Criticism of the Paper:'The association between early career informal mentorship in academic collaborations and junior author performance'

Yiling Song

December 22, 2020

Criticism of the Paper:'The association between early career informal mentorship in academic collaborations and junior author performance'

Yiling Song

December 22, 2020

Link to Github: https://github.com/yilingsong1008/Critique-of-a-Paper

Data supporting this dataset available at: https://github.com/bedoor/Mentorship (Mentorship/Repository_Data/Data_7yearcutoff.csv)

Abstract

This report is based on the following paper: 'The association between early career informal mentorship in academic collaborations and junior author performance'[1]. The paper focuses on mentoring relationships in scientific collaborations, and by investigating a random sample, concludes that the quality of the mentoring relationship predicts the impact of scientificity in papers written by protégé without a mentor, with respect to the impact of gender regarding mentors on women pursuing scientific careers. On these premises, I use the data provided by the authors of the paper to go further in the analysis and to identify possible problems.

Keywords

mentor–protégé pairs; female protégés; statistical analysis; mentors; informal mentorship

Introduction

Mentorship helps individuals develop their careers[2-4], and it has become commonplace for senior colleagues to mentor their juniors, but they are not necessarily their formal supervisors[5-6]. In the context of academic collaboration, the role of mentorship in supporting early career scientists is widely recognized[7]. Therefore the definition of mentorship is worth studying, and it is also worthwhile to analyze the mentorship experience

from the perspective of both female and male scientists. In the paper 'The association between early career informal mentorship in academic collaborations and junior author performance', AlShebli, Bedoor, Kinga Makovi & Talal Rahwan gave the following insights: 1)The scientific impact of mentors is more important than the number of their collaborators. 2)The association between the big-shot experience and the post-mentorship outcome is persistent regardless of discipline, the affiliation rank, the number of mentors, the average age of mentors, the protégé's gender, and the protégé's first year of publication. 3) The gender of the mentor and their protégé predicts not only the influence of the the protégé, but also the gain of the mentors.
This report is based on paper 'The association between early career informal mentorship in academic collaborations and junior author performance', the authors show that their study has the following advantages. Firstly, they study mentorship in a broader sense, which may involve multiple senior collaborators who may or may not assume formal supervisory roles. Secondly, they analyze the actual scientific impact of the collaboration, which also allows them to avoid sample selectivity as well as recall and reproducibility bias. Thirdly, they analyze thousands of journals spanning multiple disciplines. Fourthly, they carefully compare millions of mentor–protégé pairs in order to better understand the association between mentorship quality and scientific careers. Finally, they consider students who remain scientifically motivation after their mentorship period end. These are advantages over previous research on mentorship in academia that were not available[8-15]. I will use the data provided by the authors of the paper to go further in the analysis, such as identifying the relationship of number of mentors corresponding to the protégé gender, and the possible problems from the data[16].
One data set will be used in this report, I will talk more about the data background in the Methodology section(section 2), as well the models I use in the report for further analysis and identifying problems. All the figures and tables results will be displayed in the Results section(section 3), and overall discussions, summary & conclusions, weakness & next steps will be in the Discussion section(section 4). Reference materials will be listed in the final section(section 5).

Methodolody

This report is about some critiques of the paper 'The association between early career informal mentorship in academic collaborations and junior author performance'. In this section I will further talk about the data and the models use.

Data

```
## # A tibble: 6 x 8
##   Disambiguated_P~ numMentors ProtegeFirstPub~ ProtegeGender AffiliationRank
##              <dbl>      <dbl>            <dbl> <chr>         <chr>
## 1               19          9             2005 male          101-150
## 2               61          2             2003 male          >1000
## 3               65          4             2004 male          >1000
## 4              180          4             2007 female        >1000
## 5              182          1             1976 male          >1000
## 6              235          1             1998 female        76
## # ... with 3 more variables: AvgMentorsAcAges <dbl>,
## #   NumYearsPostMentorship <dbl>, if_gender_F <dbl>
```

The data I use is provided by the authors and can be found and downloaded on website https://github.com/bedoor/Mentorship, the specific route is: Mentorship/Repository_Data/Data_7yearcutoff.csv from the 'bedoor/Mentorship' GitHub repository. According to the Methods section in 'The association between early career informal mentorship in academic collaborations and junior author performance', they indicates that "the data used for this study consists of all the papers included in the Microsoft Academic Graph (MAG)

dataset up to December 31st, 2019[17-18]". The dataset includes records of scientific publications that specify the date of publication, the author's name and affiliations, and the place of publication. The authors use the information provided by the MAG dataset to derive two key measures: the scientists' disciplines and their impact. In addition to this, they also come up with other measures, such as the gender of the scientists, the ranking of each university, etc.
Their analysis excludes the following types of scientists, (i) who had never published anything without mentorship because we can not analyze their scientific impact at the senior level independently of their mentors; (ii) who had only individually authored papers or collaborated with junior colleagues or seniors at other universities because we can not clearly identify who their mentors were; (iii) who had not published anything more than five years apart; and (iv) who had only collaborated with senior scientists outside their primary discipline.
I then select some specific variables for the analysis, and create a new variable which determines whether the protégé is female or not.

Model

In this report, I only use two models: the simple linear regression and the logit regression.
A simple linear model is a linear regression model with a single explanatory variable. A logit model, also called a logistic model, is used to model the probability of a certain class or event existing, such as the event of being female protégé or not in this report.
The general model of simple linear regression is:

$$Y_i = \alpha + \beta x_i$$

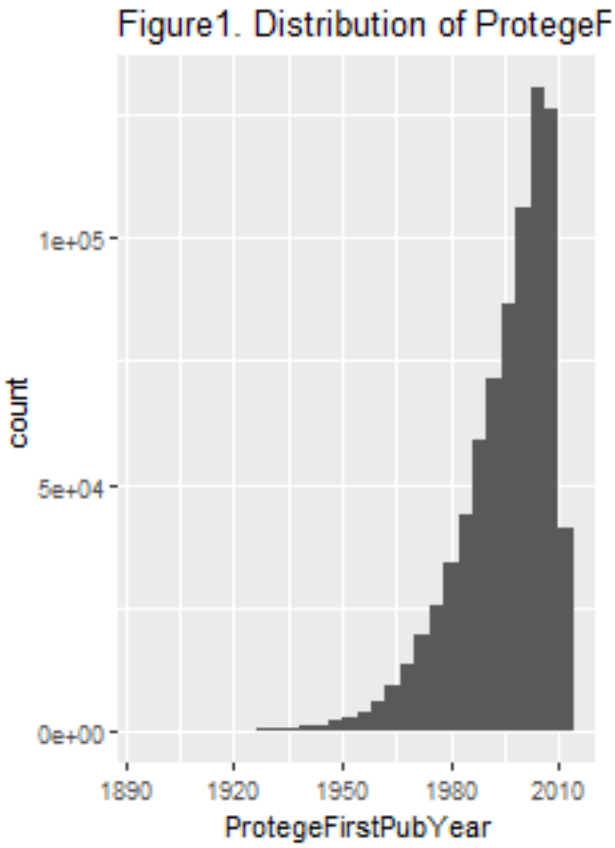
which describes a line with slope β and y-intercept α.
The general model of logit regression is:

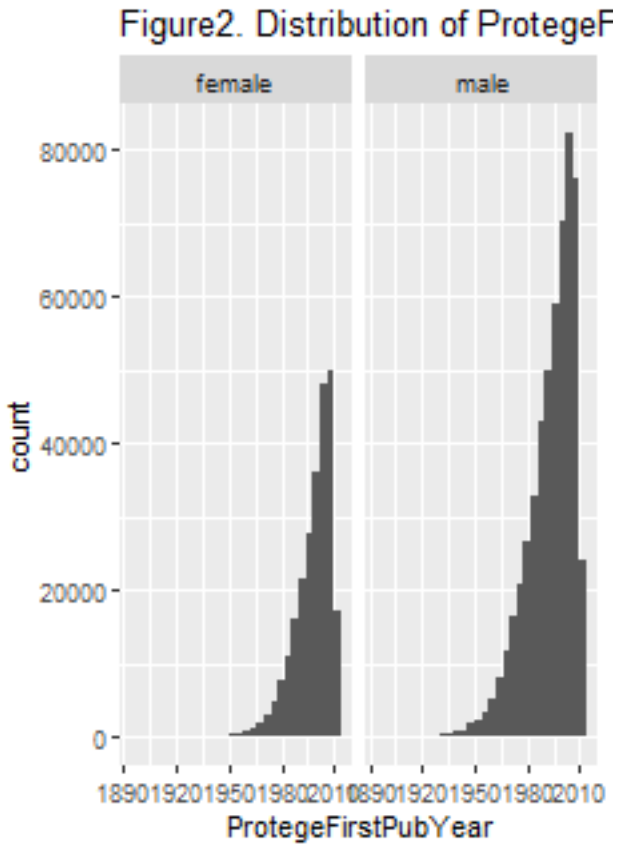
$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

While p is the probability of the event of interest occurring, coefficients like $\beta_1, ..., \beta_k$ represent change in log odds.
I use simple linear regression between variables ProtegeFirstPubYear & AvgMentorsAcAges, ProtegeFirstPubYear & numMentors, and logit regression with predictor variables ProtegeFirstPubYear & AvgMentorsAcAges & numMentors. For the linear model part, the variables are separately corresponding to Y_i and x_i depends on which is the independent variable and which is the dependent variable. For the logit model part, the variables are separately corresponding to $x_{ProtegeFirstPubYear}, x_{AvgMentorsAcAges}, x_{numMentors=1\cdots}$ in our model, and the probability of protégé is female or not is corresponding to p.

Results

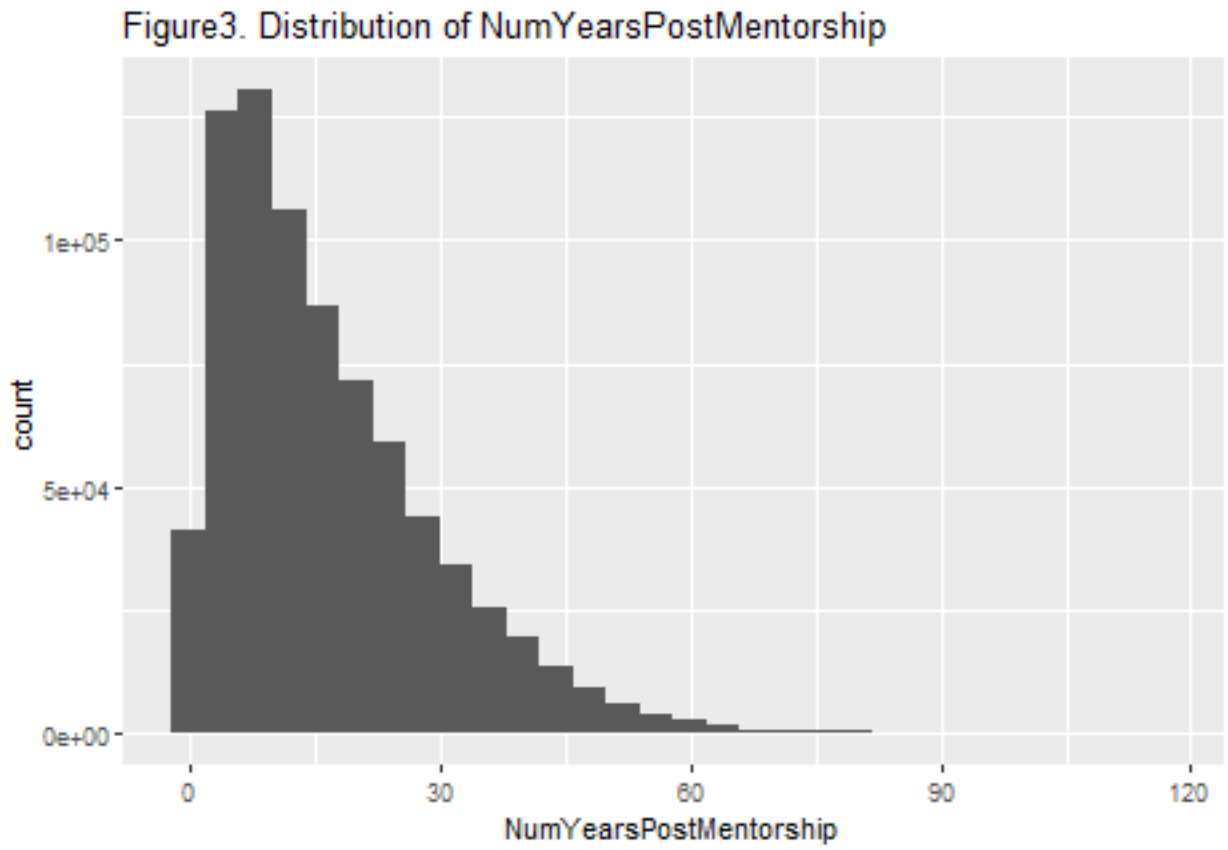
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1897    1989    1999    1996    2006    2013
```





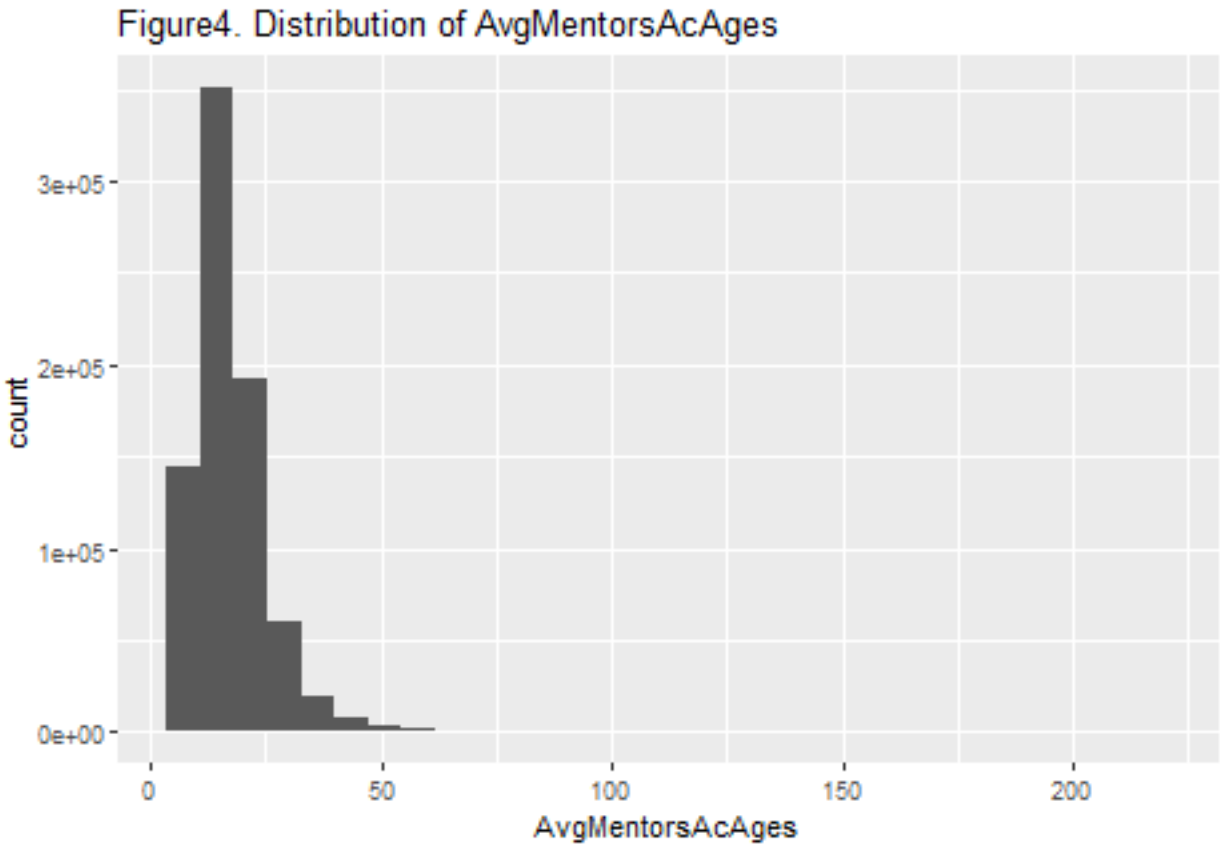
- Figure1. & Figure2 are histograms corresponding to the ProtegeFirstPubYear which is the variable corresponding to the year in which the protégé published their first mentored paper. Figure1. is the histogram for all the protégé no matter the gender, Figure2. is the histogram for protégé group by gender. According to these figures, it is obviously that the range of the year in which the protégé published their first mentored paper is from 1897 to 2013(specific numbers obtained from the table of the quantiles, minimum and maximum of the variable ProtegeFirstPubYear).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   14.00   17.29   24.00  116.00
```



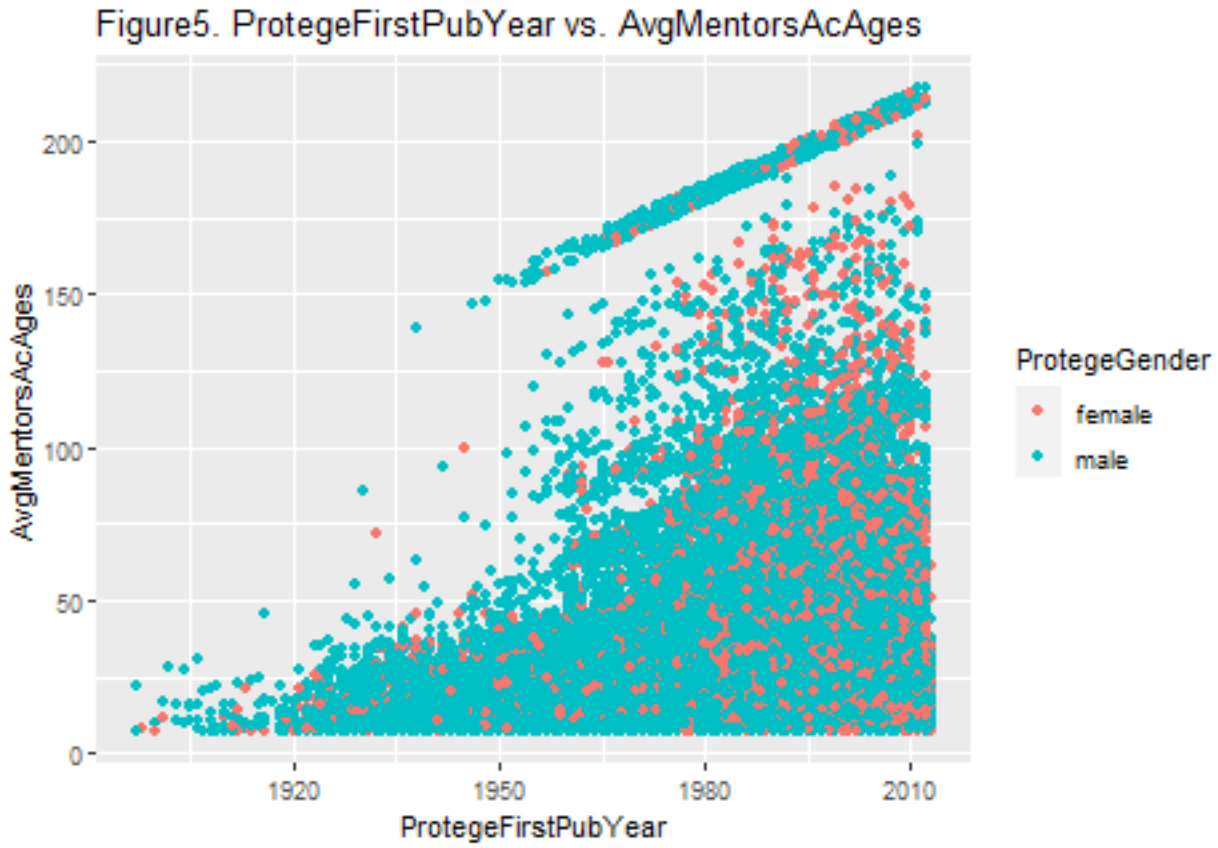
- Figure3. is the histogram of the distribution of variable NumYearsPostMentorship which is the variable corresponding to the number of years post mentorship. According to this figure, it is obviously that the range of the number of years post mentorship is from 0 to 116(specific numbers obtained from the table of the quantiles, minimum and maximum of the variable NumYearsPostMentorship).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    12.0    16.0    17.8    21.0   217.0
```



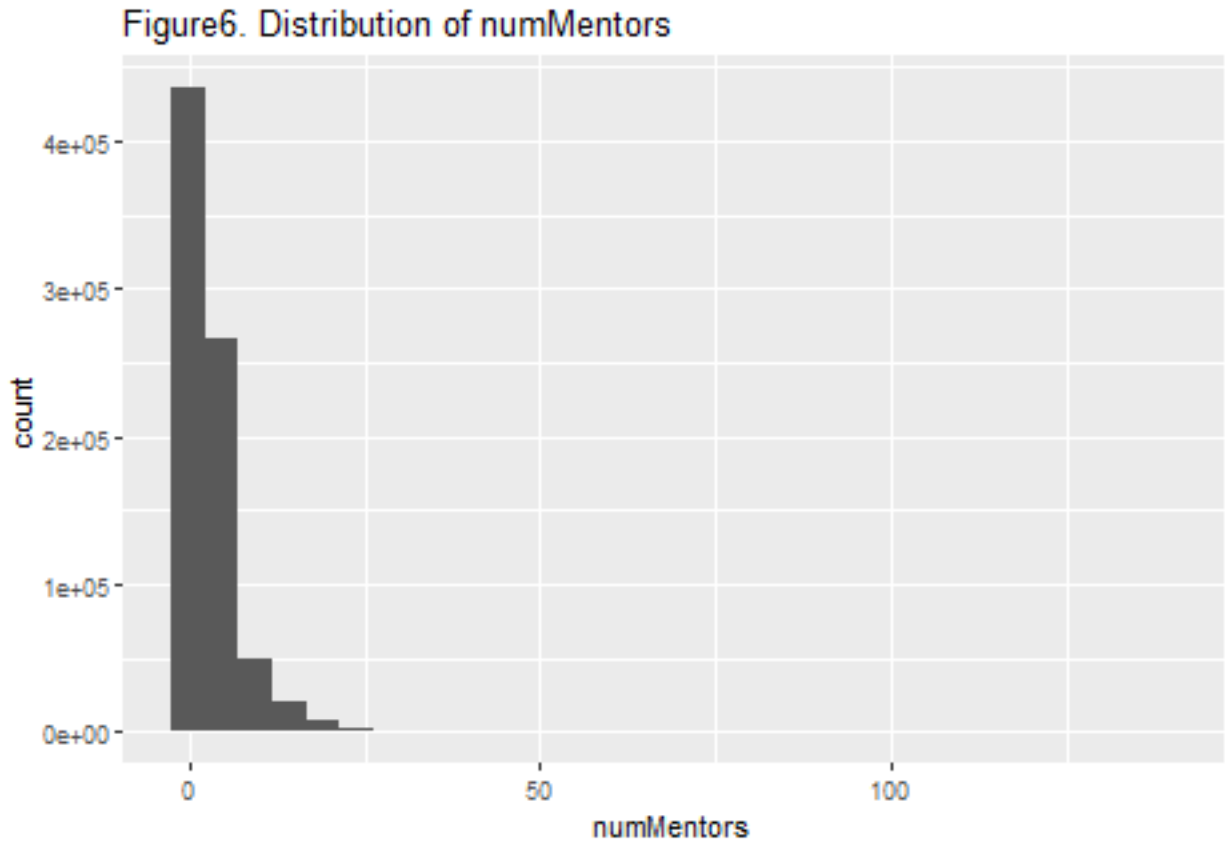
- Figure4. is the histogram of the distribution of variable AvgMentorsAcAges which is the variable corresponding to the average academic age of mentors. According to this figure, it is obviously that the range of the average academic age of mentors is from 7 to 217(specific numbers obtained from the table of the quantiles, minimum and maximum of the variable AvgMentorsAcAges).

```
##
## Call:
## lm(formula = data3$AvgMentorsAcAges ~ data3$ProtegeFirstPubYear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.166  -5.718  -1.850   2.835 197.992
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.398e+02  1.893e+00  -73.83   <2e-16 ***
## data3$ProtegeFirstPubYear  7.895e-02  9.485e-04   83.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.97 on 784746 degrees of freedom
## Multiple R-squared:  0.008752,   Adjusted R-squared:  0.008751
## F-statistic:  6929 on 1 and 784746 DF,  p-value: < 2.2e-16
```



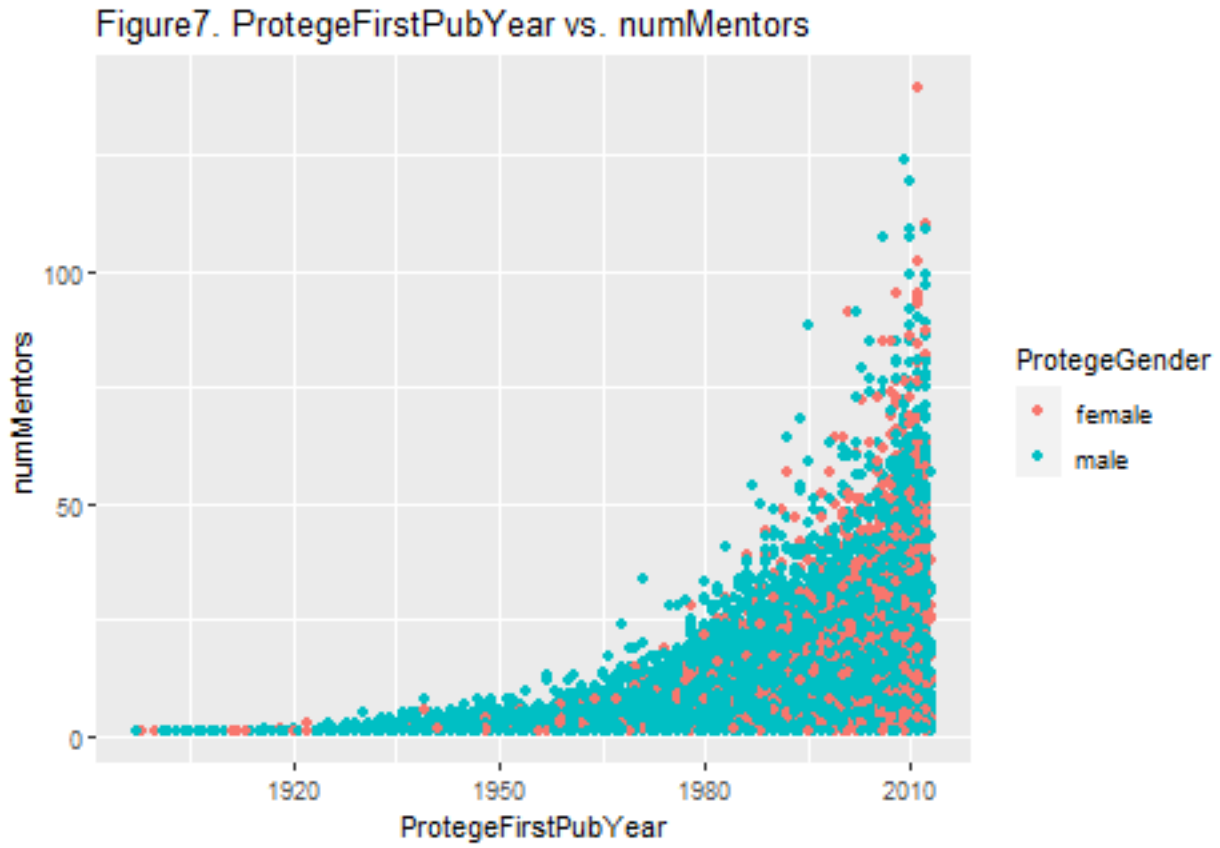
- Figure5. is a scatterplot between variables ProtegeFirstPubYear and AvgMentorsAcAges and coloured by gender. According to the linear regression between these variables, as ProtegeFirstPubYear increases 1 unit, AvgMentorsAcAges increases 7.895e-02.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   3.561   4.000 139.000
```



- Figure6. is the histogram of the distribution of variable numMentors which is the variable corresponding to the number of mentors. According to this figure, it is obviously that the range of the number of mentors is from 1 to 139(specific numbers obtained from the table of the quantiles, minimum and maximum of the variable numMentors).

```
##
## Call:
## lm(formula = data3$numMentors ~ data3$ProtegeFirstPubYear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.794  -2.295  -1.081   0.776 134.349
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.388e+02  6.876e-01  -201.8   <2e-16 ***
## data3$ProtegeFirstPubYear  7.132e-02  3.445e-04   207.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.983 on 784746 degrees of freedom
## Multiple R-squared:  0.05178,    Adjusted R-squared:  0.05178
## F-statistic: 4.286e+04 on 1 and 784746 DF,  p-value: < 2.2e-16
```



- Figure7. is a scatterplot between variables ProtegeFirstPubYear and numMentors and coloured by gender. According to the linear regression between these variables, as ProtegeFirstPubYear increases 1 unit, numMentors increases 7.132e-02.

```
##
## Call:
## glm(formula = if_gender_F ~ ProtegeFirstPubYear + AvgMentorsAcAges +
##     as.factor(numMentors), family = "binomial", data = data_filter)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3138  -0.9213  -0.7682   1.3673   2.7603
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)              -5.996e+01  4.516e-01 -132.761  < 2e-16 ***
## ProtegeFirstPubYear       2.958e-02  2.265e-04  130.575  < 2e-16 ***
## AvgMentorsAcAges          3.493e-03  2.209e-04   15.812  < 2e-16 ***
## as.factor(numMentors)2    3.279e-02  6.993e-03    4.690 2.74e-06 ***
## as.factor(numMentors)3    8.675e-02  8.096e-03   10.716  < 2e-16 ***
## as.factor(numMentors)4    1.156e-01  9.415e-03   12.283  < 2e-16 ***
## as.factor(numMentors)5    1.244e-01  1.092e-02   11.389  < 2e-16 ***
## as.factor(numMentors)6    1.502e-01  1.255e-02   11.970  < 2e-16 ***
## as.factor(numMentors)7    1.260e-01  1.455e-02    8.661  < 2e-16 ***
## as.factor(numMentors)8    1.636e-01  1.649e-02    9.922  < 2e-16 ***
```

```
## as.factor(numMentors)9   1.295e-01  1.883e-02    6.878 6.08e-12 ***
## as.factor(numMentors)10  1.438e-01  2.118e-02    6.789 1.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 922317  on 742276  degrees of freedom
## Residual deviance: 899792  on 742265  degrees of freedom
## AIC: 899816
##
## Number of Fisher Scoring iterations: 4
```

- This is our logit regression analysis, this model is used to help us determine how ProtegeFirstPubYear, AvgMentorsAcAges, and numMentors affect whether protégé is female or not. According to the logit regression above, for every additional unit increase in ProtegeFirstPubYear, we expect the log odds of protégé is female to increase by 2.958e-02, for every additional unit increase in AvgMentorsAcAges, we expect the log odds of protégé is female to increase by 3.493e-03.

Discussion

In this section I will summarize the conclusion and the weakness and further steps for the report.

Summary

Overall, I examine the data provided by the authors of this paper, and I only examine the data file Mentorship/Repository_Data/Data_7yearcutoff.csv from the 'bedoor/Mentorship' GitHub repository on website https://github.com/bedoor/Mentorship. I also build a logit model to determine how ProtegeFirstPubYear, AvgMentorsAcAges, and numMentors affect whether protégé is female or not.

Conclusions

From Figure1. & Figure2., it is clear that the time range of the protégé published their first mentored paper is very large(from 1897 to 2013), which will lead the comparison between the impact of papers published in 1897 and papers published in 2013 more difficult. From Figure3., the range of the number of years post mentorship is from 0 to 116; from Figure4., the range of the average academic age of mentors is from 7 to 217; from Figure6., the range of the number of mentors is from 1 to 139; the above three figures show unrealistically large value separately corresponding to variables NumYearsPostMentorship; AvgMentorsAcAges; numMentors. This can be considered as problems or errors of the data.
According to the two linear regression, they show weak linear relationship between the year that the protégé published their first mentored paper and the average academic age of mentors or the number of mentors, and no matter the protégé's gender, the correlations are similar in identical linear regression. From the logit regression, the model can be used to determine how the variables affect whether protégé is female or not, as the year that protégé published their first mentored paper and the average academic age of mentors increase, the probability of protégé being female also increases. However, the data provided by the authors do not give the information of mentor's gender.

Weakness & Next Steps

This report only concerns on several variables from the dataset provided by the authors, since the data provided by the authors do not give the information of mentor's gender and some variables have unrealistically

large value, though I filter the data to control the number of mentors being under 10 when dealing with the logit regression, it definitely is still influenced by other unrealistically large values, which lead the logit model poorly to believe or use. Meanwhile, the logit model is aim to determine how the variables affect whether protégé is female or not, as mentioned before, the data provided by the authors do not give the information of mentor's gender, so I cannot make connection between the genders of protégé and mentor - which is one of the focus in the paper 'The association between early career informal mentorship in academic collaborations and junior author performance'.

For the work that can be done afterwards, one is to further clean the data to ensure that all unreasonably large values are filtered out, and another is to further obtain information of mentor gender, then relate mentor gender to protégé gender and fit suitable models to observe association between. Paper 'The association between early career informal mentorship in academic collaborations and junior author performance' also provides two measurements of mentorship quality, one is "big-shot", the other is "hub", I do not discuss these two variables in the report, while these two measurements are also worthy to discuss.

References

[1]. AlShebli, Bedoor, Kinga Makovi & Talal Rahwan, 2020, 'The association between early career informal mentorship in academic collaborations and junior author performance', Nature Communications. https://www.nature.com/articles/s41467-020-19723-8/ [2]. Kram, K. E. Mentoring at Work: Developmental Relationships in Organizational Life. (University Press of America, Lanham, MD, 1988).

[3]. Allen, T. D., Eby, L. T., Poteet, M. L. & Lentz, E. Career benefits associated with mentoring for protégé: a meta-analysis. J. Appl. Psychol. 89, 127–136 (2004).

[4]. Scandura, T. A. Mentorship and career mobility: an empirical investigation. J. Organ. Behav. 13, 169–174 (1992).

[5]. Higgins, M. C. & Kram, K. E. Reconceptualizing mentoring at work: a developmental network perspective. Acad. Manage. Rev. 26, 264–288 (2001).

[6]. Higgins, M. C. & Thomas, D. A. Constellations and career: toward understanding the effects of multiple developmental relationships. Organ. Behav. 22, 223–247 (2001).

[7]. Editorial. Science needs mentors. Nature 573,5https://doi.org/10.1038/ d41586-019-02617-1 (2019).

[8]. Reskin, B. F. Academic sponsorship and scientists' careers. Sociol. Educ. 52, 129–146 (1979).

[9]. Kirchmeyer, C. The effects of mentoring on academic careers over time: testing performance and political perspectives. Hum. Relat. 58, 637–660 (2005).

[10]. Paglis, L. L., Green, S. G. & Bauer, T. N. Does adviser mentoring add value? a longitudinal study of mentoring and doctoral student outcomes. Res. High. Educ. 47, 451–476 (2006).

[11]. Malmgren, R. D., Ottino, J. M. & Amaral, L. A. N. The role of mentorship in protégé performance. Nature 465, 622–626 (2010). [12]. Chariker, J. H., Zhang, Y., Pani, J. R. & Rouchka, E. C. Identification of successful mentoring communities using network-based analysis of mentormentee relationships across nobel laureates. Scientometrics 111, 1733–1749 (2017).

[13]. Rossi, L., Freire, I. L. & Mena-Chalco, J. P. Genealogical index: a metric to analyze advisor-advisee relationships. J. Inform. 11, 564–582 (2017).

[14]. Linèard, J. F., Achakulvisut, T., Acuna, D. E. & David, S. V. Intellectual synthesis in mentorship determines success in academic careers. Nat. Commun. 9, 1733–1749 (2018).

[15]. Liu, J. et al. Understanding the advisor-advisee relationship via scholarly data analysis. Scientometrics 116, 161–180 (2018).

[16]. Daniel E. Weeks November 21, 2020, 14:11. Mentorship. https://danieleweeks.github.io/Mentorship/#summary

[17]. Sinha, A. et al. An overview of microsoft academic service (mas) and applications. In Proceedings of the 24th International Conference on World Wide Web, 243–246 (ACM, New York, 2015).

[18]. Wang, K. et al. A review of microsoft academic services for science of science studies. Front. Big Data 2, 45 (2019).