Project Part 2: Exploratory Analysis of Collaborative Study in Genomics Revolution - Modelling Success with Faculty and Network Records

Md. Rafiul Amin and MD Tanim Hasan
University of Houston
Texas, USA

January 22, 2021

Contributions from Each Member

Md. Rafiul Amin

1) Pooled cross-sectional model - robustness check.
2) Regression table generation.
3) Critical thinking over all the generated plots and models in the report.
4) About 50% of the report writing.

MD Tanim Hasan

1) Linear modeling of career data set pooled cross-sectional model.
2) The plot of the cross-sectional regression model.
3) Critical thinking over all the generated plots and models in the report.
4) About 50% of the report writing.

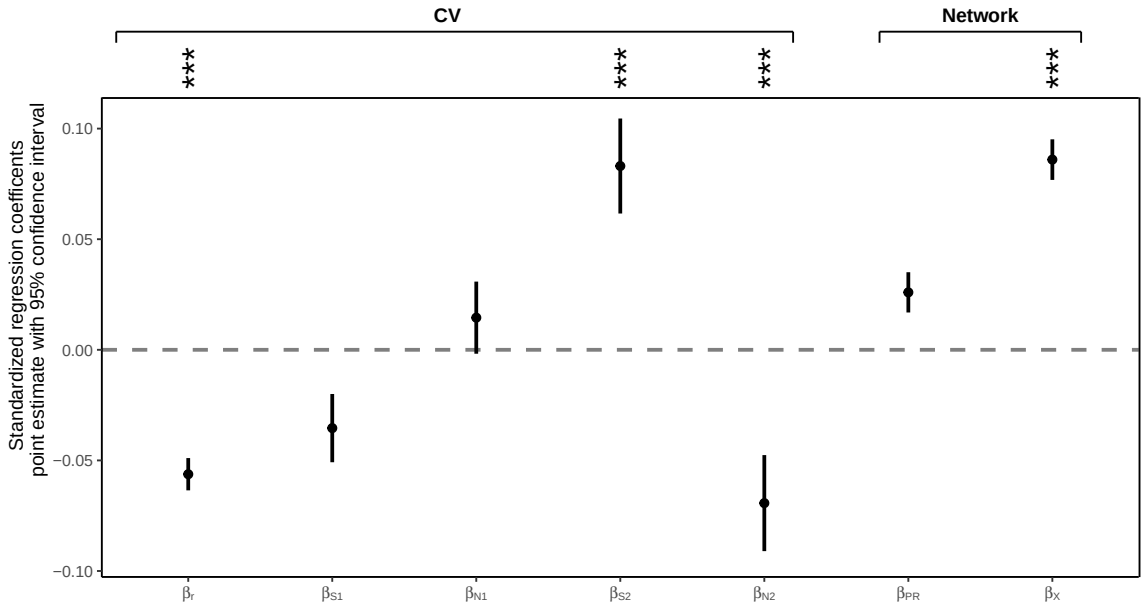

Fig. 4. Career cross-sectional regression model. This figure shows the coefficients of the independent variables for the linear model in Eq. 1, which is the standardized model of CV+Network. The coefficients for the relevant covariates split into two categories CV or Network. To facilitate comparison of the relative strength of the parameter estimates, the standardized beta coefficients are shown, representing the change in the dependent variable $\ln C_i$ that corresponds to a 1-SD shift in a given covariate. See Table S2 for the complete list of parameter estimates. The levels of statistical significance are as follows: $^{***}p \leq 0.001$.

Figure 4 shows that among all the CV parameters, department ranking (r_i), total amount of NIH funding ($\$_i^{NIH}$), and the number of NIH grants (N_i^{NIH}) are very significant with $p \leq 0.001$. The small variance in β_r shows that the slope of the linear relationship between the log of citation and log of citation is very consistent among all the faculty information we have. This is because of the fact that either the department ranking r_i is sometimes dependent on faculties success fullness in the field and every other researcher tend to follow the research directions of faculties from high-rank departments.

Table S2. Career data set: Pooled cross-sectional model: Table 2 depicts the linear model of faculties for CV and CV+ Network attributes. The dependent variable is career achievement, measured as a logarithm of Google Scholar citations, lnC_i as of 2017. The regression model is specified in Eq. (1) and estimated using standard OLS; there are 4,190 $\mathscr{F}_i$ (observations) for the pure CV model and 3,900 observations for the other two models that include network attributes, as in these cases, we exclude from consideration disconnected $\mathscr{F}_i$nodes. Natural logs were used to obtain variables that are approximately normally distributed. Thus, when the independent variable enters in ln, then β corresponds to the change in $\mathscr{C}_i$ following a 1% change in the independent variable; in the case of the cross-disciplinarity fraction, β_χ represents the % change in C_i following a 0.01 shift increase in χ_i. The first column cluster shows the estimates using only standard CV variables. The combined CV + Network model demonstrates that $\mathscr{F}_i$with larger χ correlate with higher net citation impact. For the combined model we also report the standardized beta coefficients – useful for comparing the relative strength of covariates within the regression. Standard errors were calculated using the clustered sandwich estimator, clustering on $\mathscr{F}_i$ age-cohort $y^0_{i,5}$ (based on 14 nonoverlapping 5-year career birth year groups, e.g., 1940-1944, 1945-1950, etc.) to account for within-age-cohort correlation. Y indicates additional fixed effects included in the regression model.

	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^{***}	(0.011)
Total NSF funding, $\beta_{\$1}$	-0.005	(0.002)	-0.005^{*}	(0.002)	-0.035^{*}	(0.015)
# of NSF grants, β_{N1}	0.024	(0.015)	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-disciplinary, β_χ			-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)	7.743^{***}	(0.216)
n	$4,190$		$3,900$		$3,900$	
adj. R^2	0.883		0.882		0.882	

Standard errors in parentheses,

$^{*}\ p \leq 0.05$, $^{**}\ p \leq 0.01$, $^{***}\ p \leq 0.001$

The estimated CV parameters $\beta_{\$1}$ and β_{N1} is not very significant only based on the CV/CV-Network/CV-Network [Standardized] data. That indicates that faculty success is not dependent on the total amount of NSF funding or the total number of NSF grants. While the slope related to the total amount of NIH funding and the total number of NIH grants. This might be an indication that the total amount of NIH funding information is patronizing the research towards the right direction and faculties tend to be more successful who have more NIH funding, not NSF funding. In other words, the consistency with NIH funding information is more evident in terms of the citation.

The inclusion of the parameter PageRank $\mathscr{C}^{PR}$ and the Cross-disciplinary fractional χ shows that their effect in faculty's success is very evident. The evidence of the effect of $\mathscr{C}^{PR}$ means if more connected they tend to be more successful. Same is true of χ, i.e., if the faculties have more cross-disciplinary collaboration, they tend to be more successful.

From the fitted linear models, there is also an effect of faculty affiliation and year of first publication in their number of citation.

Table S3. Career data set: Pooled cross-sectional model - robustness check: Table S3 summarizes the robustness of the cross-sectional model with CV and Network parameters. Here we have five models: (a) Model with PageRank centrality. (b) Model with betweenness centrality. (c) Model with degree centrality; (d) Model without the number of grants variables NSF & NIH. (e) Model without the departmental rank variable. Results are not significantly different with respect to the primary covariate of interest, that is, cross-disciplinarity (β_χ).

	(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	0.008	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)	(0.019)	(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-disciplinary, β_χ	0.571^{***}	0.562^{***}	0.530^{***}	0.579^{***}	0.555^{***}
	(0.061)	(0.054)	(0.061)	(0.061)	(0.062)
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,

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

The obtained results from three different selected centrality show that other than the centrality parameters ($\beta_{\mathscr{C}^{PR}}$, $\beta_{\mathscr{C}^{B}}$ and $\beta_{\mathscr{C}^{D}}$), the change in the other parameters are not significant. However, there just one difference in betweenness centrality $\beta_{\mathscr{C}^{B}}$ case. The effect of betweenness centrality is not that evident in the linear model where other centralities have significant contribution in the corresponding linear regression models.

Removal of the number of NSF and NIH grants also does not show any significant difference in all other parameters. Moreover, removal of the department rank r_i also does not make a significant change in other parameters.

In all cases, the results of the modified regression estimates are not significantly different. Furthermore, R^2's hardly changes in the third digit after the decimal point. These indicate the robustness of specification of the linear regression with respect to these adjustments in this table. This is an indication of having similar information sharing in different variables.

Possible Reason of Slight Discrepancy with the Parameters Obtained by Petersen *et al.* [1]

Most of the data in the given data set is exponentially distributed. Hence, the linear model is obtained based on log transformation of all of the exponentially distributed variables. Thus the linear model (general setting) is as follows,

$$\begin{aligned}\ln C_i = \beta_r \ln r_i + \beta_h \ln h_i + \beta_{\$1} \ln \$_i^{NSF} + \beta_{N1} \ln N_i^{NSF} + \beta_{\$2} \ln \$_i^{NIH} + \beta_{N2} \ln N_i^{NIH} \\ + \beta_{\mathscr{C}} \ln \mathscr{C}_i^X + \beta_\chi \ln \chi_i + D(O(\mathscr{F}_i)) + D(O(y_{i,5}^0)) + \beta_0 + \varepsilon \end{aligned} \quad (1)$$

where $X \in \{PR, B, D\}$ denotes different centralities. However, to obtain log transformation for some of the variables we need to avoid taking log transformation zero. To avoid that, we add a small value with all the instances of that particular variable that has any zero elements. Now, the question come is how to choose that small value. If we take a value very close to zero then after log transformation that particular instance will seem like an outlier to the linear model. And the least square regression model will have a very poor performance in estimating different parameters in the linear model. We added minimum possible value other than zero to all the elements of the variables where there is an element with zero. For example, for h_i, we added 1 with all the elements before log transformation as there are zero elements. On the contrary, we did not add any value to the elements of the variables that do not have any zero value.

Now, the choice of such kind of scheme before log transformation may lead to slightly different estimations. The result obtained in [1] might have a different scheme for the log transformation and thus have slightly different values of the parameters. Regardless, the conclusion is drawn from this report and the ones in [1] is the same.

References

[1] Alexander M Petersen, Dinesh Majeti, Kyeongan Kwon, Mohammed E Ahmed, and Ioannis Pavlidis. Cross-disciplinary evolution of the genomics revolution. *Science advances*, 4(8):eaat4211, 2018.