PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Rafael Felipe Bressan

INFERÊNCIA CAUSAL COM MACHINE LEARNING
uma aplicação para evasão fiscal

Belo Horizonte
2021

Rafael Felipe Bressan

INFERÊNCIA CAUSAL COM MACHINE LEARNING
uma aplicação para evasão fiscal

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1 Introdução

1.1 Contextualização

Métodos desenvolvidos na literatura de *Machine Learning* – ML – têm sido particularmente bem-sucedidos em configurações de *big data*, onde observamos informações sobre um grande número de unidades, ou muitas informações de cada unidade, ou ambos, e muitas vezes fora da configuração habitual, para um economista, de seção transversal de unidades. Para tais configurações, as ferramentas de ML estão se tornando o padrão em todas as disciplinas e, portanto, o *kit* de ferramentas do economista precisa se adaptar de acordo, preservando sempre que possível os pontos fortes da econometria aplicada.

Métodos de aprendizado supervisionado de máquina se concentram principalmente em **problemas de previsão**, dado um conjunto de dados rotulados, um resultado Y_i e alguns preditores $\mathbf{X}_i$, chamados de *features* no meio de *Machine Learning*, o objetivo é estimar um modelo em um subconjunto dos dados e avaliá-lo através de métricas de erro de previsão sobre uma amostra guardada apenas para este fim. Suponha que o verdadeiro processo gerador seja dado por $Y_i = f(\mathbf{X}_i, \epsilon_i; \theta)$, uma função estrutural f contendo os argumentos, características $\mathbf{X}_i$ do indivíduo i e idiossincrasias ϵ_i, parametrizadas internamente pelo vetor θ. Métodos de ML que fazem previsão estão preocupados em estimar $\hat{y}_i = \hat{f}(\mathbf{X}_i; \hat{\theta})$ através da minimização do erro de previsão, $\hat{e}_i = Y_i - \hat{y}_i$.

Em economia, pesquisadores usam uma ampla variedade de estratégias para tentar extrair **inferência causal** de dados observacionais. Essas estratégias são frequentemente chamadas de estratégias de identificação ou estratégias empíricas. A abordagem de resultados potenciais, às vezes referida como o Modelo Causal de Rubin[1], ganhou grande aceitação entre os economistas como a estrutura para se analisar problemas causais e tipicamente envolvem **estimação de parâmetros** de uma distribuição populacional. Acurácia da predição fora da amostra para este tipo de modelo não é relevante, o interesse do pesquisador está sobre a in-

[1] Abordagem estatística para a avaliação de causa-efeito através de resultados potenciais. Contemplada primeiramente por Neyman no início da década de 1920, Rubin (1974) ampliou seu escopo para tratamentos randomizados e também análise em observáveis.

ferência do parâmetro estimado, $\hat{\theta}$ e não sobre a previsão do resultado, $\hat{y}_i$. Desta forma, em inferência causal, deseja-se algoritmos que rendam intervalos de confiança sobre parâmetros da distribuição populacional, tarefa tipicamente realizada por modelos estatísticos.

Um fato que se deve ter em mente é que algoritmos típicos de ML como por exemplo árvores de decisão, florestas aleatórias, regressões penalizadas (LASSO e cia.), redes neurais, *boosting*, etc. usam algum tipo de **regularização** para evitar *overfitting*. Estes algoritmos se ajustam de forma bastante flexível aos dados e é esta flexibilidade que pode dar origem a um ajuste perfeito aos dados da amostra mas que em contra-partida, terá uma péssima capacidade de generalização. Para evitar este tipo de problema conhecido como *overfitting*, os algoritmos contam com esquemas de regularização, por exemplo a profundidade de uma árvore, o fator de penalização em um LASSO ou o número de neurônios de uma camada em uma rede, (MULLAINATHAN; SPIESS, 2017).

Seria tentador utilizar um modelo de ML ajustado aos dados e, considerando a qualidade de suas previsões, **supor que a estrutura modelada** esteja próxima do verdadeiro processo gerador dos dados. Infelizmente esta não é uma suposição válida, e o motivo é justamente a flexibilidade dos modelos de ML aliada a regularização pela qual estes modelos passam. Uma vez que as formas funcionais as quais os algoritmos de ML se adaptam são inúmeras, para um mesmo conjunto de dados o algoritmo pode chegar a dois modelos completamente distintos mas que ainda assim possuem poder preditivo equivalentes. Por exemplo, podemos ter os modelos $\hat{f}_1(\cdot;\hat{\theta}_1) \approx \hat{f}_2(\cdot;\hat{\theta}_2)$ e ainda assim $\hat{\theta}_1 \neq \hat{\theta}_2$. Neste caso a regularização do modelo, que é uma técnica puramente baseada nos dados, acabará determinando o modelo final a ser selecionado. Caso tenhamos um novo conjunto de dados com apenas algumas observações modificadas, ou em caso mais extremo, apenas com a mudança da semente da geração de números aleatórios para algoritmos como florestas ou redes neurais, a regularização pode optar por outro modelo com estrutura completamente diferente.

Como a inferência causal está intimamente ligada a estimação de parâmetros estruturais de um processo gerador de dados, os algoritmos tipicamente utilizados em técnicas de ML não devem ser diretamente interpretados para esta finalidade, já que estes são modelos *correlacionais*, descrevendo o que ocorre com uma variável de interesse (*outcome*) quando uma ou mais variáveis explicativas (*features*) se alteram **dado a distribuição conjunta** destes dados. A grande diferença está no fato que em modelos causais nós desejamos **intervir diretamente** em alguma *feature* quebrando, portanto, a distribuição conjunta dos dados obtida até então.

Recentemente o interesse de econometristas e estatísticos voltou-se para a extensão de modelos de ML para que estes possam estimar consistentemente parâmetros causais. O trabalho inovador e verdadeiramente um *ground breaking* de Chernozhukov et al. (2018) abriu de forma ampla as portas dos algoritmos de ML para o uso em inferência causal. Os autores propõe um método que chamam de *Double/Debiased Machine Learning* – DML, o qual faz uso de dois passos essenciais para a remoção do viés na estimação do parâmetro de interesse, i) ortogonalização de Neyman nos scores/momentos a serem ajustados, e ii) *cross-fitting*, a quebra da amostra em partições e estimações para cada uma destas, tomando a média ao final. O método resulta em estimações pontuais que são assintoticamente normais e convergem a taxa $\sqrt{N}$, o que possibilita a construção de intervalos de confiança válidos. Seguindo este trabalho, outros vêm surgindo e compondo a literatura de *Causal Machine Learning*, Chang (2020) aplica o procedimento de DML para problemas que podem ser modelados por diferença-em-diferenças, um antigo conhecido dos econometristas e método bastante utilizado para inferência causal. Singh e Sun (2020) e Syrgkanis et al. (2019) utilizam DML em conjunto com variáveis instrumentais, outro conhecido ferramental econométrico e bastante utilizado em problemas causais. Em uma linha paralela mas relacionada de trabalho, Athey e Imbens (2016), Wager e Athey (2018) e Athey, Tibshirani, Wager et al. (2019) adaptam o conhecido algoritmo de árvore de decisão e posteriormente de florestas aleatórias para serem otimizados a fazer a partição dos dados de forma "honesta" e estimar efeitos causais (e heterogêneos) livres de viés e intervalos de confiança válidos. Este algoritmo passou a ser conhecido como árvores causais ou **florestas causais**. Uma das vantagens das árvores causais é sua interpretabilidade, derivada diretamente das regras de partição da amostra de uma árvore de decisão.

1.2 O problema proposto

Considere um contribuinte que possui uma determinada renda e está sujeito ao pagamento de tributos auto-declarados e espontaneamente recolhidos. Caso este contribuinte não reporte seu tributo devido ele pode economizar parte de sua renda, entretanto, estará sujeito ao risco de detecção da evasão o qual o sujeitará a cobrança do tributo devido com juros e multa. Indivíduos racionais analisam a relação de custo-benefício desta evasão e optam pela escolha ótima, Allingham e Sandmo (1972). Para garantir o cumprimento da lei, a fiscalização deve

ser suficientemente presente: para um dado nível de sanções legais, o contribuinte deve esperar uma alta probabilidade de ser fiscalizado, e as sanções devem ser altas para uma determinada cobertura fiscal. Uma outra vertente da literatura, Sandmo (2005), argumenta que o cumprimento das leis também é baseado em instituições informais como normas sociais. A questão central levantada é portanto, se a fiscalização pode se valer destas alternativas comportamentais para ampliar o nível de cumprimento das obrigações tributárias.

Devido ao sigilo fiscal presente na legislação brasileira e a escassez de dados fiscais de livre acesso e estudos experimentais sobre o assunto, optou-se neste trabalho por utilizar os dados do experimento realizado por Fellner, Sausgruber e Traxler (2013). Eles realizam um experimento de campo na Áustria para avaliar estratégias alternativas para fazer cumprir as leis locais. O experimento varia o texto de correspondências enviadas para possíveis evasores das taxas de licença de TV, uma espécie de imposto de caráter auto-declaratório e obrigatório a todos os residentes que possuem aparelho de TV em sua moradia.

Nosso objetivo é obter novos *insights* sobre essa questão testando diferentes estratégias para induzir ao cumprimento legal em um experimento de campo natural. O problema de fiscalização existe pois os sinais de transmissão de TV e rádio são públicos e podem ser recebidos sem o pagamento da taxa. Os evasores encaram um risco de detecção não desprezível e a ameaça de multas consideráveis. A autoridade de fiscalização da Áustria (equivalente a Receita Federal do Brasil) concedeu aos autores o acesso aos dados de mais de 50.000 indivíduos que foram identificados como evasores em potencial. Em cooperação com a autoridade, foram enviadas correspondências para 95 % desta amostra **sorteados aleatoriamente** durante o ano de 2005. Os demais indivíduos serviram como grupo de controle e não receberam nenhuma correspondência.

Dada a configuração deste experimento, aleatoriedade no recebimento da correspondência e diferentes tipos de textos no conteúdo desta, será possível analisar os efeitos de diferentes tipos de "tratamento" (conteúdo da carta), como ameaça, persuasão moral e informação e suas interações no comportamento evasivo dos contribuintes.

De fato, esta análise foi realizada pelos autores por meio de regressões lineares e no presente trabalho, proponho o uso do mesmo conjunto de dados aliado a métodos de ML para revisar os efeitos causais originalmente estimados e também estender os resultados para a análise de efeitos heterogêneos dentro de sub-grupos da população.

2 Coleta e Tratamento de Dados

Neste trabalho optou-se por utilizar dados de um estudo fiscal realizado e publicado em periódico de destaque internacional, *Journal of the European Economic Association*, devido ao sigilo fiscal presente na legislação brasileira e a escassez de dados fiscais de livre acesso e estudos experimentais sobre o assunto. Fellner, Sausgruber e Traxler (2013) realizaram um estudo experimental de campo, essencial para se extrair a inferência causal que se deseja, e publicaram os resultados, incluindo a base de dados no seguinte link: `https://doi.org/10.1111/jeea.12013` sob a seção *Supporting Information*.

Estes dados foram baixados e então trabalhados localmente. A seguir a tabela original dos autores sobre as variáveis e seus tipos (em Inglês).

Tabela 1 – Variáveis e descrições

Variável	Descrição
treatment	Treatment variable: 0 (No_mailing), 1 (Baseline), 2 (Threat), 3 (Info), 4 (Info&Threat), 5 (Moral), 6 (Moral&Threat)
mailing	0 (No_mailing), 1 (treatment 1-6)
threat	Dummy variable equal to 1 if treatment is 2, 4, or 6.
info	Dummy variable equal to 1 if treatment is 3 or 4.
appeal	Dummy variable equal to 1 if treatment is 5 or 6.
i_tinf	Dummy variable equal to 1 if treatment is 4.
i_tapp	Dummy variable equal to 1 if treatment is 6.
resp_A	Dummy variable equal to 1 if individual registered, i.e. started paying the fee within the first 50 days after sending the mailing.
resp_A_25	Dummy variable equal to 1 if individual registered within the first 25 days after sending the mailing.
resp_A_75	Dummy variable equal to 1 if individual registered within the first 75 days after sending the mailing.
resp_A_100	Dummy variable equal to 1 if individual registered within the first 100 days after sending the mailing.
resp_B	Dummy variable equal to 1 if individual updated contract details within 50 days.
resp_all	Dummy variable equal to 1 if individual registered or updated contract details.
deregistration	Dummy variable equal to 1 if individual deregistered (available only for a subgroup of 2,291 individuals, see Section 5.3)
delivered	Dummy variable equal to 1 if the mailing was delivered to the individual (see Footnote 13)
evasion_1	Dummy variable equal 1 if individual lives in municipality with a high evasion rate (top quartile), see Section 5.5.

Tabela 1 – Variáveis e descrições (continuação)

Variável	Descrição
evasion_2	Dummy variable equal 1 if individual lives in municipality with a high evasion rate (top tercile), see Section 5.5.
threat_evasion_D1	Interaction threat × evasion_1
appeal_evasion_D1	Interaction appeal × evasion_1
info_evasion_D1	Interaction info × evasion_1
threat_evasion_D2	Interaction threat × evasion_2
appeal_evasion_D2	Interaction appeal × evasion_2
info_evasion_D2	Interaction info × evasion_2
gender	Dummy variable equal to 1 if individual is male.
age	Age of the individual (only available for a subsample of 16,281 recipients, see Table 1).
pop2005	Number of inhabitants in residence municipality 2005.
pop_density2005	Number of inhabitants normalized by size (in km2) in residence municipality 2005.
compliance	Pre-experimental compliance rate (registered no. of households normalized by total no. of households per municipality, see Table 1), excl. secondary residences.
compliance_t	Pre-experimental compliance rate (registered no. of households normalized by total no. of households per municipality), incl. secondary residences.
vo_r	Vote share right parties (parliamentary election) per municipality 2006.
vo_cr	Vote share center right parties (parliamentary election) per municipality 2006.
vo_cl	Vote share center left parties (parliamentary election) per municipality 2006.
vo_l	Vote share left parties (parliamentary election) per municipality 2006.
inc_aver	Average yearly gross income in Euro per municipality 2003.
edu_aver	Average level of education per municipality 2001.
edu_lo	Share of inhabitants with low education per municipality 2001.
edu_mi	Share of inhabitants with middle education per municipality 2001.
edu_hi	Share of inhabitants with high education per municipality 2001.
age_aver	Average age of inhabitants per municipality 2001.
age0_30	Share of inhabitants with age below 30 per municipality 2001.
age30_60	Share of inhabitants with age between 30 and 60 per municipality 2001.
fam_singl	Share of single households per municipality 2001.
fam_marri	Share of households with married couples per municipality 2001.
fam_divor_widow	Share of households with divorced or widowed inhabitants per municipality 2001.
nat_A	Share of inhabitants with Austrian Nationality per municipality 2001.
nat_EU	Share of inhabitants with Nationality from a European Union member country per municipality 2001.
nat_nonEU	Share of inhabitants with Nationality from country outside European Union per municipality 2001.

Tabela 1 – Variáveis e descrições (continuação)

Variável	Descrição
nat_nonA	Share of inhabitants with Nationality outside Austria per municipality 2001.
j_employ	Share of employed inhabitants per municipality 2001.
j_unempl	Share of unemployed inhabitants per municipality 2001.
j_retire	Share of retired inhabitants per municipality 2001.
j_house	Share of homemakers per municipality 2001.
j_studen	Share of students per municipality 2001.
rel_kath	Share of Roman Catholic denomination per municipality 2001.
rel_evan	Share of Evangelic denomination per municipality 2001.
rel_isla	Share of Islamic denomination per municipality 2001.
rel_obk	Share of no religious denomination per municipality 2001.
rel_orth_other	Share of other Christian Orthodox and other denominations per municipality 2001.
pers2	Share of 2-person households per municipality 2001.
pers3	Share of 3-person households per municipality 2001.
pers4	Share of 4-person households per municipality 2001.
pers5more	Share of 5 and more-person households per municipality 2001.
coverage	Fraction of mailings relative to the population per municipality 2005.
bgld	Dummy variable equal to 1 if Burgenland.
kaern	Dummy variable equal to 1 if Carinthia.
noe	Dummy variable equal to 1 if Lower Austria.
ooe	Dummy variable equal to 1 if Upper Austria.
salzbg	Dummy variable equal to 1 if Salzburg.
steierm	Dummy variable equal to 1 if Styria.
tirol	Dummy variable equal to 1 if Tirol.
vlbg	Dummy variable equal to 1 if Vorarlberg.
wien	Dummy variable equal to 1 if Vienna.
schober	Dummy variable equal to 1 if individual's address is not from residential register but from private marketing company.

2.1 Tratamento

Uma vez coletados os dados e descompactados em máquina local, a base de dados disponibilizada estava em formato compatível com o *software* Stata, tipicamente utilizado por econometristas.

No presente trabalho, a parte de análise exploratória dos dados e uma parcela da modelagem, especialmente a parte de replicação dos resultados dos autores originais, foi utilizada

a linguagem R, enquanto que a modelagem causal através de *Machine Learning* foi utilizado Python. Felizmente, ambas as linguagens possuem pacotes disponíveis que fazem a leitura de arquivos no formato do Stata (i.e. *.dta*), o pacote *haven* para R e o próprio *pandas* no Python foram utilizados.

Uma vez importados os dados para o ambiente de programação e análise, a base de dados não precisou de maiores tratamentos uma vez que os autores já disponibilizaram a base devidamente limpa e formatada. Apenas a variável *treatment* necessitou ser convertida em variável categórica pois, possui mais de dois níveis diferentes. Demais variáveis quando binárias (i.e. *dummy variable*) foram tratadas como inteiras, porém com apenas dois valores, zero e um. Outras variáveis assumiram o tipo numérico (i.e. *double*).

A seguir apresenta-se uma tabela com as quantidades de indivíduos que receberam cada um dos níveis de "tratamento"[1], sendo que o grupo de controle é aquele que não recebeu correspondência alguma e o tratamento básico (*baseline*, que recebeu a carta padrão) é referido por "Correio".

Tabela 2 – Distribuição dos tratamentos na amostra.

Tratamento	Descrição	Observações	Proporção
T0	Sem Correio	2586	0.0512099
T1	Correio	7984	0.1581053
T2	Ameaça	7821	0.1548774
T3	Info	7998	0.1583825
T4	Info&Ameaça	8101	0.1604222
T5	Moral	8084	0.1600855
T6	Moral&Ameaça	7924	0.1569171

Também foi realizada uma análise de dados faltantes na amostra, na Tabela 3 apresenta-se todas as variáveis que não possuem alguma observação. Destas, destaca-se a idade, que será uma variável relevante para analisar possíveis efeitos heterogêneos do tratamento e deve ser devidamente tratada posteriormente. Outra variável com dados faltantes que merece atenção é a entrega da correspondência (*delivered*), isto porque esta variável é necessária para a verificação de possível atrito[2] no experimento. A variável *delivered* é binária, indicando se a pessoa recebeu

[1] Termo tipicamente utilizado em inferência causal. No presente trabalho o tratamento refere-se a receber uma correspondência da autoridade fiscal.

[2] Atrito em inferência causal refere-se a falta de observação na resposta de indivíduos que foram selecionados no experimento. No presente caso, contribuintes que deveriam receber a correspondência mas não foram encontrados. Atrito pode comprometer a aleatorização do experimento e gerar viés na inferência.

a correspondência (código 1) ou não. Um valor NA para esta variável não está bem definido a partir da Tabela 1 e será investigado no próximo capítulo.

Tabela 3 – Dados faltantes na amostra.

Variável	No. Faltantes	Completude
deregistration	48207	0.05
delivered	2586	0.95
evasion_1	9491	0.81
evasion_2	9491	0.81
threat_evasion_D1	9491	0.81
appeal_evasion_D1	9491	0.81
info_evasion_D1	9491	0.81
threat_evasion_D2	9491	0.81
appeal_evasion_D2	9491	0.81
info_evasion_D2	9491	0.81
age	32466	0.36
coverage	9491	0.81

Nota: Completude refere-se a proporção de linhas preenchidas contra faltantes, e varia de zero a um.

Desta forma, a base de dados para o presente trabalho conta com 50.498 observações e 73 variáveis de tipos numéricos, binárias e categóricas. Dentre as variáveis relevantes temos a variável de resposta ao tratamento, *resp_A*, binária e que indica se o indivíduo se registrou junto a autoridade para o pagamento da taxa no prazo de cinquenta dias após o envio das correspondências. As variáveis de tratamento são ao todo sete e representam o grupo de controle (T0) que não recebeu correspondência, e os grupos de tratamento (T1-T6) que receberam carta padrão (T1) e cartas que **acrescentam** a padrão uma ameça, informação, ameaça e informação, apelo moral ou, ameaça e apelo moral (T2-T6). As demais variáveis são controles (conhecidas também como covariadas ou atributos) e são utilizadas tanto para verificar a aleatorização do experimento quanto para subdividir a amostra por características observáveis e então analisar possíveis efeitos heterogêneos do tratamento.

3 Análise e Exploração dos Dados

Como é de costume na literatura de inferência causal, o primeiro procedimento exploratório dos dados realizado foi o aferimento da aleatorização do experimento. Uma boa aleatorização dos indivíduos que receberão o tratamento deve refletir em distribuições iguais das demais covariadas entre os níveis de tratamento. Na prática o que se faz é um teste de médias ou uma comparação através de diferença normalizada entre grupos de tratamento para diversas variáveis de controle que se julga importantes para explicar a variável de resultado, (IMBENS; RUBIN, 2015). A Tabela 4 apresenta os resultados.

Tabela 4 – Balanceamento de características individuais e por município por tipo de tratamento.

Tratamento	Descrição	Gênero	Idade	Renda	População	Dens. pop.	Compliance
T0	Sem Correio	0.6458	48.0170	20928.4068	45815.2715	8.1711	0.9355
T1	Correio	0.6338	47.9969	20878.9958	43377.1935	8.5625	0.9352
T2	Ameaça	0.6367	47.9931	20901.1614	44542.5883	7.9605	0.9346
T3	Info	0.6260	48.0300	20882.6636	43903.0189	8.1142	0.9347
T4	Info&Ameaça	0.6335	48.0051	20879.6138	43319.4736	8.3540	0.9352
T5	Moral	0.6251	47.9982	20888.4584	44301.3718	8.4832	0.9343
T6	Moral&Ameaça	0.6422	47.9904	20876.3062	43610.1972	8.0468	0.9343
Anova:	p-values	0.1715	0.3993	0.9393	0.7577	0.5795	0.8614

Nota: Gênero igual a zero para mulher. Demais variáveis são denominadas em nível municipal, por exemplo Idade refere-se a idade média dos habitantes do município de residência do indivíduo.

As covariadas escolhidas foram: gênero (0 para mulher), idade, população média do município de residência, densidade populacional e nível médio de *compliance* com a legislação neste município, com dados referentes ao ano de 2005. A última linha da tabela apresenta os p-valores para um teste anova de diferença de médias entre todos os grupos de tratamento. Pode-se verificar que em nenhuma das covariadas escolhidas rejeita-se a hipótese nula de igualdade de médias e portanto, conclui-se que a aleatorização do experimento foi bem realizada.

Voltando-se para o problema de atrito apontado anteriormente, foram analisados se os valores faltantes da variável indicadora de entrega da correspondência estavam associados a indivíduos que efetivamente haviam sido escolhidos para algum tipo de tratamento e o resultado foi negativo, ou seja, apenas os indivíduos no grupo de controle T0 estão associados a valores faltantes da variável *delivered*. Conforme Tabela 5 não houve dados faltantes para a entrega de correspondências nos tratamentos, todos os valores NA anteriormente detectados eram de

indivíduos no grupo de controle, e portanto, NA está corretamente associado a este grupo ao qual não foram enviadas correspondências.

Tabela 5 – Taxa de atrito por tratamento.

Tratamento	Descrição	Cartas	Entregues NA	Não Entregues	Taxa Atrito
T1	Correio	7984	0	1126	0.1410
T2	Ameaça	7821	0	1127	0.1441
T3	Info	7998	0	1173	0.1467
T4	Info&Ameaça	8101	0	1141	0.1408
T5	Moral	8084	0	1164	0.1440
T6	Moral&Ameaça	7924	0	1174	0.1482

Nota: Na média total a taxa de atrito foi de 0.1441 e não houve diferença entre tratamentos, como aponta o teste qui-quadrado de Pearson para dados de contagem, com p-valor de 0.7935.

Entretanto, como se pode verificar na coluna "Não Entregues" da tabela acima, **houve de fato atrito** na entrega das correspondências. Uma média de 14,41% das correspondências não pode ser entregue aos destinatários. As taxas de atrito são estatisticamente iguais entre todos os grupos de tratamento e, portanto, caso o atrito ocorra de forma aleatória na amostra, não deve impactar significativamente nos resultados alcançados pelo experimento, (DUFLO; GLENNERSTER; KREMER, 2007).

Infelizmente, mesmo que as taxas de atrito sejam iguais entre os grupos de tratamento, esta pode não ser totalmente aleatória, mas sim determinada em função de variáveis, observadas ou não, dos indivíduos. Embora a aleatorização do experimento garanta inicialmente a independência de resultados potenciais nos grupos de tratamento e controle, ela não se mantém após o atrito não aleatório, o que pode causar viés nas estimativas do efeito causal se não for levada em conta. Neste ponto faz-se necessária a investigação da distribuição das covariadas selecionadas em relação aos **indivíduos que atritaram**.

Tabela 6 – Análise de atrito. Balanceamento de variáveis selecionadas

Tratamento	Gênero	Idade	Renda	População	Dens. pop.	Compliance
T0	0.6458	48.0170	20928.4068	45815.2715	8.1711	0.9355
T1	0.6403	47.7868	21100.3921	52084.9822	7.6001	0.9322
T2	0.6211	47.7127	21106.0117	48882.0302	6.5860	0.9337
T3	0.6138	47.8580	21077.8894	51027.8338	6.6317	0.9313
T4	0.6240	47.8056	20945.2352	48251.5259	6.5957	0.9318
T5	0.6177	47.7952	20864.3756	43273.7019	6.3919	0.9308
T6	0.6320	47.8117	20966.9995	46539.3467	6.4614	0.9324
Anova p-valor	0.4319	0.0000	0.0095	0.0936	0.0094	0.1122

Nota: Gênero igual a zero para mulher. Demais variáveis são denominadas em nível municipal, por exemplo Idade refere-se a idade média dos habitantes do município de residência do indivíduo.

De fato a Tabela 6 demonstra que a idade média, renda e a densidade populacional do município de residência possuem influência significativa na determinação do atrito, e portanto, este não ocorreu de forma aleatória.

Uma outra forma que se pode utilizar para averiguar a distribuição das covariadas entre os grupos de tratamento é através de histogramas. Na Figura 1 é apresentado no painel superior o histograma do grupo controle, T0 versus tratamento T6 para a covariada de densidade populacional determinado pela aleatorização do experimento. No painel inferior o mesmo histograma porém contra os indivíduos designados para o tratamento T6 mas que atritaram. A diferença nas distribuições fica evidente.

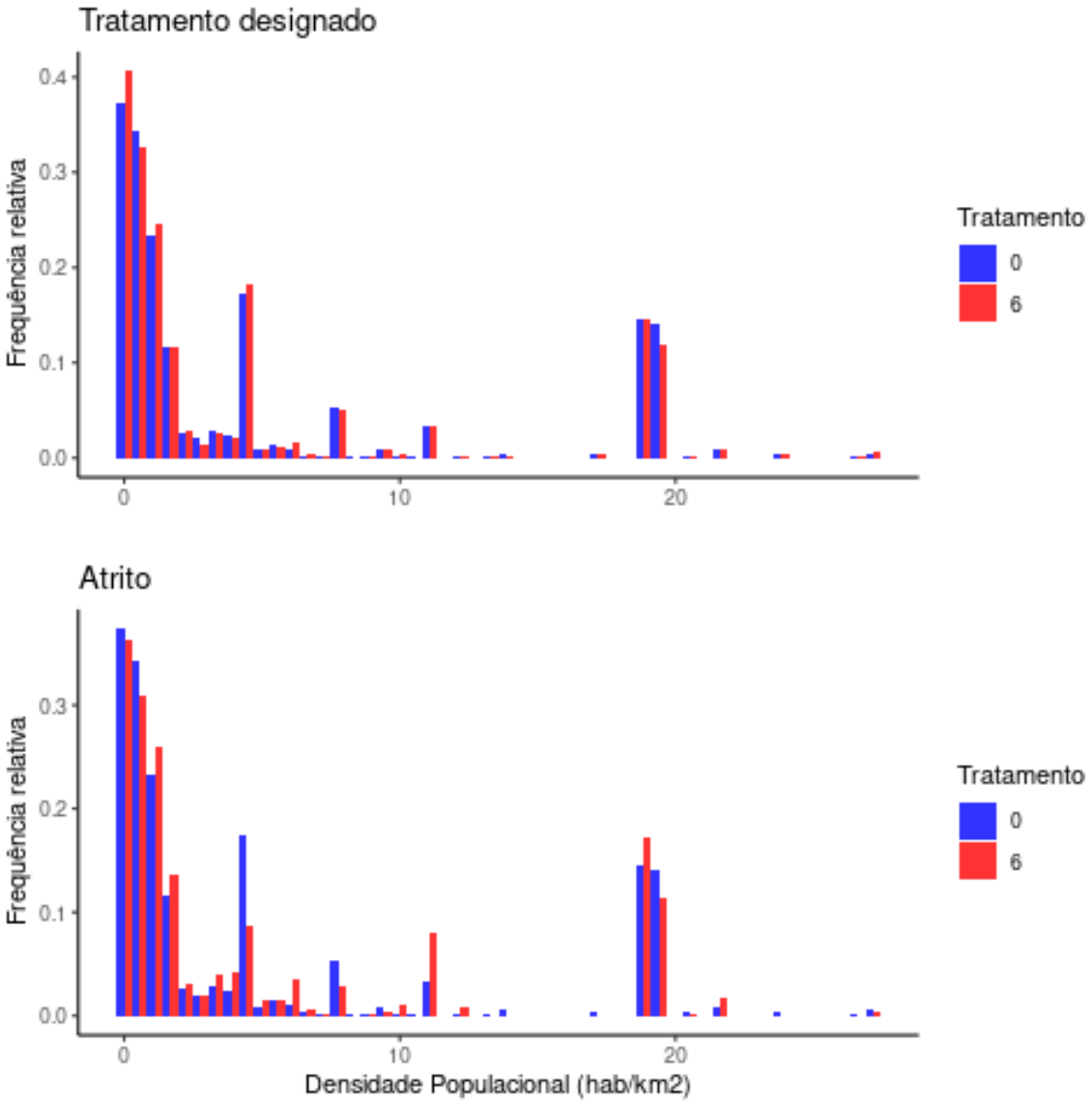

Figura 1 – Histogramas comparativos da distribuição de designação de tratamento contra atrito.

4 Criação de Modelos de Machine Learning

Neste capítulo será formulado o modelo de inferência causal o qual iremos estimar através de métodos de *machine learning*. Primeiramente traremos alguns conceitos essenciais de causalidade e definições que serão necessárias para a boa compreensão do restante deste trabalho. Começamos com a formulação de resultados potenciais e definição dos estimandos de interesse.

4.1 Resultados Potenciais

A primeira noção é a de resultados potenciais para um indivíduo, cada resultado correspondendo a um dos níveis de tratamento ou manipulação. Cada um desses resultados potenciais é *a priori* observável, no sentido de que poderia ser observado se o indivíduo recebesse o nível de tratamento correspondente. Porém, *a posteriori*, ou seja, uma vez aplicado um tratamento, no máximo um resultado potencial pode ser observado.

O fato que apenas um resultado potencial pode ser observado nos dados leva ao que é chamado de **problema fundamental da inferência causal**, ou seja, nunca observamos o resultado contrafactual de um nível de tratamento em algum indivíduo. Entretanto, para se aferir efeitos de tratamento é necessária a comparação entre o resultado potencial do tratamento ministrado, que é observável, contra o resultado potencial que o indivíduo apresentaria *caso fosse submetido a outro nível de tratamento*, resultado contrafactual que é, portanto, não observável.

Vamos estabelecer a notação necessária para nosso problema de inferência causal. Seja $\mathcal{I}$ o conjunto de indivíduos na população de tamanho N, $\mathcal{I} = \{1, \ldots, N\}$, com um indivíduo típico, $i \in \mathcal{I}$. Inicialmente simplificando, assumimos um tratamento binário para o indivíduo i, $D_i \in \{0, 1\}$. Resultado potencial em função do nível de tratamento ao qual o indivíduo foi submetido é denotado por $Y_i(D_i)$. Portanto, os resultados potenciais observados e não observados são determinados por:

$$Y_i^O = Y_i(D_i) = \begin{cases} Y_i(0) & \text{se } D_i = 0, \\ Y_i(1) & \text{se } D_i = 1. \end{cases}$$

$$Y_i^{NO} = Y_i(1 - D_i) = \begin{cases} Y_i(0) & \text{se } D_i = 1, \\ Y_i(1) & \text{se } D_i = 0. \end{cases}$$

assim sendo, o valor observado da variável de interesse (i.e. resultado) para um indivíduo pode ser escrito em função de seus resultados potenciais.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0) \tag{4.1}$$

4.2 Estimandos

Ao falar sobre os efeitos do tratamento, o conceito de resultados potenciais talvez seja o mais importante. Teoricamente, se fosse possível observar o mesmo indivíduo em dois estados diferentes (com e sem tratamento), seria possível falar sobre o efeito para o i-ésimo indivíduo. Porém em muitos casos isso não é viável e, normalmente de menor interesse para políticas públicas ou economia. Desta forma, definimos os efeitos causais em termos médios, como por exemplo efeito médio do tratamento (ATE - *average treatment effect*). É preciso então obter um grupo de indivíduos que são tratados e outro grupo que não (os controles). Um grupo de controle bem definido é essencial para estas comparações. Suponha que um determinado tratamento afeta toda a população de interesse, mesmo que indiretamente. Neste caso não temos um grupo de controle bem definido e portanto, não há como definir um efeito do tratamento. É necessária uma hipótese sobre os efeitos de tratamentos que garanta a existência de um grupo de controle bem definido, esta hipótese é conhecida por SUTVA - *Stable unit treatment value assumption*. A SUTVA incorpora a ideia que não existe interferência entre os indivíduos tratados e não tratados, em outras palavras, não pode haver efeitos de transbordamento (*spillover*) do tratamento de algum indivíduo para outro que esteja no grupo de controle.

A hipótese SUTVA é essencial para a identificação de efeitos médios de tratamento e será mantida ao longo deste trabalho. Formalmente, seja $\mathbf{D}$ um vetor de alocação de tratamento para os N indivíduos do programa. Utilizando a notação de resultados potenciais, a SUTVA exige que, para quaisquer duas alocações $\mathbf{D}$ e $\mathbf{D}'$,

$$Y_i(\mathbf{D}) = Y_i(\mathbf{D}') \quad \text{se } D_i = D_i' \tag{4.2}$$

onde D_i é o nível do tratamento recebido pelo i-ésimo indivíduo. Ou seja, o resultado potencial do indivíduo i depende somente do nível de tratamento recebido por este indivíduo. Uma vez que aceitamos como válida a hipótese SUTVA, podemos definir alguns estimandos de interesse.

Os principais efeitos médios de tratamento que em geral estamos interessados são: Efeito Médio do Tratamento – ATE – (*Average Treatment Effect*), Efeito Médio do Tratamento nos Tratados – ATT – (*Average Treatment Effect on the Treated*) e suas versões condicionadas a um conjunto de covariadas (ou *features*), CATE e CATT, onde a inicial "C" vem de condicional. Suponha uma população de interesse onde a expectância sobre uma determinada medida de probabilidade esteja bem definida, resultados potenciais para um tratamento binária estão definidos como anteriormente e existe um vetor de covariadas $\mathbf{X}$, então, em termos populacionais podemos definir os estimandos.

$$\theta_{ate} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{4.3}$$

$$\theta_{att} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \tag{4.4}$$

$$\theta_{cate} = \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}] \tag{4.5}$$

$$\theta_{catt} = \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}, D_i = 1] \tag{4.6}$$

Suponha agora que apenas a designação ao tratamento foi aleatorizada, mas sua efetiva administração pode depender de características individuais, exatamente como ocorreu em nosso experimento. Seja $\mathcal{T} = \{0, 1\}$ o conjunto binário para designação com $t \in \mathcal{T}$ um nível de designado. Quando a conformidade ao tratamento escolhido é cumprida apenas parcialmente, existirá uma diferença entre ser designado para um nível de tratamento t e receber o nível de tratamento $d \in \mathcal{D} = \{0, 1\}$ e neste caso teremos quatro combinações identificando o comportamento dos indivíduos com relação ao tratamento.

4.2.1 Conformidade parcial ao tratamento

No experimento que estamos analisando, (FELLNER; SAUSGRUBER; TRAXLER, 2013), conforme mostrado no Capítulo 3, a simples designação para um dos grupos de tratamento não garantiu que os indivíduos efetivamente receberam sua correspondência, ou seja, houve atrito no experimento. Nesta situação a designação ao tratamento, que foi aleatorizada e

é desconhecida dos recipientes, pode ser utilizada como **variável instrumental**[1], enquanto que recebimento de tratamento passa a ser a variável de interesse primário.

O atrito em experimentos aleatórios é um caso particular de não conformidade com a designação ao tratamento. Enquanto que no caso geral pode-se ter indivíduos designados ao grupo controle recebendo tratamento, o atrito que estamos analisando permite apenas que indivíduos designados ao tratamento não o recebam. A probabilidade de alguém do grupo controle receber tratamento é nula, $E[D_i(0)] = E[D_i|T_i = 0] = 0$. Esta condição tem implicações para a interpretação do efeito causal, como será visto adiante.

Vamos expandir a notação de resultados potenciais para acomodar tanto a designação ao tratamento $T_i \in \{0, 1\}$ quanto o recebimento de tratamento $D_i \in \{0, 1\}$. Formalmente pode-se entender que o recebimento potencial de tratamento depende da designação exatamente da mesma forma que resultados potenciais. Assim $D_i(0)$ é o nível de tratamento **recebido** pelo indivíduo i quando este **não foi designado** para o grupo de tratados. Importante ressaltar que o recebimento potencial continua a ser uma variável binária neste *setup* simplificado, ou seja, $D_i(T_i) \in \{0, 1\}$ para todo $T_i \in \{0, 1\}$. Os resultados potenciais passam a depender tanto da designação do tratamento, quanto do recebimento deste (que por sua vez depende da designação), ou seja, $Y_i(T_i, D_i)$.

Retornando ao recebimento potencial de tratamento, $D_i(T_i)$ este pode assumir dois valores para uma dada designação, portanto o indivíduo pode ou não receber o tratamento se for designado para tal, e também pode ou não receber tratamento mesmo que não seja designado. Temos, portanto, quatro situações distintas no comportamento do indivíduo em receber o tratamento.

$$\text{Indivíduo } i \begin{cases} \text{Always taker se } D_i = 1, T_i \in \{0, 1\}, \\ \text{Never taker se } D_i = 0, T_i \in \{0, 1\}, \\ \text{Defier se } D_i = 0, T_i = 1 \text{ e } D_i = 1, T_i = 0 \\ \text{Complier se } D_i = 0, T_i = 0 \text{ e } D_i = 1, T_i = 1 \end{cases}$$

Um indivíduo *always taker* é aquele que sempre recebe o tratamento, independentemente de sua designação. De modo oposto o *never taker* é aquele indivíduo que nunca recebe o tratamento. Já o *defier*, ou desafiante, é aquele indivíduo que contraria sua prescrição e não toma o tratamento quando designado mas o faz assim que não está no grupo de selecionados ao

[1] O uso de variáveis instrumentais tem longa tradição na econometria. Para um tratamento sobre sua utilização no contexto de inferência causal em experimentos aleatorizados os textos de Angrist e Pischke (2008) e Imbens e Rubin (2015) podem ser consultados.

tratamento. Por fim o caso de maior interesse, o *complier* ou **obediente** é aquele indivíduo que respeita sua designação ao tratamento.

Nesta situação podemos definir outros dois estimandos de interesse, o Efeito da Intenção de Tratar – ITT – (*Intention to Treat Effect*) e o Efeito Médio do Tratamento Local – LATE – (*Local Average Treatment Effect*).

$$ITT = \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \tag{4.7}$$

$$LATE = \frac{ITT}{\mathbb{E}[D_i(1) - D_i(0)]} \tag{4.8}$$

O ITT é facilmente calculado ignorando-se o *status* de cumprimento ao tratamento designado e concentrando-se apenas na designação, ou seja, o ITT é calculado da mesma forma que o ATE no caso de conformidade perfeita ao tratamento.

Caso a conformidade seja parcial de modo que ninguém do grupo de controle tenha acesso ao tratamento, o denominador na equação do $LATE$ se reduz a $E[D_i(1) - D_i(0)] = E[D_i|T_i = 1] = P(D_i|T_i = 1) = \pi_{co}$, ou seja, a proporção de obedientes na população. Assim, no caso do presente trabalho, o $LATE$ pode ser calculado diretamente a partir da intenção de tratar dividindo-a pela proporção de obedientes, que nada mais é que $\pi_{co} = 1 -$ taxa de atrito.

No experimento em análise não existem *always takers*. Não existe a possibilidade de alguém que não tenha sido escolhido para algum dos grupos de tratamento T1-T6 receber uma correspondência. Este caso configura o que se chama de *one-sided partial compliance* e a estimativa de $LATE$ se torna igual ao efeito do tratamento sobre os tratados, ATT, em um resultado demonstrado em Bloom (1984).

Suponha que $E[D_i(0)] = E[D_i|T_i = 0] = 0$ e portanto $E[D_i(1) - D_i(0)] = P(D_i|T_i = 1) = \pi_{co}$. O ITT pode ser decomposto da seguinte forma:

$$
\begin{aligned}
ITT &= E[Y_i(1, D_i) - Y_i(0, D_i)] \\
&= E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \\
&= E\left[Y_i(T_i, 0) + [Y_i(T_i, 1) - Y_i(T_i, 0)]D_i|T_i = 1\right] - E[Y_i(0, 0)|T_i = 0] \\
&= E\left[Y_i(0, 0) + [Y_i(T_i, 1) - Y_i(T_i, 0)]D_i|T_i = 1\right] - E[Y_i(0, 0)|T_i = 1] \\
&= E[[Y_i(T_i, 1) - Y_i(T_i, 0)]D_i|D_i = 1, T_i = 1]\underbrace{P(D_i = 1|T_i = 1)}_{\pi_{co}} \\
&+ E[[Y_i(T_i, 1) - Y_i(T_i, 0)]\cancelto{0}{D_i}|D_i = 0, T_i = 1]P(D_i = 0|T_i = 1) \\
&= E[Y_i(T_i, 1) - Y_i(T_i, 0)|D_i = 1]\pi_{co}
\end{aligned}
$$

onde fizemos uso de $Y_i(T_i, D_i) \perp T_i$ uma vez que a designação ao tratamento foi aleatorizada e da hipótese de exclusão da variável instrumental, $Y_i(1, D_i) = Y_i(0, D_i)$, a ser detalhada abaixo.

Assim, o efeito local do tratamento dado pela Equação 4.8 se resumirá a $LATE = E[Y_i(T_i, 1) - Y_i(T_i, 0)|D_i = 1] = ATT$. Ou seja, ao calcularmos o $LATE$ em nossas estimativas, estaremos de fato interpretando o efeito do tratamento nos tratados, um resultado mais forte e desejável do ponto de vista de política pública.

4.3 Hipóteses de identificação

Infelizmente a hipótese SUTVA é necessária, porém não suficiente para a identificação de efeitos causais. É necessária também a Hipótese de Independência Condicional – CIA – (*Conditional Independence Assumption*), por vezes também conhecida como hipótese de *unconfoundedness*.[2] A CIA afirma que, condicionado às características observadas, $\mathbf{X}_i$, os resultados potenciais são independentes do tratamento D_i:

$$\{Y_i(1), Y_i(0)\} \perp D_i|\mathbf{X}_i \tag{4.9}$$

ou seja, condicional a $\mathbf{X}_i$ a escolha do tratamento é tão boa quanto uma escolha aleatória em relação aos resultados potenciais. Logicamente, a melhor maneira de garantir esta hipótese é

[2] Não encontrei uma boa tradução para *unconfoundedness* a não ser "não confusão". Uma variável *confounding* é aquela que está relacionada tanto ao tratamento quanto ao resultado e portanto, se não for levada em conta no modelo causal, poderá enviesar as estimativas de efeito.

efetivamente aleatorizar o tratamento. Não é para menos que os experimentos randomizados são conhecidos como o *padrão-ouro* para inferência causal.

No caso de variáveis instrumentais, como a designação do tratamento quando este apresenta atrito, devemos fazer algumas hipóteses adicionais para a correta identificação do efeito local do tratamento.

Primeiramente uma hipótese de exclusão sobre todos aqueles indivíduos com comportamento não obediente (*always takers*, *never takers* e desafiantes). A designação para tratamento não afeta a variável de interesse quando o indivíduo não recebe tratamento.

$$Y_i(0,0) = Y_i(1,0), \text{ para todo indivíduo não obediente} \tag{4.10}$$

É importante notar que esta hipótese não se confunde com a aleatorização da designação ao tratamento, ou seja, $Y_i \perp T_i$ não implica na relação dada em 4.10. Esta hipótese deve ser defendida em termos peculiares a aplicação em análise e será adequada em algumas situações, enquanto que em outras pode não ser plausível.

O grupo de obedientes também possui sua própria hipótese de exclusão. O instrumento deve ser independente dos resultados potenciais para todos os níveis de tratamento.

$$Y_i(0,d) = Y_i(1,d), \forall d \in \mathcal{D} \text{ e indivíduos obedientes.} \tag{4.11}$$

Esta restrição assume que $Y_i(0,0) = Y_i(1,0)$ **e** $Y_i(0,1) = Y_i(1,1)$ para os *compliers* e portanto, o efeito de tratamento se deve somente ao efetivo recebimento deste e não a intenção de tratar.

Estas restrições de exclusão, assim como as hipóteses SUTVA e CIA possibilitam, a partir dos dados observados, estimar os efeitos causais da seção 4.2. Sem elas, não é possível determinar efeito causal de algum tratamento em uma variável de interesse pois, pode existir uma terceira variável U que interfere tanto no tratamento quanto no resultado e portanto os dados irão refletir uma mistura do efeito causal e desta interferência.

4.4 *Machine Learning* sem Viés

Mostraremos a seguir como um modelo *off-the-shelf* de ML pode causar viés em estimativas de efeitos causais e a solução derivada por Chernozhukov et al. (2018) para eliminá-

lo, conhecida como *Double/Debiased Machine Learning*. Suponha que nosso modelo para a variável de interesse seja uma regressão parcialmente linear, (ROBINSON, 1988):

$$Y = \theta_0 T + g_0(\mathbf{X}) + U, \qquad E[U|\mathbf{X}, T] = 0, \tag{4.12}$$

$$T = f_0(\mathbf{X}) + V, \qquad E[V|\mathbf{X} = 0,] \tag{4.13}$$

onde Y é a variável de interesse, T o tratamento, o vetor de covariadas $\mathbf{X}$ explica tanto o resultado quando a seleção para tratamento e U, V são erros aleatórios. As funções g_0 e f_0 mapeiam o vetor de covariadas em $\mathbb{R}$ e são potencialmente não lineares. Considerando que a equação 4.12 tem interpretação causal (i.e. é derivada de uma formulação de resultados potenciais), o parâmetro θ_0 é o efeito causal que estamos interessados em estimar.

Uma abordagem **equivocada** para estimar o efeito causal de interesse utilizando-se técnicas de ML seria construir um sofisticado estimador $\hat{\theta}_0 T + \hat{g}_0(\mathbf{X})$[3] e imaginar que o parâmetro estimado $\hat{\theta}_0$ converge em probabilidade para o verdadeiro efeito causal. Suponha que repartimos a amostra em duas parcelas iguais, a principal com n observações e $i \in \mathcal{I}$ e uma partição auxiliar com $N - n$ observações e indivíduos $i \in \mathcal{I}^c$. Utiliza-se a partição auxiliar para estimar $\hat{g}_0$ e a partição principal para $\hat{\theta}_0$. Desta forma:

$$\hat{\theta}_0 = \left(\frac{1}{n} \sum_{i \in \mathcal{I}} T_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in \mathcal{I}} T_i (Y_i - \hat{g}_0(\mathbf{X}_i)). \tag{4.14}$$

Em um caso geral onde $\hat{g}_0$ foi estimado a partir de modelos de *machine learning* onde faz-se uso de regularização para controlar o *trade-off* viés-variância e portanto, minimizar a raiz do erro quadrático médio – RMSE – o estimador da Equação 4.14 não converge em probabilidade para o verdadeiro θ_0, $\sqrt{n}(\hat{\theta}_0 - \theta_0) \xrightarrow{P} 0$. Podemos decompor esta expressão como

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n} \sum_{i \in \mathcal{I}} T_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} T_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n} \sum_{i \in \mathcal{I}} T_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} T_i (g_0(\mathbf{X}_i) - \hat{g}_0(\mathbf{X}_i))}_{:=b}.$$

O termo denotado por a é bem comportado assintoticamente, uma vez que o erro U é independente do tratamento T e possui média igual a zero, portanto $a \xrightarrow{d} N(0, \Sigma)$ para alguma variância Σ. O segundo termo, b, é o **viés de regularização** que não converge para zero uma vez que $\hat{g}_0 \xrightarrow{P} g_0$ pois o estimador faz uso de regularização, que introduz viés.

[3] Por exemplo utilizar uma floresta aleatória para estimar a função não linear $\hat{g}_0$ e então inserir esta estimativa em uma regressão linear e estimar $\hat{\theta}_0$.

Chernozhukov et al. (2018) mostram como criar um novo estimador para o efeito causal no modelo das equações 4.12 e 4.13. O estimador proposto faz uso de uma (quasi) ortogonalização de T com relação ao vetor de covariadas $\mathbf{X}$ de forma que a técnica passa a fazer duas tarefas de predição com métodos de ML, e portanto o nome *Double Machine Learning* – DML. Ao fazer a ortogonalização e utilizar *cross-fitting* com várias partições da amostra, os autores demonstram que o estimador será assintoticamente não viesado.

Suponha que obtemos um estimador dos resíduos da equação 4.13, retirando os efeitos de $\mathbf{X}$ em T, $\hat{V} = T - \hat{f}_0(\mathbf{X})$ com a amostra auxiliar. De maneira análoga a anterior, continuamos a estimar um modelo para $\hat{g}_0$. O novo estimador será então dado por:

$$\check{\theta}_0 = \left(\frac{1}{n} \sum_{i \in \mathcal{I}} \hat{V}_i T_i \right)^{-1} \frac{1}{n} \sum_{i \in \mathcal{I}} \hat{V}_i (Y_i - \hat{g}_0(\mathbf{X}_i)). \tag{4.15}$$

O ponto central da estratégia é que ao fazer $\hat{V}$ (quasi) ortogonal a $\mathbf{X}$[4], este também será ortogonal a qualquer função de $\mathbf{X}$, especificamente g_0 e $\hat{g}_0$. Donde se extrai que o novo termo para o viés de regularização:

$$\check{b} := \left(\frac{1}{n} \sum_{i \in \mathcal{I}} \hat{V}_i T_i \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} \hat{V}_i (g_0(\mathbf{X}_i) - \hat{g}_0(\mathbf{X}_i)) \tag{4.16}$$

será aproximadamente zero em valor esperado.

4.5 Modelos Utilizados

Para fins de replicação do artigo original de Fellner, Sausgruber e Traxler (2013), utilizamos como um dos modelos de ML supervisionado a regressão linear. Além desta, e com o intuito de analisar heterogeneidade do tratamento, modelos do tipo *Double Machine Learning*, florestas causais e variáveis instrumentais duplamente robustas, (CHERNOZHUKOV et al., 2018; WAGER; ATHEY, 2018; ATHEY; TIBSHIRANI; WAGER et al., 2019; OPRESCU; SYRGKANIS; WU, 2019; SYRGKANIS et al., 2019) foram também especificados e estimados.

Florestas aleatórias ortogonais são uma combinação de florestas causais e *double machine learning*, daí o nome **ForestDML**, que permitem controlar um conjunto de alta dimen-

[4] O resíduo seria ortogonal se $\hat{f}_0$ fosse estimado através da projeção linear em $\mathbf{X}$, ou seja, através de uma regressão linear. Caso $\hat{f}_0$ seja não linear, esta projeção é apenas aproximadamente ortogonal.

sionalidade de variáveis de confusão $\mathbf{W}$, enquanto ao mesmo tempo estima de forma não paramétrica o efeito de tratamento heterogêneo $\theta(\mathbf{X})$, em um conjunto de dimensão inferior (e finita) de variáveis $\mathbf{X}$. Além disso, as estimativas são assintoticamente normais e, portanto, os intervalos de confiança baseados em bootstrap são válidos. O método assume as seguintes hipóteses estruturais sobre o processo de geração dos dados:

$$Y = \theta(\mathbf{X}) \cdot T + g(\mathbf{X}, \mathbf{W}) + \epsilon \qquad \mathbb{E}[\epsilon \mid \mathbf{X}, \mathbf{W}] = 0 \tag{4.17}$$

$$T = f(\mathbf{X}, \mathbf{W}) + \eta \qquad \mathbb{E}[\eta \mid \mathbf{X}, \mathbf{W}] = 0 \tag{4.18}$$

$$\mathbb{E}[\eta \cdot \epsilon \mid \mathbf{X}, \mathbf{W}] = 0 \tag{4.19}$$

nenhuma outra hipótese é feita sobre as funções θ, g e f podendo ser não-lineares. O modelo é parcialmente linear, ou seja, os efeitos são lineares **no tratamento**, mas qualquer outra característica X pode mediar este tratamento de forma não-linear. A identificação de θ se dá pela seguinte equação de momento que respeita a condição de ortogonalidade de Neyman[5]

$$\mathbb{E}[(Y - q(\mathbf{X}, \mathbf{W}) - \theta(\mathbf{X}) \cdot (T - f(\mathbf{X}, \mathbf{W}))) \cdot (T - f(\mathbf{X}, \mathbf{W}))] = 0 \tag{4.20}$$

onde $q(\mathbf{X}, \mathbf{W}) = \mathbb{E}[Y \mid \mathbf{X}, \mathbf{W}]$ e $f(\mathbf{X}, \mathbf{W}) = \mathbb{E}[T \mid \mathbf{X}, \mathbf{W}]$.

O análogo amostral para a função θ ótima é descrito por:

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} K_x(X_i) \cdot \left(Y_i - \hat{q}(X_i, W_i) - \theta \cdot \left(T_i - \hat{f}(X_i, W_i)\right)\right)^2 \tag{4.21}$$

onde o kernel $K_x(X_i)$ é uma métrica de similaridade calculada pela construção de uma floresta aleatória "honesta" com critério causal, (WAGER; ATHEY, 2018; ATHEY; TIBSHIRANI; WAGER et al., 2019). Este critério é ligeiramente modificado de modo a incorporar a residualização quando do cálculo da pontuação de cada candidato a divisão da amostra na criação da árvore[6].

Além disso, precisamos estimar em um primeiro estágio as funções perturbação $\hat{q}(\mathbf{X}, \mathbf{W})$ e $\hat{f}(\mathbf{X}, \mathbf{W})$, o que é feito através de técnicas tradicionais de ML. O método divide os dados e

[5]Em Chernozhukov et al. (2018) a ortogonalidade de Neyman para uma condição de momento é definida em termos da derivada de Gateaux, que deve ser zero em valor esperado no ponto dos verdadeiros parâmetros. $\partial_\xi E[\psi(W; \theta_0, \xi_0)[\xi - \xi_0]] = 0$

[6]Uma árvore causal possui critério de particionamento modificado em relação a árvores de decisão tipicamente encontradas nas técnicas de ML. O particionamento para descoberta de efeitos causais heterogêneos está descrito em (ATHEY; IMBENS, 2016)

realiza o *cross-fitting*: isto é, ajusta os modelos de expectativa condicional, $\hat{q}$ e $\hat{f}$ e o próprio kernel (uma floresta aleatória causal) na primeira metade e prevê os efeitos heterogêneos na segunda metade e vice-versa. Posteriormente estima $\theta(\mathbf{X})$ em todos os dados.

A equação 4.21 é portanto o nosso estimador do efeito causal condicional a uma realização x, $CATE(x) = \theta(x)$, através do método de **ForestDML**. O valor do ATE será dado pela média ponderada de $\theta(x)$ sobre toda a distribuição das covariadas $\mathbf{X}$, $ATE = \mathbb{E}_{\mathbf{X}}[\theta(x)]$ enquanto que o ATT é o valor esperado sobre a subpopulação tratada, $ATT = \mathbb{E}_{\mathbf{X}}[\theta(x) \mid T = 1]$.

Consideramos agora o problema onde o efetivo tratamento T é endógeno, porém, temos a disposição uma variável instrumental Z que nos auxilia a fazer a inferência. Neste *setup* desconsideramos a presença de variáveis puramente para fins de controle, assumindo que todas as variáveis que interferem no resultado Y também podem mediar o tratamento e causar heterogeneidade. Suponha que temos o seguinte modelo estrutural:

$$Y = \theta(\mathbf{X}) \cdot T + g(\mathbf{X}) + \epsilon, \qquad \mathbb{E}[\epsilon \mid \mathbf{X}, Z] = 0 \tag{4.22}$$

$$Z = m(\mathbf{X}) + \eta, \qquad \mathbb{E}[\eta \mid \mathbf{X}] = 0 \tag{4.23}$$

$$\mathbb{E}[\eta \cdot \epsilon \mid \mathbf{X}, Z] = 0 \tag{4.24}$$

$$\mathbb{E}[T \cdot \epsilon \mid \mathbf{X}] \neq 0 \tag{4.25}$$

Este modelo é semelhante ao anterior com exceção que agora o tratamento pode ser correlacionado com variáveis não observadas que interferem no resultado, entretanto, a variável instrumental é completamente exógena ao processo gerador de Y. A condição de momento satisfazendo a ortogonalidade de Neyman é dada por:

$$\mathbb{E}[(Y - q(\mathbf{X}) - \theta(\mathbf{X}) \cdot (T - f(\mathbf{X}))) \cdot (Z - m(\mathbf{X}))] = 0 \tag{4.26}$$

onde $q(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$, $f(\mathbf{X}) = \mathbb{E}[T \mid \mathbf{X}]$ e $m(\mathbf{X}) = \mathbb{E}[Z \mid \mathbf{X}]$. Considere as seguintes transformações, $\tilde{Y} = Y - q(\mathbf{X})$, $\tilde{T} = T - f(\mathbf{X})$ e $\tilde{Z} = Z - m(\mathbf{X})$. Chamamos de conformidade (ao tratamento) heterogênea o fato de $\beta(\mathbf{X}) = \mathbb{E}[\tilde{T}\tilde{Z} \mid \mathbf{X}]$ ser uma função de $\mathbf{X}$. Para dar conta dessa conformidade heterogênea, precisamos mudar nossa equação de momento para reponderá-la com base em $\beta(\mathbf{X})$, que é desconhecido e também precisa ser estimado a partir dos dados. Dado que essa função pode ser arbitrariamente complexa, queremos que nossa estimativa final seja robusta a erros na estimativa de $\beta(\mathbf{X})$. Podemos conseguir isso considerando uma abordagem duplamente robusta para estimar $\hat{\theta}(x)$. Suponha que tenhamos algum outro método

de calcular uma estimativa preliminar do efeito de tratamento heterogêneo $\theta_{\text{pre}}(x)$, então podemos combinar ambas as estimativas para obter um método mais robusto para o LATE e o estimador para $\theta(x)$ neste caso será dado por:

$$\hat{\theta}_{DR}(x) = \underset{\theta}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} \left(\theta_{\text{pre}}(x) + \frac{\left(\hat{\tilde{Y}}_i - \theta_{\text{pre}}(x)\hat{\tilde{T}}_i\right)\hat{\tilde{Z}}_i}{\hat{\beta}(X_i)} - \theta(X_i) \right)^2 \tag{4.27}$$

onde todas as funções denotadas com circunflexo são estimativas de suas respectivas funções populacionais.

A equação 4.27 fornece o estimador *Doubly Robust Instrumental Variable* – DRIV – que deve ser utilizado em situações onde o tratamento recebido é endógeno a variável de resultado, ou seja, existem fatores comuns e não observados que afetam ambos, tratamento e resultado. Cabe ressaltar que a **interpretação deste estimador é um LATE**, o efeito do tratamento somente sobre a subpopulação dos indivíduos obedientes.

5 Apresentação dos Resultados

Primeiramente apresentamos uma replicação dos resultados originais obtidos através de regressão linear **desconsiderando** o atrito no experimento. A regressão realizada foi:

$$\begin{aligned} y_i = \alpha + \theta_1 \text{Correio}_i + \theta_2 \text{Ameaça}_i + \theta_3 \text{Info}_i \\ + \theta_4 \text{Info\&Ameaça}_i + \theta_5 \text{Moral}_i + \theta_6 \text{Moral\&Ameaça}_i + \varepsilon_i \end{aligned} \tag{5.1}$$

A Tabela 7 abaixo apresenta os mesmos resultados obtidos pelos autores do artigo original, demonstrando que a metodologia aplicada foi compreendida.

Tabela 7 – Efeito do tratamento nos registros, atualizações contratuais, and resposta geral para o modelo de regressão linear.

Dep. Var.	Registro		Atual. Contratual		Resposta Geral	
Modelo	(1)	(2)	(3)	(4)	(5)	(6)
Variáveis						
Correio	0.065***	0.066***				
	(0.003)	(0.003)				
Ameaça	0.010***	0.009**	0.011**	0.006	0.027***	0.019**
	(0.002)	(0.004)	(0.004)	(0.008)	(0.005)	(0.009)
Moral	-0.004	-0.004	-0.024***	-0.033***	-0.032***	-0.043***
	(0.003)	(0.004)	(0.005)	(0.008)	(0.006)	(0.008)
Info	-0.002	-0.004	-0.018***	-0.017**	-0.023***	-0.024***
	(0.003)	(0.004)	(0.005)	(0.008)	(0.006)	(0.008)
Ameaça x Moral		0.0005		0.018		0.021*
		(0.006)		(0.011)		(0.012)
Ameaça x Info		0.004		-0.002		0.001
		(0.006)		(0.011)		(0.012)
Constante	0.008***	0.008***	0.289***	0.291***	0.427***	0.431***
	(0.002)	(0.002)	(0.004)	(0.005)	(0.005)	(0.006)
Estatísticas de diagnóstico						
Observações	50,498	50,498	41,007	41,007	41,007	41,007

Erro-padrão robusto a heterocedasticidade em parenteses
*Nível Significância: ***: 0.01, **: 0.05, *: 0.1*

A Tabela 7 apresenta os valores estimados do ATE para cada um dos tratamentos, Correio, Ameaça, Info, Info&Ameaça, Moral e Moral&Ameaça sobre os resultados de interesse, registro para o pagamento do tributo dentro de 50 dias após o envio da correspondência,

atualização contratual e resposta através de uma carta explicando a situação individual do contribuinte. Para os tratamentos T2-T6 cabe ressaltar que o valor estimado é o *acréscimo* com relação ao tratamento básico, T1-Correio, uma vez que em um modelo de regressão linear a interpretação dos coeficientes é parcial.

O fato de o contribuinte receber uma carta cobrança tem efeito positivo no registro sendo estatisticamente significativo ao nível de confiança de 1%. Caso a cobrança seja acompanhada do texto contendo a ameaça, o efeito é ainda maior, com incremento de 0,01 e novamente significativo a 1%. Os demais tratamentos e interações não se mostraram significativos.

Continuando a replicação do trabalho original, a Tabela 8 abaixo apresenta a Tabela do apêndice C onde os autores investigam efeitos heterogêneos para alguns estratos da população que representam vários subgrupos dividindo a amostra de acordo com a mediana de i) tamanho da população, ii) densidade populacional, iii) renda familiar média no município, e iv) parcela dos votos dados a partidos de direita no município.

Tabela 8 – Efeito heterogêneo do tratamento. Modelo de regressão linear.

Dep. Var.:	Registro							
	População		Densidade		Renda		Votantes Direita	
	≥mediana	<mediana	≥mediana	<mediana	≥mediana	<mediana	≥mediana	<mediana
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tratamento								
Ameaça	0.007**	0.017***	0.011***	0.014***	0.005	0.019***	0.018***	0.007*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)
Moral	-0.008*	-0.002	-0.010**	-0.0008	-0.005	-0.004	-0.002	-0.007*
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
Info	-0.002	-0.002	-0.005	0.0004	-0.004	-0.0002	-0.003	-0.002
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
Diagnóstico								
Observações	20,714	20,293	20,281	20,726	19,606	21,401	20,588	20,419

Erro-padrão robusto a heterocedasticidade em parenteses
*Nível Significância: ***: 0.01, **: 0.05, *: 0.1*

A ameaça feita nas cartas-cobrança manteve sua significância estatística em todos os estratos analisados, havendo de fato algum efeito heterogêneo com relação a este tratamento. Um resultado que chama a atenção foi o efeito **negativo** do apelo moral para municípios com população e densidade populacional abaixo da mediana e que votam mais a direita que a mediana dos municípios Austríacos. Entretanto, estes resultados devem ser interpretados considerando que os autores ignoraram o atrito no experimento e que este é maior justamente nos municípios de mais baixa população e densidade.

5.1 Modelos de *Machine Learning*

Partimos para os principais modelos deste estudo, aqueles que fazem uso de técnicas de *Machine Learning* para inferência causal. Para estudos aleatorizados os modelos de ML oferecem como vantagem a possibilidade de estudar heterogeneidade do efeito do tratamento sem especificar uma estrutura *a priori* para esta, como fizemos anteriormente ao replicarmos o resultado dos autores originais na Tabela 8. Desta forma, podemos deixar que os dados nos mostrem a relação das covariadas e suas interações com o efeito do tratamento. Em todos os resultados apresentados a seguir, estaremos estimando os efeitos causais dos tratamentos sobre a variável de *Registro* apenas. Não foram realizadas análises sobre as variáveis de atualização contratual ou de reposta geral.

Começamos com o modelo *Double Machine Learning* fazendo uso de florestas causais – ForestDML – *sem levar em conta o atrito do experimento*. Apresentamos o ATE e ATT estimados para cada um dos tratamentos, seguidos de interpretações da heterogeneidade com o uso de uma árvore de decisão rasa treinada a partir dos efeitos causais *individuais* estimados para o tratamento T2 - Ameaça como exemplo.

Em seguida, introduzimos modelos de variáveis instrumentais que possam inferir corretamente os efeitos causais na presença de atrito. O primeiro modelo estimado é uma regressão linear em dois estágios, porém esta especificação não apresenta efeitos heterogêneos. O principal modelo deste trabalho, que captura tanto a presença de atrito no experimento quanto a possibilidade de heterogeneidade nos efeitos, é o modelo *Doubly Robust Instrumental Variable* descrito em (SYRGKANIS et al., 2019). O DRIV é flexível o bastante para utilizar uma variável instrumental na estimação, eliminando o viés causado pelo atrito, assim como a heterogeneidade dos efeitos pode ser estruturada de forma não-paramétrica, através de uma floresta aleatória. A interpretação desta heterogeneidade, assim como no modelo ForestDML, pode ser feita através de uma árvore de decisão rasa.

A Tabela 9 apresenta estimativas de efeitos heterogêneos feitas a partir de um modelo ForestDML para quantis especificados das covariadas escolhidas, $\mathbf{X}$[1]. Pode-se perceber que, para cada um dos tratamentos, o efeito estimado varia consideravelmente entre quantis, caracterizando a heterogeneidade.

[1] A Tabela 12 apresenta estas covariadas escolhidas para identificação de efeitos heterogêneos assim como seus valores para os quantis especificados.

Tabela 9 – Efeitos heterogêneos do tratamento estimados por Double Machine Learning.

X	Correio	Ameaça	Info	Moral
Média	0.0680***	0.0810***	0.0620***	0.0690***
	(0.0110)	(0.0110)	(0.0090)	(0.0110)
Mínimo	0.0780***	0.0790***	0.1150***	0.0790***
	(0.0140)	(0.0210)	(0.0320)	(0.0190)
25%	0.0700***	0.0690***	0.0590***	0.0650***
	(0.0120)	(0.0100)	(0.0080)	(0.0140)
50%	0.0670***	0.0800***	0.0570***	0.0650***
	(0.0090)	(0.0130)	(0.0110)	(0.0110)
75%	0.0840***	0.1120***	0.0780***	0.0770***
	(0.0110)	(0.0210)	(0.0130)	(0.0130)
Máximo	0.1010***	0.1090***	0.0840***	0.0820***
	(0.0190)	(0.0240)	(0.0150)	(0.0180)

*Nível Significância: ***: 0.01, **: 0.05, *: 0.1*
Nota: Os estágios de previsão foram floresta aleatória para $E[Y|\mathbf{X}]$ e regressão logística para $E[T|\mathbf{X}]$. O modelo final para o efeito condicional do tratamento, $\theta(\mathbf{X})$, é uma floresta aleatória.

Este modelo ainda podem apresentar viés em suas estimativas, uma vez que foi identificado atrito não aleatório no experimento. Faremos uso, portanto, da técnica de variáveis instrumentais para estimar corretamente o efeito causal dos tratamentos sobre o subgrupo dos indivíduos obedientes. Uma vez que o não cumprimento a designação ao tratamento ocorreu apenas para indivíduos que deveriam ser tratados, conforme demonstrado na seção 4.2.1 o LATE estimado pode ser interpretado como ATT. O ATE não pode ser recuperado em tratamentos com descumprimento a designação.

A variável instrumental natural a ser escolhida em um experimento aleatório com cumprimento parcial ao tratamento é a própria designação ao tratamento. Pelo próprio desenho do experimento, esta variável cumpre com a restrição de exclusão (i.e. é independente de qualquer variável *confounder* e não interfere diretamente no resultado) e a condição de relevância (i.e. é altamente correlacionada com o efetivo recebimento de tratamento).

Escolhemos primeiramente realizar uma estimação de variáveis instrumentais pelo tradicional método de Mínimos Quadrados em 2 Estágios – 2SLS[2] – para fins de comparação com o modelo de ML. O método 2SLS segue uma especificação linear, sem outras variáveis mediadoras dos tratamentos e portanto, estima apenas efeitos médios para a população, sem

[2]Ou IV2SLS, usaremos de forma intercambiável.

heterogeneidade. O método 2SLS recebe este nome pois, pode ser entendido como um método de estimação realizado em duas etapas. Na primeira estima-se o valor do efetivo tratamento, T em razão das variáveis instrumentais, Z, conhecido como primeiro estágio. Na segunda etapa a forma reduzida é obtida através da regressão de Y contra os instrumentos Z, conforme equações 5.2 e 5.3.

$$T_i = \pi_0 + \pi_1 Z_i + u_i \tag{5.2}$$

$$Y_i = \gamma_0 + \gamma_1 Z_i + v_i \tag{5.3}$$

Por fim, o coeficiente de interesse, que relaciona T a Y pode ser obtido pela divisão γ_1/π_1. Os resultados desta regressão estão apresentados na Tabela 11 coluna (3).

Os valores descritos na Tabela 11 coluna (3) **não devem** ser interpretados como o ATE, mas sim como LATE, uma vez que estamos levando em conta o atrito do experimento. Conforme a seção 4.2.1, no caso específico deste experimento, onde o não cumprimento a designação de tratamento se dá apenas para aqueles que foram escolhidos para algum tratamento ativo (em outros termos, não existem *always takers*), o LATE será numericamente igual ao ATT e portanto, podemos interpretar os resultados deste trabalho obtidos através dos métodos com variáveis instrumentais como o efeito de tratamento naqueles indivíduos que efetivamente receberam suas correspondências.

Uma vez que desejamos avaliar possíveis efeitos heterogêneos dos tratamentos, substituímos o método IV2SLS pelo DRIV, descrito na seção 4.5. Com a utilização deste modelo é possível eliminar algum viés que ocorra em função do atrito e ainda obter efeitos heterogêneos do tratamento em função das covariadas $\mathbf{X}$.

A Tabela 10 apresenta a mesma análise de heterogeneidade feita anteriormente, agora para o modelo DRIV. A primeira observação a ser feita é uma redução geral nos efeitos estimados para todos os tratamentos, alguns inclusive perdendo significância. Outra importante observação é a variabilidade do efeito estimado com relação ao quantil. Tomamos como exemplo o tratamento *Correio* que possui efeito negativo (-0,1140) para um indivíduo hipotético que possui todas as componentes de $\mathbf{X}$ no valor mínimo da amostra, enquanto que outro indivíduo no quantil máximo teria uma reação positiva ao tratamento no valor de 0,444. Esta grande variabilidade não foi encontrada nos outros métodos, ForestDML (Tabela 9) ou na regressão linear da Tabela 8, apesar da metodologia diferente.

Tabela 10 – Efeitos heterogêneos do tratamento estimados por Doubly Robust IV.

X	Correio	Ameaça	Info	Moral
Média	0.0650**	0.0570***	0.0320***	0.0480*
	(0.0250)	(0.0140)	(0.0120)	(0.0250)
Mínimo	-0.1140*	0.1910	0.3500**	-0.2570***
	(0.1980)	(0.2040)	(0.2020)	(0.1480)
25%	0.0570*	0.0260	0.0280	0.0110
	(0.0430)	(0.0470)	(0.0290)	(0.0370)
50%	0.0420***	0.0700***	0.0360***	0.0410**
	(0.0100)	(0.0150)	(0.0150)	(0.0150)
75%	0.0420	0.1530***	0.0540	0.1370***
	(0.0600)	(0.0570)	(0.0550)	(0.0550)
Máximo	0.4440**	0.1660	0.0740	0.2680
	(0.2220)	(0.2810)	(0.1720)	(0.1890)

*Nível Significância: ***: 0.01, **: 0.05, *: 0.1*
Nota: os estágios de previsão foram gradient boosted tree (regressão) para $E[Y|\mathbf{X}]$ e gbm (classificação) $E[T|\mathbf{X}]$. O modelo final para o efeito condicional do tratamento, $\theta(\mathbf{X})$, também foi uma gbm-regressão, porém mais rasa, com apenas 3 níveis de profundidade.

Mesmo com toda esta variabilidade nos resultados, o modelo DRIV providenciou estimativas condizentes para o LATE, conforme Tabela 11 abaixo.

Tabela 11 – Efeitos médios dos tratamentos estimados pelos métodos de ForestDML, IV2SLS e DRIV.

	ForestDML		IV2SLS	DRIV
	ATE	ATT	LATE	LATE
	(1)	(2)	(3)	(4)
Correio	0,0766	0,0766	0,0767	0,0588
Ameaça	0,0850	0,0848	0,0872	0,0650
Info	0,0762	0,0760	0,0728	0,0547
Moral	0,0695	0,0695	0,0724	0,0513

Esta tabela condensa todos os efeitos médios de tratamento estimados pelos modelos ForestDML, IV2SLS e DRIV, onde apenas o primeiro, por não levar em consideração o atrito, pode estimar ambos ATE e ATT. Os resultados são semelhantes entre os métodos, demonstrando um grau de robustez para as especificações que envolvem técnicas de *Machine Learning* **corrigidas para estimação de efeitos causais**.

De fato, todas as especificações apontam que, em média, o tratamento com a ameaça é

o mais efetivo, sendo que não há diferença estatisticamente significativa entre a carta-cobrança padrão e aquelas acrescidas de maiores informações ou apelo moral.

Finalmente, a heterogeneidade dos efeitos estimados pode ser melhor compreendida através de uma árvore de decisão criada a partir da estimação de $\theta(\mathbf{X})$, nosso efeito heterogêneo. Dado um vetor X_i com as covariadas específicas do indivíduo i, esta árvore a partir de regras estipuladas como base no valor das componentes de X_i pode "separar" este indivíduo para um determinado subgrupo e então estimar o efeito do tratamento aplicado. Indivíduos que recaiam em diferentes subgrupos apresentarão efeitos diferentes.

As figuras 2 e 3 apresentam árvores de três níveis de profundidade estimadas para o efeito do tratamento *Ameaça* pelos modelos ForestDML e DRIV respectivamente.

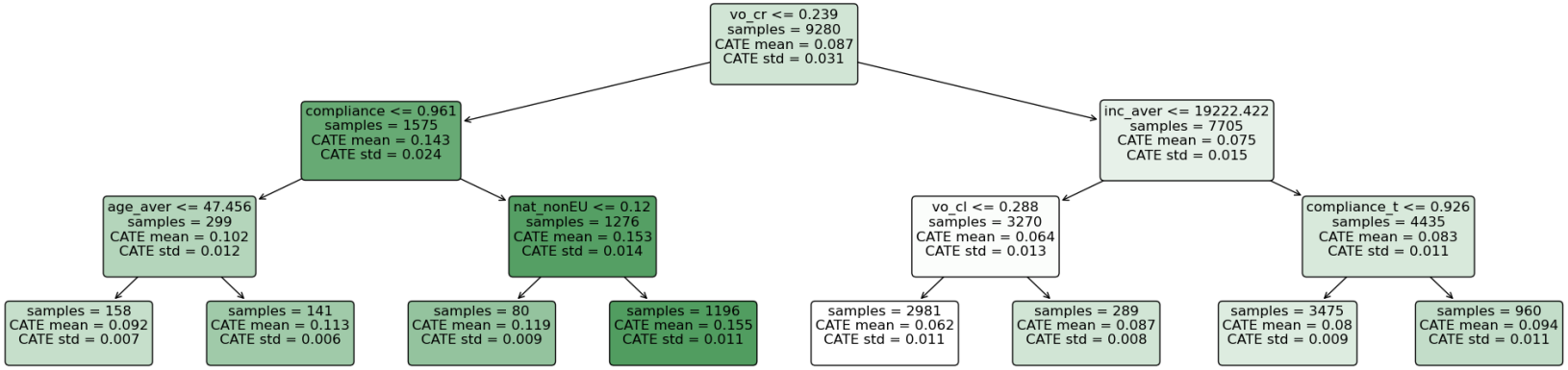

Figura 2 – Interpretação do efeito heterogêneo do tratamento através de árvore de decisão. Modelo ForestDML desconsiderando atrito.

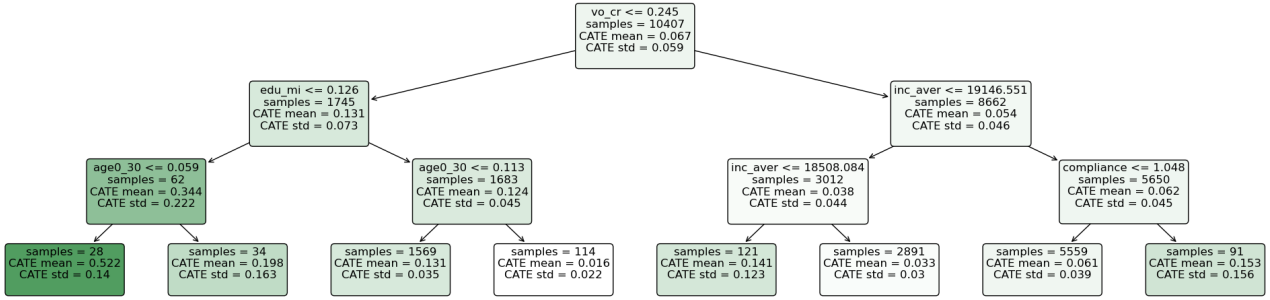

Figura 3 – Efeito heterogêneo no modelo DRIV, que leva em consideração o atrito.

Tomamos a figura 3 para exemplificar a interpretação dos resultados. A variável *vo_cr* (fração de votantes na centro-direita no município em 2006) foi escolhida a mais relevante para determinar **heterogeneidade** do tratamento e seu valor de corte foi 0,245. Se o município do indivíduo possui uma fração menor que o valor de corte, este indivíduo é levado para os subgrupos a esquerda, caso contrário a direita. Assim é feito sucessivamente, comparando o valor da variável do indivíduo contra o valor de corte indicado. Resultados verdadeiros correm a esquerda e falsos a direita.

Interpretando esta árvore verificamos que os indivíduos com a maior resposta ao tratamento (i.e. maior CATE mean) são aqueles que residem em municípios com uma fração menor de jovens, com idade entre 0 e 30 anos, fração de pessoas com ensino médio é menor que 0,126 e a fração de votantes na centro-direita é inferior a 0,245. Já os indivíduos com a menor resposta ao tratamento também residem em municípios com poucos votantes na centro-direita porém com uma fração da população com ensino médio mais alta que 0,126 e fração daqueles com idade inferior a 30 anos mais alta que 0.113.

A estrutura da árvore também fornece informação sobre as variáveis mais relevantes para a heterogeneidade dos efeitos, de fato, os nós de hierarquia superior (i.e. *vo_cr*) foram assim escolhidos justamente porque são aquelas variáveis que trazem uma maior diversidade na estimação do efeito causal. Desta forma, verificando as árvores das figuras 2 e 3 qualitativamente podemos dizer que alinhamento político, educação e renda (todas ao nível municipal) foram as variáveis que mais contribuíram para os efeitos heterogêneos verificados neste experimento. Contrastando com a Tabela 8, que traz as seguintes variáveis: população, densidade populacional, renda e alinhamento político, verificamos que as duas últimas estão presentes também nas árvores de decisão. E mais, estas variáveis presentes em ambos os modelos, são justamente aquelas que apresentam maior heterogeneidade de acordo com as estimativas de Fellner, Sausgruber e Traxler (2013) presentes na Tabela 8.

Destacamos o fato que no modelo dos autores originais, estes **supõem** variáveis passíveis de causar heterogeneidade para então dividir a amostra com base na mediana destas variáveis e estimar os efeitos para cada subamostra, ratificando ou não a hipótese de heterogeneidade. Já nos modelos de *Machine Learning* para efeitos causais aqui apresentados as variáveis relevantes para **heterogeneidade derivam diretamente** dos dados, os algoritmos são especializados em detectar os padrões de efeitos heterogêneos na amostra, sem a necessidade de uma hipótese *a priori*.

De posse de uma árvore de decisão como esta para cada tratamento a autoridade tributária pode escolher adequadamente o tipo de carta-cobrança a ser enviada a um contribuinte de modo que esta tenha o maior efeito estimado na recuperação do crédito tributário.

Referências

ALLINGHAM, Michael G; SANDMO, Agnar. Income tax evasion: A theoretical analysis. **Journal of public economics**, North-Holland, v. 1, n. 3-4, p. 323–338, 1972.

ANGRIST, Joshua D; PISCHKE, Jörn-Steffen. **Mostly harmless econometrics: An empiricist's companion**. [S.l.]: Princeton university press, 2008.

ATHEY, Susan; IMBENS, Guido. Recursive partitioning for heterogeneous causal effects. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 113, n. 27, p. 7353–7360, 2016.

ATHEY, Susan; TIBSHIRANI, Julie; WAGER, Stefan et al. Generalized random forests. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 47, n. 2, p. 1148–1178, 2019.

BLOOM, Howard S. Estimating the effect of job-training programs, using longitudinal data: Ashenfelter's findings reconsidered. **Journal of Human Resources**, JSTOR, p. 544–556, 1984.

CHANG, Neng-Chieh. Double/debiased machine learning for difference-in-differences models. **The Econometrics Journal**, Oxford University Press, v. 23, n. 2, p. 177–191, 2020.

CHERNOZHUKOV, Victor et al. Double/debiased machine learning for treatment and structural parameters. **The Econometrics Journal**, v. 21, n. 1, p. c1–c68, jan. 2018. DOI: 10.1111/ectj.12097.

DUFLO, Esther; GLENNERSTER, Rachel; KREMER, Michael. Using randomization in development economics research: A toolkit. **Handbook of development economics**, Elsevier, v. 4, p. 3895–3962, 2007.

FELLNER, Gerlinde; SAUSGRUBER, Rupert; TRAXLER, Christian. Testing enforcement strategies in the field: Threat, moral appeal and social information. **Journal of the European Economic Association**, Oxford University Press, v. 11, n. 3, p. 634–660, 2013.

IMBENS, Guido W; RUBIN, Donald B. **Causal inference in statistics, social, and biomedical sciences**. [S.l.]: Cambridge University Press, 2015.

MULLAINATHAN, Sendhil; SPIESS, Jann. Machine learning: an applied econometric approach. **Journal of Economic Perspectives**, v. 31, n. 2, p. 87–106, 2017.

OPRESCU, Miruna; SYRGKANIS, Vasilis; WU, Zhiwei Steven. Orthogonal Random Forest for Causal Inference. In: _____. **Proceedings of the 36th International Conference on Machine Learning**. [S.l.]: PMLR, 2019. v. 97. (Proceedings of Machine Learning Research), p. 4932–4941. Disponível em: ¡http://proceedings.mlr.press/v97/oprescu19a.html¿.

ROBINSON, Peter M. Root-N-consistent semiparametric regression. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 931–954, 1988.

RUBIN, Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of educational Psychology**, American Psychological Association, v. 66, n. 5, p. 688, 1974.

SANDMO, Agnar. The theory of tax evasion: A retrospective view. **National tax journal**, JSTOR, p. 643–663, 2005.

SINGH, Rahul; SUN, Liyang. **De-biased Machine Learning in Instrumental Variable Models for Treatment Effects**. [S.l.: s.n.], 2020. arXiv: 1909.05244 [stat.ML].

SYRGKANIS, Vasilis et al. **Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments**. [S.l.: s.n.], 2019. arXiv: 1905.10176 [econ.EM].

WAGER, Stefan; ATHEY, Susan. Estimation and inference of heterogeneous treatment effects using random forests. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 523, p. 1228–1242, 2018.

A APÊNDICE

Tabela 12 – Variáveis utilizadas para identificar heterogeneidade nos efeitos.

	Média	Mínimo	25%	50%	75%	Máximo
gender	0,634896	0,000000	0,000000	1,000000	1,000000	1,000000
pop_density2005	8,495767	0,011677	0,362956	1,260130	5,922626	256,298584
compliance	0,935165	0,582418	0,904187	0,945887	0,967879	1,435407
compliance_t	0,886377	0,540201	0,845865	0,896116	0,934060	1,396313
vo_r	0,185270	0,040000	0,154268	0,172742	0,204835	0,459963
vo_cr	0,425843	0,115640	0,303905	0,411067	0,545568	0,920000
vo_cl	0,270831	0,020000	0,164189	0,266549	0,357143	0,784148
vo_l	0,118056	0,000000	0,081181	0,110133	0,148869	0,372320
inc_aver	20867,302734	17497,126953	18715,361328	20378,222656	22280,625000	36105,191406
edu_aver	11,169192	9,720000	10,839196	11,109369	11,431844	13,226902
edu_lo	0,708511	0,266667	0,662312	0,708918	0,785747	0,978261
edu_mi	0,150024	0,000000	0,126961	0,143243	0,174451	0,511111
edu_hi	0,141465	0,000000	0,079417	0,133014	0,194444	0,465745
age_aver	48,035931	42,813904	47,124645	48,228565	48,641762	54,185158
age0_30	0,073386	0,008032	0,055845	0,071653	0,090299	0,183333
age30_60	0,684685	0,525842	0,649355	0,682824	0,713466	0,829338
nat_A	0,906111	0,496835	0,874880	0,906192	0,945038	1,000000
nat_EU	0,002884	0,000000	0,001120	0,002317	0,003959	0,030108
nat_nonEU	0,091006	0,000000	0,051177	0,091558	0,122511	0,496835

A.1 Links

Este trabalho foi realizado utilizando o GitHub como ferramenta de repositório de código e o Overleaf para escrever a monografia utilizando LaTeX. O vídeo de apresentação do trabalho está hospedado no YouTube.

Link para o vídeo: https://youtu.be/qs8nG0l4Vk4

Link para o repositório: https://github.com/rfbressan/curso_big_data

Link para a monografia: https://www.overleaf.com/read/njxdsjrtxqnm

A.2 Script Python

```
# Metodos de Machine Learning para inferencia causal
# com dados de Fellner et all (2013)
# Autor: Rafael Felipe Bressan
# Arquivo: script_py.py
# Script Python para o trabalho monografico: Inferência Causal com
# Machine Learning uma aplicacao para evasao fiscal
# Pós-Graduacao Lato Sensu em Ciencia de Dados e Big Data PUC-MG
```

```
# Ano: 2021

# importa bibliotecas necessarias
import gc
from copy import deepcopy
import numpy as np
import pandas as pd

# Regressao linear e IV
import statsmodels.api as sm
from linearmodels import IV2SLS

# Modelos de ML
import lightgbm as lgb
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LogisticRegression, Lasso
from sklearn.ensemble import RandomForestRegressor

# EconML
from econml.ortho_iv import DMLATEIV, IntentToTreatDRIV
from econml.cate_interpreter import SingleTreeCateInterpreter
from econml.dml import ForestDML, DML, SparseLinearDML
from econml.causal_forest import CausalForest

# Graficos
import matplotlib.pyplot as plt

# Funcoes auxiliares
def stars(x, levels=[0.1, 0.05, 0.01]):
    assert (len(levels)==3), "Comprimento de levels deve ser 3."

    if x>levels[0]:
        return ''
    elif x>levels[1]:
        return '*'
    elif x>levels[2]:
        return '**'
    else:
        return '***'

def format_float(x, digits):
    return '{:.{dig}f}'.format(x, dig=digits)

def surr_parenthesis(x, digits):
    return '('+'{:.{dig}f}'.format(x, dig=digits)+')'


# Carregando os dados
data = pd.read_stata("data_final.dta").sort_values("treatment")
X_cols = ["gender", "pop_density2005",
    "compliance", "compliance_t", "vo_r", "vo_cr", "vo_cl", "vo_l",
    "inc_aver", "edu_aver", "edu_lo", "edu_mi", "edu_hi",
    "age_aver", "age0_30", "age30_60", "nat_A", "nat_EU", "nat_nonEU"]

#############################################################
# Ignorando o atrito e estimando os efeitos apenas para
# delivered == 1
#############################################################
```

```
deliv = data[data["delivered"]!=0].fillna(0)
# Para avaliacao de efeitos heterogeneos
X_eval = (deliv[X_cols]
    .describe()
    .loc[["mean", "min", "25%", "50%", "75%", "max"]]
)
# Tratamentos
t1_deliv = deliv[deliv["treatment"].isin([0,1])]
t2_deliv = deliv[deliv["treatment"].isin([0,2])]
t3_deliv = deliv[deliv["treatment"].isin([0,3])]
t5_deliv = deliv[deliv["treatment"].isin([0,5])]

# Variaveis de interesse
Y1 = t1_deliv["resp_A"].values
T1 = t1_deliv["delivered"].values
X1 = t1_deliv[X_cols].values
X1_treat=X1[T1 == 1]

Y2 = t2_deliv["resp_A"].values
T2 = t2_deliv["delivered"].values
X2 = t2_deliv[X_cols].values
X2_treat=X2[T2 == 1]

Y3 = t3_deliv["resp_A"].values
T3 = t3_deliv["delivered"].values
X3 = t3_deliv[X_cols].values
X3_treat=X3[T3 == 1]

Y5 = t5_deliv["resp_A"].values
T5 = t5_deliv["delivered"].values
X5 = t5_deliv[X_cols].values
X5_treat=X5[T5 == 1]

# ForestDML() = CausalForest()??
# ForestDML eh muito mais rapido que CausalForest com resultados
# semelhantes para a interpretacao via arvore

# DML com regressao logistica para E[T|X,W] e Floresta Aleatoria para
# E[Y|X,W]. Modelo final para Theta(X) eh escolhido por floresta
dml=ForestDML(
    model_t=LogisticRegression(),
    model_y=RandomForestRegressor(),
    discrete_treatment=True,
    n_estimators=1000,
    subsample_fr=0.7,
    min_samples_leaf=20,
    n_crossfit_splits=3,
    n_jobs=-1
)
# DML para T1
dml.fit(Y1, T1, X1, inference='auto')
dml1_eff=dml.effect(X1, T0=0, T1=1)
dml1_eff_treat=dml.effect(X1_treat, T0=0, T1=1)
print(f"ATE T1 por DML: {np.mean(dml1_eff)}\nATT T1 por DML: {np.mean(dml1_eff_treat)}
dml1_inf=dml.effect_inference(X_eval.values)
dml1_summary=dml1_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr"]
dml1_summary.index=X_eval.index
dml1_summary["star"]=dml1_summary["pvalue"].apply(stars)
```

```
dml1_summary["point_estimate"]=dml1_summary["point_estimate"].apply(format_float, digi
dml1_summary["point_estimate"]=dml1_summary["point_estimate"].str.cat(dml1_summary["st
dml1_summary["stderr"]=dml1_summary["stderr"].apply(surr_parenthesis, digits=4)
dml1_summary=dml1_summary[["point_estimate", "stderr"]].stack()
dml1_summary.name="Correio"

# DML para T2
dml.fit(Y2, T2, X2, inference='auto')
dml2_eff=dml.effect(X2, T0=0, T1=1)
dml2_eff_treat=dml.effect(X2_treat, T0=0, T1=1)
print(f"ATE T2 por DML: {np.mean(dml2_eff)}\nATT T2 por DML: {np.mean(dml2_eff_treat)}
dml2_inf=dml.effect_inference(X_eval.values)
dml2_summary=dml2_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr"]
dml2_summary.index=X_eval.index
dml2_summary["star"]=dml2_summary["pvalue"].apply(stars)
dml2_summary["point_estimate"]=dml2_summary["point_estimate"].apply(format_float, digi
dml2_summary["point_estimate"]=dml2_summary["point_estimate"].str.cat(dml2_summary["st
dml2_summary["stderr"]=dml2_summary["stderr"].apply(surr_parenthesis, digits=4)
dml2_summary=dml2_summary[["point_estimate", "stderr"]].stack()
dml2_summary.name="Ameaça"

# Interpretacao por arvore de decisao para T2
interp = SingleTreeCateInterpreter(
    include_model_uncertainty=False,
    max_depth=3,
    min_samples_leaf=10)
interp.interpret(dml, X2)
fig, ax1 = plt.subplots(figsize=(25,6))
interp.plot(feature_names=X_cols, fontsize=12, ax=ax1)
fig.savefig("Figs/fig_tree_dml.png")

# DML para T3
dml.fit(Y3, T3, X3, inference='auto')
dml3_eff=dml.effect(X3, T0=0, T1=1)
dml3_eff_treat=dml.effect(X3_treat, T0=0, T1=1)
print(f"ATE T3 por DML: {np.mean(dml3_eff)}\nATT T3 por DML: {np.mean(dml3_eff_treat)}
dml3_inf=dml.effect_inference(X_eval.values)
dml3_summary=dml3_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr"]
dml3_summary.index=X_eval.index
dml3_summary["star"]=dml3_summary["pvalue"].apply(stars)
dml3_summary["point_estimate"]=dml3_summary["point_estimate"].apply(format_float, digi
dml3_summary["point_estimate"]=dml3_summary["point_estimate"].str.cat(dml3_summary["st
dml3_summary["stderr"]=dml3_summary["stderr"].apply(surr_parenthesis, digits=4)
dml3_summary=dml3_summary[["point_estimate", "stderr"]].stack()
dml3_summary.name="Info"

# DML para T5
dml.fit(Y5, T5, X5, inference='auto')
dml5_eff=dml.effect(X5, T0=0, T1=1)
dml5_eff_treat=dml.effect(X5_treat, T0=0, T1=1)
print(f"ATE T5 por DML: {np.mean(dml5_eff)}\nATT T5 por DML: {np.mean(dml5_eff_treat)}
dml5_inf=dml.effect_inference(X_eval.values)
dml5_summary=dml5_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr"]
dml5_summary.index=X_eval.index
dml5_summary["star"]=dml5_summary["pvalue"].apply(stars)
dml5_summary["point_estimate"]=dml5_summary["point_estimate"].apply(format_float, digi
dml5_summary["point_estimate"]=dml5_summary["point_estimate"].str.cat(dml5_summary["st
dml5_summary["stderr"]=dml5_summary["stderr"].apply(surr_parenthesis, digits=4)
```

```
dml5_summary=dml5_summary[["point_estimate", "stderr"]].stack()
dml5_summary.name="Moral"

# Melhora consumo de memoria
del dml
collected=gc.collect()
# ATE
dml_ate=[np.mean(x) for x in [dml1_eff, dml2_eff, dml3_eff, dml5_eff]]
# ATT
dml_att=[np.mean(x) for x in
    [dml1_eff_treat, dml2_eff_treat, dml3_eff_treat, dml5_eff_treat]]

treatments_list=["Correio", "Ameaça", "Info", "Moral"]
dml_effect=pd.DataFrame({"ATE": dml_ate, "ATT": dml_att}, index=treatments_list)

# Quantis das variaveis utilizadas para aferir heterogeneidade
X_eval.transpose().to_latex(
    buf="Tables/tab_het_vars.tex",
    decimal=",",
    caption="Variáveis utilizadas para identificar heterogeneidade nos efeitos.",
    label="tab:het-vars"
)

# Sumario com os resultados para os 4 tratamentos
dml_summary=(
    pd.concat(
        [dml1_summary, dml2_summary, dml3_summary, dml5_summary],
        axis=1)
    .reset_index()
    .rename(columns={"level_0": "X"})
    .drop(columns="level_1")
)

dml_summary.to_latex(
    buf="Tables/tab_dml_summary.tex",
    decimal=",",
    caption="Efeitos heterogêneos do tratamento estimados por Double Machine Learning
    label="tab:dml-summary",
    index=False
)
# Nota: os estágios de previsão foram floresta aleatória para $E[Y|\bfx]$
# e regressão logística para $E[T|\bfx]$. O modelo final para o efeito
# condicional do tratamento, $\theta(\bfx)$, é uma floresta aleatória.

#############################################################
# Metodos com variaveis Instrumentais
#############################################################

#############################################################
# Considerando o atrito e estimando os efeitos
#############################################################
iv=data.copy()
iv["delivered"] = iv["delivered"].fillna(0)
# Para avaliacao de heterogeneidade
X_eval = (iv[X_cols]
    .describe()
    .loc[["mean", "min", "25%", "50%", "75%", "max"]]
)
```

```
# Tratamentos
t1_iv = iv[iv["treatment"].isin([0,1])]
t2_iv = iv[iv["treatment"].isin([0,2])]
t3_iv = iv[iv["treatment"].isin([0,3])]
t5_iv = iv[iv["treatment"].isin([0,5])]

# Definindo as variaveis
Z1 = t1_iv["treatment"]
T1 = t1_iv["delivered"]
Y1 = t1_iv["resp_A"]
X1 = t1_iv[X_cols]

Z2 = t2_iv["treatment"]
T2 = t2_iv["delivered"]
Y2 = t2_iv["resp_A"]
X2 = t2_iv[X_cols]

Z3 = t3_iv["treatment"]
T3 = t3_iv["delivered"]
Y3 = t3_iv["resp_A"]
X3 = t3_iv[X_cols]

Z5 = t5_iv["treatment"]
T5 = t5_iv["delivered"]
Y5 = t5_iv["resp_A"]
X5 = t5_iv[X_cols]

# Modelos para E[Y|X] e E[T|Z,X]
lgb_YX_par = {
    "metric": "rmse",
    "learning_rate": 0.1,
    "num_leaves": 30,
    "max_depth": 5
}

lgb_TXZ_par = {
    "objective": "binary",
    "metric": "auc",
    "learning_rate": 0.1,
    "num_leaves": 30,
    "max_depth": 5
}

lgb_theta_par = {
    "metric": "rmse",
    "learning_rate": 0.1,
    "num_leaves": 30,
    "max_depth": 3
}

modelTXZ = lgb.LGBMClassifier(**lgb_TXZ_par)
modelYX = lgb.LGBMRegressor(**lgb_YX_par)
modelZX = lgb.LGBMClassifier(**lgb_TXZ_par)
# Modelo inicial para o efeito heterogeneo
# Sera melhorado pelo algoritmo Double Robust IV
pre_theta = lgb.LGBMRegressor(**lgb_theta_par)

## Modelo 2-stages Least Squares
```

```
# Variaveis com amostra completa
Z = iv[["mailing", "threat", "info", "appeal", "i_tinf", "i_tapp"]]
T = Z.multiply(iv["delivered"], axis=0)
Y = iv["resp_A"]
model_2sls = IV2SLS(Y, exog=np.ones(len(Y)), endog=T, instruments=Z)
iv2sls_fit = model_2sls.fit(debiased=True)
iv2sls_params=iv2sls_fit.params[["mailing", "threat", "info", "appeal"]]
# Soma os valores para ter mesma interpretacao que DML e DRIV
iv2sls_params["threat"]=iv2sls_params["threat"]+iv2sls_params["mailing"]
iv2sls_params["info"]=iv2sls_params["info"]+iv2sls_params["mailing"]
iv2sls_params["appeal"]=iv2sls_params["appeal"]+iv2sls_params["mailing"]
treatments_list=["Correio", "Ameaça", "Info", "Moral"]
iv2sls_effect=pd.DataFrame({"LATE": iv2sls_params.values}, index=treatments_list)

## Modelo DRIV

# Treina o modelo DRIV mais flexivel. Theta(X) pode ser um modelo
# flexivel (nao parametrico, ie. floresta aleatoria) de X
# ATENCAO: leva bastante tempo para rodar
driv1 = IntentToTreatDRIV(
    model_Y_X=modelYX,
    model_T_XZ=modelTXZ,
    flexible_model_effect=pre_theta,
    n_splits=3,
    featurizer=None #PolynomialFeatures(degree=1, include_bias=False)
)
# Mesmo modelo para cada tratamento
driv2=deepcopy(driv1)
driv3=deepcopy(driv1)
driv5=deepcopy(driv1)

# DRIV para T1
print("Iniciando fit de DRIV T1\n")
driv1.fit(Y1, T1, Z=Z1, X=X1, inference="bootstrap")
print("Fim do fit de DRIV T1\n")
driv1_eff=driv1.effect(X1, T0=0, T1=1)
print(f"LATE T1 por DRIV: {np.mean(driv1_eff)}")
print("Iniciando inferencia de DRIV T1\n")
driv1_inf=driv1.effect_inference(X_eval)
print("Fim da inferencia de DRIV T1\n")
driv1_summary=driv1_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr
driv1_summary.index=X_eval.index
driv1_summary["star"]=driv1_summary["pvalue"].apply(stars)
driv1_summary["point_estimate"]=driv1_summary["point_estimate"].apply(format_float, di
driv1_summary["point_estimate"]=driv1_summary["point_estimate"].str.cat(driv1_summary
driv1_summary["stderr"]=driv1_summary["stderr"].apply(surr_parenthesis, digits=4)
driv1_summary=driv1_summary[["point_estimate", "stderr"]].stack()
driv1_summary.name="Correio"
# Melhora o consumo de memoria
del driv1
collected=gc.collect()

# DRIV para T2
print("Iniciando fit de DRIV T2\n")
driv2.fit(Y2, T2, Z=Z2, X=X2, inference="bootstrap")
print("Fim do fit de DRIV T2\n")
driv2_eff=driv2.effect(X2, T0=0, T1=1)
```

```
print(f"LATE T2 por DRIV: {np.mean(driv2_eff)}")
print("Iniciando inferencia de DRIV T2\n")
driv2_inf=driv2.effect_inference(X_eval)
print("Fim da inferencia de DRIV T2\n")
driv2_summary=driv2_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr
driv2_summary.index=X_eval.index
driv2_summary["star"]=driv2_summary["pvalue"].apply(stars)
driv2_summary["point_estimate"]=driv2_summary["point_estimate"].apply(format_float, di
driv2_summary["point_estimate"]=driv2_summary["point_estimate"].str.cat(driv2_summary
driv2_summary["stderr"]=driv2_summary["stderr"].apply(surr_parenthesis, digits=4)
driv2_summary=driv2_summary[["point_estimate", "stderr"]].stack()
driv2_summary.name="Ameaça"
# Interpretacao causal por arvore de decisao
interp = SingleTreeCateInterpreter(
    include_model_uncertainty=False,
    max_depth=3,
    min_samples_leaf=10
)
interp.interpret(driv2, X2)
fig, ax1 = plt.subplots(figsize=(25,6))
interp.plot(feature_names=X2.columns, fontsize=12, ax=ax1)
fig.savefig("Figs/fig_tree_driv.png")
# Melhora o consumo de memoria
del driv2
collected=gc.collect()

# DRIV para T3
driv3.fit(Y3, T3, Z=Z3, X=X3, inference="bootstrap")
driv3_eff=driv3.effect(X3, T0=0, T1=1)
print(f"LATE T3 por DRIV: {np.mean(driv3_eff)}")
driv3_inf=driv3.effect_inference(X_eval)
print("Fim da inferencia de DRIV T3\n")
driv3_summary=driv3_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr
driv3_summary.index=X_eval.index
driv3_summary["star"]=driv3_summary["pvalue"].apply(stars)
driv3_summary["point_estimate"]=driv3_summary["point_estimate"].apply(format_float, di
driv3_summary["point_estimate"]=driv3_summary["point_estimate"].str.cat(driv3_summary
driv3_summary["stderr"]=driv3_summary["stderr"].apply(surr_parenthesis, digits=4)
driv3_summary=driv3_summary[["point_estimate", "stderr"]].stack()
driv3_summary.name="Info"
# Melhora o consumo de memoria
del driv3
collected=gc.collect()

# DRIV para T5
driv5.fit(Y5, T5, Z=Z5, X=X5, inference="bootstrap")
driv5_eff=driv5.effect(X5, T0=0, T1=1)
print(f"LATE T5 por DRIV: {np.mean(driv5_eff)}")
driv5_inf=driv5.effect_inference(X_eval)
print("Fim da inferencia de DRIV T5\n")
driv5_summary=driv5_inf.summary_frame(alpha=0.05)[["point_estimate", "pvalue", "stderr
driv5_summary.index=X_eval.index
driv5_summary["star"]=driv5_summary["pvalue"].apply(stars)
driv5_summary["point_estimate"]=driv5_summary["point_estimate"].apply(format_float, di
driv5_summary["point_estimate"]=driv5_summary["point_estimate"].str.cat(driv5_summary
driv5_summary["stderr"]=driv5_summary["stderr"].apply(surr_parenthesis, digits=4)
driv5_summary=driv5_summary[["point_estimate", "stderr"]].stack()
```

```
driv5_summary.name="Moral"
# Melhora o consumo de memoria
del driv5
collected=gc.collect()

# LATE
driv_late=[np.mean(x) for x in
    [driv1_eff, driv2_eff, driv3_eff, driv5_eff]]
treatments_list=["Correio", "Ameaça", "Info", "Moral"]
driv_effect=pd.DataFrame({"LATE": driv_late}, index=treatments_list)
# driv_effect.to_latex(
#     buf="Tables/tab_driv_late.tex",
#     decimal=",",
#     caption="LATE estimado por Doubly Robust IV para diferentes tratamentos.",
#     label="tab:driv-late"
# )

# Junta todos os efeitos em uma tabela unica
all_effects=(dml_effect
    .merge(iv2sls_effect, left_index=True, right_index=True)
    .merge(driv_effect, left_index=True, right_index=True)
)
mindex=pd.MultiIndex.from_tuples([
    ("ForestDML", "ATE"),
    ("ForestDML", "ATT"),
    ("IV2SLS", "LATE"),
    ("DRIV", "LATE")],
    names=["Modelo", "Efeito"])
all_effects.columns=mindex
all_effects.to_latex(
    buf="Tables/tab_all_effects.tex",
    decimal=",",
    caption="Efeitos médios dos tratamentos estimados pelos métodos de ForestDML, IV2S
    label="tab:all-effects",
    float_format="%.4f",
    multirow=True,
    multicolumn=True,
    multicolumn_format="c"
)

# Sumario com os resultados para os 4 tratamentos
driv_summary=(
    pd.concat(
        [driv1_summary, driv2_summary, driv3_summary, driv5_summary],
        axis=1)
    .reset_index()
    .rename(columns={"level_0": "X"})
    .drop(columns="level_1")
)

driv_summary.to_latex(
    buf="Tables/tab_driv_summary.tex",
    decimal=",",
    caption="Efeitos heterogêneos do tratamento estimados por Doubly Robust IV.",
    label="tab:driv-summary",
    index=False
)
# Nota: os estágios de previsão foram gradient boosted tree (regressão) para $E[Y|\bfx
```

```
# e gbm (classificação) $E[T|\bfx]$. O modelo final para o efeito
# condicional do tratamento, $\theta(\bfx)$, também foi uma gbm-regressão, porém mais
# com apenas 3 níveis de profundidade.
```

A.3 Script R

```
#' Analise e replicacao de alguns resultados em Fellner et all (2013)
#' Autor: Rafael Felipe Bressan
#' Arquivo: script_R.R
#' Script R para o trabalho monografico: Inferência Causal com Machine Learning
#' uma aplicacao para evasao fiscal
#' Pós-Graduacao Lato Sensu em Ciencia de Dados e Big Data PUC-MG
#' Ano: 2021
#'
#' Carrega as bibliotecas
library(sandwich)
library(fixest)
library(tidyverse)
library(glue)
library(skimr)
library(knitr)
library(kableExtra)
library(stargazer)
library(texreg)

#' Tabela 1 com a descricao das variaveis
desc_tbl <- read_csv("descricao.csv")

kbl(desc_tbl, booktabs = TRUE, longtable = TRUE, format = "latex",
    col.names = c("Variável", "Descrição"),
    caption = "Variáveis e descrições",
    label = "descricao") %>%
  kable_styling(full_width = FALSE,
                latex_options = c("repeat_header"),
                repeat_header_text = "(continuação)") %>%
  column_spec(2, width = "30em") %>%
  save_kable(file = "./Tables/table_descricao.tex")

#' Carregando os dados
data <- haven::read_dta("data_final.dta") %>%
  as.data.frame()

# Descrevendo o desenho do experimento ---------------------------------

#' Tabela contando o numero de recipientes de cada tratamento
#'
data$treatment <- factor(data$treatment)
t_sizes <- as.vector(table(data$treatment))

buckets <- data.frame(treatment = sort(unique(data$treatment)),
                      Buckets = c("T0", "T1", "T2", "T3", "T4", "T5", "T6"),
                      Description = c("Sem Correio", "Correio", "Ameaça", "Info",
```

```
                                        "Info&Ameaça", "Moral", "Moral&Ameaça"),
                      Size = t_sizes) %>%
  mutate(Prop = Size / nrow(data))

#' Tabela 2 recipientes de cada tratamento
kbl(buckets[-1], format = "latex", booktabs = TRUE, label = "descritivas1",
    col.names = c("Tratamento", "Descrição", "Observações", "Proporção"),
    caption = "Distribuição dos tratamentos na amostra.") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  kable_classic(full_width = FALSE) %>%
  save_kable("./Tables/table_descritivas1.tex")

#' Junta de volta as inforamacoes de bucket e descricao para os dados
data <- data %>%
  left_join(buckets[, c("treatment", "Buckets", "Description")], by = "treatment")
#' Estatisticas descritivas da base
#' Variaveis com valor NA
skim_df <- data %>%
  skim_without_charts()

na_df <- skim_df %>%
  filter(n_missing > 0) %>%
  select(skim_variable, n_missing, complete_rate)

#' Tabela 3 dados faltantes
kbl(na_df, digits = 2, format = "latex", booktabs = TRUE, label = "missings",
    col.names = c("Variável", "No. Faltantes", "Completude"),
    caption = "Dados faltantes na amostra.") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  kable_classic(full_width = FALSE) %>%
  footnote(general_title = "Nota:",
           general = "Completude refere-se a proporção de linhas preenchidas contra fa
           threeparttable = TRUE) %>%
  save_kable("./Tables/table_missings.tex")

#' Analise Exploratoria
#' Problema de atrito
attrition_level <- data %>%
  filter(mailing == 1) %>%
  group_by(treatment, Buckets, Description) %>%
  summarise(mail_count = n(),
            deliv_na_count = sum(is.na(delivered)),
            deliv_0_count = sum(delivered == 0),
            attr_rate = (deliv_na_count + deliv_0_count) / nrow(cur_data()))
chi_test <- chisq.test(attrition_level$deliv_0_count,
                       p = attrition_level$mail_count, rescale.p = TRUE)
atrito_foot <- glue("Na média total a taxa de atrito foi de {format(mean(attrition_lev
#' Tabela 5 detalhando o nivel de atrito por tratamento
kbl(attrition_level[-1], digits = 4, format = "latex", booktabs = TRUE,
    label = "atrito-level",
    col.names = c("Tratamento", "Descrição", "Cartas", "Entregues NA",
                  "Não Entregues", "Taxa Atrito"),
    caption = "Taxa de atrito por tratamento.") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  kable_classic(full_width = FALSE) %>%
  footnote(general_title = "Nota:", general = atrito_foot,
           threeparttable = TRUE) %>%
  save_kable("./Tables/table_atrito_level.tex")
```

```r
#' Todos os NAs se referem ao grupo de controle, mas houve um pouco de atrito.
#' Verificar balanceamento de variaveis para aqueles que atritaram
atrito_controle <- data %>%
  filter(mailing == 0 | (mailing == 1 & delivered == 0))

attr_bal <- atrito_controle %>%
  group_by(treatment, Buckets) %>%
  summarise(across(c(gender, age_aver, inc_aver, pop2005, pop_density2005,
                     compliance),
                   mean, na.rm = TRUE))
#' Teste anova para diferenca de medias
attr_anova <- tibble(var = c("gender", "age_aver", "inc_aver", "pop2005",
                                    "pop_density2005", "compliance")) %>%
  rowwise() %>%
  mutate(anov = list(anova(lm(paste0(var, "~treatment"), data = atrito_controle))),
         pval = anov["treatment", "Pr(>F)"]) %>%
  select(var,pval) %>%
  pivot_wider(names_from = var, values_from = pval) %>%
  add_column(Buckets = "Anova p-valor", .before = 1)

atr_bal_foot <- "Gênero igual a zero para mulher. Demais variáveis são denominadas em
#' Tabela 6 balanceamento com atrito
attr_bal[-1] %>%
  bind_rows(attr_anova) %>%
  kbl(digits = 4, format = "latex", booktabs = TRUE, label = "atrito-bal",
      col.names = c("Tratamento", "Gênero", "Idade", "Renda", "População",
                    "Dens. pop.", "Compliance"),
      caption = "Análise de atrito. Balanceamento de variáveis selecionadas") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  kable_classic(full_width = FALSE) %>%
  footnote(general_title = "Nota:",
           threeparttable = TRUE,
           general = atr_bal_foot) %>%
  save_kable("./Tables/table_atrito_bal.tex")
#' Histograma com designacao de tratamento e atrito
hist1 <- data %>%
  filter(treatment %in% c("0", "6"), pop_density2005 < 30) %>%
  select(treatment, pop_density2005) %>%
  ggplot(aes(x = pop_density2005, y = ..density.., fill = treatment)) +
  geom_histogram(bins = 50, alpha = 0.8, position = "dodge") +
  labs(x = "",
       y = "Frequência relativa",
       title = "Tratamento designado") +
  guides(fill = guide_legend(title = "Tratamento")) +
  scale_fill_discrete(type = c("blue", "red")) +
  theme_classic()

hist2 <- atrito_controle %>%
  filter(treatment %in% c("0", "6"), pop_density2005 < 30) %>%
  select(treatment, pop_density2005) %>%
  ggplot(aes(x = pop_density2005, y = ..density.., fill = treatment)) +
  geom_histogram(bins = 50, alpha = 0.8, position = "dodge") +
  labs(x = "Densidade Populacional (hab/km2)",
       y = "Frequência relativa",
       title = "Atrito") +
  guides(fill = guide_legend(title = "Tratamento")) +
  scale_fill_discrete(type = c("blue", "red")) +
  theme_classic()
```

```
png("./Figs/fig_atr_hist.png")
gridExtra::grid.arrange(hist1, hist2)
dev.off()

#' Replicacao das tabelas 1 e 2 de Fellner et al.
#'
#' Tabela 1
gtab1 <- data %>%
  group_by(treatment, Buckets, Description) %>%
  summarise(across(c(gender, age_aver, inc_aver, pop2005,
                     pop_density2005, compliance),
                   mean, na.rm = TRUE))

anova_results <- data.frame(var = c("gender", "age_aver", "inc_aver", "pop2005",
                                    "pop_density2005", "compliance")) %>%
  rowwise() %>%
  mutate(anov = list(anova(lm(paste0(var, "~treatment"), data = data))),
         pval = anov["treatment", "Pr(>F)"])
anov_row <- anova_results %>%
  select(-anov) %>%
  pivot_wider(names_from = var, values_from = pval) %>%
  mutate(Buckets = "Anova: ",
         Description = "p-values") %>%
  select(Buckets, Description, everything())

#' Tabela 4 balanceamento. Table 1 de Fellner et. al
tbl1_cap <- "Balanceamento de características individuais e por município por tipo de
gtab1[-1] %>%
  bind_rows(anov_row) %>%
  kbl(digits = 4, booktabs = TRUE, format = "latex", label = "tab1",
      col.names = c("Tratamento", "Descrição", "Gênero", "Idade", "Renda",
                    "População", "Dens. pop.", "Compliance"),
      caption = tbl1_cap) %>%
  kable_styling(latex_options = "HOLD_position", font_size = 10) %>%
  kable_classic(full_width = FALSE) %>%
  footnote(general_title = "Nota:",
           threeparttable = TRUE,
           general = atr_bal_foot) %>%
  save_kable(file = "./Tables/table1.tex")

#' Regressoes da Tabela 7
#'
reg_21 <- feols(resp_A~mailing+threat+appeal+info, data = data)
reg_22 <- feols(resp_A~mailing+threat+appeal+info+i_tinf+i_tapp, data = data)

delivered <- data %>%
  filter(delivered == 1)
reg_23 <- feols(resp_B~threat+appeal+info, data = delivered)
reg_24 <- feols(resp_B~threat+appeal+info+i_tinf+i_tapp, data = delivered)
reg_25 <- feols(resp_all~threat+appeal+info, data = delivered)
reg_26 <- feols(resp_all~threat+appeal+info+i_tinf+i_tapp, data = delivered)

#' Dicionário para o nome das variáveis nas tabelas
fixest::setFixest_dict(c(resp_A = "Registro",
                         resp_B = "Atual. Contratual",
                         resp_all = "Resposta Geral",
                         mailing = "Correio",
                         threat = "Ameaça",
```

```
                          appeal = "Moral",
                          info = "Info",
                          i_tinf = "Ameaça x Info",
                          i_tapp = "Ameaça x Moral",
                          threat_evasion_D1 = "Ameaça x Evasão",
                          appeal_evasion_D1 = "Moral x Evasão",
                          info_evasion_D1 = "Info x Evasão",
                          evasion_1 = "Evasão",
                          threat_evasion_D2 = "Ameaça x Evasão",
                          appeal_evasion_D2 = "Moral x Evasão",
                          info_evasion_D2 = "Info x Evasão",
                          evasion_2 = "Evasão",
                          "(Intercept)" = "Constante"))
#' Ajusta o estilo das tabelas
est_style = list(depvar = "title:Dep. Var.",
                 model = "title:Modelo",
                 var = "title:\\emph{Variáveis}",
                 stats = "title:\\emph{Estatísticas de diagnóstico}",
                 notes = "title:\\emph{\\medskip Notas:}")
#' Cria a Tabela 7. Table 2 in Fellner et. al
esttex(reg_21, reg_22, reg_23, reg_24, reg_25, reg_26,
       file = "./Tables/table2.tex",
       label = "tab:tab2",
       style = est_style,
       replace = TRUE,
       se = "White",
       digits = 3,
       fitstat = "",
       order = c("Correio", "^Ameaça$", "^Moral$", "^Info$", "Ameaça x Moral",
                 "Ameaça x Info", "Constante"),
       title = "Efeito do tratamento nos registros, atualizações contratuais, and resp

#' Efeitos Heterogeneos
#' Replicacao da Table C1
#' Mediana da população dos municípios, da densidade, da renda e de votantes
#' a direita
med_pop <- median(data$pop2005, na.rm = TRUE)
med_den <- median(data$pop_density2005, na.rm = TRUE)
med_rend <- median(data$inc_aver, na.rm = TRUE)
med_vot <- median(data$vo_cr + data$vo_r, na.rm = TRUE)
#' Dataframe com indicadores de acima da mediana
efeito_het_df <- data %>%
  mutate(pop_hi = pop2005 >= med_pop,
         popdens_hi = pop_density2005 >= med_den,
         rend_hi = inc_aver >= med_rend,
         vot_hi = (vo_cr + vo_r) >= med_vot)
#' Conjunto de variáveis de controle
Z_extended <- gsub("\\s+", "+",
                   "pop_density2005 pop2005 nat_EU nat_nonEU fam_marri fam_divor_widow
                   edu_hi edu_lo rel_evan rel_isla rel_orth_other rel_obk pers2 pers3
                   pers5more  vo_r vo_cl vo_l  j_unempl j_retire j_house j_studen
                   inc_aver age0_30 age30_60  bgld kaern noe ooe salzbg steierm
                   tirol vlbg wien schober")
#' Regressões para efeitos heterogêneos
#' formulas
form_str <- paste0("resp_A~threat+appeal+info+gender+compliance_t+", Z_extended)
#' Default do erro padrão é ser robusto a heterocedasticidade
fixest::setFixest_se(no_FE = "white")
```

```
het_reg_pop <- efeito_het_df %>%
  filter(delivered == 1) %>%
  group_by(pop_hi) %>%
  summarise(lm_model = list(feols(as.formula(form_str), data = cur_data())))

het_reg_den <- efeito_het_df %>%
  filter(delivered == 1) %>%
  group_by(popdens_hi) %>%
  summarise(lm_model = list(feols(as.formula(form_str), data = cur_data())))

het_reg_rend <- efeito_het_df %>%
  filter(delivered == 1) %>%
  group_by(rend_hi) %>%
  summarise(lm_model = list(feols(as.formula(form_str), data = cur_data())))

het_reg_vot <- efeito_het_df %>%
  filter(delivered == 1) %>%
  group_by(vot_hi) %>%
  summarise(lm_model = list(feols(as.formula(form_str), data = cur_data())))

#' Tabela 8
esttex(het_reg_pop$lm_model, het_reg_den$lm_model, het_reg_rend$lm_model,
       het_reg_vot$lm_model,
       file = "./Tables/tablec1.tex",
       label = "tab:tabc1",
       style = est_style,
       replace = TRUE,
       se = "White",
       digits = 3,
       fitstat = "",
       keep = c("^Ameaça$", "^Moral$", "^Info$"),
       order = c("^Ameaça$", "^Moral$", "^Info$"),
       title = "Efeito heterogêneo do tratamento. Modelo de regressão linear.")
```