Data Jujutsu III – Papers from UofC*

Stefano Allesina & Graham Smith *University of Chicago*

Description of the data

I have collected information on the 13,140 papers published between 2011 and 2020 by researchers with a UofC affiliation in the field of biology—including all papers in multidisciplinary journals (as such, some papers from other fields are included). We are going to explore this data set to highlight the variety of research programs and publications produced by our faculty. We will use these data to test the hypothesis that `open access` publishing leads to more citations than paywalled publishing.

The challenge

1. Read the data The data are stored in the file `All_UofC_Bio_2011-20.csv`. Read the file, and rename the columns for easier typing:

```
au = Authors
au_ids = ‘Author(s) ID‘
year = Year
journal = ‘Source title‘
cits = ‘Cited by‘
article = ‘Document Type‘
oa = ‘Access Type‘
```

2. Distribution of citations by year Several studies have shown that the number of citations per paper in a given year is approximately lognormal (with better fit for older years). Remove the papers with 0 citations, and plot the histogram for the log number of citations faceting by year. You should see that older years yield (approximately) a normal distribution. With this transformation at hand, you can attempt modeling the citations as a function of other bibliometric variables.

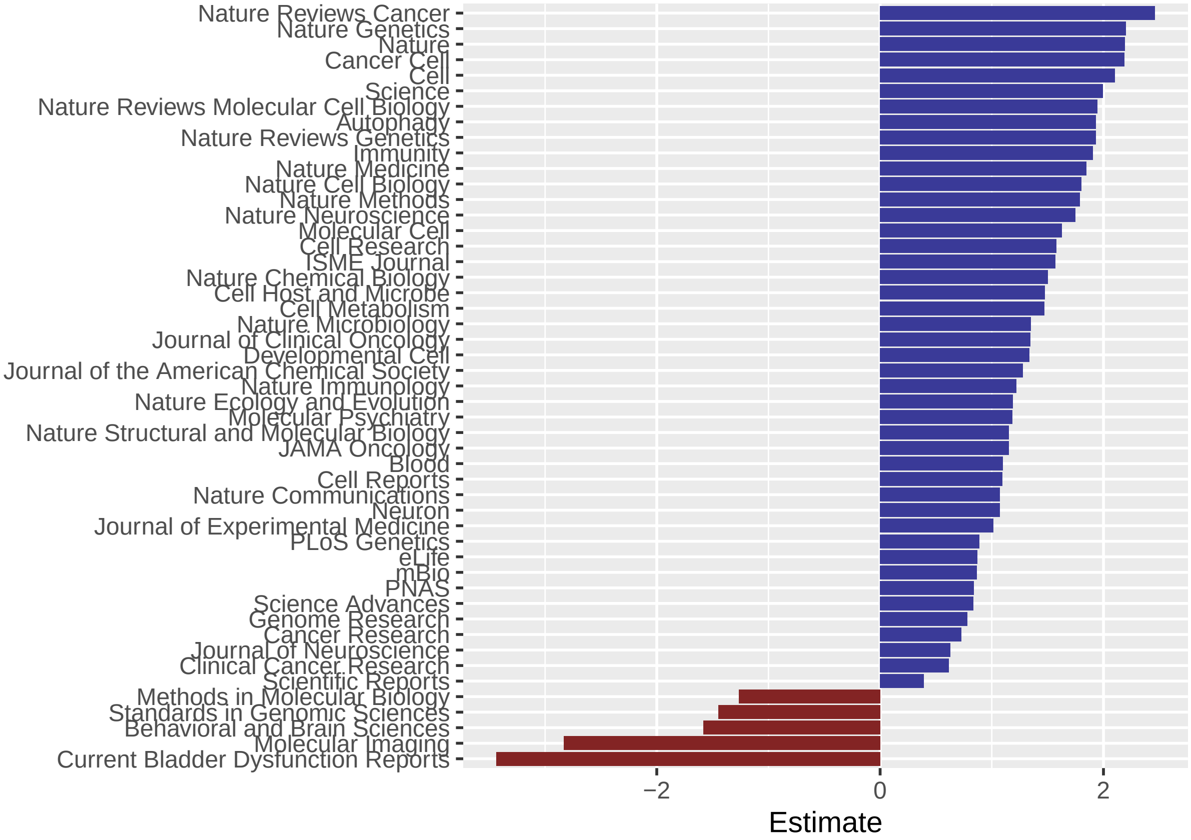
To start off, fit a linear model `log(cits) ~ I(2010 - year)` where we use `I(2010 - year)` to make the associated coefficient measure the growth in number of citations for each year.

3. Multi-authored papers The number of authors per paper varies dramatically. Count the number of authors per paper, and show that including a covariate specifying whether the paper has more than 12 authors improves the fit (store this variable in the column `multi`). Similarly, including whether the article is a research article or a review improves the fit (column `article`).

4. Top journals Of course, research papers published in high-visibility journals (such as Nature and Science) tend to receive many more citations. We can model each journal separately, and look at the distribution of effects on citations: fit the model `log(cits) ~ I(2010 - year) + multi +`

[*] This document is included as part of the Advanced Computing I packet for the U Chicago BSD qBio6 boot camp 2020. **Current version**: August 09, 2020; **Corresponding author**: sallesina@uchicago.edu.

`article + journal`, and plot the effects from most positive to most negative. Include only the journals with a strongly significant effect (e.g., pvalue $< 10^{-6}$ to avoid problems with multiple hypothesis testing), and draw a barplot such as the one below (obtained setting the baseline journal as "PLoS ONE"):



4. Effects of open access Some of the journals have an "open access option", which typically costs top dollars. Will this give your article more visibility (and therefore citations)? To test this effect, find all the journals that have published both open access and paywalled articles in the same year, and contrast the citation counts between the two subsets.

Extract all the papers for the journal/year combinations identified above. Test the effect of open access by regressing:

```
log(cits) ~ year:journal:multi:article + open
```

where `year:journal:multi` fits the mean number of log citations for each year, journal and accounting for multiauthored papers and reviews, and `open` tests the effect of having published with the open access option. Is the effect positive?

5. Most productive researchers [Optional] Count how many times does each author ID appears in the data. Find the most productive authors in biology, and try extracting their names from the `au` column.

Hints & Nifty tricks

- When performing a regression with a covariate that is a factor, you can use `relevel(my_factor, ref = "my_value")` to set the desired reference ("PLoS ONE" for the graph above).