Examining the association between grant application panel scores and subsequent citations

Adrian Barnett

8 October 2018

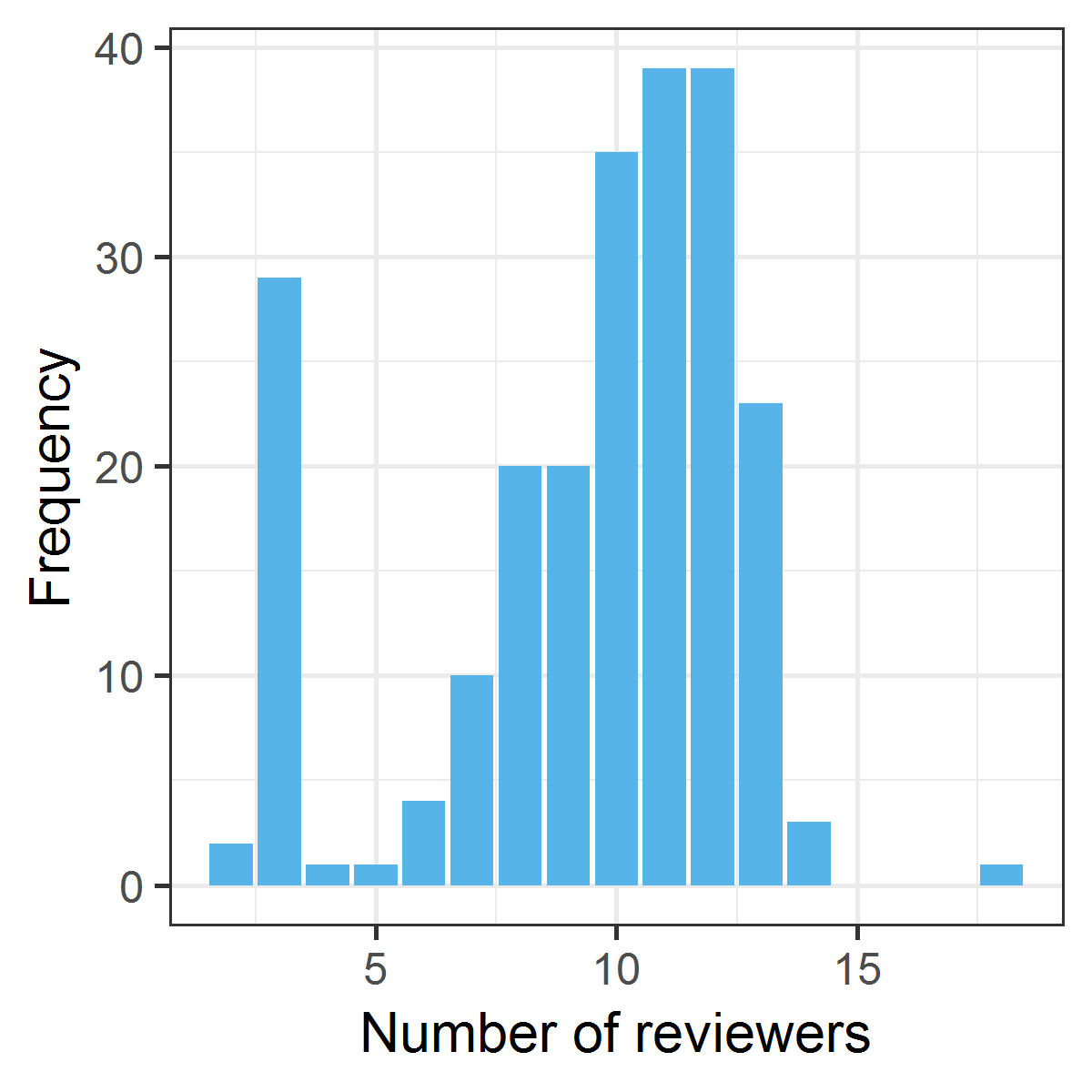
## Introduction

The data are 227 successful grant applications for the years 1999 to 2006 from the *American Institute of Biological Sciences*. The outcome is total relative citations (larger is better). The key predictor is the application score, which ranges from 1 (best) to 5 (worst). Funding decisions typically use the mean score to rank applications, but there may be value in also using the standard deviation in a score which reflects an application where there was some disagreement between panel members.

### Missing data

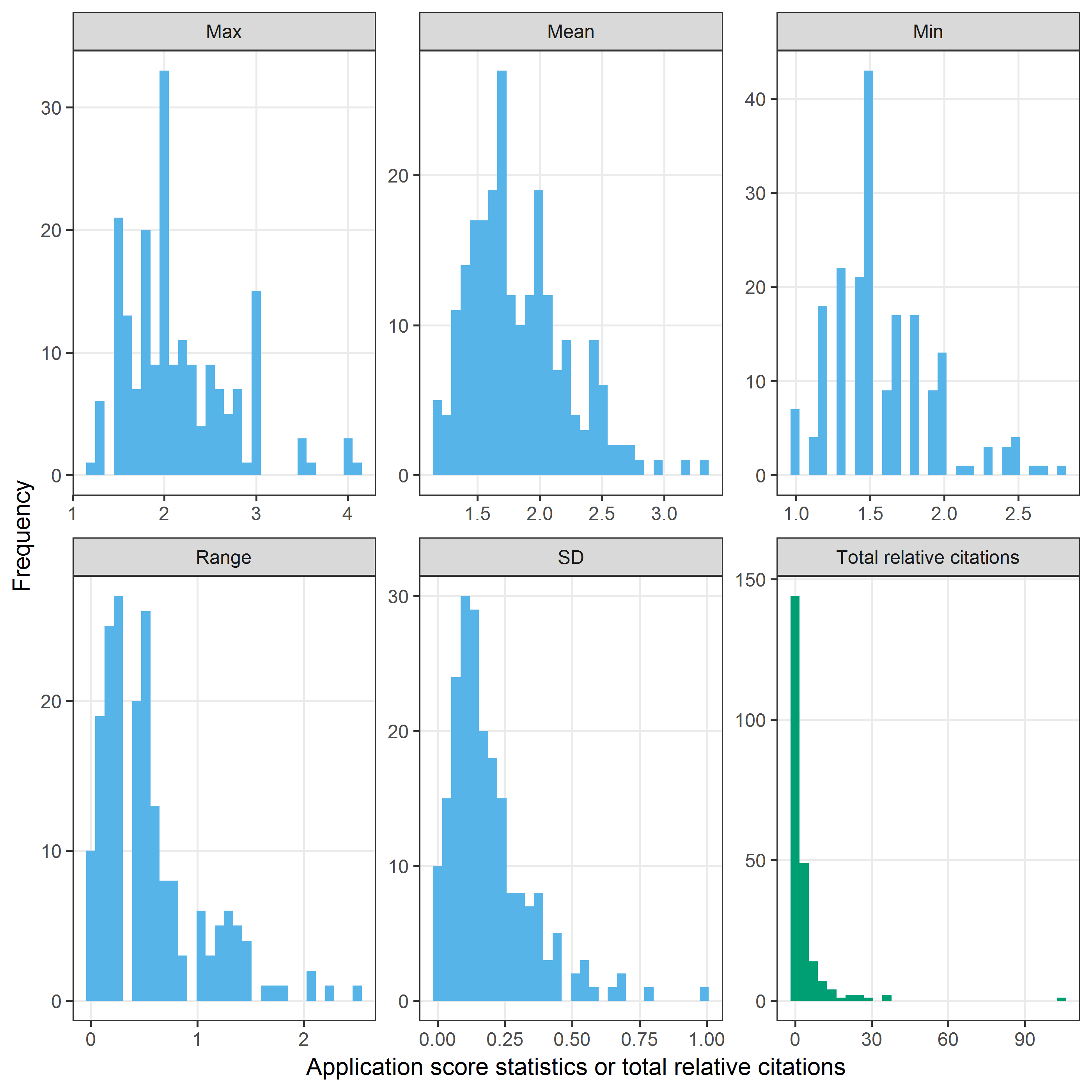
16 (7%) observations were missing the score standard deviation and 32 (14%) observations were missing the score range. These missing observations were imputed before any analysis.

### Bar plot of reviewer numbers



This is a bimodal distribution with a mode at 3 reviewers for the teleconference reviews and second mode at 11 to 12 for the larger panels.

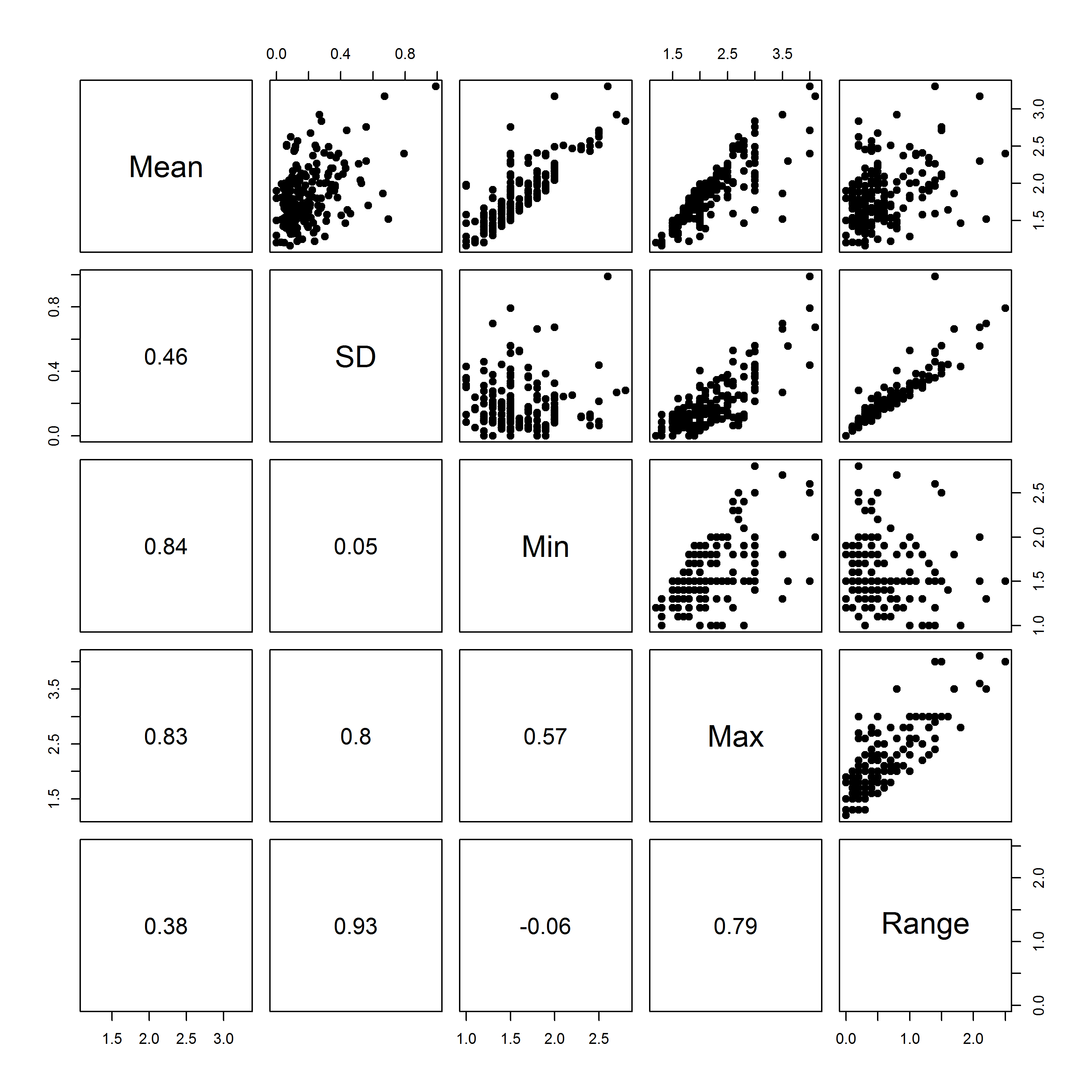
#### Histograms of the application score statistics and citations (primary outcome)



The histograms above show the mean, standard deviation, minimum, maximum and range (maximum minus minimum). The lower the mean score, the better the application did in peer review. The standard deviation and range have a positive skew.

The main outcome is strongly positively skewed and there was one outlier with a total relative citation of over 90. There were 74 (33%) grants with zero citations.

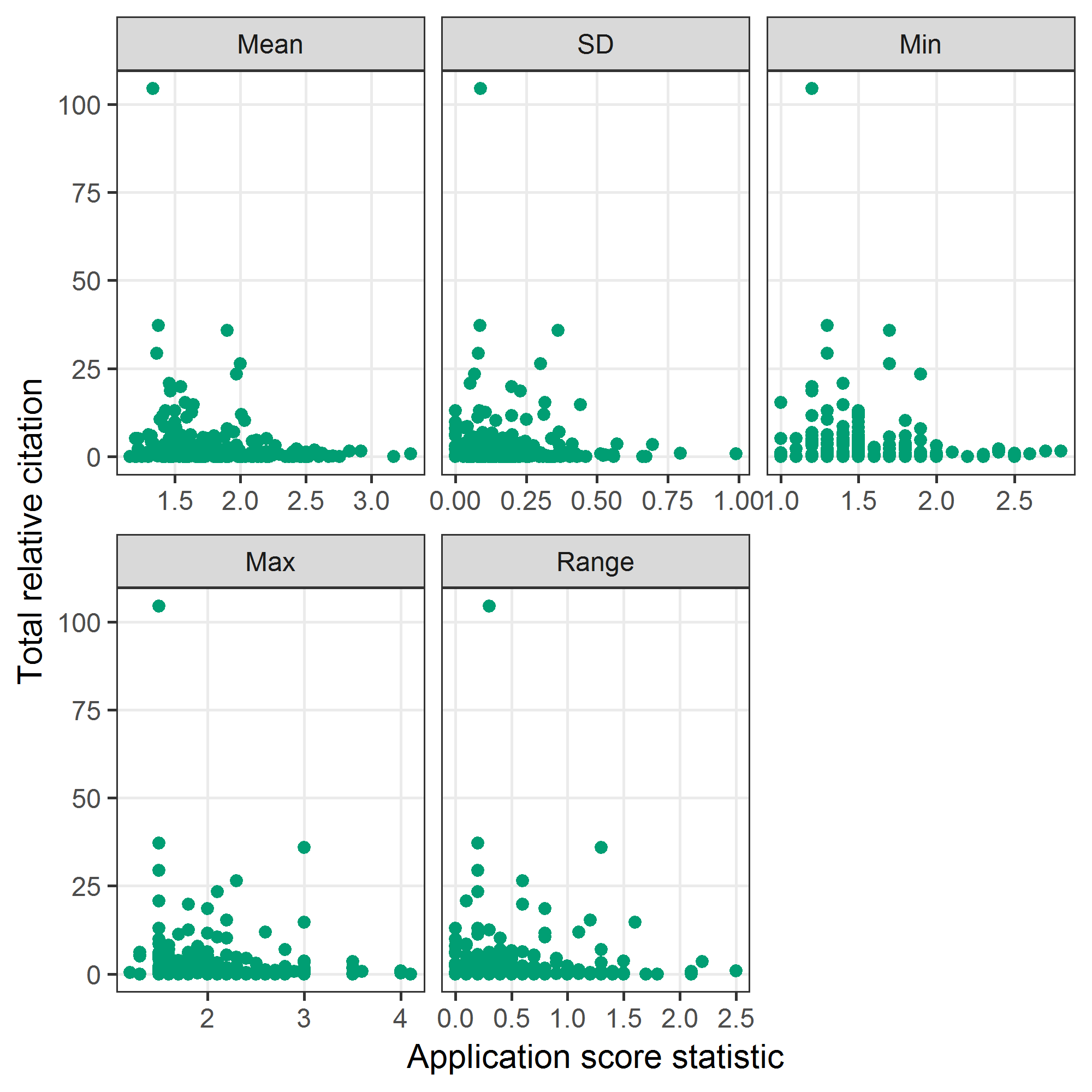
#### Scatter-plots and correlations between application score statistics



The numbers in the bottom-left half of the plot matrix are the Pearson correlations. We expect some positive correlation between a score’s mean and standard deviation (0.46). There is a relatively strong correlation between the standard deviation and maximum (0.80), but not between the standard deviation and minimum (0.05). This suggests the largest disagreement is where at least one panel member has given a poor score. Applications where there was one dissenting panel member with a good score were probably not funded as their mean score would not be competitive, and hence are not in this sample.

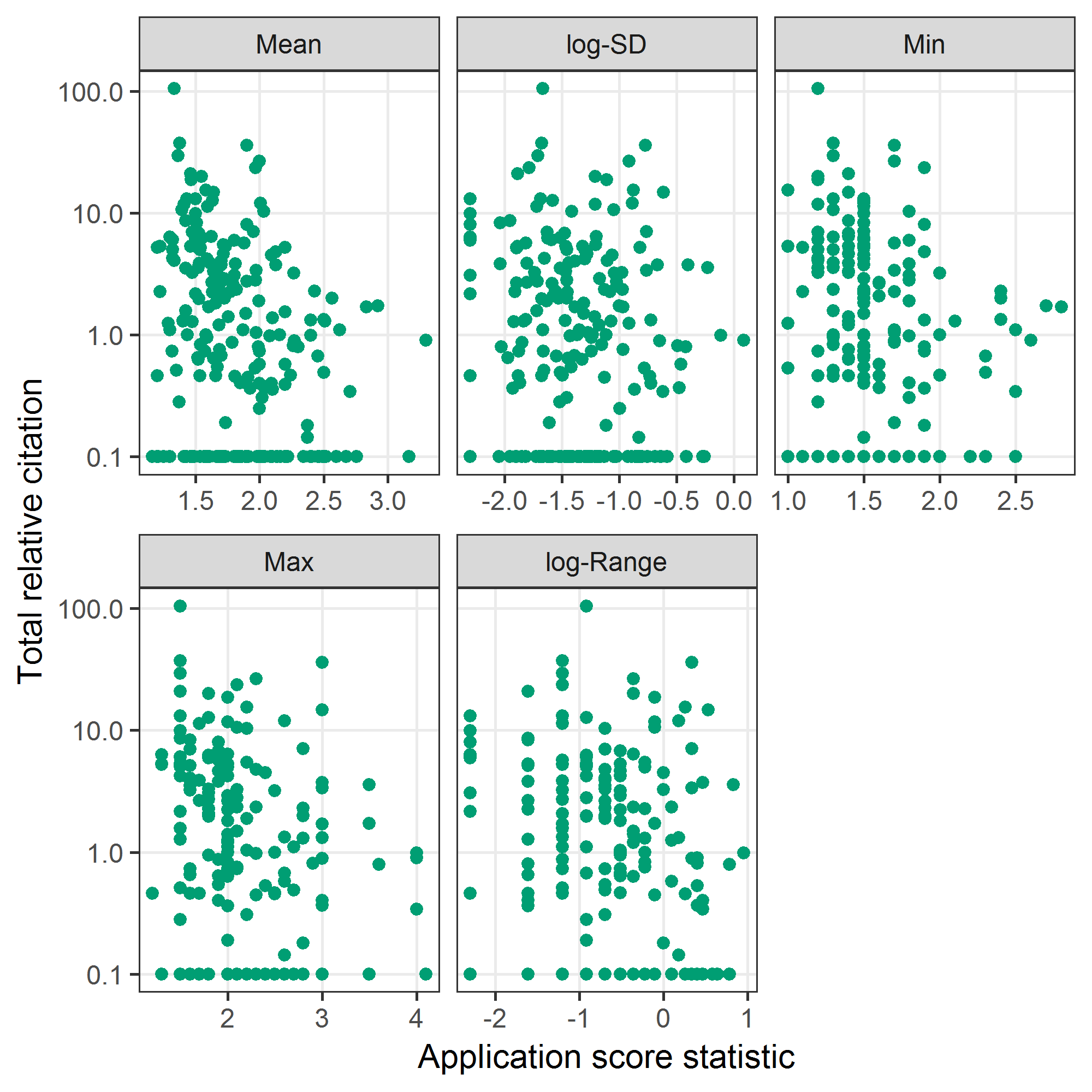
There is a strong positive correlation between the standard deviation and the range (0.93), and both are measures of panel disagreement.

### Scatter plots of citations against application score statistics



The top-left panel showing the mean is a repeat of Figure 2 from the *PLOS ONE* paper that used this data to examine the association between mean score and citations (Gallo et al, 2014).

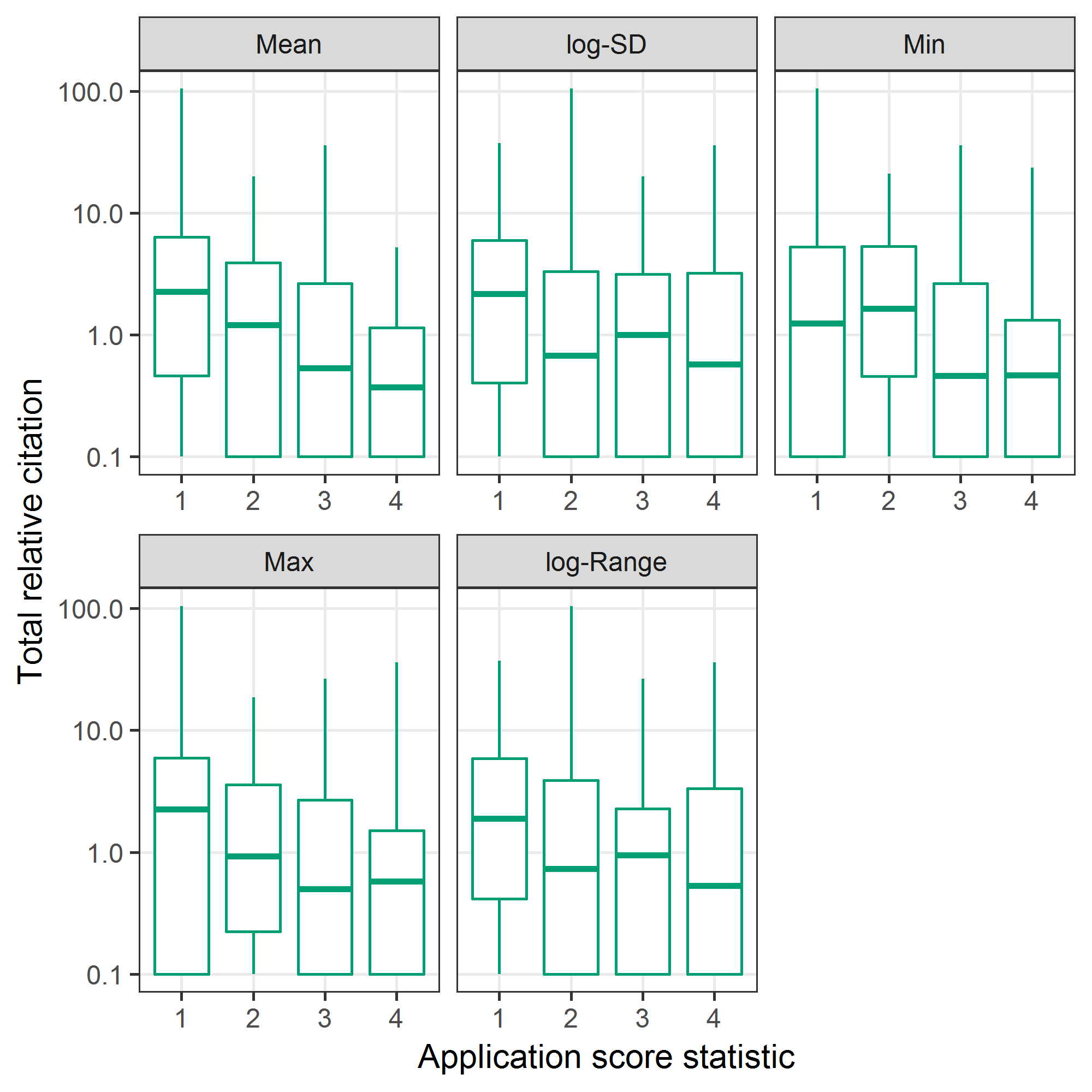
### Scatter plots of log-citations against application score statistics



The above scatter plot uses log-transformed citations in an attempt to show a clearer association with the score statistics. Some association between mean score and citations is visible, with a generally downward pattern in citations for increasing score.

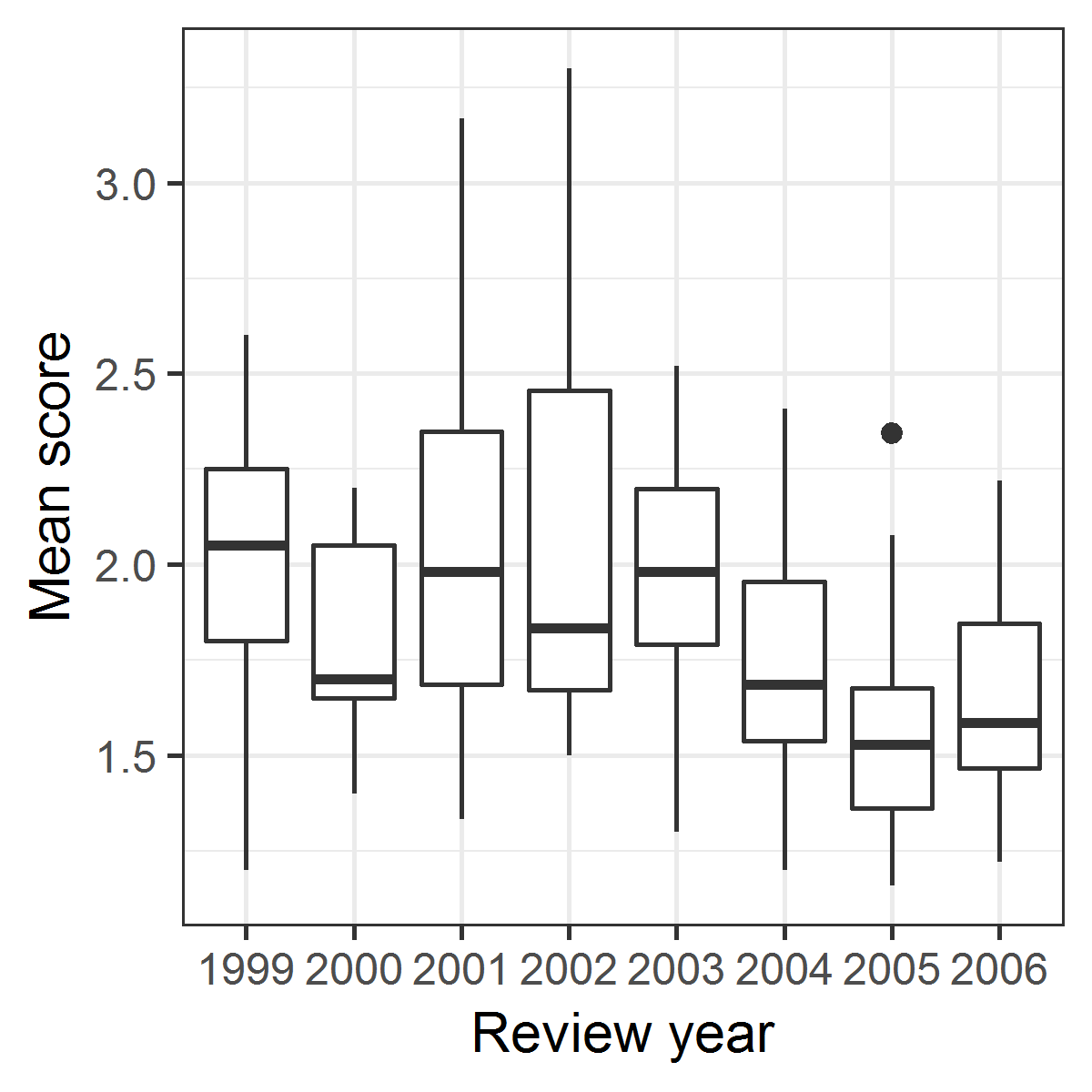
#### Checking for constant variance

The score statistics have been grouped into quintiles to check the variance.

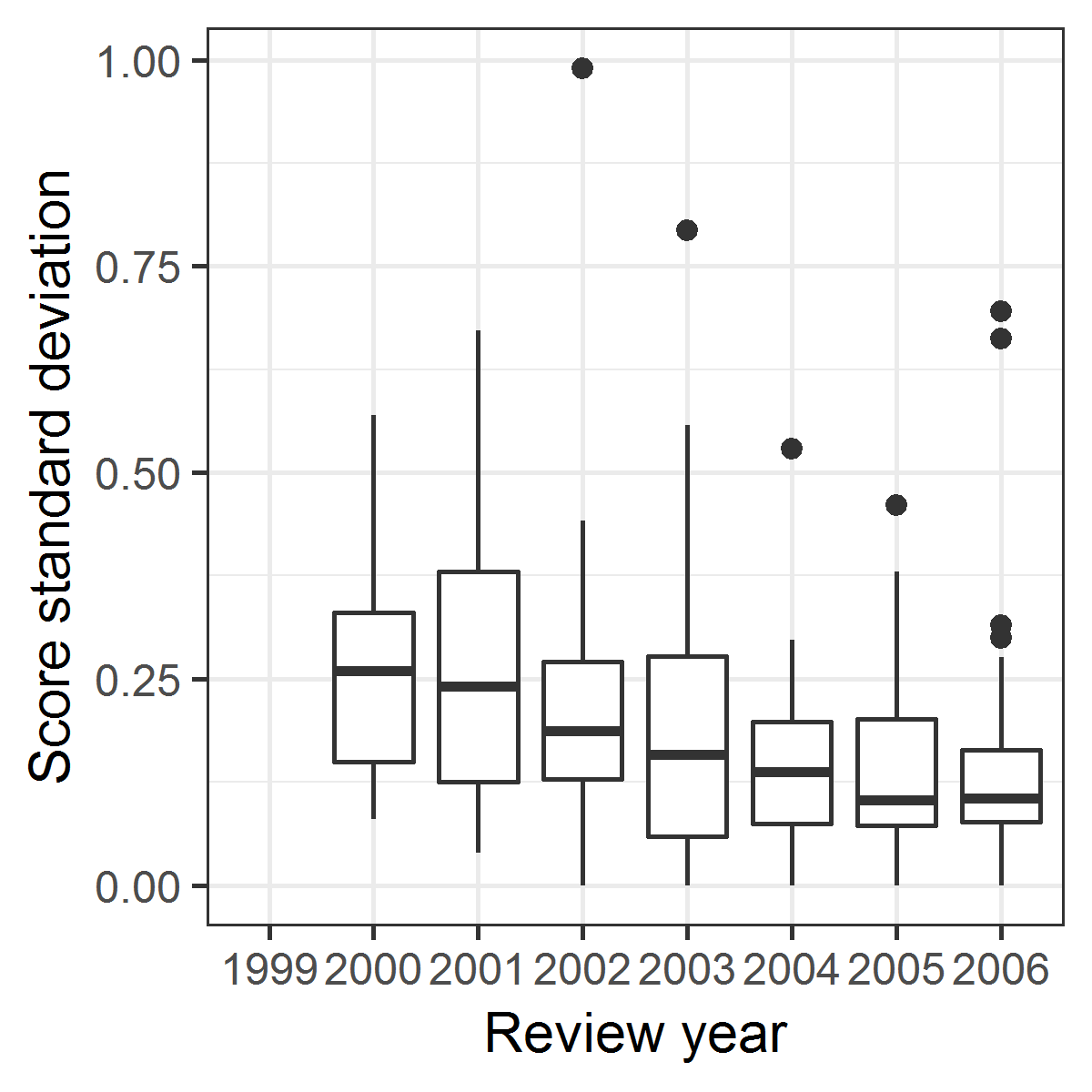


#### Changes in application scores over time

##### Mean scores



##### Score standard deviation



There has been some reduction in the mean and standard deviation over time. Hence it would be wise for our regression models to adjust for review year. This change was highlighted in the original *PLOS ONE* paper and the plot of the mean score over time was also in that paper.

## Regression models

The total relative citations were modelled using a multiple regression model. The citations were first log-transformed because of their positive skew. We added a small positive constant of 0.1 prior to using the log-transform because some citations were zero. The predictions were back-transformed to show the estimates on the original citations scale.

We used fractional polynomials to examine a range of non-linear associations between the scores and citations.

Using equations the regression model was:

The log-transformed citations (*Y*) were modeled using a Normal distribution. The mean () had a constant and the application score statistics (*X*) which were first transformed by the fractional polynomial function. The random intercept () adjusted for a potential within-panel correlation in citations.

The 16 missing score standard deviations and 32 missing ranges were accounted for using multiple imputation (details in later section). We also show results that excluded these data using a complete case analysis.

### Model with just mean score (complete case data)

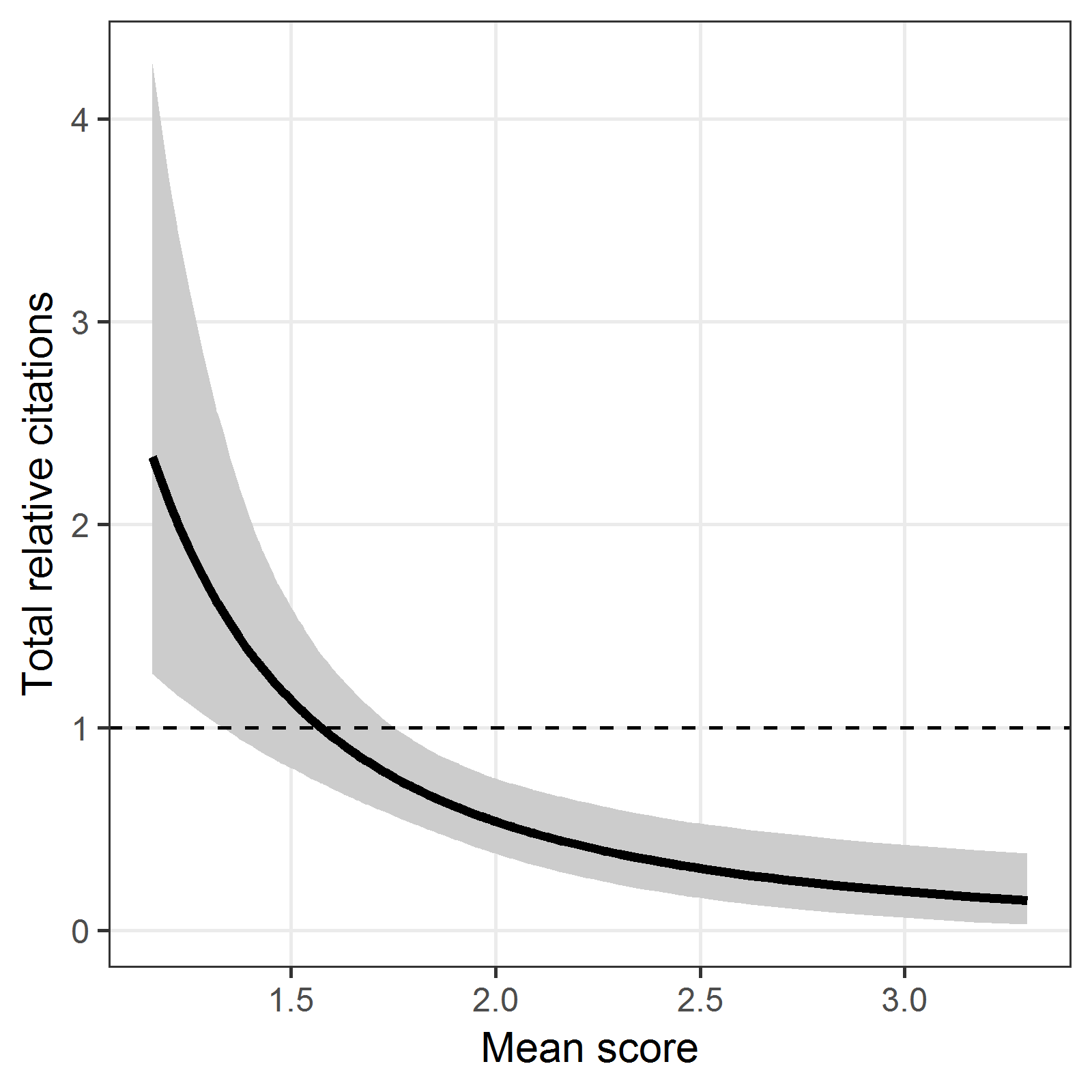
In an initial analysis, we fit just the mean application score to confirm the findings from the previous paper.

#### Model fit - deviance (complete case data)

|  |  |  |
| --- | --- | --- |
| Fractional polynomial | Deviance | Difference in deviance |
| -2 | 886.9 | 3.2 |
| -1 | 885 | 1.3 |
| -0 | 883.6 | 0 |
| 0 | 885.4 | 1.7 |
| 0 | 884.6 | 0.99 |
| 1 | 887 | 3.3 |
| 2 | 891.1 | 7.4 |
| 3 | 895.5 | 12 |

The table shows the lowest deviance is for a fractional polynomial of -0.5, although the difference between the three best models is small.

#### Predictions from best model (complete case data)



The plot shows the predicted relative citations using the best fractional polynomial model. The solid line is the mean and the grey area is a 95% confidence interval.

#### Table of parameter estimates (complete case data)

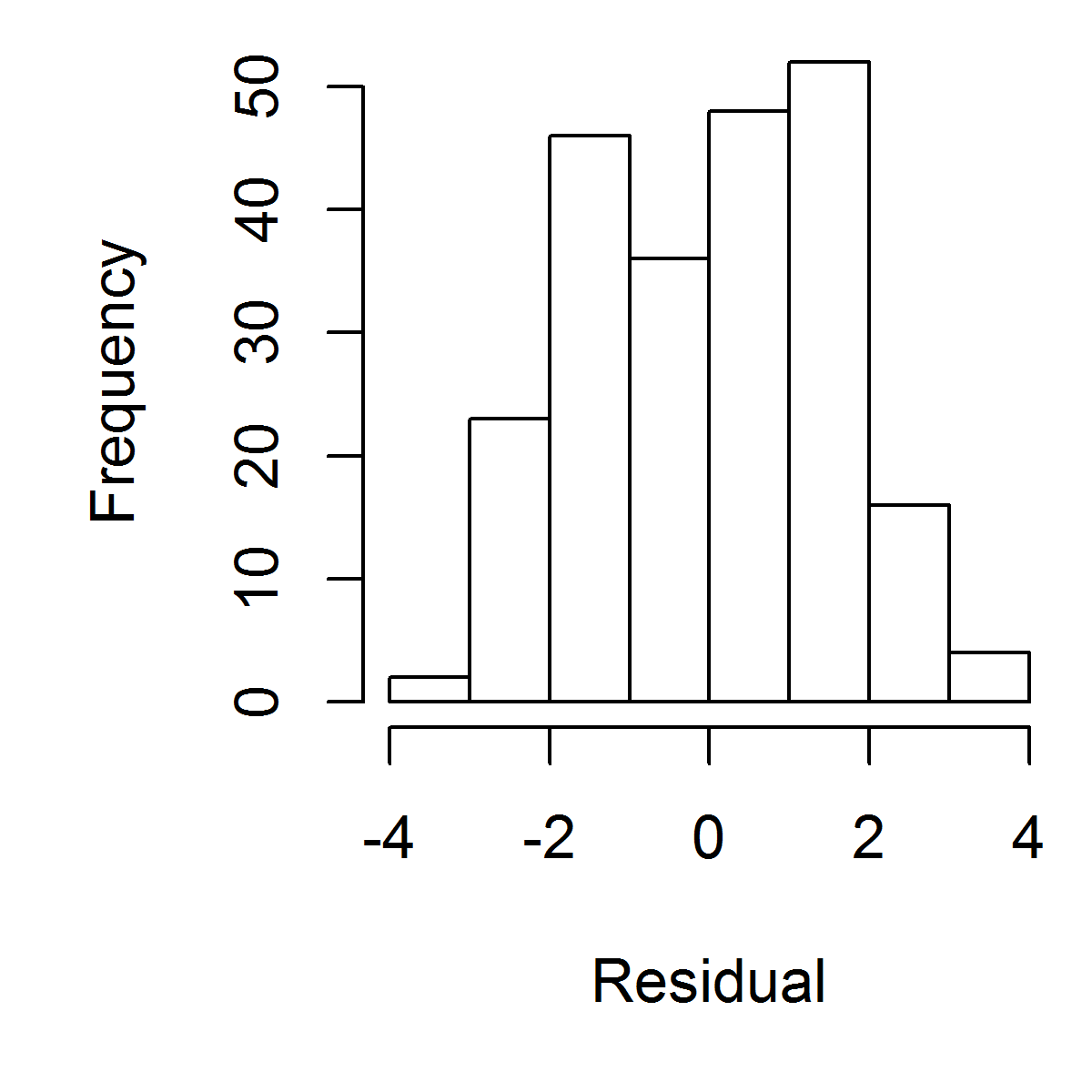
========================================  
 Dependent variable:   
 --------------------  
 log(`Web TRC` + 0.1)  
----------------------------------------  
Mean score, f(-0.5) 6.1 (3.1 to 9.0)   
Intercept -4.7 (-7.0 to -2.5)   
----------------------------------------  
Observations 227   
Log Likelihood -441.8   
========================================  
Note:

The table shows the mean and 95% confidence interval for the parameter estimates (). The “f()” in the variable label is the best fractional polynomial transformation. There was a strongly statistically significant effect of the mean score as the 95% confidence interval does not contain zero. However, it is difficult to interpret these estimates because of the fractional polynomial transformation. Hence it is better to interpret the effect using the previous figure.

The within-panel correlation was 0.11, signifying a small positive correlation in the citations of applications assessed by the same panel.

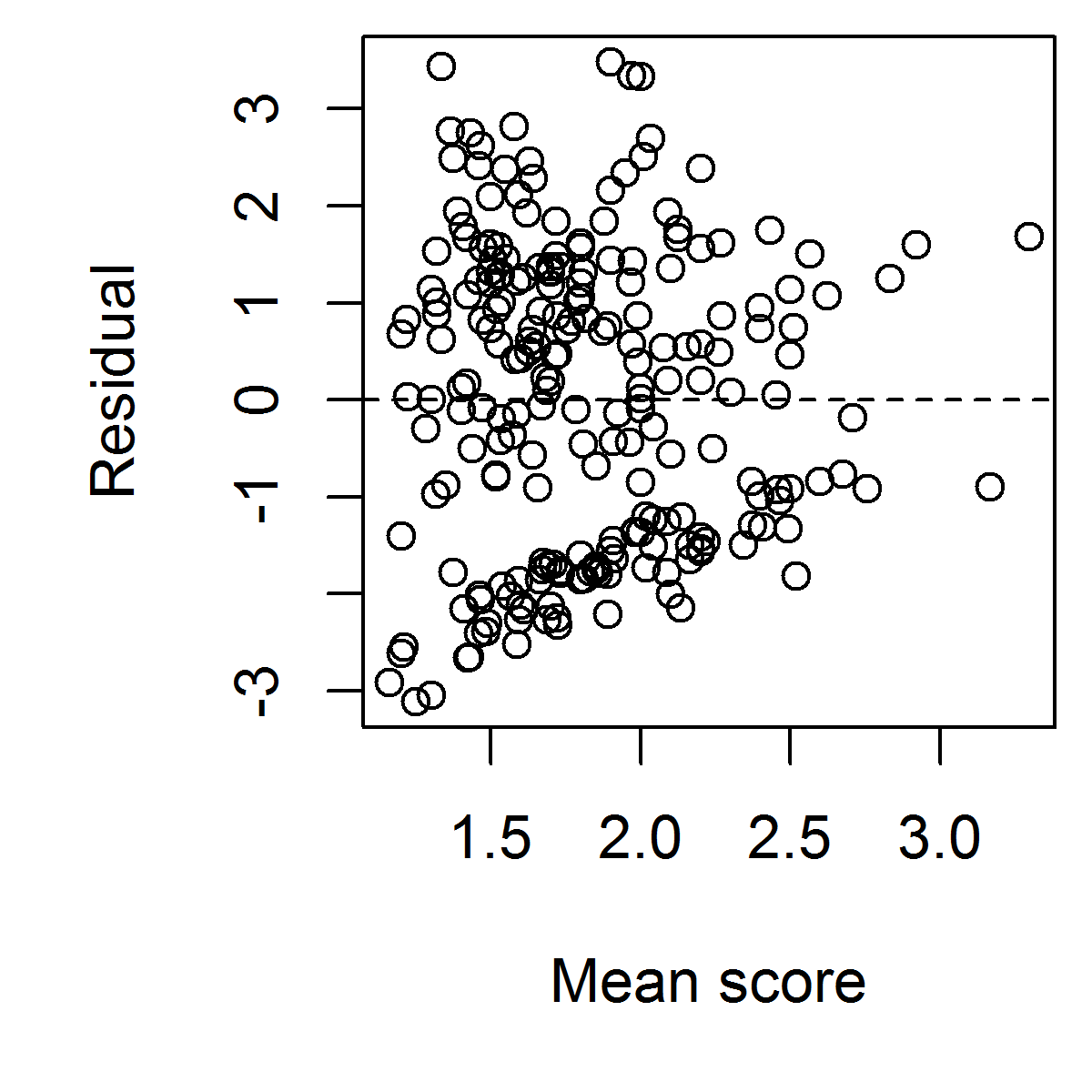
#### Model checking (complete case data)

##### Histogram of residuals (complete case data)

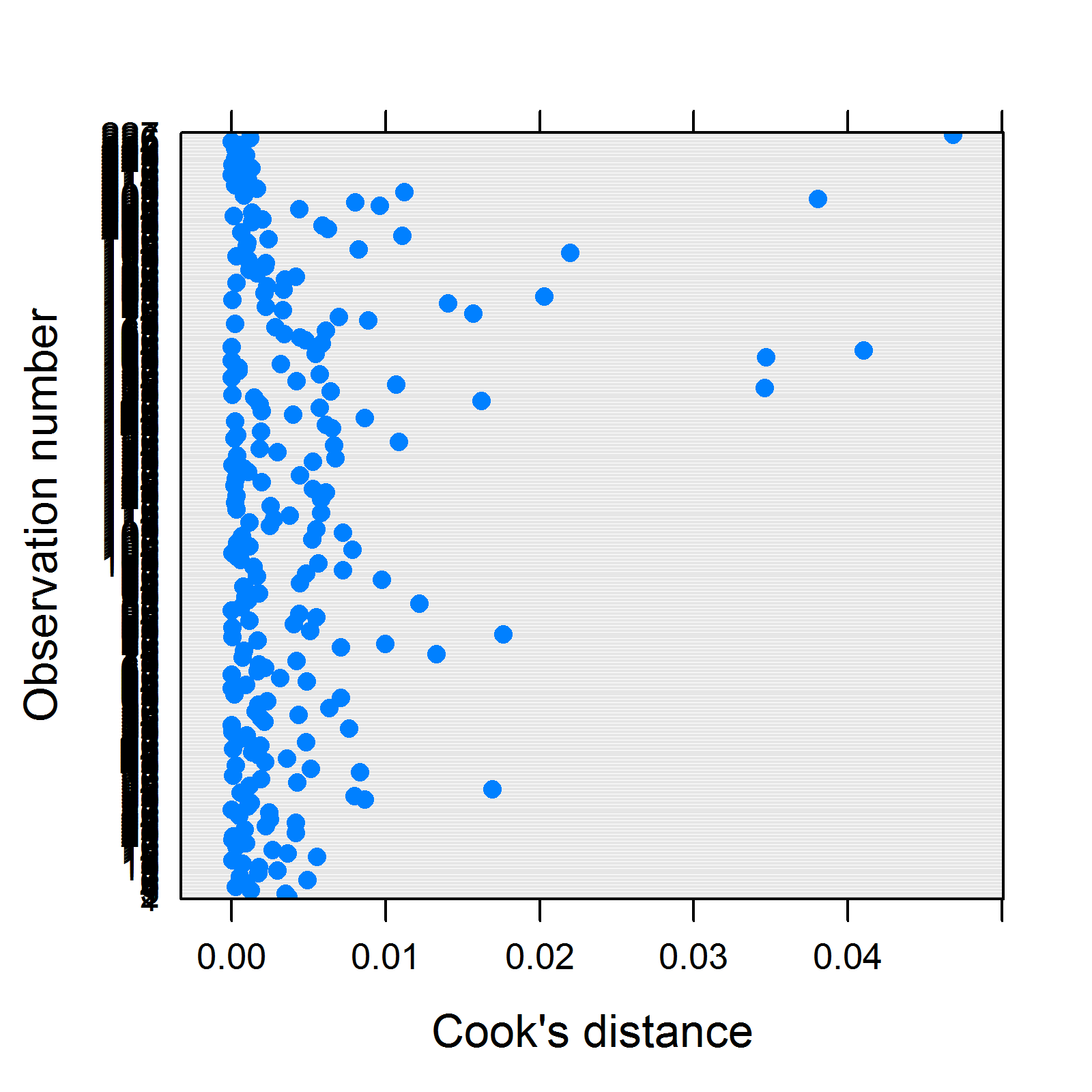


There are no clear outliers in the residuals.

##### Scatter plot of residuals against mean (complete case data)



##### Cook’s distance for finding influential observations (complete case data)



### Models for standard deviation and range (complete case data)

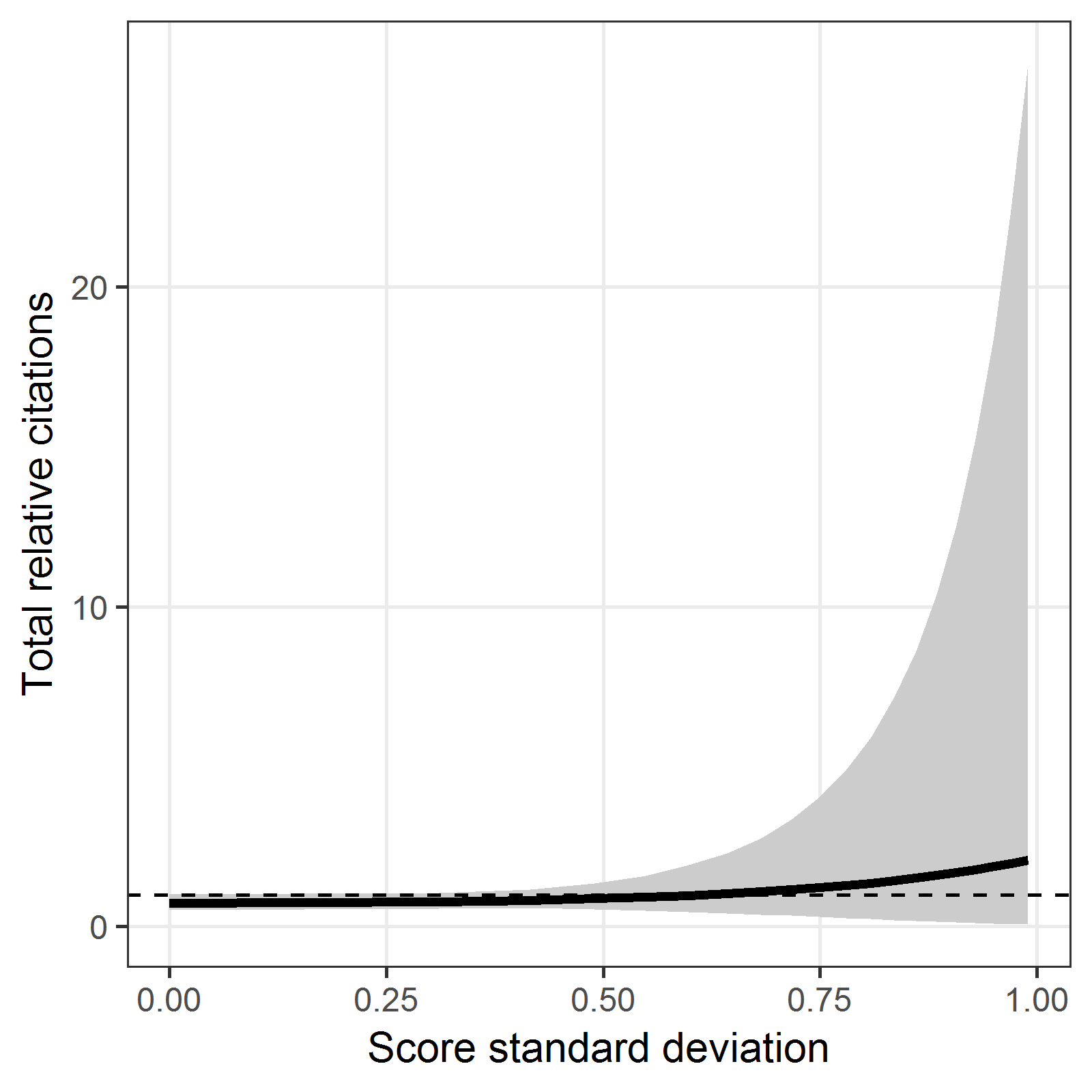
Now we examine the alternative predictors of the application score standard deviation and range using the complete case data.

#### Model fit - deviance (complete case data)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FP1 | FP2 | FP3 | Deviance | Difference in deviance |
| -0 | 3 | -2 | 820.7 | 0 |
| -0 | 0 | -2 | 821.1 | 0.42 |
| -0 | 2 | -2 | 821.2 | 0.46 |
| -0 | 1 | -2 | 821.5 | 0.81 |
| -0 | 3 | -0 | 821.6 | 0.87 |

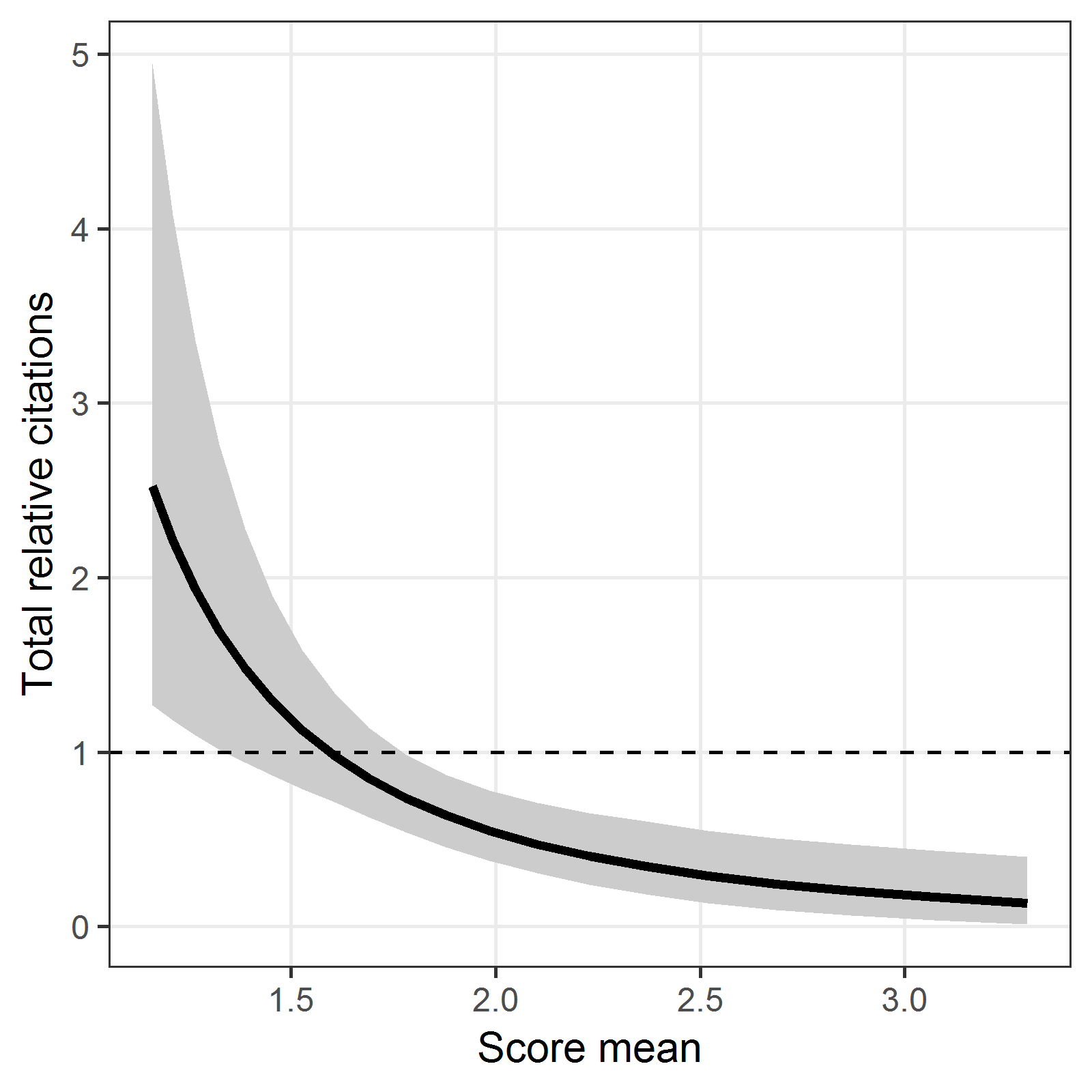
FP1 = application mean score, FP2 = application score standard deviation, FP3 = review year.

#### Predictions for score standard deviation (complete case)



The plot shows the predicted relative citations using the best fractional polynomial model. The solid line is the mean and the grey area is a 95% confidence interval.

#### Predictions for mean score (complete case)



The plot shows the predicted relative citations using the best fractional polynomial model. The solid line is the mean and the grey area is a 95% confidence interval.

#### Table of parameter estimates (complete case data)

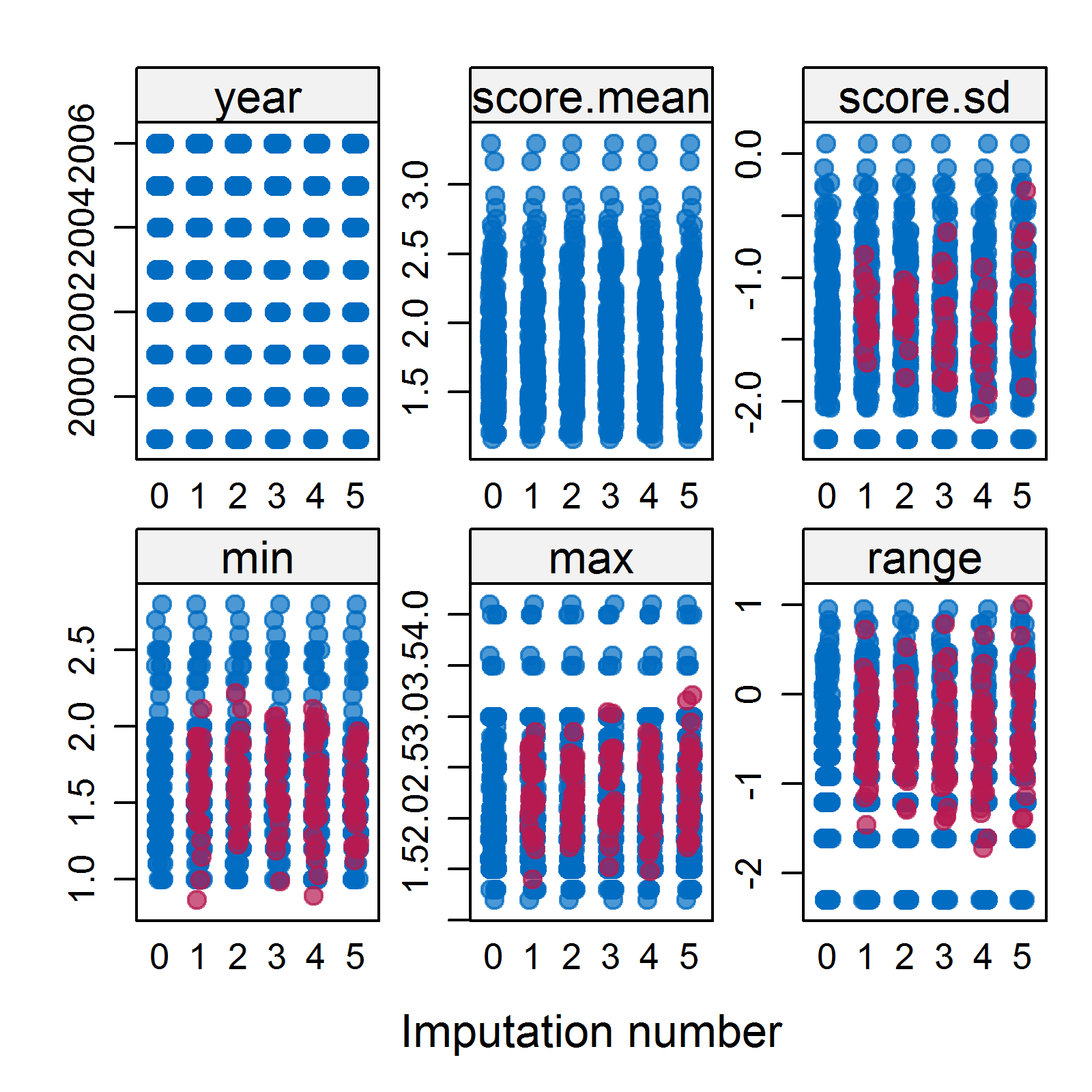
========================================  
 Dependent variable:   
 --------------------  
 log(`Web TRC` + 0.1)  
----------------------------------------  
Mean score, f(-0.5) 6.3 (2.9 to 9.8)   
Score SD, f(3) 0.7 (-1.3 to 2.8)   
Review year, f(-2) -1.5 (-5.7 to 2.8)   
Intercept -4.9 (-7.6 to -2.2)   
----------------------------------------  
Observations 211   
Log Likelihood -410.4   
========================================  
Note:

This table can be compared to the later table that uses the imputed data. Overall, there was very little difference in the estimates between the results using the complete case and imputed data.

The “f()” in the variable label is the best fractional polynomial transformation.

# Imputing missing standard deviations and ranges

We used regression imputation to impute the missing standard deviations and ranges based on the mean, minimum and maximum score. We use the *mice* library in R. We log-transformed the standard deviation and range because of their positive skew. We created five imputed data sets. The plot shows the observed (blue) and imputed (red) observations.



### Model with mean score and score standard deviation (imputed data)

#### Model fit - Top five deviances (FP = Fractional Polynomial) (imputed data)

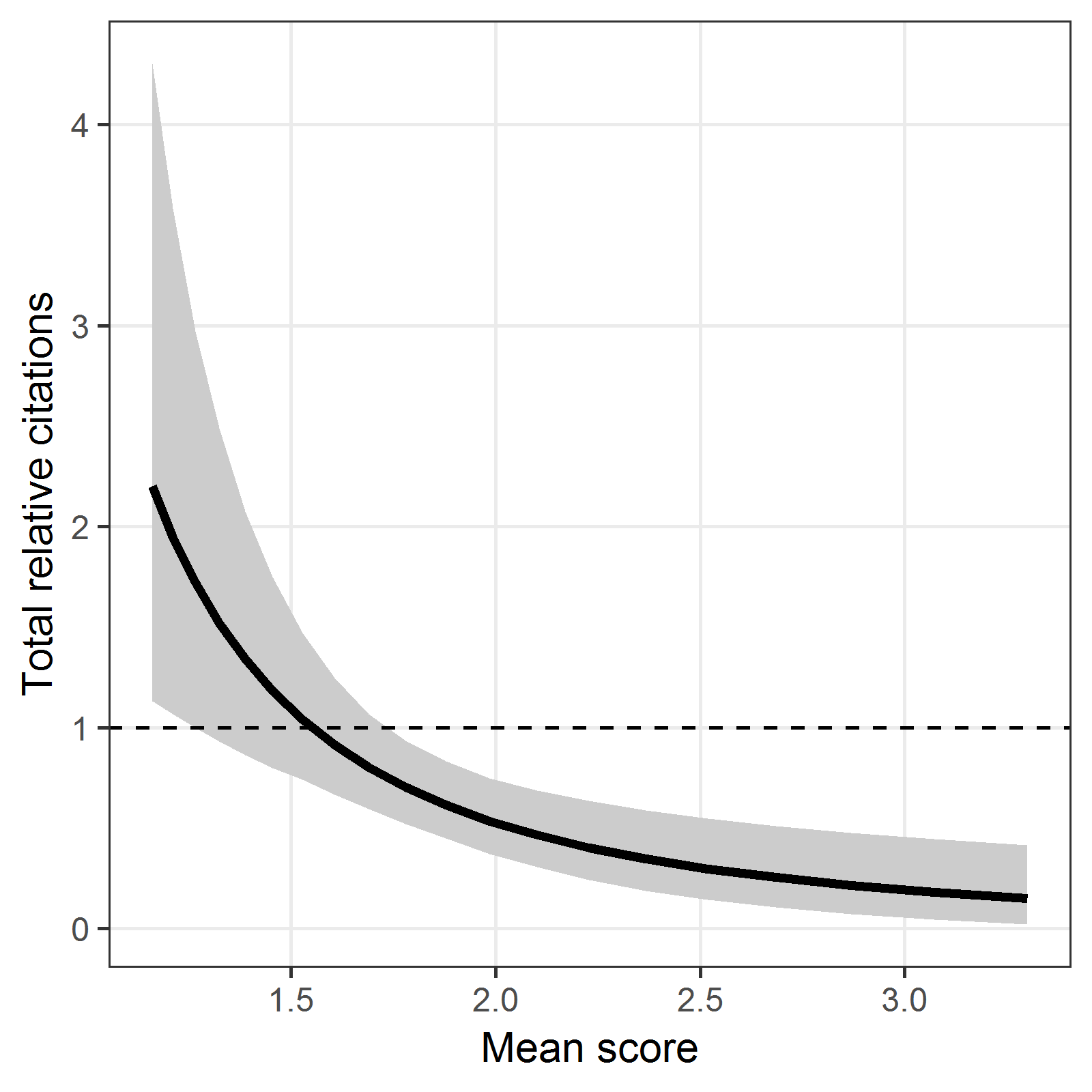
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FP1 | FP2 | FP3 | Deviance | Difference in deviance |
| -0 | 3 | -0 | 878 | 0 |
| -0 | 0 | -0 | 878.4 | 0.4 |
| -0 | 3 | -1 | 878.4 | 0.44 |
| -0 | 2 | -0 | 878.4 | 0.46 |
| 0 | 3 | -0 | 878.6 | 0.65 |

FP1 = application mean score, FP2 = application score standard deviation, FP3 = review year.

As before the best fractional polynomial for the mean was -0.5.

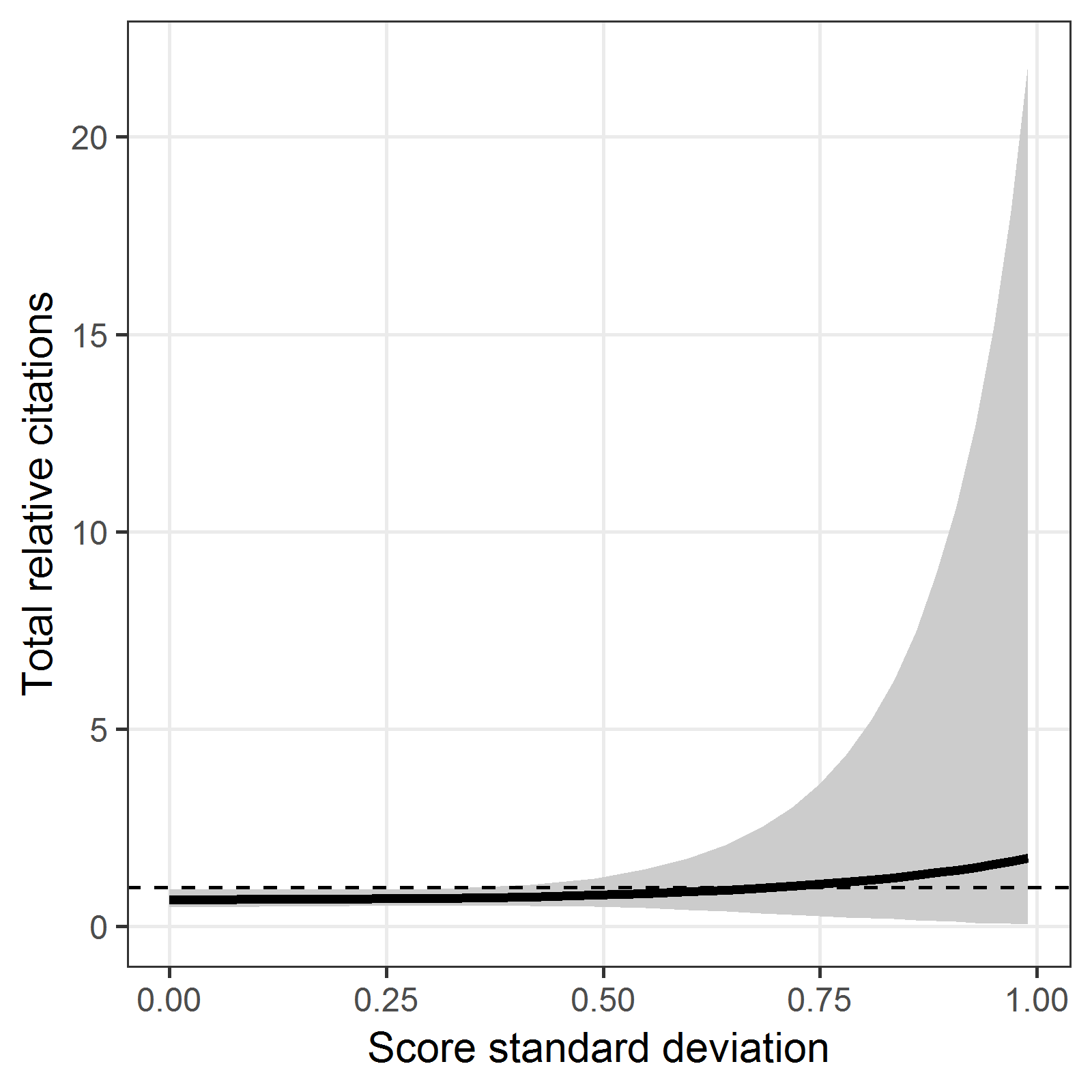
#### Predictions from best model (imputed data)

##### Predictions for mean score (imputed data)



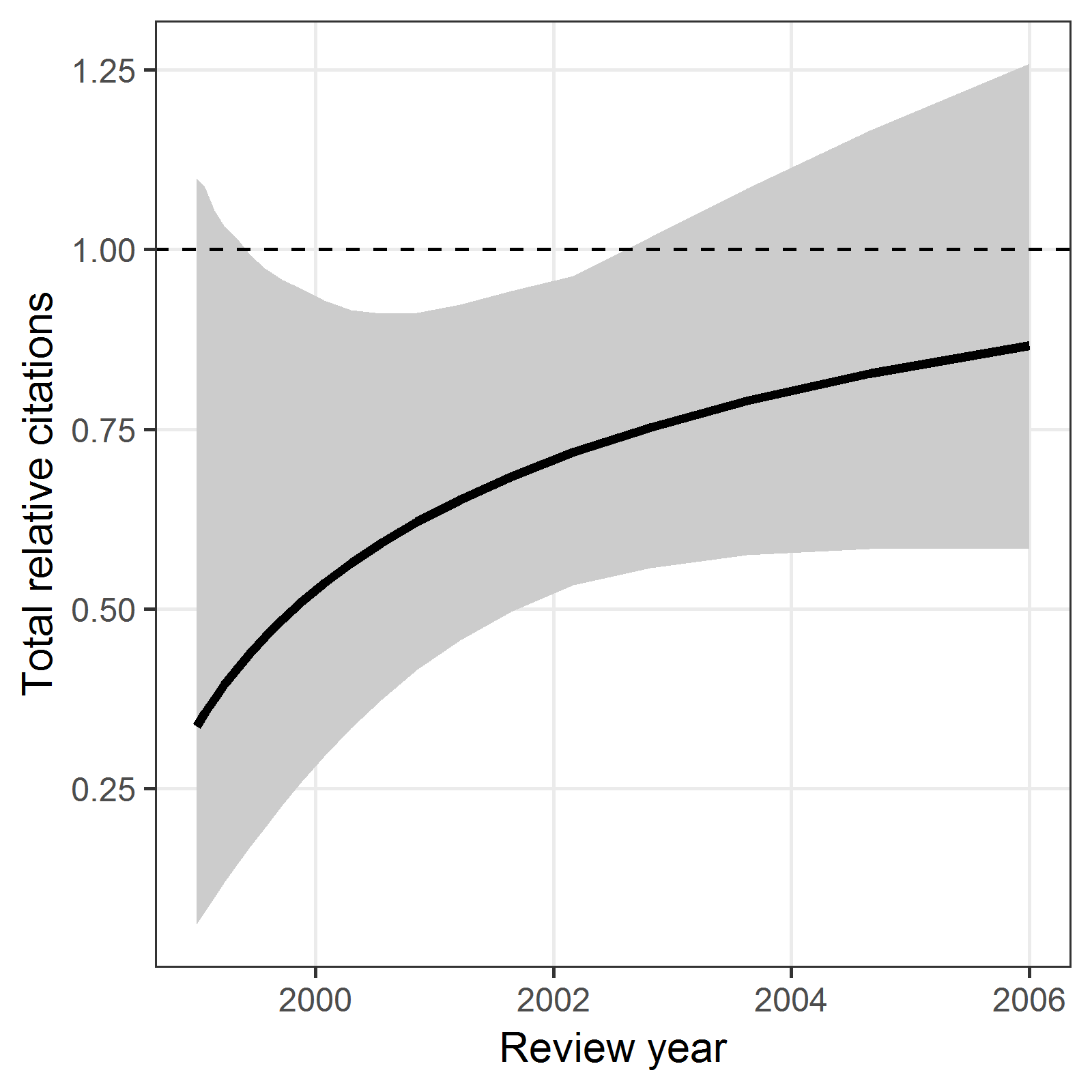
The mean shows a reduction in total citations for higher scores. The reduction for scores between 1 to 2 is steeper than the reduction for scores between 2 to 3.

##### Predictions for standard deviation (imputed data)



The mean line is flat, showing no clear association between the standard deviation and total citations. The very wide confidence interval indicates great uncertainty for high standard deviations.

##### Predictions for review year (imputed data)



There was a relatively small increase in citations by review year.

#### Table of parameter estimates (imputed data)

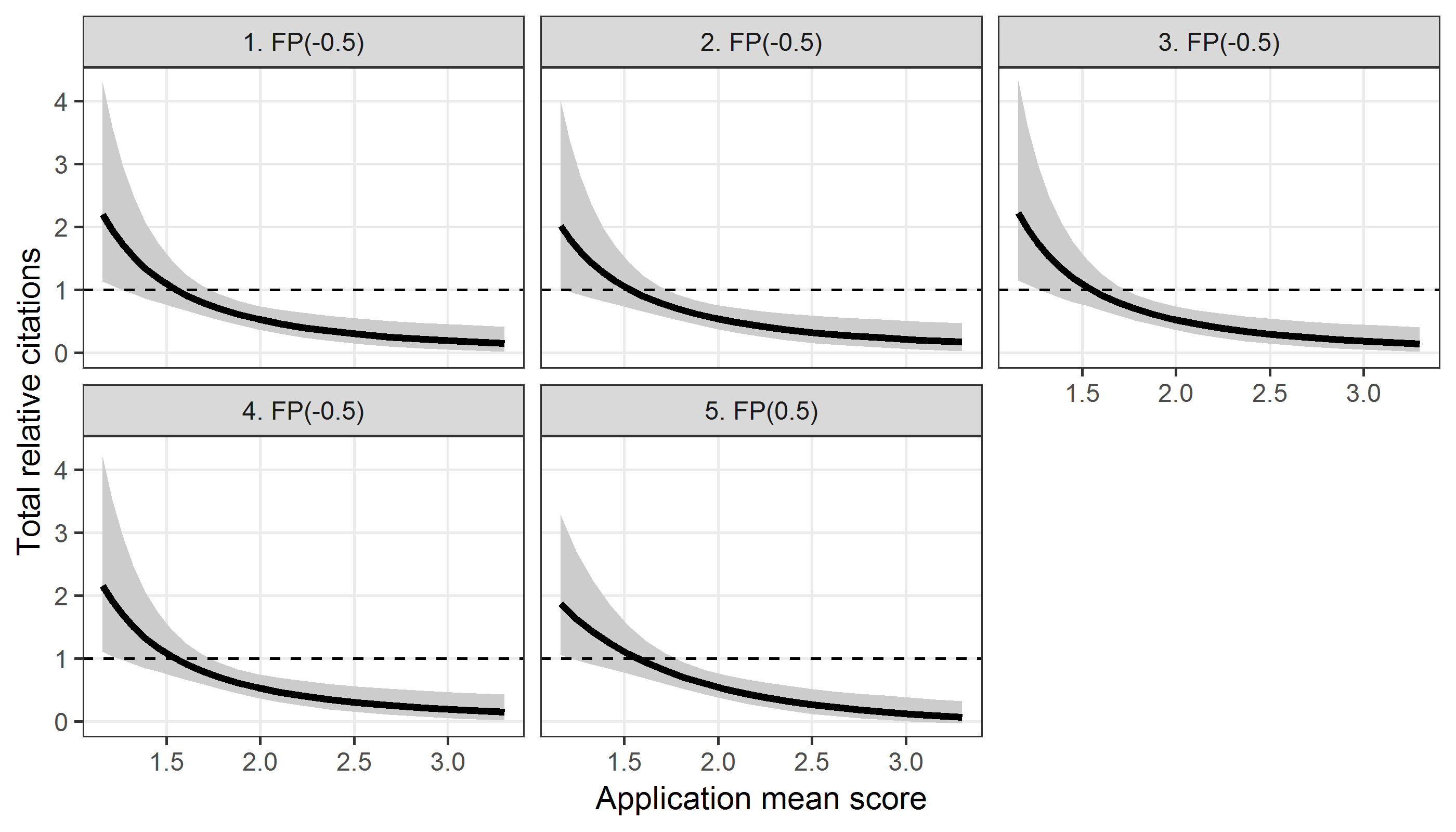
=========================================  
 Dependent variable:   
 --------------------  
 log(`Web TRC` + 0.1)  
-----------------------------------------  
Mean score, f(-0.5) 5.9 (2.6 to 9.1)   
Score SD, f(3) 0.7 (-1.3 to 2.6)   
Review year, f(-0.5) -1.2 (-3.0 to 0.6)   
Intercept -4.1 (-6.9 to -1.2)   
-----------------------------------------  
Observations 227   
Log Likelihood -439.0   
=========================================  
Note:

Only the mean score had a statistically significant effect on citations. As before it is better to use the plots rather than the parameter estimates to interpret the estimated effects. The “f()” in the variable label is the best fractional polynomial transformation.

#### Examining the top five models (imputed data)

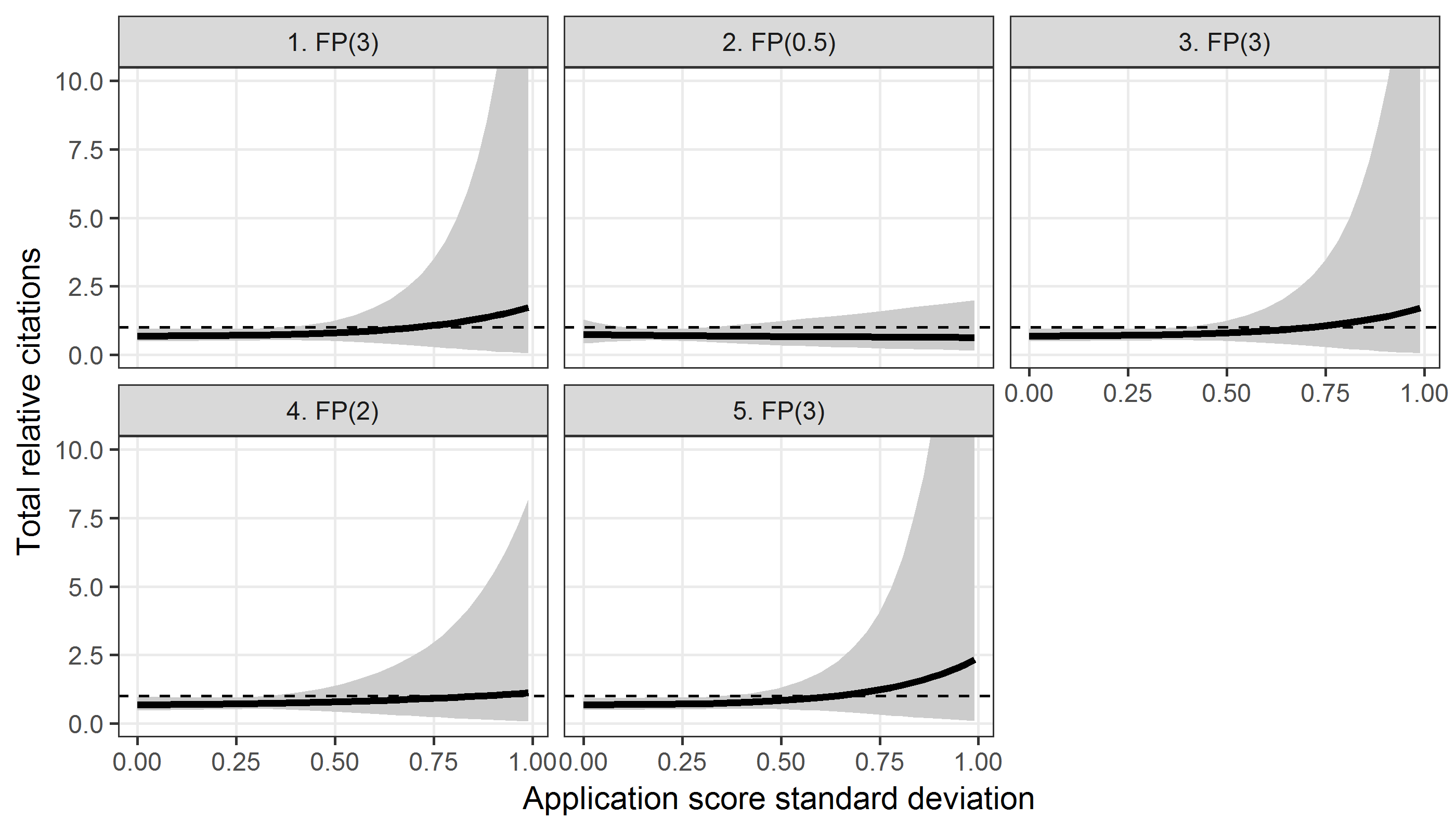
Here we examine the top five models in terms of the deviance to check whether there is any variation in the estimates for models that have a similarly good fit.

##### Predictions for top five models for mean application score



The headings for each plot show the deviance ranking (1 to 5) and the fractional polynomial (FP).

##### Predictions for top five models for application score standard deviation

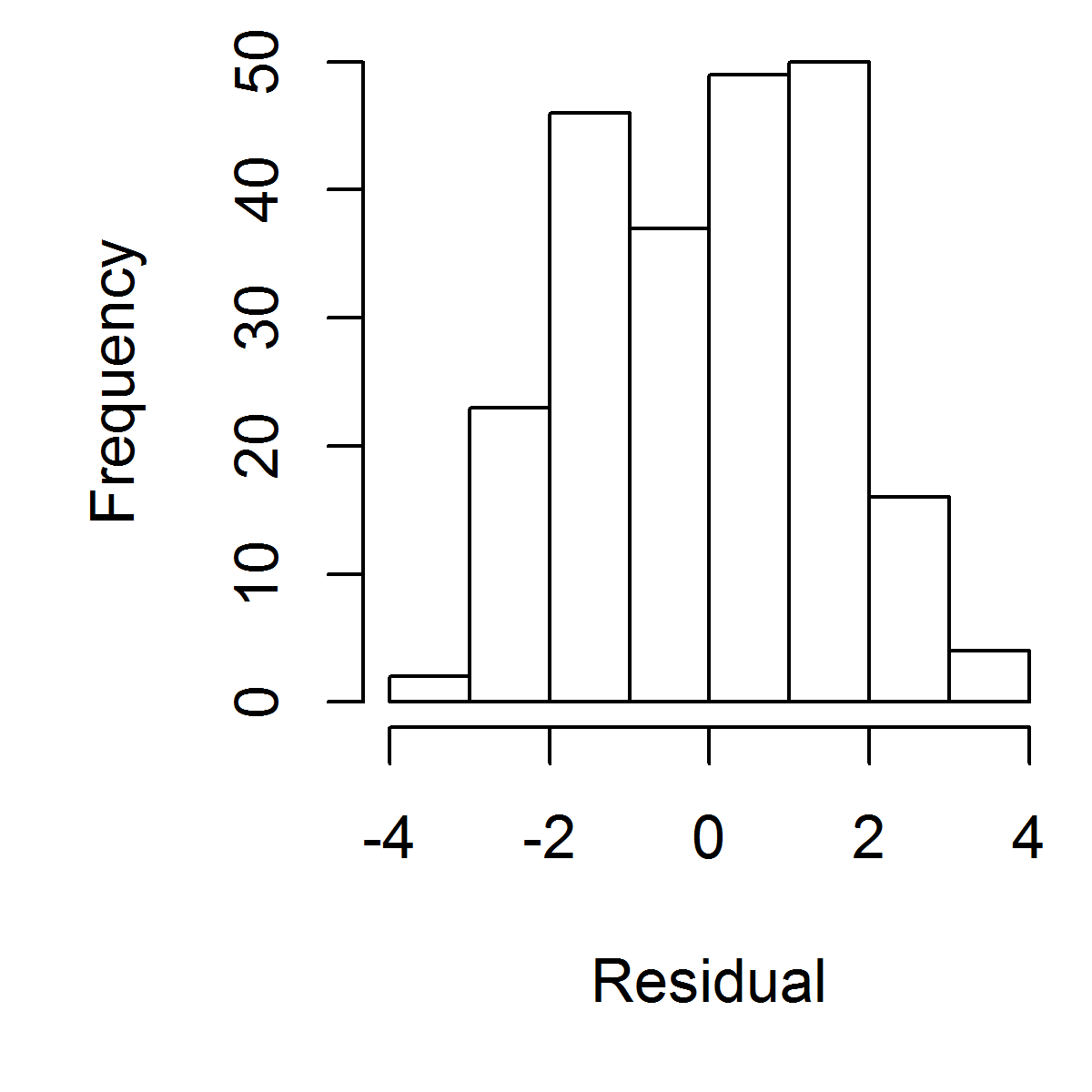


The headings for each plot show the deviance ranking (1 to 5) and the fractional polynomial (FP).

To help focus the plots on the most interesting predictions, we truncated the upper limits of the y-axis at 10. The predictions for the second best model with a fractional polynomial of 0.5 were interesting as they had a much smaller confidence interval for larger standard deviations.

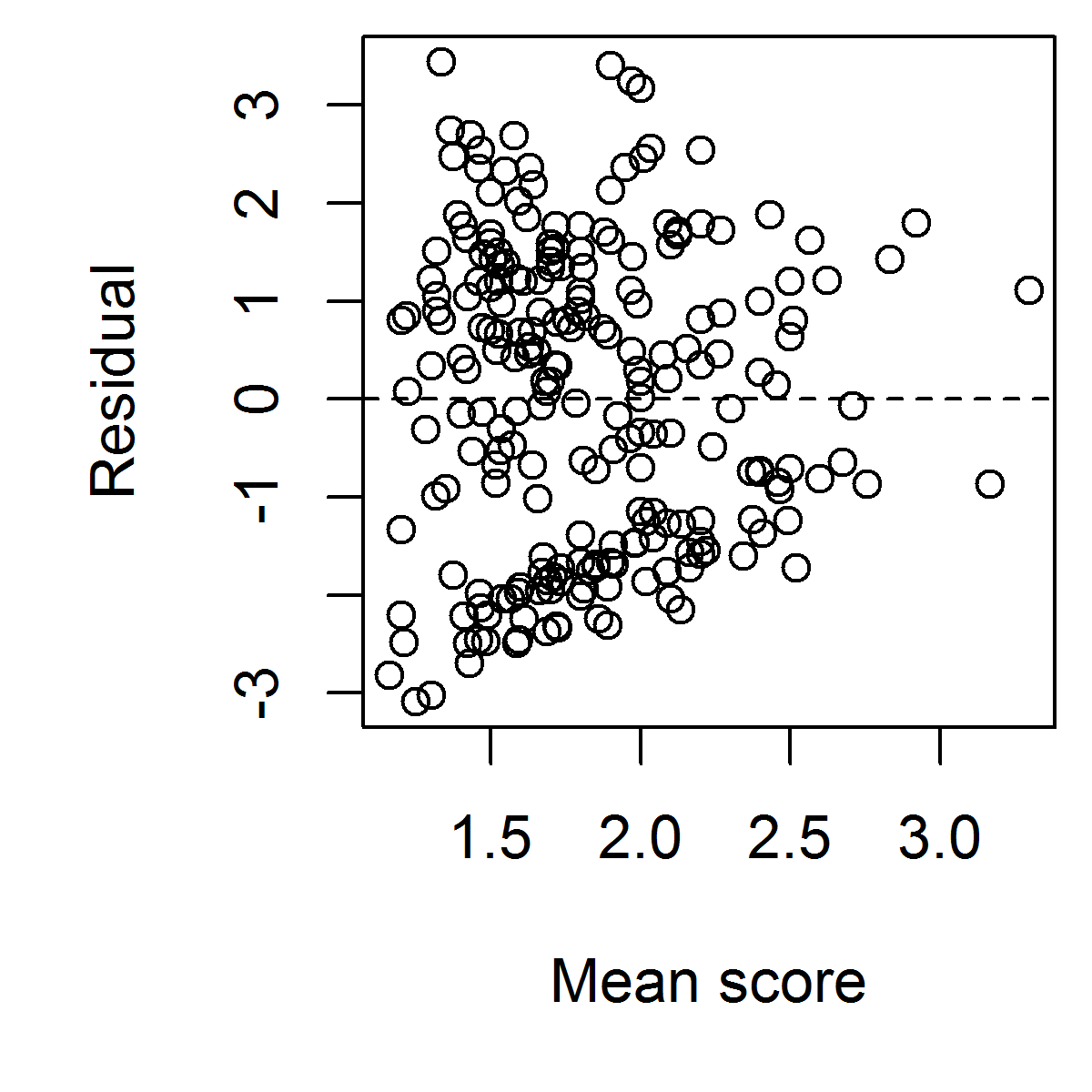
#### Model checking (imputed data)

##### Histogram of residuals (imputed data)

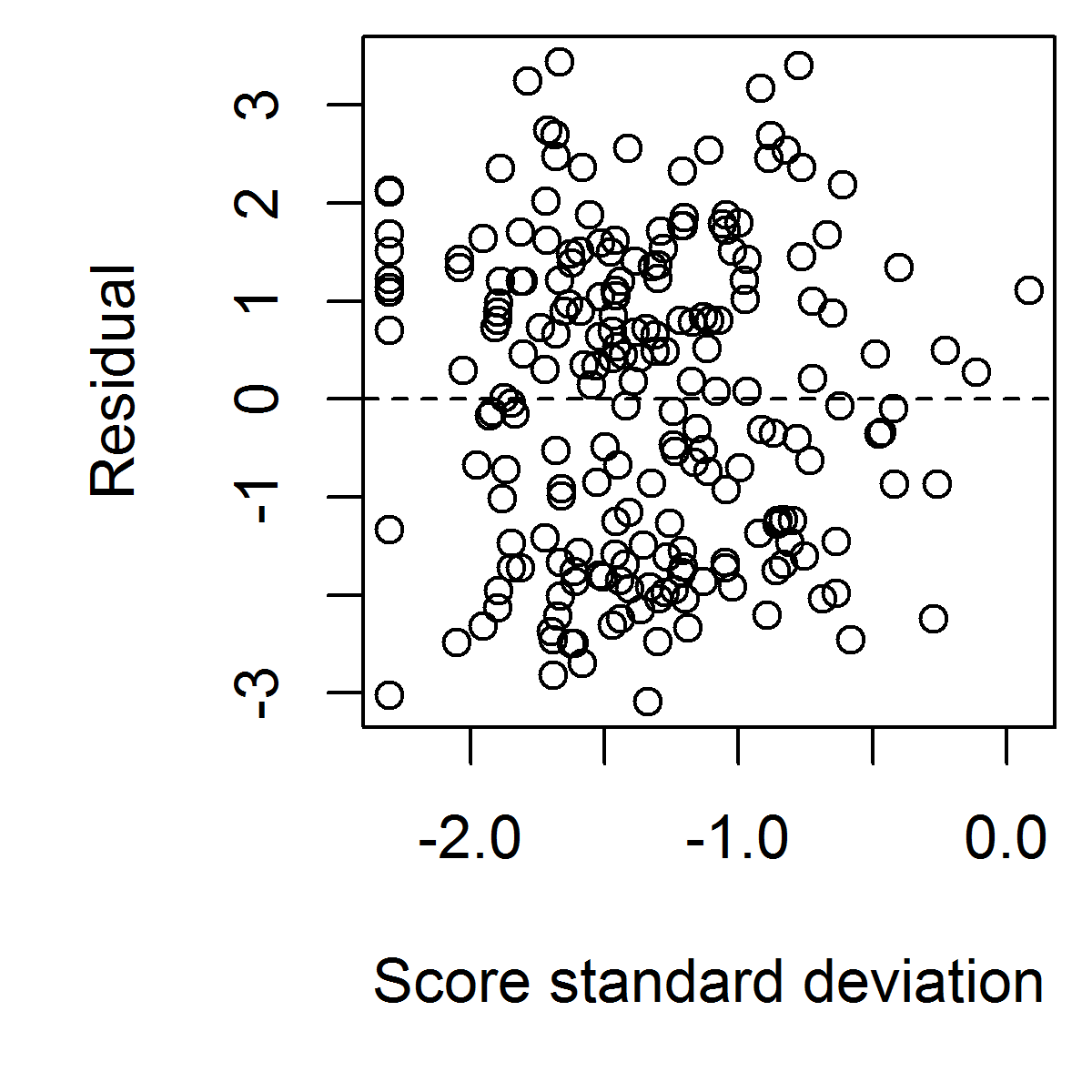


There are no clear outliers in the residuals.

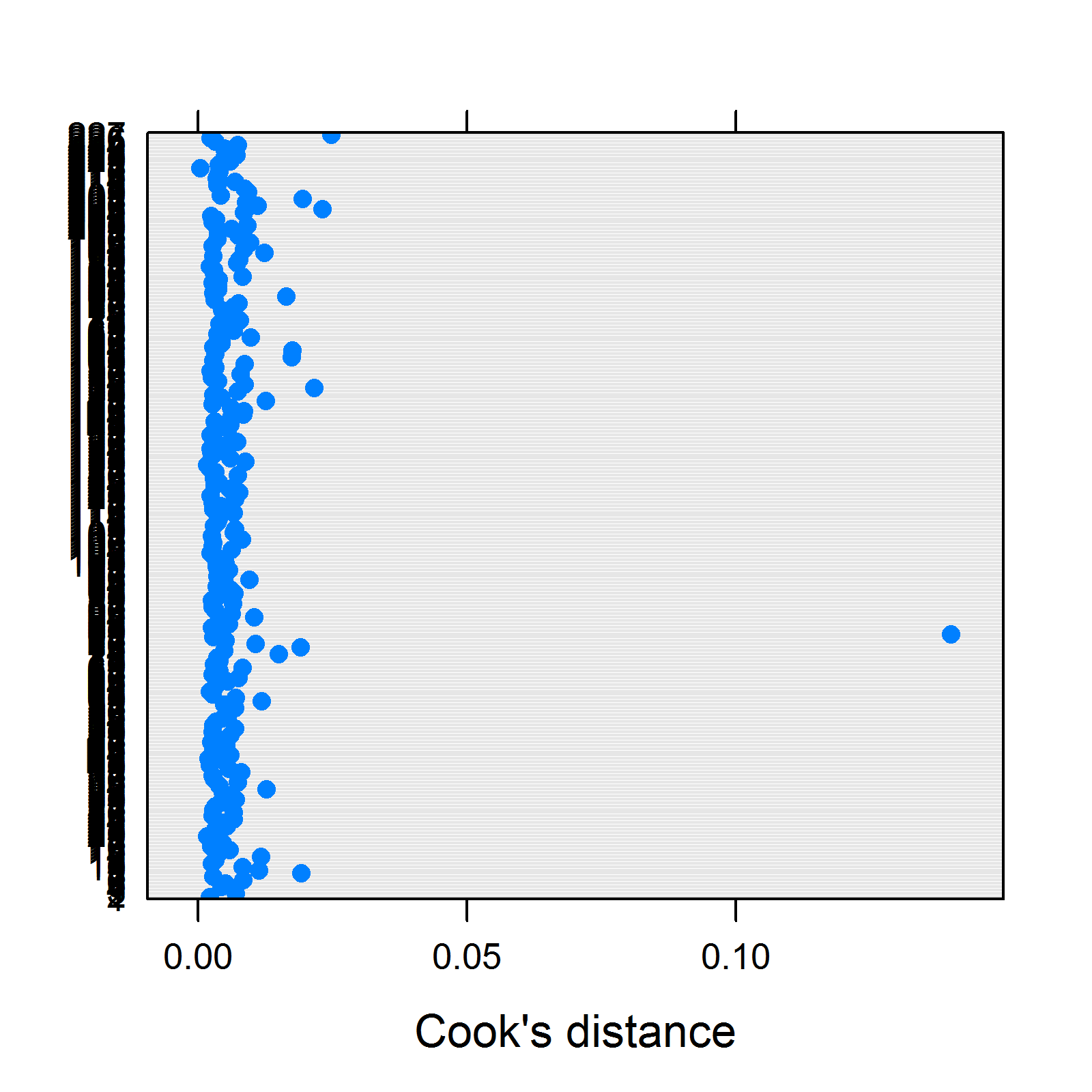
##### Scatter plot of residuals against mean (imputed data)



##### Scatter plot of residuals against log standard deviation (imputed data)



##### Cook’s distance for finding influential observations (imputed data)



There is one clear outlier using Cook’s distance which may be an influential observation. This is the application with the largest standard deviation which had a small citation number just above zero.

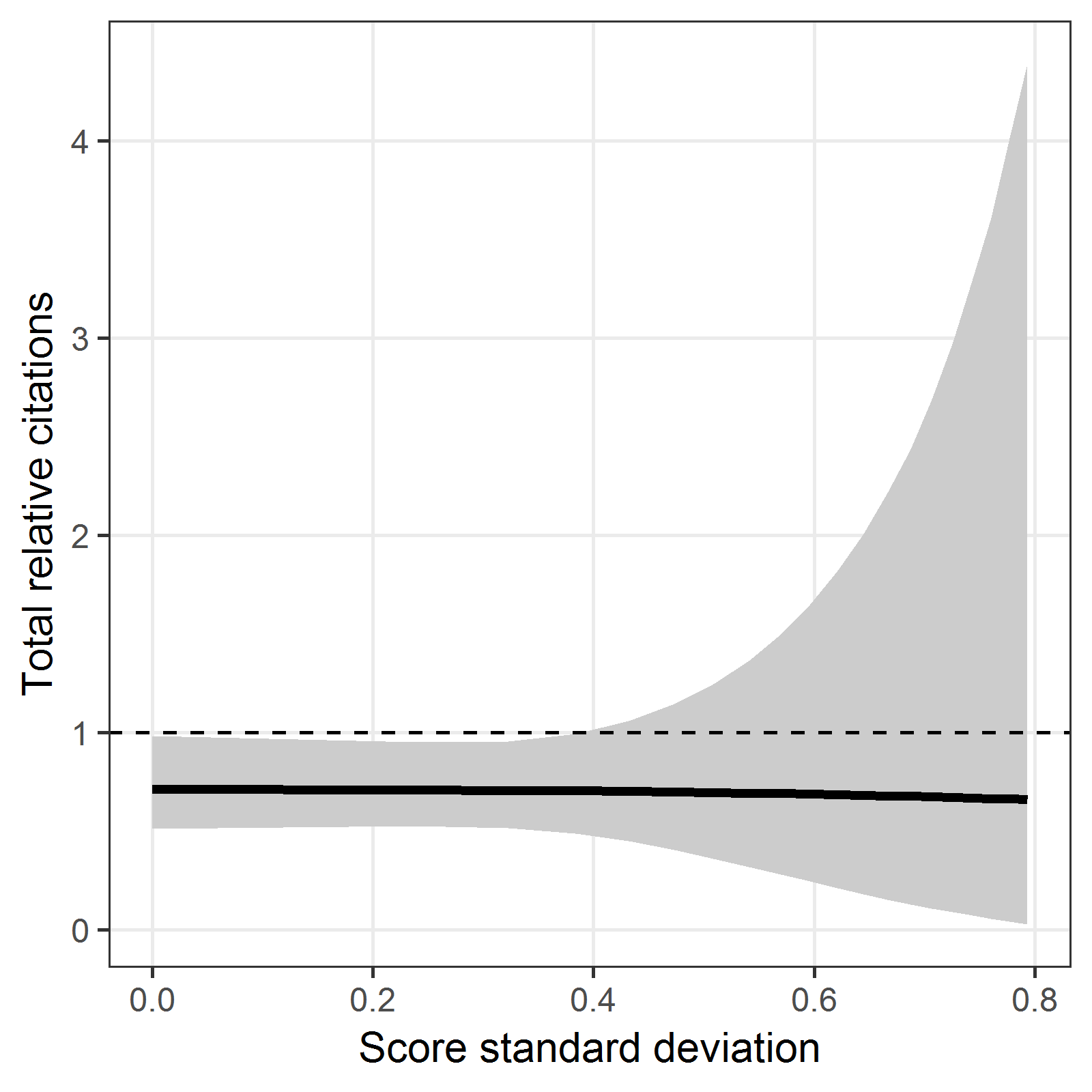
#### Results excluding one influential observation (imputed data)

=========================================  
 Dependent variable:   
 --------------------  
 log(`Web TRC` + 0.1)  
-----------------------------------------  
Mean score, f(-0.5) -3.3 (-5.1 to -1.5)   
Score SD, f(3) -0.1 (-2.8 to 2.6)   
Review year, f(-0.5) -1.2 (-3.0 to 0.7)   
Intercept 4.7 (2.4 to 7.1)   
-----------------------------------------  
Observations 226   
Log Likelihood -437.1   
=========================================  
Note:

The estimate for the score standard deviation is very close to zero, indicating little influence of the standard deviation on citations.

The “f()” in the variable label is the best fractional polynomial transformation.

##### Predictions for standard deviation (excluding influential observation; imputed data)



The mean line is flat, showing no clear association between the standard deviation and total citations. There is still uncertainty for high standard deviations, even after excluding the influential observation, although this has been reduced compared with the previous figure.

### Model using score range (imputed data)

Here we use the range in place of the standard deviation as the measure of panel disagreement.

#### Model fit - Top five deviances for score range (FP = Fractional Polynomial) (imputed data)

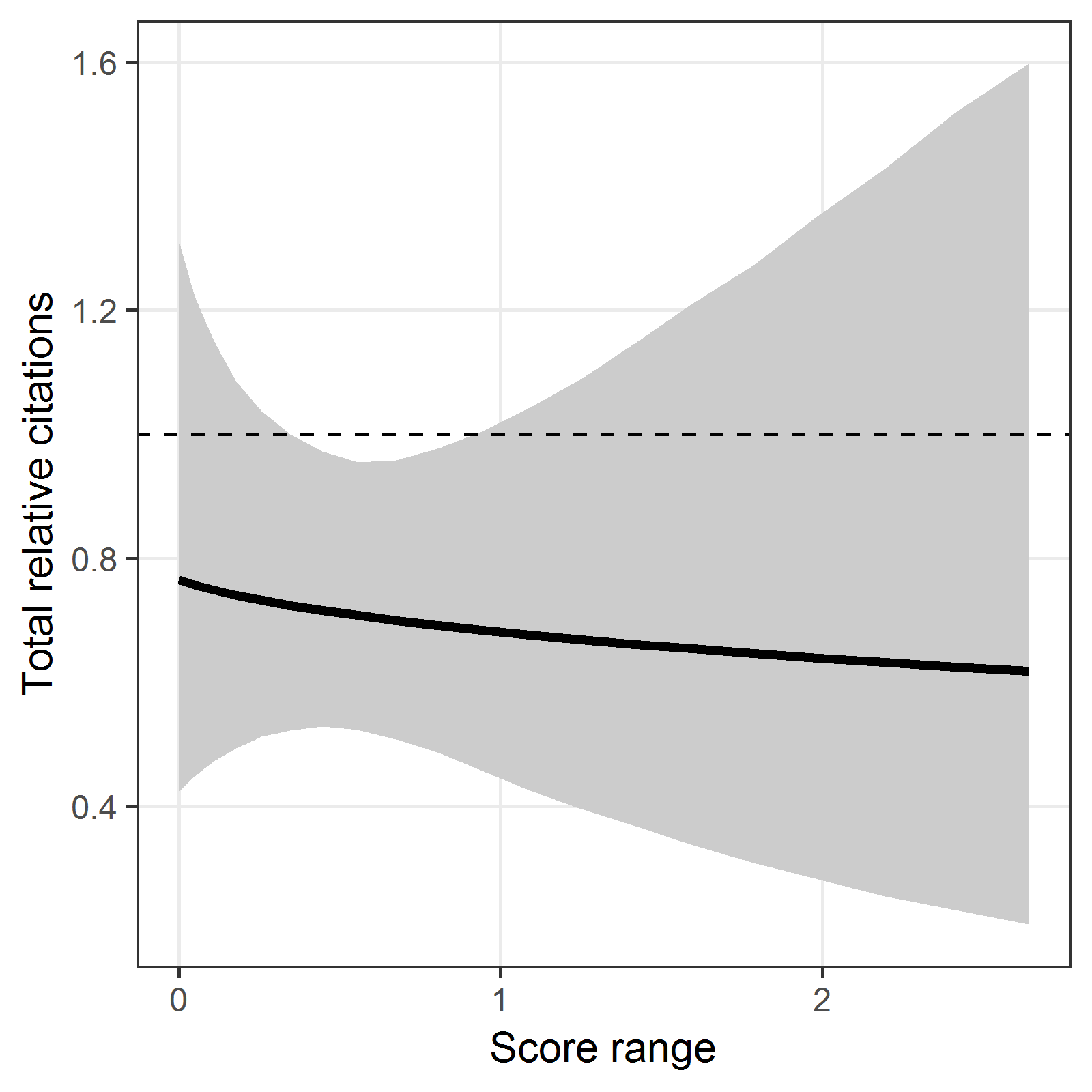
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FP1 | FP2 | FP3 | Deviance | Difference in deviance |
| -0 | 0 | -0 | 879.9 | 0 |
| -0 | 0 | -1 | 880.3 | 0.41 |
| -0 | 0 | -2 | 880.5 | 0.66 |
| -0 | -0 | -0 | 880.6 | 0.7 |
| 0 | 0 | -0 | 880.9 | 1 |

FP1 = application mean score, FP2 = application score range, FP3 = review year.

#### Table of parameter estimates (imputed data)

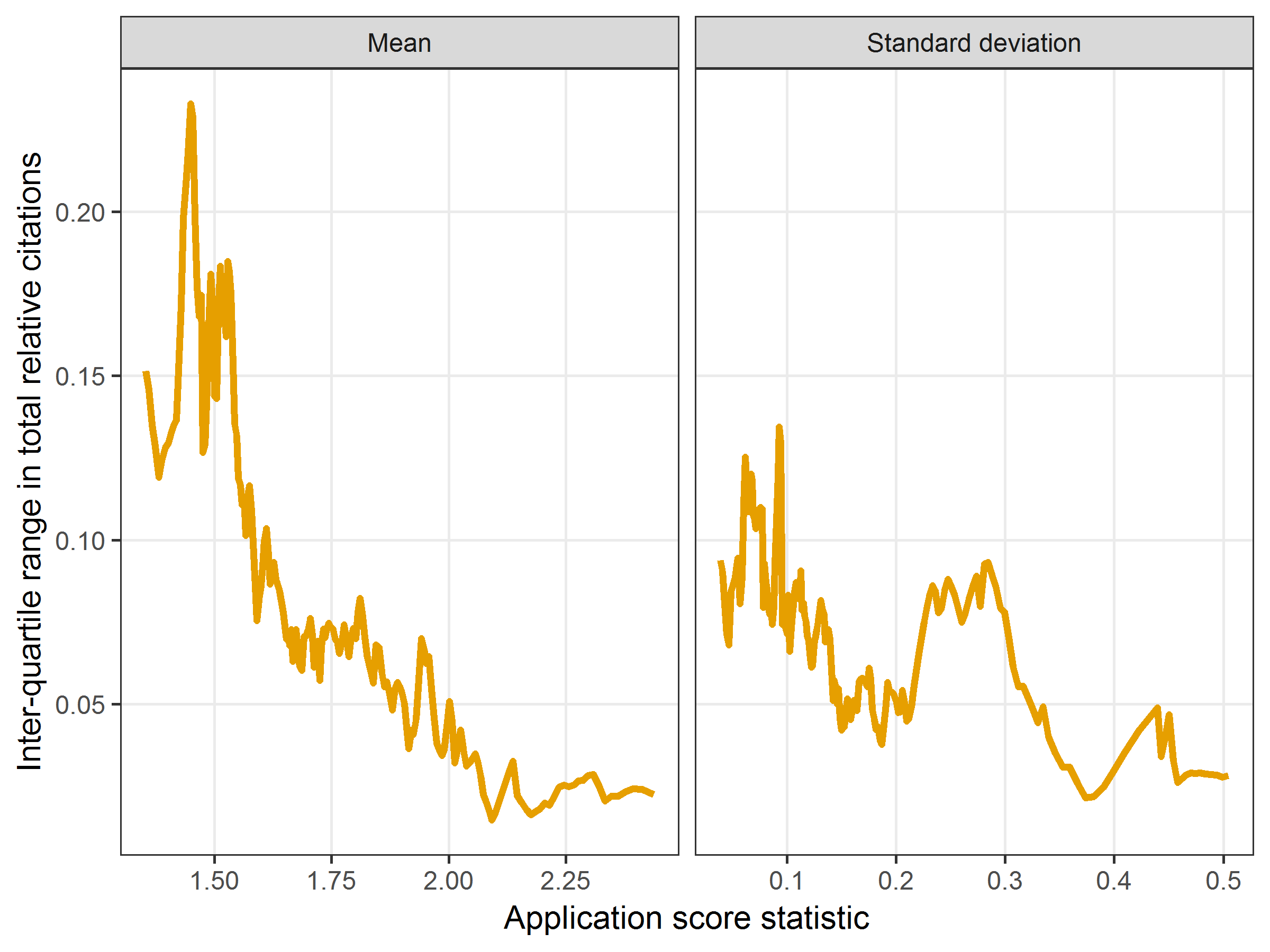
=========================================  
 Dependent variable:   
 --------------------  
 log(`Web TRC` + 0.1)  
-----------------------------------------  
Mean score, f(-0.5) 5.4 (2.1 to 8.7)   
Score range, f(0.5) -0.1 (-1.0 to 0.8)   
Review year, f(-0.5) -1.2 (-3.0 to 0.6)   
Intercept -3.6 (-6.6 to -0.5)   
-----------------------------------------  
Observations 227   
Log Likelihood -439.9   
=========================================  
Note:

##### Predictions for score range (imputed data)



There was no clear association between the range in an application’s score and its subsequent citations.

# Plots of inter-quartile range in citations against score statistics (complete case)



We used a scatterplot smoother to examine the association between the inter-quartile range (IQR) in citations and the application score statistics (Diggle et al, 2002, chapter 3). The aim is to see if there is a change in the variance in citations by the score statistics. This could occur because of a high volatity in risky proposals, with some failures (with no citations) and some huge successes (with many citations). This pattern would increase the variance in citations. We did not examine the variance or standard deviation because the citations counts are skewed and the estimates were strongly influenced by the application with the highest total relative citation.

The plots show a general decrease in the IQR in citations with increasing mean, suggesting more volatility in ciations for applications with the better scores. However, the scale on the y-axis is relatively small. There is also a slight reduction in the IQR in citations for larger standard deviations, which directly contradicts the idea of having more volatility in citations where there was greater disagreement between reviewers.

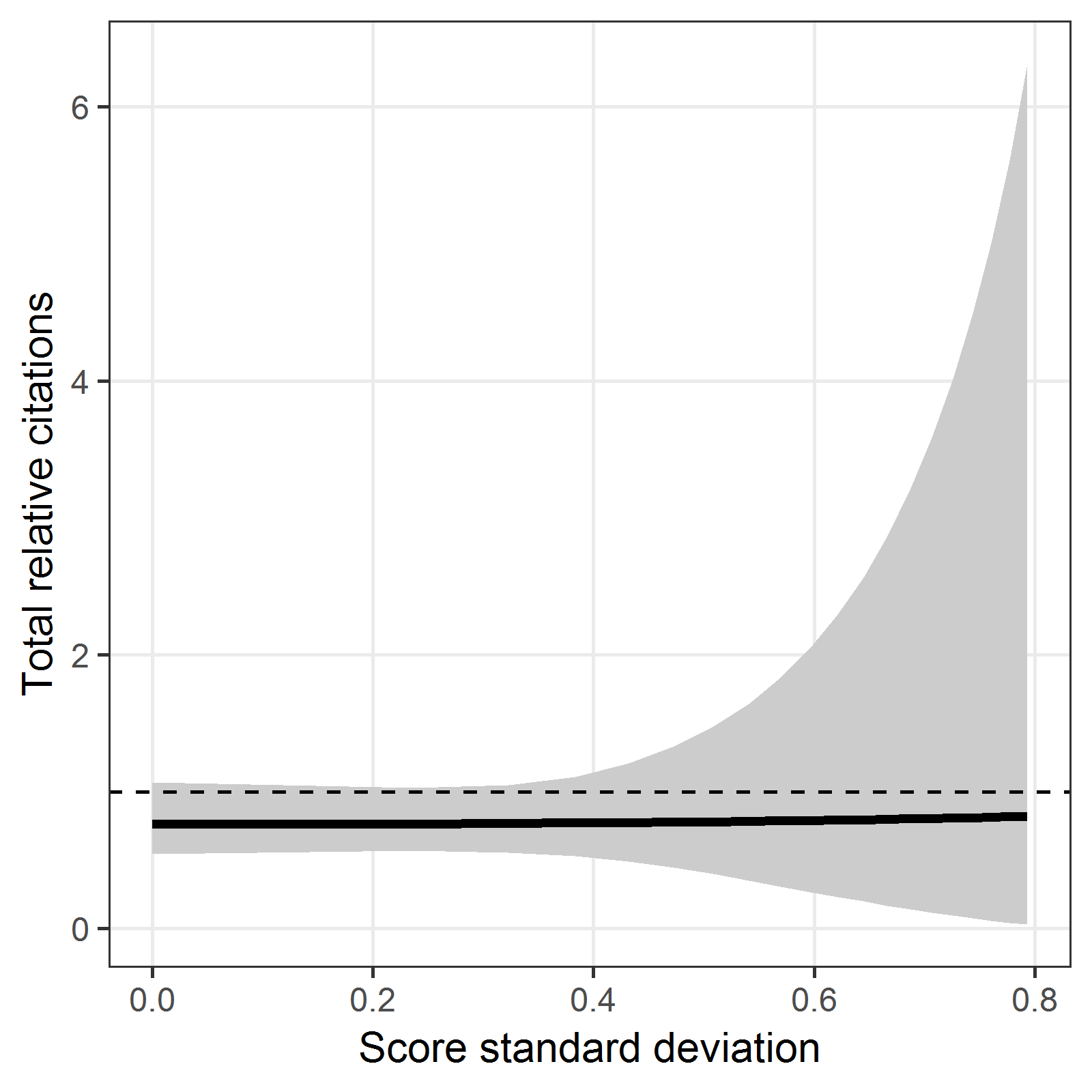
# Sensitivity analysis: just look at sub-sample with a larger number of reviewers (imputed data)

Around 10% of applications had only a few panel members and were reviewed by mail. To examine the influence of these applications, we ran a sensitivity analysis just using the larger sample of standard peer review panels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FP1 | FP2 | FP3 | Deviance | Difference in deviance |
| -0 | 3 | -2 | 758 | 0 |
| -0 | 0 | -2 | 758.5 | 0.49 |
| -0 | 2 | -2 | 758.5 | 0.51 |
| -0 | 1 | -2 | 758.9 | 0.84 |
| 0 | 3 | -2 | 758.9 | 0.91 |

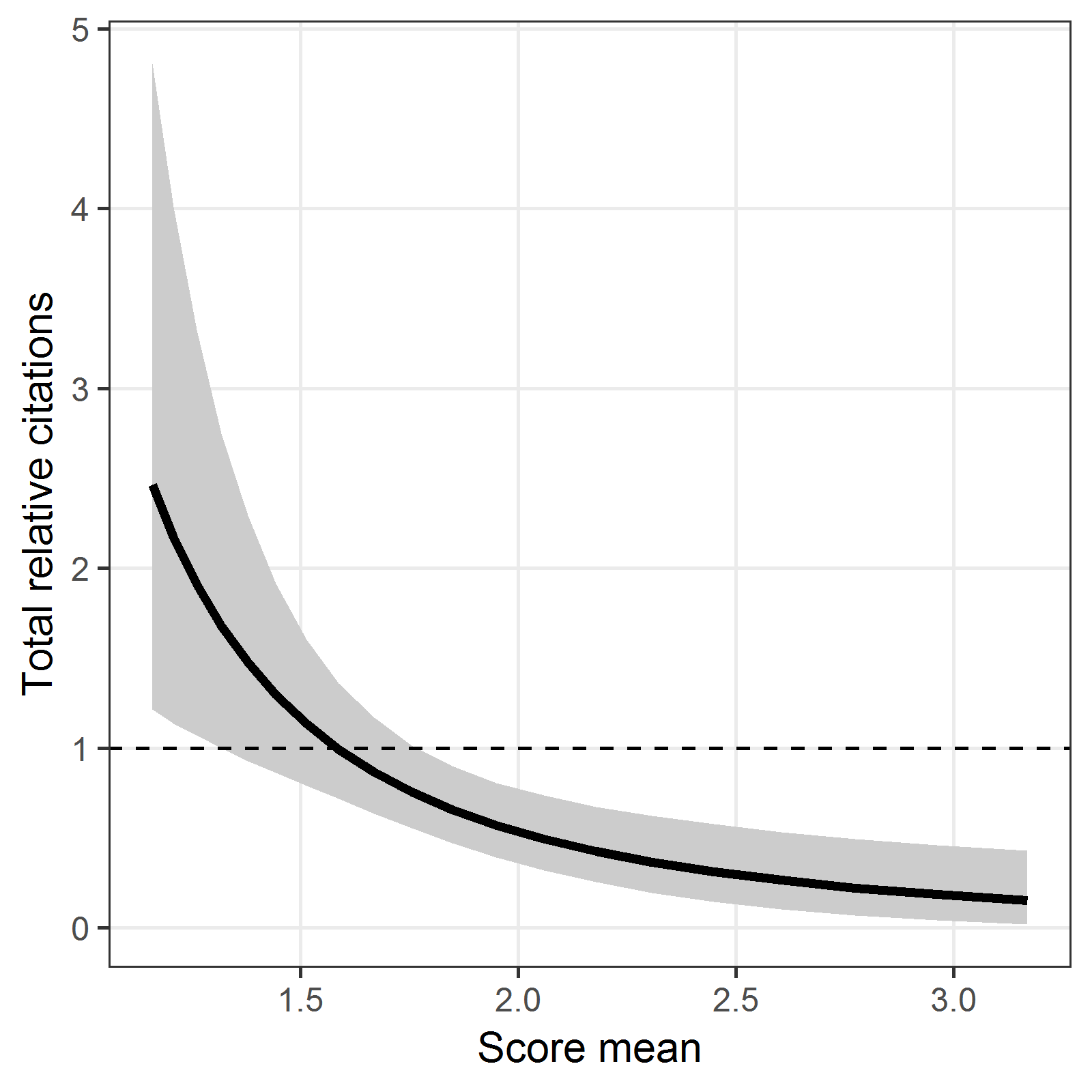
FP1 = application mean score, FP2 = application score standard deviation, FP3 = review year.

#### Predictions for standard deviation for sensitivity analysis (imputed data)



The results from this model were very similar to the model using all panels.

#### Predictions for mean for sensitivity analysis (imputed data)



The results from this model were very similar to the model using all panels.

# Scoring approach suggested by our peer reviewers (complete case)

The peer reviewers for our *F1000Research* paper suggested using the mean application score to fund a top-half of applications in a first “bucket”, then ranking the next half of funded applications using the application score standard deviation in a second “bucket”.

Funding the second bucket by ranking using the mean (so the same as the first bucket) gave the following results in that bucket:

* a total relative citation of 224.1
* a maximum total relative citation of 19.9
* an inter-quartile range in relative citation of 4.3

Funding the second bucket by ranking using the standard deviation (largest to smallest) gave the following results in that bucket:

* a total relative citation of 195.5.
* a maximum total relative citation of 35.9.
* an inter-quartile range in total relative citation of 2.9

# References

* Diggle PJ, Heagerty PJ, Liang KY, Zeger S. (2002) Analysis of Longitudinal Data. *OUP Oxford*
* Gallo SA, Carpenter AS, Irwin D, McPartland CD, Travis J, et al. (2014) The Validation of Peer Review through Research Impact Measures and the Implications for Funding Strategies. *PLOS ONE* **9**(9): e106474. <https://doi.org/10.1371/journal.pone.0106474>
* Stef van Buuren, Karin Groothuis-Oudshoorn (2011) mice: Multivariate Imputation by Chained Equations in R. *J Stat Software* **45**(3)