Panel Data and Multilevel Analyses of Academic Publishing Success

Richard Williams, University of Notre Dame

Lutz Bornmann, Max Plank Society

Andreas Thor, University of Applied Sciences

Last updated July 12, 2018

Overview

- What factors contribute to academic publishing success? What causes some works to be more widely cited and influential than others?
- The quality of the research and the researcher are obvious factors. But those interested in the Sociology of Science and Bibliometrics raise several other questions and possibilities.

- Do scientists do their most successful and important work early in their careers, when their ideas are fresh and innovative? Or do they do it later when they are more established and experienced?
- How important is the outlet for the work? What role does journal prestige play? Does co-authorship increase success or diminish it?
- Does success breed success? As cumulative advantage theory would suggest, does early success lead to later success, perhaps because those who are successful early on receive greater resources later?

- Do determinants of success differ by characteristics of the individual? For example, do women benefit less from having papers in prestigious journals than do men?
- Are the determinants of publishing success different in different academic fields? Is the effect of co-authorship in a field like Sociology (where papers tend to have only a small number of co-authors) different than it is in other fields where co-authorship is far more common?

- We will examine these issues using a unique 30+ year longitudinal dataset of nearly 400,000 publications and over 13,000 of their authors. We used three data sources:
  - ResearcherID: The web site http://www.researcherid.com provides a solution to the author ambiguity problem within the scholarly research community. Each member is assigned a unique identifier to enable researchers to manage their publication lists and to avoid author misidentification.
  - We used a gender name dictionary (see https://ideas.repec.org/c/wip/eccode/10.html) to determine the likely gender of authors.
  - The bibliometric data are from an in-house database developed and maintained by the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics, formerly the IP & Science business of Thomson Reuters (Philadelphia, Pennsylvania, USA).

- Panel data and multilevel modeling will be employed. We will discuss
  - Panel Data/ Generalized Linear Mixed Models (especially logit)
  - Fixed Effects Models
  - Random Effects/ Random Intercepts Models
  - Hybrid Models
  - Random Slopes Models

- NOTE: All analyses are very preliminary!!! We are in the early stages of this work, and know that several analysis issues still need to be resolved. But this should give a good idea of how these models can be used and where our work will be going in the future.

- Among other things, we will show how these techniques:
  - Can sometimes allow us to estimate and/or control for the effects of characteristics of authors, even when those characteristics are not directly measured
  - Allow us to examine how effects of variables may randomly differ across individuals, e.g. the importance of career stage for success (or even the direction of its effect) may not be the same for everyone

The Data

- Data were obtained for several hundred thousand publications for the years 1980-2016.
- Measures included:
  - The number of times the piece was cited. From this, a percentile ranking for how often the piece was cited, standardized for field and publication year, was created. The higher the percentile ranking, the higher the number of citations within the field and publication year. Papers less than three years old do not have reliable percentile rankings yet and are excluded from the analysis.
    - From this we computed various dependent variables. In particular, we were able to determine each author's most cited piece, as well as which papers ranked in the top quartile of papers cited.
  - The number of authors, the page length, and the number of references cited for each paper
  - The prestige of the journal, as measured by the Journal Impact Factor standardized for field of study and publication year

- This information was then merged with information on over 13,000 authors. Information that was available or computed included
  - The author's gender (computed using another database that linked names with likely gender)
  - The author's career stage at the time the paper was published, i.e. how many years had s/he been publishing when each paper appeared?
  - The nationality of the author. For now, we treat that as a dichotomy, USA versus not USA
  - The author's self-reported fields of study. These include natural science, engineering, medical, agriculture, social science, and the humanities. Authors can select more than one area.

Variables

- Dependent Variables
  - Highest ranked paper. Coded 1 for one paper per author, 0 for all other papers
  - Highly ranked (i.e. top quartile) papers. Coded 1 for highly ranked, 0 otherwise. There is no limit to how many or how few highly ranked papers an author can have.
  - The above variables are computing by using percentile rankings of citations. A highly ranked paper in one field may need many more citations than a highly ranked paper in another field. Standardized percentiles take care of such interdisciplinary differences.

- Independent Variables Measuring Paper Characteristics
  - Nauthors – Number of authors on the paper. This varies widely, ranging from one author to over 3,000! For now we have deleted papers with more than 25 co-authors, which still leaves about 99% of the papers.
  - Npages – The length of the paper, in pages
  - Nrefs – Number of references cited
  - Jifperc – The Journal Impact Factor percentile ranking for the publishing source, standardized by field and publication year

- Independent Variables Measuring Author Characteristics
  - Careerstage – The estimated number of years since the author began publishing when the current paper was published
  - Female – Coded 1 if the author is female, 0 otherwise
  - USA – coded 1 if the author is from the USA, 0 otherwise.
  - SocialScience – Coded 1 if the author lists the social sciences as a field, 0 otherwise. Information is available for several other fields but for now we are keeping it simple.

- For author characteristics, only careerstage varies across time (at least with these data, since we only have author data from one point in time. Information on nationality and field of study is from when the researcher included that information in the ResearcherID database. Thus, we do not know how up-to-date the information is or how the values of these variables may have changed across time.)
  - NOTE: Variables like gender and race have typically been treated as being invariant across time. This is not always true (at least with self-reports) and analyses will eventually have to adapt to this.

Generalized Linear Mixed Models (GLMMs).

- All models estimated will be Generalized Linear Mixed Models (GLMMs). These models also go by many other names, e.g. multilevel models, hierarchical models.
- Many of the equations and the descriptions of the model are adapted (sometimes verbatim) from
  - Schunk and Perales, Stata Journal 2017
  - Allison, Sage, 2009
- We will present the general model and then discuss special cases of it as we use them

- We often have data clustered within groups. We sometimes call these multilevel data or hierarchical data. For example,
  - We can have a sample of schools (level 2), and within each school we can have a sample of children (level 1).
  - We can have a sample of individuals (level 2) and for each individual we can have data collected annually (level 1). This is also called panel or longitudinal data.
  - At a minimum, we have to make sure the standard errors are correct in our analyses. Records are not independent of each other. Data on 100 students from each of 100 schools should not be treated the same as a sample of 10,000 students drawn at random from all schools.
  - BUT, GLMMs allow us to do much more than that.

Schunk & Perales Description of GLMMs (p. 91)

GLMs can be extended to include random effects and are thus suited for analyzing clustered data, such as multilevel and panel data. These models are known as generalized linear mixed models (GLMM). Consider a situation where we have data with two hierarchical levels. Let i denote level two (for example, schools) and j denote level one (for example, students). y_{ij} is the response (dependent) variable, x_{ij} is a level-one variable that varies within and between clusters, c_i is a level-two variable that varies only between clusters, and u_i is the random intercept. A GLMM is specified as

$$g\{E(y_{ij}|x_{ij}, c_i, u_i)\} = g(\mu_{ij}) = \beta x_{ij} + \gamma c_i + u_i \quad (1)$$

The "mixing" outlined above becomes obvious: this model "mixes" a fixed part (the fixed coefficients β and γ) and a random part (the random intercept u_i). To relax the

- g(.) is the so called link function. The dependent variable is some sort of function of E(y)
  - For linear models, it is often called the identity link. E(y) (the expected value of y) is estimated by the model.
  - For logistic regression it is the logit link. The dependent variable is actually the log odds of success.
- x_{ij} and y_{ij} can have different values across individuals and across clusters e.g. student grades and family income.
- c_i only differs across clusters, e.g. schools can be public or private, but within a school all students are attending either a private school or a public one.

- In our data, we have papers (level 1) written by authors (level 2).
- Things like the success a paper has and the prestige of the journal that published it can vary across papers and across publications. These are y_{ij} and x_{ij} variables in the model.
- However, things like gender and nationality will differ across authors, but will always be the same for all papers written by that author. These are examples of c_i papers.

- The error term ui may reflect level 2 (cluster or group) variables not included in the model. *It is assumed to be uncorrelated with the variables that are in the model.* If this assumption is violated, there will be omitted variable bias and coefficients will be biased.
  - Note: For their purposes, Schunk and Perales assume that only level 2 variables are omitted and that there is no omitted variable bias at level 1, and hence do not include an ε_{ij} term in their models. For example, a model might not include the income of individuals at each time period, which may bias estimates. Hybrid and FE models cannot control for these types of omitted variables.

- BUT, suppose the omitted variables always have the same values for all cases within a cluster/group. For example, for annual data collected on an individual, the gender of the individual, or the year the respondent was born, will be the same at each time the data was collected. With panel data, these are often called time-invariant variables. For other types of data they might be called group-invariant or cluster-invariant.
- In such cases, when we have multiple records for cases, subjects can sometimes serve as their own controls. The idea is that, whatever effect an omitted variable has on one group record, it will have the same effect on all the other group records.

- *Fixed effects models* make this possible. When they are appropriate, FE models control for the effects of omitted variables, and make the coefficients for the variables that are in the model unbiased.
- The course notes provide more technical details on how exactly this is done. An even better and more thorough explanation appears in Allison's 2009 book.
- The next slide shows how a fixed effects model can be estimated in Stata using the xtlogit command with the fe option.

```
. xtlogit paprbest  nauthors npages nrefs jifperc careerstage i.female i.usa
i.socialscience, nolog fe
note: 799 groups (799 obs) dropped because of all positive or
      all negative outcomes.
note: 1.female omitted because of no within-group variance.
note: 1.usa omitted because of no within-group variance.
note: 1.socialscience omitted because of no within-group variance.

Conditional fixed-effects logistic regression   Number of obs     =    373,535
Group variable: id                              Number of groups  =     13,330

                                                Obs per group:
                                                              min =          2
                                                              avg =       28.0
                                                              max =        683

                                                LR chi2(5)        =    7158.64
Log likelihood  = -33211.053                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     paprbest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     nauthors |   .1020616   .0030442    33.53   0.000      .0960951     .108028
       npages |   .0046791   .0012776     3.66   0.000       .002175    .0071832
        nrefs |   .0049528    .000256    19.35   0.000      .0044511    .0054546
      jifperc |   .0386638   .0006544    59.09   0.000      .0373812    .0399463
  careerstage |  -.0576846   .0028195   -20.46   0.000     -.0632107   -.0521586
              |
       female |
      Female  |          0  (omitted)
              |
          usa |
          Yes |          0  (omitted)
              |
socialscience |
          Yes |          0  (omitted)
------------------------------------------------------------------------------
```

- Several results are obvious and fairly easy to interpret.
  - The more authors a paper has, the more pages it has, and the more references it cites, the more likely it is to be the author's most cited.
  - Placement really does matter. The prestige of the publishing outlet has the most statistically significant effect on an author's paper being his or hers most cited.
  - Careerstage has probably the most interesting effect. The negative coefficient suggests that authors have their greatest success earlier in their careers rather than later. (However we suspect that forthcoming analyses will reveal that the relationship between career stage and publishing success is more complicated than that.)

- Of course, we are also interested in how variables like gender, nationality, and field of study are related to publishing success. We therefore included these variables in the model. And the results show us…
- Absolutely nothing! All three variables are omitted. Why?
- As previously noted, FE models can control for time-invariant variables that are not measured and/or not included in the model.
- Unfortunately, the same process that makes this possible also makes it impossible to estimate the effects of these variables, even if we do measure them. That is, their effects are controlled for but not actually estimated.

- Luckily, with an FE model it is still possible to estimate the effects of interactions between time-invariant variables and other variables in the model.
- In Stata, I did selected interactions with the command

```
xtlogit paprbest  nauthors npages nrefs jifperc ///
    careerstage i.female i.usa i.socialscience ///
    (i.female i.usa i.socialscience)# ///
    (c.nauthors c.jifperc c.careerstage ) ///
    , nolog fe
```

Highest Ranked paper – Fixed Effects Model with Interactions

Variable	Main effects	Female Intr	USA intr	SocSci intr
# authors	0.1020***	-0.0109	0.0103	0.0107
# pages	0.00471***			
# refs	0.00494***			
JIF percentile	0.0400***	0.0007	0.0001	-0.0096***
Career Stage	-0.0653***	0.0208**	0.0200*	0.0109

- The interaction effects suggest that women, and those from the United States, tend to have their most successful paper later in their careers than do others (although they still tend to have them earlier in their careers, just not as early as others do).
- For Social Scientists, the success of their most cited piece is somewhat less dependent on the journal's prestige than it is for those in other fields.

Critique of the last analysis and of fe models in general

- A key concern with the highest ranked variable is that it is a moving target: the highest ranked paper at the time the data were collected might not be the highest ranked paper later. Some paper not yet even written may eventually be the author's most successful, or some existing paper may become more prominent with time. The current strategy would be better if we only had scholars who had finished their publishing career, but that is not the case.
- In our subsequent analyses the dependent variable will indicate whether the paper was "highly ranked," i.e. was in the top quartile of papers cited within a field and publication year.
- In theory at least, one researcher could have no papers in the top quartile, while another could have all of his/her papers there.

- Fixed effects models have also attracted criticisms and concerns.
  - As noted, they can help avoid omitted variable bias, by controlling for time-invariant variables that may not have even be measured.
  - BUT, the tradeoff is that, while these variables can be controlled for, their effects cannot be estimated.
  - As Schunk and Perales note, in multilevel analysis this is often a major concern, because (p. 94) "the interest often lies in these effects, for example, how the characteristics of neighborhoods, schools, workplaces, or geographical areas influence individuals' outcomes."

- Another example: suppose your dissertation examined the effects of gender on earnings, and the model you were using did not allow you to estimate the effects of gender!
- Schunk and Perales add that "Because the fixed-effects approach discards all contextual (level-two) information, some argue that it is generally less preferable than the random-effects approach for multilevel analysis." [Emphasis added.]

- Put another way, some researchers would prefer to put up with some omitted-variable bias if, in exchange, they could examine the effects of critical variables they were especially interested in.
- Besides, there are many types of omitted variable bias that FE models cannot deal with anyway. Allison's 2009 book and my course notes elaborate further.
- We will therefore consider several other types of GLMMs, beginning with a basic random effects model. In RE models, the questionable assumption is made that omitted variables are NOT correlated with the variables in the model. However, they also allow for the constant terms to vary across clusters, which a regular logistic regression would not do.

```
Random-effects logistic regression              Number of obs     =    374,334
Group variable: id                              Number of groups  =     14,129

Wald chi2(8)      =   36801.57
Log likelihood  = -209470.86                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        topq |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nauthors |   .0729849   .0013875    52.60   0.000     .0702656    .0757043
       npages |    .014749   .0009566    15.42   0.000      .012874     .016624
        nrefs |    .006277   .0001762    35.63   0.000     .0059316    .0066223
      jifperc |   .0399781   .0002369   168.75   0.000     .0395138    .0404425
  careerstage |  -.0394571   .0008703   -45.34   0.000    -.0411628   -.0377514
              |
       female |
       Female |  -.2309206   .0204889   -11.27   0.000     -.271078   -.1907632
              |
          usa |
          Yes |   .2643758   .0267614     9.88   0.000     .2119244    .3168272
              |
socialscience |
          Yes |   .1206236   .0276417     4.36   0.000     .0664469    .1748004
        _cons |  -3.993144   .0237151  -168.38   0.000    -4.039625   -3.946663
-------------+----------------------------------------------------------------
     /lnsig2u |  -.6285584   .0224981                     -.6726538    -.584463
-------------+----------------------------------------------------------------
      sigma_u |   .7303151   .0082153                      .7143895    .7465957
          rho |   .1395052   .0027007                      .1342955    .1448832
------------------------------------------------------------------------------
LR test of rho=0: chibar2(01) = 1.6e+04                  Prob >= chibar2 = 0.000
```

- We see similar patterns as before. Once again, the most highly cited papers researchers have tend to be published earlier in their careers, rather than later.
- Also, we see that women are less likely to have their papers be highly cited than do men. Conversely, US scholars, and those in the social sciences, are more likely to have highly cited papers.
- Interaction effects (not shown) can also be estimated. None of the gender interactions are significant. The effect of career stage is less negative for US scholars than it is for others (meaning US scholars are somewhat more likely to have highly rated papers later in their careers). Conversely, the success of papers by Social Scientists is somewhat less dependent on the prestige of the journal.
- The test statistic at the end of the output indicates that it would be a major mistake to ignore the way the data are clustered, e.g. using regular logistic regression would probably result in (even more) biased parameter estimates and incorrect standard errors.

- We next consider a hybrid model, proposed by Allison (2009) and others. A hybrid model combines some of the best features of FE and RE models
- A hybrid model makes it possible to get unbiased estimates for some variables (indeed, estimates are nearly identical to estimates from an FE model) while at the same time being able to estimate effects for time-invariant or group-invariant variables like gender.
- Conceptually, the procedure is

- Within each group, calculate the mean for each independent time-varying variable. The means will represent the *between-group differences* (i.e. group means will differ between clusters but not within them).
- Then, again within each group, subtract the mean for the group from each variable. These deviations from the group mean will represent the *within-group variability*.
- Estimate an RE (not FE) model that includes both the means of the variables and the difference-from-the-means variables.
- Unlike a regular FE model, you can also include time-invariant variables like gender and estimate their effects.

- Schunk and Perales write the model as

$$g(\mu_{ij}) = \beta_W(x_{ij} - \overline{x}_i) + \beta_B\overline{x}_i + \gamma c_i + u_i$$

- The following Stata code shows one way this can be done. When computing group means, it is important to use only the cases that are included in the model, e.g. if listwise deletion causes some records to be dropped from the analysis, they should also be dropped when computing group means.

```
*** Hybrid Model
gen mysample = !missing(topq, nauthors, npages, nrefs, jifperc, ///
        careerstage, female, usa, socialscience, id)

foreach var of varlist nauthors npages nrefs jifperc careerstage {
        egen m`var' = mean(`var') if mysample, by (id)
}

foreach var of varlist nauthors npages nrefs jifperc careerstage {
        gen d`var' = `var' - m`var' if mysample
}

xtlogit topq dnauthors-dcareerstage mnauthors-mcareerstage i.female i.usa i.socialscience , nolog re
```

- However, life is much simpler if you use Perales and Schunk's xthybrid command, available from SSC, which automates the whole process.

```
. xthybrid topq female usa socialscience, use(nauthors npages nrefs jifperc careerstage) ///
>         family(binomial) link(logit) clusterid(id) star

Hybrid model. Family: binomial. Link: logit.

+-------------------------------------+
|             Variable |     model    |
|----------------------+--------------|
| topq                 |              |
|            R__female |    -0.1694*** |
|     R__socialscience |     0.0924*** |
|               R__usa |     0.1860*** |
|          W__nauthors |     0.0824*** |
|            W__npages |     0.0106*** |
|             W__nrefs |     0.0068*** |
|           W__jifperc |     0.0388*** |
|       W__careerstage |    -0.0534*** |
|          B__nauthors |     0.0207*** |
|            B__npages |     0.0428*** |
|             B__nrefs |     0.0071*** |
|           B__jifperc |     0.0518*** |
|       B__careerstage |    -0.0063*** |
|                _cons |    -5.1434*** |
|----------------------+--------------|
|       var(_cons[id])|              |
|                _cons |     0.4649*** |
|----------------------+--------------|
| Statistics           |              |
|                  ll | -2.088e+05    |
|                chi2 | 37783.1969    |
|                   p |     0.0000    |
|                 aic |  4.177e+05    |
|                 bic |  4.178e+05    |
+-------------------------------------+
   legend: * p<.05; ** p<.01; *** p<.001
Level 1: 374334 units. Level 2: 14129 units.
```

- You are primarily interested in the variables that start with R_ (the coefficients for the time-invariant variables) and those that start with W_ (which show the effects of within-group variability).
- If the assumptions of the random effects model are true, the coefficients for the B_ variables (between-group) should equal the coefficients for the corresponding W_ variables. xthybrid has a test option that lets you test whether or not the assumptions hold.
- If you don't like the way the results are displayed, xthybrid has options for changing their appearance.

- The xthybrid command has some limitations that might sometimes make you prefer to compute all the necessary variables yourself.
  - Factor variable notation (e,g. i.gender) is not supported. You need to create any dummy variables yourself.
  - Temporary variables are created but then deleted. As a result, some post-estimation commands (e.g. predict) will not work.

More importantly, hybrid models themselves have some limitations.

- You may not be able to estimate marginal effects correctly with them (however, estimating marginal effects after any FE model can be problematic)
- Schunk (Stata Journal, 2013) notes various other limitations, e.g. including interaction terms can be cumbersome.
- Nevertheless, Schunk concludes "[hybrid] models are useful extensions to the standard random-effects and fixed-effects approaches."

- Finally, we consider a random slopes model.
  - In previous models, the intercept terms could differ across groups.
  - In random slopes models, the slopes of selected variables can differ across groups too.
  - Schunk and Perales write the model as

$$g(\mu_{ij}) = (\beta + u_{i2})x_{ij} + \gamma c_i + u_{i1}$$

- Each group has a value for u_{i2}. These reflect how the effect of xij differs for that group, e.g. the effect of x_{ij} might be stronger in that group, or it may be weaker.
- We next show a random slopes model where the effect of careerstage is free to vary across clusters. (These models can take very long to estimate, so we drew a random subsample of 10% of all the authors.)

```
. *** 5. Advanced random effects - 10% sample
. melogit topq  nauthors npages nrefs jifperc careerstage i.female i.usa i.socialscience
if sample10  || id: careerstage, nolog

Mixed-effects logistic regression               Number of obs     =     37,391
Group variable:              id                 Number of groups  =      1,413
------------------------------------------------------------------------------
            topq |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
        nauthors |   .0707621   .0044313    15.97   0.000     .0620768    .0794474
          npages |   .0222111   .0031477     7.06   0.000     .0160418    .0283804
           nrefs |   .0053515   .0005671     9.44   0.000     .0042399     .006463
         jifperc |    .039661   .0007399    53.60   0.000     .0382108    .0411112
     careerstage |  -.0413714    .003255   -12.71   0.000     -.047751   -.0349918
                 |
          female |
          Female |  -.1289583   .0633318    -2.04   0.042    -.2530862   -.0048303
                 |
             usa |
             Yes |   .3708106   .0844497     4.39   0.000     .2052922    .5363291
                 |
   socialscience |
             Yes |   .2172029    .083411     2.60   0.009     .0537203    .3806855
           _cons |   -4.02058   .0746733   -53.84   0.000    -4.166937   -3.874223
-----------------+------------------------------------------------------------
id               |
 var(careerstage)|   .0008324   .0001953                      .0005255    .0013185
       var(_cons)|   .4294519   .0362121                      .3640321    .5066281
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(2) = 1601.45             Prob > chi2 = 0.0000
```

- The main thing that is new is var(careerstage). This shows us how much the effect of careerstage varies across clusters. The estat sd command shows the corresponding standard deviations.
- It is also possible, and informative, to actually estimate the random effects for each group, i.e. the u_{i2} values. This can be done using the predict command.

```
. estat sd
------------------------------------------------------------------------------
           topq |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
id              |
 sd(careerstage)|   .0288514   .0033854                      .0229238    .0363117
       sd(_cons)|   .6553258   .0276291                      .6033507    .7117781
------------------------------------------------------------------------------


. predict re1 re2 if e(sample), reffects
(calculating posterior means of random effects)
(using 7 quadrature points)
(336943 missing values generated)

. sum re1 re2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         re1 |     37,391    .0010369    .0145811   -.041775   .0541951
         re2 |     37,391    .1278538    .5251565  -1.351337   2.557217
```

- re1 gives the estimated random effects for the slope coefficients. Note that the estimated effects range from -.0418 to .0542.
- The beta coefficient for careerstage was -.0413. So, when you add the random effect to the Beta coefficient, the estimated effect of careerstage for each cluster ranges anywhere from -.0831 to .0129.
- That is, the effect of careerstage is much more negative for some authors, while for others it is actually slightly positive!
- While these results may be valid, such extreme variations raise concerns that the effects of careerstage are not being modeled correctly.

Directions for future work

- Though flawed, these early analyses indicate the types of additional work that may be worth pursuing.
- The effect of careerstage needs to be better modeled. Our own random slopes model suggests that the effect of careerstage differs greatly across authors. Some things to consider:
  - Quadratic terms or spline functions might be added to the model. (Such transformations may be appropriate for other variables as well.)
  - Latent growth curve models show how trajectories vary across time, e.g. some children may develop more quickly than do others. We might be able to differentiate between those who peak early in their career and those who do so later.

- Number of authors also has to be modeled better.
  - I have never published a paper with more than 3 or 4 authors.
  - But in our sample, some papers had over 3,000 authors. Co-authorship practices differ radically across fields, and we need to determine how to handle that.
  - We've considered using a variable coded 1 if single-authored, 0 if co-authored. While this might work well for a field like Sociology, it would not work well for fields where single-authorship is far less common.
  - Co-authorship also makes it more difficult to examine the effects of careerstage. The most highly cited papers for some people may have been co-authored in graduate school when they worked with a distinguished senior scholar. The more co-authors there are, the more difficult it is to assess the contributions of any one of them.

- Cumulative advantage theory needs to be better modeled. Does early success lead to later success? To model this we might add a variable like careerstage, where the success of earlier papers is included as a variable in models of subsequent success.
- Differences by field need to be considered more closely. Does it even make sense to try to develop a model which applies to all fields, or would it be better to develop models for each field separately?

- While there is still much work to be done, we are confident that more refined theory and Generalized Linear Mixed Models will lead to our eventual success.

- For more information on the teaching and research of the authors, see

https://www3.nd.edu/~rwilliam/

https://www.researchgate.net/profile/Lutz_Bornmann

https://dbs.uni-leipzig.de/en/person/andreas_thor