Efeitos de gênero no desempenho de artigos

Ricardo S. Castro

O presente trabalho busca medir a diferença de desempenho acadêmico entre homens e mulheres. Mais especificamente, utilizo dados de repositórios de artigos da web para entender como o gênero dos autores afeta o número de citações que um trabalho obteve.

Embora fuja do escopo desse trabalho diferenciar as causas desse possível *gap*, a cargo de motivação, vale a pena elencar algumas de suas possíveis explicações. Paloma 2000 faz um resumo de vários efeitos de gênero presentes no mercado de trabalho e perfil educacional: mulheres recebem menos incentivos a perseguir carreira acadêmica; a diferenciação de planos de carreira por gênero; menor representatividade feminina nas *hard science's*; a jornada dupla; e a proporção de mulheres no topo de cada carreira é sempre menor, mesmo em setores bem representados na base. Muitos desses motivos, alguns mais diretamente que outros, indicam a existência de algum tipo de discriminação, porém, sem uma análise empírica robusta, não é possível descartar que uma diferença de desempenho possa ser explicada por outros motivos.

Dados

A base de dados foi construída a partir as API's de alguns repositórios. A lista de artigos foi retirada do ArXiv, repositório focado em artigos de matérias exatas, e o mais amigável para obter grandes quantidades de dados. Esse site nos dá acesso aos nomes dos autores, instituição e jornal associados, e o tema do artigo. Foram captados 165 artigos do ano de 2019.

A seguir foi preciso criar a variável de gênero a partir de modelos preditivos sobre o nome dos autores. Foram usadas duas bases de dados: uma com 40 mil nomes dos países norte americanos, europeus, e dos maiores países asiáticos, publicada pela revista alemã C'T; e outra com 32.5 mil nomes dos EUA, da Social Security Administration. Os resultados foram rodados com as duas bases e comparados. Nomes de países fora do escopo de cada base foram excluídos. Foram removidos da amostra autores com nomes considerados andrógenos, isto é, os que aparecem com proporções parecidas em homens e mulheres[1].

Os dados do Semantic Scholar foram utilizados para obter o número de citações, e uma estatística de "número de citações influentes"(desenvolvida por Valenzuela, Ha e Etzioni 2015). A API do Altmetrics, serviço que coleta dados sobre artigos nas redes sociais, foi utilizado para dados de desempenho no Twitter. Os dados da ORCID, base de dados sobre pesquisadores, foi utilizada para obter dados acerca da educação dos autores. Essa variável deixa a desejar, é uma dummie de nível de diploma (PhD, MBA, etc.), e está disponível para poucas observações. Seria interessante ter mais dados sobre a carreira dos autores, algo disponibilizado pelo Google Scholar, que tem dados sobre número de artigos, número de jornais em que publicou, e o h-index dos autores. Porém, não foi possível obter acesso a esse repositório.

A seguir são apresentadas algumas estatísticas descritivas para motivar a análise.

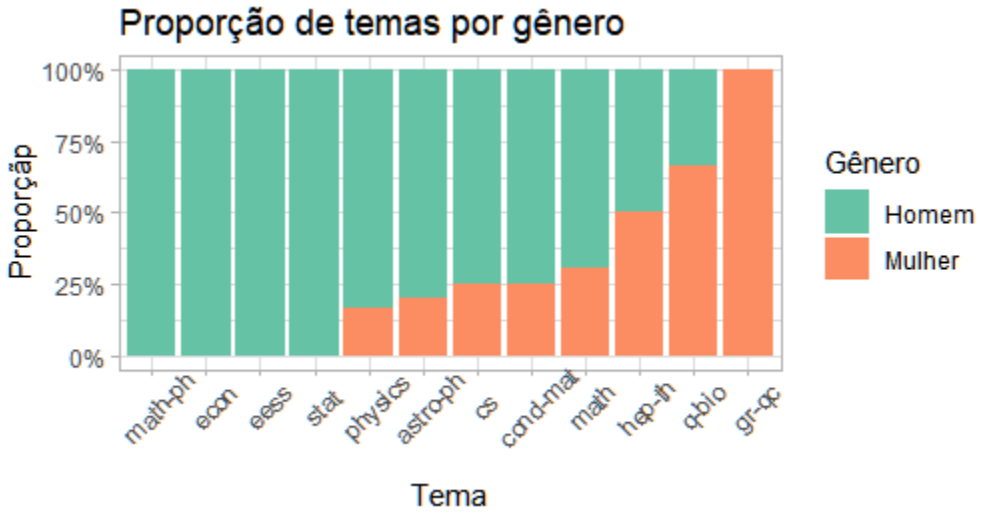
A tabela 1 mostra as diferenças de médias entre os gêneros para as seguintes variáveis, em ordem: número de artigos associado à uma universidade e a um jornal; número médio de autores no grupo; número de citações e de citações influentes; número de diplomas no grupo, grupos com algum diploma, grupos com algum PhD; média de citações do jornal onde o artigo foi publicado; número de Tweets sobre o artigo, e

[1]Os resultados foram obtidos com uma diferença mínima de 50% entre as proporções. Diferenças de 60% e 90% foram computadas e os resultados se mantiveram

Tabela 1: Diferenças nas covariadas

Variável	Mulheres	Homens	P-valor
Universidade	0	0.01	2.4e-01
Jornal	0.32	0.6	2.4e-06
#Autores	1.43	3.83	1.3e-28
Citações	6.07	11.12	2.1e-04
Cit. nfluentes	0.71	1.04	9.9e-02
Educ média	0.58	1.03	5.1e-07
Educ cat.	0.32	0.59	1.3e-08
Educ PhD	0.06	0.19	1.9e-05
Alt. journal	59.29	118.62	7.2e-28
Tweets	6.75	15.37	7.0e-18
Público	5.96	12.21	1.0e-13
Alt. score	7.36	19.05	4.3e-10

Figura 1



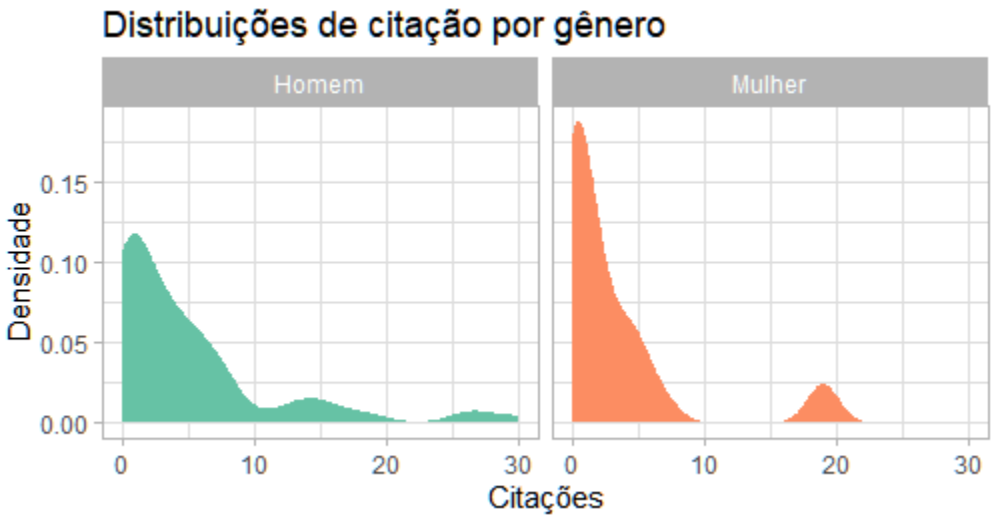
número desses feitos pelo público não-pesquisador; Score de popularidade feito pelo Altmetric. Vemos que os homens tem valores significantemente maiores em basicamente todas as métricas.

A diferença das distribuições de citações é melhor analisada na figura 2. A diferença de médias, junto com as diferentes proporções entre autores homens e mulheres através dos diferentes tema (Figura 1), mostra que de fato a distribuição de homens e mulheres difere em várias dimensões importantes para o sucesso acadêmico.

Metodologia

Como muitos artigos são escritos em grupos de autores, nossa análise foi separada em duas: primeiro utilizando a porcentagem de homens no grupo como variável explicativa; e segunda excluindo os grupos mistos e lidando apenas com uma *dummy* de gênero. Como dito anteriormente, também são utilizadas as

Figura 2



duas variáveis dependentes obtidas pelo Semantic Scholar, bem como duas bases de dados distintas.

A variável de educação também precisa ser agregada. Foram escolhidas três medidas: número de diplomas médio do grupo, proporção do grupo com ao menos um diploma, e proporção do grupo com ao menos PhD.

Estimaremos a seguinte regressão por MQO:

$$Y_i = \beta_0 + \beta_1 D_i + X_i \Gamma + \varepsilon_i$$

Onde Y_i é uma das duas variáveis dependentes obtidas pelo Semantic Scholar, D_i é a proporção ou a dummie de gênero, e X_i são os controles supracitados.

Dentre as hipóteses-padrão de MQO, vale apontar as seguintes: assumimos que podemos utilizar a análise assintótica sobre os estimadores, por mais que nossa amostra seja pequena; não temos heterocedasticidade[2], de modo que o modelo perde em eficiência, e os erros padrão reportados foram corrigidos por White; e temos motivos para desconfiar de endogeneidade no modelo. Essas três hipóteses serão discutidas mais a fundo na seção de limitações.

Resultados

Alterar a variável dependente não mudou os resultados, portanto serão reportados apenas as regressões feitas com o número total de citações. A tabela de resultados está dividida entre as duas bases de dados, e para cada base, foi feita a regressão usando a proporção de mulheres no grupo, e a dummie de gênero.

Os coeficientes de interesse foram significativos na base SSA, mas não na C'T. Uma possível explicação para isso é o fato de que a C'T tenta abranger um número muito maior de países, de modo que a chance de erro é maior, e a variável de gênero não é tão confiável[3]. Não podemos garantir que esses efeitos sejam *ceteris paribus*, mas a estimação com a base SSA aponta na direção de melhor desempenho para homens: 10% mais homens em um grupo geram 1.105 mais citações, e um grupo somente masculino gera 8.7 mais citações que grupos femininos.

[2]Identificado através do teste de Breush-Pagan sobre $\hat{\varepsilon}_i$

[3]Quanto pior a variável de gênero, mais subestimado o efeito será (mais discutido na seção de limitações).

Vale citar que o fato de ambas as variáveis dependentes terem retornados os mesmos resultados indica que o efeito de gênero deve ser o mesmo para artigos de diferentes qualidades.

Para defender que a variável de gênero da SSA é melhor, vale analisar quais variáveis poderiam estar capturando o efeito de gênro na estimação com a base C'T.

O número de diplomas médio no grupo só foi significativo na esmiação C'T, o que faz sentido dada a tendência de menor educação nas *hard science's* citada na introdução.

A quantidade dos Tweets feitos pelo público geral (não-pesquisadores) deixa de ser significante, isso poderia ser explicado pelo fato estilizado de haver menos mulheres no topo da carreira, logo menos mulheres com contas influentes no Twitter. Grupos masculinos terão mais pesquisadores influentes, e se o controle de gênero não for bom - como nas duas primeiras regressões não era - a variável "Público"deveria estar capturando esse fato erroneamente.

Porém, ainda existem muitas lacunas. Alguns resultados não são tão fáceis de explicar, como Tweets serem negativamente relacionados, e o fato de termos algumas variáveis bem importantes para o modelo omitidas. Em suma, temos motivos para acreditar que o resultado de C'T está subestimado, mas não temos robustez suficiente para aceitar os resultados do modelo SSA como verdadeiros.

Resultados

	Dependent variable: # Citações			
Dados:	C'T		SSA	
% homem	1.02 (2.38)		11.05* (5.65)	
Grupo masculino		−0.09 (1.77)		8.70** (3.48)
Universidade	5.55 (5.53)	5.83 (5.55)	−4.74 (20.31)	2.67 (20.15)
Jornal	0.92 (2.58)	0.75 (2.58)	0.61 (3.99)	−0.80 (3.91)
#Autores	0.18 (1.02)	0.18 (1.05)	3.72** (1.77)	5.28*** (1.83)
Educ média	8.27** (4.14)	8.12* (4.14)	5.49 (5.37)	4.70 (5.18)
Educ cat.	−12.99 (7.81)	−12.61 (7.81)	−10.03 (10.33)	−10.16 (9.92)
Educ PhD	4.26 (5.10)	4.33 (5.11)	4.60 (7.65)	6.29 (7.41)
Alt. journal	−0.004 (0.03)	−0.01 (0.03)	0.01 (0.05)	0.02 (0.05)
Tweets	−0.59 (0.41)	−0.56 (0.41)	−1.39** (0.60)	−1.44** (0.57)
Público	0.75** (0.31)	0.75** (0.31)	0.28 (0.48)	0.14 (0.46)
Alt. score	0.15 (0.15)	0.14 (0.15)	0.25 (0.41)	0.33 (0.40)
Intercepto	1.93 (10.05)	3.56 (10.00)	−21.23 (21.60)	−24.30 (20.54)
Dummies tema	Sim	Sim	Sim	Sim
Observações	165	165	96	96
R^2	0.70	0.70	0.85	0.87
Estatística F	1.90***	1.89***	2.22**	2.43***

Limitações

Mesmo antes de discutir as limitações à isolar o efeito causal do gênero, este trabalho encontra algumas limitações técnicas. Mesmo removendo nomes andrógenos, a variável de gênero não é 100% confiável, o

que faz com que o efeito seja subestimado, uma vez que o grupo que performaria pior está parcialmente sendo levado em conta nos resultados do grupo que performaria melhor, jogando tal resultado para baixo, e vice versa. Outra limitação técnica é o tamanho e abrangência da amostra: os repositórios permitem utilizar um número muito maior de dados do que a amostra desse trabalho; outras bases sobre de gêneros de nomes poderiam ter sido utilizadas para aumentar o número de países e culturas analisadas. Nessa mesma linha de amostra limitada, a lista de artigos está limitada à alguns temas específicos, algo problemático dado que a distribuição de proporção de mulheres escrevendo sobre cada tema não é uniforme.

O modelo também apresenta problemas de endogeneidade, existem algumas variáveis omitidas que devem estar relacionadas com gênero e afetam o desempenho dos papers. Escolaridade, cujo controle no modelo foi sensivelmente fraco, provavelmente afeta a variável dependente, e existem teorias acerca de menor nível de educação sobre *hard-sciences* dentre as mulheres.

O tema de cada artigo também deve estar relacionado com gênero, uma vez que existe literatura apontando que algumas carreiras/matérias são mais perseguidas por mulheres que homens. Ao mesmo tempo, artigos de algumas matérias são muito mais citados do que de outras, por exemplo, é possível imaginar que um artigo de matemática pura deva ser utilizado para a prova de um numero pequeno de outros artigos, enquanto um tema muito mais aplicado deve ter mais citações em média.

As variáveis acerca da carreira dos autores, provavelmente relacionada com gênero dada a motivação feita na introdução, também estão omitidas.

Conclusões

A análise presente nesse artigo buscou medir diferenças no desempenho de artigos de acordo com o gênero dos autores. Existem pesquisas teóricas e empíricas sobre diferenças entre os gêneros em âmbitos diretamente conectados com a academia, e este trabalho encontrou aluma evidência, por mais que limitada, da existência desse *gap* nesse âmbito também.

Os resultados apontam que 10% mais homens em um grupo geram 1.105 mais citações, e um grupo somente masculino gera 8.7 mais citações que grupos femininos.

Embora os resultados devem ser vistos com cautela, dada das limitações empíricas, creio que tanto o modelo quanto a variedade de covariadas, obtidas através da junção de vários repositórios diferentes, pode ser reaproveitada em uma nova análise que consiga uma maior quantidade e melhor qualidade de dados.

Referências

Palomba Rosella. *Figlie di Minerva*, 2000. Disponível em: `https://www.galileonet.it/figlie-di-minerva/`.