COSC 6323 - Statistical Methods in Research
Project Phase - 3

Members: Team-8

1. Farah Naz Chowdhury, ID:1798957, fchowdhury4@uh.edu
2. Md Rafiqul Islam Rabin, ID:1797648, mrabin@central.uh.edu
3. S M Salah Uddin Kadir , ID:1800503, ssalahuddinkadir@uh.edu

May 02, 2019.

Contributions: We sit together to discuss the requirement and task distribution. We divided task equally among us and agreed to support each other. The contribution status - Data Analysis (All Faculty - Rabin, XD only - Prity, Matched XD - Salah Uddin), Model (TableS4 - Rabin, TableS5 - Prity, TableS6 - Salah Uddin), Fig5 (5A+5B - Salah Uddin, 5C+5D - Rabin), and Report (Rabin, Prity, Salah Uddin). We always shared our progress/problem with each other and helped during implementation. We distributively contributed to our created git repository for the project. As we all completed our tasks and actively involved with each other all the times, we equally contributed to the project.

Fig. 5:

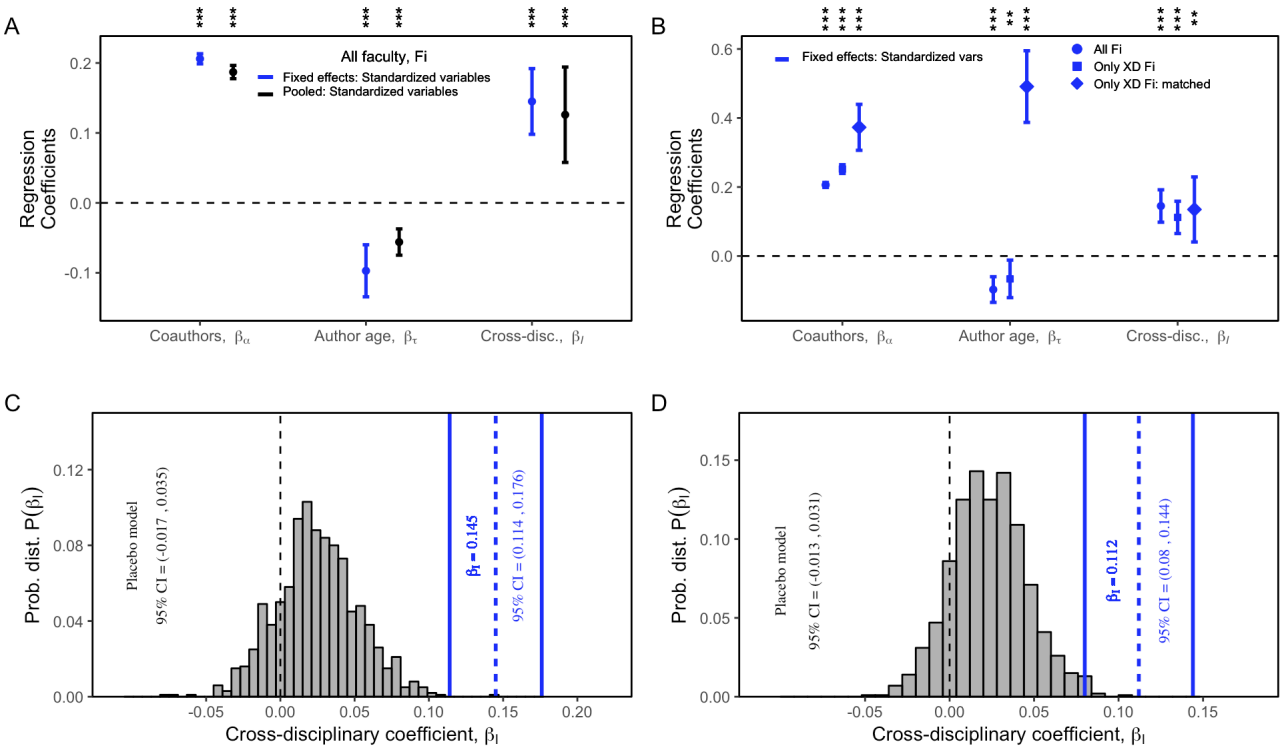

Fig. 5. Career panel regression model.

Description of figure content:

(**5A, 5B**) Co-authors (β_a), Author age (β_τ), and Cross-disciplinary indicator of publication(β_I) are three principal explanatory variables included in the fixed effect $\mathcal{F}$ career model. Regression coefficients are given in y-axis and explanatory variables are given in x-axis. Parameter estimates are given in TableS4, TableS5 and TableS6. Fig 5A shows the model estimates for those three explanatory variables and represents the relationship between explanatory variables and citation impact. The green color indicates fixed standardized effect and black color is pooled standardized effect. Fig 5B shows the fixed effect model estimates for (i) all faculty, (ii) only cross-disciplinary faculty, and (iii) cross-disciplinary matched pair faculty. The circle shape for all faculty, square shape for xd only faculty and diamond shape for matched xd faculty. (**5C, 5D**) The placebo randomization has been taken place to explore the spurious correlations. To check the robustness of panel regression model, in this regards, we shuffled the $I_{i,p}^{XD}$ value across the data set without replacement. For random 1000 faculty, we ran placebo randomization and repeat this 1000 times and calculate the value of β_I. The solid blue line in left side indicates the β_I value of third column of TableS4 and dashed blue line indicates the 95% confidence interval (CI). The x-axis shows the β_I value and y-axis shows the probability distribution in specified binwidth, calculated by count divided by total observation. Fig 5C shows the distribution of the placebo estimates $P(\beta_I)$ for all faculty $\mathcal{F}_i$, on the other hand, Fig 5D shows the distribution of the placebo estimates $P(\beta_I)$ for cross-disciplinary faculty $\mathcal{F}_{XD}$. The levels of statistical significance ($p - value$) are as follows: $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Observations, conclusions, and hypotheses:

- Fig 5A. The relationship between career model explanatory variables and citation impact are shown in this figure: (i) There is a positive relationship between cross-disciplinary and citation impact that means the average cross-disciplinary publication is more highly cited than the average disciplinary publication. (ii) We also observe a positive relationship between team size and citation impact that means the increase of citation is associated with the increase of team size. (iii) But we observe a negative relationship between career age and citation impact that means the citation decrease with each career year. Additionaly, the estimates with xd only faculty and matched xd faculty are also consistent which indicates the sign of robustness.

- Fig 5B. The parameter estimates are consistent for co-authors and cross-disciplinary indicator, both positive with (i) all

faculty, (ii) xd faculty, and (iii) matched xd faculty. On the other hand, the parameter estimates are not consistent for author age, as it is negative with (i) all faculty and (ii) xd faculty, but positive with (iii) matched xd faculty. This inconsistency of author age is because of the bias introducing by matched data of faculty's longitudinal profile.

- Fig 5C. The distribution of the placebo estimates $P(\beta_I)$ is for all faculty $\mathcal{F}_i$. The paper claimed 0% placebo estimates were larger than original estimates, but unfortunately we observed that around 10% placebo estimates cross the original estimates. We believe this is for the random selection and we plot the distribution with filtering those overlap.

- Fig 5D. The distribution of the placebo estimates $P(\beta_I)$ is for cross-disciplinary faculty $\mathcal{F}_{XD}$. The paper also claimed 0% placebo estimates were larger than original estimates, but unfortunately we again observed that around 5% placebo estimates cross the original estimates. We similarly believe this is for the random selection and we plot the distribution with filtering those overlap, as well.

Table S4:

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of co-authors, β_α	0.284***	0.187***	0.312***	0.206***
	(0.00718)	(0.00474)	(0.00547)	(0.00361)
Career age, β_τ	-0.00547***	-0.0560***	-0.00949***	-0.0971***
	(0.000919)	(0.00940)	(0.00182)	(0.0186)
Cross-disciplinary indicator, β_I	0.126***	0.126***	0.145***	0.145***
	(0.0341)	(0.0341)	(0.0235)	(0.0235)
Network characteristics				
PageRank centrality, $\beta_{\mathscr{C}^{PR}}$	0.0440**	0.0284**	X	X
	(0.0142)	(0.00920)		
Bridge fraction, β_λ	0.334***	0.121***	X	X
	(0.0256)	(0.0093)		
Discipline ($\mathscr{F}$) dummy	-0.00790	-0.00790	X	X
	(0.0139)	(0.0139)		
Constant	0.458**	0.164	-0.293***	-0.0670**
	(0.142)	(0.102)	(0.0528)	(0.0203)
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.
* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table S4. Career data set: Panel model on all faculty $\mathscr{F}$.

Description of figure content:

In phase 3 of our project we have implemented a panel regression model. The model leverages the longitudinal dimension of the career data which is disaggregated at the publication level. Following is the hierarchical panel regression model,

$$z_{i,p} = \beta_i + \beta_\alpha ln a_{i,p} + \beta_\tau \tau_{i,p} + \beta_I I_{i,p}^{XD} + D(t) + \epsilon_{i,p}$$

In the above equation coefficient, the response variable $z_{i,p}$ indicates the normalization citation score, that is mapped from the GS citation count $c_{i,p,s,t}$ for an article p that was published in year t by a faculty F_i from discipline s. Other control variables are: β_i stands for the author specific fixed effects, β_I represents the subset of cross-disciplinary publications, $a_{i,p}$ measures the total number of coauthors listed on each publication p, $\tau_{i,p}$ refers to the number of years since the researcher's first publication. $D(t)$ is the dummy year variable controlling for year specific shocks, and $\epsilon_{i,p}$ is the white noise.

Here, $z_{i,p} = [ln(1 + c_{i,p,s,t}) - \mu_t]/\sigma_t$, where μ_t is the mean and σ_t is the SD of the citation distribution, also +1 has been added to handle uncited publication. And $\tau_{i,p} = t_p - y_i^0$, where t_p publication year of current paper and y_i^0 is the first publication year.

The model without fixed effects incorporates time-independent author-level characteristics with following additional terms $[\beta_{\mathscr{C}^{PR}} \ln \mathscr{C}_i^{PR} + \beta_\lambda ln\lambda_i + D(\mathscr{F}_i)]$. We only analyzed the 3,900 scholars connected within the network as $\mathscr{C}_i^{PR}$ is defined. Here these additional variables are absorbed into β_i in the fixed effects model. The additional connectivity variable λ_i is the fraction of the total pollinators that are bridge pollinators. And finally $D(\mathscr{F}_i)$ represents the dummy discipline. This description are same for Table S4, Table S5, and Table S6.

In table S4, the robustness of panel model is checked with or without effects using all 3900 connected scholars. Here, each column represents estimted coefficients for a specific model. The dependent variable is the $z_{i,p}$, which is the normalized citation impact of an individual article of a faculty. The first two column represents a panel regression without F_i fixed effects and the last two represents the same for F_i fixed effects. The second and fourth column values are calculated using standardized variables. Each β coefficient indicates the change in $z_{i,p}$, which is associated with a one standard deviation shift in the corresponding independent variable. The robust standard errors are shown in parenthesis, X denotes time-independent variables absorbed by the fixed effects model and Y indicates additional fixed effects included in the regression model.

Observations, conclusions, and hypotheses:

In this Table S4, the robustness of the panel model has been tested on all faculty by exploring several variations, through without/with fixed effect and without/with standardized.

- The reason to normalization citation score is mainly taking care of following three statistical bias: (1) larger profile produce more citations, (2) older publication get more time than new publication, and (3) increasing publication rates and references list has significant change in citation over time.
- The panel model on the all faculty $\mathcal{F}_i$ has been implemented using all the 4,190 $\mathcal{F}_i$ with $\mathscr{O}(\mathcal{F}_i) = [BIO_\mathcal{F}, CS_\mathcal{F}, XD_\mathcal{F}]$.
- We observe the publication count, n = 413,565 by using $nrow(df)$.
- We also count the $I_{i,p}^{XD} = 1$ and get 3915.
- If we take a look at the significance level of estimate, then we can interpret that the parameters are significant to all model as $p - value$ is small.
- Here, for no fixed effects (first two columns) Adjusted R-squared = 0.055 and for fixed effects (last two columns) Adjusted R-squared = 0.039.

Therefore, in all cases, the results of the observed regression estimates are not significantly different, thus indicates the robustness of the panel model.

Table S5:

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of co-authors, β_α	0.329***	0.236***	0.351***	0.252***
	(0.0123)	(0.00884)	(0.00880)	(0.00632)
Career age, β_τ	-0.00499**	-0.0536**	-0.00616*	-0.0662*
	(0.00181)	(0.0194)	(0.00253)	(0.0272)
Cross-disciplinary indicator, β_I	0.109***	0.109***	0.112***	0.112***
	(0.0328)	(0.0328)	(0.0234)	(0.0234)
Network characteristics				
Author centrality, $\beta_{\mathscr{C}}$	0.0526*	0.0333*	X	X
	(0.0265)	(0.0168)		
Bridge fraction, β_λ	0.319***	0.112***	X	X
	(0.0493)	(0.0172)		
Discipline ($\mathscr{F}$) dummy	0.0383	0.0383	X	X
	(0.0256)	(0.0256)		
Constant	0.210	-0.0293	-0.409***	-0.0372*
	(0.239)	(0.170)	(0.0778)	(0.0291)
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.
* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table S5. Career data set: Panel model on the $XD_{\mathscr{F}}$ faculty.

Description of figure content:

In table S5, the robustness of panel model is checked with or without effects using only the 1,247 F_i with orientation $O(F_i) = XD_F$. Here, each column represents estimted coefficients for a specific model. The dependent variable is the $z_{i,p}$, which is the normalized citation impact of an individual article of a faculty. The first two column represents a panel regression without F_i fixed effects and the last two represents the same for F_i fixed effects. The second and fourth column values are calculated using standardized variables. Each β coefficient indicates the change in $z_{i,p}$, which is associated with a one standard deviation shift in the corresponding independent variable. The standard errors are shown in parenthesis, X denotes time-independent variables absorbed by the fixed effects model and Y indicates additional fixed effects included in the regression model. The description for panel model are same for Table S4, Table S5, and Table S6.

Observations, conclusions, and hypotheses:

In this Table S5, the robustness of the panel model has been tested on xd only faculty by exploring several variations, through without/with fixed effect and without/with standardized.

- The panel model on the $XD_{\mathcal{F}}$ faculty has been implemented using only the 1,247 $\mathcal{F}_i$ by filtering with $\mathscr{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$.
- We observe the publication count, n = 166,621 by using $nrow(df)$.
- We also count the $I_{i,p}^{XD} = 1$ and get 3915.
- If we take a look at the significance level of estimate, then we can interpret that the parameters are significant to all model as $p - value$ is small.
- Here, for no fixed effects (first two columns) Adjusted R-squared = 0.067 and for fixed effects (last two columns) Adjusted R-squared = 0.052.

Therefore, in all cases, the results of the observed regression estimates are not significantly different, thus indicates the robustness of the panel model.

Table S6:

	No Fixed Effects	**No Fixed Effects [Standardized]**	**Fixed Effects**	**Fixed Effects [Standardized]**
Publication characteristics				
# of co-authors, β_α	0.515***	0.374***	0.506***	0.387***
	(0.0452)	(0.0332)	(0.0440)	(0.0581)
Career age, β_τ	-0.0023	-0.0228	0.0471***	0.489***
	(0.0057)	(0.0612)	(0.00387)	(0.0653)
Cross-disciplinary indicator, β_I	0.133**	0.133**	0.135**	0.135**
	(0.0472)	(0.0472)	(0.0470)	(0.0470)
Network characteristics				
Author centrality, $\beta_{\mathscr{C}}$	0.218*	0.162*	X	X
	(0.0871)	(0.0557)		
Bridge fraction, β_λ	0.745**	0.221**	X	X
	(0.231)	(0.0711)		
Discipline ($\mathscr{F}$) dummy	-0.223	-0.205	X	X
	(0.125)	(0.113)		
Constant	-0.423	-1.545***	-2.257***	-0.436*
	(0.635)	(0.172)	(0.0881)	(0.0765)
Year dummy	Y	Y	Y	Y
n	1987	1987	1987	1987
adj. R^2	0.253	0.253	0.092	0.092

Standard errors in parentheses, below estimate.
* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table S6. Career data set: Panel model on the $XD_{\mathscr{F}}$ faculty with matched pairs.

Description of figure content:

The table describes the robustness of XD faculty members for matched pairs. We selected each google scholar (gs) who has at least 10 matched pairs of publications where each matched pair was selected based on three criteria. (1) All publications that are within two years from each other, (2) all publications from the first criteria set of data with the combinations of XD(1) and not XD(0), and also (3) all publications whose number of coauthors(ap) that do not differ more than 20%.

Observations, conclusions, and hypotheses:

In this Table S6, the robustness of the panel model has been tested on xd only faculty with matched pairs by exploring several variations, through without/with fixed effect and without/with standardized.

- The panel model on the $XD_{\mathcal{F}}$ faculty has been implemented using only the 54 $\mathcal{F}_i$ by filtering with $\mathscr{O}(\mathcal{F}_i) = XD_{\mathcal{F}}$ who have at least 10 matched pairs of publications.
- We observe the publication count, n = 1987 by using $nrow(df)$.
- If we take a look at the significance level of the estimate, then we can interpret that the parameters are significant to all model as $p - value$ is small.
- Here, for no fixed effects (first two columns) Adjusted R-squared = 0.253 and for fixed effects (last two columns) Adjusted R-squared = 0.092.

Therefore, in all cases, the results of the observed regression estimates are not significantly different, thus indicates the robustness of the panel model.