COCS 6323: Statistical Methods in Research

Group Project

Group 2

Department of Computer Science

University of Houston

May 2, 2019

Contents

List of Tables

List of Figures

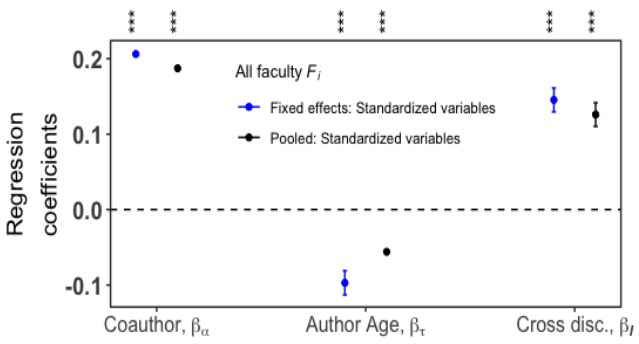
1 Contribution

Member	**Contribution**
Bradley Macdonald	Analyze regression models of Figure 4
Tung Huynh	Preprocess Data, create and analyze regression models of Table S2, Table S3
Yifan Zhang	Preprocess Data, and draw plot of Figure 4

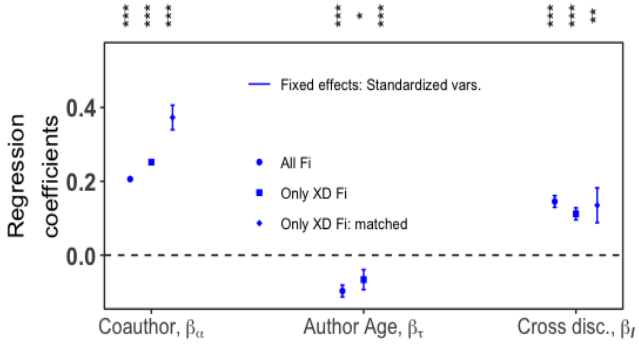
Table 1: Contribution of group members of the second milestone

2 Figure 5

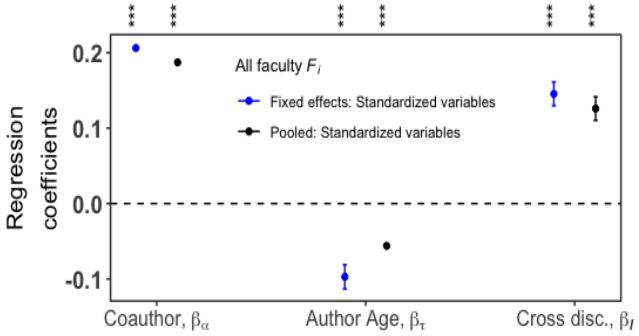
A



B



C



D

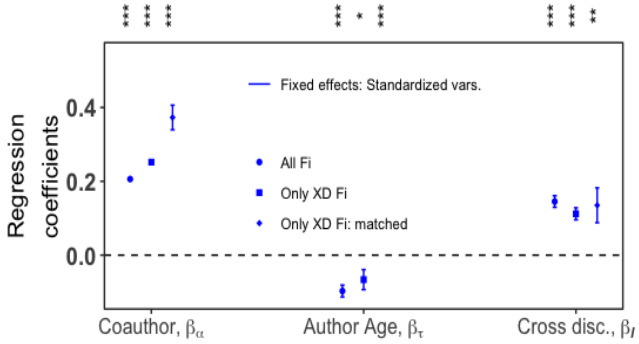


Figure 1: Career panel regression model

Figure 5A displays model estimates for Coauthors, Author age and Crossdisciplinarity. Interestingly for the purposes of this project, there is a significant effect of crossdisciplinarity on citation impact. Thus, it can be concluded that cross disciplinary papers are more heavily cited than biology or computer science genomics papers and therefore cross-disciplinarity positively effects the impact which publications make on the body of knowledge in the field of genomics. Significant effects for Co-authorship and Author age are also observed. The differences between fixed effects models and pooled models appear to be minimal, with similar significant relationships being observed in each case regardless of fixed effects. Figure 5B also examines the relationships of Coauthors, Author age and Crossdisciplinarity on citation impact, with fixed effects models being compared across different subsets of the overall data set. The figure compares all F_i, all XD F_i, and the pair matched subset of 53 matched F_i. Significant effects of all three characteristics are observed for each of the three F_i subsets. It certainly makes sense that the more coauthors a paper has, the greater the citation impact would be. Large research teams tend to attract more attention than small ones in many fields. The important relationship between crossdisciplinarity and citation impact remains in all three F_i subsets as well, demonstrating that

the robust model is able to demonstrate that cross disciplinary publications are more likely to make a large impact in the literature when compared to biology or computer science genomics papers. Figure 5C displays the probability distribution for the placebo model in order to test for effects greater than the set Cross-disciplinarity coefficient shown in the blue dashed line. This shows that it is not down to chance that cross disciplinary papers made a greater impact in the literature than single-discipline papers in genomics. Figure 5D repeats the result displayed in Figure 5C, with the difference being that only XD F_i are included in the experiment. Once again, a similar result is observed, no value within the cross-disciplinarity probability distribution is as large as the set coefficient. This further confirms the overarching result of this paper, that cross-disciplinarity is related to and directly impacts the citation impact which a paper makes in the genomic literature.

3 Supplementary Table S4

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of author, β_α	0.2836***	0.1872***	0.312***	0.2061***
	(0.0025)	(0.0016)	(0.00262)	(0.00173)
Career age, β_τ	-0.00547***	-0.0560***	-0.00949***	-0.0971***
	(0.0002)	(0.0018)	(0.00156)	(0.01601)
Cross-disciplinary indicator, β_I	0.1259***	0.1259***	0.1453***	0.1453***
	(0.0157)	(0.0157)	(0.01578)	(0.01578)
Network characteristics				
Author centrality, β_ζ	0.0440***	0.0284***	X	X
	(0.0029)	(0.0018)		
Bridge ratio, β_λ	0.3338***	0.1210***	X	X
	(0.0055)	(0.0020)		
Discipline (F) dummy	0.00790*	0.0079*	X	X
	(0.0033)	(0.0033)		
Constant	0.4586***	0.1638*	-0.2932***	-0.0669**
	(0.0761)	(0.0728)	(0.0456)	(0.0176)
Year dummy	Y	Y	Y	Y
n	413,565	413,565	413,565	413,565
adj. R^2	0.055	0.055	0.036	0.036

Standard errors in parentheses below estimate * p $\leq$ 0.05, ** p $\leq$ 0.01, *** p $\leq$ 0.0001

Table 2: Career data set: Panel model on all falcuty

All faculty F were included in the panel model. Columns included in the table are divided by the inclusion of fixed effects, with no fixed effects included in the first two columns and fixed effects included in the following two. The second and fourth columns are standardized using the effect of a one standard deviation shift in the independent variable on the dependent variable. Although all faculty were meant to be included in the model, n = 3900 due to the inclusion of F_i who have defined page rank centrality within the data set. All relevant publication characteristics and network characteristics are shown to have a significant effect of the scholarly impact which publications within the data set have. In fact, all characteristics within the table can be considered highly significant due to a p-value of less than or equal to 0.001.

4 Supplementary Table S5

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of author, β_α	0.329***	0.236***	0.351***	0.252***
	(0.0037)	(0.0027)	(0.00392)	(0.00282)
Career age, β_τ	-0.00499***	-0.0536***	-0.00617*	-0.0663*
	(0.0003)	(0.0030)	(0.00253)	(0.0271)
Cross-disciplinary indicator, β_I	0.1095***	0.1095***	0.112***	0.112***
	(0.0165)	(0.0165)	(0.0162)	(0.0162)
Network characteristics				
Author centrality, β_ζ	0.0526***	0.0333***	X	X
	(0.0046)	(0.0029)		
Bridge ratio, β_λ	0.3192***	0.1116***	X	X
	(0.0092)	(0.0032)		
Discipline (F) dummy	-0.0383***	-0.0383***	X	X
	(0.0052)	(0.0052)		
Constant	0.2113	-0.0287	-0.408***	-0.0370*
	(0.1236)	(0.1187)	(0.0778)	(0.0285)
Year dummy	Y	Y	Y	Y
n	166,621	166,621	166,621	166,621
adj. R^2	0.067	0.067	0.049	0.049

Standard errors in parentheses below estimate * p $\leq$ 0.05, ** p $\leq$ 0.01, *** p $\leq$ 0.0001

Table 3: Career data set: Panel model on all XD_F falcuty

In order to conduct a robustness check for the panel model, all F_i with XD orientation (n = 1247) were included. The stricter inclusion criteria for this model appears to have impacted the effect of the characteristics included in the panel model when compared to the results in Table S4. Although all characteristics still appear to have a significant impact on the impact which researchers in the data set have on the body of genomics literature, some of these effects appear to be diminished in the new model. The p-values associated with the different models for Career age and Author centrality are no longer less than or equal to 0.001. This perhaps means that the impact of XD publications on the literature is not as dependent on these two characteristics.

5 Supplementary Table S6

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of author, β_α	0.329***	0.236***	0.351***	0.252***
	(0.0037)	(0.0027)	(0.00392)	(0.00282)
Career age, β_τ	-0.00499***	-0.0536***	-0.00617*	-0.0663*
	(0.0003)	(0.0030)	(0.00253)	(0.0271)
Cross-disciplinary indicator, β_I	0.1095***	0.1095***	0.112***	0.112***
	(0.0165)	(0.0165)	(0.0162)	(0.0162)
Network characteristics				
Author centrality, β_ζ	0.0526***	0.0333***	X	X
	(0.0046)	(0.0029)		
Bridge ratio, β_λ	0.3192***	0.1116***	X	X
	(0.0092)	(0.0032)		
Discipline (F) dummy	-0.0383***	-0.0383***	X	X
	(0.0052)	(0.0052)		
Constant	0.2113	-0.0287	-0.408***	-0.0370*
	(0.1236)	(0.1187)	(0.0778)	(0.0285)
Year dummy	Y	Y	Y	Y
n	166,621	166,621	166,621	166,621
adj. R^2	0.067	0.067	0.049	0.049

Standard errors in parentheses below estimate * p $\leq$ 0.05, ** p $\leq$ 0.01, *** p $\leq$ 0.0001

Table 4: Career data set: Panel model on all XD_F falcuty: matching

This table contains the results of a panel model robustness check with and without fixed effects. In order to perform the robustness check, F_i were selected for inclusion based on matched pairs. Criteria for matched pairs included the matching of an XD publication to a non-XD publication which were published within two years of each other, with within 20% similar number of coauthors. F_i included in the model (n = 53) had at least 10 such matched pairs. Using the matching process, a counterfactual was able to be properly defined for each XD publication. Thus, underlying connections between XD publications and greater impact on the body of literature could be explored. Similar to the comparison between Tables S4 and S5, with the even stricter criteria for inclusion in the model for Table S6 the effect of certain characteristics within the model on a publication's scholarly impact changes in this model. The number of coauthors which a publication has remains strongly related to scholarly impact with a p-values of less than or equal to 0.001. There is no significant effect of career age observed for the model without fixed effects, whereas career age with fixed effects remains highly significant. The significant effects of author centrality and bridge fraction remain, although they are only significant to

alpha levels of 0.05 and 0.01, respectively. As the data set is more and more limited, most of the significant effects remain with regards to the impact which publications make on the body of knowledge in genomics.