When Do A Scientists Do Their Best Work? Finding Associations Between Scientific Impact and Timing of A Scientist's Career

Haoda Li

January 30, 2020

Summary

Based on the research on Nobel Prize Laureates and their prize papers, this report examines the relationship between the scientific impact of a paper and the timing of a paper during a scientist's career. The report also investigates on the effect of different factors on the scientific impact and the timing. By building Poisson regression models and using stratified analysis, the analysis shows some evidence of relationship between the scientific impact and timing of a paper, and the relationship is higher dependent on the publication year.

1. Introduction

Albert Einstein developed the theory of relativity at the age of 26, Ludwig Boltzmann predicted the properties of atoms at the age of 28. A common belief on scientists is that there is an early peak on productivity and then quickly declines. With the examinations on the some of the most genius scientists, the Nobel Prize winners of natural science, we try to model how scientific impact is influenced by the timing of the scientist's career.

The paper's scientific impact will be measured by its number of citations. Rather than other Scientometrics measurements such as IF, citation number is the most accessible and universal for our time range. The paper's timing during a scientist's career will be measured in the ratio as

$$\text{ratio} = \frac{\text{Number of papers before the scientist's fist prize-winning paper}}{\text{Number of papers the scientist produced up to winning Nobel Prize}}$$

Using this ratio allows us to focus on the stage of a scholar's career rather than his/her overall life.

The scope of the research is on prize-winning papers from laureates of Nobel Prize in Physics, Chemistry, and Medicine, from 1880 to 2010. After cleaning on the original datasets on papers [1], the report is analyzed on 713 prize-winning papers and 453 laureates.

2. Methodology

2.1. Investigations on factors

First,we tried a log transformation, which is a common link function for Poisson regression. By looking at the scatter plots in Figure 1, the log number of citations gives more constant variance over the number of citations. Therefore, we assume that the number of citations follows a Poisson distribution rather than a Normal distribution. However, there is no obvious relationship between the two variables.

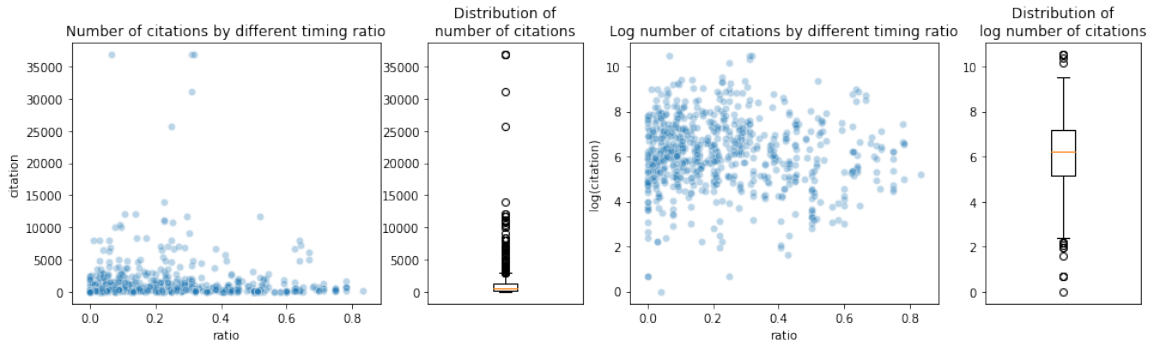

Figure 1. Number of Citations by Timing Ratio

I proposed some possible factors that may affect the relationship. Consider gender, prize category, and team size side-by-side box plots are used to see the differences of the variable's distribution within each group of the researched factor.

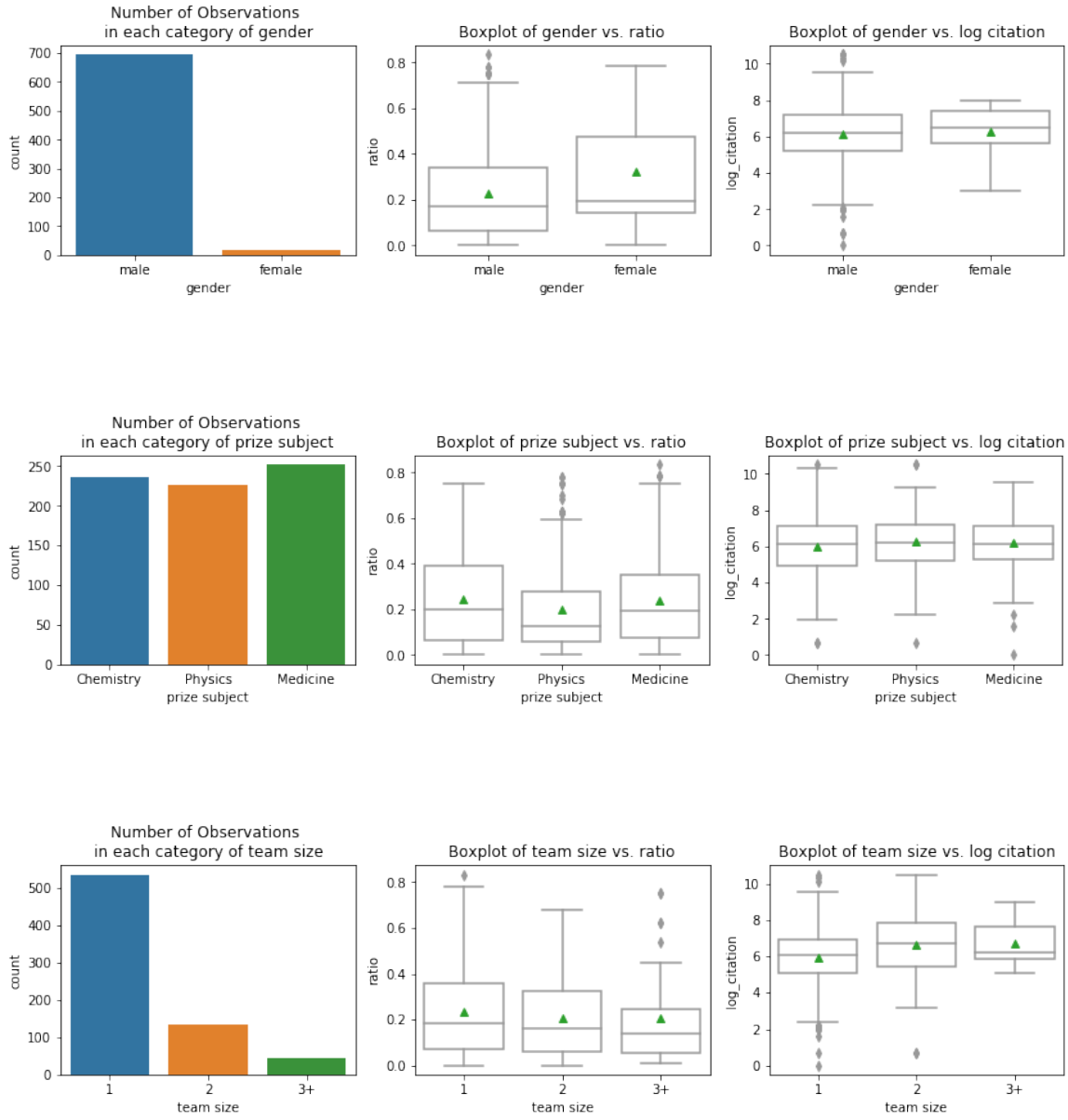

Figure 2. Effects of each categorical factor on ratio and log citation

For gender, although there is some minor differences on the statistics of ratio and log citation between males and females. Due to the extremely imbalanced number within each group, the difference is statistically insignificant.

Prize category seems to be associated with the timing ratio, where physicists seems to have their most significant work earlier in their career. However, the difference is not very significant. Also, prize category seems to be independent of citation number.

Team size might be a confounding variable for its impact on both ratio and log citation. However, due to the imbalanced sample sizes, there are only weak evidences.

I excluded the analysis on age because there is definitely a strong association between the ratio and the scientist's age at the publication. Including age will cause perfect multicollinearity on predicting the log number of citation.

Finally, notice that publication years will have strong positive association with citation numbers. Due to the publication explosion, the number of scientific publications grew by 4.6% annually [2]. Shown in Figure 2, there is a strong evidence that log citation number has a positive linear relationship with the publication year, which suggests a exponential growth of citations across the years. However, we are not clear about the relationship between the timing ration and the publication year.

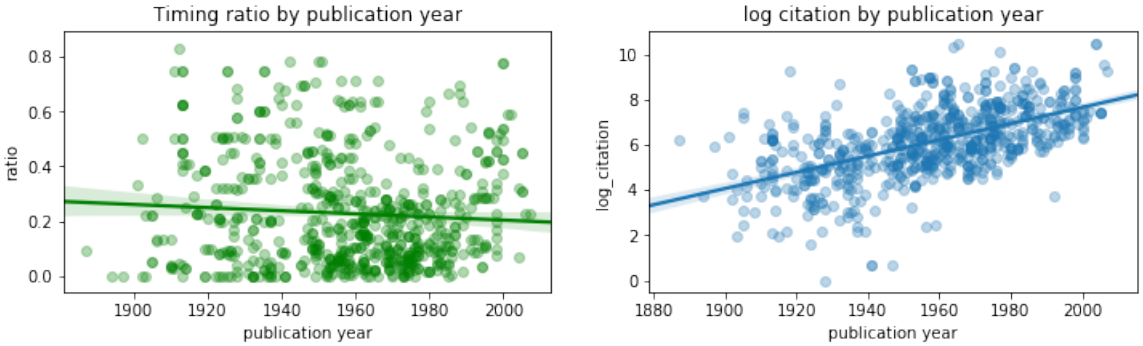

Figure 3. Effects of each continuous factor on ratio and log citation

2.2. Models

Based on the investigations, we are interested in whether team size has impact on citation number and how publication year is related to the timing ratio and citation. From the investigations, the Poisson model will be appropriate for the data. In addition, notice that there are multiple papers that belong to the same scholar. We make the assumption that each laureate will have similar behavior on their papers, i.e. random effect from laureates on citation numbers. Therefore, we assume a underlying Poisson mixed effect models. 3 candidate models are proposed after some considerations. note that we use year $=$ publication year $-$ 1880 so that the baseline is set to the minimum publication year in our research subject, and ratio is multiplied by 100 so that one unit change will be 1% change in the actual ratio.

Model 1 is the additive model with the hypothesis that the timing ratio is independent of publication year. The model is defined as

$$\begin{aligned}\log(\text{citation}) = \beta_0 + \beta_1 R + \beta_2 Y + \beta_3 T2 + \beta_4 T3 \\ + \gamma_1 \text{laureate} + \epsilon\end{aligned}$$

Model 2 is the interaction model with the hypothesis that there are two-way interactions among the three factors. The model is defined as

$$\begin{aligned}\log(\text{citation}) = \beta_0 &+ \beta_1 R + \beta_2 Y + \beta_3 T2 + \beta_4 T3 \\ &+ \beta_5 R:T2 + \beta_6 R:T3 \\ &+ \beta_7 Y:T2 + \beta_8 R:T3 \\ &+ \beta_7 R:Y \\ &+ \gamma_1 \text{laureate} + \epsilon\end{aligned}$$

Model 3 is the stratified analysis. I divided the publication years into 3 groups, $1880 \sim 1930, 1931 \sim 1980, 1981 \sim 2010$, and perform regression on each group with additive models. The year intervals are suggested by Larsen & von Ins [2], which are 3 time intervals with different growth rate. Each model is defined as

$$\begin{aligned}\log(\text{citation}|YG) = \beta_0 &+ \beta_1 R|YG + \beta_2 T2|YG + \beta_3 T3|YG \\ &+ \gamma_1 \text{laureate}|YG + \epsilon\end{aligned}$$

where R is for ratio, Y is for year, and $T2$ is for team size of 2, $T3$ is team size more than or equal to 3.

```
The rpy2.ipython extension is already loaded. To reload it, use:
  %reload_ext rpy2.ipython
```

3. Results

The fitted coefficients and standard errors of the models are shown in Figure 4; and the fit statistics are shown in the table below. Model31, Model32, Model33 are for Model3 for 1880-1930, 1931-1980, 1981-201 groups, respectively.

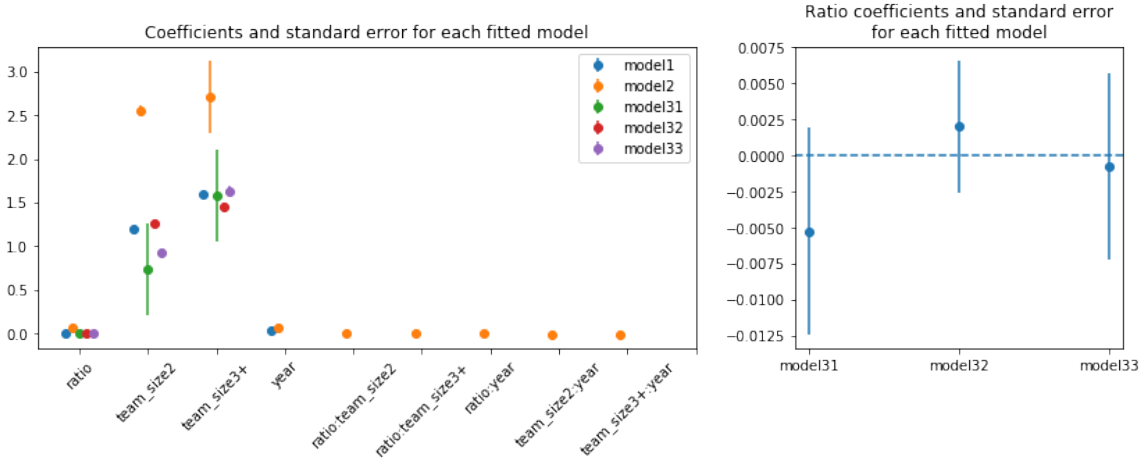

Figure 4. Fitted parameters for each model

First, notice that the fitted value for ratio and interaction terms with ratio are all approximately 0 for all fitted models. Therefore, 1% change in the timing ratio is likely to have no effect on the number of citations. The standard errors and p-values are quite small due to the large sample size, while the extremely small coefficients hints that there is no relationship between citation and timing ratio.

Second, shown by the coefficients on team sizes, there are evidences that laureates that worked in teams have higher citation numbers than laureates that worked alone. However, the differences between team size of 2 and 3

model	AIC	BIC	log likelihood	deviance
model1	118392.664780	118420.073247	-59190.332390	118380.664780
model2	117344.255367	117394.504224	-58661.127683	117322.255367
model31	30024.162862	30037.973731	-15007.081431	30014.162862
model32	69901.776874	69922.572350	-34945.888437	69891.776874
model33	14224.480962	14238.501067	-7107.240481	14214.480962

is not significant. From Model3, the large p-value may suggest that the association between team size and citation numbers is less likely to be significant.

Moreover, Model2 has slightly higher log-likelihood and BIC than Model1, hence the interaction model is probably a better fit than the additive model. Also, consider the models on the 3 year groups, the coefficients for ratios are different and the signs overturn for Model32. This gives some evidence that publication year is a confounding variable for ratio and number of citations.

4. Conclusion

This report investigates the relationship between a paper's scientific impact, as measured in citation numbers, and the timing of the paper during a scientist's career, as measured in a ratio. By proposing and analyzing several Poisson mixed effect models, we conclude that there are strong evidence that timing is independent of the scientific impact. The scientific impact is more likely to be effected by the team size and publication years. Also, the way that scientific research is conducted was evolved through time, as the publication years is likely to be a confounding variable for scientific impact and timing.

In this analysis, we have to note that the AIC, BIC, and log-likelihoods are quite large for all the examined models. These fitted statistics hint that the fitted models are likely to be invalid. Further research should consider different models and how publication year impacts other factors.

References

[1] Jichao Li, Yian Yin, Santo Fortunato, and Wang Dashun. A dataset of publication records for Nobel laureates, 2018.

[2] Peder Larsen and Markus Von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603, 2010.