COSC 6323 - Statistical Methods in Research
Project Phase - 2

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

April 05, 2019.

Contributions: We sit together several times to discuss about the requirement and task distribution. Fortunately, there was no task break down needed as we agreed to sit together to implement the tasks (i.e. data, figure, model and table) equally. So, the contribution status - Fig 4 (Salah, Rabin, Prity), Table S2 (Rabin, Prity, Salah) and Table S3 (Rabin, Prity, Salah). 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 this phase of the project.

Fig. 4:

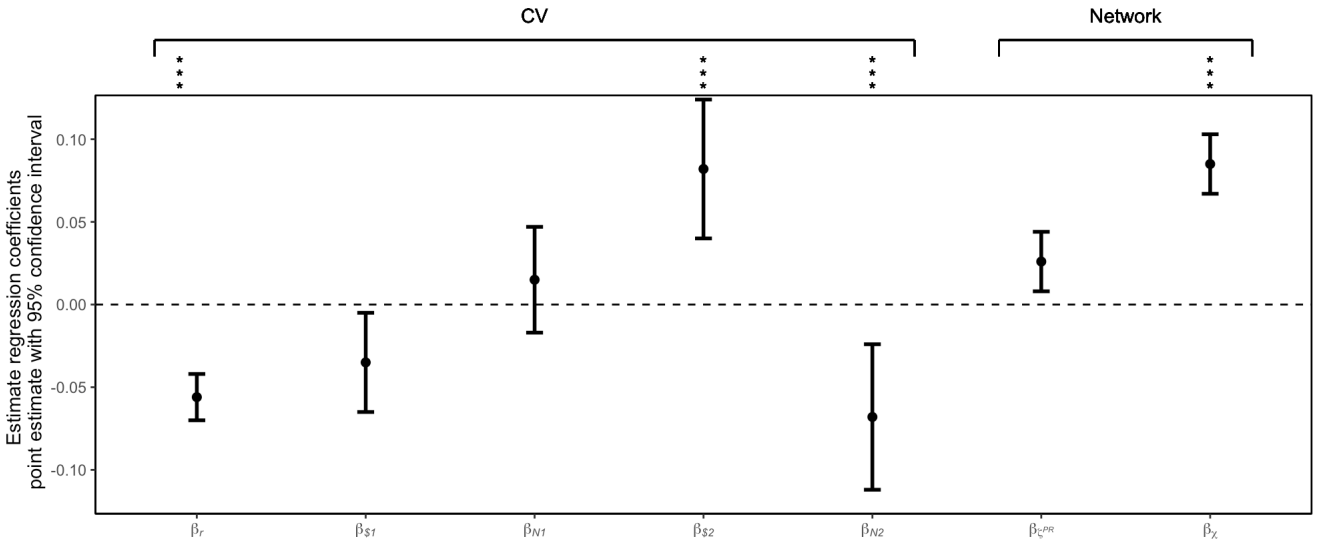


Fig. 4. Career cross-sectional regression model.

Description of figure content:

The figure shows the standardized coefficient values of (CV + Network) model. Each coefficient values has corresponding error ranges and split into two groups: researchers's CV and collaboration network. We have specified the CV coefficients and Network coefficients using the annotation. We have used the three star (***) annotation to specify the coefficients which probability of $p - value$ is as follows: $***p \leq .001$.

Observations, conclusions, and hypotheses:

From the figure, we can see that the total amount of funding ($\beta_{\$2}$) of NIH is greater than the total amount of NSF funding ($\beta_{\$1}$). The $p - value$ of NIH is less than 0.001 which indicates the NIH coefficients are significant. Although, the correlation with the number of NIH grants (β_{N2}) is negative, so the cost is related to the management of several smaller grants versus fewer bigger grants. On the other hand, the $p - values$ of NSF coefficients are greater than 0.001, so we can say that the estimates of NSF variables are not significant in the OLS model. Therefore, there is different levels of dependency on the NIH/NSF funding between the biology and computer science faculty. The main result of this plot shows that the higher degrees of cross-disciplinary activity ($\beta_\chi > 0, P < 0.001$) correlate with the higher career citations.

Table S2:

	CV		CV + Network		CV + Network [Standardized]	
CV parameters						
Departmental rank, β_r	-0.052***	(0.006)	-0.047***	(0.006)	-0.056***	(0.007)
Productivity (h-index), β_h	1.857***	(0.016)	1.866***	(0.018)	1.179***	(0.012)
Total NSF funding, $\beta_{\$1}$	-0.005*	(0.002)	-0.005*	(0.002)	-0.035*	(0.015)
# of NSF grants, β_{N1}	0.024	(0.014)	0.013	(0.015)	0.015	(0.016)
Total NIH funding, $\beta_{\$2}$	0.016***	(0.004)	0.014***	(0.003)	0.083***	(0.021)
# of NIH grants, β_{N2}	-0.067***	(0.019)	-0.061***	(0.019)	-0.069***	(0.022)
Network parameters						
PageRank Centrality, $\beta_{\mathscr{C}^{PR}}$			0.041**	(0.014)	0.026**	(0.009)
Cross-disciplinarity, β_x			0.571***	(0.061)	0.086***	(0.009)
Discipline ($\mathscr{O}$) dummy	Y		Y		Y	
5-year cohort ($y^0_{i,5}$) dummy	Y		Y		Y	
Constant	1.398***	(0.234)	1.706***	(0.271)	7.744***	(0.216)
n	4190		3900		3900	
adj. R^2	0.883		0.882		0.882	

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

Table S2. Career data set: Pooled cross-sectional model.

Description of figure content:

The following is the cross-sectional ordinary least squares (OLS) regression model:

$$lnC_i = \beta_r lnr_i + \beta_h lnh_i + \beta_{\$1} ln\$_i{}^{NSF} + \beta_{N1} lnN_i{}^{NSF} + \beta_{\$2} ln\$_i{}^{NIH} + \beta_{N2} lnN_i{}^{NIH} + \beta_{\mathscr{C}} ln\mathscr{C}_i{}^{PR} + \beta_\chi \chi_i + D(O(\mathcal{F}_i)) + D(y^0_{i,5}) + \beta_o + \epsilon$$

Here, C_i is the total citation number of $\mathcal{F}_i$, r_i is the department ranking of $\mathcal{F}_i$, h_i is the $h-index$ productivity metric, N_i^{NSF} and N_i^{NIH} are the total counts of National Science Foundation (NSF) and National Institutes of Health (NIH) grants, N_i^{NSF} and N_i^{NIH} are the total monies of $\mathcal{F}_i$ from the NSF and NIH grants deflated to constant 2010 USD, $\mathscr{C}_i{}^{PR}$ is the PageRank centrality of $\mathcal{F}_i$ within the $\mathcal{F}$ network, χ_i is the fraction of the total K_i co-authors who are cross-disciplinary. Additionally, there are two dummy variables: the XDIndicator $\mathscr{O}(\mathcal{F}_i) = BIO_{\mathcal{F}}$ or $CS_{\mathcal{F}}$ or $XD_{\mathcal{F}}$ capturing the three possible disciplinary orientations and the year of the faculty's first publication grouped into nonoverlapping 5-year intervals $y^0_{i,5}$ capturing the age cohort variation, and ϵ is the white noise.

The dependent variable is the natural logarithm of the Google Scholar citations which is considered as the quantitative measure of career achievement. The independent variables are classified into two groups: the standard CV attributes (h-index, funding, and school rank) and the network attributes (centrality and degree of cross-disciplinarity).

In table, the ***** represent the significance level, the Std Error is shown for each parameters indise **()** followed by the Estimate value, and the **Y** indicates additional fixed effects included in the regression model.

Observations, conclusions, and hypotheses:

- The logarithmic transformation (except the dummy factors) pushes the data into the direction of normality. Natural log is used to obtain the approximately normally distributed variables. Here, the β corresponds to the % change in C_i by 1% change in the independent variable.
- The first model shows the estimates using only standard CV variables. The second model shows the combined estimates of researchers's CV variables and collaboration network variables. And the third model shows the standardized beta coefficients of CV+Network model. The third model has been computed by scaling the independent variables (except the dummy factors) into 1-SD range, where the standardized beta coefficients representing the change in the dependent variable corresponding to the 1-SD shift in a given covariate.

- The main-takeaway of model (CV+Network) is that larger χ_i correlate with higher net citation impact. On the other hand, the standardized (CV+Network) model is useful for comparing the relative strength of covariates within the regression.

- The principal target is to see whether $\mathcal{F}_i$ with higher cross-disciplinary orientation (χ_i) results in higher career citation (C_i). The result of Fig 4 shows that the higher degrees of cross-disciplinary activity ($\beta_\chi > 0, P < 0.001$) correlate with the higher career citations.

- If we take a look at the significance level of estimate, then we can interpret that the department rank, $h - index$, NIH variables, and cross-disciplinarity are more significant to the model as $p - value$ is than 0.001.

- Here, for CV model, observations = 4190 and Adjusted R-squared = 0.883. And, for CV+Network model, observations = 3900 and Adjusted R-squared = 0.882; ($df\$PRCentrality > 0$). red = 0.882 ($df\$PRCentrality > 0$); for model (e), observations = 3900 and Adjusted R-squared = 0.881 ($df\$PRCentrality > 0$). The R-squared value is large for all model that indicates the data are closer to the fitted regression line for all model.

- The residuals seems to be strongly randomly scattered for all model and it is almost linear line. Additionally, all the values of VIF (computed by **car::vif(model_x)**) ensures that multicollinearity isn't available here. Thus represents the goodness of the models.

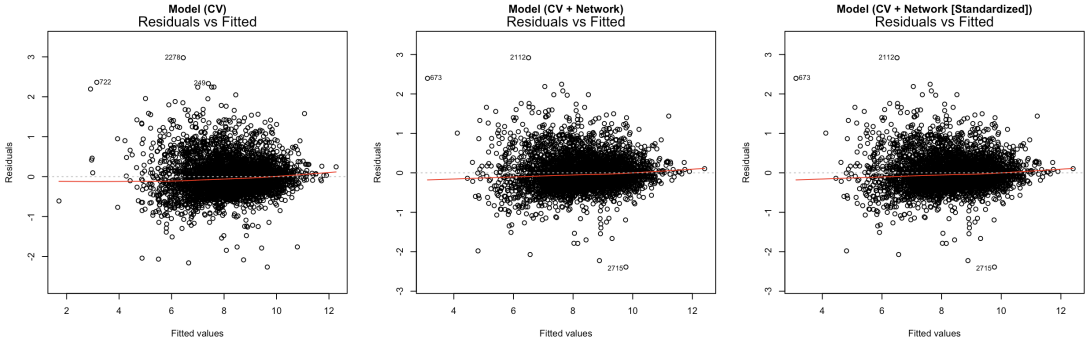


Table S3:

	(a) $\mathscr{C}^{PR}$	(b) $\mathscr{C}^{B}$	(c) $\mathscr{C}^{D}$	(d) $\cancel{\beta_{N1}, \beta_{N2}}$	(e) $\cancel{\beta_r}$
CV parameters					
Departmental rank, β_r	-0.047***	-0.042***	-0.044***	-0.046***	
	(0.006)	(0.006)	(0.006)	(0.006)	
Productivity (h-index), β_h	1.866***	1.901***	1.848***	1.862***	1.892***
	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)
Total NSF funding, $\beta_{\$1}$	-0.005*	-0.004	-0.005*	-0.003**	-0.005*
	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
# of NSF grants, β_{N1}	0.013	0.009	0.007		0.006
	(0.015)	(0.015)	(0.015)		(0.015)
Total NIH funding, $\beta_{\$2}$	0.014***	0.014***	0.014***	0.003*	0.013***
	(0.003)	(0.004)	(0.003)	(0.001)	(0.004)
# of NIH grants, β_{N2}	-0.061***	-0.065***	-0.062***		-0.059**
	(0.019)	(0.020)	(0.019)		(0.019)
Network parameters					
PageRank centrality, $\beta_{\mathscr{C}^{PR}}$	0.041**			0.042**	0.057***
	(0.014)			(0.014)	(0.014)
Betweenness centrality, $\beta_{\mathscr{C}^{B}}$		-0.000			
		(0.005)			
Degree centrality, $\beta_{\mathscr{C}^{D}}$			0.052***		
			(0.010)		
Cross-disciplinarity, β_{χ}	0.571***	0.562***	0.530***	0.579***	0.555***
	(0.061)	(0.062)	(0.061)	(0.061)	(0.061)
Discipline ($\mathscr{O}$) dummy	Y	Y	Y	Y	Y
5-year cohort ($y^0_{i,5}$) dummy	Y	Y	Y	Y	Y
Constant	1.706***	1.200***	1.344***	1.711***	1.615***
	(0.271)	(0.225)	(0.226)	(0.271)	(0.273)
n	3900	3387	3900	3900	3900
adj. R^2	0.882	0.873	0.883	0.882	0.881

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

Table S3. Career data set: Pooled cross-sectional model — robustness check.

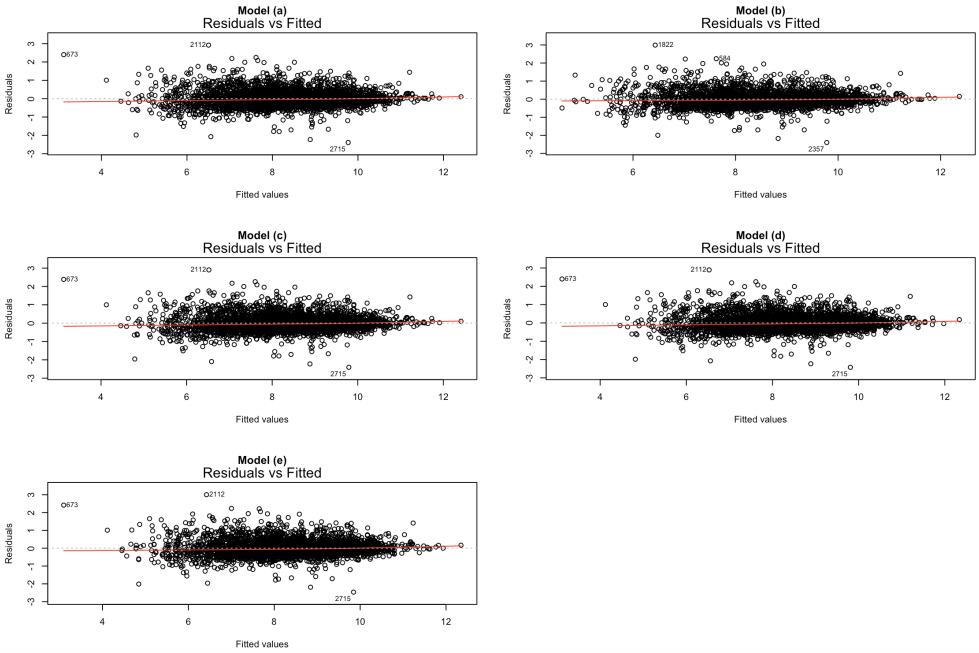
Description of figure content:

The meaning of parameters and symbols are similar like Table S2. Here, the parameter estimates for the variants of the 'CV + Network' pooled cross-sectional models reported in Table S3: (a) Model with PageRank centrality ($df\$PRCentrality > 0$), (b) Model with betweenness centrality ($df\$BetCentrality > 0$), (c) Model with degree centrality ($df\$KDirect > 0$), (d) Model without the number of grants variables ($df\$PRCentrality > 0$), and (e) Model without the departmental rank variable ($df\$PRCentrality > 0$). The model (a) of the Table S3 is same as the (CV+Network) model of the Table S2. In next two models (b) and (c), the PageRank Centrality ($\mathscr{C}^{PR}$) has been replaced respectively by the Betweenness Centrality ($\mathscr{C}^{B}$) and the Degree Centrality ($\mathscr{C}^{D}$), those Betweenness and Degree Centrality are two alternative centrality along with the PageRank Centrality. In the model (d), the variables (N^{NSF} & N^{NIH}) related to the number of grants have been removed. In the model (e), the department rank variable (β_r) has been removed.

Observations, conclusions, and hypotheses:

In this Table S3, the robustness of the cross-sectional model has been tested by exploring several variations, through model (a) to (e).

- Model (a,b,c): The Betweenness and Degree Centrality are two alternative centrality along with the PageRank Centrality. Replacing the PageRank Centrality with those two centralities respectively doesn't indicate any significant differences.
- Model (d): This model suspects the correlation effects, and so consider only the effect of total funding with removing the effect of number of grants. The result also isn't significantly different.
- Model (e): This model assumes that the most recent university affiliation could inaccurately represent the career of faculties. The estimates after removing the department rank variable doesn't indicate any significant differences too.
- If we take a look at the significance level of estimate, then we can interpret that the department rank, $h-index$, NIH variables, and cross-disciplinarity are more significant to the model as $p-value$ is than 0.001.
- Here, for model (a), observations = 3900 and Adjusted R-squared = 0.882 ($df\$PRCentrality > 0$); for model (b), observations = 3387 and Adjusted R-squared = 0.873 ($df\$BetCentrality > 0$); for model (c), observations = 3900 and Adjusted R-squared = 0.883 ($df\$KDirect > 0$); for model (d), observations = 3900 and Adjusted R-squared = 0.882 ($df\$PRCentrality > 0$); for model (e), observations = 3900 and Adjusted R-squared = 0.881 ($df\$PRCentrality > 0$). The R-squared value is large for all model (to be specific, model (b) is slightly larger than others) that indicates the data are closer to the fitted regression line for all model.
- The residuals seems to be strongly randomly scattered for all model and it is almost linear line. Additionally, all the values of VIF (computed by **car::vif(model_x)**) ensures that multicollinearity isn't available here. Thus represents the goodness of the models.



Therefore, in all cases, the results of the modified regression estimates are not significantly different with respect to the primary covariate of interest (cross-disciplinarity (β_χ)), thus indicates the robustness of the cross-sectional model.