Entendendo a opinião política na eleição presidencial dos EUA de 1988

Aluno: Ezequiel de Braga Santos
Professor: Luiz Max Fagundes de Carvalho
Escola de Matemática Aplicada, FGV/EMAp
Rio de Janeiro - RJ.

1 Introdução

Ao longo dos anos, as pesquisas de opinião têm desempenhado um papel fundamental na compreensão do comportamento das pessoas. No cenário eleitoral, em especial, as pesquisas eleitorais ganham cada vez mais destaque para obter prévias do resultado de uma eleição. Assim, é comum dados serem coletados regularmente em cada período eleitoral, considerando-se diversas características do entrevistado, como sexo, educação, renda, etnia e local. Dessa maneira, é desejável entender como essas covariáveis influenciam na resposta do indivíduo: quais fatores determinam a opinião política? É possível prever como um indivíduo votaria apenas analisando essas características?

Nesse contexto, o presente trabalho busca entender como o comportamento da população é influenciado tanto pelas particularidades individuais quanto pelos aspectos coletivos. Para isso, foram considerados dados coletados da última pesquisa que antecedeu a eleição presidencial dos Estados Unidos de 1988, realizada pela *CBS News*. Inicialmente, os dados possuem as seguintes colunas:

- *org*: órgão que realizou a pesquisa;
- *year*: ano de realização da pesquisa;
- *survey*: qual a pesquisa;
- *bush*: se é favorável ao George Bush (1), Michael Dukakis (0), outros/indeciso (NA);
- *state*: id do estado do indivíduo;
- *edu*: grupo educacional do indivíduo (1: *less than high school*, 2: *high school*, 3: *some college* e 4: *college graduate*);
- *age*: grupo de idade do indivíduo (1: 18–29, 2: 30–44, 3: 45–64 e 4: 65+);
- *female*: sexo feminino (1) ou masculino (0);

- *black*: etinia afro-americana (1) ou não (0);
- *weight*: coluna desconhecida.

É importante salientar que o entendimento da opinião política é fundamental para a formulação de políticas públicas, uma vez que os resultados de uma pesquisa eleitoral refletem (em geral) as preferências e preocupações dos eleitores. Ademais, a compreensão de padrões é fundamental para candidatos traçarem estratégias de campanha mais eficazes. Dessa forma, analisar quais fatores são determinantes para o voto é de extrema relevância para desenvolver formas de combater possíveis opiniões negativas e conquistar o eleitorado.

2 Metodologia

Propõe-se ajustar os modelos com regressão logística: denotando por Y_i a resposta da pesquisa do indivíduo i e X_i as covariáveis associadas aos coeficientes β, a proposta consiste em modelar a probabilidade de $Y_i = 1$ como

$$P(Y_i = 1) = logit^{-1}(X_i\beta).$$

2.1 Modelo 1

Como primeira abordagem, foi proposto o seguinte modelo simples:

$$P(Y_i = 1) = logit^{-1}(\alpha + \beta_1 edu_i + \beta_2 age_i + \beta_3 female_i + \beta_4 black_i).$$

2.2 Modelo 2

Em seguida, propõe-se um modelo multinível com intercepto variando por estado, eliminando a covariável *age*:

$$\begin{aligned} P(Y_i = 1) &= logit^{-1}(\alpha_{j[i]} + \beta_1 edu_i + \beta_2 female_i + \beta_3 black_i), \\ \alpha_j &= \mu_\alpha + \eta_j, \\ \eta_j &\sim N(0, \sigma^2_{state}). \end{aligned}$$

2.3 Modelo 3

Posteriormente, foi adicionada interação entre *black* e *female* e retirada a covariável *edu*:

$$
\begin{aligned}
P(Y_i = 1) &= logit^{-1}(\alpha_{j[i]} + \beta_1 black_i + \beta_2 female_i \cdot black_i), \\
\alpha_j &= \mu_\alpha + \eta_j, \\
\eta_j &\sim N(0, \sigma^2_{state}).
\end{aligned}
$$

Para a realização da modelagem foram utilizadas a função *glm()* e a função *glmer()*, do pacote **lme4**, presentes na linguagem de programação R. Ambas realizam o ajuste de modelos lineares generalizados por máxima verosimilhança.

Posteriormente, para avaliação dos modelos foram analisados a significância dos parâmetros estimados, as predições realizadas, bem como algumas métricas relacionadas, como AUC, acurácia, recall, precisão e F1-score.

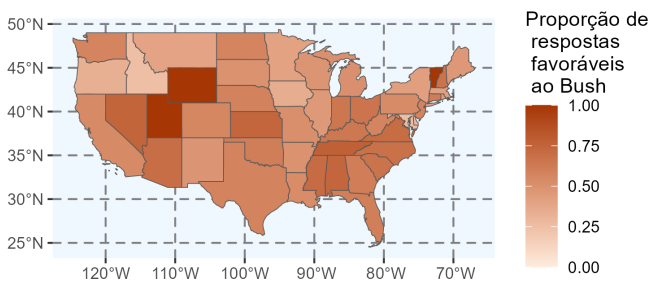
3 Resultados

3.1 Análise exploratória dos dados

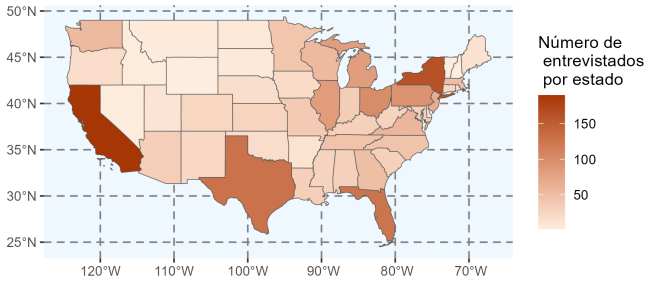
Para as análises propostas, os indivíduos indecisos ou que preferem outros candidatos (NA) foram desconsiderados. Além disso, a coluna *weight* foi desconsiderada e foram adicionadas as siglas dos estados associados aos seus respectivos identificadores.

Ao considerar a proporção de respostas favoráveis ao Bush, a figura abaixo sugere que os estados Utah, Wyoming e Vermont possuem o maior número de pessoas favoráveis proporcionalmente ao número de entrevistados. No entanto, há poucos dados referentes a esses estados: apenas 2 pessoas são favoráveis ao Bush em Wyoming e Vermont, e 11 pessoas em Utah. Observa-se também que há vários estados com poucos indivíduos entrevistados, especialmente nas regiões Oeste e Centro-Oeste.

Figura 1: Respostas ao nível estadual



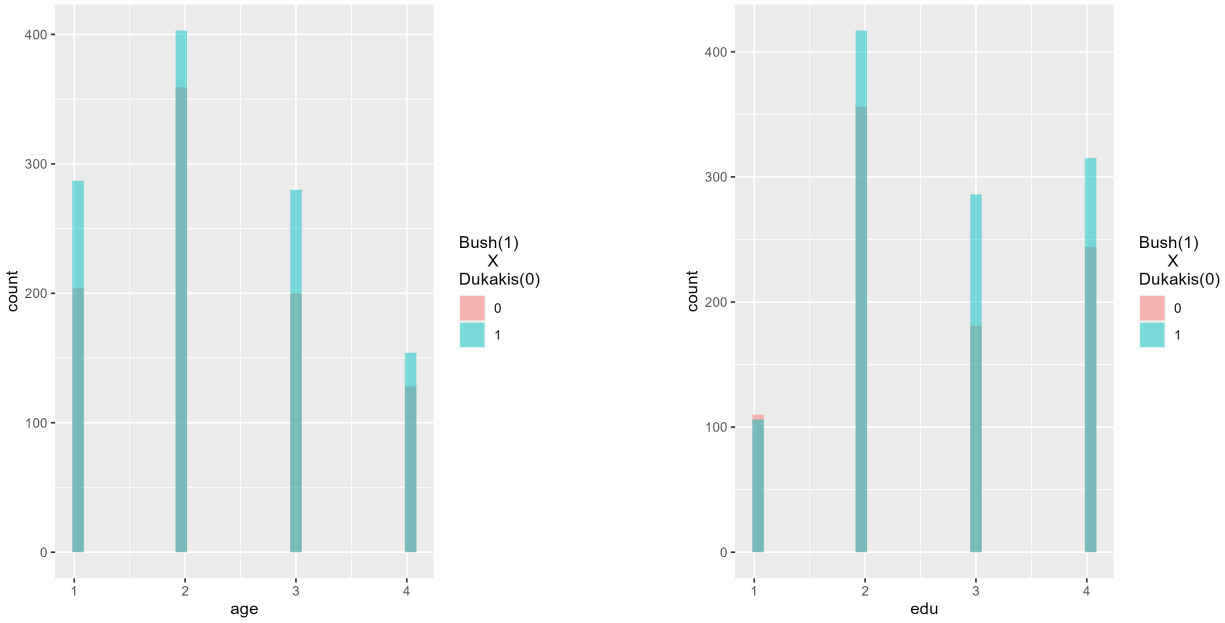
(a) Proporção de respostas por estado



(b) Número de entrevistados por estado

Por outro lado, ao analisar as respostas por grupo de idade e educação, observa-se uma boa vantagem do candidato republicano (Bush), apenas no primeiro grupo educacional há um equilíbrio, conforme os histogramas abaixo.

Figura 2: Número de respostas por grupo



(a) Grupo de idade

(b) Grupo educacional

Ao analisar sexo e etnia, o Bush também leva vantagem com todos os sexos. No entanto, entre os afro-americanos o Dukakis tem uma larga vantagem, conforme as tabelas a seguir.

Tabela 1: Número de respostas por sexo

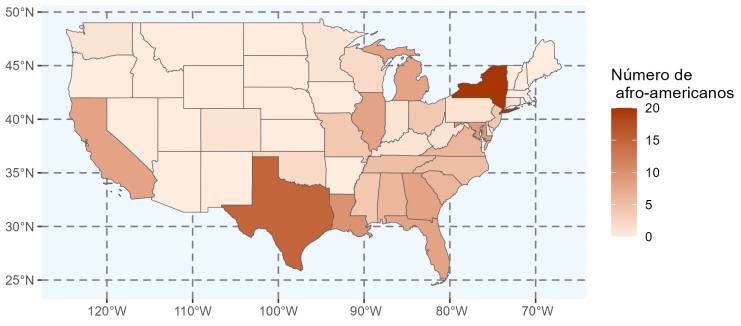
	M	F
Dukakis	355	536
Bush	476	648

Tabela 2: Número de respostas por etnia

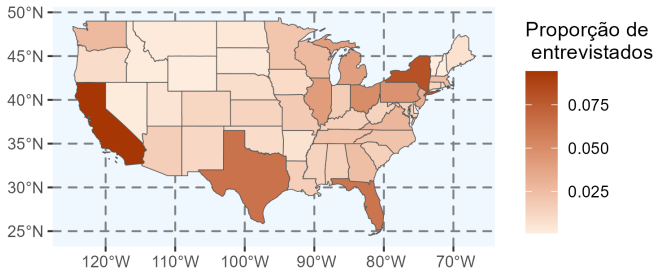
	Outros	Afro-americano
Dukakis	770	121
Bush	1091	33

Além disso, a figura abaixo mostra que a região Oeste possui praticamente nenhum negro entrevistado.

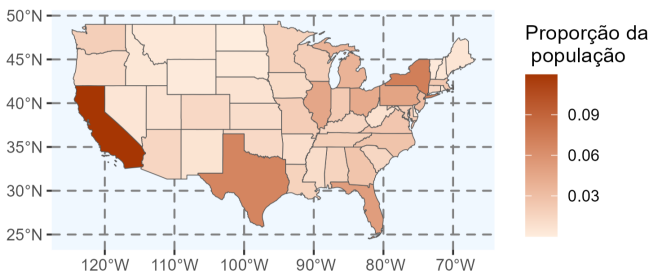
Figura 3: Número de afro-americanos por estado



Por fim, ao comparar a proporção de entrevistados com a proporção da população por estado, observa-se nos gráficos abaixo que a amostra foi bem distribuída:



(a) Proporção de entrevistados por estado



(b) Proporção da população por estado

3.2 Ajustes

A tabela a seguir contêm os parâmetros ajustados com suas respectivas incertezas para cada um dos modelos:

Tabela 3: Sumário das estimativas

	Parâmetro	Estimativa	Erro padrão	Int. confiança (95%)
Modelo 1	α	0.316	0.191	$(-0.059,\ 0.692)$
	β_1	0.061	0.047	$(-0.030,\ 0.153)$
	β_2	-0.037	0.047	$(-0.129,\ 0.055)$
	β_3	-0.080	0.093	$(-0.263,\ 0.102)$
	β_4	-1.641	0.203	$(-2.053, -1.255)$
Modelo 2	μ_α	0.216	0.165	$(-0.108,\ 0.540)$
	β_1	0.084	0.048	$(-0.009,\ 0.178)$
	β_2	-0.088	0.095	$(-0.275,\ 0.099)$
	β_3	-1.720	0.210	$(-2.146, -1.321)$
Modelo 3	μ_α	0.388	0.084	$(0.220,\ 0.561)$
	β_1	-1.586	0.320	$(-2.248, -0.985)$
	β_2	-0.259	0.408	$(-1.055,\ 0.553)$

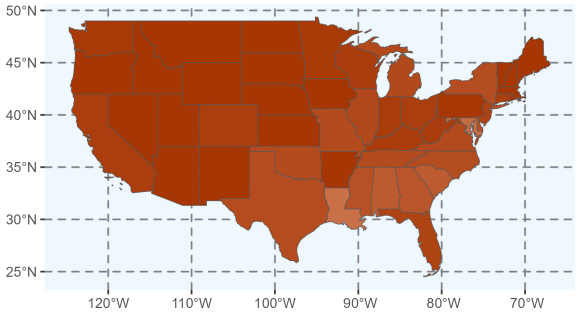
A partir da tabela, nota-se que:

- O modelo 1 apresenta os quatro primeiros parâmetros com incerteza razoável, considerando suas escalas. Isso faz sentido, já que há equilíbrio na opinião por idade, grupo educacional e sexo;

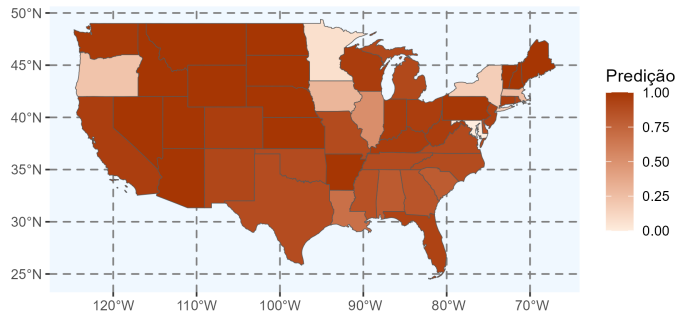
- O modelo 2 também apresenta algo semelhante ao primeiro modelo, com uma incerteza alta da média do intercepto;
- No modelo 3, ao desconsiderar idade e educação os interceptos são estimados com menos incerteza, mas o coeficiente da interação entre sexo e etnia é bem incerto;
- Em todos os modelos o coeficiente de etnia foi estimado com pouca incerteza, uma vez que com os dados disponíveis é evidente que a grande maioria dos afro-americanos são favoráveis ao Dukakis.

3.3 Predições

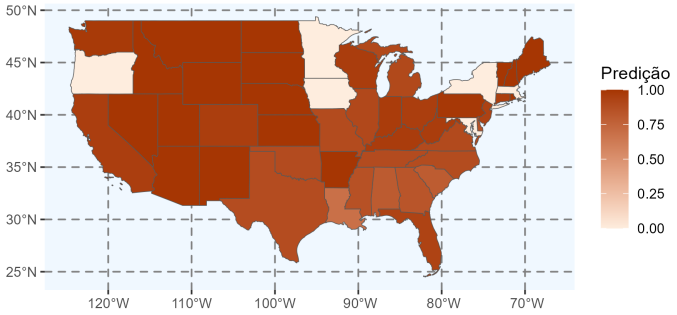
Ao realizar as predições da proporção de respostas favoráveis ao Bush dos três modelos no nível estadual, obtêm-se os seguintes resultados:



(a) Modelo 1



(b) Modelo 2



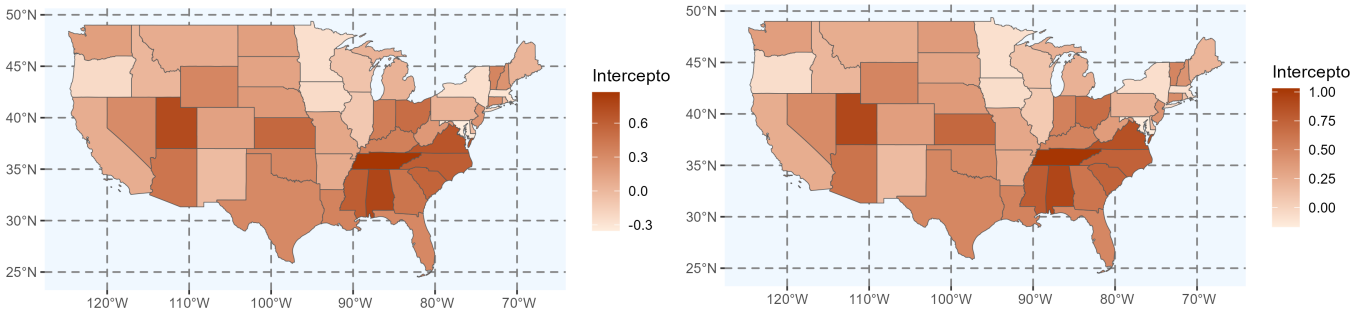
(c) Modelo 3

Figura 5: Predição por estado

Observa-se que:

- O primeiro modelo é bem ruim para diferenciar as predições estaduais, já que leva em conta essa variação apenas na média;
- Ao adicionar interceptos variando por estado, os modelos 2 e 3 conseguem captar as características inerentes de cada estado, mas aqueles que possuem poucos dados

continuam com predições ruins, uma vez que é difícil perceber o que influencia a opinião das pessoas favoráveis ao Dukakis nesses estados.



(a) Modelo 2 (b) Modelo 3

Figura 6: Intercepto estimado por estado

Por outro lado, ao avaliar somente os interceptos estimados acima, os modelos representam o que ocorre em geral: estados com maior intercepto tendem a ter maior número de respostas favoráveis ao Bush.

Por fim, ao considerar as métricas calculadas abaixo, adicionar intercepto por estado melhora a maior parte delas. No primeiro modelo, um recall alto indica baixa quantidade de falso negativo e uma precisão baixa indica alta quantidade de falso positivo, isto é, o primeiro modelo prevê poucas pessoas a favor do Dukakis que são favoráveis ao Bush, mas prevê muitos indivíduos favoráveis a este último que são favoráveis ao candidato democrata. A medida que considera-se intercepto variando por estado e as covariáveis mais relevantes, as quantidades de falso positivo e falso negativo vão se balanceando, mas os modelos ainda continuam com dificuldade de prever opiniões favoráveis ao Dukakis.

Tabela 4: Métricas dos modelos

	AUC	Acurácia	Recall	Precisão	F1-Score
Modelo 1	0.5737	0.6015	0.9706406	0.586244	0.7309883
Modelo 2	0.666	0.6228	0.866548	0.614899	0.7193501
Modelo 3	0.6622	0.6253	0.86121	0.6177409	0.7194352

4 Discussão e Conclusão

Neste trabalho, foram apresentados três modelos logísticos, desde um mais simples considerando todas as covariavéis relevantes para a modelagem, ao mais robusto, com efeitos aleatórios ao nível estadual e com as covariáveis consideradas mais significativas. Com o primeiro modelo foram obtidas predições mais próximas do que ocorre na média, enquanto

os modelos posteriores captaram as diferenças presentes nos estados juntamente com suas covariáveis.

Ao analisar os coeficientes estimados, apenas o coeficiente relacionado a etnia é estimado com menor incerteza, já que é possível perceber a diferença de opinião nesse grupo com os dados analisados. Considerando-se as métricas calculadas na tabela 4, observa-se que cada um dos modelos possui suas qualidades e, a medida que as alterações foram feitas até o modelo 3, algumas características foram melhoradas e outras pioradas, sendo necessário entender aquele que melhor atende aos objetivos traçados, balanceando os prós e contras.

Sendo assim, a pequena quantidade de dados presentes em alguns estados limita a previsão das respostas: aqueles que possuem poucos entrevistados ou que possuem poucos entrevistados a favor do candidato democrata possuem previsões ruins. Todavia, o modelo é razoável para entender o comportamento geral da população, uma vez que permite perceber a grande influência da etnia na diminuição da probabilidade da resposta ser favorável ao Bush, isto é, negros e tendem a ter preferência pelo candidato democrata. Na região Oeste, especialmente, como há poucos entrevistados negros, os modelos não conseguem prever opiniões favoráveis ao Dukakis, uma vez que sua probabilidade é determinada por essa covariável.

Para um melhor resultado, pode-se tentar uma regressão bayesiana, desde que seja possível encontrar prioris informativas. Além disso, pode-se modelar a probabilidade com outra função, com um modelo probit, por exemplo. No entanto, pela natureza do assunto, grandes melhorias são improváveis com os dados analisados, sobretudo na predição individual, uma vez que é difícil determinar a opinião política somente com as covariáveis disponíveis.

Referências

[1] Andrew Gelman and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models, 2007.