Negative Binomial Regression Fit Report

20250610

```
csv_report
```

```
## # A tibble: 13 x 6
##    journal_abrev     n all_data_rsq no_1percent_rsq five_years_rsq ten_years_rsq
##    <chr>         <dbl>        <dbl>           <dbl>          <dbl>         <dbl>
##  1 aac            3237       0.362           0.394          0.556         0.476
##  2 aem            8638       0.486           0.496          0.628         0.620
##  3 genomea        6578       0.0636          0.0636        NA             0.0321
##  4 iai            1854       0.390           0.394          0.598         0.606
##  5 jb             4867       0.315           0.322          0.577         0.558
##  6 jcm            4374       0.188           0.188          0.514         0.388
##  7 jvi            4583       0.417           0.411          0.526         0.517
##  8 mbio           2498       0.668           0.668          0.644         0.663
##  9 mra            5738       0.371           0.371          0.368         0.371
## 10 msphere        1041       0.652           0.651          0.651         0.652
## 11 msystems       1436       0.717           0.704          0.696         0.717
## 12 spectrum       2957       0.542           0.542          0.542         0.542
## 13 all_journals  47808       0.678           0.683          0.660         0.680
```

All Data Together

- The first attempt at fitting the negative binomial regression model to all results
- Use model format 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)
- N = 47, 808 (NSD = Yes papers)
- R^2 value = 0.678
- Next, removal of the top 1% of data to see if model fit changes by filtering
  - filter(is.referenced.by.count < quantile(nsd_yes_metadata$is.referenced.by.count, na.rm = TRUE, prob = 0.99))
  - R^2 value = 0.682
- Truncate data at only data from the last 5 years
  - filter(age.in.months <= 60)
  - R^2 value = 0.660
- Truncate data from the last 10 years
  - filter(age.in.months <= 120)
  - R^2 value = 0.680
- **Summary :** Model fit does not change by removing the top 1% of data or truncating to data from the last 5 or 10 years.

Data by Journal

- Next, looking at data for each journal on its own, removal of container.title variable from model and associated combination terms
- Removal of journal of microbiology and biology education as it only has 7 papers with nsd = yes
- Use model format MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + log(age.in.months)*da_factor, data = <each journal>, link = log)
- **Summary:** 4/12 journals have overall model fit comparable to the full model (R^2 > 0.5)
  - mbio, N = 2498, R^2 = 0.668
  - msphere, N = 1041, R^2 = 0.652
  - msystems, N = 1436, R^2 = 0.717
  - spectrum, N = 2957, R^2 = 0.542
- Remove the top 1% of each journal to see how the model changes
  - **Summary:** Once again, only 4/12 journals have fit comparable to the full model with top 1% of data removed (R^2 > 0.5)
  - mbio, N = 2498, R^2 = 0.668
  - msphere, N = 1041, R^2 = 0.651
  - msystems, N = 1436, R^2 = 0.704
  - spectrum, N = 2957, R^2 = 0.542
- Truncate data at 5 years
  - **Summary:** All but 2 journals (10/12) have model fits >0.5, so they are better than their fit overall when truncated to the last 5 years (see table col five_years_rsq)
  - genome announcements, N = 6578, R^2 = NA (I don't think there were any papers published in this journal in this period)
  - mra, N = 5738, R^2 = 0.368
- Truncate data at 10 years
  - **Summary:** 8/12 journals have model fits >0.5, so they are better than their fit overall when truncated to the last 10 years (see table col ten_years_rsq)
  - aac, N = 3237, R^2 = 0.476
  - genome announcements, N = 6578, R^2 = 0.032
  - jcm, N = 4374, R^2 = 0.388
  - mra, N = 5738, R^2 = 0.371