#20251020 - figuring out why that figure looks crazy

Looking at the predicted number of citations figure for Spectrum

```
#setup dataset and model
nsd_yes_metadata <-
  metadata %>%
  filter(nsd == "Yes") %>%
  filter(., age.in.months != "NA" & da != "NA" & container.title != "NA") %>%
  mutate(da_factor = factor(da),
         container.title = factor(container.title))


nsd_yes_model <-
  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)
```

What are the average citations for spectrum compared to other journals?

What about for papers with data at >= 100 months?

```
#ok let's look at avg citations for spectrum compared to other journals

nsd_yes_metadata  %>%
summarize(mean_citations = mean(is.referenced.by.count),
median_citations = median(is.referenced.by.count), .by =
container.title)
```

```
## # A tibble: 13 x 3
##    container.title                                mean_citations median_citations
##    <fct>                                                   <dbl>            <dbl>
##  1 Antimicrobial Agents and Chemotherapy                    56.7             34
##  2 Applied and Environmental Microbiology                   64.4             40
##  3 Genome Announcements                                      7.11             5
##  4 Infection and Immunity                                   52.7             36
##  5 Journal of Bacteriology                                  50.2             30
##  6 Journal of Clinical Microbiology                         59.6             37
##  7 Journal of Microbiology &amp;amp; Biology Educat~         2.25              2.5
##  8 Journal of Virology                                      49.2             29
##  9 mBio                                                     43.0             22
## 10 Microbiology Resource Announcements                      2.21             1
## 11 mSphere                                                 20.9             12
## 12 mSystems                                                23.8             11
## 13 Microbiology Spectrum                                    5.93             3
```

```
#is it like something weird in papers with ages over 100 months
nsd_yes_metadata %>%
  filter(age.in.months >= 100) %>%
  summarize(mean_citations_100 = mean(is.referenced.by.count),
            median_citations_100 = median(is.referenced.by.count),
            .by = container.title)
```

```
## # A tibble: 10 x 3
##    container.title                          mean_citations_100 median_citations_100
##    <fct>                                                 <dbl>                 <dbl>
##  1 Antimicrobial Agents and Chemotherapy                  73.4                  47
##  2 Applied and Environmental Microbiolo~                  83.1                  57
##  3 Genome Announcements                                    8.26                  5
##  4 Infection and Immunity                                 62.4                  45
##  5 Journal of Bacteriology                                55.5                  36
##  6 Journal of Clinical Microbiology                       64.4                  42
##  7 Journal of Virology                                    65.5                  44
##  8 mBio                                                  101.                   70
##  9 mSphere                                                57.5                  37
## 10 mSystems                                              123.                   54.5
```

What kind of spectrum data was included in this dataset?

- data is included for spectrum for 2021-2024
- and there's nothing older than 41 months

```
#there's only data from 2021-2024
nsd_yes_metadata %>%
  filter(journal_abrev == "spectrum") %>%
  count(year.published)
```

```
## # A tibble: 4 x 2
##   year.published     n
##            <dbl> <int>
## 1           2021   238
## 2           2022   915
## 3           2023  1014
## 4           2024   569
```

```
#ok there's nothing older than 41 months
nsd_yes_metadata %>%
  filter(journal_abrev == "spectrum") %>%
  count(age.in.months) %>%tail()
```

```
## # A tibble: 6 x 2
##   age.in.months     n
##           <dbl> <int>
## 1            31   163
## 2            33   117
## 3            35   111
## 4            37    92
## 5            39    93
## 6            41    53
```

Is there a weird maximum in the spectrum data that's messing with something?

- no the max number of citations is 99

```
#what's the max in spectrum data
nsd_yes_metadata %>%
  filter(journal_abrev == "spectrum") %>%
  .$is.referenced.by.count %>%
  max()
```

```
## [1] 99
```

Based on this- it has to do with the model itself

Getting predicted data out of the model

```
# getting data from the model using get_model_data()

age_values <- seq(5, 120, 5)
  p <-  get_model_data(model = nsd_yes_model, type = "pred",
                    terms = c("da_factor", "age.in.months[age_values]", "container.title"),
                    colors = "bw") %>%
      tibble(da_factor = ifelse(.$x == 1, "Data not available", "Data available"), predicted_citations =
          age.in.months = .$group, container.title = .$facet) %>%
          filter(container.title != "Journal of Microbiology &amp;amp; Biology Education" &
        container.title != "Genome Announcements" &
        container.title != "Microbiology Resource Announcements")

  p %>%
    filter(facet == "Microbiology Spectrum") %>%
    dplyr::select(container.title, da_factor, age.in.months, std.error, conf.low, conf.high) %>%
    print(n = Inf)
```

```
## # A tibble: 48 x 6
##    container.title       da_factor    age.in.months std.error conf.low conf.high
##    <fct>                 <chr>        <fct>              <dbl>    <dbl>     <dbl>
##  1 Microbiology Spectrum Data not av~ 5                 0.155     0.201     0.369
##  2 Microbiology Spectrum Data not av~ 10                0.0923    0.868     1.25
##  3 Microbiology Spectrum Data not av~ 15                0.0596    2.03      2.56
##  4 Microbiology Spectrum Data not av~ 20                0.0438    3.65      4.33
##  5 Microbiology Spectrum Data not av~ 25                0.0415    5.64      6.64
##  6 Microbiology Spectrum Data not av~ 30                0.0475    7.94      9.56
##  7 Microbiology Spectrum Data not av~ 35                0.0565   10.5      13.1
##  8 Microbiology Spectrum Data not av~ 40                0.0661   13.3      17.3
##  9 Microbiology Spectrum Data not av~ 45                0.0754   16.5      22.1
## 10 Microbiology Spectrum Data not av~ 50                0.0842   19.8      27.6
## 11 Microbiology Spectrum Data not av~ 55                0.0923   23.5      33.7
## 12 Microbiology Spectrum Data not av~ 60                0.1000   27.4      40.5
## 13 Microbiology Spectrum Data not av~ 65                0.107    31.5      47.9
## 14 Microbiology Spectrum Data not av~ 70                0.114    35.9      56.0
## 15 Microbiology Spectrum Data not av~ 75                0.120    40.5      64.8
## 16 Microbiology Spectrum Data not av~ 80                0.126    45.3      74.3
## 17 Microbiology Spectrum Data not av~ 85                0.132    50.4      84.5
## 18 Microbiology Spectrum Data not av~ 90                0.137    55.8      95.3
## 19 Microbiology Spectrum Data not av~ 95                0.142    61.3     107.
## 20 Microbiology Spectrum Data not av~ 100               0.147    67.1     119.
## 21 Microbiology Spectrum Data not av~ 105               0.151    73.0     132.
```

```
## 22 Microbiology Spectrum Data not av~ 110            0.155    79.2     146.
## 23 Microbiology Spectrum Data not av~ 115            0.160    85.6     160.
## 24 Microbiology Spectrum Data not av~ 120            0.164    92.3     175.
## 25 Microbiology Spectrum Data availa~ 5              0.0844    0.179     0.249
## 26 Microbiology Spectrum Data availa~ 10             0.0496    0.853     1.04
## 27 Microbiology Spectrum Data availa~ 15             0.0317    2.12      2.40
## 28 Microbiology Spectrum Data availa~ 20             0.0232    4.00      4.38
## 29 Microbiology Spectrum Data availa~ 25             0.0225    6.48      7.08
## 30 Microbiology Spectrum Data availa~ 30             0.0262    9.53     10.6
## 31 Microbiology Spectrum Data availa~ 35             0.0315   13.2      14.9
## 32 Microbiology Spectrum Data availa~ 40             0.0369   17.4      20.1
## 33 Microbiology Spectrum Data availa~ 45             0.0421   22.1      26.1
## 34 Microbiology Spectrum Data availa~ 50             0.0470   27.5      33.1
## 35 Microbiology Spectrum Data availa~ 55             0.0515   33.5      41.0
## 36 Microbiology Spectrum Data availa~ 60             0.0558   40.1      49.9
## 37 Microbiology Spectrum Data availa~ 65             0.0597   47.3      59.7
## 38 Microbiology Spectrum Data availa~ 70             0.0634   55.0      70.6
## 39 Microbiology Spectrum Data availa~ 75             0.0669   63.4      82.5
## 40 Microbiology Spectrum Data availa~ 80             0.0702   72.4      95.4
## 41 Microbiology Spectrum Data availa~ 85             0.0732   82.1     109.
## 42 Microbiology Spectrum Data availa~ 90             0.0762   92.3     124.
## 43 Microbiology Spectrum Data availa~ 95             0.0789  103.      141.
## 44 Microbiology Spectrum Data availa~ 100            0.0816  115.      158.
## 45 Microbiology Spectrum Data availa~ 105            0.0841  127.      176.
## 46 Microbiology Spectrum Data availa~ 110            0.0865  139.      196.
## 47 Microbiology Spectrum Data availa~ 115            0.0888  153.      216.
## 48 Microbiology Spectrum Data availa~ 120            0.0910  167.      238.
```