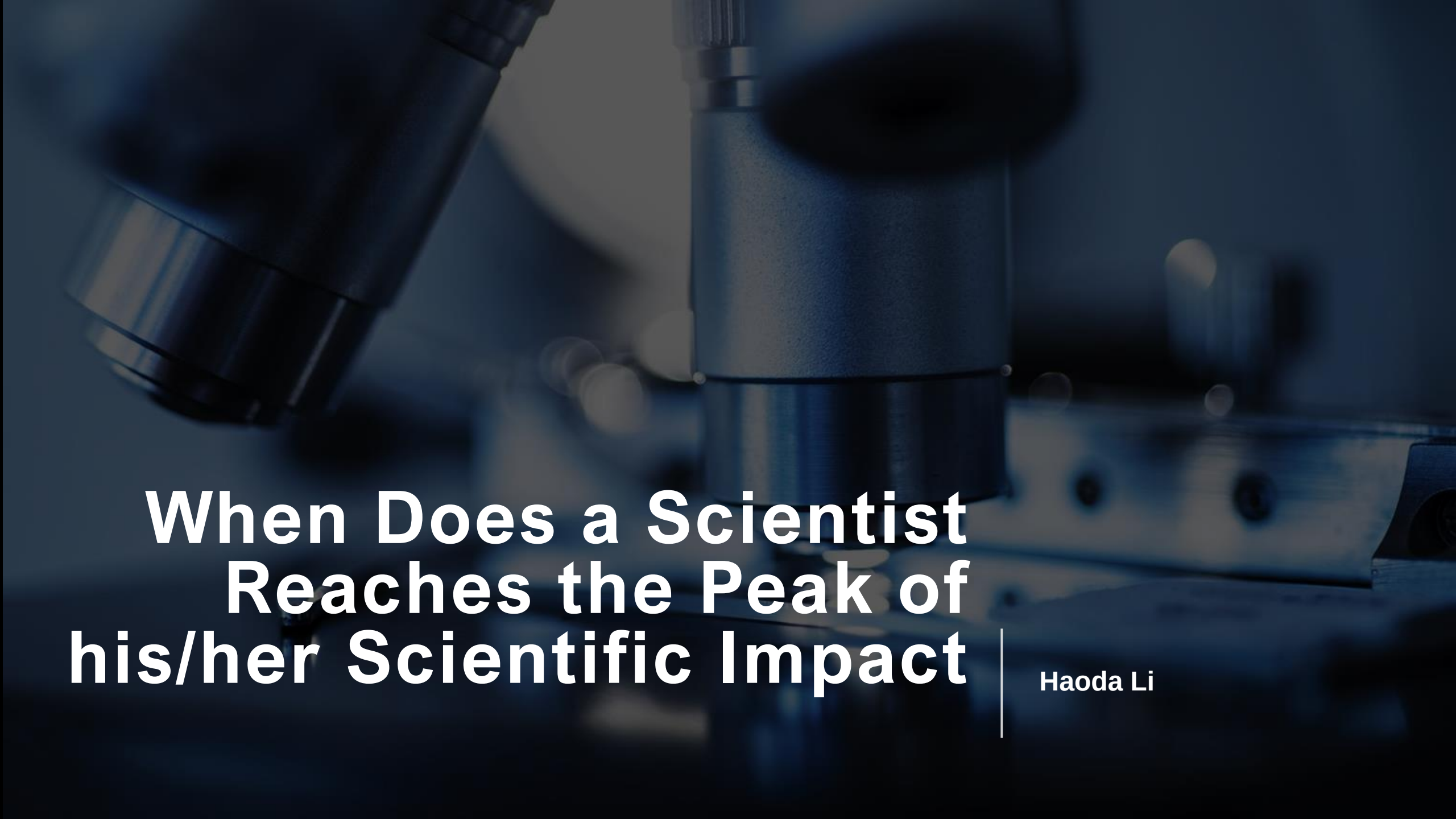


When Does a Scientist Reaches the Peak of his/her Scientific Impact

Haoda Li

Problem Definition

Is there any correlation between **Scientific impact** of a paper and **timing** of the paper during a scholar's career in natural science fields?

Research Subject

All papers that win Nobel Prize of Chemistry, Physics, and Medicine from 1880 to 2010.

Scientific Impact

Number of citations. The most direct and commonly used measurement consider the range of year.

Timing

Measured as

$$\text{ratio} = \frac{\text{Number of papers up to the given paper}}{\text{Number of papers in total}}$$

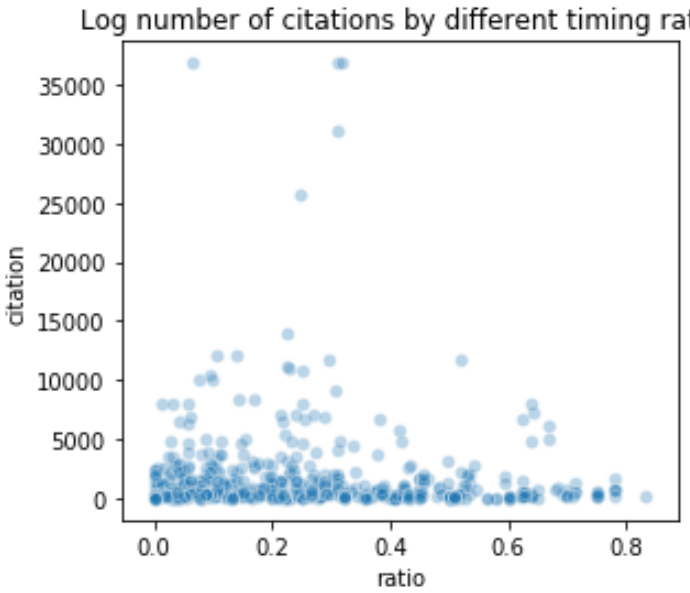
Only consider the stage of the career as a scholar, not the whole life.

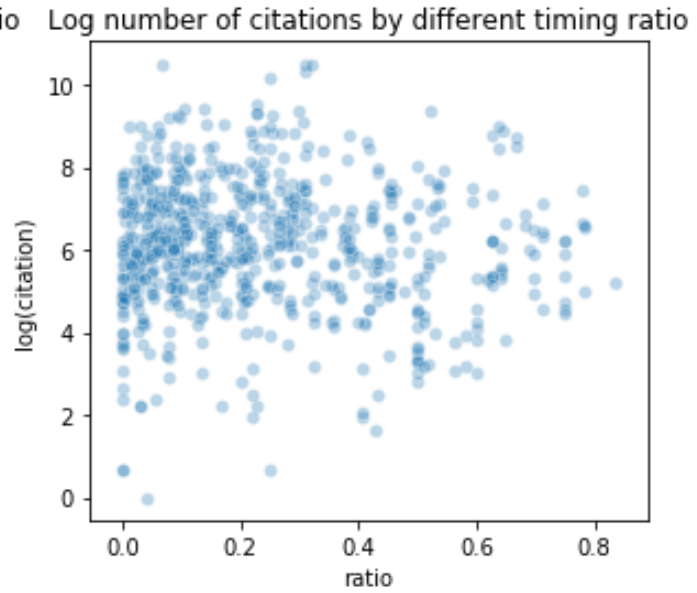
Dataset

874 prize-winning paper from *A dataset of publication records for Nobel laureates* [1].

The dataset is not perfectly correct. Some Laureates do not correspond to his/her paper.

713 papers from 453 Laureates after cleaning and merging.





LID	name	prize_year	title	pub_year	paper_id	DOI
20148	fischer, h	1930	Einfluss der configuration auf die wirkung der...	1894.0	1.992788e+09	10.1002/cber.18940270364

This paper is written by Emil Fischer and won Nobel Prize in Chemistry in 1902.

[1] *Li, Jichao; Yin, Yian; Fortunato, Santo; Wang Dashun, 2018, "A dataset of publication records for Nobel laureates", https://doi.org/10.7910/DVN/6NJ5RN,*

Factors not to consider

Gender

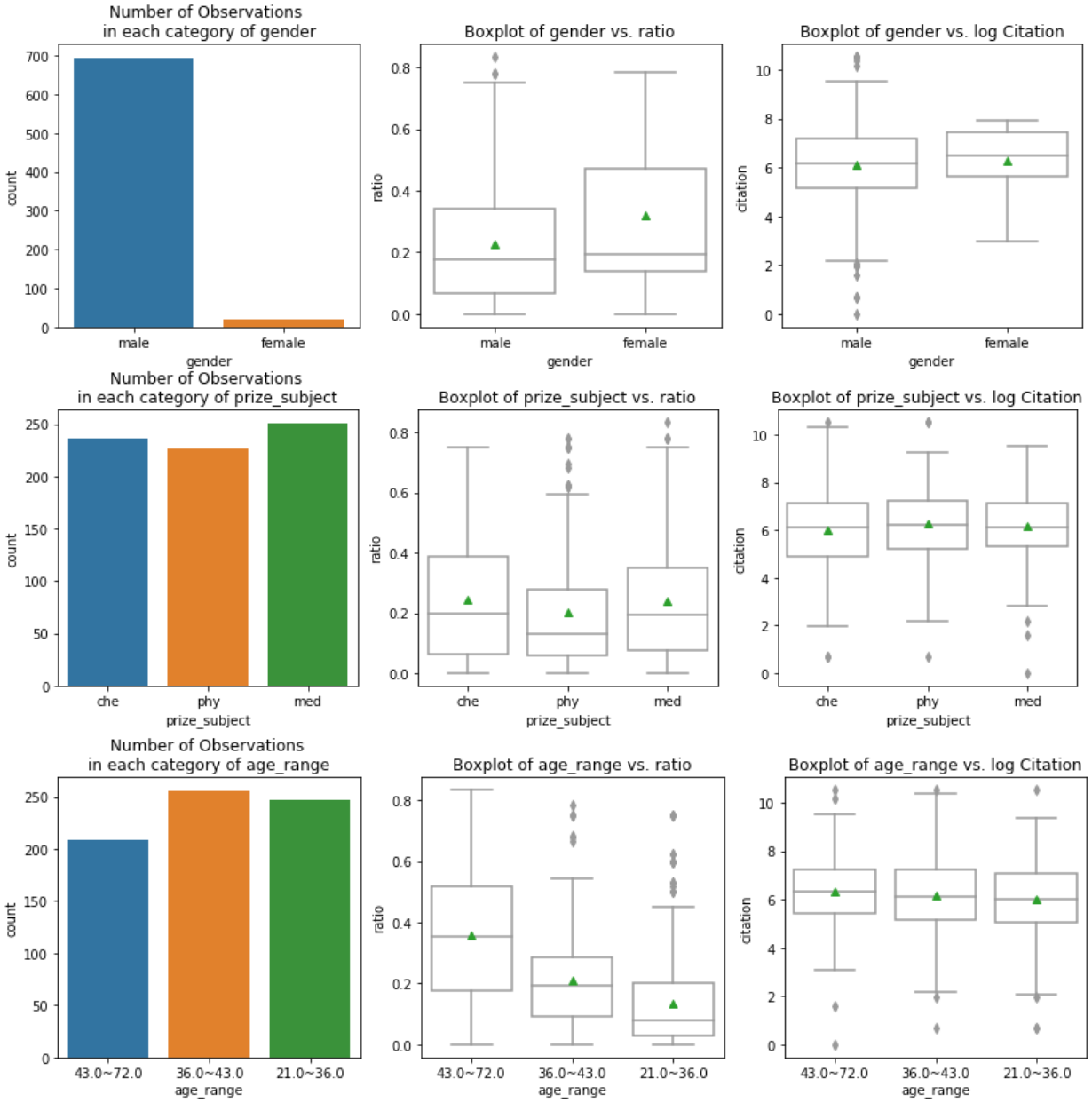
Extremely imbalanced sample size, weak evidence of association between ratio and gender.

Prize Category

Evidence of no relationship. Prize category may be independent from ratio and citation.

Age at the publication

The boxplot shows evidence of relation-ship between age and ratio. However, age is definitely highly correlated with ratio.



Factors to consider

Team Size

Imbalanced, might be a confounding variable

Publication Year

Shows strong evidence of correlations on **pub year vs. citation** and weak evidence of **pub year vs. citation**

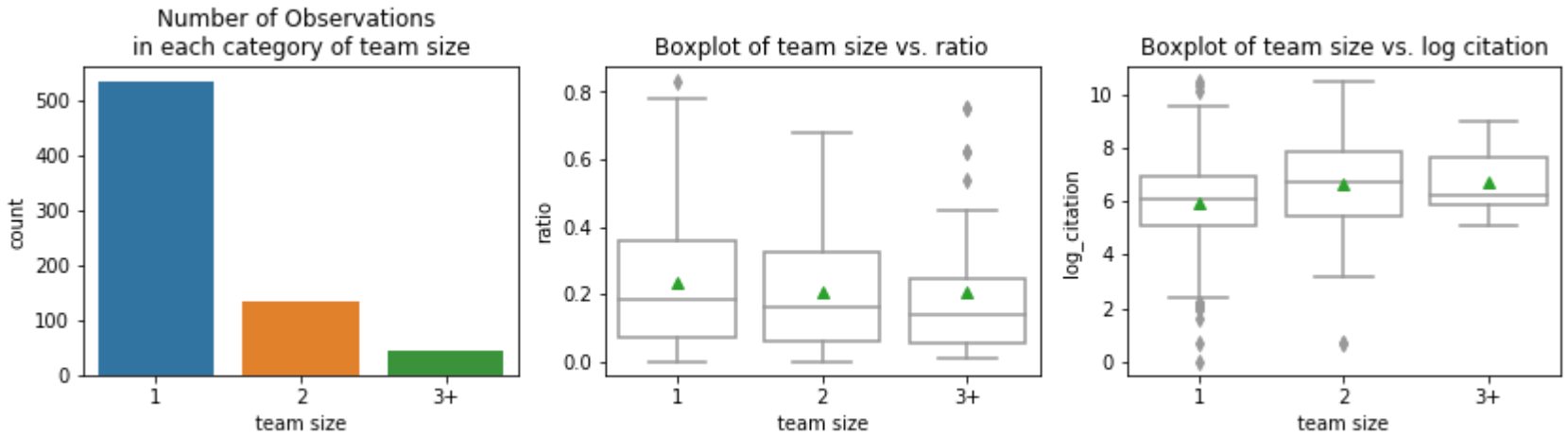
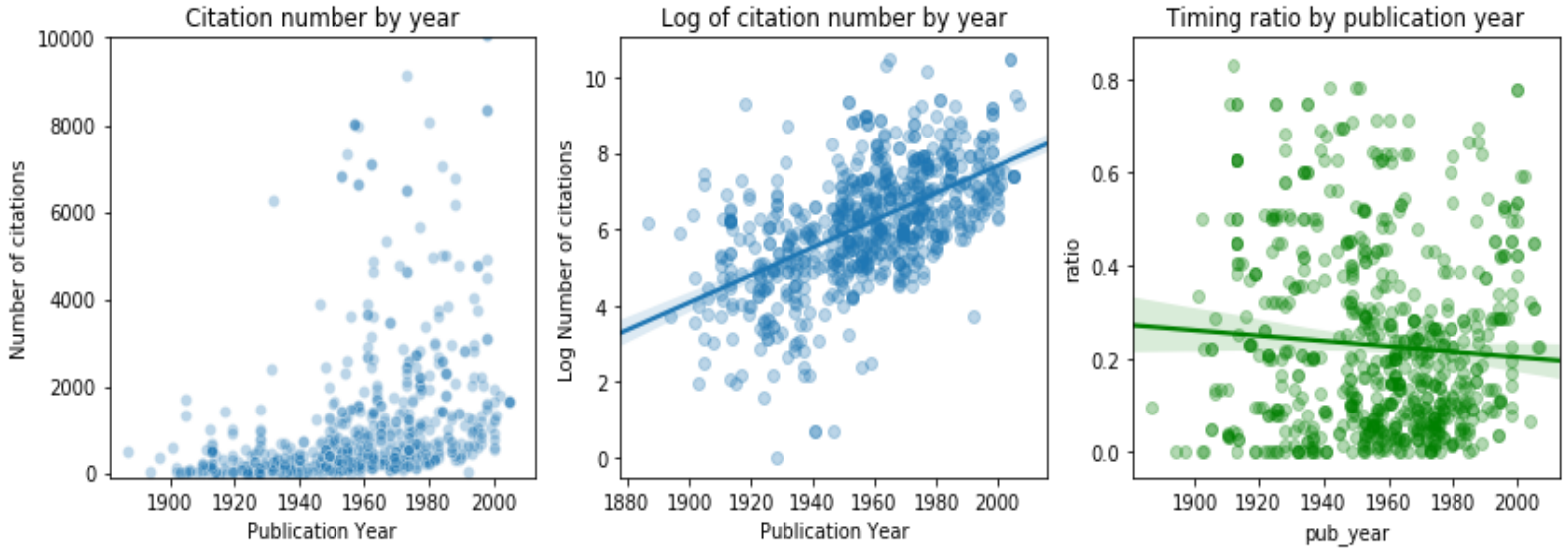


Figure 2. Effects of each categorical factor on ratio and log citation



Models

Assume a underlying Poisson distribution for all models

Publication year have much effect on citation, but does it effect other variables?

Is there any evidence that **team size** has any effect on our variables.

Model 1 Assume no relationship between publication year and other independent variables

$$\log(Citation) = \beta_0 + \beta_1 Year + \beta_2 Ratio + \beta_3 TeamSize + \epsilon$$

Model 2 Assume interactions between publication year and other independent variables

$$\log(Citation) = \beta_0 + \beta_1 Year + \beta_2 Ratio + \beta_3 TeamSize + \beta_4 Ratio{:}\,Year + \beta_5 TeamSize{:}\,Year + \epsilon$$

Model 3 Creating subgroups using publication year and build regression model on each subgroup

$$\log(Citation|YearGroup) = \beta_0 + \beta_2 Ratio|YearGroup + \beta_3 TeamSize|YearGroup$$

* $Year$ = Publication Years – 1880

Results

Consider the scale of ratio (0~1), its coefficient is too small to be significant.

Model 2 gives better log-likelihood result. However, the log-likelihood is still extremely large.

Publication year is a confounding variable on all other variables.

Model	Coef. Ratio	Coef. Team Size	Log-likelihood
Model 1	$\mathbf{0.0296}$	$\mathbf{0.168}$	$\mathbf{-8.49 \times 10^{-5}}$
Model 2	$\mathbf{1.77}$	$\mathbf{0.292}$	$\mathbf{-8.41 \times 10^{-5}}$
Model 3 (1880-1930)	$\mathbf{-0.950}$	$\mathbf{-0.0705}$	N.A.
Model 3 (1931-1980)	$\mathbf{1.11}$	$\mathbf{0.464}$	N.A.
Model 3 (1981-2010)	$\mathbf{0.433}$	$\mathbf{0.153}$	N.A.

Other Coefficients and summary statistics are omitted here
All the coefficient have p-value <0.01

Conclusions

There is no significant association between scientific impact and the timing of the paper.

The field of scientific research had been greatly evolved throughout the 20th century. Rather than God given genius, scientists nowadays work in large teams and produce more impactful results later in his/her career as a scholar.