Regression models for narrator analysis

Adrian Barnett

02 September 2022

# Regression model

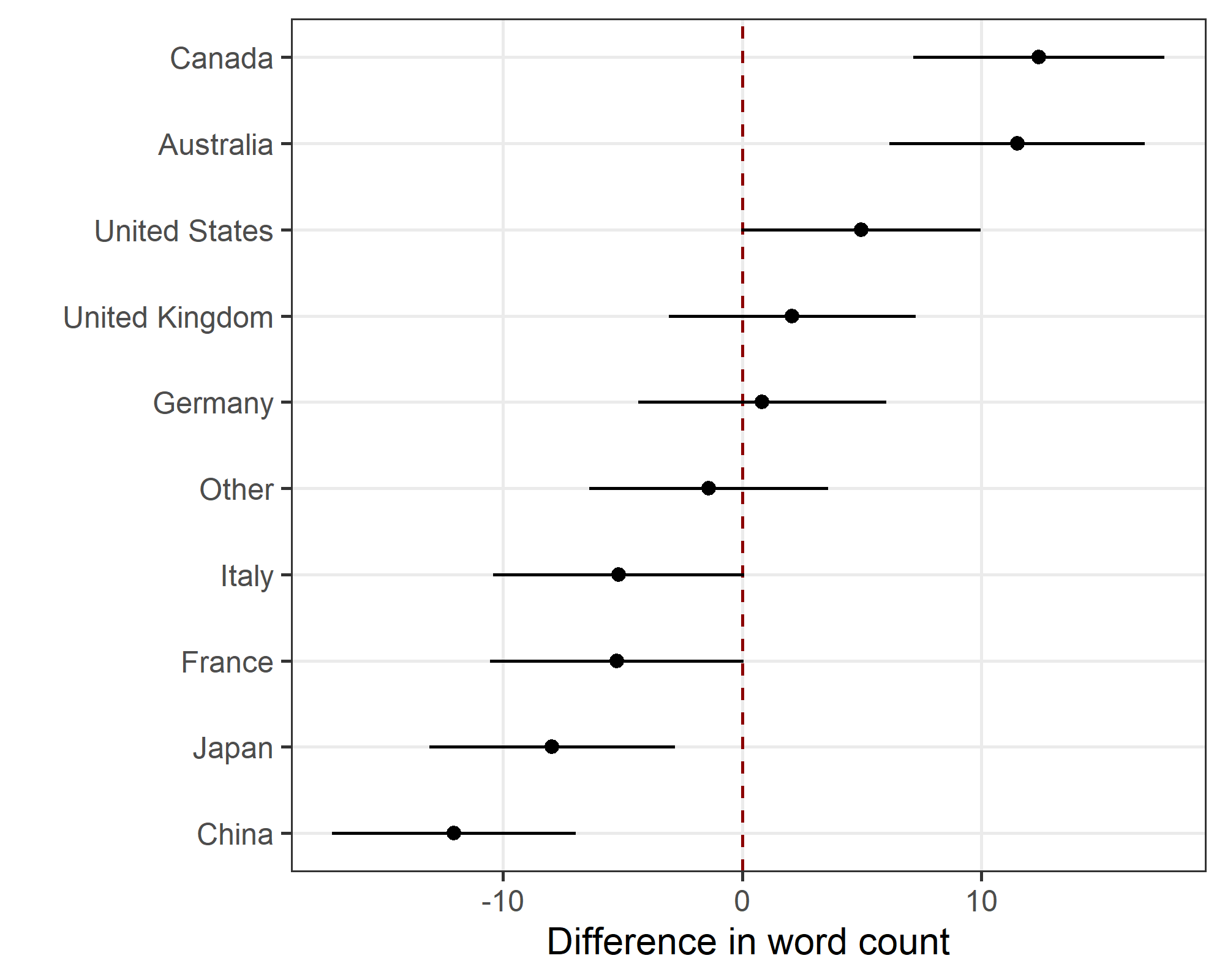
| **term** | **estimate** | **conf.low** | **conf.high** |
| --- | --- | --- | --- |
| (Intercept) | 83.3 | 50.2 | 116.5 |
| bs(author\_transform, df = 3)1 | 194.4 | 181.6 | 207.2 |
| bs(author\_transform, df = 3)2 | 160.4 | 137.9 | 182.9 |
| bs(author\_transform, df = 3)3 | -104.1 | -164.6 | -43.7 |
| Year (+10 years) | 12.2 | 11.8 | 12.5 |

We used a regression model to examine the number of words in the abstract.

The regression model used random effects for country and article type (plotted below). The effect of author numbers was strongly non-linear and so used a spline. These are the “bs” terms which are visualised below.

The table shows the estimated means and 95% confidence interval. The effect for year is per 10 year increase. So every 10 years, the average abstract increases by 12 words.

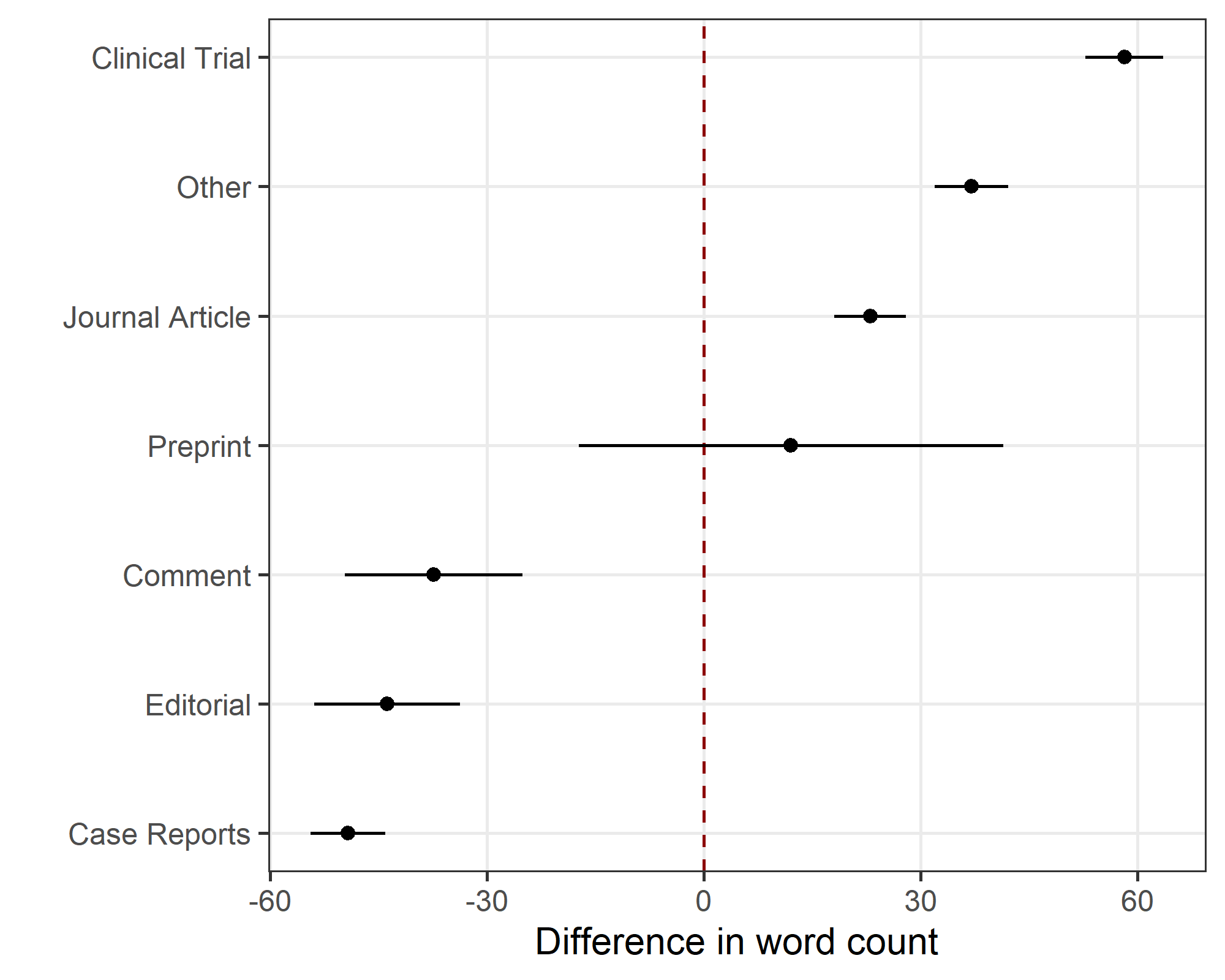
# Effect for country



The plot shows the mean difference in word count and 95% confidence interval.

English-speaking countries tend to have longer abstracts.

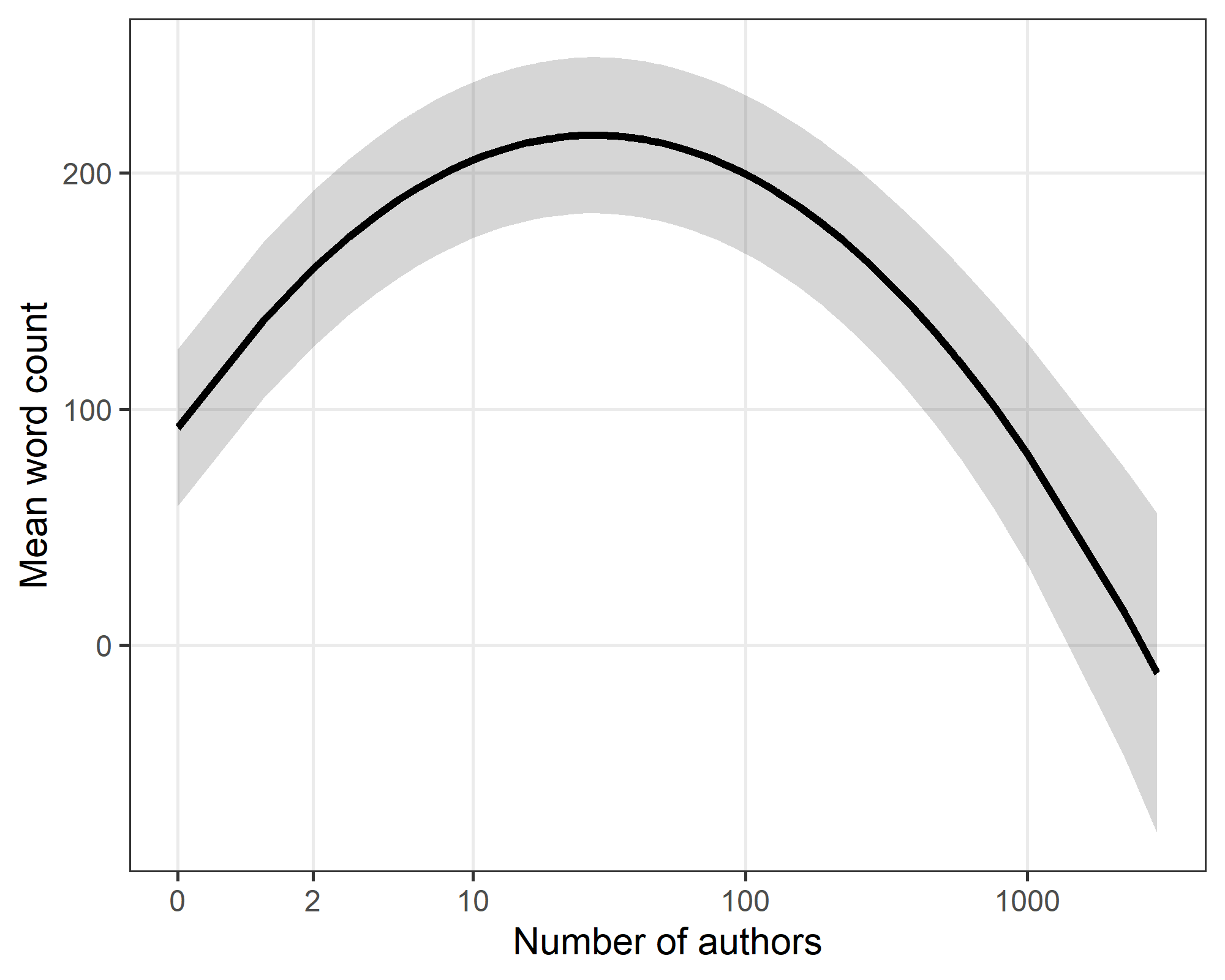
# Effect for article type



The plot shows the mean difference in word count and 95% confidence interval.

Clinical trials have the longest abstracts on average.

# Word count by author numbers

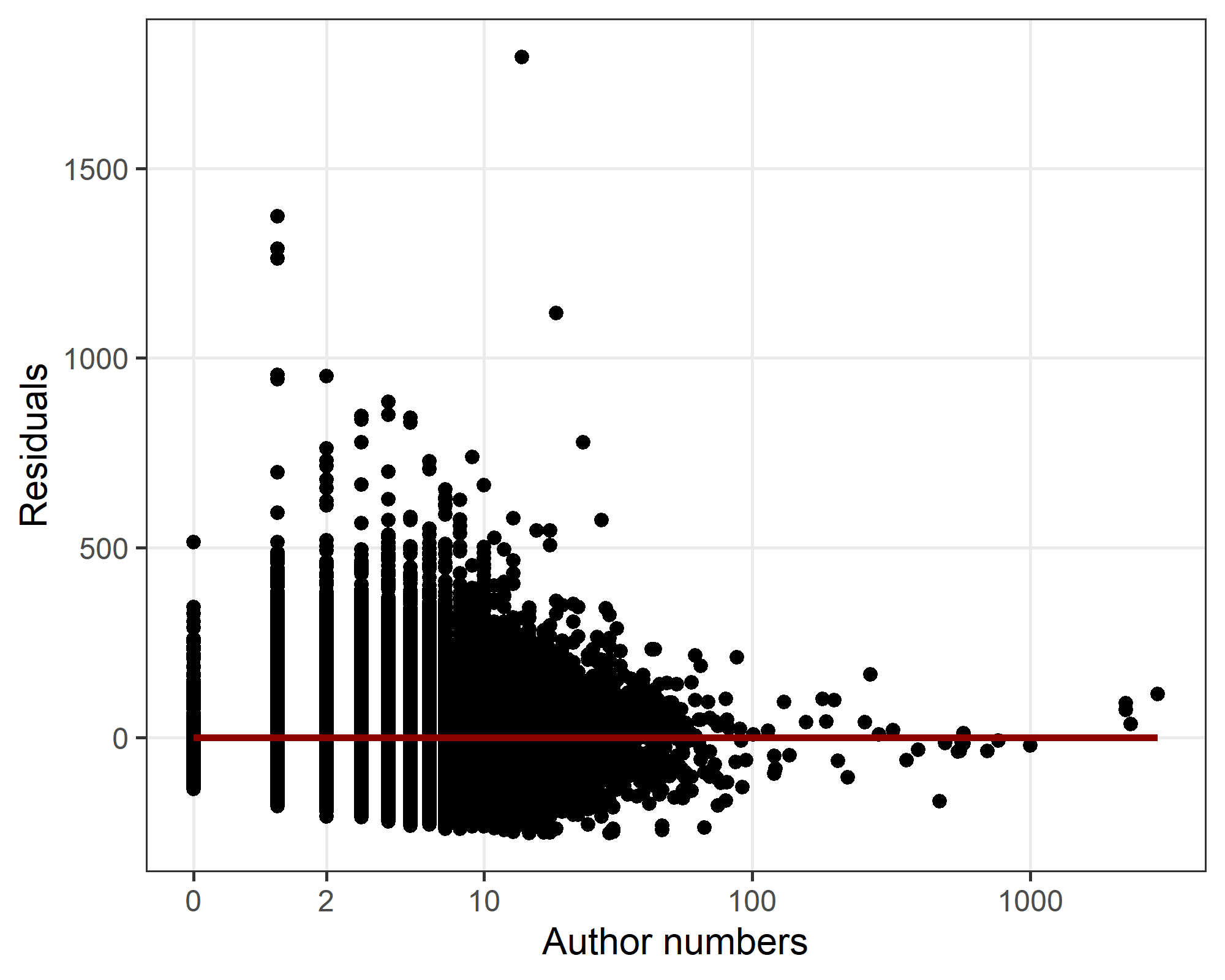


The axis for the number of authors is on a log-scale.

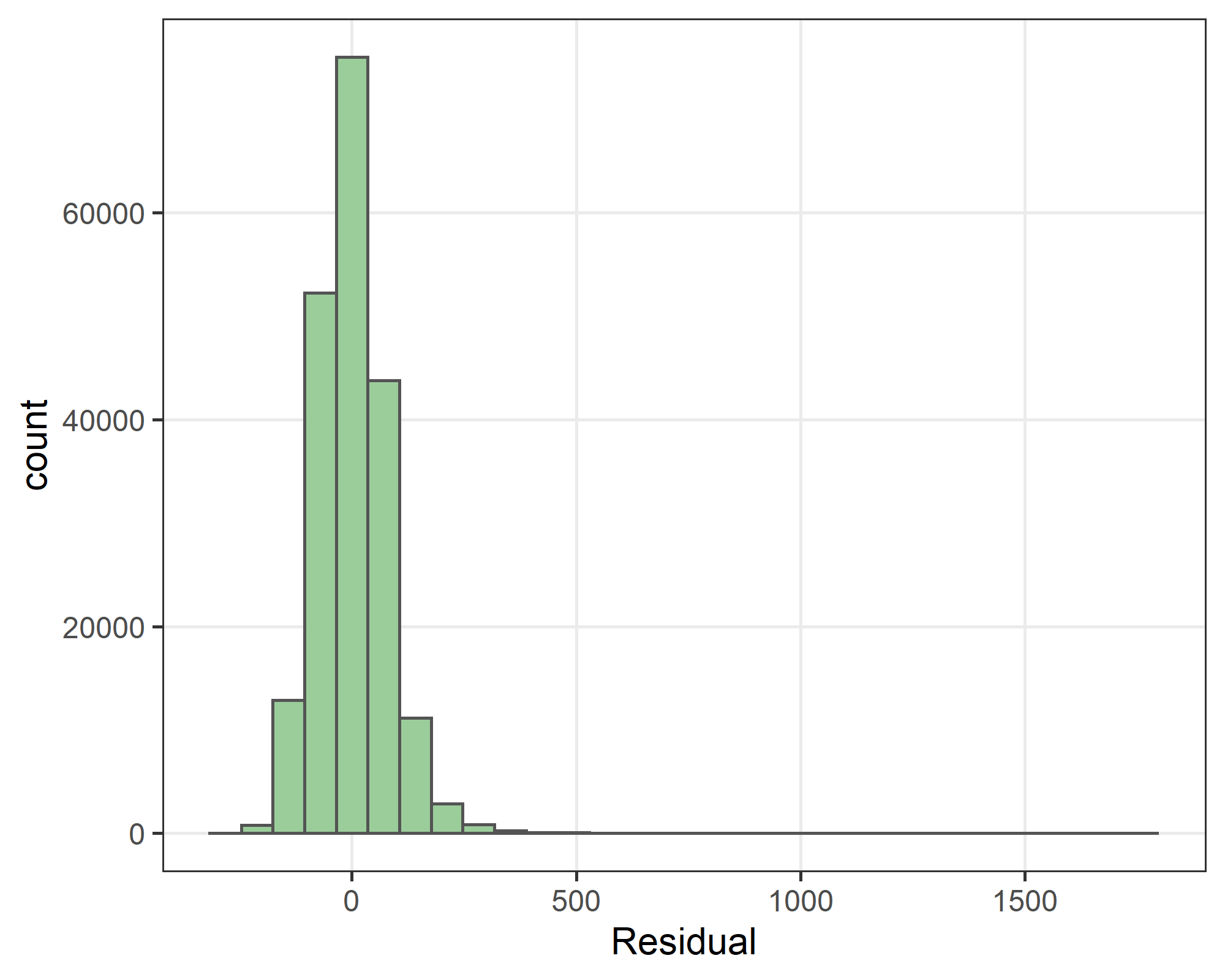
The longest abstracts have 28 authors (on average).

# Residuals

## Residuals by author numbers



## Residuals histogram



There are a small number of very long abstracts, for example this one which over 1500 words: <https://pubmed.ncbi.nlm.nih.gov/23074487/>. These outliers have a large positive residual, but are unlikely to over-influence the model given the large size of the data.