PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO


DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

# AVALIANDO EMPIRICAMENTE OS DETERMINANTES DE ERRO NAS PESQUISAS ELEITORAIS BRASILEIRAS PARA PRESIDENTE

**João Felipe Gomes de Almeida**

*Matrícula* 1411932

**Arthur Bragança (arthurbraganca@gmail.com)**

*Orientador*

Junho de 2019

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO


DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

# AVALIANDO EMPIRICAMENTE OS DETERMINANTES DE ERRO NAS PESQUISAS ELEITORAIS BRASILEIRAS PARA PRESIDENTE

**João Felipe Gomes de Almeida**

Matrícula: 1411932

**Arthur Bragança (arthurbraganca@gmail.com)**

*Orientador*

Junho de 2019

"Declaro que o presente trabalho é de minha autoria e que não recorri para realizá-lo, a nenhuma forma de ajuda externa, exceto quando autorizado pelo professor tutor"

“As opiniões expressas neste trabalho são de responsabilidade única e exclusiva do autor”

Dedico esta monografia à minha família que me apoiou sempre ao longo desta grande jornada; desde meu ingresso ao curso de engenharia até minha mudança de curso para economia que me levou a ser quem eu sou hoje.

Agradeço os meus amigos e amigas que sempre se fizeram disponíveis para conselho e companheirismo a respeito de tudo que eles poderiam ajudar.

Agradeço todos os professores do Departamento de Economia da PUC-RJ, pelo conhecimento que foram capazes de transmitir.

Por fim, agradeço muito o meu orientador, Arthur Bragança, por toda sua contribuição neste trabalho. Sem dúvidas, não seria capaz de efetuar este estudo não fosse sua enorme função como guia na conduta desta monografia.

# Sumário

# Índice de Tabelas

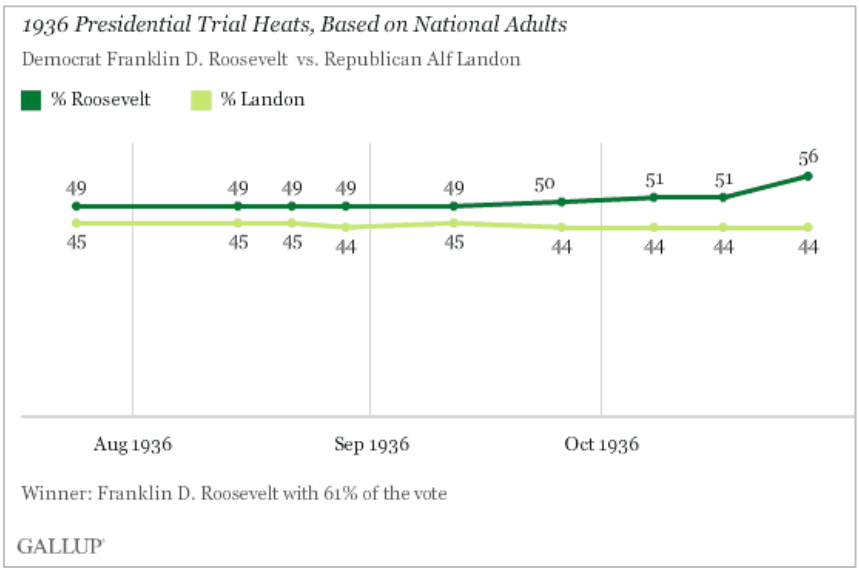
# Índice de figuras

# 1. Introdução

Estados Unidos da América, 1936, ano de eleição entre o democrata Franklin Delano Roosevelt e o republicano Alf Landon. A revista *The Literary Digest* envia 10 milhões de correspondências para seus assinantes perguntando em quem eles votariam. A revista recebe de volta aproximadamente 2,3 milhões de respostas e conclui que o vencedor será Alf Landon com 370 dos 539 *electoral votes*[1], e 57% do voto popular[2].

Entretanto, George Gallup, fundador do *Gallup Institute* que hoje ainda está ativo e presente em diversos países, discordava fortemente. Gallup afirmava que o método utilizado pelo *The Literary Digest* iria prover resultados enviesados, pois em sua coleta de informações eles não haviam incluído os milhões de Americanos mais pobres que apoiavam Roosevelt fortemente (em especial, sua proposta do *New Deal).* Gallup, então, fez o que nenhum homem havia feito até aquele momento; utilizou da estatística moderna, desenvolvida não muito tempo antes, para conduzir as primeiras pesquisas eleitorais probabilísticas da história. Foram conduzidas nove pesquisas entrevistando cinquenta mil pessoas, número risível à época quando comparado aos 2,3milhões de seu concorrente, e, em todas essas, previu-se uma vitória de Roosevelt, levando 56% dos votos populares na boca de urna, à margem de erro de 4 pontos percentuais[3].

**Figura 1.1**




[1] No sistema eleitoral dos EUA cada Estado vale um determinado número de 'votos eleitorais' e tem sua área geográfica divididas em regiões. O vencedor no Estado leva todos os votos eleitorais.
[2] https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/.
[3] https://news.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4.

Resultado Final: Franklin Delano Roosevelt – 61%, Alf Landon – 37%. Apesar do erro de 5%, acima da margem, este resultado foi considerado uma vitória completa não só para o Instituto Gallup, como também, e mais importantemente, para o início da implementação estatística nas pesquisas eleitorais. Este evento, incitado por um único homem e seu instituto recém-fundado (1935), foi suficiente para iniciar uma cultura de pesquisas estatísticas frente à política, que culminou para o que vemos hoje onde temos veículos de mídia, cientistas e até participantes do mercado financeiro contratando pesquisas eleitorais.

No Brasil, a primeira eleição presidencial democrática desde que Getúlio Vargas assumiu poder foi em 1945. Nesta eleição temos o candidato general Eurico Gaspar Dutra, do exército, contra o brigadeiro Eduardo Gomes, da aeronáutica. O Eduardo Gomes seria o representante da oposição às ideologias Varguistas, visto que a aeronáutica foi o setor militar que mais se opôs ao Estado Novo de Vargas, mas a principio Dutra não possuía apoio de Getúlio devido às complicações de sua destituição.

O, à época recém-criado, Instituto Brasileiro de Opinião e Pesquisa previu, em maio, uma vitória do candidato da oposição com margens consideráveis[4]. Porém, diferentemente do que ocorreu com Gallup, o Ibope errou. E errou muito. O resultado final foi vitória de Dutra com 20% de margem, completamente inverso a previsão do instituto.

Porém, a comparação com Gallup aqui não é justa; o Ibope acabava de ser criado pelo Estado durante uma ditadura e pôde conduzir apenas uma única pesquisa para as eleições, meses antes do dia de votação. Entre a pesquisa e a eleição foi se vendo diversas mudanças no cenário político; em especial, Vargas, ainda muito popular, acaba por apoiar Dutra na reta final. Temos que levar em consideração também que o Ibope ainda não tinha recursos para conduzir uma pesquisa em larga escala, tendo que se limitar a 1000 pessoas, todas da região de São Paulo.

Sendo assim, um erro como o que foi observado não deveria surpreender ninguém. Sabemos que as pesquisas comumente erram em relação ao resultado final, mas o objetivo delas não é prever o resultado e sim retratar a situação atual da sociedade frente uma eleição que está por vir. É extremamente possível que em maio de 1945 a opinião do povo de São Paulo, aqui proxy para o Brasil, era de fato uma forte preferência por Eduardo

[4] http://www.aberta.org.br/educarede/2013/05/21/historia-da-pesquisa/

Gomes, mas com todas as mudanças no cenário político e as novas informações absorvidas pelo povo até o dia da eleição Novembro, Dutra tomou a frente.

Muita coisa mudou desde a divulgação desta primeira pesquisa eleitoral no Brasil. Este fato é exemplificado pelas cento e dez pesquisas eleitorais conduzidas por 15 institutos diferentes para as eleições de 2018. Hoje os institutos melhoraram seus métodos e questionários de forma à, em tese, melhor retratar a opinião social em relação a seus candidatos. Pesquisas são frequentes e sempre são capazes de abranger a nação, e não só São Paulo.

A maior motivação para o desenvolvimento deste estudo vem a partir do peso que a sociedade atribui às pesquisas eleitorais. Em um estudo divulgado pelo Ibope em 26 de setembro de 2018 tem-se que 28% da população julgava ser "alta" ou "muito alta" a probabilidade de efetuarem o chamado voto útil[5]. Voto útil é aquele no qual não se opta pelo candidato que você julga ser o melhor, mas sim pelo melhor candidato (ou "menos pior") que ainda considera-se ter chances de vitória. Essa chance de vitória por sua vez, é em grande parte determinada pelos resultados das pesquisas eleitorais.

Em outras palavras, as pessoas usam o retrato contemporâneo da eleição para definir o retrato futuro da eleição em uma relação retro alimentícia. É fácil imaginar como receber a informação, por exemplo, de que seu candidato favorito tem 5% dos votos e margem de erro de 2% poderia o fazer desviar seu voto dele, visto que tem um outro candidato qual você rejeita muito com 15% das intenções de voto, à beira do segundo turno. A preocupação que teremos aqui é: mas e se soubéssemos que a margem de erro frente o resultado, dado que estamos à "T" dias antes da eleição e o candidato/pesquisa tem o conjunto de características "C", não for 2% mas sim (2+X)%; será que as pessoas teriam menos incentivo para aderir ao voto útil e mais incentivo a fazer campanha para aquele candidato qual elas julgam ser melhor?

O objetivo desse estudo é investigar empiricamente o erro das pesquisas presidenciais e seus principais determinantes. Para isso é construída uma base de dados de 324 pesquisas de eleições presidenciais dos anos de 2006, 2010, 2014 e 2018 e utilizados métodos de regressão linear para entender seus determinantes. Os resultados indicam que os principais fatores geradores de erro são (i) a quantidade de dias até a eleição, (ii) o tipo do questionário e (iii) o alinhamento ideológico do partido do candidato em questão.

[5] https://politica.estadao.com.br/noticias/eleicoes,cniibope-tres-em-cada-10-eleitores-dizem-que-podem-aderir-ao-voto-util-no-primeiro-turno,70002520385.

O restante desta monografia esta dividido em 5 capítulos sendo estes a Revisão de Literatura, Metodologia, Apresentação dos Dados, Analise dos Resultados e Conclusão.

# 2. Revisão de Literatura

Este capítulo será dedicado à revisão de literatura existente quanto ao tópico de analise empírica e teórica do processo de condução das pesquisas eleitorais, em virtude de contextualizar a profundidade na qual foi estudado os erros atribuídos a elas.

Como já previamente mencionado e amplamente conhecido, as pesquisas eleitorais dispõem de uma "margem de erro" em torno de seus resultados. Esta margem costuma oscilar geralmente entre dois ou três pontos percentuais, ao nível de significância de 95%. Isto significa, por exemplo, que se determinado candidato tem sua intenção de voto apresentada como 42%, existe uma probabilidade de 95% do resultado verdadeiro se encontrar entre 40% e 44% (assumindo uma margem de erro de dois pontos percentuais).

Porém, como evidenciado por Ferraz (1996), este cálculo da margem de erro pode não haver base estatística, dependendo da maneira em que é coletada a amostragem. A razão para tal vem do fato que a maioria dos institutos utilizam a técnica de amostragem por quotas, no lugar da amostragem aleatória, devido a sua maior facilidade e menor custo. Na amostragem aleatória, determina-se um número "N" de pessoas a serem entrevistadas, e escolhe-se aleatoriamente esta quantidade de pessoas da população. Com isso, espera-se, estatisticamente, que por ter sido feita a amostragem de maneira aleatória, ela irá representar adequadamente a população, dado um N grande suficiente. Por outro lado, na amostragem por quotas, são escolhidos os entrevistados de maneira a preencher determinadas "quotas" que, em tese, representam o todo. Por exemplo, se o censo de um país determinou que a percentagem de mulheres na população é de 54%, serão escolhidos entrevistados de modo a atingir esta razão entre homens e mulheres. Após atingir a quota, todos os demais entrevistados que excederem o valor da quota terão suas respostas descartadas. O problema do método vem do fato de que uma vez que a amostra não foi feita de maneira probabilística, o cálculo da margem de erro e nível de significância, que são probabilísticos em base, se tornam enviesados.

Deste modo, pode-se concluir que o erro apresentado na divulgação das pesquisas é um intrinsecamente enganoso, pois o cálculo é feito utilizando o N da amostra e calculando como se esta fosse aleatória.

Entretanto, deve-se perguntar o porquê da divulgação desta margem de erro estatisticamente fictícia? A principal razão seria que

> "essa discrepância está na aplicação da Lei 9.504/1997, que diz que as pesquisas pré-eleitorais, para serem divulgadas, devem estar registradas na Justiça Eleitoral cinco dias antes, contendo informações sobre o plano amostral, o intervalo de confiança e a margem de erro, entre outras (Art. 33, IV). Porém, o cálculo preciso da margem de erro só pode ser feito após a realização do trabalho de campo. Aplicada à risca, a regra levaria institutos e meios de comunicação a divulgarem pesquisas cujos dados foram coletados com, por exemplo, uma semana ou mais de antecedência, correndo o risco de noticiar retratos já defasados da corrida eleitoral. Portanto, os institutos informam antes mesmo de ir a campo uma margem de erro fictícia, calculada a partir de uma amostra aleatória simples, que nunca se cumpre uma vez que os institutos usam, na maioria das vezes, amostras por cotas ou conglomerados." (Gramacho, 2013, p. 5)

Tendo em mente o que foi exposto quanto à natureza do erro divulgado, torna-se ainda mais importante o estudo do erro verdadeiro; como não se pode confiar nas margens apresentadas pelos institutos, o maior agregador de conhecimento para interpretação das pesquisas seria então encontrar a margem real, aqui objetivado através da analise dos dados empíricos divulgados pelos institutos.

Quanto ao estudo da precisão empírica das pesquisas eleitorais, este é menos desenvolvido do que se esperaria. Dentre o que foi publicado sobre este tema, ressalta-se dois; um relatório desenvolvido pelo Itaú Asset Management (2018) e um artigo escrito por Gramacho (2013), já mencionado anteriormente. Enquanto o relatório do Itaú (2018) cobre uma janela temporal mais similar à que aqui será estudada, por não se tratar de um artigo acadêmico ele não se aprofunda na base teórica por trás das discrepâncias. Deste modo, o relatório não tem como foco principal observar o que causa os desvios do valor verdadeiro das percentagens de voto, mas sim buscar entender se esta margem de erro verdadeira tem aumentado ou diminuído ao longo dos anos.

Por outro lado, o artigo de Gramacho (2013) elabora uma análise mais profunda justamente dos erros empíricos e seus possíveis fatores geradores. Gramacho (2013) baseia-se na literatura especializada de Crespi (1988); Magalhães E Moreira (2007); Groß (2007); Callegaro & Gasperoni (2008), que aponta diversas possíveis fontes de erro para uma pesquisa eleitoral. Estas fontes podem ser divididas em dois grupos, e serão apresentadas de maneira resumida, baseadas no exposto por Gramacho (2013).

O primeiro grupo possui três fatores, todos quais o pesquisador tem influência sobre, e são:

i. Tempo entre a coleta dos dados e a eleição; quanto mais perto da eleição for à coleta, maior será a precisão.
ii. Técnica empregada pelo instituto condutor; atributos da pesquisa escolhidos como: tipo de amostragem, tipo de pesquisa (telefone, face a face...),

elaboração do questionário, treinamento dos entrevistadores e etc. Dependendo da conduta, pode-se chegar mais ou menos próximo do resultado verdadeiro.

iii. Tamanho da amostra: como ensinado em qualquer aula básica de estatística, uma maior amostragem deveria retornar um menor erro.

O segundo grupo de fatores é composto por aqueles em que o pesquisador não tem como influenciar, e inclui:

i. Indecisão eleitoral: Maior indecisão eleitoral às vésperas da eleição causaria uma maior volatilidade nos resultados, podendo diminuir a precisão das pesquisas.
ii. Abstenção eleitoral: quanto maior a abstenção, supõe-se que um maior numero de pessoas que responderam o questionário não irão votar, causando volatilidade. Se a abstenção for mais forte pra um candidato que outro, o resultado será enviesado.
iii. Competitividade da eleição: Quanto mais perto está uma eleição maior a probabilidade de um determinado voto ser decisório, assim maior o incentivo a votar, reduzindo o poder dos fatores (i) e (ii).
iv. Multipartidarismo: Aqui interpretado como numero de candidatos viáveis (pois o Brasil sempre foi de sistema multipartidário). O efeito sobre a precisão deste fator é incerto, e Gramacho (2013) utiliza esta variável como uma forma de controle.
v. 2º Turno: No segundo turno os institutos podem revisar suas técnicas pelo feedback quanto à sua precisão nas pesquisas do 1º turno. Existe também o efeito do eleitor não só conhecer melhor suas opções, como também ter menos chance de se confundir ou errar o número, dado que no segundo turno os cargos a serem votados são no máximo dois, versus possivelmente seis no primeiro (CAVALLARI, 2010).
vi. Cargo: Quão mais importante for o cargo, maior atenção, tanto do eleitor como da mídia, a disputa irá chamar. Isto causa um maior conhecimento dos candidatos, aumentando a precisão.
vii. Desigualdade: Dependendo do contexto social que se trata a pesquisa (e.g. estados do país nas pesquisas de Governador), a população pode ter falta de acesso à informação e educação em todos os níveis, causando tomada de

decisões mais erráticas. Além disso, no caso das entrevistas, existem diversas regiões no Brasil onde o acesso é bastante restrito.

O artigo de Gramacho (2013) utiliza estes fatores mencionados acima e calcula uma média dos erros para cada um deles. Ele faz uma comparação entre alguns cenários, dente eles: a média dos erros de pesquisas efetuadas quando se tinha dois candidatos viáveis na disputa versus quando se tinha três, os erros presentes no primeiro turno comparados aos erros do segundo turno e os erros das pesquisas feitas pelo Datafolha, Vox Populi, IBOPE. Porém, Gramacho (2013) foca apenas nas eleições de Governador e Presidente ocorridas em 2010, o que significa que ele detém de um número relativamente baixo de observações para cada fator. Por exemplo, na avaliação de precisão entre institutos, existem apenas 13 observações para o Vox Populi. Ao utilizar um N tão baixo, o estudo conduzido por Gramacho (2013) pode também ter resultados enviesados, diminuindo a força de suas conclusões.

Por fim, após compreender melhor a análise existente referente à que fatores afetam os níveis de precisão das pesquisas eleitoras, pode se prosseguir para os métodos desenvolvidos e utilizados no passado para medição destes erros. Dentre os artigos publicados sobre o assunto, destaca se o escrito por Frederick Mosteller em 1949 após o fracasso das pesquisas na previsão da eleição do então Presidente Truman sobre Thomas Dewey. Neste artigo, feito também em livro, Mosteller (1949) destaca 8 métodos de análise dos erros, explicitados brevemente abaixo:

i. Estimação de erros de forma proporcional, em percentual.
ii. Estimação de erros de forma proporcional, em percentual considerando apenas os dois primeiros colocados.
iii. Estimação do erro em pontos percentuais; resultado esperado subtraído do resultado verdadeiro.
iv. Erro percentual médio.
v. Estimação de erros em pontos percentuais, considerando apenas os dois primeiros colocados.
vi. Erro máximo observado, em pontos percentuais.
vii. Utilização do teste qui-quadrado.
viii. Erro da taxa de participação.

Em face destes métodos apresentados, iremos nos basear no Método Mosteller 3, ou, MM3. Este é o mesmo método utilizado pelo Itaú (2018) e por Gramacho (2013), pois, de acordo com o próprio estudo de Mosteller (1949), é o de melhor encaixe com o sistema multipartidário.

$$Erro_{it} = \sum_{n=1}^{N} \frac{|Resultado_n - Pesquisa_{nit}|}{N}$$

Entretanto, como a estimação que será feita será ao nível do candidato, faremos modificações à formula no capítulo seguinte.

Resumindo o que foi exposto, é possível entender que a divergência entre a margem de erro divulgada e a margem real advém de calculá-la como se a amostragem fosse sempre aleatória, quando maioria das vezes não é o caso. Foi visto também que já existem estipulações embasadas quanto à quais poderiam ser os fatores geradores de erro, porém que a análise desses fatores ainda não foi efetuada profundamente, sobre uma janela adequada.

Deste modo, se torna clara a estimação desejada; desenvolver uma regressão qual seja capaz de estimar consistentemente os determinantes do erro verdadeiro de pesquisas eleitorais. A existência de uma regressão nestes moldes, desenvolvida a partir de toda base de dados existente para as pesquisas presidenciais desde o ano de 2002, tornaria muito mais fácil e acurada a interpretação das informações apresentadas pelos institutos em suas pesquisas eleitorais.

# 3. Metodologia

Conforme mencionado anteriormente, utilizamos como formula base a do MM3, porém faremos alterações para que seja possível à especificação do erro à nível do candidato. Além disso, quando um instituto divulga, em uma mesma apresentação de pesquisa eleitoral, resultados que utilizaram tanto o método de questionário espontâneo quanto o de estimulado, trataremos estes como duas pesquisas diferentes. Com isso, temos a seguinte formula para o erro:

$$Erro_{citra} = |Resultado_{ca} - Pesquisa_{citra}|$$

Onde "c" indica o candidato e "r" o tipo da pesquisa feita pelo instituto "i" no tempo "t" para a eleição do ano "a". Inclui-se aqui o termo "a" pois o mesmo candidato pode concorrer em mais do que uma eleição da base de dados.

Este "$Erro_{citra}$" é a variável explicada em nossa regressão; o quanto foi estimado versus qual foi o resultado de fato, em pontos percentuais, por candidato. E, uma vez que está definido qual será a variável dependente, podemos apresentar a equação da regressão a ser estimada, conforme segue abaixo:

$$\begin{aligned} Erro_{citra} = {} & \theta \mathrm{T}_{citra} + \sigma S_{citra} + \gamma E_{citra} + \omega_1 I_{Datafolha_{citra}} \\ & + \omega_2 I_{Voxpopuli_{citra}} + \ldots + \omega_{11} I_{Ibope_{citra}} + \beta_1 P_{1_{citra}} + \beta_2 P_{2_{citra}} \\ & + \tau M_{citra} + \delta_1 A_{2010_{citra}} + \delta_2 A_{2014_{citra}} + \delta_3 A_{2018_{citra}} + \varepsilon \end{aligned}$$

Onde:

T = Tempo até eleição (medido em dias), variável linear

S = Tamanho da amostra, variável linear.

E = Tipo de pesquisa, dummy com valor 1 caso seja espontânea

$I_x$ = Instituto onde "x" denota qual o instituto em questão. Variável dummy que assume valor 1 quando x = i

$P_1$ = Dummy de ideologia de partido com valor 1 caso o partido seja de esquerda, 0 caso contrário.

$P_2$ = Dummy de ideologia de partido com valor 1 caso o partido seja de direita, 0 caso contrário.

M = Dummy com valor 1 caso a pesquisa tenha sido efetuada pelo menos uma semana após o início do período de propaganda eleitoral.

$A_t$ = Dummy referente ao ano da eleição. Assume valor 1 quando t = a

Enfim, definida nossa regressão, podemos detalhar os aspectos e as hipóteses relativas as variáveis explicativas. Estas seriam:

1. Tempo. Certamente a variável com efeitos mais óbvios no processo de estimação de qualquer resultado. Espera-se que quanto maior for a distância entre a data de colheita de informações e a data da eleição, maior será o erro. A princípio, utilizaremos uma variável linear na estimação.

   Deve-se notar, porém, que o fator tempo é um que não necessariamente implica na má condução de uma pesquisa eleitoral, pois é natural esperar que as opiniões das pessoas podem mudar conforme são lançadas novas propagandas, ocorridos novos debates e efetuadas novas pesquisas a respeito dos candidatos.
2. Tamanho da amostra. Em termos estatísticos, uma maior amostra causa maior acurácia com tudo o mais constante. Porém, não é impossível que os critérios de seleção na definição da amostragem estejam estruturalmente falhos, de modo que não haja o devido aumento de precisão com o aumento na amostra. Seria surpreendente um resultado diferente de maior acurácia com maior amostra, mas vale a verificação. A estimação desta variável se dará de forma linear.
3. Tipo de pesquisa; espontânea versus estimulada. Neste caso será utilizada uma variável dummy com valor 1 caso a pesquisa seja espontânea e 0 caso seja estimulada. A razão para tal vem do fato de que a grande maioria das pesquisas presentes na base de dados são do tipo estimulado. Esta variável é uma com efeito bastante turvo; as eleições brasileiras são do tipo espontâneo (não são dadas as opções de voto quando se entra na urna), o que poderia indicar que as pesquisas efetuadas com o mesmo método seriam mais próximas da realidade. Porém, há de se imaginar que exista uma maior concentração de pesquisa e busca por informações dos candidatos mais perto do dia da eleição. Assim, um entrevistado pode não saber o nome de seu candidato para responder à pesquisa, mas saberia em quem votar (possivelmente devido ao partido) se tivesse os nomes à sua frente.
4. Instituto. Conforme foi visto nos capítulos anteriores, já existem estudos referentes ao que é chamado de “house effect”[6], como por exemplo o de

[6]Efeitos causados por vieses sistemáticos de um instituto. www.jota.info/dados/agregador-de-pesquisas/house-effects-institutos-pesquisas-19092018.

Gramacho (2013). Em primeira instância, não existem fatores estruturais para suportar que um instituto seria mais acurado que outro em relação às suas estimativas, porém é possível a existência de fatores não observáveis recorrentes presentes em uma casa, mas não em outra. Temos 12 institutos na base de dados, e para estimar estes fatores serão utilizadas onze variáveis dummy (o Datafolha será utilizado como instituto base).

5. Ideologia Partidária. Esta variável irá testar se partidos que se encontram à esquerda ou à direita do centro em seu alinhamento socioeconômico vis a vis a sociedade tem suas estimativas menos precisas. Será testado também se há algum viés erro preponderante frente a esquerda ou à direita. A intuição aqui não é clara, mas um cenário possível seria onde candidatos (e eleitores) considerados mais radicais (fora do centro) sejam mais resilientes em suas ideologias, e assim, tanto os partidos de esquerda quanto os de direita poderiam ter um menor nível de erro por serem mais previsíveis. Pode ser também que a tentativa de captura dos votos com a conversão ao centro por parte dos partidos fora dele faça com que estes tenham um erro maior em média. A estimação desta variável será feita através de duas dummies, uma com valor 1 caso o partido seja de direita e 0 caso contrário, e outra análoga para partidos de esquerda.

   A definição a respeito da posição ideológica de cada partido virá daqueles que costumam estar no centro do debate eleitoral de esquerda x direita; PT x PSDB. Vale notar aqui que o PSDB não é necessariamente um partido de direita, mas foi aquele que capturou grande parte deste eleitorado por fazer oposição ao PT nas urnas em todas as eleições presidenciais entre 1994 e 2014. O restante dos partidos será definido de acordo com a pesquisa de Power & Zucco (2009), e complementaremos a base de dados para os partidos recém-formados com base em suas próprias afirmações públicas a respeito de seu posicionamento. Abaixo a relação dos partidos e suas ideologias.

**Tabela 3.1**

| Partido | Posicionamento Ideológico |
|---|---|
| MDB | Centro |
| Patriota* | Centro |
| Podemos* | Centro |
| PRP* | Centro |
| PRTB* | Centro |
| PSC* | Centro |
| PV | Centro |
| REDE* | Centro |
| DC | Direita |
| Novo* | Direita |
| PSDB | Direita |
| PSL | Direita |
| PCB | Esquerda |
| PCO* | Esquerda |
| PDT | Esquerda |
| PPL* | Esquerda |
| PSB | Esquerda |
| PSOL | Esquerda |
| PSTU | Esquerda |
| PT | Esquerda |

*Partidos que não fazem parte do estudo de Power and Zucco (2009).

6. Período de Propaganda. Para esta variável será utilizado também uma dummy, com valor 1 caso a disputa já esteja há pelo menos uma semana em período de propaganda e valor 0 caso contrário. Supõe-se que após o início do período de propaganda os votos dos eleitores se tornem melhor informado, e assim apresente maior previsibilidade.
7. Ano da eleição. Novamente utilizaremos variáveis dummy para testar este efeito, com valor 1 caso a estimação seja para o ano em questão e zero caso contrário. O ano de 2006 será utilizado como base na regressão, e então teremos 3 dummies. A ideia desta variável é para testar a presença de tendências para as eleições específicas.

O modelo utilizado será aditivo linear. A estimação será feita através de mínimos quadrados ordinários e os valores dos coeficientes então representarão pontos percentuais; onde valores positivos denotam um erro verdadeiro maior, e negativos denotam um erro verdadeiro menor. Serão consideradas apenas as apresentações das pesquisas em votos totais.

Para efeito da estimação, será utilizada uma base de dados que contém todas as pesquisas de 1º turno efetuadas desde as eleições de 2002. Nesta base de dados temos informações sobre quando a pesquisa foi conduzida, quais eram os candidatos presentes e seus partidos, qual tipo do questionário (espontâneo vs. estimulado), quantas pessoas foram entrevistadas, qual instituto foi responsável e quem foi o contratante, além da margem de erro divulgada. Serão desconsideradas todas as pesquisas que utilizam cenários que possuem candidatos que acabaram não concorrendo. Faremos isso pois quando se tem opções não representativas das verdadeiras opções dos eleitores, suas escolhas de voto ficam claramente alteradas. Alguns exemplos mais proeminentes seriam as pesquisas efetuadas com questionários que ainda incluíam o então-candidato Eduardo Campos para as eleições de 2014 (candidatura qual teve de ser substituída após seu falecimento) e aquelas com Luís Inácio Lula da Silva que acabou tendo sua candidatura impugnada nas eleições de 2018. Outra exclusão que faremos será quando temos ausência de alguma variável estimada na regressão, o que, infelizmente, é o caso de todas as pesquisas eleitorais divulgadas para a eleição de 2002.

# 4. Apresentação dos Dados

A base de dados utilizada, conforme já mencionado, virá a partir do banco de dados da instituição Poder360 que fica disponível em seu website[7]. Neste banco de dados são apresentadas informações sobre todas as pesquisas registradas junto ao TSE desde 2002. Porém como já previamente mencionado, devido à ausência da quantidade de entrevistados (uma de nossas variáveis explicativas) para as pesquisas efetuadas para a eleição de 2002, utilizaremos em nossa base de dados apenas as pesquisas a partir da eleição de 2006. O resultado disso é uma base de dados com um total de 324 pesquisas eleitorais.

Entretanto, para efeitos de nossa regressão, o fator limitante não é o número de pesquisas. Cada cenário estipulado e cada variação no tipo (espontâneo ou estimulado) será considerado individualmente. Assim, 324 não é o número que utilizaremos como referência para a quantidade de pesquisas, mas sim 884.

Contudo, ainda não chegamos na base de dados sobre a qual iremos rodar nossa regressão. A razão para tal vem do fato de que pesquisas eleitorais são conduzidas desde muito tempo antes da eleição; em muitos casos temos pesquisas sendo conduzidas até mesmo fora sequer do ano da eleição. O quão preditiva é uma pesquisa eleitoral feita com tanto tempo de antecedência é algo para ser discutido no próximo capítulo, mas temos que muitas dessas 884 pesquisas foram efetuadas com cenários considerando candidatos que nem chegam a concorrer. Nesses cenários, não podemos atribuir o erro entre a pesquisa e o resultado final a nenhum fator palpável, pois como as opções do questionário não eram as mesmas do dia do voto, os resultados apresentados tratam de um cenário fictício. Deste modo, todas as pesquisas considerando cenários fictícios serão desconsideradas da análise.

Assim, após este filtro, chegamos a nossa base de dados final. As 324 pesquisas gerais que tínhamos anteriormente se tornam 193 e as 884 considerações individuais caem drasticamente para 226. O motivo dessa diferença de 33 pesquisas vem do fato de que alguns cenários omitiam candidatos que chegaram a concorrer, mas sem adicionar nenhum que de fato concorreu. Nos casos em que o candidato omitido em nenhum momento teve intenções de voto superiores à 0,20% os cenários foram considerados validos.

[7] https://pesquisas.poder360.com.br/

Com isso, a base de dados a ser utilizada tem as propriedades como demonstrado na tabela abaixo:

**Tabela 4.1**

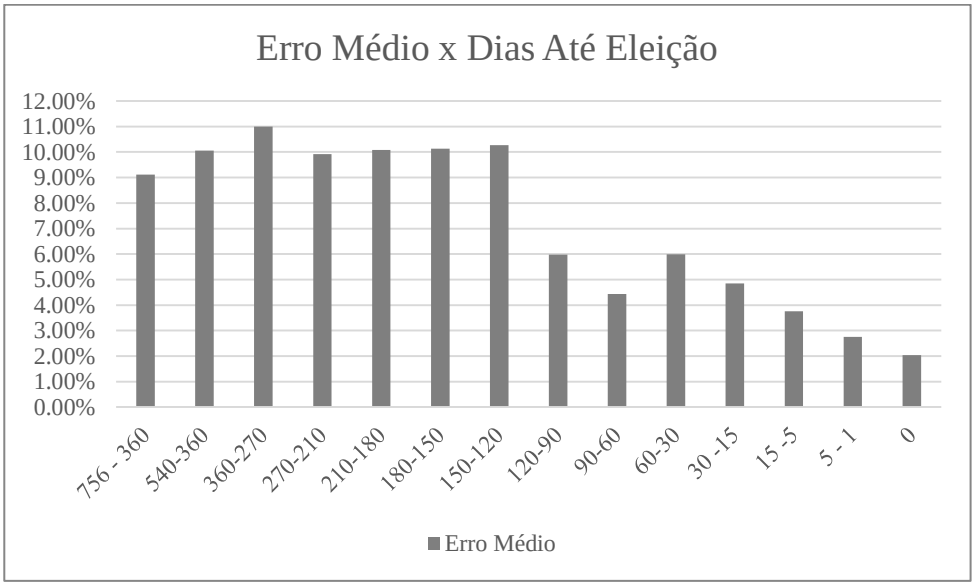
| Indicador | Base de dados Filtrada |
|---|---|
| Total | 226 |
| Espontânea | 26 |
| Estimulada | 200 |
| 2006 | 21 |
| 2010 | 75 |
| 2014 | 67 |
| 2018 | 63 |
| Fora do Período de Propaganda - 2006 | 6 |
| Fora do Período de Propaganda - 2010 | 46 |
| Fora do Período de Propaganda - 2014 | 22 |
| Fora do Período de Propaganda - 2018 | 22 |
| Tamanho Médio de Amostra | 2940 |
| Margem de Erro Média | 2.3% |
| Erro Verdadeiro Médio | 5.2% |
| Média de Dias Até Eleição | 110 |
| Ibope | 56 |
| Vox Populi | 52 |
| Datafolha | 48 |
| MDA | 18 |
| Sensus | 13 |
| DataPoder360 | 10 |
| Ipespe | 10 |
| FSB | 6 |
| Paraná Pesquisas | 4 |
| Real Time Big Data | 4 |
| Virtu Analise | 4 |
| Verita | 1 |

As 226 pesquisas têm um número médio de entrevistados de 2940 pessoas com uma margem de erro apresentada média de 2,3% e o tempo médio até eleição de 110 dias, quase 4 meses. Porém, observamos também uma maior concentração de pesquisas por volume conforme nos aproximamos da eleição. Dentre as 226 pesquisas que iremos utilizar, mais da metade (116) são conduzidas no último mês.

O erro verdadeiro médio é de 5,2%. Entretanto, antes que possamos concluir que as pesquisas tem um alto grau de imprecisão, devemos considerar que existem pesquisas na amostra no qual os erros chegam a níveis superiores à 20% (em geral são pesquisas efetuadas com um ano ou mais de antecedência que, parcialmente por sorte, usaram o

cenário correto) puxando para cima esse erro. A figura 4.1 mostra o valor do erro médio atribuído às pesquisas eleitorais por período no tempo.

**Figura 4.1**




Vemos que a tendência de queda no erro conforme este se aproxima da eleição é clara, e possivelmente não linear; ela decai mais rapidamente próxima da eleição. Em uma pesquisa divulgada pelo Datafolha a respeito da tendência das decisões de voto ao longo do tempo[8], foi afirmado que 36% da população decide seu voto nos últimos 15 dias, sendo impressionantes 12% no dia da eleição. Estes dados corroboram fortemente com o observado no gráfico; parcelas muito relevantes de eleitores decidem seus candidatos apenas na reta final, aumentando fortemente a previsibilidade das pesquisas eleitorais. Importante ressaltar também que as pesquisas de boca de urna (no dia da eleição) tem um erro médio de tamanho igual à margem de erro comumente divulgada. Isto é extremamente relevante, pois sugere que as pesquisas de fato fazem boa amostragem, e são capazes de representar adequadamente a população como um todo no instante de sua conduta.

Quanto à distribuição das pesquisas por ano, um ponto a se notar é o numero bastante reduzido da quantidade de pesquisas efetuadas para o ano de 2006; 21 versus as 60-80 disponíveis para os demais anos. Um dos fatores nesta discrepância veio a partir da candidatura de Rui Costa Pimenta, que foi indeferida pelo TSE devido à problemas na prestação de contas de seu partido, PCO (Partido da Causa Trabalhadora). A questão aqui foi que o partido recorreu, e assim foi permitido manter sua campanha eleitoral até o

[8]https://g1.globo.com/politica/eleicoes/2018/eleicao-em-numeros/noticia/2018/10/10/12-dos-eleitores-decidiram-o-voto-para-presidente-no-dia-da-eleicao-diz-datafolha.ghtml.

ultimo mês, quando esta foi de fato barrada. Isso causou a inclusão de Rui Costa Pimenta em muitas pesquisas; pesquisas quais foram descartadas pelas razões mencionadas anteriormente.

Existe também uma concentração frente o tipo do questionário. Em nossa amostra, temos 200 pesquisas do tipo estimulada frente apenas 26 do tipo espontânea (11,5%). Um possível fator para essa discrepância pode ser que as pesquisas feitas de forma espontânea tiveram como resposta candidatos que não concorreram, sendo assim eliminadas na filtragem. Comparando com a base de 904 pesquisas que tínhamos anteriormente, vemos que nela 755 eram do tipo estimulada versus 149 espontânea (16,4% do total); ou seja, por mais que a filtragem possa ter sido um fator, de fato pesquisas do tipo estimulada são muito mais comuns. Observa-se também que as pesquisas feitas com o questionário do tipo Espontâneo têm um erro médio bastante elevado frente às demais; 7,0% para as espontâneas contra 5,0% para as estimuladas. Considerando o que foi dito acima a respeito da concentração de decisão do eleitor muito próxima do dia da eleição, este resultado faz sentido; a ausência de apresentação das opções à um eleitor que ainda não sabe em quem vai votar irá naturalmente gerar mais erro, mesmo que seja por simples falta de recordação a respeito de quem ele pode votar.

Em relação aos institutos, são 12 ao todo. Contudo, neste quesito também há uma grande concentração; Ibope, Datafolha e Vox Populi são responsáveis sozinhos por 156 das 226 pesquisas. O instituto Verita tem apenas uma observação, e FSB, Paraná Pesquisas, Real Time Big Data e Virtu Analise também não chegam à 10 pesquisas cada. Este baixo número de amostras pode se provar um problema na regressão. Comparando os erros médios dos três maiores institutos, dos quais temos observações suficiente para analisar, vemos que o erro médio do Datafolha é 4,6%, do Ibope é 5,0% e do Vox Populi é 6,90%. Sem os demais dados relativos às pesquisas (como quanto tempo antes da eleição elas foram conduzidas) não podemos tomar conclusões definitivas, mas a principio o Vox Populi parece apresentar um erro elevado.

Na questão dos candidatos viáveis não há muito o que se analisar, utilizando aqui o nível de 2% no resultado final como determinante do que era "viável" de voto, temos que 2006, 2010, 2014 e 2018 tiveram 4, 3, 3 e 5 candidatos viáveis em suas eleições, respectivamente.

Dentre nossas observações, temos que, com exceção de 2010, a proporção de pesquisas efetuadas até menos de uma semana depois do início do período de propaganda (a variável usada na regressão) gira em torno de 30%. Por razão ainda desconhecida, a

eleição de 2010 teve 61% de suas pesquisas fora do período de propaganda, o dobro do comum.

**Tabela 4.2**

| Eleição | Último Dia de Registro | Início da Propaganda | Data da Eleição |
|---|---|---|---|
| 2006 | 05-Jul-06 | 15-Aug-06 | 04-Oct-06 |
| 2010 | 05-Jul-10 | 17-Aug-10 | 03-Oct-10 |
| 2014 | 05-Jul-14 | 19-Aug-14 | 05-Oct-14 |
| 2018 | 15-Aug-18 | 31-Aug-18 | 07-Oct-18 |

Ademais, conforme era o esperado, as pesquisas feitas a partir de uma semana depois do inicio do período de propaganda apresentam um erro menor; 4,0% versus 7.7%. Nesta variável temos também influência do fator tempo, então não é possível associar desde já esta diferença às propagandas eleitorais (isso só poderá ser feito quando rodarmos a regressão).

Contudo, essas 226 pesquisas também não são exatamente o objeto de nossa regressão. Lembremos a formula apresentada anteriormente;

$$Erro_{citra} = |Resultado_{ca} - Pesquisa_{citra}|$$

Assim, vemos que cada estimativa para a porcentagem de intenção de voto apresentada para um candidato será tratada como uma observação.

O resultado disso é a utilização de 1360 observações em nossa regressão, mitigando fortemente qualquer problema associado ao baixo número de observações na amostra (este 1360 desconsidera todas as estimativas para votos brancos, nulos e afins). Ainda assim, devemos considerar que nos casos como o do instituto Verita, o problema persiste, pois de 1 observação passamos a ter 13; número ainda bastante baixo.

Em vista do acima exposto, podemos finalmente tratar da variável de alinhamento ideológico do partido, pois esta se aplica de candidato em candidato. Das 1360 estimativas presentes na base de dados, 579 são de esquerda, 364 são de centro e 417 são de direita. Observando os erros médios dessa variável temos que os erros apresentados para o Centro são notavelmente menores, com os da direita consideravelmente maiores que os demais.

**Tabela 4.3**

| Alinhamento | Erro Médio |
|---|---|
| Esquerda | 4.36% |
| Centro | 3.77% |
| Direita | 7.67% |

# 5. Analise dos Resultados

## 5.1. Método 1 – Modelo Apresentado

Uma vez que temos nossa base de dados adequadamente filtrada, podemos finalmente rodar nossa regressão e analisar seus resultados.

Em nossa base de dados aplicamos também o filtro sobre qualquer outra pesquisa que possua sua quantidade de entrevistados igual à zero, para manter a consistência frente nossa eliminação de todas as pesquisas da eleição de 2002. Devemos lembrar aqui também que não consideramos em nosso modelo as estimativas para os níveis de votos brancos, nulos e afins.

Com isso, temos os nossos resultados através do que chamaremos de Método 1; aquele já explicado profundamente nas paginas anteriores. Chamaremos este de Método 1 pois mais adiante iremos fazer alterações aos modelos para melhor testar a robustez de nossas estimativas. Na tabela 5.1, as colunas numeradas de 1 a 7 denotam a quantidade de variáveis do modelo inclusas naquelas regressões, sendo a coluna 7 nosso modelo completo.

A partir da tabela vemos primeiramente que a base de dados do modelo completo, após as linhas descartadas, possui 1235 observações.

Quantidade de entrevistas. Variável essencialmente nula e insignificante em todas as regressões apresentadas. Abrindo as casas decimais, vemos que o coeficiente aqui foi de -0.000063 pontos percentuais. Ou seja, são necessários 10.000 entrevistados para uma redução no erro de 0,6 p.p. (a média de entrevistados é de 2783), e mesmo assim o valor da estatística t não chega nem à 1. Este resultado é certamente inesperado pois as margens de erro apresentadas pelos institutos são baseadas exatamente neste quesito. Mesmo que, conforme dito anteriormente, este método de cálculo para a margem de erro não seja correto no caso de amostragem por quotas, seria esperado pelo menos um resultado não-nulo para esta variável. Uma possível causa para isso pode ser o fato de que nenhuma das pesquisas foi efetuada com um número excessivamente baixo de entrevistados, de modo que todas teriam um grau de previsibilidade dos resultados baseado na amostragem similar. Este resultado tem uma forte implicação: como não faz sentido continuar estimando o modelo com esta variável presente, visto que ela claramente não tem participação na explicação do erro, podemos reincluir as observações descartadas por ausência de dados para a variável de quantidade de entrevistas. Para este efeito,

regredimos a variável explicada novamente, excluindo esta variável; demonstrado pela coluna 8. Este modelo tem 1360 observações e a partir deste ponto iremos considera-lo como o "modelo base".

Tempo até a Eleição. Variável positiva e significante desde sua introdução. Esta variável demonstra uma queda em seu coeficiente conforme adicionamos novas variáveis explicativas ao modelo, mas seu resultado é bastante consistente. Temos que a cada dia que nos distanciamos da eleição vemos um aumento no erro de 0,01 pontos percentuais nas regressões completas (colunas 7 e 8) e 0,02 p.p. nas regressões que possuem de 1 até 6 variáveis. Este valor, quando comparado a figura 4.1 do capítulo anterior, parece ser bastante reduzido. Uma possível causa para isso pode vir a partir das observações muito distantes da eleição. No gráfico 4.1 vemos que após uma determinada quantidade de dias (possivelmente por volta de 120) o erro deixa de aumentar com a distância da eleição. Na próxima seção iremos restringir a base de dados para períodos mais próximos das eleições.

Tipo de questionário. Em relação à esta variável vemos que as pesquisas do tipo estimulada tem um erro significativamente menor frente aquelas do tipo espontânea. Esta redução no erro gira em torno de 3 p.p. com alto grau de significância a todo momento, atingindo um ápice de 3,7 p.p. no modelo da coluna (6) e terminando em 2,7 p.p. no modelo base. Na apresentação desta variável foi mencionado seu fator ambíguo; as eleições brasileiras de fato são do tipo espontâneo, mas devido à ausência de informações até os períodos próximos da eleição os cidadãos poderiam não saber muito de suas opções para responder ao questionário devidamente. Os modelos apresentados até agora certamente indicam que este segundo fator possui peso maior na questão. Posteriormente, será rodada a regressão limitando a quantidade de dias até eleição; será interessante ver como será a reação desta variável.

Institutos. O instituto ocultado na regressão, aquele usado como base, foi o Datafolha, e temos que de uma maneira geral não há fortes tendências para concluirmos que um instituto erra mais que outro consistentemente. Os únicos coeficientes significantes foram do DataPoder360 no modelo completo e do VirtuAnálise antes da inclusão das variáveis de ano. Observando apenas o modelo base, temos que o instituto DataPoder360 erra 2,9 p.p. a mais que o Datafolha.

Alinhamento. A variável de alinhamento é uma com resultados significantes e consistentes em todos os modelos apresentados, com a observação de que o grau de significância do alinhamento à esquerda cai para o nível de 10% no modelo base. De

maneira geral, as estimações para candidatos de partidos de esquerda erram por volta de 1% a mais frente candidatos de centro, e aquelas para candidatos de partidos de direita; 4%. No modelo base estes números são de 0,7 p.p. e 4,1 p.p. respectivamente. Na apresentação desta variável foi mencionado que seu efeito era intuitivamente turvo, mas que poderia se aplicar tanto como um fator geral sobre o fato de os partidos estarem longe do centro (serem mais radicais) como um fator especificamente sobre alguma das ideologias. Nossos resultados apresentados na tabela indicam que de fato ambos os efeitos existem e, líquidos, são positivos. Uma questão interessante de se aprofundar seria a respeito do vetor destes erros; eles estariam sendo gerados por alguma tendência de superestimação? Subestimação? Ou será que estes erros são simplesmente escalar, sem apresentar um viés frente alguma ideologia? Infelizmente, essas são informações que vão além do escopo deste estudo.

Propaganda. Conforme o esperado, temos que as pesquisas efetuadas a partir de uma semana após o início do período de propaganda tem um erro 2,1 p.p. menor do que aquelas efetuadas fora do período. Estes resultados sugerem que a permissão à propaganda eleitoral aumenta o nível de conhecimento da população a respeito de seus candidatos, reduzindo a volatilidade dos resultados da eleição.

Ano da eleição. Estas variáveis dummy atuam de certa forma como constantes específicas a cada ano, e tem coeficientes significativos e positivos. Como o ano de 2006 é utilizado como base, e então omitido, o fato destes coeficientes serem positivos significa que as pesquisas eleitorais de 2006 foram as que apresentaram menor erro.

# Tabela 5.1

| | Erro Verdadeiro | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | Base |
| dias_ate_eleicao | 0.0002*** | 0.0002*** | 0.0002*** | 0.0002*** | 0.0002*** | 0.0002*** | 0.0001*** | 0.0001*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| qtd_entrevistas | | 0 | 0 | 0 | 0 | 0 | 0 | |
| | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| tipoEstimulada | | | -0.031*** | -0.036*** | -0.036*** | -0.037*** | -0.028*** | -0.027*** |
| | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoDataPoder360 | | | | 0.012 | 0.013 | 0.013 | 0.021** | 0.029*** |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoFSB | | | | 0.009 | 0.009 | 0.007 | 0.011 | 0.013 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoIbope | | | | -0.002 | -0.003 | -0.004 | -0.002 | -0.0002 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoIpespe | | | | -0.005 | -0.005 | -0.009 | -0.004 | -0.002 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoMDA | | | | -0.004 | -0.003 | -0.004 | -0.004 | -0.002 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoParana Pesquisas | | | | -0.006 | -0.006 | -0.006 | -0.002 | 0.0001 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoReal Time Big Data | | | | 0.011 | 0.009 | 0.004 | 0.007 | 0.007 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoSensus | | | | -0.01 | -0.01 | -0.009 | -0.004 | -0.001 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoVerita | | | | -0.019 | -0.019 | -0.017 | -0.015 | -0.015 |
| | | | | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| institutoVirtu Analise | | | | -0.025** | -0.025** | -0.023** | -0.014 | -0.01 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoVox Populi | | | | 0.001 | 0.0003 | 0.001 | -0.003 | -0.004 |
| | | | | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| alinhamentoDireita | | | | | 0.043*** | 0.043*** | 0.045*** | 0.041*** |
| | | | | | (0.01) | (0.01) | (0.01) | (0.00) |
| alinhamentoEsquerda | | | | | 0.010** | 0.010** | 0.011*** | 0.007* |
| | | | | | (0.00) | (0.00) | (0.00) | (0.00) |
| propaganda | | | | | | -0.018*** | -0.017*** | -0.021*** |
| | | | | | | (0.01) | (0.01) | (0.00) |
| as.factor(ano)2010 | | | | | | | 0.040*** | 0.043*** |
| | | | | | | | (0.01) | (0.01) |
| as.factor(ano)2014 | | | | | | | 0.033*** | 0.035*** |
| | | | | | | | (0.01) | (0.01) |
| as.factor(ano)2018 | | | | | | | 0.019*** | 0.018*** |
| | | | | | | | (0.01) | (0.01) |
| Constant | 0.039*** | 0.041*** | 0.068*** | 0.078*** | 0.061*** | 0.077*** | 0.043*** | 0.046*** |
| | -0.002 | -0.003 | -0.006 | -0.009 | -0.009 | -0.01 | -0.012 | -0.01 |
| Observations | 1,235 | 1,235 | 1,235 | 1,235 | 1,235 | 1,235 | 1,235 | 1,360 |
| $R^2$ | 0.092 | 0.093 | 0.111 | 0.122 | 0.19 | 0.198 | 0.22 | 0.218 |
| Adjusted $R^2$ | 0.092 | 0.092 | 0.109 | 0.112 | 0.179 | 0.187 | 0.207 | 0.207 |
| Residual Std. Error | 0.063 (df = 1233) | 0.063 (df = 1232) | 0.063 (df = 1231) | 0.063 (df = 1220) | 0.060 (df = 1218) | 0.060 (df = 1217) | 0.059 (df = 1214) | 0.058 (df = 1340) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

## 5.2. Método 2 – Restringindo os períodos das pesquisas

Apesar dos resultados satisfatórios com as regressões apresentadas nas seções anteriores, existem ainda alguns pontos quais podem render frutos nos aprofundar. Um destes pontos é referente à variável de dias até a eleição. Como mencionado na apresentação dos dados, existem algumas pesquisas que são conduzidas centenas de dias antes da eleição. Estas pesquisas costumam utilizar de diversos cenários diferentes com diversos candidatos diferentes. Em geral, a maioria destas pesquisas foi excluída quando tiramos aquelas que apresentavam candidatos que acabaram por não concorrer, mas não todas. Na realidade, muitas pesquisas desse tipo ainda estão presentes na base de dados utilizada, e devemos nos perguntar qual o impacto delas sobre a regressão. Afinal, mal se houve falar sobre eleições até pelo menos poucos meses antes da mesma; o quão relevante é para o objetivo deste estudo incluir pesquisas conduzidas com tanta antecedência?

Em virtude disso, rodaremos dois novos modelos; um no qual utilizaremos apenas as pesquisas efetuadas quando se faltava menos de 90 dias até a eleição e outro mais extenso no qual o limite será 150 dias. A escolha do numero 90 vem a partir deste ser a quantidade usual de dias antes da eleição no qual se fecha o período de registro das candidaturas (sendo o ano de 2018 uma exceção, fechando o registro à apenas 52 dias da eleição). Lembremos o cronograma:

**Tabela 5.2**

| Eleição | Ultimo Dia de Registro | Inicio da Propaganda | Data da Eleição |
|---|---|---|---|
| 2006 | 05-Jul-06 | 15-Aug-06 | 04-Oct-06 |
| 2010 | 05-Jul-10 | 17-Aug-10 | 03-Oct-10 |
| 2014 | 05-Jul-14 | 19-Aug-14 | 05-Oct-14 |
| 2018 | 15-Aug-18 | 31-Aug-18 | 07-Oct-18 |

É natural pensar também que uma vez definidas as candidaturas, há um aumento na conversa cotidiana a respeito do tema, e o inicio do período de maior especulação a respeito do resultado. Por outro lado, a escolha da limitação em 150 vem a partir da figura 4.1, onde vemos que este foi o ponto de corte para o aumento do erro frente a quantidade de dias até a eleição. Os resultados das regressões nestes novos moldes estão apresentados na tabela 5.3.

Com as novas restrições, temos algumas alterações significantes em nossos resultados. A quantidade de observações foi pouco reduzida; de 1360 para 1189 quando restrito à 150 dias (coluna 2) e 1137 quando restrito à 90 dias (coluna 3), reforçando o fato de que há concentração de pesquisas próximo a eleição.

Além disso, imediatamente, observamos que em ambos os casos o coeficiente da regressão de dias até eleição foi elevado. O coeficiente, que antes era de 0,01 p.p. ao dia no modelo da coluna 2, se multiplicou por seis para 0,06 p.p. ao dia. Enquanto isso, no modelo restrito a 90 dias antes da eleição, este coeficiente foi multiplicado pela ordem de 10, se tornando 0,1 pontos percentuais para cada dia que se distancia das eleições.

Quanto as demais variáveis, temos que as únicas que apresentaram mudanças notáveis foram Institutos, Alinhamento Esquerda (apenas ganhou nível de significância) e Propaganda. Houve também variação substancial nas constantes, que se tornaram não-significantes em ambos os modelos e até negativa no modelo da coluna 3. Possivelmente seus efeitos passaram a ser capturados pela variável de dias até a eleição.

Em relação aos institutos, vemos que o DataPoder360 perdeu significância e valor em seu coeficiente com as restrições aos dias, chegando até a ser não-significante. Isto sugere que as pesquisas do instituto conduzidas longe das eleições estavam puxando seu erro para cima. Por outro lado, o coeficiente do instituto FSB ganhou significância, indicando a possibilidade de um caso inverso ao DataPoder360.

Por último, e talvez mais curioso, a variável de "propaganda" apresentou variações inesperadas; ela passou de ser negativa para ser positiva e deixou de ser significante para o modelo restrito à 150 dias. No modelo restrito a 90 dias ela caiu um grau de significância. É possível que isto seja efeito simplesmente da janela, mas também não devemos desconsiderar a possibilidade de ela estar capturando parte do efeito da variável de dias até a eleição nos modelos sem a restrição aos dias. Visto a enorme variação no coeficiente desta variável quando aplicadas as restrições, essa seria uma conclusão bastante plausível.

## Tabela 5.3

| | Base | Erro Verdadeiro (2) | (3) |
|---|---|---|---|
| dias_ate_eleicao | 0.0001*** | 0.0006*** | 0.0009*** |
| | (0.00) | (0.00) | (0.00) |
| tipoEstimulada | -0.027*** | -0.027*** | -0.027*** |
| | (0.01) | (0.01) | (0.01) |
| institutoDataPoder360 | 0.029*** | 0.007 | 0.007 |
| | (0.01) | (0.01) | (0.01) |
| institutoFSB | 0.013 | 0.016* | 0.015* |
| | (0.01) | (0.01) | (0.01) |
| institutoIbope | -0.0002 | 0.002 | 0.002 |
| | (0.01) | (0.01) | (0.01) |
| institutoIpespe | -0.002 | 0.003 | 0.003 |
| | (0.01) | (0.01) | (0.01) |
| institutoMDA | -0.002 | -0.004 | -0.007 |
| | (0.01) | (0.01) | (0.01) |
| institutoParana Pesquisas | 0.0001 | 0.0004 | -0.001 |
| | (0.01) | (0.01) | (0.01) |
| institutoReal Time Big Data | 0.007 | 0.01 | 0.008 |
| | (0.01) | (0.01) | (0.01) |
| institutoSensus | -0.001 | -0.009 | -0.009 |
| | (0.01) | (0.01) | (0.01) |
| institutoVerita | -0.015 | -0.01 | -0.007 |
| | (0.02) | (0.02) | (0.02) |
| institutoVirtu Analise | -0.01 | -0.012 | -0.012 |
| | (0.01) | (0.01) | (0.01) |
| institutoVox Populi | -0.004 | 0.003 | 0.003 |
| | (0.01) | (0.01) | (0.01) |
| alinhamentoDireita | 0.041*** | 0.047*** | 0.049*** |
| | (0.00) | (0.00) | (0.00) |
| alinhamentoEsquerda | 0.007* | 0.009** | 0.008** |
| | (0.00) | (0.00) | (0.00) |
| propaganda | -0.021*** | 0.009 | 0.015** |
| | (0.00) | (0.01) | (0.01) |
| as.factor(ano)2010 | 0.043*** | 0.038*** | 0.047*** |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2014 | 0.035*** | 0.044*** | 0.047*** |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2018 | 0.018*** | 0.025*** | 0.027*** |
| | (0.01) | (0.01) | (0.01) |
| Constant | 0.046*** | 0.0003 | -0.012 |
| | (0.01) | (0.01) | (0.01) |
| Observations | 1,360 | 1,189 | 1,137 |
| $R^2$ | 0.218 | 0.21 | 0.213 |
| Adjusted $R^2$ | 0.207 | 0.198 | 0.2 |
| Residual Std. Error | 0.058 (df = 1340) | 0.055 (df = 1169) | 0.054 (df = 1117) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

## 5.3. Método 3 – Testando a linearidade da variável de dias até a eleição

O comportamento da variável de dias até a eleição apresentado na seção anterior unido aos que foi possível observar na figura 4.1 nos faz pensar sobre a possibilidade desta variável ter comportamento não linear. Afinal, há claros indicativos que de o efeito do distanciamento da eleição é mais forte nos dias mais próximos do que nos mais distantes, onde há uma estagnação do nível de erro. Desta forma, esperamos que o coeficiente da variável quadrática será negativo.

Com isso em mente, rodamos duas novas regressões, em uma adicionamos a variável de dias até eleição ao quadrado. Na outra adicionamos tanto dias até eleição ao quadrado quanto dias até eleição ao cubico. Os resultados são apresentados na tabela 5.4.

Imediatamente, observamos que ambas as novas variáveis (representadas por dias2 e dias3; quadrática e cúbica respectivamente) são significantes ao nível de 1%. Além disso, temos que, conforme o esperado, a variável quadrática é negativa.

Outro ponto interessante é que agora que temos o efeito redutor quando distantes da eleição, os coeficientes lineares tiveram aumentos substanciais. O efeito da adição da variável quadrática foi multiplicar o coeficiente linear por 4, e quando adicionamos a variável cubica o efeito foi de uma multiplicação por 12. Estes resultados em muito são similares aos apresentados na seção anterior, indicando que de fato o coeficiente da estimação linear estava sendo penalizado pela inclusão de dias distantes, sem coeficientes polinomiais para compensar o efeito.

Todos os demais coeficientes mantiveram comportamentos similares ao do modelo base, com exceção de alinhamento a esquerda, propaganda e a constante, todos no modelo da coluna 3. Estes três coeficientes se tornaram não significantes, tendo o de propaganda novamente se tornado positivo.

## Tabela 5.4

| | Base | Erro Verdadeiro (2) | (3) |
|---|---|---|---|
| dias_ate_eleicao | 0.000065478*** | 0.000272024*** | 0.000807076*** |
| | (0.00) | (0.00) | (0.00) |
| dias2 | | -0.000000323*** | -0.000002340*** |
| | | (0.00) | (0.00) |
| dias3 | | | 0.000000002*** |
| | | | 0.00 |
| tipoEstimulada | -0.0266*** | -0.0289*** | -0.0294*** |
| | (0.01) | (0.01) | (0.01) |
| institutoDataPoder360 | 0.0287*** | 0.0217** | 0.0219** |
| | (0.01) | (0.01) | (0.01) |
| institutoIbope | -0.0002 | 0.0005 | 0.0011 |
| | (0.00) | (0.00) | (0.00) |
| institutoIpespe | -0.0016 | 0.0007 | 0.0036 |
| | (0.01) | (0.01) | (0.01) |
| institutoMDA | -0.0017 | -0.0023 | -0.0016 |
| | (0.01) | (0.01) | (0.01) |
| institutoParana Pesquisas | 0.0001 | 0.0005 | 0.0003 |
| | (0.01) | (0.01) | (0.01) |
| institutoReal Time Big Data | 0.0069 | 0.0090 | 0.0098 |
| | (0.01) | (0.01) | (0.01) |
| institutoSensus | -0.0013 | -0.0026 | -0.0034 |
| | (0.01) | (0.01) | (0.01) |
| institutoVerita | -0.0151 | -0.0137 | -0.0084 |
| | (0.02) | (0.02) | (0.02) |
| institutoVirtu Analise | -0.0097 | -0.0119 | -0.0132 |
| | (0.01) | (0.01) | (0.01) |
| institutoVox Populi | -0.0043 | -0.0005 | 0.0016 |
| | (0.01) | (0.01) | (0.01) |
| alinhamentoDireita | 0.0410*** | 0.0404*** | 0.0401*** |
| | (0.00) | (0.00) | (0.00) |
| alinhamentoEsquerda | 0.0071* | 0.0066* | 0.0064 |
| | (0.00) | (0.00) | (0.00) |
| propaganda | -0.0213*** | -0.0085* | 0.0094 |
| | (0.00) | (0.01) | (0.01) |
| as.factor(ano)2010 | 0.0432*** | 0.0399*** | 0.0409*** |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2014 | 0.0347*** | 0.0335*** | 0.0378*** |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2018 | 0.0179*** | 0.0197*** | 0.0242*** |
| | (0.01) | (0.01) | (0.01) |
| Constant | 0.0463*** | 0.0323*** | 0.0043 |
| | (0.01) | (0.01) | (0.01) |
| Observations | 1,360 | 1,360 | 1,360 |
| $R^2$ | 0.218448 | 0.2314144 | 0.2425705 |
| Adjusted $R^2$ | 0.2073663 | 0.2199344 | 0.2306826 |
| Residual Std. Error | 0.057793030 (df = 1340) | 0.057333020 (df = 1339) | 0.056936660 (df = 1338) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

## 5.4. Método 4 – Restringindo apenas para candidatos viáveis

A última alteração que faremos ao nosso modelo base será a restrição da base de dados apenas a candidatos viáveis. A razão para tal vem do fato de que muitos candidatos que aparecem com estimativas em nenhum momento atingem sequer 1% de intenção de votos. Uma vez que o candidato nunca esteve muito na consideração do eleitor, não há como o erro de sua estimativa ser muito diferente de 0. Sendo assim, há uma possibilidade muito grande destes candidatos estarem enviesando os coeficientes para baixo, como evidenciado pela tabela 5.5:

**Tabela 5.5**

| Base de dados | Erro Médio |
|---|---|
| Base de dados - Filtro Qtd. Entrevistas | 5.05% |
| Base de dados - Completa | 5.21% |
| Base de dados – Filtro 90 Dias até Eleição | 4.30% |
| Base de dados - Filtro Viáveis | 7.95% |
| Base de dados - Filtro 90 Dias + Viáveis | 7.22% |

A comparação nesta seção será feita utilizando tanto a base de dados completa (coluna 2), quanto a restrição de 90 dias até a eleição (coluna 3), porém iremos excluir todos os candidatos que não atingiram 2% dos votos totais em suas respectivas eleições. O efeito desta restrição é causar com que haja bem menos candidatos por eleição, conforme a tabela 5.6 abaixo.

**Tabela 5.6**

| Eleição | Candidatos Viáveis |
|---|---|
| 2006 | 4 |
| 2010 | 3 |
| 2014 | 3 |
| 2018 | 5 |

As novas regressões são apresentadas na tabela 5.7 e imediatamente observamos que muita coisa mudou, especialmente para o modelo da coluna 2. Neste modelo, diversos coeficiente como o de dias até eleição, DataPoder360, alinhamento a esquerda e as dummies para o ano de 2010 e 2018 se tornaram não significantes. O $R^2$ caiu 6% e as observações foram reduzidas em 495. Claramente, as interações para as observações restantes foram anormais quando comparadas aos outros modelos que utilizamos, sendo

grande parte de seus efeitos agora capturados na constante com valor de 10,2 pontos percentuais de erro.

O modelo da coluna 3, restrito a observações com 90 dias até a eleição, também sofreu alterações na nova especificação, porém em menor escala. O $R^2$ se manteve essencialmente não alterado, mas suas principais mudanças foram as quedas nas significâncias dos coeficientes das variáveis de propaganda, FSB, alinhamento à esquerda e ano de 2010. Fora estas alterações, não houve movimentos notáveis.

Deste modo, apesar de a remoção dos candidatos inviáveis fazer sentido conceitualmente, vemos que esta remoção em muito pouco agrega conhecimento sobre os fatores de erro.

# Tabela 5.7

| | Base | Erro Verdadeiro (2) | (3) |
|---|---|---|---|
| dias_ate_eleicao | 0.0001*** | 0.00002 | 0.001*** |
| | (0.00) | (0.00) | (0.00) |
| tipoEstimulada | -0.027*** | -0.029*** | -0.025*** |
| | (0.01) | (0.01) | (0.01) |
| institutoDataPoder360 | 0.029*** | 0.014 | -0.004 |
| | (0.01) | (0.01) | (0.02) |
| institutoFSB | 0.013 | -0.003 | -0.002 |
| | (0.01) | (0.01) | (0.01) |
| institutoIbope | -0.0002 | 0.00005 | 0.002 |
| | (0.01) | (0.01) | (0.01) |
| institutoIpespe | -0.002 | 0.007 | 0.01 |
| | (0.01) | (0.01) | (0.01) |
| institutoMDA | -0.002 | 0.003 | -0.003 |
| | (0.01) | (0.01) | (0.01) |
| institutoParana Pesquisas | 0.0001 | 0.004 | 0.002 |
| | (0.01) | (0.01) | (0.01) |
| institutoReal Time Big Data | 0.007 | -0.008 | -0.008 |
| | (0.01) | (0.01) | (0.01) |
| institutoSensus | -0.001 | 0.014 | 0.003 |
| | (0.01) | (0.01) | (0.01) |
| institutoVerita | -0.015 | -0.026 | -0.015 |
| | (0.02) | (0.03) | (0.03) |
| institutoVirtu Analise | -0.01 | 0.001 | 0.002 |
| | (0.01) | (0.02) | (0.01) |
| institutoVox Populi | -0.004 | 0.001 | 0.004 |
| | (0.01) | (0.01) | (0.01) |
| alinhamentoDireita | 0.041*** | 0.031*** | 0.043*** |
| | (0.00) | (0.01) | (0.01) |
| alinhamentoEsquerda | 0.007* | -0.008 | -0.006 |
| | (0.00) | (0.01) | (0.01) |
| propaganda | -0.021*** | -0.033*** | 0.006 |
| | (0.00) | (0.01) | (0.01) |
| as.factor(ano)2010 | 0.043*** | 0.009 | 0.017 |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2014 | 0.035*** | 0.022** | 0.041*** |
| | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2018 | 0.018*** | 0.011 | 0.021** |
| | (0.01) | (0.01) | (0.01) |
| Constant | 0.046*** | 0.102*** | 0.027 |
| | (0.01) | (0.01) | (0.02) |
| Observations | 1,360 | 865 | 647 |
| $R^2$ | 0.218 | 0.155 | 0.23 |
| Adjusted $R^2$ | 0.207 | 0.136 | 0.207 |
| Residual Std. Error | 0.058 (df = 1340) | 0.063 (df = 845) | 0.059 (df = 627) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

## 5.5. Método 5 – Apresentação ano a ano

Por fim, como forma de testar a robustez de nossos coeficientes, iremos rodar as regressões com a base de dados dividida ano a ano. Os resultados são apresentados na tabela 5.8.

A primeira coisa que devemos perceber é que houve uma forte redução na quantidade de observações. Temos 158 para 2006, 227 para 2010, 287 para 2014 e, com um numero bem mais elevado que os demais, 688 para 2018. Além disso, temos que o $R^2$ apresenta bastante volatilidade entre os modelos.

Esta redução do numero de observações pode ser problemática quanto a significância de nossas variáveis, como observado pelo aumento do número de coeficientes não-significantes

Na comparação dos resultados ano a ano com o modelo base temos algumas questões; algumas das variáveis não possuem observações em determinados anos de eleição. Essas variáveis em geral são as de instituto, porém, temos que todas as pesquisas em nossa base de dados para o ano de 2006 foram do tipo estimulada, impossibilitando o uso desta. Entretanto, ainda é possível observar algumas tendências gerais em linha com o modelo base.

Um exemplo seria a variável de dias até a eleição, que apesar de perder sua significância nos anos de 2006 e 2010, sempre se manteve positiva. Inclusive, vemos que para o ano de 2018 seu coeficiente foi até superior.

Adicionalmente, observamos que novamente os coeficientes de alinhamento à direita obtiveram níveis de significância extremamente elevados. Os coeficientes também foram, em geral, positivos e relevantes. A exceção neste caso foi o ano de 2010 onde os partidos com maior erro em média foram aqueles de centro, e os candidatos de direita obtiveram o menor grau de erro entre os três.

Outra tendência mantida foi na variável de Tipo, onde as pesquisas do tipo estimulada continuaram demonstrando coeficientes negativos, significativos e da mesma ordem de grandeza que anteriormente. O modelo rodado para o ano de 2018 foi exceção, onde o coeficiente não teve significância relevante com valor bastante reduzido.

Quanto aos institutos, observamos que ainda não há uma tendência clara de que um erra mais do que outro consistentemente.

Por fim, vemos que a variável de propaganda foi negativa em todos os casos que foi significante, e positiva nos demais. Isto sugere que a tendência geral do coeficiente é, de fato, negativa.

**Tabela 5.8**

| | | Erro Verdadeiro | | | |
|---|---|---|---|---|---|
| | Base | 2006 | 2010 | 2014 | 2018 |
| dias_ate_eleicao | 0.0001*** | 0.0004 | 0.00001 | 0.0001*** | 0.0002*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| tipoEstimulada | -0.027*** | | -0.033** | -0.026*** | -0.004 |
| | (0.01) | | (0.02) | (0.01) | (0.01) |
| institutoDataPoder360 | 0.029*** | | | | 0.004 |
| | (0.01) | | | | (0.01) |
| institutoIbope | -0.0002 | 0.002 | 0.002 | -0.0001 | -0.003 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| institutoIpespe | -0.002 | | 0.045* | | -0.003 |
| | (0.01) | | (0.03) | | (0.01) |
| institutoMDA | -0.002 | | | -0.007 | 0.004 |
| | (0.01) | | | (0.01) | (0.01) |
| institutoParana Pesquisas | 0.0001 | | | | -0.002 |
| | (0.01) | | | | (0.01) |
| institutoReal Time Big Data | 0.007 | | | | 0.004 |
| | (0.01) | | | | (0.01) |
| institutoSensus | -0.001 | 0.003 | 0.029** | -0.011 | 0.0003 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) |
| institutoVerita | -0.015 | | | | -0.013 |
| | (0.02) | | | | (0.02) |
| institutoVirtu Analise | -0.01 | | | | 0.002 |
| | (0.01) | | | | (0.01) |
| institutoVox Populi | -0.004 | -0.004 | 0.009 | -0.004 | -0.008 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) |
| alinhamentoDireita | 0.041*** | 0.038*** | -0.019*** | 0.150*** | 0.033*** |
| | (0.00) | (0.01) | (0.01) | (0.01) | (0.01) |
| alinhamentoEsquerda | 0.007* | 0.007 | -0.012* | 0.029*** | 0.016*** |
| | (0.00) | (0.01) | (0.01) | (0.01) | (0.01) |
| propaganda | -0.021*** | 0.014 | -0.038*** | 0.006 | -0.022*** |
| | (0.00) | (0.02) | (0.01) | (0.01) | (0.01) |
| as.factor(ano)2010 | 0.043*** | | | | |
| | (0.01) | | | | |
| as.factor(ano)2014 | 0.035*** | | | | |
| | (0.01) | | | | |
| as.factor(ano)2018 | 0.018*** | | | | |
| | (0.01) | | | | |
| Constant | 0.046*** | -0.018 | 0.130*** | 0.023** | 0.039** |
| | (0.01) | (0.03) | (0.02) | (0.01) | (0.02) |
| Observations | 1,360 | 158 | 227 | 287 | 688 |
| $R^2$ | 0.218 | 0.144 | 0.256 | 0.722 | 0.157 |
| Adjusted $R^2$ | 0.207 | 0.104 | 0.225 | 0.713 | 0.137 |
| Residual Std. Error | 0.058 (df = 1340) | 0.042 (df = 150) | 0.042 (df = 217) | 0.037 (df = 277) | 0.063 (df = 671) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

# 6. Conclusão

Sabemos que entre a divulgação da primeira pesquisa eleitoral brasileira em 1945 (a única para o ano) e as 110 efetuadas para a eleição de 2018, certamente muita coisa mudou. O que antes era uma novidade, qual pouco se confiava e entendia, hoje é uma presença constante no cotidiano brasileiro em época de eleição. A cada novo resultado que é divulgado temos novas conversas e debates a respeito dos possíveis resultados que estão por vir, e novas considerações a respeito de aplicar ou não o famoso "voto útil".

No início deste estudo determinamos que seu objetivo seria a melhor compreensão dos fatores causadores de erro nas previsões de resultado das pesquisas eleitorais, a fim de gerar uma melhor interpretação dos dados que nos são apresentados. Afinal, se iremos tomar decisões com base naquilo que é apresentado pelas pesquisas eleitorais, pode ser socialmente benéfico saber o quanto peso devemos atribuir aos números observados.

Após a analise dos dados obtidos através das diferentes regressões, ficou claro que existem 4 variáveis principais na explicação dos fatores geradores de erro sobre as pesquisas eleitorais. Estas variáveis por sua vez são: (i) quantidade de dias até a eleição, (ii) tipo do questionário da pesquisa, (iii) alinhamento ideológico e (iv) variável do ano.

A variável de quantidade de dias até a eleição apresentou coeficiente claramente positivo em todos os modelos utilizados, obtendo altos níveis de significância em quase todos. O maior ponto de observação aqui seria a variação deste coeficiente em respeito à sua ordem de grandeza. Através do método 2 foi levantado suspeitas de que a estimação linear desta variável não era ideal, pois o coeficiente dela aumentava conforme restringiam-se os dias para mais próximos da eleição. Através do método 3 foi possível confirmar estas suspeitas; tanto o modelo quadrático quanto o cúbico demonstraram melhor encaixe à base de dados completa. As implicações destes resultados são que existe de fato uma aceleração da queda no erro verdadeiro conforme nos aproximamos mais e mais da data de eleição. No modelo cúbico temos que a cada dia extra até a eleição desde a divulgação da pesquisa temos um efeito sobre o erro, medido em pontos percentuais, de:

$$0{,}08 * dias - 0{,}00023 * dias^2 + 0.0000002 * dias^3$$

Ou seja, para uma pesquisa efetuada a 60 dias da eleição, espera-se que este fator seja causador de 4,02 pontos percentuais de erro; em linha com o observado na figura 4.1.

Quanto à variável do tipo do questionário empregado na pesquisa, observamos que pesquisas efetuadas com questionários do tipo estimulado erram 2,7 pontos

percentuais a menos do que aquelas do tipo espontâneo, tudo o mais constante. Esta variável foi certamente a mais consistente dentre todas utilizadas. Os coeficientes foram sempre relevantes, unidirecionais e significativos à casa de 1%. O valor do coeficiente variou de modelo em modelo, mas sempre se manteve na casa dos 2 ou 3 pontos percentuais.

Em relação a variável de alinhamento, tivemos que esta demonstrou que candidatos com alinhamento à direita do centro tendem a ter suas estimativas de voto menos precisas do que aqueles situados ao centro ou à esquerda. Os coeficientes aqui também foram sempre relevantes, unidirecionais e significativos à casa de 1%, e tem-se que em geral, se erra 4,1% a mais para candidatos de direita frente ao centro. O valor aqui também não foi o mesmo entre todos os modelos, mas sempre se manteve na mesma ordem de grandeza.

Em contrapartida, não podemos dizer o mesmo para os coeficientes gerados para os candidatos alinhados à esquerda. Neste quesito, vimos que em diversos modelos o coeficiente foi ou não-significativo ou significativo apenas ao nível de 10%. Com certeza parece existir uma tendência, onde há também um erro maior para os candidatos da esquerda, mas esta não pode ser confiantemente determinada com os dados que temos hoje (e em todo o caso o erro parece ser inferior ao da direita).

Por último tivemos a variável de ano, que se manteve bem consistente ao longo dos resultados. De maneira geral, tivemos que as pesquisas para as eleições de 2006 foram as com menor nível de erro, seguidas das eleições de 2018, 2014 e 2010, respectivamente. Estas variáveis, como já mencionado, atuam similar à constantes, absorvendo parte das explicações dos erros particulares àquelas eleições. Quando estimamos a regressão ano a ano, vimos que muitos dos resultados sofreram variações, porém este efeito pode ser explicado parcialmente ao numero reduzido de observações, possivelmente enviesando os resultados.

Quanto aos demais aprendizados com o estudo, pudemos ver que não há nenhuma tendência clara a respeito dos níveis de erro de cada instituto. Os únicos coeficientes que apesentaram significância eram específicos à certos modelos, indicando a ausência de erro sistêmico maior intrínseco à um instituto ou outro, conforme já demonstrava Gramacho (2013). Esse resultado foi surpreendente, dado que os resultados dos institutos maiores (Ibope, Datafolha) tendem a ser mais debatidos no cotidiano.

Outro resultado extremamente surpreendente foi em relação ao coeficiente da quantidade de pessoas entrevistadas, que apresentou resultados nulos. Como foi

anteriormente mencionado, as margens de erro divulgadas pelos institutos quando apresentam suas pesquisas são baseadas puramente na quantidade de entrevistados. Desta forma, seria de imaginar que este coeficiente seria ao menos minimamente relevante na explicação do erro.

Por fim, temos a variável de propaganda. Na apresentação desta variável no capítulo de metodologia tinha-se a hipótese de que uma vez que já se havia passado ao menos uma semana desde o inicio do período de propaganda eleitoral, haveria uma queda substancial nos erros devido ao maior conhecimento da base eleitoral a respeito dos candidatos. Em muitos de nossos modelos os coeficientes corroboraram com esta teoria, porém, este comportamento não foi completamente consistente. Nos modelos em que restringimos a quantidade de dias até eleição (colunas 2 e 3 da tabela 5.3), os coeficientes se tornaram positivos. Além disso, nos modelos em que foram utilizadas variáveis quadráticas e cúbicas para os dias até a eleição (tabela 5.4, colunas 2 e 3 respectivamente), observamos um comportamento similar; onde na coluna 2 houve redução na magnitude e significância e na coluna 3 houve inversão da direção e não-significância. Estes resultados indicam que há um comportamento entrelaçado entre a variável do período de propaganda e a variável de dias até a eleição. Esta relação pode valer ser estudada mais a fundo em analises futuras.

Os resultados apresentados neste estudo ajudaram a trazer a luz informações sobre quais são os determinantes das divergências entre os resultados apresentados pelas pesquisas eleitorais e os resultados finais de uma eleição; aqui denominados como os erros verdadeiros. Entretanto, a democracia brasileira é uma ainda bastante jovem, de modo que nem sequer seria possível ter uma base de dados hoje muito extensiva, em comparação às democracias mais velhas como as da Europa e Estados Unidos. Desta forma, fico esperançoso quanto ao prosseguimento deste estudo ao longo do tempo, para que seja possível uma compreensão ainda mais profunda dos tópicos aqui apresentados.

# 7. Referências bibliográficas

FERRAZ, C. *Crítica Metodológica às Pesquisas Eleitorais no Brasil.* 1996. 87f. Dissertação de Mestrado – Universidade Estadual de Campinas, Campinas, 1996. Disponível em http://repositorio.unicamp.br/bitstream/REPOSIP/307378/1/Ferraz_Cristiano_M.pdf. Acesso em: 10 dez. 2018.

GRAMACHO, W.G. À margem das margens? A precisão das pesquisas pré-eleitorais brasileiras em 2010. OPINIÃO PÚBLICA, Campinas, vol. 19, nº 1, p. 65-80, junho, 2013. Disponível em http://www.scielo.br/scielo.php?pid=S0104-62762013000100004&script=sci_arttext&tlng=es. Acesso em: 10 dez. 2018

MOSTELLER, F. Measuring the error. In: MOSTELLER, F.H.; HYMAN, P.J.; MCCHARTY, E.S.; MARKS & TRUMAN, D.B. (Eds.). The pre-election polls of 1948. Report to the committee on analysis and pre-election polls and forecast. New York: Social Science Research Council, p. 54-80, 1949

PESQUISAS ELEITORAIS NO BRASIL – 2018 - https://www.itauassetmanagement.com.br/content/dam/itau-asset-management/content/pdf/white-papers/Pesquisas%20Eleitorais%20no%20Brasil%20-%20White%20Paper.pdf - Acesso em: 10 dez. 2018

POWER, T.J. and ZUCCO, Jr., C. (2009), Estimating Ideology of Brazilian Legislative Parties, 1999-2005. In: *Latin American Research Review*, 44, 1, 218-246. Disponível em http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.475.7589&rep=rep1&type=pdf - Acesso em: 03 mai. 2019.

ROSSI, Amanda. ‘O eleitor decide o voto cada vez mais tarde‘, diz diretora do Ibope. São Paulo, 12 out. 2018. Disponível em: https://www.bbc.com/portuguese/brasil-45845475. Acesso em: 11 jun. 2019.

DUARTE, Guilherme Jardim; MELLO, Fernando. **Modelo de ‘house effects’ mostra vieses de institutos para cada candidato**. São Paulo, 19 set. 2018. Disponível em: https://www.jota.info/dados/agregador-de-pesquisas/house-effects-institutos-pesquisas-19092018. Acesso em: 20 nov. 2018.

JORDÃO, Rogério. **Modelo de ‘house effects’ mostra vieses de institutos para cada candidato**. [*S. l.*], 19 set. 2018. Disponível em: História da pesquisa. Acesso em: 13 jun. 2019.

GALLUP. **Gallup Presidential Election Trial-Heat Trends, 1936-2008**. Disponível em: https://news.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4. Acesso em: 14 jun. 2019.

G1. **12% dos eleitores decidiram o voto para presidente no dia da eleição, diz Datafolha**. [*S. l.*], 10 out. 2018. Disponível em: https://g1.globo.com/politica/eleicoes/2018/eleicao-em-numeros/noticia/2018/10/10/12-dos-eleitores-decidiram-o-voto-para-presidente-no-dia-da-eleicao-diz-datafolha.ghtml. Acesso em: 11 jun. 2019

HAUBERT, Mariana; TRUFFI, Renan. **Três em cada 10 eleitores dizem que podem aderir ao 'voto útil' no primeiro turno, diz pesquisa**. São Paulo, 26 set. 2018. Disponível em: https://politica.estadao.com.br/noticias/eleicoes,cniibope-tres-em-cada-10-eleitores-dizem-que-podem-aderir-ao-voto-util-no-primeiro-turno,70002520385. Acesso em: 15 jun. 2019.

QUALTRICS. **The 1936 Election – A Polling Catastrophe**. [*S. l.*], 12 out. 2010. Disponível em: https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/. Acesso em: 14 jun. 2019.