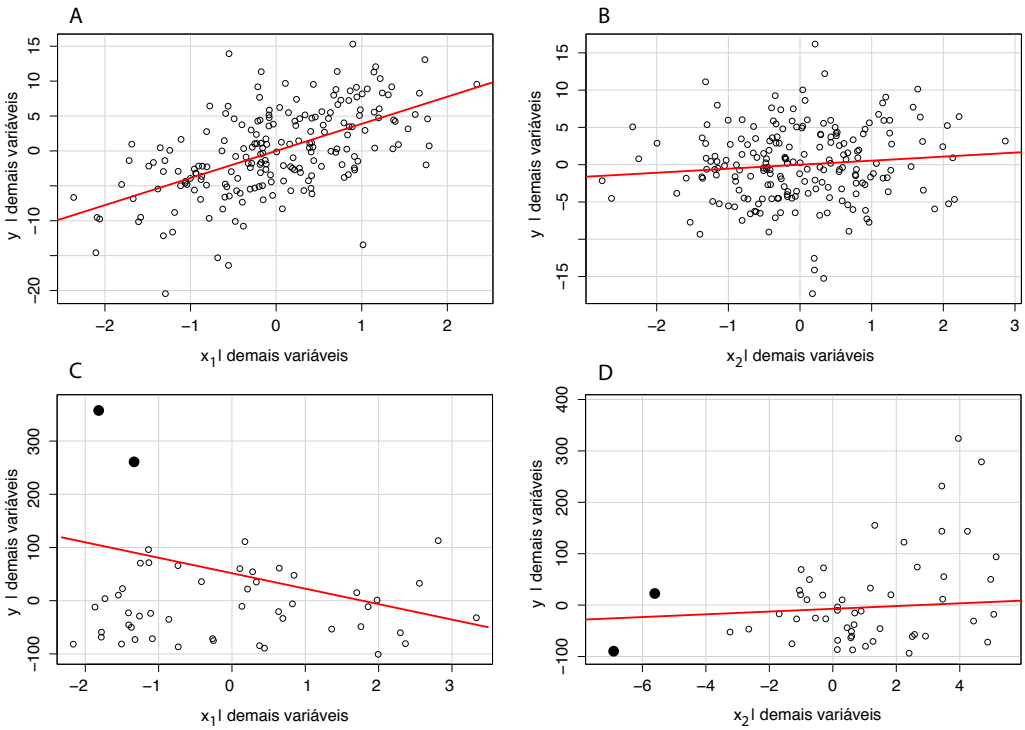

Figura 8: Gráficos de regressão parcial

Na primeira linha mostramos os gráficos para um banco de dados aparentemente normal, enquanto que na segunda apresentamos um caso em que pontos de alavancagem sérios problemas para o uso da regressão de MQO. A diferenciação aqui é também visual. É fácil

observar que nos dois primeiros gráficos os pontos se comportam muito bem e não há qualquer um deles que possa representar uma ameaça séria ao ajuste da reta. No segundo caso, ao contrário, é possível identificar alguns casos que funcionam como alavancas trazendo a reta de regressão ou mais para baixo (primeiro caso) ou mais para cima (segundo caso). Nessa situação, talvez fosse mais apropriado o uso de modelos não lineares ou corrigir a distribuição dessas variáveis usando transformações de *Box-Cox* (Weisberg, 2005).

Os mesmos problemas identificados acima podem ser observados usando uma medida mais comum para se capturar graus de alavancagem, os chamados *'hat values'* (ver mais em Fox, 2008). Tais valores determinam, basicamente, quanto cada observação de y se distância dos seus respectivos valores preditos de y. A idéia é capturar estas distâncias através de um indicador, os *hat-values*, já que os valores preditos são combinações lineares das observações. Outra opção é calcular as 'distâncias de Cook' que representam a importância de cada observação para os coeficientes de regressão quando um caso específico é retirado da análise. Em outras palavras, o procedimento para se calcular tais distâncias é avaliar o poder relativo de cada observação caso ela fosse retirada da amostra. O 'gráfico de influência' pode ser uma alternativa para avaliar a influência dos casos sobre as estimativas combinando essas duas idéias. A figura 9 abaixo ilustra tal procedimento.

No eixo y tem-se os valores dos resíduos padronizados (*studentized residuals*) em relação aos valores esperados em uma distribuição t. As linhas pontilhadas horizontais demarcam os limites aceitáveis da influência exercida por cada ponto sobre as estimativas da regressão. O eixo x representa os valores de alavancagem (*hat-values*) que mostram os níveis de influencia de cada observação analisada. O tamanho das circunferências, por sua vez, é proporcional às distâncias de Cook, que indicam o impacto que nossos estimadores sofrem quando um determinado caso é excluído da amostra analisada.

O gráfico de influencia 1 apresenta a situação em que o pesquisador tem menos motivos para se preocupar com a influência dos casos destoantes. Ou seja, a escolha dos casos tem menos

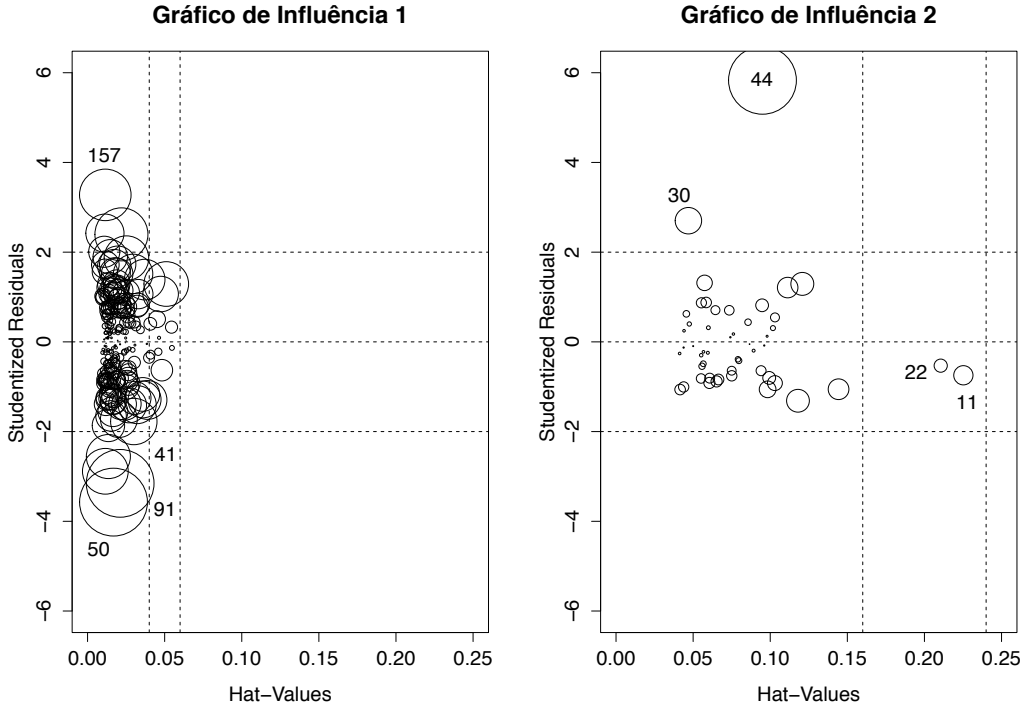

Figura 9: Gráficos de Influência

impacto sobre as estimativas produzidas. No gráfico de influência 2, ao contrário, os casos 11, 22, 30 e 44 devem ser observados com mais cuidado. Note que todos eles ultrapassam os limites estabelecidos nos eixos X e Y. Em relação à medida de Cook (magnitude dos círculos), observamos alto grau de influência em ambos os casos. No gráfico e influência 1, a influência é expressa nas extremidades da distribuição padronizada dos resíduos (ver casos 41, 50, 91 e 157). No gráfico de influência 2, apenas a observação 44 apresenta influência quando o teste de exclusão dos casos é feito[33].

O próximo diagnóstico necessário é o da normalidade dos resíduos. O método de mínimos quadrados opera melhor quando os erros são normalmente distribuídos. Erros substancialmente assimétricos podem comprometer a eficiência dos MQO e podem levantar dúvidas quanto a razoabilidade de se estimar a média condicional de y a partir de x. A distribuição

[33]E o que deve ser feito para lidar com os pontos destoantes? Temos 3 principais recomendações: (1) Excluir os casos desviantes (reportando o procedimento no paper); (2) recodificar os casos destoantes a partir de valores menos extremos e (3) escrever uma seção do artigo descrevendo e explicando os motivos pelos quais os casos são destoantes (Geddes, 2003)

dos resíduos da regressão é a melhor *proxy* para a distribuição dos erros, e é com ela que vamos investigar se o erro amostral satisfaz os pressupostos para o uso do método de MQO. O gráfico 10 abaixo ilustra tal procedimento.

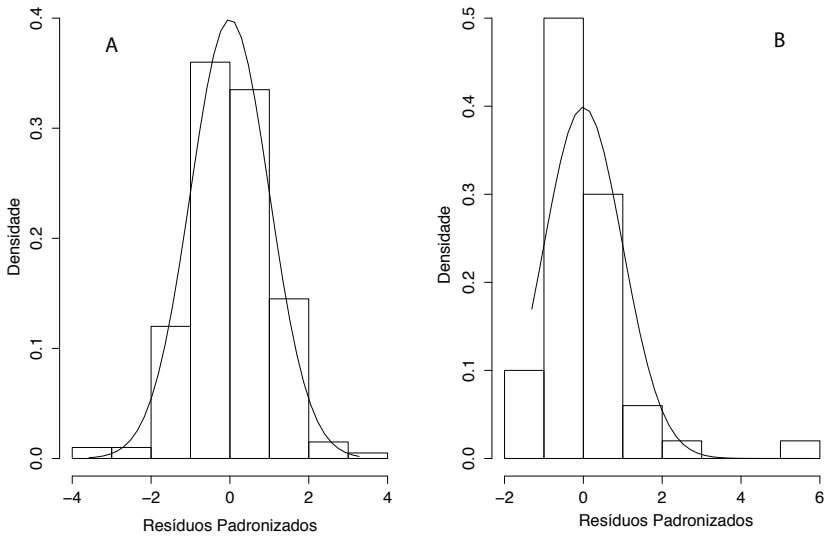

Figura 10: Distribuição dos resíduos da regressão linear

O gráfico A apresenta distribuição normal dos resíduos. Por sua vez, o gráfico B apresenta uma distribuição assimétrica. A solução para esse problema é transformar as variáveis incluídas no modelo de regressão com o objetivo de conseguir um melhor ajuste, satisfazendo os pressupostos. Recomendamos duas principais transformações: (1) logarítima e (2) extrair a raiz quadrada de y. Como regra geral, o pesquisador deve começar transformando a sua variável dependente e, se julgar necessário, deve transformar as variáveis independentes.

É importante destacar o pressuposto da homocedasticidade, ou seja, a variância de y (a variância do erro) deve ser aproximadamente constante para qualquer valor de x. Dado que nossas regressões envolvem o uso de muitas variáveis é impossível observar diretamente a distribuição dos resíduos em volta da superfície estimada. Os livros de estatística trazem tal ilustração quando se usa uma regressão bivariada (ver Fox, 2008). Um padrão comum da variância do erro não-constante, contudo, é o aumento da dispersão dos pontos de y com o aumento do seu nível. Esse padrão pode ser detectado plotando os resíduos da regressão

contra os valores preditos. A figura 11 abaixo ilustra esse procedimento.

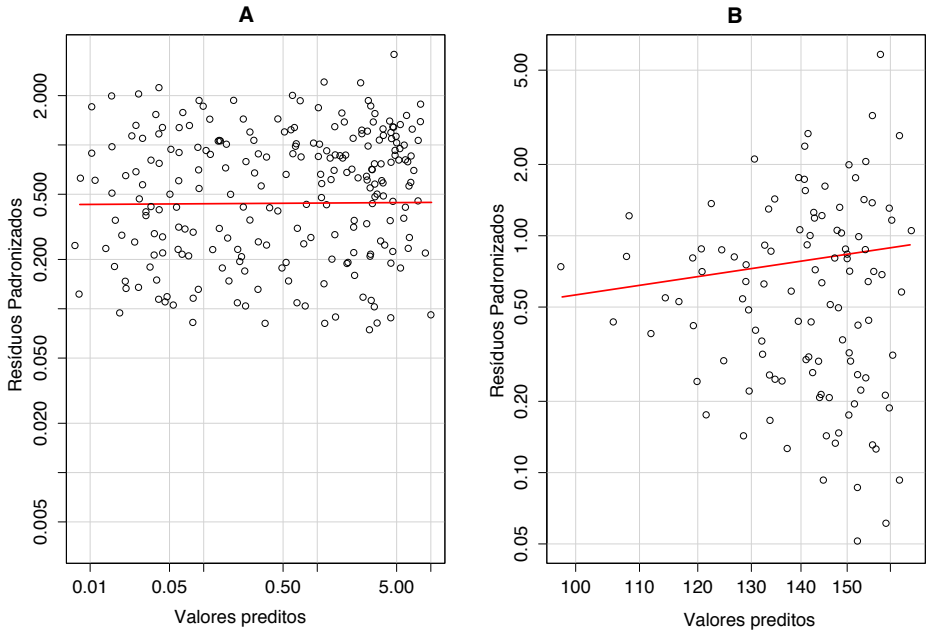

Figura 11: Distribuição dos pontos estimados em torno da reta de resíduos

O gráfico A apresenta uma distribuição homocedástica, ou seja, os valores preditos e os resíduos padronizados estão distribuídos aleatoriamente, o que significa dizer que nenhum padrão de variância é observada. Por sua vez, o gráfico B apresenta uma distribuição heterocedástica na medida em que a variância dos valores preditos de y aumenta com o incremento dos resíduos do modelo ajustado. Uma outra forma de distinguir os padrões é observando a inclinação da reta ajustada aos pontos. Quanto mais inclinada estiver a reta, maior é a heterocedasticidade dos nossos dados. Quanto maior a heterocedasticidade, menor tende a ser a eficiência dos coeficientes, o que produz erros de estimativa maiores.

Nós finalmente chegamos ao diagnóstico de multicolinearidade entre as variáveis independentes. A forma mais simples de avaliar se a premissa da independência entre os vetores de X é observada é plotando um gráfico de dispersão para todas as variáveis independentes utilizadas. Na figura 12 abaixo nós mostramos o grau e a direção da correlação entre 4 variáveis hipotéticas.

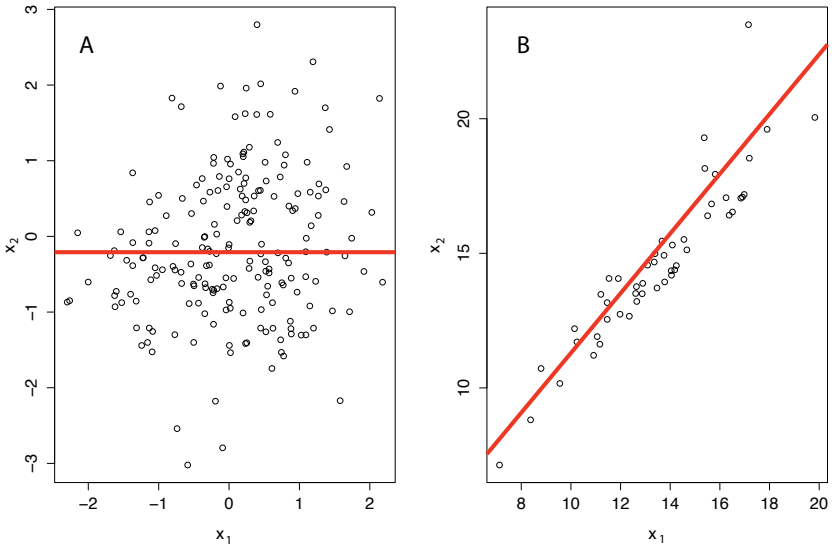

Figura 12: Gráfico de Correlação das Variáveis Independentes

O gráfico A ilustra uma situação em que x_1 e x_2 não apresentam nenhuma correlação linear. Observe que a reta ajustada é horizontal já que a dispersão dos pontos é completamente aleatória. O segundo caso, no entanto, mostra a situação inversa. Observe como no gráfico B as duas variáveis tem um grau muito alto de correlação, o que prejudicaria nossos estimadores, e por consequência os resultados da regressão linear ajustada. A recomendação para casos como esse é excluir x_1 ou x_2 do modelo e testar seus efeitos de forma independente. No exemplo ilustrado acima, não faria nenhuma diferença usar uma variável (x_1) ou outra (x_2) na regressão. Se as duas fossem usadas ao mesmo tempo, no entanto, o resultado seria a absorção do efeito de uma pela outra.

Muitos outros diagnósticos poderiam ser apresentados neste texto. Mostramos nessa seção o poder dos gráficos para a realização dos diagnósticos, mas vale lembrar que há muitos outros testes de significância que também nos auxiliam nesse trabalho (ver Cook, 1994; Fox, 2008). Nossa opção pelos cinco acima se deu tendo em vista o que de mais importante se tem discutido na ciência política atualmente[34]. A principal lição dessa seção é que o pesquisador

[34]Além desses, também é importante avaliar a auto-correlação entre as variáveis independentes, o que pode ser feito com testes de VIF ('variance-inflation factors'); a linearidade entre as variáveis, o que pode ser feito usando gráficos que combinam resíduos e valores preditos de cada variável (CERES plots); e a

precisa conhecer seus dados e reconhecer seus limites para que a escolha do modelo de análise seja a mais bem informada possível.

8 Conclusão

A análise de regressão de mínimos quadrados ordinários (MQO) é o modelo estatístico mais usualmente empregado na ciência política contemporânea. No Brasil, no entanto, a realidade é bem diferente. Decidimos elaborar esse artigo assumindo o seguinte pressuposto: a resistência à sua utilização pode ser explicada pela ausência de formação metodológica em geral e pelo limitado conhecimento da referida técnica em particular. Nosso principal objetivo foi reduzir a escassez de produção sobre metodologia em português, introduzindo a análise de regressão linear de mínimos quadrados ordinários de forma intuitiva.

Na primeira parte do texto apresentamos a estrutura básica do modelo e alguns dos principais pressupostos que precisam ser satisfeitos, bem como as consequências de sua violação sobre a consistência das estimativas. Na segunda parte mostramos a aplicação prática da análise de regressão utilizando dados simulados. Essa estratégia possibilitou controlar os parâmetros e, portanto, verificar a eficiência das estimativas produzidas. Por fim, discutimos superficialmente alguns cuidados que os pesquisadores devem tomar durante a utilização do modelo, destacando alguns dos testes que podem ser realizados para verificar em que medida os dados são adequados à utilização do modelo de mínimos quadrados ordinários.

Algumas lições podem ser tiradas desse exercício. Primeiro, embora nossa principal meta seja incentivar a utilização da regressão linear em ciência política, não estamos defendendo aqui o uso indiscriminado da referida técnica. Como os bons livros de metodologia nos ensinam, o que orienta a pesquisa científica é a pergunta que se pretende responder. Em muitos casos, no

independência das observações que pode ser feita testando a auto-correlação dos resíduos (Durbin-Watson test). Recomendamos o uso de tais diagnósticos para que a análise dos dados possa ser feita com o maior cuidado possível.

entanto, a aplicação do modelo linear é equivocada e não ajuda o pesquisador a atingir seus objetivos. Acreditamos que o uso indistinto e displicente de regressão linear pode induzir o pesquisador a cometer erros graves na interpretação de seus resultados, prejudicando o avanço do conhecimento científico.

Segundo, defendemos também que a melhor forma de apresentar os resultados e os diagnósticos é através de gráficos (Kastellec e Leoni, 2007). As tabelas dever ser utilizadas, mas como complemento. Queremos reforçar aqui a importância da visualização como instrumento analítico. A audiência científica, assim como a que está fora da academia, tem muito mais facilidade de compreender os resultados de pesquisa quando eles são explorados graficamente. O pacote estatístico usado e sugerido neste artigo (R) é uma ótima ferramenta para tanto. Além de gratuito, o R tem uma comunidade crescente e que desenvolve suas aplicações na internet. Ou seja, qualquer pesquisador pode ter acesso às atualizações, assim como pode contribuir para o aprimoramento do programa[35].

Terceira lição, o modelo linear de MQO não permite concluir nada sobre a causalidade entre variáveis. Na verdade, não existe nenhuma técnica estatística capaz de determinar causalidade entre os fenômenos de interesse do pesquisador. É o desenho da pesquisa escolhido que vai nos possibilitar inferir causalidade a partir dos dados coletados. Relações causais devem ser estabelecidas a partir da teoria disponível sobre o assunto. Como mostrado, a regressão linear trata-se de uma ferramenta que analisa a correlação entre variáveis, emulando o efeito de controle observado em experimentos de pesquisa. Há uma crescente literatura sobre como usar desenhos de pesquisa para solucionar tal limitação (Bartels, 1991; Angrist, Imbens, e Rubin, 1996; Dunning, 2008; Imbens e Lemieux, 2008; Angrist e Pischke, 2009). Além disso, a realização de experimentos tem se tornado cada vez mais comum em ciência política (ver Habyarimana, 2009), o que abre um leque imenso de oportunidades para o avanço do conhecimento na nossa disciplina. Quando bem empregada, a regressão linear auxilia, portanto,

[35]Há outras opções interessantes. Sugerimos, por exemplo, *DataDesk*, *GGobi*, *InfoVis*, e *Infographics*.

na identificação dos efeitos sistemáticos que conduzem nossas variáveis a se relacionarem.

Finalmente, nem o teste de significância nem o valor do r^2 devem ser o foco do pesquisador. É a importância de cada variável independente ou o tamanho do seu efeito que determina a relevância ou a significância substantiva dos resultados de pesquisa. Defendemos aqui a utilização dos intervalos de confiança ao invés dos testes padrão de significância, já que a partir dos intervalos de confiança é possível avaliar diferentes testes de hipótese ao mesmo tempo.

King, Keohane e Verba (1994) afirmam que os mesmos problemas de inferência assolam a pesquisa quantitativa e a qualitativa, argumentando que apenas é possível entender a realidade social se as pesquisas seguirem a lógica da inferência científica. Se quantitativa, qualitativa ou se combinando ambas as perspectivas, o importante é que o método seja um componente irredutível na produção do conhecimento. Do ponto de vista crítico, é inaceitável que não exista um único artigo publicado em periódicos nacionais que discuta a elaboração de desenho de pesquisa e/ou a aplicação de técnicas de pesquisa de forma mais aplicada. É inconcebível que a ciência social brasileira permaneça a-metodológica como diferentes diagnósticos sugerem. Acreditamos fortemente que o distanciamento do método está mais relacionado ao problema de formação de nossos pesquisadores do que propriamente a uma posição ontológica e epistemológica definida a respeito de como o conhecimento deve ser produzido. Como lembrado por Abraham Maslow, "se a única ferramenta a disposição é um martelo, é bem tentador tratar tudo como se fosse um prego". Talvez, no caso da ciência social brasileira, pior do que tratar todo problema como um prego, é a incapacidade de distinguir o martelo do prego.

9 Referências

ANGRIST, Joshua A., IMBENS, Guido W. and RUBIN, Donald. (1996) "Identification of Causal Effects Using Instrumental Variables."Journal of the American Statistical Association 91(434): 444-455.

ANGRIST, Joshua; PISCHKE, Jörn-Steffen (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

BARTELS, Larry M. (1991) "Instrumental and 'Quasi-Instrumental' Variables."American Journal of Political Science 35(3): 777-800.

BECK, Nathaniel; KATZ, Jonathan (1995). "What to do (and not to do) with Times-Series Cross-Section". American Political Science Review, vol 89, no 3: 634-647.

COLLIER, David; BRADY, Henry; SEAWRIGHT, Jason (2004). "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology."In Henry E. Brady and David Collier, eds. Rethinking Social Inquiry: Diverse Tools, Shared Standards. Lanham, MD: Rowman and Littlefield.

COOK, R. Dennis. (1998). Regression Graphics: Ideas for Studying Regressions Through Graphics. New York: Wiley.

COOK, R. Dennis; WEISBERG, Sanford (1994). An Introduction to Regression Graphics. New York: Wiley.

DUNNING, Thad. (2008) "Model Specification in Instrumental-Variables Regression."Political Analysis.

FOX, John (2008). Applied Regression Analysis and Generalized Linear Models. Second Edition, Sage Publications.

GARSON, David (2011). Statnotes: Topics in Multivariate Analysis, by G. David Garson. Disponível em: http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm. Acessado em 24 de agosto de 2011.

GEDDES, Barbara (2003). Paradigms and sand castles: theory building and research design in comparative politics. Ann Arbor, University of Michigan Press.

GELMAN, Andrew (2004). "Exploratory Data Analysis for Complex Models". Journal of Computational and Graphical Statistics, vol 13, no 4: 755779.

GERBER, Alan; GREEN, Donald; NICKERSON, David (2001). "Testing for Publication Bias in Political Science". Political Analysis, vol 9: 385-392.

GOLDBERGER, Arthur (1989). "The ET Interview: Arthur S. Goldberger". Econometric Theory, vol 5: 133-160.

GUJARATI, Damodar (2000). Econometria Básica. São Paulo: Macron Books, Pearson Education do Brasil.

HAIR Jr., Joseph; ANDERSON, Ralph; TATHAM, Ronald; BLACK, Willian (2009), Multivariate data analysis. 17 Edição. Prentice-Hall.

HABYARIMANA, James; HUMPHREYS, Macartan; POSNER, Daniel; WEINSTEIN, Jeremy. (2009) "Coethnicity: diversity and the dilemmas of collective action."Russell Sage Foundation Publications.

IMBENS, Guido e LEMIEUX, Thomas. (2008) "Regression discontinuity designs: A guide to practice."Journal of Econometrics, 142, 615-635.

IP, Edward (2001). "Visualizing Multiple Regression,"Journal of Statistics Education, vol 9, no 1.

KASTELLEC, Jonathan P. e LEONI, Eduardo. (2007) "Using Graphs Instead of Tables in Political Science." *Perspectives in Politics*, vol. 5, n. 4.

KENNEDY, Peter (2002). "Sinning in the Basement: What Are the Rules? The Ten Commandments of Applied Econometrics,"Journal of Economic Surveys, Wiley Blackwell, vol. 16: 569-589.

KENNEDY, Peter. (2009), "A Guide to Econometrics". Boston: MIT Press.

KING, Gary (1986). "How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science."American Journal of Political Science, vol 30: 666-687.

KING, Gary. (1995). "Replication, Replication."Political Science and Politics. no 28: 443-499.

KING, Garry.; KEOHANE, Robert. e VERBA, Sidney. (1994). "Designing social inquiry: scientific inference in qualitative research."Princeton: Princeton University Press.

KRUEGER, James; LEWIS-BECK, Michael. (2008). "Is OLS Dead?"The Political Methodologist, vol 15, no 2: 24.

LEWIS-BECK, Michael (1980). Applied Regression: an introduction. Series Quantitative Applications in the Social Sciences. SAGE University Paper.

MOORE, David; McCABE, George. (2009), Introduction to the practice of statistics. New York, Freeman.

PALLANT, Julie (2007). SPSS Survival Manual: A Step by Step Guide to Data Analysis using SPSS for Windows. Open University Press.

PEARSON, Karl (1982). The Grammar of Science. London: J.M. Dent and Sons Ltd.

SANTOS, Maria Helena; COUTINHO, Marcelo (2000), "Política Comparada: estado das

artes e perspectivas no Brasil", BIB, no 54: 3-146.

SOARES, Gláucio (2005). "O Calcanhar Metodológico da Ciência Política no Brasil". Sociologia, Problemas e Práticas, no 48: 27-52.

STEVENS, James (1996). Applied Multivariate Statistics for the Social Sciences. Terceira Edição. Mahwah, NJ: Lawrence Erlbaum Associates.

TABACHNICK, Barbara; FIDELL, Linda. (2007), Using multivariate analysis. Needham Heights, Allyn e Bacon.

TUFTE, Edward (1990). Envisioning Information. Cheshire, CT: Graphics Press.

VALLE SILVA, Nelson (1999), Relatório de Consultoria sobre Melhoria do Treinamento em Ciência Social Quantitativa e Aplicada no Brasil, Rio de Janeiro, Laboratório Nacional de Computação Científica, 15 de Abril de 1999, 22 pág.

VIANNA, Luiz Werneck; CARVALHO, Maria Alice Rezende de; MELO, Manuel Palacios Cunha; BURGOS, Marcelo Baumann (1999), "Doutores e teses em ciências sociais", Dados, vol 41, n 3: 453-515.

WEISBERG, Sanford (2005). Applied linear regression. Hoboken NJ: John Wiley.

WOOLDRIDGE, Jeffrey (2009). Econometrics: a modern approach. 4ª Edição. South-Western, Cengage Learning.