Modelos de Resposta para Experimentos Randomizados em Redes Sociais de Larga Escala*

Francisco Galuppo Azevedo, Bruno Demattos Nogueira, Fabricio Murai, Ana Paula Couto da Silva

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{franciscogaluppo, bruno.demattos, murai, ana.coutosilva}@dcc.ufmg.br

Abstract. *A/B tests are randomized experiments frequently used by companies that offer services on the Web for assessing the impact of new features. During an experiment, each user is randomly redirected to one of two versions of the website, called treatments. Several response models were proposed to describe the behavior of a user in a social network website, where the treatment assigned to her neighbors must be taken into account. However, there is no consensus as to which model should be applied to a given dataset. In this work, we propose a new response model, derive theoretical limits for the estimation error of several models, and obtain empirical results for cases where the response model was misspecified.*

Resumo. *Testes A/B são experimentos randomizados muito utilizados por empresas que oferecem serviços na Web para avaliar o impacto de novas funcionalidades. Durante um experimento, cada usuário é redirecionado aleatoriamente para uma de duas versões do site, chamadas tratamentos. Diversos modelos de resposta foram propostos para descrever o comportamento de um usuário em sites de redes sociais, onde o tratamento atribuído aos seus vizinhos deve ser considerado. Porém, não há consenso sobre qual modelo deve ser aplicado a um conjunto de dados. Neste trabalho, propomos um novo modelo de resposta, derivamos limites teóricos para o erro de estimação de diversos modelos, e obtemos resultados empíricos para o caso de especificação incorreta do modelo.*

1. Introdução

Há alguns anos grandes empresas que oferecem serviços na Web (como Amazon, e-Bay, Facebook, Google e LinkedIn) vêm percebendo a importância de se conduzir pesquisas sobre a experiência dos usuários para a tomada de decisões, a nível de desenvolvimento (p. ex., mudanças no funil de compra ou no layout do site) e também a nível de negócio (p. ex., novas funcionalidades e diferenciais em relação aos concorrentes) [Kohavi et al. 2013]. Este impacto é quantificado através de medidas de interesse, tais como a fração de usuários que retorna ao site, o número médio de cliques por usuário, o lucro obtido através de anúncios no site etc. Para cada medida, deve-se estimar o efeito médio do tratamento, conhecido como *average treatment effect* (ATE).

*The authors' work has been partially funded by the EUBra-BIGSEA project by the European Commission under the Cooperation Programme (MCTI/RNP 3rd Coordinated Call), Horizon 2020 grant agreement 690116, CAPES, CNPq and FAPEMIG.

Dentre as técnicas existentes, testes A/B se destacam como uma das mais proeminentes devido a sua capacidade de quantificar mudanças comportamentais objetivamente, calcular resultados e determinar a significância estatística de forma automática (sem que seja necessário feedback explícito do usuário) [Kohavi et al. 2009]. Testes A/B consistem em experimentos randomizados com duas variantes de um tratamento. Usuários que visitam um website são aleatoriamente atribuídos ao grupo de controle (versão atual) ou ao grupo de tratamento (versão sendo testada).

Os testes A/B têm como premissa a *Stable Unit Treatment Value Assumption* (SUTVA). A SUTVA é a suposição de que não há interferência entre os indivíduos do experimento, i.e., que o comportamento de um indivíduo depende apenas do tratamento que lhe foi atribuído. Enquanto esta suposição é válida para testes clínicos envolvendo placebos e medicamentos, experimentos em redes sociais normalmente violam esta condição (efeito conhecido como *spillover*) [Xu et al. 2015].

Idealmente, gostaríamos de comparar os resultados de dois experimentos simultâneos: no primeiro, toda a população está no grupo de controle e, no segundo, todos estão no grupo de tratamento. Assim, mesmo que a SUTVA fosse violada, conseguiríamos estimar o ATE. Como isso não é possível, para se estimar o ATE é preciso assumir um modelo que descreve como a resposta de um indivíduo varia conforme o tratamento que é atribuído a ele e ao restante da população. No caso de redes sociais, assume-se que a resposta de um indivíduo é influenciada apenas pelos seus relacionamentos diretos.

Modelos de resposta são importantes não só para estimar o ATE, como também para comparar vários métodos de estimação quanto a sua acurácia. Como a função de resposta em um experimento real é desconhecida, a comparação entre métodos é feita a partir de dados sintéticos, gerados utilizando-se modelos de resposta. Tradicionalmente, os modelos de resposta utilizados são diferentes daqueles assumidos pelos métodos propostos. Contudo, para que os resultados em relação ao erro de estimação sejam colocados em perspectiva, é fundamental conhecer os limites teóricos do erro de estimação, caso o modelo correto fosse empregado.

Este trabalho possui três contribuições principais. A primeira consiste em propor um novo modelo de resposta, chamado τ-exposure, baseado em limiar de exposição. Neste modelo um indivíduo é considerado "exposto" se ele recebe o controle (tratamento) e uma fração de seus vizinhos maior que τ também recebe controle (tratamento). A segunda contribuição é a derivação analítica dos limites inferiores para o erro de estimação quando o modelo correto de interferência é utilizado. Finalmente, iremos utilizar diversos modelos de resposta para realizar um estudo empírico do erro resultante ao se especificar o modelo incorreto durante a estimação. A versão estendida deste trabalho está disponível em [Azevedo et al. 2018].

2. Trabalhos relacionados

Com o auxílio de um grafo é possível definir modelos de resposta computacionalmente tratáveis. Estes modelos fazem parte da classe *neighborhood treatment response*. Em alguns destes modelos, assume-se que um indivíduo é considerado "exposto" ao tratamento se ele e pelo menos k de seus vizinhos (ou uma fração maior que q) recebem o mesmo tratamento [Backstrom and Kleinberg 2011, Ugander et al. 2013]. Apesar de apresentarem resultados empíricos, nenhum destes trabalhos estuda os limites teóricos de estimação.

Em outro trabalho propõe-se um modelo linear para a resposta de um nodo em função do tratamento dele e dos seus vizinhos [Gui et al. 2015]. Para reduzir o impacto de supor incorretamente tal relação linear, os autores usam randomização a nível de cluster. Por outro lado, o modelo, que produz respostas reais, é aplicado a respostas binárias sem que haja um estudo sobre o impacto desta inconsistência. Embora diversos modelos de resposta tenham sido propostos para capturar a interferência, não existe uma regra para escolher aquele que melhor representa o conjunto de dados, tampouco um entendimento do erro que a especificação incorreta pode ocasionar. Atualmente, existe apenas um método que permite testar se a SUTVA é válida [Saveski et al. 2017], assim como calcular a probabilidade de erro do Tipo I (rejeitar a SUTVA quando ela é válida).

3. Modelos de resposta

Descrevemos a seguir a notação usada nos modelos de resposta. Seja cada unidade (ou indíviduo) do experimento indexada por $i \in 1, \ldots, N$. Defina $\mathbf{Z} \in \mathcal{Z}^N$ e $\mathbf{Y} \in \mathcal{Y}^N$ como sendo os vetores de tratamento e de resposta, onde $Z_i \in \mathcal{Z}$ e $Y_i \in \mathcal{Y}$ são, respectivamente, o tratamento atribuído ao e a resposta do usuário i. Defina $g_i(\mathbf{Z})$ como sendo uma função específica à unidade i, calculada sobre a atribuição de tratamento $\mathbf{Z}$. Quando as unidades estão conectadas segundo uma topologia de rede, exemplo típicos de tais funções são a fração ou o número de vizinhos a que foram atribuídos um dado tratamento. Defina o vetor coluna $\mathbf{x}_i = [1, Z_i, g_i(\mathbf{Z})]^\top$. Em geral, existe um componente estocástico ϵ_i associado à unidade i, amostrado independentemente para $i = 1, \ldots, N$ a partir da distribuição Gaussiana com média zero e variância desconhecida σ^2. A seguir, iremos considerar apenas experimentos envolvendo dois tratamentos (testes A/B), logo $Z_i \in \mathcal{Z} \equiv \{0, 1\}$.

Modelo linear de resposta. Quando as respostas são números reais (i.e., $Y_i \in \mathcal{Y} \equiv \mathbb{R}$), o modelo linear de resposta costuma ser utilizado:

$$Y_i \;\; = \;\; \mathbf{x}_i^\top \boldsymbol{\beta} \; + \epsilon_i, \tag{1}$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ são parâmetros do modelo.

Modelo probit. Uma forma de se produzir respostas binárias (i.e., $Y_i \in \mathcal{Y} \equiv \{0, 1\}$) a partir de uma função real (p. ex., $\mathbf{x}_i^\top \boldsymbol{\beta} \; + \epsilon_i$) consiste em usar o sinal da função para determinar a resposta. Este é caso do modelo probit:

$$\Pr(Y_i = 1 | \mathbf{x}_i) \;\; = \;\; \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}) \tag{2}$$

onde $\Phi(.)$ é a função de distribuição cumulativa da Normal de média 0 e variância σ^2.

Modelo logístico. Outra forma de se produzir respostas binárias é mapear o valor de uma função real $f(.)$ em uma probabilidade, usada para amostrar o valor da resposta. Nesse caso, $f(.)$ precisa ser determinística para que seja possível estimar os parâmetros do modelo. Um exemplo desse tipo de modelo é o modelo logístico:

$$\Pr(Y_i = 1 | \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})}. \tag{3}$$

Modelo τ-Exposure: um novo modelo baseado em limiar de exposição. No modelo *fractional q-neighborhood response* de [Ugander et al. 2013], um indivíduo é considerado "exposto" a um tratamento (A ou B) se ele e uma fração maior que q de seus vizinhos for atribuída am mesmo tratamento. A suposição implícita feita por este modelo é

de que existe um limiar de exposição: uma vez que um nodo v é considerado "exposto", aumentar a fração de vizinhos atribuídos ao mesmo tratamento não afeta a distribuição da resposta Y_i. Este modelo não pode ser usado para geração de respostas sintéticas, pois não especifica a distribuição das respostas (seja o nodo exposto ou não).

Assim, propomos o modelo de resposta τ-Exposure baseado na suposição da existência de um limiar de exposição $\tau \geq 0.5$. Seja $g_i(\mathbf{Z})$ a fração de vizinhos do nodo i no grupo de tratamento. A resposta de um indivíduo é dada por

$$Y_i = \epsilon_i + \begin{cases} \beta_0 & \text{se } Z_i = 0 \text{ e } g_i(\mathbf{Z}) \leq 1-\tau, \\ \beta_0 + \beta_2(g_i(\mathbf{Z}) - (1-\tau)) & \text{se } Z_i = 0 \text{ e } g_i(\mathbf{Z}) > 1-\tau, \\ \beta_0 + \beta_1 & \text{se } Z_i = 1 \text{ e } g_i(\mathbf{Z}) \geq \tau, \\ \beta_0 + \beta_1 + \beta_2(g_i(\mathbf{Z}) - \tau) & \text{se } Z_i = 1 \text{ e } g_i(\mathbf{Z}) < \tau. \end{cases} \tag{4}$$

Para que o modelo seja realista, é necessário que $\beta_1\beta_2 > 0$, ou seja, que impacto do tratamento do nodo e daquele dos seus vizinhos na resposta esperada tenha o mesmo sinal. Além disso, é necessário que $|\beta_2\tau| \leq |\beta_1|$ para que as curvas de $E[Y_i]$ quando $Z_i = 0$ e $Z_i = 1$ não se cruzem.

Modelo τ-exposure com respostas binárias. As respostas ($Y_i \in \mathbb{R}$) deste modelo podem ser transformadas em respostas binárias de maneira similar ao modelo probit ou ao modelo logístico. Neste artigo, para a geração de respostas binárias, iremos considerar o sinal de Y_i, assim como é feito no modelo probit.

Efeito Médio do Tratamento. Novamente, assumimos que Y_i é uma variável aleatória condicionada em $\mathbf{Z}$. Neste caso, o efeito médio do tratamento (*average treatment effect* ou ATE) é o valor esperado da diferença em médias (*difference in-means*) entre a resposta de uma unidade quando a população está em tratamento (i.e., $\mathbf{Z} = \mathbf{1}$) e a resposta da unidade quando a população está em controle (i.e., $\mathbf{Z} = \mathbf{0}$):

$$\text{ATE} = \text{E}\left[\frac{1}{N}\sum_{i=1}^{N} Y_i(\mathbf{Z} = \mathbf{1}) - \frac{1}{N}\sum_{i=1}^{N} Y_i(\mathbf{Z} = \mathbf{0})\right]. \tag{5}$$

Esta equação pode ser especializada para cada um dos modelos apresentados nesta seção (maiores detalhes em [Azevedo et al. 2018, Seção 5]).

Estimadores. Os estimadores para cada um dos modelos apresentados nesta seção são detalhados em [Azevedo et al. 2018, Seção 6].

4. Limites inferiores para o erro de estimadores não-enviesados

Nesta seção derivamos limites inferiores (alguns assintóticos) para o erro de qualquer estimador não-enviesado do ATE, para cada um dos modelos de resposta. Para isto, usamos o *Cramér-Rao Lower Bound* (CRLB), que relaciona o erro médio quadrático (MSE) com a quantidade de informação contida nos dados à respeito dos parâmetros a serem estimados, medida pela matriz de informação de Fisher (FIM). Denotamos por $T_1(\mathbf{X})$, $T_2(\mathbf{X})$, $T_3(\mathbf{X})$ e $T_4(\mathbf{X})$ estimadores não-enviesados do ATE para cada um dos respectivos modelos. As demonstrações podem ser encontradas no nosso relatório técnico [Azevedo et al. 2018].

Teorema 1. *O limite inferior do erro de estimação para o modelo linear é dado por* $MSE(T_1(\mathbf{X})) \geq [0\ 1\ 1]\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}[0\ 1\ 1]^\top$. *Este limite é válido mesmo que o termo da variância σ^2 seja desconhecido.*

No caso do modelo probit, foi provado em [Demidenko 2001] que a informação contida em uma amostra sobre $\boldsymbol{\beta}$, medida pela matriz de informação de Fisher (FIM) é $\mathcal{I}(\mathbf{X}) = \sum_{i=1}^{N} \frac{\phi^2(s_i)}{\Phi(S_i)(1-\Phi(s_i))}\mathbf{x}_i\mathbf{x}_i^\top$, onde $s_i = \mathbf{x}_i^\top\boldsymbol{\beta}$. Com isso, provaremos o Teorema 2.

Teorema 2. *O limite inferior* assintótico *do erro de estimação para o modelo probit* (2) *é dado por MSE*$(T_2(\mathbf{X})) \geq (\nabla_{\boldsymbol{\beta}} h)^\top \mathcal{I}^{-1}(\mathbf{X})\nabla_{\boldsymbol{\beta}} h$*, onde*

$$\nabla_{\boldsymbol{\beta}} h = \left(\phi(\boldsymbol{\beta}^\top\mathbf{1}) - \phi(\beta_0), \phi(\boldsymbol{\beta}^\top\mathbf{1}), \phi(\boldsymbol{\beta}^\top\mathbf{1})\right).$$

Segundo [Erhardt 2017], a FIM do modelo logístico é dada por $\mathcal{I}(\mathbf{X}) = \mathbf{X}^\top\mathbf{W}\mathbf{X}$, onde $\mathbf{W} = [W_{ij}]_{N\times N}$ é a matriz diagonal tal que $W_{ii} = \exp(s_i)/(1+\exp(s_i))^2$. Em [Azevedo et al. 2018], usamos este fato para provar o seguinte teorema.

Teorema 3. *O limite inferior* assintótico *do erro de estimação para o modelo logit* (3) *é dado por MSE*$(T_3(\mathbf{X})) \geq (\nabla_{\boldsymbol{\beta}} g)^\top \mathcal{I}^{-1}(\mathbf{X})\nabla_{\boldsymbol{\beta}} g$*, onde*

$$\nabla_{\boldsymbol{\beta}} g = \left(\frac{1}{e^{\boldsymbol{\beta}^\top\mathbf{1}}+1} - \frac{1}{(e^{\boldsymbol{\beta}^\top\mathbf{1}}+1)^2} - \frac{e^{\beta_0}}{(e^{\beta_0}+1)^2}, \frac{e^{\boldsymbol{\beta}^\top\mathbf{1}}}{(e^{\boldsymbol{\beta}^\top\mathbf{1}}+1)^2}, \frac{e^{\boldsymbol{\beta}^\top\mathbf{1}}}{(e^{\boldsymbol{\beta}^\top\mathbf{1}}+1)^2}\right).$$

Teorema 4. *O limite inferior do erro de estimação para o modelo* τ*-exposure é dado por MSE*$(T_4(\mathbf{X})) \geq [0\ 1\ 1]\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}[0\ 1\ 1]^\top$*, onde a matrix* $\mathbf{X}$ *é definida em [Azevedo et al. 2018].*

5. Metodologia de Avaliação

Nesta seção descrevemos a metodologia que adotamos no estudo empírico do erro resultante ao se assumir, durante a estimação, um modelo de interferência incorreto. Para tanto, geramos respostas com os modelos apresentados anteriormente e estudaremos o valor do Erro Quadrático Médio (MSE) ao estimar com os diferentes estimadores, incluindo o MLE do modelo correto. Os resultados obtidos foram omitidos por razões de espaço, mas podem ser encontrados em [Azevedo et al. 2018].

Datasets. Utilizamos três redes sociais Bitcoin OTC, Enron Email e Wiki-Vote disponíveis na coleção de datasets SNAP, de Stanford[1].

Parâmetros. Cada um dos modelos é parametrizado por um vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ cujas coordenadas tem uma interpretação semelhante. O parâmetro β_0 quantifica a propensão intrínseca da população e será fixado em zero. O parâmetro β_1 quantifica a influência do tratamento atribuído ao indivíduo na sua resposta. Finalmente, o parâmetro β_2 quantifica a influência da fração dos vizinhos de um nodo em tratamento na sua resposta. Consideramos os seguintes casos: (i) SUTVA é válida: $\boldsymbol{\beta} = (0, 1, 0)$; (ii) tratamento do nodo é irrelevante: $\boldsymbol{\beta} = (0, 0, 1)$; (iii) tratamento dos vizinhos é menos relevante que do nodo: $\boldsymbol{\beta} = (0, 1, 0.5)$; (iv) tratamento dos vizinhos é tão relevante quanto do nodo: $\boldsymbol{\beta} = (0, 1, 1)$; e (v) tratamento dos vizinhos é mais relevante que do nodo: $\boldsymbol{\beta} = (0, 1, 2)$. Neste trabalho, o parâmetro adicional do modelo τ-exposure será fixado em $\tau = 0.85$.

Vetores de tratamento. Para cada combinação de grafo, vetor $\boldsymbol{\beta}$ e modelo de resposta, geramos um vetor de tratamento $\mathbf{Z}$ aleatório (binário), em que $P(Z_i = 1) = 0.5, i \in \mathcal{V}$.

[1] http://snap.stanford.edu/

Vetores de resposta. Para cada vetor de tratamento, geramos 1000 vetores de resposta $\mathbf{Y}$ para um dado modelo.

Estimadores. Para cada par $(\mathbf{Z}, \mathbf{Y})$, calculamos diversas estimativas para o ATE. Usamos o estimador de mínimos quadrados do modelo linear para estimar o ATE nos casos em que a resposta é real (i.e., $Y_i \in \mathbb{R}$). Usamos o MLE dos modelos probit e logístico para estimar o ATE nos casos em que a resposta é binária (i.e., $Y_i \in \{0, 1\}$). Além disso, os estimadores $\widehat{\text{ATE}}_{\text{SUTVA}}$ e $\widehat{\text{ATE}}_{\tau}$ foram aplicados a todos os pares $(\mathbf{Z}, \mathbf{Y})$.

6. Conclusões

Embora já exista um método para aceitar ou rejeitar a SUTVA [Saveski et al. 2017], não existe um método para determinar qual modelo de resposta melhor descreve os dados. Quando a SUTVA é rejeitada, é preciso assumir um modelo para estimar o ATE. Mesmo que o modelo assumido descrevesse perfeitamente a função de resposta dos usuários, existe um erro inerente a flutuações estatísticas, que derivamos analiticamente neste trabalho. Usando como referência este erro inerente, avaliamos os erros obtidos ao se especificar incorretamente o estimador. Observamos que alguns erros de especificação não elevaram muito o MSE (p. ex., assumir Probit, quando o modelo de resposta é Logístico, ou ainda, assumir o τ-exposure com τ elevado, quando o modelo de resposta é Linear). Contudo, o erro depende, em geral, da rede e dos parâmetros do modelo de resposta.

Referências

Azevedo, F. G., Nogueira, B. D., Murai, F., and da Silva, A. P. C. (2018). Modelos de resposta para experimentos randomizados em redes sociais de larga escala. arXiv:1803.03497.

Backstrom, L. and Kleinberg, J. M. (2011). Network bucket testing. *WWW*, pages 615–624.

Demidenko, E. (2001). Computational aspects of probit model. *Mathematical Communications*, 6(2):233–247.

Erhardt, E. B. (2009 (acessado 22 de janeiro de 2017)). *Logistic Regression and Newton-Raphson*.

Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network A/B Testing: From Sampling to Estimation. *WWW*, pages 399–409.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In *KDD*, pages 1168–1176.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., and Airoldi, E. M. (2017). Detecting Network Effects. In *KDD*, pages 1027–1035.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *KDD*, pages 329–337.

Xu, Y., Chen, N., Fernandez, A., Sinno, O., and Bhasin, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *KDD*, pages 2227–2236.