Artigo apresentado no X Encontro Da Associação Brasileira de Ciência Política (ABCP). Belo Horizonte, 30 de Agosto a 02 de Setembro – 2016.

Área temática:
Ensino e Pesquisa em Ciência Política e Relações Internacionais
Coordenadores:
Lorena Barberia/USP e Marcelo Valença/UERJ

PASSA OU REPASSA? QUESTÕES DE REGRESSÃO QUE VOCÊ TEM QUE SABER[β]

Ranulfo Paranhos (ICS/UFAL)
E-mail: ranulfoparanhos@me.com
Lattes: http://lattes.cnpq.br/5670542372454107

José Alexandre Silva Junior (ICS/UFAL)
E-mail: jasjunior2007@yahoo.com.br
Lattes: http://lattes.cnpq.br/5100245942386773

Willber Nascimento (DCP/UFPE)
E-mail: nascimentowillber@gmail.com
Lattes: http://lattes.cnpq.br/6856013993591304

Resumo

Como utilizar a análise de regressão linear de mínimos quadrados ordinários? Qual é o efeito da multicolinariedade sobre a eficiência das estimativas? Quais são as limitações do coeficiente de determinação (r^2)? Como interpretar o coeficiente quando a variável independente é *dummy*? Qual é a utilidade dos coeficientes não padronizados? Quais os perigos dos *outliers*? O que é e como se interpreta o p-valor? O objetivo desse trabalho é responder essas questões. Metodologicamente, o desenho de pesquisa utiliza simulação básica e replica dados secundários. Com esse artigo esperamos facilitar a compreensão de conceitos fundamentais de análise de dados e melhorar a qualidade dos resultados empíricos reportados em revistas científicas.

Palavras-chave: estatística; análise de dados; métodos quantitativos.

[β] Esse trabalho foi elaborado como material complementar da disciplina de Métodos Quantitativos I ofertada no departamento de Ciência Política da Universidade Federal de Pernambuco (DCP - UFPE). Agradecemos aos membros do grupo de Métodos de Pesquisa pelo apoio logístico. Eventuais limitações nem delegamos, nem dividimos: são todas nossas. Material de replicação está disponível em: {LINK}

Statistics is the grammar of science

Karl Pearson

It is the mark of a truly intelligent person to be moved by statistics

George Bernard Shaw

INTRODUÇÃO

No documentário *The Joy of Statistics*, Rosling (2010) afirma que "o mundo em que vivemos está repleto de dados que chovem por todos os lados. Sozinhos, esses dados são apenas barulho e confusão. Para dar sentido e achar significado precisamos de um ramo poderoso da ciência: a Estatística"[1]. Na verdade, é impossível passar um único dia sem ser exposto aos dados. Eles estão nos jornais, no noticiário, nos relatórios governamentais e até mesmo na quantidade de curtidas da sua postagem no *Facebook*. Em particular, a quantidade de informações digitais cresce exponencialmente de modo que a análise de dados é um procedimento central na gestão de organizações, no planejamento estratégico de empresas e na formulação de políticas públicas.

Este trabalho apresenta uma introdução à análise de dados a partir de sete questões sobre a utilização da regressão linear. Nosso público alvo são estudantes de graduação e pós-graduação em fases iniciais de treinamento. O trabalho tem três motivações principais. A primeira é a escassez de material intuitivo em português voltado especificamente para Ciência Política[2]. A segunda é a utilização incorreta de termos e técnicas em congressos e revistas especializadas. É comum observar a ocorrência de erros primários que comprometem a robustez das inferências realizadas. A terceira diz respeito a própria importância da regressão. De acordo com Krueger e Lewis-Beck (2008), a regressão linear de mínimos quadrados ordinários é a técnica mais utilizada na Ciência Política contemporânea. Dessa forma, é importante compreender seus pressupostos e aplicações.

[1] Ver <https://www.youtube.com/watch?v=U5Q9zdIHbRU>.
[2] Por exemplo, não existe no Brasil uma revista específica sobre metodologia de pesquisa como a *Political Analysis*, que atualmente é um dos periódicos mais influentes da área. Ver: <http://pan.oxfordjournals.org/>.

O restante do artigo está dividido em oito seções. A primeira parte apresenta uma introdução ao modelo de regressão linear de mínimos quadrados ordinários (MQO). A segunda discute o efeito da multicolinariedade sobre a eficiência das estimativas. A terceira introduz as principais limitações do coeficiente de determinação (r^2). A quarta parte ensina a interpretar os coeficientes do modelo de regressão linear quando a variável independente é dicotômica (*dummy*). A quinta seção discute a diferença dos coeficientes não padronizados e padronizados. A sexta discute o efeito de observações destoantes (*outliers*) sobre a consistência das estimativas. A sétima explica o que é e como se interpreta o p-valor. A última parte sumariza as conclusões.

1. PARA QUE SERVE E COMO UTILIZAR A ANÁLISE DE REGRESSÃO?

> *Regression is a powerful tool for forecasting. Economists using it successfully predicted ten out of the last two recessions*
>
> Nate Silver

Apesar da conotação negativa da referência acima, a regressão é uma poderosa ferramenta de análise de dados. Na verdade, regressão é uma denominação genérica para um conjunto de modelos matemáticos que são utilizados para avaliar o nível de associação entre variáveis independentes e uma variável dependente (Y) (TRIOLA, 2005; TABACHINCK e FIDELL, 2007; GUJARATI, 2012). É possível estimar o efeito individual de cada variável sobre a variação de Y e fazer previsões do valor da variável dependente a partir de valores conhecidos das variáveis independentes (HAIR et al, 2009; FIGUEIREDO FILHO et al 2011; WOOLDRIDGE, 2012). Em sua notação mais simples, o modelo pode ser descrito da seguinte forma:

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

Y representa a variável dependente, ou seja, aquilo que procuramos entender/explicar/predizer. A constante (α) representa o valor de Y na ausência de variáveis independentes. β_1 representa a variação observada em Y ao se elevar a variável independente (X_1) em uma unidade. Por fim, ε representa o termo de erro. Desde que os pressupostos sejam respeitados, as estimativas produzidas a partir de uma amostra aleatória serão não viesadas e

eficientes[3]. A ausência de viés significa que nem sobrestima nem subestima sistematicamente o valor do parâmetro populacional. E eficientes porque apresenta a menor variância entre todas as possíveis estimativas (LEWIS-BECK, 1980; KENNEDY, 2005).

E quando os pressupostos forem violados? Existem correções e aplicações que permitem modelar a variação de Y quando não for possível satisfazer os pressupostos do modelo de MQO. Por exemplo, quando a variável dependente for categórica binária, é mais adequado utilizar a regressão logística ou Probit. Em alguns desenhos de pesquisa, a variável dependente é ordinal ou até mesmo politocômica (mais de duas categorias), tem-se então a necessidade de modelos de regressão ordinal e multinomial, respectivamente. Tem-se ainda os modelos específicos para dados de contagem como Poisson e Binominal Negativo ou quando a variável dependente é censurada, caso em que deve-se optar pelo modelo Tobit e afins. Em outras oportunidades, o pesquisador está interessado em observar a variação da variável dependente ao longo do tempo e, por esse motivo, deve-se trabalhar com modelos específicos que levem em conta essa dimensão (regressão de painel e séries temporais). O Quadro 1 sumariza algumas recomendações sobre como utilizar o modelo de regressão na pesquisa empírica.

Quadro 1 - Como utilizar o modelo de regressão

ITEM	O QUE OBSERVAR
Conhecer teoricamente o campo de pesquisa	Como trabalhos anteriores especificaram o modelo explicativo? Quais são os principais problemas de mensuração? Os dados são válidos e confiáveis?
Definir as hipóteses *ex ante*	Testar hipóteses teoricamente orientadas. Postular a direção e a magnitude dos coeficientes a partir do conhecimento acumulado na área. Não utilizar a regressão em estudos exploratórios.
Conhecer seu banco de dados	Deve-se explorar exaustivamente o banco de dados antes de estimar os modelos explicativos. Erros de digitação, problemas de importação e casos destoantes podem comprometer a consistência das estimativas. Em *surveys,* deve-se analisar o questionário e a codificação das questões. Para escalas e índices, deve-se observar os pesos de cada variável. Para distribuições assimétricas, deve-se procurar a correção mais adequada.
Estimar o modelo	O desenvolvimento tecnológico facilitou a implementação computacional de modelos matemáticos por pesquisadores que prescindem de treinamento matemático avançado.
Analisar os resíduos	O erro é a diferença entre os valores observados e os estimados. Um dos pressupostos do modelo de MQO é de que o erro é aleatório e tem média zero. A análise dos

[3] (1) linearidade; (2) ausência de erro sistemático de mensuração; (3) a expectativa da média do termo de erro igual a zero; (4) homocedasticidade; (5) ausência de autocorrelação; (6) a variável independente não deve ser correlacionada com o termo de erro; (7) nenhuma variável teoricamente relevante para explicar Y foi excluída e nenhuma variável irrelevante foi incluída no modelo (erro de especificação); (8) ausência de multicolinearidade perfeita; (9) distribuição normal do termo de erro e (10) proporção adequada entre o número de casos e a quantidade de parâmetros estimados.

	resíduos é uma etapa fundamental já que existe muita informação dentro dos erros.
Apresentar os resultados	Os resultados devem ser comunicados de forma eficiente. Deve-se evitar linguagem técnica e enfatizar a compreensão intuitiva dos dados.

Fonte: elaboração dos autores (2016)

O primeiro passo é conhecer substantivamente o campo de pesquisa. Deve-se examinar como trabalhos anteriores têm especificado os modelos explicativos. Além disso, é importante observar como as variáveis estão sendo mensuradas. As medidas são válidas e confiáveis? Existe oferta de dados para o período da análise? Qual é a proporção de casos ausentes? Antes de estimar qualquer modelo de regressão, o pesquisador deve se familiarizar com contribuições teóricas e aplicações empíricas do campo de pesquisa com o objetivo de minimizar eventuais problemas de especificação e mensuração (HAIR et al, 2009).

O segundo é definir as hipóteses. Elas devem ser teoricamente orientadas e definidas *ex ante*. Não basta postular que o efeito será diferente de zero, deve-se informar a direção e a magnitude esperadas dos coeficientes. Em alguns trabalhos fica evidente que as hipóteses foram definidas depois da análise dos dados. Isso é errado. Deve-se evitar também a utilização da regressão em estudos exploratórios uma vez que o objetivo da técnica é fornecer estimativas sobre o padrão de associação entre variáveis a partir de alguma teoria. Na perspectiva explicativa, o pesquisador está interessado em estabelecer relações causais entre variáveis. Já na exploratória o objetivo é desbravar campos desconhecidos e/ou desenvolver novas hipóteses/questões (SPECTOR, 1982).

A terceira etapa é conhecer o banco de dados. O pesquisador deve saber a fonte das informações, a codificação das variáveis, as limitações dos dados, etc. Deve-se explorar descritivamente o banco de dados antes de estimar os modelos explicativos. Erros de digitação, problemas de importação, casos destoantes, entre outros pequenos problemas podem comprometer a consistência das estimativas (FIGUEIREDO et al, 2011). Por exemplo, durante uma junção de bases de dados imagine que em uma base os nomes dos municípios estão acentuados e em outra não. Essa incompatibilidade pode desorganizar a junção e comprometer a qualidade das informações. Enfim, apenas deve-se utilizar o modelo de regressão após ganhar familiaridade com a base de dados e eliminar problemas associados à coleta das informações.

O quarto procedimento é estimar o modelo. O desenvolvimento tecnológico facilitou a implementação computacional de modelos matemáticos por pesquisadores que não possuem treinamento específico na área. Esse avanço tem proporcionado o desenvolvimento de programas amigáveis que permitem realizar análises sofisticadas com uma curva de

aprendizagem relativamente simples. Comparativamente, o SPSS e STATA aparecem como os *softwares* mais utilizados (MUECHEN, 2015)[4]. Recentemente, alguns pesquisadores começaram a utilizar o R *Statistical* que além de ser um programa livre permite o desenvolvimento de novos algoritmos (DALGART, 2008; ZUUR, LENO e MEESTERS, 2009)[5]. Além disso, existe uma comunidade de usuários que disponibiliza códigos já prontos e ajuda na superação de problemas técnicos[6].

O quinto é passo é analisar os resíduos ou também chamados de erro, que nada mais é do que a diferença entre os valores observados e os estimados. Quanto maior o erro, pior é o modelo. Um dos pressupostos do modelo de MQO é de que o erro é aleatório e tem média zeo. Adicionalmente, o erro amostral deve ter uma distribuição normal para que os estimadores sejam não-enviesados e eficientes. A análise dos resíduos é uma etapa fundamental da pesquisa já que existe muita informação dentro dos erros (FREUND, VAIL e CLUNIES-ROSS, 1961; LARSEN e MCCLEARLY, 1972). Por exemplo, é possível identificar problemas na forma funcional da relação esperada entre os parâmetros da variável dependente e o conjunto de variáveis explicativas (ZYSKIND, 1963). A partir dos resíduos também é possível identificar casos extremos que podem comprometer a robustez dos resultados (COOK e WEISBERG, 1999; NELSON, 1973).

Por fim, os resultados deve ser comunicados de forma eficiente. Deve-se evitar linguagem técnica e enfatizar a compreensão intuitiva dos dados. Por exemplo, em artigos científicos, deve-se prezar por tabelas objetivas que reportem apenas as informações essenciais. Dados adicionais devem ser reportados nos anexos e a base original deve ser sempre disponibilizada em algum repositório público com o objetivo de aumentar a transparência e garantir a replicabilidade dos resultados (KING, 1995; JANZ, 2015)[7].

2. ENTENDENDO E SUPERANDO A MULTICOLINARIEDADE

Um dos pressupostos do modelo de regressão é de que não existe colinearidade perfeita entre as variáveis explicativas, ou seja, correlação de 1, independente do sinal (KENEEDY, 2005; HAIR ET AL, 2009). A colinariedade ocorre quando o nível de correlação entre duas variáveis independentes é muito alta (em geral, acima de 0,9, independente do sinal). No entanto, a gravidade do problema depende do tamanho da amostra. Quanto maior o número de

[4] Ver <http://r4stats.com/articles/popularity/>
[5] Ver <https://www.r-project.org/>
[6] Ver <http://www.inside-r.org/what-is-r>, <http://www.r-statistics.com/>, < http://www.r-bloggers.com/>
[7] Em relação à transparência, ver a iniciativa promovida pelo *Berkeley Initiatiave for Transparency in Social Sciences* (BITTS). Para compartilhar os dados, sugerimos o Consórcio de informações Sociais (CIS) da Universidade de São Paulo e o DATAVERSE da Universidade de Harvard.

observações, menos os coeficientes são afetados por altos níveis de correlação entre as variáveis independentes. A Tabela 1 apresenta um exemplo de colinearidade.

Tabela 1 - Colinearidade com três variáveis independentes

	VD	**X_1**	**X_2**	**X_3**
VD	1			
X_1	0,720	1		
X_2	0,670	0,750	1	
X_3	0,600	0,740	0,960	1

Fonte: elaboração dos autores (2016)

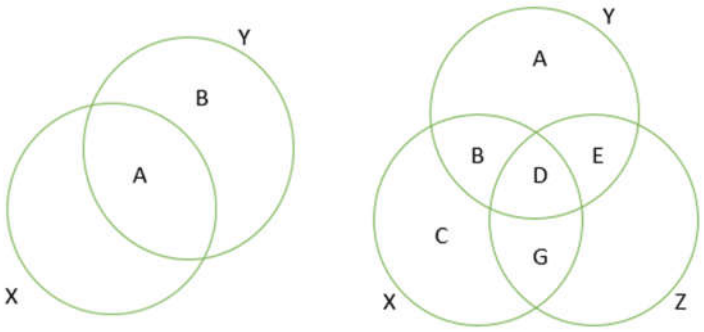
Matematicamente, a correlação de +1 ou -1 entre as variáveis explicativas impossibilita o cálculo do erro padrão e alguma variável será automaticamente excluída da análise[8]. No caso acima, a correlação entre X_2 e X_3 é de 0,960, superando tanto a correlação entre X_2 e a VD (0,670), quanto a associação entre X_3 com a VD (0,600). A multicolinariedade ocorre quando altos níveis de correlação são observados entre mais de duas variáveis independentes. Aqui tanto X_1 quanto X_2 estão altamente correlacionadas com X_3. Ou seja, tem-se uma alta correlação múltipla entre as variáveis explicativas.

E por que o analista de dados deve se preocupar com isso? Primeiro, a multicolinariedade tende a sobrestimar a magnitude dos erros padrão dos coeficientes, prejudicando a confiabilidade dos testes de significância. Isso quer dizer que tanto o p-valor quanto os intervalos de confiança serão afetados. Além disso, altos níveis de correlação entre as variáveis independentes podem produzir um modelo cujo coeficiente de determinação (r^2) é alto (ver próxima seção), mesmo quando nenhum dos coeficientes de regressão é estatisticamente diferente de zero. Em outras palavras, os coeficientes não serão significativos, mas o ajuste do modelo será alto, o que não faz sentido do ponto de vista substantivo.

De forma mais recorrente, altos níveis de correlação entre as variáveis independentes irão inverter o verdadeiro sinal do coeficiente de regressão. Ou seja, ao invés de observar um efeito positivo, o pesquisador vai concluir que as variáveis estão negativamente correlacionadas. Por fim, outro problema gerado por variáveis independentes colineares é a instabilidade dos coeficientes. Tanto a inclusão e/ou exclusão de um único caso ou a acréscimo de uma nova variável pode mudar dramaticamente a magnitude e, de forma mais preocupante, a direção dos coeficientes. A figura 1 ilustra como esse problema afeta a consistência das estimativas.

[8] No entanto, é raro observar correlações perfeitas entre variáveis observadas. Em geral, isso pode ocorrer com variáveis simuladas ou com erros de recodificação. Ver <http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis>.

Figura 1 - Multicolinariedade



Fonte: elaboração dos autores (2016)

No primeiro modelo (figura da esquerda) tem-se duas variáveis. A interseção entre X e Y está ilustrada pela letra A. B representa a variação de Y que independe da variação em X, denotada por B. O segundo modelo tem duas variáveis independentes (X e Z) e a mesma variável dependente, Y. A área D + G representa a correlação entre as variáveis independentes. Se apenas a variável X for utilizada para entender/explicar/predizer Y, tem-se informação referente à área B + D. Se apenas a variável Z for utilizada para entender/explicar/predizer Y tem-se informação referente a área D + E. Mas o que acontece se forem utilizadas as variáveis X e Z ao mesmo tempo? A regressão linear de mínimos quadrados ordinários utiliza apenas a variância única entre cada variável independente e a variável dependente. Ou seja, toda a informação da área D seria perdida (área comum entre X e Z). Portanto, quanto maior for a correlação entre as variáveis independentes (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 (KENNEDY, 2009).

Como detectar problemas de multicolinariedade? A forma mais simples é estimar uma matriz de correlação entre as variáveis independentes. Quanto maior a magnitude dos coeficientes, independente do sinal, maiores os problemas de multicolinariedade. Deve-se tratar cada variável independente como se ela fosse uma variável dependente e estimar um modelo explicativo a partir das demais variáveis independentes. A literatura indica duas principais medidas para diagnosticar problemas de colinariedade: (1) Tolerância e (2) *Fator de Inflação da Variância (Variance Inflation Factor – VIF).*

A tolerância é a quantidade de variabilidade de uma variável independente que *não é explicada pelas demais variáveis independentes.* Ela é calculada a partir de $1\text{-}R^2$. Por exemplo,

se o modelo explica 30% da variável independente, então a tolerância de X_1 é de 0,70 (1 – 0,3). Quanto maior a tolerância, menor nível de colinariedade.

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

Até 1 = ausência de multicolinariedade
Entre 1 e 10 = multicolinariedade aceitável
Acima de 10 = multicolinariedade problemática

Por fim, uma vez detectado os problemas de multicolinariedade, como superá-los? Nesse artigo sugerimos quatro procedimentos: (1) checar a codificação e transformações das variáveis independentes; (2) aumentar o tamanho da amostra; (3) utilizar alguma técnica de redução de dados e (4) consultar a literatura.

A primeira recomendação é checar as codificações e transformações realizadas nas variáveis independentes. Um simples deslize de atenção pode produzir efeitos perversos sobre a qualidade do modelo, principalmente em amostras pequenas. Por exemplo, uma recodificação mal renomeada pode criar problemas de colinariedade já que a mesma variável vai ser duplamente incluída. Após a checagem dos dados, se os problemas persistirem, a próxima recomendação é aumentar o tamanho da amostra. Kennedy (2005) sugere que os problemas de multicolinariedade são especialmente recorrentes em amostras pequenas (micronumerosidade). Dessa forma, sugerimos elaborar um desenho de pesquisa que maximize a quantidade de observações. Se a unidade de análise é a unidade federativa, uma forma de aumentar a amostra é desagregar os dados por município. Todavia, muitas vezes a adição de novas observações ou não resolve o problema ou não é possível. O que fazer então nesses casos?

Nossa terceira sugestão é utilizar alguma técnica de redução de dados. Essas técnicas são especialmente adequadas para lidar com situações em que as variáveis independentes são fortemente correlacionadas (HAIR et al, 2009). Dessa forma, é possível reduzir a

dimensionalidade dos dados e criar um índice que carrega a informação das variáveis originais. Esse novo indicador pode ser utilizado como variável dependente ou independente em novos modelos explicativos (FIGUEIREDO FILHO e SILVA JUNIOR, 2010; FIGUEIREDO FILHO et al 2014). Uma desvantagem desse procedimento é a impossibilidade de observar o efeito individual de cada variável explicitava. Aqui vale a máxima: ninguém pode ter tudo[9].

Nossa última recomendação é consultar a literatura sobre o fenômeno de interesse e identificar que variáveis são as mais relevantes. Muitos modelos incluem variáveis por comodidade ou simplesmente para "ver o que acontece". Nesses casos, uma saída é literalmente excluir a variável do modelo. Lembrando que a excluso de uma variável teoricamente importante pode gerar problemas de especificação, que são mais graves do que os gerados por variáveis altamente correlacionadas. Dessa forma, a exclusão de variáveis colineares apenas deve ser utilizada como último recurso.

3. VANTAGENS E LIMITAÇÕES DO COEFICIENTE DE DETERMINAÇÃO[10]

O coeficiente de determinação (r^2) é uma medida da qualidade do ajustamento da equação de regressão. Ele fornece a proporção da variação da variável dependente explicada pela variação das variáveis independentes. Em geral, serve como indicador para avaliar em que medida a relação entre as variáveis pode ser descrita por uma função linear. Em um cenário de ajuste perfeito todas as observações estariam situadas na reta de regressão. Como isso raramente acontece, é comum a ocorrência de erros positivos (observações acima da reta) e negativos (observações abaixo da reta). Quanto menor a distância entre as observações e a reta maior o coeficiente de determinação (r^2), melhor o ajuste do modelo. Dessa forma, o r^2 pode ser calculado a partir da razão entre soma dos quadrados da regressão (explicados) e a soma total de quadrados (GUJARATI E PORTER, 2011; HAIR ET AL, 2009).

O coeficiente de determinação é uma das medidas de ajuste mais utilizadas na pesquisa empírica (MOKSONY, 1990). Por outro lado, para Anderson-Sprecher (1994), "o coeficiente de determinação múltipla R^2 é uma medida que muitos estatísticos amam odiar. Essa animosidade existe primariamente porque o uso generalizado do r^2 leva inevitavelmente ao seu ocasional uso indevido" (ANDERSON-SPRECHER, 1994: 113). Embora a controvérsia a

[9] Agradecemos a Dona Neusa Américo dos Santos por esse alerta.

[10] Essa seção se baseia no artigo *What is R^2 all about?,* publicado pela Revista Leviathan no seguinte endereço eletrônico: <http://www.fflch.usp.br/dcp/leviathan/index.php/leviathan/article/view/85>. Luskin (1984; 1991a; 1991b) e King (1986; 1990; 1991) fornecem uma excelente introdução a respeito do papel do R^2 na Ciência Política.

respeito do r^2 tenha origem na Estatística, o debate é importante para todas as áreas de conhecimento que utilizam modelos de regressão linear. Portanto, já que o uso de modelos de regressão linear tem crescido na Ciência Política, é importante entender o papel controverso do r^2, e o significado substantivo que pesquisadores podem extrair dessa medida.

Para King (1986), tanto o coeficiente de correlação de Pearson quanto o coeficiente de determinação possuem sérias limitações. Ele argumenta que "na maioria das situações práticas da ciência política, não faz muito sentido utilizar essas estatísticas. Elas não medem aquilo que aparentam medir, e podem ser bastante enganosas" (KING, 1986: 669). Achen (1977) afirma que uma das principais limitações do coeficiente de correlação é a sua incapacidade de ser comparado entre amostras. King (1986) argumenta que "todas as críticas feitas aos coeficientes de correlação e de regressão padronizados aplicam-se igualmente à estatística r^2" (KING, 1986: 675).

Kavalseth (1985) indica duas situações em que o r^2 não é confiável: (1) quando o modelo é estimado sem constante e (2) quando o pressuposto da linearidade é violado. Scott e Wild (1991) alertam que a interpretação do coeficiente de determinação é inapropriada quando os modelos são estimados utilizando transformações da escala da variável resposta. McGuirk e Driscoll (1995) argumentam que o r^2 é importante, mas "o tamanho do R^2 e do R^2 ajustado são indicadores de especificação precários, já que modelos especificados corretamente podem apresentar um R^2 'baixo', e modelos mal especificados frequentemente apresentam R^2 'altos'" (MCGUIRK e DRISCOLL, 1995 p. 319).

Hair et al (2009) apontam que a adição de qualquer variável – seja ela significante ou não – tende a aumentar o valor do r^2. O impacto dessa adição será maior a partir do momento em que o tamanho da amostra for igual ao número de variáveis independentes. Para solucionar esse problema, os autores recomendam a utilização do r^2 ajustado, que minimiza o impacto da adição de variáveis ao modelo e permite a "[...] comparação entre equações de regressão que envolvem diferentes números de variáveis independentes ou diferentes tamanhos de amostra" (HAIR ET AL, 2009: 182). Em suma, a literatura indica que:

a) como o coeficiente de determinação (R^2) depende do coeficiente de correlação (r), ele pode ser influenciado pelas diferentes variâncias entre as amostras (ACHEN, 1977). Por essa razão, a estatística r^2 não pode ser utilizada para comparar amostras diferentes;
b) o r^2 não deve ser empregado para analisar modelos sem constante (KAVALSETH, 1985);

c) o r^2 não deve ser usado quando os pressupostos básicos da regressão de mínimos quadrados ordinários forem violados (SCOTT E WILD, 1991);
d) o r^2 não garante um bom ajuste. Do mesmo modo, um r^2 baixo não garante que o modelo é mal especificado (ACHEN, 1977; MCGUIRK e DRISCOLL, 1995).

Em nossa opinião, a principal limitação do r^2 é a sua incapacidade de oferecer ao pesquisador uma estimativa do efeito de uma determinada variável independente sobre a variação da variável dependente (ASCOMBE, 1973; ACHEN, 1977; KING, 1986). Dessa forma, essa estatística é pouco informativa quando o objetivo da pesquisa é encontrar relações causais entre variáveis. Por outro lado, o r^2 é uma ferramenta importante quando o objetivo é examinar o ajuste geral de um modelo de previsão (LEWIS-BECK e SKALABAN, 1990).

4. E QUANDO A VARIÁVEL INDEPENDENTE É *DUMMY?*

Usualmente, as variáveis *dummies* são codificadas como "0" e "1", lembrando que não faz diferença qual categoria recebe cada valor já que a magnitude dos coeficientes permanece constante, o que muda é o sinal. Como foi visto, o modelo de regressão pode ser utilizado para entender/explicar/predizer a variação de uma variável dependente a partir de um conjunto de variáveis independentes. Por exemplo:

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

O coeficiente de regressão β_1 representa a variação observada em Y ao se elevar X_1 em uma unidade. No caso específico das variáveis *dummies*, β_1 representa a diferença média entre o grupo codificado como "1" e o grupo codificado como "0". Ou seja, o coeficiente informa o mesmo valor do que seria observado ao se realizar um teste t para amostras independentes entre os dois grupos. Para fixar a interpretação, testamos a hipótese de que o time mandante apresenta melhor desempenho do que o visitante. O modelo estimado é o seguinte:

$$\text{Número de pontos} = \alpha + \beta_1 \text{ Mandante} + \varepsilon$$

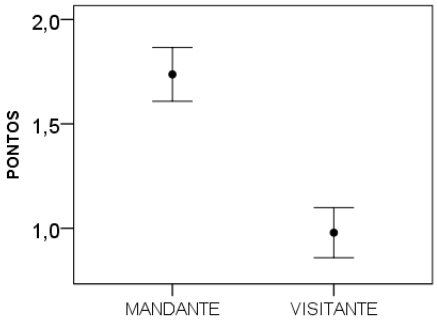
A variável dependente é o número de pontos conquistados em uma determinada partida e a variável independente é a condição de mandante (1) ou visitante (0). Dessa forma, espera-se que β_1 seja positivo. O Gráfico 1 e a Tabela 2 sumarizam a comparação de médias.

Gráfico 1 - Intervalo de confiança (95%)

Tabela 2 - Comparação de médias (n = 380)

Time	Média	Desvio padrão	Coef. Var
Mandante	1,737	1,277	0,736
Visitante	0,979	1,189	1,214

* Diferença média = 0,758 (p-valor<0,001)



Fonte: elaboração dos autores (2016)

Em média, o time mandante conquista 1,737 pontos por partida, enquanto o visitante aufere 0,979 pontos. A diferença média entre mandantes e visitantes é de 0,758 pontos (p-valor<0,001), ou seja, é exatamente a subtração da média dos mandantes (1,737) da média dos visitantes (0,979). Outra forma de observar a diferença entre os grupos é analisar o intervalo de confiança das médias (ver Gráfico 1). Quando não há interseção entre os intervalos, deve-se rejeitar a hipótese nula de igualdade de médias. Por outro lado, quanto maior a interseção entre os intervalos de confiança, menor a probabilidade de rejeitar a hipótese nula. A Tabela 3 sumariza as principais estatísticas de interesse do modelo linear de mínimos quadrados.

Tabela 3 - Modelo linear de mínimos quadrados ordinários

	Coef. não padronizados		Coef. padronizados	t	p-valor
	B	**Erro padrão**	**BETA**		
Constante	0,979	0,063		15,469	0,000
Mandante	0,758	0,089	0,294	8,469	0,000

VD: número de pontos

Fonte: elaboração dos autores (2016)

$$\text{Número de pontos} = 0{,}979 + 0{,}758\ \text{Mandante}$$

A constante (0,979) representa o valor esperado de Y para o grupo codificado como zero, ou seja, representa a média de pontos esperada do time visitante. O time mandante tem uma vantagem média de 0,758 pontos em relação ao time visitante. A média de pontos esperada do time mandante é calculada a partir da soma da constante (0,979) mais a diferença média entre mandantes e visitantes (0,758) = 1,737 (ver Tabela 2).

Essa seção apresentou uma introdução "relâmpago" sobre como o coeficiente de regressão linear deve ser interpretado quando a variável independente é categórica binária. Os resultados sugerem que quando os times jogam em casa eles apresentam melhor desempenho

do que quando visitam. Isso o torcedor já sabia. O que ele não sabia era como fazer *smart regression*s with *dummy variables*.

5. QUAL É A UTILIDADE DOS COEFICIENTES DE REGRESSÃO?

Em termos estatísticos, o coeficiente de regressão representa a variação em Y devido a alteração de X. Portanto, corresponde a uma medida de compartilhamento de variância entre as duas variáveis (Y, X). Matematicamente, o coeficiente pode ser chamado de taxa de variação ou crescimento da função[11]. A partir dele é possível saber qual o ritmo de crescimento ou redução de Y em função de X. Por exemplo, se na equação $Y = \alpha + \beta_1 X_1$, $\beta_1 = 3$ então Y aumentará três vezes mais rápido que X. Em termos gráficos, o coeficiente é representado pelo coeficiente angular ou, simplesmente, inclinação da reta. Considere o seguinte exemplo:

Equação 01

$$Y_{VINC} = 48{,}159 + 2{,}177\, X_{Emennda} + 5{,}733 X_{Mídia}$$

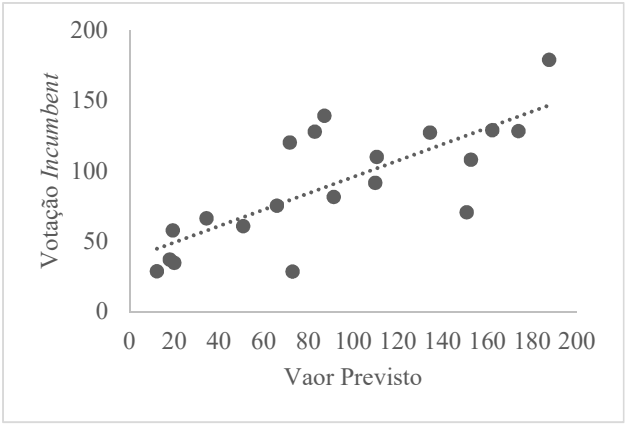
Onde,

Y_{VINC} = Votação do *Incumbent* (100mil);

$X_{Mídia}$ = Tempo de Exposição na Mídia (Horas);

X_{Emenda} = Valor Executado em Emendas Individuais ao Orçamento (Milhões)

Gráfico X – Modelo de Regressão Linear



O modelo prevê que cada hora a mais de exposição na mídia deve elevar em 5.733 a votação do *incumbent*. Já cada milhão a mais gastos em emendas individuas deverá render 2.177 votos para o candidato. Caso desprezemos a unidade de medida, pode-se dizer que o total de votos cresce 2,177 mais rápido do que o total de recursos executados em emendas individuais ao orçamento. Similarmente, é possível dizer que esse ritmo de crescimento é 5,733 mais rápido do que a exposição na mídia. Graficamente,

[11] O modelo de regressão linear está baseado em uma função linear do tipo y = ax + b, para todo x ϵ R e a ≠ 0. A principal especificidade desta função é que a variação é constante e igual **a**.

o efeito total das duas variáveis sobre a votação dos *incumbents* é representado pela inclinação da reta.

Tecnicamente, os coeficientes de regressão sumarizam a importância de cada variável independente na explicação da variação da dependente. Além disso, eles indicam a correspondência entre a distribuição dos dois tipos de variáveis. Para fornecer essas informações os coeficientes são apresentados em dois formatos: não padronizado e padronizado.

Coeficientes Não Padronizados

O não padronizado é baseado na *covariância* entre a variável independente (X) e a dependente (Y). Isso significa que sua magnitude cresce na medida em que escores de X acima da média são correspondentes aos escores de Y na mesma condição (WRIGHT, 1934; ACHEN, 1977; HARGENS, 1976)[12]. Além disso, ele possui a característica peculiar de ser apresentado na unidade de medida da variável original. Logo, representa a variação de Y ao se elevar X em uma unidade, ambas em suas respectivas medidas originais. Segundo a literatura, o coeficiente não padronizado é indicado para verificar relações causais (BLALOCK, 1964; DUCAN, 1975; KING, 2001; KIM E MULLLER, 2002). Isso porque é uma medida menos afetada pela variância amostral. Em outras palavras, é mais estável que coeficiente padronizado entre as subamostras da população. Segundo Kim e Muller (1976), sob certas condições, apenas o coeficiente padronizado é afetado por conjunto de variâncias-covariâncias, com fonte na: 1) variável independente; 2) variáveis que são explicitamente identificadas no sistema causal e 3) variáveis não explicitamente incluídas no sistema causal (KIM E MULLER, 1976: 436). Para Duncan (1975), a análise de causalidade pode, em grande medida, desprezar a variação das variáveis oriundas da replicação de amostras aleatórias. No limite, a utilização do coeficiente não padronizado pode evitar que o pesquisador descarte uma relação de causalidade por conta da variância amostral. Adicionalmente, o coeficiente não padronizado é o único que fornece informações por unidade de medida. Portanto, é recomendado quanto o objetivo é estimar alteração na variável dependente (Y) ligada a adições pontuais na independente (X).

[12] Embora relacionados os termos covariância e correlação não são intercambiáveis. A covariância corresponde ao valor esperado do produto do desvio entre os escores e as respectivas médias de X e Y, a correlação é o quociente da divisão entre a covariância (X, Y) pelo produto dos desvios padrões (X, Y). Precisamente: $cov\,(x,y) = \frac{\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{n}$; $r = \frac{Cov^2(y,x)}{var\,(y)var\,(x)}$

Coeficiente Padronizado

A base de cálculo do coeficiente padronizado é a *correlação* entre X e Y. Isso significa que ele cresce na mesma direção da covariância, mas é afetado pelos desvios padrões de X e Y. Na verdade, o desvio padrão é a sua unidade de medida. Portanto, ele representa a variação de Y ao se elevar X em desvios padrões. De acordo com a literatura, a principal função do coeficiente padronizado é assegurar comparabilidade. Segundo Hunter e Hamilton (2002), ele permite comparar a relação causal replicada por estudos diferentes. Além disso, permite ao pesquisador determine qual preditor tem efeito maior sobre a variável dependente. O coeficiente padronizado coloca as variáveis na mesma escala, dessa forma, estabelece comparabilidade mesmo quando elas são mesuradas por medidas distintas (HUNTER E HAMILTON, 2002). Segundo Bentler (2001), quanto menor o consenso sobre a melhor medida para mensurar o fenômeno maior a importância desse artifício. Também por isso, existe a possibilidade do pesquisador ponderar a importância de cada variável independente. Precisamente, o maior coeficiente irá pertencer a variável que possui uma distribuição mais próxima da variável dependente. Portanto, a informação oferecida não dirá respeito as unidades da variável, mas ao seu conjunto (TUKEY, 1954; HARGENS, 1976; ACHEN, 1977; RICHARDS, 1986). Voltemos ao exemplo anterior.

Equação 01 – Coeficiente não Padronizado

$$Y_{VINC} = 48{,}159 + 2{,}177\ X_{Emenda} + 5{,}733 X_{Mídia}$$

Equação 02 – Coeficiente Padronizado

$$Y_{VINC} = 0{,}528\ X_{dpEmenda} + 0{,}515 X_{dpMídia}$$

Onde,

Y_{VINC} = Votação do *Incumbent* (100mil);

$X_{Mídia}$ = Tempo de Exposição na Mídia (Horas);

X_{Emenda} = Valor Executado em Emendas Individuais ao Orçamento (Milhões)

Como a variável exposição a mídia e emendas executadas são medidas em unidades diferentes, não é possível dizer qual delas têm o maior peso para explicar a variação da votação dos *incumbents*. Afinal é difícil precisar a correspondência entre tempo e recursos orçamentários. Seria um erro acreditar que a exposição na mídia é mais importante por ter um coeficiente não padronizado maior. No entanto, parte da literatura acredita que essa comparação pode ser feita a partir do coeficiente padronizado (HUNTER E HAMILTON, 2002). No nosso exemplo, esses coeficientes têm magnitudes muito próximas com alguma

vantagem para a variável emenda executada. Dessa forma, é correto dizer que a emenda executada tem uma distribuição mais próxima da votação dos *incumbentes* e está mais correlacionada com a variável dependente. Na prática, essa seria uma forma de ponderar a importância das variáveis independentes, ainda que indiretamente.

A critica de King (1986)

No entanto, King (1991) sugere cautela na interpretação dos coeficientes não padronizados. Um primeiro problema é que eles são menos intuitivos já que é interpretação é realizada em termos de desvio padrão. Ou seja, a cada unidade adicional no desvio padrão da variável independente, a variável dependente varia em B_n desvio padrão. Principalmente para o leitor não especializado, informar que a variável dependente aumentou em 2 desvio padrão é pouco intuitivo. Algumas variáveis independentes, como gênero e raça, não podem ser alteradas. Dessa forma, deve-se evitar a comparação entre variáveis manipuláveis e não manipuláveis. Por exemplo, em um estudo sobre mercado de trabalho, o pesquisador pode observar que o efeito da raça é maior do que o impacto da escolaridade. O problema é que não é possível elaborar uma política para elevar a raça, ao passo que a escolaridade é passível de intervenção.

Por fim, em uma perspectiva de políticas públicas, é importante observar o custo associado à modificação das variáveis inseridas no modelo. Para ilustrar esse argumento, replicamos o exemplo elaborado por King (1991) que propõe explicar o número de visitas ao médico por ano em função de duas variáveis: (1) número de maças ingeridas e (2) número de laranjas consumidas. Vejamos:

$$Y = 10 – 1{,}5\ X_1 – 0{,}25\ X_2$$

Cada maçã adicional está associada a uma redução média de 1,5 visitas ao médico por ano, enquanto a ingestão de uma laranja reduz, em média, em 0,25 a variação da variável dependente. Então deve-se concluir que o efeito da maçã é maior do que o efeito da laranja? Resposta: depende. Se você tem apenas dinheiro para comprar uma unidade de fruta, é melhor comprar a maçã. No entanto, suponha que a maçã custa R$ 0,50, enquanto a laranja custa R$ 0,05. Com R$ 1 é possível comprar duas maçãs e reduzir a quantidade de visitas em 3, enquanto o mesmo dinheiro pode comprar 20 laranjas, o que diminuiria o número de visitas em

5, em média. Dessa forma, a interpretação dos coeficientes padronizados depende também da questão específica que será respondida.

Imagine agora que outra variável é adicionada ao modelo (X_3), no entanto, ela é medida em outro nível de mensuração, vejamos:

$$Y = 10 – 1{,}5\ X_1 – 0{,}25\ X_2 + 2\ X_3$$

Ainda que todas as variáveis independentes tenham o objetivo de explicar a variação da mesma variável dependente, a diferença no nível de mensuração impossibilita qualquer comparação substantiva dos efeitos observados. Para King (1991), a padronização dos coeficientes não melhora a qualidade da informação disponível sobre o modelo. Para ele, se não existia sentido na comparação antes da padronização, não existe justificativa para comparar os coeficientes após a padronização.

Em síntese, sugerimos quatro principais alertas na interpretação dos coeficientes padronizados: (1) eles são menos intuitivos do que os coeficientes não padronizados; (2) evitar a comparação de coeficientes entre variáveis manipuláveis e não manipuláveis; (3) observar o custo de modificação das variáveis de interesse e (4) a padronização não garante necessariamente comparações inteligíveis.

6. *OUTLIERS*: O QUE SÃO, COMO IDENTIFICÁ-LOS E O QUE FAZER COM ELES?[13]

Essa seção responde essas questões. O Quadro 2 sumariza diferentes definições.

Quadro 2 - Definições de *outliers*

Autor (ano)	Definição
Grubbs (1969)	an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which occurs
Hawkins (1980)	an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism

[13] Chandola, Banerjee e Kumar (2007) e Hodge e Austin (2004) apresentam um *survey* dessa literatura. Berton e Zhao (2011) discutem a detecção de *outliers* em redes complexas. Ben-gal (2005) apresenta uma introdução à lógica de detecção de observações atípicas. Seo (2002) compara diferentes metodologias de detecção de *outliers* em dados univariados. Comparativamente, Barnett e Lewis (1994) apresentam uma das mais completas abordagens sobre o tema. Para um tratamento intuitivo em português sobre casos discrepantes e seus efeitos sobre o coeficiente de correlação ver Figueiredo Filho et al (2014).

Johnson (1992)	An observation in a data set which appears to be inconsistent with the remainder of that set of data
Mendenhall *et al* (1993)	Observations whose values lies very far from the middle of the distribution in either direction
Barnett e Lewis (1994)	Indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
Pyle (1999)	An outlier is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable

Fonte: elaboração dos autores (2016)

Comparativamente, preferimos a definição de Hawkins (1980) já que o autor faz referência ao processo gerador das observações. Dessa forma, entendemos *outlier* como uma observação que se desvia tanto dos demais casos que desperta desconfiança sobre a sua origem. A literatura identifica quatro principais motivos que explicam a presença de casos destoantes: (1) atividade maliciosa; (2) erro de instrumento; (3) mudança no meio ambiente e (4) erro humano.

Um exemplo típico da atividade maliciosa como gerador de casos desviantes é quando a central do cartão de crédito liga para o indivíduo para confirmar uma determinada compra. Ou seja, sempre que as observações se distanciarem muito bruscamente da média, é um sinal de que algo anormal pode estar acontecendo.

O *outlier* causado por erro de instrumento é aquele que recebe uma determinada característica/atributo muito diferente das demais observações por falha do instrumento. Imagine uma balança desregulada ou um termômetro defeituoso. Se o pesquisador depende desses instrumentos para atribuir valores para os seus casos, ele corre o risco de incluir uma observação destoante em sua amostra.

Mudanças abruptas no meio ambiente, como uma tempestade, tendem a produzir observações muito diferentes do esperado. É comum observar esse tipo de ocorrência em reportagens sobre os índices pluviométricos, quando o jornalista afirma que em determinada semana choveu mais do que o esperado para todo o mês. Em Ciências Sociais, podemos citar revoluções, guerras civis e crises econômicas como exemplos de eventos que tendem a produzir mudanças abruptas nas variáveis de interesse.

Por fim, o caso destoante gerado por erro humano é um dos mais recorrentes na pesquisa científica. Muitos bancos de dados são efetivamente elaborados por estudantes de graduação e pós-graduação com diferentes níveis de treinamento técnico. Mesmo pesquisadores experientes muitas vezes cometem erros durante o processo de coleta e/ou tabulação de dados. Para além de

erros de digitação, *outliers* podem surgir como resultado de erros de importação de dados ou de processos automatizados de coleta. Ainda, a experiência de pesquisa indica que casos destoantes surgem simplesmente porque a fonte original de dados apresenta alguma inconsistência de mensuração. Por exemplo, o Cadastro Nacional de Improbidade Administrativa (CNJ) reúne algumas sentenças que foram julgadas antes de serem protocoladas, o que produz um tempo de julgamento negativo. No entanto, por definição, o tempo transcorrido para a ocorrência de um evento é uma variável não negativa.

E como identificar casos destoantes? Nesse artigo apresentamos três diferentes procedimentos: (1) regra do desvio padrão; (2) diferença interquartílica e (3) *trimed mean*.

7. O QUE É E COMO SE INTERPRETA O P-VALOR?

O p-valor é o parâmetro de significância mais utilizado pela ciência para testar hipóteses. Historicamente, os testes de significância começaram a ser usados há mais de 300 anos com os estudos de Laplace (1778), comparando a diferença na população de homens e mulheres na Europa do século XVIII (HUBERTY, 1993). Pearson (1900) introduz a ideia do p-valor a partir da criação do teste do chi-quadrado[14]. Porém, é a partir de Fisher (1925) que a utilização do p-valor se popularizou na Estatística com a adoção dos níveis de significância. Neyman-Pearson (1933) também geraram importantes contribuições ao introduzir a ideia de teste de hipótese. Apesar de suas particularidades, as contribuições de Fisher e Neyman-Pearson se complementam e formam um importante *framework* dentro da Estatística e, consequentemente, de todos os ramos do conhecimento científico (HUBBARD e BAYARRI, 2003).

Em um teste de significância, tem-se duas hipóteses acerca do valor de um determinado parâmetro: nula (H_o) e alternativa (H_a). Enquanto a nula aponta que o parâmetro assume um valor fixo, a alternativa assume que o parâmetro está em um intervalo de valores, possuindo algum nível de variação (AGRESTI e FINLAY, 2012). As hipóteses nula e alternativa sempre caminham em sentidos opostos: uma nega o efeito (H_o) e a outra assume o efeito (H_a). Em geral, a alternativa é a hipótese de trabalho. Por exemplo, em um estudo sobre a relação entre investimento em segurança pública e taxa de homicídios, a hipótese nula sustenta que não existe relação entre as variáveis, enquanto a hipótese alternativa supõe que quanto maior o investimento, menor a taxa de homicídios.

14

Estatisticamente, o p-valor é a probabilidade de rejeição da H_o. Quando menor o p-valor, em relação ao nível de significância, maior a plausibilidade de se rejeitar[15] a hipótese nula. De uma maneira simplificada, podemos afirmar que: "*when the p-value is low, null hypostesis most go*". Ou, como colocou um aluno nosso: "quando o p-valor é baixo, a hipótese nula deve ir por água abaixo.

Contudo, seja para rejeitar seja para não rejeitar a H_o, deve-se definir um nível de significância específico. Convencionalmente, os níveis considerados significativos são: 1%, 5% e 10%. A adoção de um determinado nível de significância é totalmente arbitrária. Não há uma regra na literatura que justifique a adoção de um valor em detrimento de outro, isso varia de cada pesquisador e por área.

Além de saber o que é e como interpretar, é necessário ter alguns cuidados ao realizar os testes de significância. Figueiredo et al (2013) apontam quatro cuidados que o pesquisador deve ter antes de interpretar o p-valor.

O primeiro deles é atenção aos gráficos. Sem eles, o pesquisador pode ter problemas na especificação da forma funcional de suas análises. Por exemplo, pode-se assumir que a relação entre duas variáveis é linear. Contudo, examinando os gráficos, constata-se que a relação entre elas é quadrática. Caso o pesquisador não esteja atento à análise gráfica, ele pode subestimar a magnitude dos parâmetros analisados.

O segundo ponto é em relação à amostragem. De acordo com Figueiredo et al (2013), não faz sentido analisar o p-valor em amostras não aleatórias, já que elas ferem os princípios da distribuição normal e do teorema do limite central e, consequentemente, levam a inferências enviesadas. A amostras devem seguir o princípio da randomização, que, além de eliminar o viés (SMITH, 1983), permite a produção de inferências válidas. Não é possível interpretar o p-valor sem a randomização das amostras, independentemente do tamanho (FIGUEIREDO FILHO et al, 2013).

O terceiro ponto é o tamanho da amostra. Em alguns casos, tende-se culpar a falta de significância estatística pelo tamanho da amostra (FIGUEIREDO FILHO et al, 2013), já que o cálculo do p-valor também é uma função inversa da quantidade de casos (*n*). Com isso, quanto maior a amostra, espera-se um menor p-valor e, com isso, a significância estatística. No entanto, os pesquisadores devem ter cuidado, pois, ao se aumentar o tamanho da amostra, pequenos efeitos podem tornar-se significantes, o que pode comprometer a robustez dos resultados esperados.

[15] Lembrando que o termo técnico é rejeitar e não rejeitar. Jamais, deve-se dizer aceitar ou não aceitar alguma hipótese.

Por fim, o último ponto estimar um p-valor quando se analisa uma população. A partir do momento em que os pesquisadores trabalham com a população, não faz sentido realizar estatística inferencial, já que as diferenças entre os casos, por mais pequenas que sejam, acontecem de fato (HAIR et al, 2009). A lógica da estatística inferencial é produzir inferências válidas para toda a população a partir de amostras. Não faz sentido realizar hipóteses quando todos os parâmetros populacionais são conhecidos pelos pesquisadores (FIGUEIREDO FILHO et al, 2011).

Para exemplificarmos, voltemos ao exemplo do Brasileirão 2013, apresentado na seção 4. Vamos testar a hipótese de que os times mandantes ganham mais pontos que os times visitantes. Adotaremos um ∞ = 0,05. O primeiro passo é montarmos nossas hipóteses nula e alternativa:

$$H_o: \bar{X}_m = \bar{X}_v$$

$$H_a: \bar{X}_m > \bar{X}_v$$

A H_o aponta que a média de pontos dos times mandantes e visitantes são iguais, ou seja, a diferença entre elas é zero. Sendo assim, não há diferença entre jogar em casa ou em jogar fora de casa. Já a hipótese alternativa afirma que a média de pontos dos mandantes é maior do que a dos visitantes, representando nossa hipótese de interesse de que os mandantes ganham mais pontos que os visitantes. O próximo passo é analisar a diferença média entre os grupos.

Tabela 4 - Estatística descritiva

Time	Média	Desvio padrão	Coef. Var
Mandante	1,737	1,277	0,736
Visitante	0,979	1,189	1,214

Fonte: elaboração dos autores (2016)

Notamos que o p-valor = 0,001 é menor do que nosso ponto de corte (0,05). Com isso, menor é a probabilidade de erro em relação a nossa hipótese alternativa, logo devemos rejeitar a hipótese nula e assumirmos que há uma diferença de pontos, estatisticamente significativa, entre mandantes e visitantes. Sendo que os primeiros ganham, em média, mais pontos que o segundo[16].

Analisaremos, agora, o modelo dos mínimos quadrados ordinários (tabela X), elaborado também na seção 4. Os coeficientes, para que sejam aplicados no modelo, devem ter

[16] Vale observar que, mesmo se adotássemos um ∞ = 0,05, a diferença entre mandantes e visitantes seria estatisticamente significativa, já que o p-valor = 0,001 < 0,01.

significância estatística. Adotaremos o mesmo ponto de corte (0,05) e montaremos nossas hipóteses nula e alternativa para a variável mandante.

Tabela 5 - Modelo linear de mínimos quadrados ordinários

	Coef. não padronizados		Coef. padronizados	t	p-valor
	β	**Erro padrão**	**BETA**		
Constante	0,979	0,063		15,469	0,000
Mandante	0,758	0,089	0,294	8,469	0,000

VD: número de pontos

Fonte: elaboração dos autores (2016)

$$H_o: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

A H_o aponta que o efeito da variável mandante sobre o número de pontos é zero. Não havendo diferença, então, entre jogar em casa ou em jogar fora de casa, assim como no exemplo anterior. Já a hipótese alternativa afirma que o efeito do time mandante é diferente de zero, ou seja, o fato do time jogar em casa irá impactar de alguma maneira na quantidade de pontos conquistados.

Ao compararmos o p-valor ao nível de significância, somos orientados a rejeitar a hipótese nula e assumirmos a hipótese alternativa. Com isso, podemos afirmar que a variável mandante é estatisticamente significativa em nosso modelo. É importante ressaltar que, caso o p-valor fosse acima do nível de significância, e, por conseguinte, tivéssemos que rejeitar a hipótese alternativa e assumirmos a hipótese nula, a variável teria que ser excluída do modelo, já que ela não possui a significância estatística.

CONCLUSÃO

Esse trabalho apresentou uma introdução à sete questões de Estatística que analista de dados devem saber. Nosso foco foi discutir os conceitos de forma intuitiva, minimizando aplicações algébricas com o objetivo de maximizar o potencial de audiência. O artigo é dirigido para estudantes de graduação e pós-graduação em fases iniciais de treinamento. Em termos substantivos, esperamos facilitar a compreensão de conceitos fundamentais de análise de dados e melhorar a qualidade dos resultados empíricos reportados em revistas científicas.

REFERÊNCIAS

Achen, C. H. (1977). Measuring representation: Perils of the correlation coefficient. *American Journal of Political Science*, 805-815.

Agresti, A., & Finlay, B. (2012). Métodos estatísticos para as ciências sociais. In *Métodos estatísticos para as ciências sociais*. Penso.

Anderson-Sprecher, R. (1994). Model comparisons and R 2. *The American Statistician*, *48*(2), 113-117.

Arceneaux, K., & Huber, G. A. (2007). What to do (and not do) with multicollinearity in state politics research. *State Politics & Policy Quarterly*,*7*(1), 81-101.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). New York: Wiley.

Berton, L., & Zhao, L. (2011). Caracterização de Classes via Otimização em Redes Complexas. *VIII Encontro Nacional de Inteligência Artificial (ENIA2011)*, 548-559.

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook* (pp. 131-146). Springer US.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). New York: Wiley.

Cook, R. D., & Weisberg, S. (1999). Graphs in statistical analysis: Is the medium the message?. *The American Statistician*, *53*(1), 29-37.

Chandola, V., Banerjee, A., & Kumar, V. (2007). *Outlier detection: A review*. Technical Report, University of Minnesota.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107.

Figueiredo Filho, D. B., Paranhos, R., Rocha, E. C. D., Batista, M., Silva Jr, J. A. D., Santos, M. L. W. D., & Marino, J. G. (2013). When is statistical significance not significant?. *Brazilian Political Science Review*, *7*(1), 31-55.

Figueiredo Filho, D. B. et al (2014). Reply on the Comments on When is Statistical Significance not Significant?. *Brazilian Political Science Review*, *8*(3), 141-150.

Figueiredo Filho, D. B., & Silva Júnior, J. A. D. (2010). Visão além do alcance: uma introdução à análise fatorial. *Opinião Pública*, *16*(1), 160-185.

Figueiredo Filho, D. et al (2011). O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). *Revista Política Hoje*, *20*(1).

Freund, R. J., Vail, R. W., & Clunies-Ross, C. W. (1961). Residual analysis.*Journal of the American Statistical Association*, *56*(293), 98-104.

Fisher, R. A. (1925, July). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 22, No. 05, pp. 700-725). Cambridge University Press.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*,*59*(2), 127-136.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Gujarati, D. N. (2012). *Basic econometrics*. Tata McGraw-Hill Education.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education*, *61*(4), 317-333.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*(3), 171-178.

Hair, J. F. et al. (2009). *Análise multivariada de dados*. Bookman.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of economics and statistics*, 486-489.

Helland, I. S. (1987). On the interpretation and use of R2 in regression analysis. *Biometrics*, 61-69.

Kavalseth, T. O. (1985). Cautionary note abaut R 2. *The American Statistican*, *39*(4), 279-285.

Kavalseth, T. O. (1985). Cautionary note abaut R 2. *The American Statistican*, *39*(4), 279-285.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 666-687.

King, G. (1990). Electoral responsiveness and partisan bias in multiparty democracies. *Legislative Studies Quarterly*, 159-181.

King, G. (1991). " Truth" Is Stranger than Prediction, More Questionable than Causal Inference. *American Journal of Political Science*, 1047-1053.

King, G. (2001). Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization*, *55*(02), 497-507.

Korn, E. L., & Simon, R. (1991). Explained residual variation, explained risk, and goodness of fit. *The American Statistician*, *45*(3), 201-206.

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

Larsen, W. A., & McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, *14*(3), 781-790.

Leamer, E. E. (1973). Multicollinearity: a Bayesian interpretation. *The review of economics and statistics*, 371-380.

Lewis-Beck, M. S., & Skalaban, A. (1990). The R-squared: Some straight talk. *Political Analysis*, 153-171.

Luskin, R. C. (1984). Looking for R^2: Measuring Explanation Outside OLS.*Political Methodology*, 513-532.

Luskin, R. C. (1991). Abusus non tollit usum: standardized coefficients, correlations, and R 2s. *American Journal of Political Science*, 1032-1046.

McGuirk, A. M., & Driscoll, P. (1995). The hot air in R2 and consistent measures of explained variation. *American Journal of Agricultural Economics*, *77*(2), 319-328.

McGuirk, A. M., & Driscoll, P. (1995). The hot air in R2 and consistent measures of explained variation. *American Journal of Agricultural Economics*, *77*(2), 319-328.

Moksony, F. (1990). Ecological analysis of suicide: problems and prospects.*Current Concepts in Suicide*.

Pearce, D. K., & Reiter, S. A. (1985). Regression strategies when multicollinearity is a problem: A methodological note. *Journal of Accounting Research*, 405-407.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157-175.

SMITH, T. M. F. (1983), On the validity of inferences from Non-random Samples, *Journal of the Royal Statistical Society – Series A (General)*, vol. 146, nº 4, pp. 394–403.

Scott, A. J., & Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 497-510.

Scott, A., & Wild, C. (1991). Transformations and R 2. *The American Statistician*, *45*(2), 127-129.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 539-552.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole.

Triola, M. F. (2005). *Introdução à estatística* (Vol. 9). Rio de Janeiro: Ltc.

Pyle, D. (1999). *Data preparation for data mining* (Vol. 1). Morgan Kaufmann.

Vaughan, T. S., & Berry, K. E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education*, *13*(1), 1-9.

Wooldridge, J. (2012). *Introductory econometrics: A modern approach*. Cengage Learning.

Zyskind, G. (1963). A note on residual analysis. *Journal of the American Statistical Association*, *58*(304), 1125-1132.