Negative Binomial Regression Fit Report

20250610

Project Summary

- We are using data from the American Society of Microbiology's (ASM) 12 published journals to investigate the relationship between the number of citations (variable 'is.referenced.by.count') a published scientific article receives and if the authors have included access to their raw sequencing data (variable 'da', data availability) in the manuscript.
- We are trying to understand if publishing raw data helps to improve citation metrics. We have data from 2000-2024, and will also adjust for time published (variable 'age.in.months'), as older papers have had the opportunity to accumulate more citations over time.

```
opts <- options(knitr.kable.NA = " ")
knitr::kable(all_journals, digits = 3, col.names = gsub("_", " ", names(all_journals)), cap = "log(time) = log(age.in.months), shortened f
```

Table 1: log(time) = log(age.in.months), shortened for ease of reading

coefficients	full	full pvalue	time adj	time adj pvalue	at one year	at one year pvalue	at five years	at five years pvalue	five years	five years pvalue	at ten years	at ten years pvalue	ten years	ten years pvalue
rsquared	0.678		0.681		0.355		0.516		0.660		0.223		0.680	
(Intercept)	- 0.946	0.000	- 0.580	0.000	- 2.334	0.100	2.539	0.000	- 3.473	0.000	3.995	0.000	- 2.696	0.000
da_Yes	- 1.508	0.000	- 0.972	0.000	1.929	0.089	0.258	0.628	- 0.178	0.763	- 0.075	0.840	- 0.267	0.448
log(time)	0.993	0.000	0.928	0.000					1.614	0.000			1.398	0.000
Applied and Environmental Microbiology	- 1.034	0.000	- 0.885	0.000	2.334	0.166	0.198	0.685	- 0.121	0.828	0.046	0.866	- 0.565	0.079
Genome Announcements	- 4.530	0.002	- 4.495	0.001									- 8.881	0.001

coefficients	full	full pvalue	time adj	time adj pvalue	at one year	at one year pvalue	at five years	at five years pvalue	five years	five years pvalue	at ten years	at ten years pvalue	ten years	ten years pvalue
Infection and Immunity	-0.793	0.002	-0.795	0.001	-13.968	0.995	0.234	0.780	-1.329	0.168	-0.337	0.399	-1.180	0.024
Journal of Bacteriology	-1.171	0.000	-1.077	0.000	2.894	0.057	-0.460	0.453	-0.336	0.635	-0.344	0.390	-0.250	0.552
Journal of Clinical Microbiology	0.273	0.196	0.222	0.261	3.433	0.048	-0.188	0.782	-0.678	0.353	-0.846	0.001	1.976	0.000
Journal of Microbiology & Biology Education	-0.226	0.900	-0.221	0.863					1.022	0.701			1.542	0.385
Journal of Virology	-0.437	0.013	-0.355	0.030	2.922	0.052	-0.033	0.946	-0.173	0.755	-0.061	0.826	0.201	0.526
mBio	-1.882	0.000	-1.114	0.000	4.819	0.001			0.012	0.983			-0.266	0.402
Microbiology Resource Announcements	-3.084	0.029	-2.174	0.055	-1.215	0.170	-1.557	0.000	-0.017	0.992			-1.325	0.347
Microbiology Spectrum	-3.555	0.000	-1.706	0.000	2.359	0.100			-1.002	0.059			-1.791	0.000
mSphere	-2.577	0.000	-1.512	0.000	2.334	0.127			-0.222	0.727			-0.827	0.049
mSystems	-2.629	0.000	-1.281	0.000	1.236	0.164			-0.029	0.964			-0.870	0.053
da_Yes:Applied and Environmental Microbiology	0.869	0.002	0.744	0.003	-1.592	0.290	0.154	0.800	-0.191	0.783	0.269	0.548	0.100	0.811
da_Yes:Genome Announcements	1.688	0.259	1.168	0.407									7.219	0.011
da_Yes:Infection and Immunity	1.085	0.011	0.890	0.023	15.067	0.994	-1.239	0.301	0.964	0.430			0.860	0.221
da_Yes:Journal of Bacteriology	1.526	0.000	1.118	0.001	-0.697	0.655	1.074	0.180	0.885	0.358			0.205	0.737
da_Yes:Journal of Clinical Microbiology	0.799	0.027	0.709	0.030	-2.111	0.209	-0.045	0.966	1.110	0.232	0.360	0.521	-1.634	0.002
da_Yes:Journal of Microbiology & Biology Education														

coefficients	full	full pvalue	time adj	time adj pvalue	at one year	at one year pvalue	at five years	at five years pvalue	five years	five years pvalue	at ten years	at ten years pvalue	ten years	ten years pvalue
da_Yes:Journal of Virology	0.802	0.009	0.630	0.023	-1.669	0.226	-0.428	0.521	-0.086	0.907	0.155	0.784	-0.347	0.440
da_Yes:mBio	1.151	0.000	0.901	0.001	-3.880	0.002			-0.234	0.728			-0.149	0.721
da_Yes:Microbiology Resource Announcements	0.621	0.664	0.131	0.909					-1.738	0.304			-0.618	0.666
da_Yes:Microbiology Spectrum	1.027	0.009	0.728	0.017	-1.525	0.189			-0.302	0.648			-0.214	0.647
da_Yes:mSphere	1.366	0.003	1.036	0.009	-1.523	0.255			-0.116	0.881			0.137	0.794
da_Yes:mSystems	0.881	0.062	0.538	0.164					-0.671	0.375			-0.358	0.500
da_Yes:log(time)	0.367	0.000	0.249	0.000					0.066	0.685			0.096	0.249
log(time):Applied and Environmental Microbiology	0.206	0.000	0.179	0.000					-0.016	0.919			0.109	0.145
log(time):Genome Announcements	0.578	0.057	0.572	0.047									1.508	0.012
log(time):Infection and Immunity	0.081	0.098	0.083	0.073					0.248	0.355			0.195	0.113
log(time):Journal of Bacteriology	0.157	0.000	0.141	0.000					-0.010	0.960			-0.057	0.565
log(time):Journal of Clinical Microbiology	-0.093	0.023	-0.083	0.030					0.310	0.128			-0.481	0.000
log(time):Journal of Microbiology & Biology Education	-0.543	0.251	-0.552	0.112					-0.717	0.390			-0.952	0.042
log(time):Journal of Virology	0.069	0.046	0.054	0.098					0.071	0.641			-0.053	0.481
log(time):mBio	0.527	0.000	0.355	0.000					0.098	0.507			0.161	0.033
log(time):Microbiology Resource Announcements	0.393	0.273	0.170	0.564					-0.408	0.376			-0.014	0.969
log(time):Microbiology Spectrum	0.939	0.000	0.420	0.000					0.309	0.040			0.530	0.000
log(time):mSphere	0.622	0.000	0.372	0.000					0.063	0.720			0.217	0.033

coefficients	full	full pvalue	time adj	time adj pvalue	at one year	at one year pvalue	at five years	at five years pvalue	five years	five years pvalue	at ten years	at ten years pvalue	ten years	ten years pvalue
log(time):mSystems	0.719	0.000	0.391	0.000					0.073	0.682			0.312	0.007
da_Yes:log(time):Applied and Environmental Microbiology	-0.170	0.005	-0.142	0.010					0.109	0.569			0.009	0.929
da_Yes:log(time):Genome Announcements	-0.392	0.213	-0.277	0.355									-1.604	0.009
da_Yes:log(time):Infection and Immunity	-0.239	0.008	-0.194	0.020					-0.219	0.519			-0.193	0.248
da_Yes:log(time):Journal of Bacteriology	-0.407	0.000	-0.317	0.000					-0.230	0.388			-0.022	0.880
da_Yes:log(time):Journal of Clinical Microbiology	-0.112	0.152	-0.089	0.208					-0.325	0.209			0.462	0.000
da_Yes:log(time):Journal of Microbiology & Biology Education														
da_Yes:log(time):Journal of Virology	-0.174	0.010	-0.137	0.026					0.033	0.871			0.098	0.364
da_Yes:log(time):mBio	-0.214	0.003	-0.167	0.010					0.109	0.558			0.076	0.452
da_Yes:log(time):Microbiology Resource Announcements	-0.246	0.498	-0.140	0.640					0.386	0.415			0.026	0.944
da_Yes:log(time):Microbiology Spectrum	-0.153	0.172	-0.098	0.262					0.148	0.433			0.118	0.354
da_Yes:log(time):mSphere	-0.293	0.010	-0.232	0.018					0.050	0.818			-0.024	0.851
da_Yes:log(time):mSystems	-0.162	0.179	-0.094	0.347					0.213	0.315			0.109	0.418

How well do the models fit (by Cragg-Uhler pseduo R-squared metric)

- See above table “rsquared” ### Model Formats
- Model format for all data from all journals MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + container.title + container.title*da_factor + log(age.in.months)*da_factor + container.title*log(age.in.months) + log(age.in.months)*da_factor*container.title, data = nsd_yes_metadata, link = log)

- Use model format for data from each journal MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + log(age.in.months)*da_factor, data = <each journal>, link = log)
- **Overall model fit with all data from all journals:**
  - R^2 value = 0.678
  - Removal of top 1% of data: R^2 value = 0.682
  - Truncate data to last 5 years: R^2 value = 0.660
  - Truncate data to last 10 years: R^2 value = 0.680
  - **Summary :** Model fit by R^2 metric does not change by removing the top 1% of data or truncating to data from the last 5 or 10 years.
- **Overall model fit for data from EACH journal individually:**
  - 4/12 journals have **overall model fit** with R^2 > 0.5
  - 4/12 journals have fit with R^2 > 0.5 with **top 1% of data removed**
  - 10/11 journals have model fits >0.5 when **truncated to the last 5 years,** so they are better than their fit overall (one journal has no data from this period)
  - 8/12 journals have model fits >0.5 when **truncated to the last 10 years,** so they are better than their fit overall
  - **Summary:** Data fits negative binomial model better with only more recent data considered.

All journal model is resistant to changes from removing top 1% of data, but less resistant to changes from truncating at 5 and 10 years.

- When working across the columns in the second table, we have coefficients on the left, followed by their values under the following conditions
  - full_model_value = all data included in the model
  - no_1percent_value = top 1% of data removed
  - five_years_value = data truncated at 5 years in age of paper
  - ten_years_value = data truncated at 10 years in age of paper
  - **Note:** Journal of Microbiology and Biology Education(jmbe) has N=7 papers with new sequence data and has been excluded for these analyses, but is a part of the model, and appears as NAs in the table above.

Each journal model are semi-resistant to changes from removing top 1% of data, and even less resistant to changes from truncating at 5 and 10 years.

- See above for mutations on these columns, but these models look less resistant to the transformation of removing the top 1% of data, and even less resistant to changes in coefficients from truncating at 5 and 10 years of data.