Popularity Prediction of Scientific Publications

Adam Bacho
(adam.bacho@gmail.com)
Supervisor:
Róbert Móro

Goal

To build a predictive model that is able to estimate citation counts of articles as soon as possible after publishing them, while using article related features enriched by **Eigenfactor** and **Journal Impact Factor**.

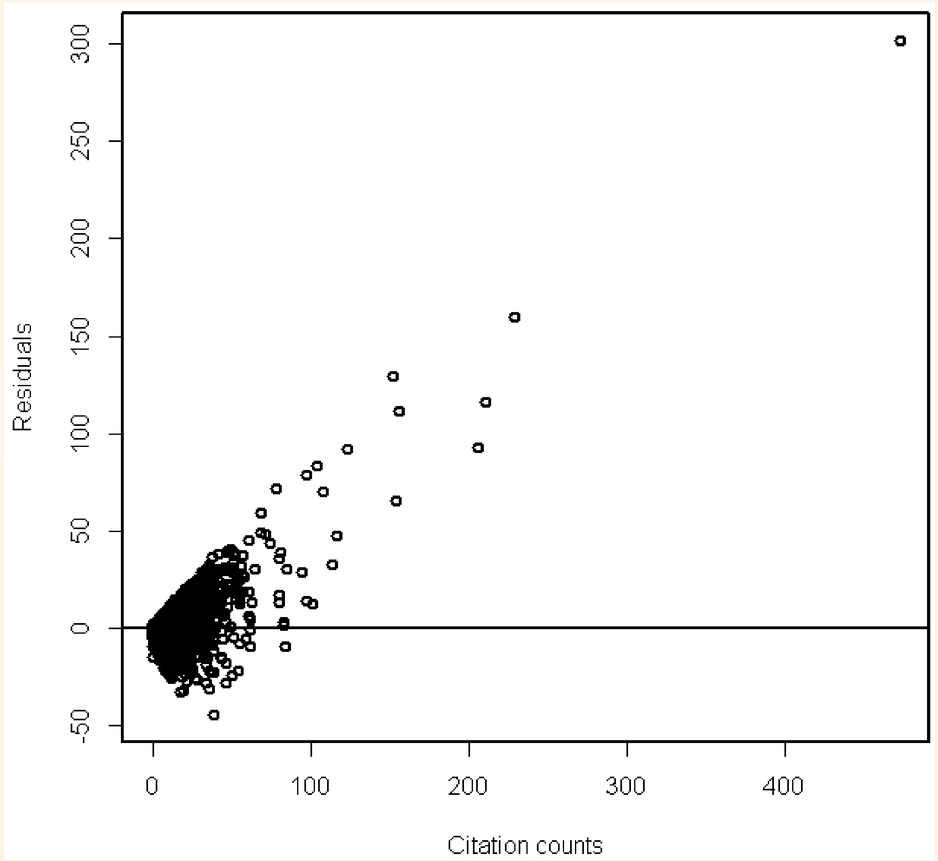
PubMed Central Dataset

- PMC Open Access Subset is a part of the total collection of articles in PMC mainly domain of medicine
- **907,851 articles**
- **4,731 journals**
- unique identifiers of articles (ISSN, PubMed ID)

8-years-after-publication prediction

- multiple linear regression was performed with only statistically significant features
- model was then enriched by citation counts after some time
- adding Journal Impact Factor resulted in worse outcome

CITATION COUNTS:	R^2 (%)	RMSE
in publication year	17.8	9.54
until 1 year after publication	42.3	8.1
until 2 years after publication	58.3	6.93
until 5 years after publication	89.8	3.45



JIF versus Eigenfactor

Tests shown **Journal Impact Factor is more powerful predictor compared to Eigenfactor**. The results are shown in the following table.

FEATURE	R^2 (%)	RMSE
JIF	17.4	8.07
Eigenfactor	10.4	8.33
JIF & Eigen.	17.8	8.01
All features	44.4	6.65

Altmetrics

- the new approach of citation counts prediction
- first tests shown **the number of Mendeley readers is statistically significant feature**
- it seems to have great potential and could be available much earlier than for example citation counts