The relationship between scientific publishing retractions and democracy: An ecological analysis

Ahmad Sofi-Mahmudi

Table of contents

1 Aim

To determine the relationship between the number of retracted papers and the level of democracy and affecting factors.

2 Preparing the data

2.1 Variables

The variables have come from various sources, as follows:

Table 1: Data sources for each of the variables

Variable	Definition	Data source
Retractions	The number of retracted articles	The Retraction Watch Database: http://retractiondatabase.org/
Democracy	The level of democracy in the country, with a score range of 0 to 10, higher values indicating better democratic settings	Democracy Index by the Economist Intelligence Unit (EIU): https://www.eiu.com/topic/democracy-index/
Published papers	The number of all published papers for each country	SCImago Journal & Country Rank: https://www.scimagojr.com/countryrank.php
Campaigns	The number of non-violent mass campaigns	NAVCO: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ON9XND Reference (3)
GDP per capita	The total output created through the production of goods and services in a country during a certain period.	The World Bank: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
HDI	Country's social and economic development.	UNDP: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
Industry's share of economy in percent	Share of manufacturing in gross domestic product	The World Bank: https://databank.worldbank.org/source/world-development-indicators
Length of executive tenure	As a measure of political (in)stability.	Archigos: http://ksgleditsch.com/archigos.html Reference (4)

Variable	Definition	Data source
Location	The continent that each country is located at.	United Nations geoscheme: https://unstats.un.org/unsd/methodology/m49/
Muslim share of population	Estimated proportion of each country that is recognized to be Muslim.	Reference (5)
The number of top universities	Among 1,000 top universities, based on the Academic Ranking of World Universities (ARWU) – commonly known as the Shanghai Ranking.	Shanghai Ranking: https://www.shanghairanking.com/rankings/arwu/2022
Plurality/majority system	Whether the political system is plural or not.	https://havardhegre.net/iaep/ Reference (6)

I have stored the cleaned version of all these variables in the *retractions.csv* file. For the dependent variable, I assign 0 to all those countries that were not listed in the Retraction Watch Database. As these countries are almost entirely small countries with low research output, I am also creating a zero-truncated dataset including exclusively countries with at least one retraction.

To perform a sensitivity analysis, it is better to have an outlier-removed dataset.

First, loading the needed packages:

```
pacman::p_load(dplyr,
               ggplot2,
               knitr,
               car,
               mice,
               gamlss,
               broom.mixed,
               miceadds,
               rnaturalearth,
               maps,
               ggpubr)
```

And then, loading the datasets:

```r
retractions = read.csv("data/retractions_democracy_data.csv")

# Factoring and releveling regions
retractions$region = relevel(as.factor(retractions$region), ref = 3)

# Factoring two other variables
retractions$ongoing_nonviolent_campaign = as.factor(retractions$ongoing_nonviolent_campaig

retractions$plurality = as.factor(retractions$plurality)


retractions$Income.Group. = NULL
```

Now, creating the zero-truncated dataset:

```r
trunc_retraction = subset(retractions, retractions!=0)
```

And outlier-removed one:

```r
Q1 = quantile(retractions$retractions, 0.25)
Q3 = quantile(retractions$retractions, 0.75)
IQR = IQR(retractions$retractions)

no_out_retraction = subset(retractions, retractions > (Q1 - 1.5*IQR) & retractions < (Q3 +
```

Let us take a quick look at the first rows of the main dataset:

```r
kable(head(retractions))
```

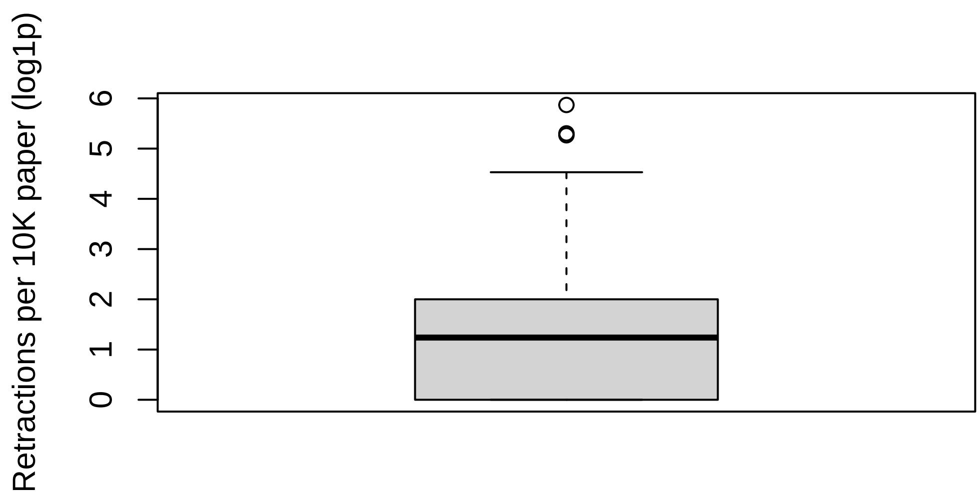
country	ISO_2023 region subregion retractions citable_documents mean_democracy_1996_2021 GDP_2021 ... plurality shanghai_											
Afghanistan	AFG	Asia	Southern Asia	0	2346	2.505385	5	300.05563949263 49491	2	99.3	0	0
Albania	ALB	Europe	Southern Europe	5	6460	5.880769	9	2558.90273167210 85183	3	20.5	0	0
Algeria	DZA	Africa	Northern Africa	61	90863	3.677692	2	2246.06207996 5884945	17	99.1	0	0
American Samoa	ASM	Oceania	Polynesia	0	224	NA	0	11479.3068 NA	NA	NA	0	NA

country	ISO_20230124	region	subregion	retractions	citabledocuments_1996_2021	mean_democracy_2008_2021	[illegible]	GDP_2015 [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] shanghai_
Andorra	AND	Europe	Southern Europe	0	310	NA	0	24620.4	0.8515	13.09153	NA	0.0	0	NA
Angola	AGO	Africa	Sub-Saharan Africa	1	1525	3.444615	5	1719.34	0.5203	24.763098	37	0.0	0	0

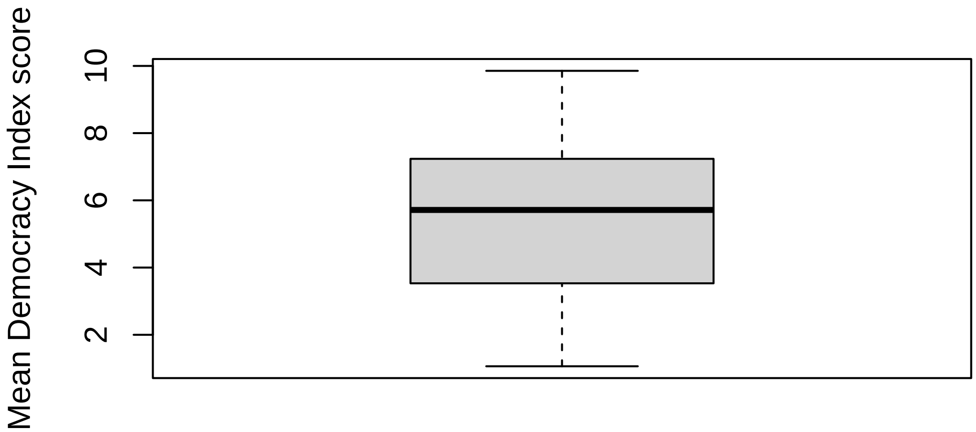
2.2 A quick look at the data

First, box plots:

```r
boxplot(log1p(retractions$retractions/retractions$citabledocuments_1996_2021*10000),
        ylab = "Retractions per 10K paper (log1p)")
```

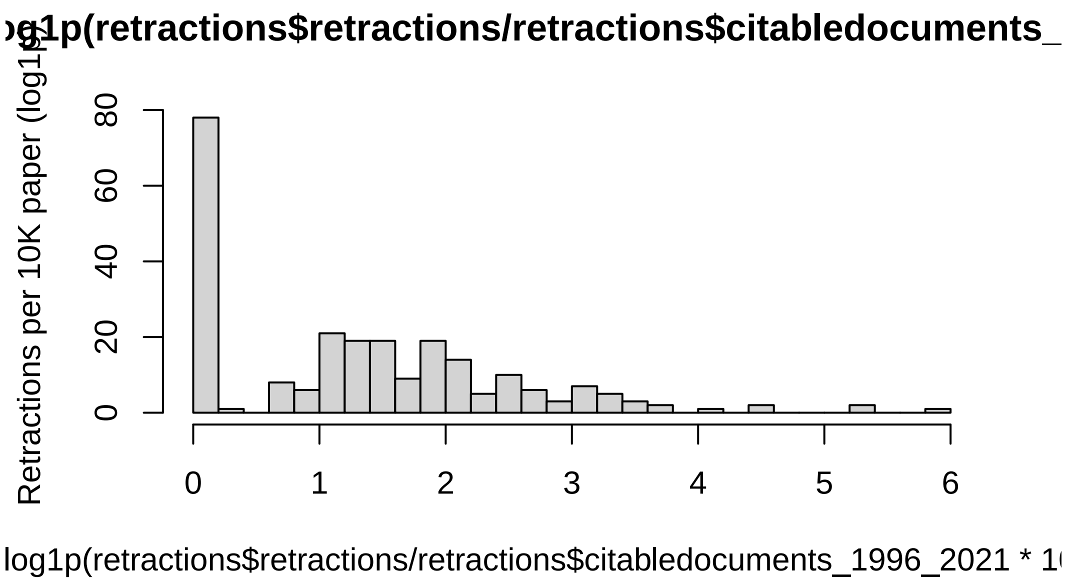



```r
boxplot(retractions$mean_democracy_2008_2021,
        ylab = "Mean Democracy Index score")
```

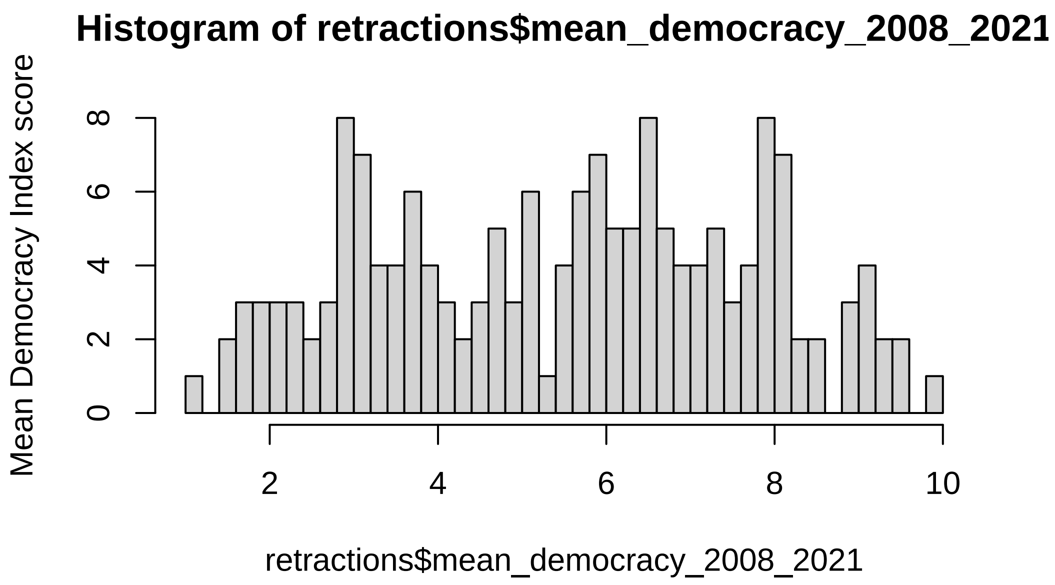


Histograms:

```r
hist(log1p(retractions$retractions/retractions$citabledocuments_1996_2021*10000),
        ylab = "Retractions per 10K paper (log1p)",
     breaks = 32)
```




```r
hist(retractions$mean_democracy_2008_2021,
        ylab = "Mean Democracy Index score",
     breaks = 32)
```



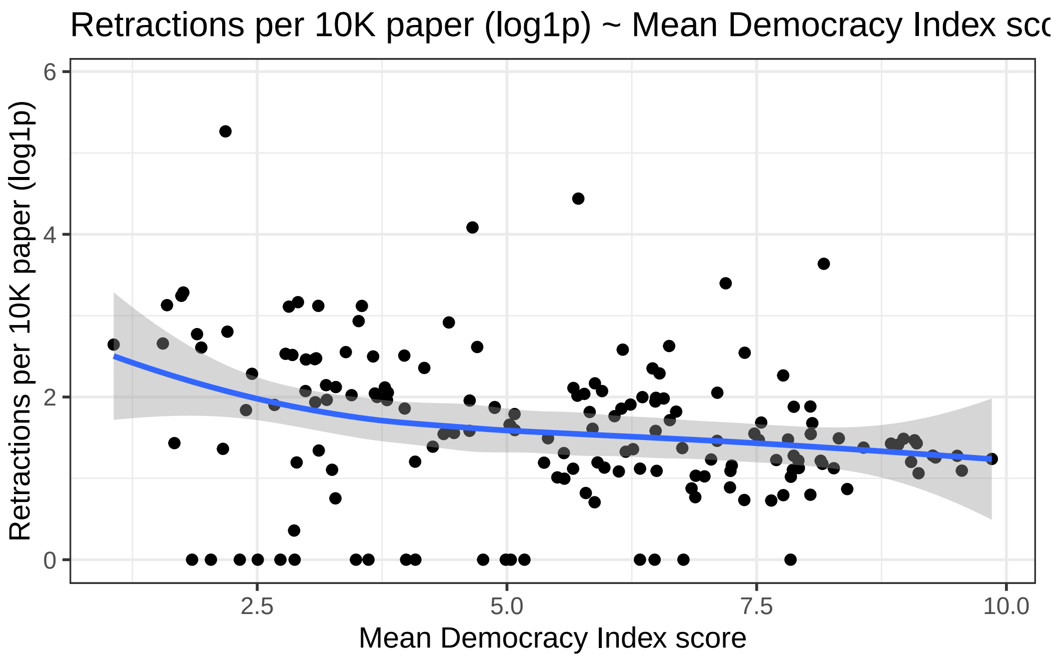
And now scatterplot:

```
ggplot(retractions) +
    aes(x = mean_democracy_2008_2021, y = log1p(retractions/citabledocuments_1996_2021*100
    geom_point() +
    labs(title = "Retractions per 10K paper (log1p) ~ Mean Democracy Index score",
         x = "Mean Democracy Index score",
         y = "Retractions per 10K paper (log1p)") +
    geom_smooth(method = "loess", se = T) + theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 74 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 74 rows containing missing values (`geom_point()`).
```




```
# tiff("Figure 1.tiff", width = 6, height = 3.73, units = "in", res = 300)
```

We can see that there is a negative relationship between the two variables. We will explore this relationship further.

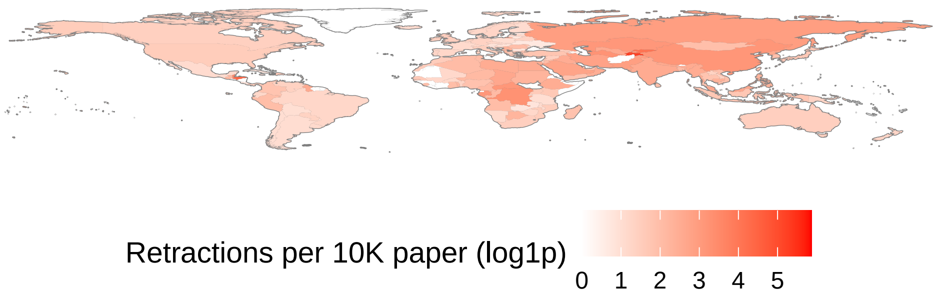
Let's also create world heat maps:

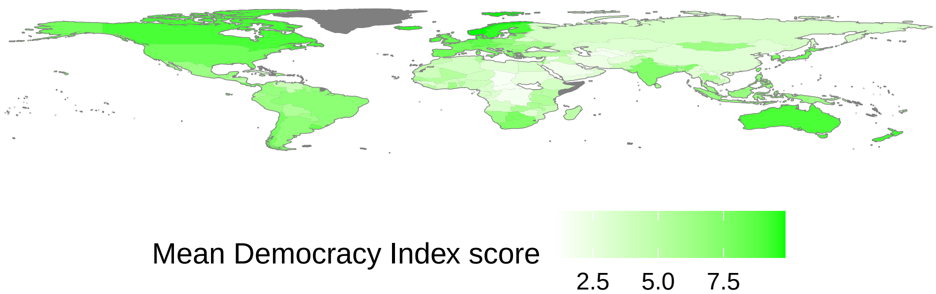
```
# Loading the world map
world_map = map_data("world")
world_map = subset(world_map, region != "Antarctica")

# Some modifications are needed
retractions$country_20230124[retractions$country_20230124 == "United States"] = "USA"
retractions$country_20230124[retractions$country_20230124 == "United Kingdom"] = "UK"
retractions$country_20230124[retractions$country_20230124 == "Russian Federation"] = "Russ
retractions$country_20230124[retractions$country_20230124 == "Republic of the Congo"] = "R
retractions[35, 1] <- "Ivory Coast"



# Drawing the map
map1 = ggplot(retractions) +
  geom_map(
```

```
    dat = world_map, map = world_map, aes(map_id = region),
    fill = "white", color = "#7f7f7f", size = 0.25
  ) +
  geom_map(map = world_map, aes(map_id = country_20230124, fill = log1p(retractions/citabl
  scale_fill_gradient(low = "white", high = "red", name = "Retractions per 10K paper (log1
  expand_limits(x = world_map$long, y = world_map$lat) + theme(legend.position="bottom",
        axis.line=element_blank(),
        axis.text=element_blank(),
        axis.ticks=element_blank(),
        axis.title=element_blank(),
        panel.background=element_blank(),
        panel.border=element_blank(),
        panel.grid=element_blank())
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
map2 = ggplot(retractions) +
  geom_map(
    dat = world_map, map = world_map, aes(map_id = region),
    fill = "white", color = "#7f7f7f", size = 0.25
  ) +
  geom_map(map = world_map, aes(map_id = country_20230124, fill = mean_democracy_2008_2021
  scale_fill_gradient(low = "white", high = "green", name = "Mean Democracy Index score")
  expand_limits(x = world_map$long, y = world_map$lat) + theme(legend.position="bottom",
        axis.line=element_blank(),
        axis.text=element_blank(),
        axis.ticks=element_blank(),
        axis.title=element_blank(),
        panel.background=element_blank(),
        panel.border=element_blank(),
        panel.grid=element_blank())

figure = ggarrange(map1, map2,
                   ncol = 1, nrow = 2, vjust = 1,
                   align = "hv", common.legend = F, legend = "bottom")

# tiff("Figure 1.tiff", width = 10, height = 10, units = "in", res = 300)
figure
```






```
#dev.off()
```

2.3 Missing values

This dataset has many missing values. Let us see their percentage for each variables in both datasets:

In the main dataset:

```
p_missing = unlist(lapply(retractions, function(x) sum(is.na(x))))/nrow(retractions)

kable(sort(p_missing[p_missing > 0], decreasing = TRUE)*100)
```

	x
plurality	35.68465
mean_democracy_2008_2021	30.70539
length_of_last_leader_tenure_2015	30.29046
HDI_mean_1990_2021	20.74689
industry_share_mean_1960_2022	15.35270
muslim_proportion	15.35270
GDP_pc_mean_1960_2022	13.27801

And in the zero-truncated one:

```
p_missing_trunc = unlist(lapply(trunc_retraction, function(x) sum(is.na(x))))/nrow(retract

kable(sort(p_missing_trunc[p_missing_trunc > 0], decreasing = TRUE)*100)
```

	x
plurality	10.788382
length_of_last_leader_tenure_2015	7.053942
mean_democracy_2008_2021	5.809129
HDI_mean_1990_2021	4.564315
industry_share_mean_1960_2022	2.489627
muslim_proportion	2.489627
GDP_pc_mean_1960_2022	2.074689

As we can see, there are many missing values, especially in the main dataset. Therefore, I performed multiple imputation for both datasets.

2.4 Multicollinearity

One problem that may arise in the process of both multiple imputation and regression analysis is high multicollinearity between the variables. Our variables are chosen based on the proposed DAG; however, we should investigate whether there are highly multicollinear variables. To do so, we run a linear regression model and assess the variance inflation factor of the covariates:

```
model_multi = lm(retractions~mean_democracy_2008_2021+ongoing_nonviolent_campaign+region+G

# Checking VIF:
kable(vif(model_multi))
```

	GVIF	Df	GVIF^(1/(2*Df))
mean_democracy_2008_2021	4.059375	1	2.014789
ongoing_nonviolent_campaign	1.277929	1	1.130455
region	5.268801	4	1.230875
GDP_pc_mean_1960_2022	3.110690	1	1.763715
HDI_mean_1990_2021	5.993550	1	2.448173
industry_share_mean_1960_2022	1.512866	1	1.229986
length_of_last_leader_tenure_2015	1.562382	1	1.249953
muslim_proportion	1.669691	1	1.292165

	GVIF	Df	GVIF^(1/(2*Df))
top_universities_shanghai_2022	1.296451	1	1.138618
plurality	1.153743	1	1.074124

As you can see, HDI, GDP, region, and democracy score are (almost) highly collinear. Removing HDI will almost solve this problem:

```
model_multi = lm(retractions~mean_democracy_2008_2021+ongoing_nonviolent_campaign+region+G

# Checking VIF:
kable(vif(model_multi))
```

	GVIF	Df	GVIF^(1/(2*Df))
mean_democracy_2008_2021	3.333916	1	1.825901
ongoing_nonviolent_campaign	1.272779	1	1.128175
region	3.135621	4	1.153561
GDP_pc_mean_1960_2022	2.479162	1	1.574536
industry_share_mean_1960_2022	1.233694	1	1.110718
length_of_last_leader_tenure_2015	1.507105	1	1.227642
muslim_proportion	1.651058	1	1.284935
top_universities_shanghai_2022	1.282764	1	1.132592
plurality	1.153732	1	1.074119

2.5 Multiple imputation

Now, we run multiple imputation using *mice* package for both datasets.

First, we rule out variables that cause problems in the imputation procedure. To do so, we should specify imputation methods manually.

```
# We run the mice code with 0 iterations
imp = mice(retractions, maxit=0)
```

```
Warning: Number of logged events: 3
```

```
# Extract predictorMatrix and methods of imputation
predM = imp$predictorMatrix
meth = imp$method
```

```r
# Setting values of variables I'd like to leave out to 0 in the predictor matrix
predM[, c("country_20230124")] = 0
predM[, c("ISO")] = 0
predM[, c("region")] = 0
predM[, c("subregion")] = 0
predM[, c("retractions")] = 0
predM[, c("citabledocuments_1996_2021")] = 0
predM[, c("HDI_mean_1990_2021")] = 0

# If you like, view the first few rows of the predictor matrix
# head(predM)
```

We will create 20 datasets, each with 50 iterations.

```r
imp = mice(retractions, m = 20, maxit = 50,
             predictorMatrix = predM,
             method = meth, print =  F, seed = 1280)


trunc_imp = mice(trunc_retraction, m = 20, maxit = 50,
            predictorMatrix = predM,
            method = meth, print =  F, seed = 1280)

no_out_imp = mice(no_out_retraction, m = 20, maxit = 50,
            predictorMatrix = predM,
            method = meth, print =  F, seed = 1280)
```

All set. Now, we move on to the analysis part.

3 Analysis

Since the number of retracted papers is a "count data", I used Poisson family regressions. Because of the different sample size for the number of all papers for each country, I used the number of citable documents as "offset". I also performed linear regression with the proportion of retractions as the dependent variable. Following codes show the results of both regression families for all three datasets.

3.1 Poisson family regression

Poisson regression uses Poisson distribution. This distribution is discrete with a single parameter, the mean, which is usually symbolized as either or . The mean is also understood as a rate parameter. It is the expected number of times that an item or event occurs per unit of time, area, or volume.

In the Poisson distribution, the mean and variance are identical, or at least nearly the same; i.e., Poisson distributions with higher mean values have correspondingly greater variability. This criterion of the Poisson distribution is referred to as the equidispersion criterion. The problem is that when modelling real data, the equidispersion criterion is rarely satisfied. Analysts usually must adjust their Poisson model in some way to account for any under- or overdispersion that is in the data.

Simply put, Poisson overdispersion occurs in data where the variability of the data is greater than the mean. A model that fails to properly adjust for overdispersed data is called an overdispersed model. As such, its standard errors are biased and cannot be trusted. Therefore, some other models have been proposed to consider overdispersion. All these models are based on the original Poisson model. These models are: 1) linear negative binomial (NB1), 2) standard negative binomial (NB2), 3) Poisson inverse Gaussian (PIG), 4) generalized negative binomial (NB-P), and 5) generalized Poisson (GP). The mean-variance relationship for each of these models is illustrated in Table below.

Table 7: Poisson regression family

Model	Mean	Variance
Poisson		
Negative binomial (NB1)		$(1 + \ ) = \ + $
Negative binomial (NB2)		$(1 + \ ) = \ + \ ^2$
Poisson inverse Gaussian (PIG)		$(1 + \ ^2) = \ + \ ^3$
Generalized negative binomial (NB-P)		$(1 + \ ) = \ + $
Generalized Poisson		$(1 + \ )^2 = \ + 2\ ^3 + \ ^2\ ^3$

3.1.1 Main dataset

In our data, retractions' mean and variance are not identical (mean=180.4, variance=1553568.0, Pearson ^{2} dispersion statistic=8302.9):

```
c(mean(retractions$retractions, na.rm = T), var(retractions$retractions, na.rm = T))
```

```
[1]     180.3693 1553568.0006
```

```
pois_model = glm(retractions~mean_democracy_2008_2021,
             data = retractions,
             family = poisson(link = "log"),
             offset = log(citabledocuments_1996_2021))

sum(residuals(pois_model, type="pearson")^2)
```

```
[1] 8302.858
```

Therefore, our dependent variable is overdispersed. To compensate for that, we should use other members of the family. We start with NB2.

3.1.1.1 Negative binomial type 2 (NB2)

```
fitimp_nb_uni = with(data = imp, gamlss(retractions~mean_democracy_2008_2021+offset(log(ci

kable(summary(pool(fitimp_nb_uni)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-5.4485437	0.0307778	-177.0283	228.9942	0
mu	mean_democracy_2008_2021	-0.2971098	0.0082056	-36.2083	232.7318	0
sigma	(Intercept)	3.9937983	0.0276270	144.5612	232.3065	0

This model shows that with each 1 unit increase in the number of retractions, the mean democracy score decreases by a factor of exp(-0.297)=0.743 (P<0.001).

NB2 model seems to have a better fit. Let's take a look at the AIC:

```
mean(sapply(fitimp_nb_uni$analyses, AIC))
```

```
[1] 1722.333
```

The AIC is also acceptable (1722.3). What about dispersion statistics?

```
sum(residuals(fitimp_nb_uni$analyses[[1]], type="simple")^2)/fitimp_nb_uni$analyses[[1]]$d
```

```
[1] 1.58386
```

I just used the first imputated dataset and it seems we have overdispersion.

Let's try PIG model:

3.1.1.2 Poisson inverse Gaussian (PIG)

```
fitimp_pig_uni = with(data = imp, gamlss(retractions~mean_democracy_2008_2021+offset(log(c

kable(summary(pool(fitimp_pig_uni)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	- 6.5064524	0.1630289	- 39.909805	219.1419	0.0000000
mu	mean_democracy_2008_2021	- 0.1200446	0.0297232	-4.038752	211.0693	0.0000752
sigma	(Intercept)	0.3240282	0.1064879	3.042863	226.6041	0.0026199

This model shows that with each 1 unit increase in the number of retractions, the mean democracy score decreases by a factor of exp(-0.120)=0.887 (P<0.001).

Let's assess AIC:

```
mean(sapply(fitimp_pig_uni$analyses, AIC))
```

```
[1] 1473.47
```

The AIC (1473.5) is lower than the NB2 model. And now dispersion statistics:

```
sum(residuals(fitimp_pig_uni$analyses[[1]], type="simple")^2)/fitimp_pig_uni$analyses[[1]]
```

```
[1] 1.018976
```

It seems we have complete equidispersion. To be sure about this choice, let's perform a log likelihood ratio test:

```
pchisq(2 * (mean(sapply(fitimp_pig_uni$analyses, logLik)) - mean(sapply(fitimp_nb_uni$anal
```

```
[1] 4.595296e-56
```

The test confirms that PIG model is better fitted with the data compared with the NB2 model. Therefore, we proceed with the PIG model. In order not to make the model more complex, I do not investigate the fitness of other members of the Poisson family (and there is no need to do so).

Now, let's perform adjusted PIG regression:

```
fitimp_pig_multi = with(data = imp, gamlss(retractions~mean_democracy_2008_2021+offset(log
```

```
Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged
```

```
Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged
```

```r
kable(summary(pool(fitimp_pig_multi)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-6.3709102	0.7377066	-8.6361028	203.8579	0.0000000
mu	mean_democracy_2008_2021	-0.0450430	0.0932492	-0.4830389	200.0130	0.6295964
mu	ongoing_nonviolent_campaign1	0.3094233	0.3243970	0.9538414	224.5186	0.3411897
mu	GDP_pc_mean_1960_2022	0.0000014	0.0000125	0.1140758	197.4623	0.9092936
mu	regionAfrica	-0.4550709	0.3786583	-1.2017981	224.1035	0.2307102
mu	regionAmericas	-0.4768720	0.3838343	-1.2423905	221.7722	0.2154046
mu	regionEurope	-0.8142575	0.4463945	-1.8240759	220.2284	0.0694958
mu	regionOceania	-0.8911059	0.5966595	-1.4934916	223.2317	0.1367206
mu	industry_share_mean_1960_2022	-0.0068191	0.0103507	-0.6588111	206.5799	0.5107507
mu	length_of_last_leader_tenure_2015	0.0070282	0.0175640	0.4001490	192.3444	0.6894908
mu	muslim_proportion	0.0014117	0.0047233	0.2988796	214.5442	0.7653212
mu	top_universities_shanghai_2022	0.0026187	0.0080100	0.3269306	224.4056	0.7440251
mu	plurality1	-0.2290695	0.2646919	-0.8654196	153.1914	0.3881618
sigma	(Intercept)	0.2761475	0.1214040	2.2746159	151.9453	0.0243283

In this model, mean democracy score decreases by a factor of exp(-0.045)=0.956 (P=0.630) with each 1 unit increase in the number of retracted papers.

3.1.2 Zero-truncated dataset

For this dataset, I only run the PIG regression models.

3.1.2.1 Poisson inverse Gaussian (PIG)

```
fit_truncimp_pig_uni = with(data = trunc_imp, gamlss(retractions~mean_democracy_2008_2021+

kable(summary(pool(fit_truncimp_pig_uni)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-6.2956290	0.1968214	-31.986506	143.3171	0.0000000
mu	mean_democracy_2008_2021	-0.1137712	0.0354537	-3.209007	133.6756	0.0016676
sigma	(Intercept)	0.4897032	0.1189476	4.116967	125.5763	0.0000691

This model shows that with each 1 unit increase in the number of retractions, the mean democracy score decreases by a factor of exp(-0.114)=0.892 (P=0.002) (compared with exp(-0.120)=0.887 in main dataset).

Let's assess AIC:

```
mean(sapply(fit_truncimp_pig_uni$analyses, AIC))
```

```
[1] 1418.264
```

The AIC is 1418.3. And now dispersion statistics:

```
sum(residuals(fit_truncimp_pig_uni$analyses[[1]], type="simple")^2)/fit_truncimp_pig_uni$a
```

```
[1] 0.9352772
```

It seems we have underdispersion which seems accepatble.

Now, let's perform adjusted PIG regression:

```
fit_truncimp_pig_multi = with(data = trunc_imp, gamlss(retractions~mean_democracy_2008_202
```

```
Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged

Warning in RS(): Algorithm RS has not yet converged
```

```
kable(summary(pool(fit_truncimp_pig_multi)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-5.7576983	0.7366929	-7.8156017	133.03862	0.0000000
mu	mean_democracy_2008_2021	-0.1117974	0.1001790	-1.1159766	127.34147	0.2665336
mu	ongoing_nonviolent_campaign	0.3057710	0.3413557	0.8957550	146.32049	0.3718549
mu	GDP_pc_mean_1960_2022	0.0000130	0.0000184	0.7064840	142.87025	0.4810377
mu	regionAfrica	-0.2750600	0.4063496	-0.6769048	146.34304	0.4995349
mu	regionAmericas	-0.1622912	0.3752653	-0.4324706	144.83967	0.6660427
mu	regionEurope	-0.7574194	0.4464423	-1.6965673	144.66939	0.0919291
mu	regionOceania	0.0100765	0.8362647	0.0120494	146.22080	0.9904026
mu	industry_share_mean_1960_2022	-0.0165683	0.0104073	-1.5919844	144.59211	0.1135721
mu	length_of_last_leader_tenure_2023	0.0027319	0.0184957	0.1477070	131.41680	0.8828004
mu	muslim_proportion	0.0019888	0.0048479	0.4102425	143.29594	0.6822408
mu	top_universities_shanghai_2022	0.0003319	0.0064697	0.0512969	146.40912	0.9591589
mu	plurality1	-0.2551337	0.2714264	-0.9399737	96.04896	0.3495890
sigma	(Intercept)	0.3131166	0.1249357	2.5062229	117.40103	0.0135719

In this model, mean democracy score decreases by a factor of exp(-0.112)=0.894 (compared with exp(-0.045)=0.956 from the main dataset) with each 1 unit increase in the number of retracted papers.

3.1.3 Outlier-removed dataset

Also for this dataset, I only run the PIG regression models.

3.1.3.1 Poisson inverse Gaussian (PIG)

```
fit_nooutimp_pig_uni = with(data = no_out_imp, gamlss(retractions~mean_democracy_2008_2021

kable(summary(pool(fit_nooutimp_pig_uni)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-6.5155447	0.2358225	-27.629015	158.6149	0.0000000
mu	mean_democracy_2008_2021	-0.1120694	0.0450876	-2.485593	140.8637	0.0141032
sigma	(Intercept)	0.7291663	0.1452664	5.019511	183.1767	0.0000012

This model shows that with each 1 unit increase in the number of retractions, the mean democracy score decreases by a factor of exp(-0.112)=0.894 (P=0.014) (compared with exp(-0.120)=0.887 in the main dataset).

Let's assess AIC:

```
mean(sapply(fit_nooutimp_pig_uni$analyses, AIC))
```

```
[1] 912.4588
```

The AIC is 912.5. And now dispersion statistics:

```
sum(residuals(fit_nooutimp_pig_uni$analyses[[1]], type="simple")^2)/fit_nooutimp_pig_uni$a
```

```
[1] 0.8677551
```

It seems we have overdispersion which seems accepatble.

Now, let's perform adjusted PIG regression:

```
fit_nooutimp_pig_multi = with(data = no_out_imp, gamlss(retractions~mean_democracy_2008_20

kable(summary(pool(fit_nooutimp_pig_multi)))
```

parameter	term	estimate	std.error	statistic	df	p.value
mu	(Intercept)	-6.3264538	0.7981126	-7.9267685	151.3446	0.0000000
mu	mean_democracy_2008_2021	-0.0299660	0.1057899	-0.2832595	132.6747	0.7774194
mu	ongoing_nonviolent_campaign	0.3104323	0.4031503	0.7700162	183.6942	0.4422794
mu	GDP_pc_mean_1960_2022	0.0000035	0.0000115	0.3038908	164.4666	0.7615952

parameter	term	estimate	std.error	statistic	df	p.value
mu	regionAfrica	-0.4763300	0.4249574	-1.1208889	183.6200	0.2637990
mu	regionAmericas	-0.5076896	0.4894351	-1.0372969	176.5678	0.3010160
mu	regionEurope	-0.9554074	0.5775942	-1.6541153	176.3726	0.0998827
mu	regionOceania	-1.0212320	0.6545162	-1.5602852	182.5658	0.1204243
mu	industry_share_mean_1960_2022	-0.0094807	0.0118116	-0.8026602	166.1885	0.4233176
mu	length_of_last_leader_tenure_2022	0.0092577	0.0181583	0.5098327	158.9141	0.6108760
mu	muslim_proportion	-0.0000520	0.0055576	-0.0093574	166.2980	0.9925452
mu	top_universities_shanghai_2022	-0.0560104	0.1143096	-0.4899887	183.2453	0.6247280
mu	plurality1	-0.2352238	0.3084402	-0.7626235	103.6765	0.4474196
sigma	(Intercept)	0.5056055	0.1468534	3.4429278	169.6405	0.0007249

In this model, mean democracy score decreases by a factor of exp(-0.030)=0.970 (compared with exp(-0.045)=0.956 from the main dataset) with each 1 unit increase in the number of retracted papers.

3.2 Linear regression

For this part, I used the number of retractions per 10K articles.

3.2.1 Main dataset

```
full.impdata = complete(imp, 'long', include = TRUE) %>%
  mutate(retraction_prop = retractions/citabledocuments_1996_2021*10000)

new_imp = as.mids(full.impdata)
```

Let's run the model and assess its fitness:

```
fitimp_linear_uni = with(data = new_imp,
               lm(retraction_prop~mean_democracy_2008_2021))
```

x
0.0020381

	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	448.3614	1	1535.118	0.3117	0.576729	0.002038	0.002038
Residual	219545.6545	NA	NA	NA	NA	NA	NA

x
2

```
kable(summary(pool(fitimp_linear_uni)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	11.6120829	5.9657105	1.9464711	151.1184	0.0534530
mean_democracy_2008_2021	-0.4572615	0.9681643	-0.4722974	140.0494	0.6374496

This model shows with each 1 unit increase in democracy score, the number of retracted papers per 10K article decreases by 0.457 unit.

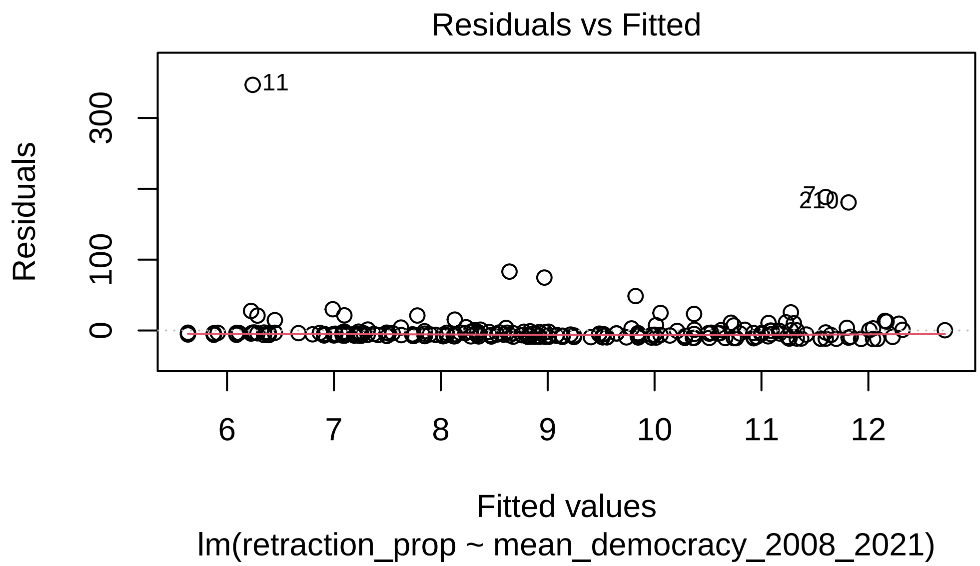
Now, let's check the model fitness:

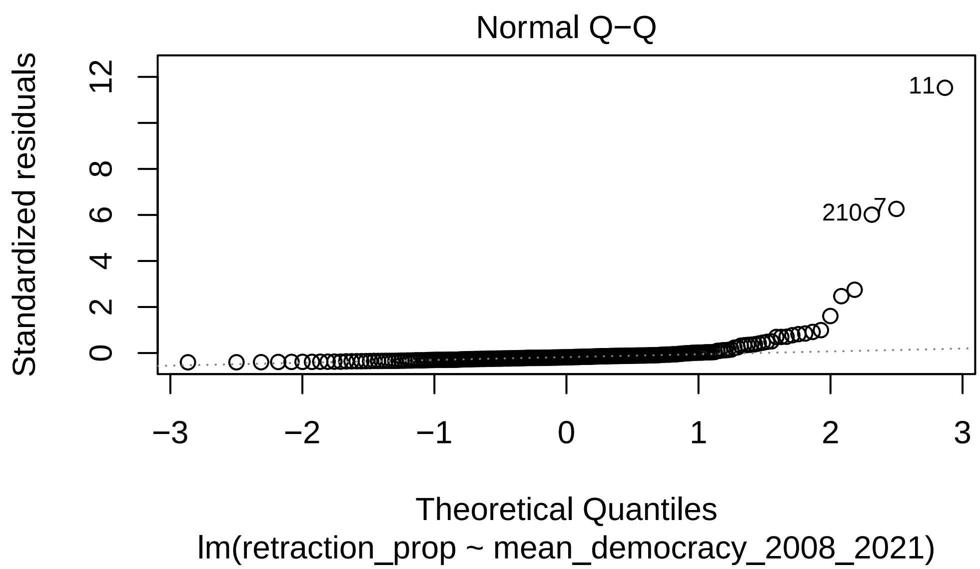
```
kable(mi.anova(mi.res=new_imp, formula="retraction_prop~mean_democracy_2008_2021"))
```

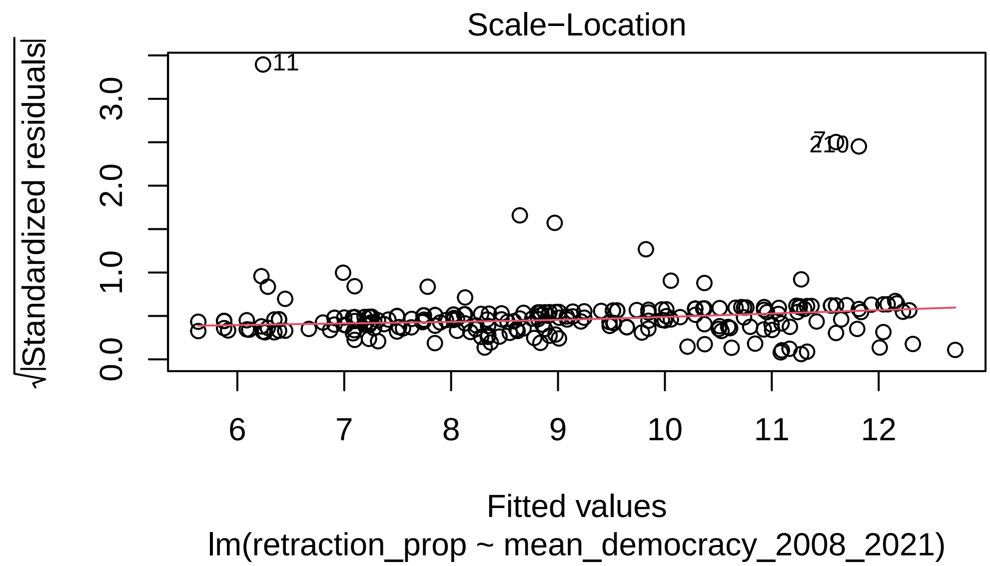
```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  retraction_prop~mean_democracy_2008_2021
R^2=0.002
..........................................................................
ANOVA Table
                                  SSQ df1      df2 F value  Pr(>F)    eta2
mean_democracy_2008_2021     448.3614   1 1535.118  0.3117 0.57673 0.00204
Residual                  219545.6545  NA       NA      NA      NA      NA
                         partial.eta2
mean_democracy_2008_2021      0.00204
Residual                           NA
```

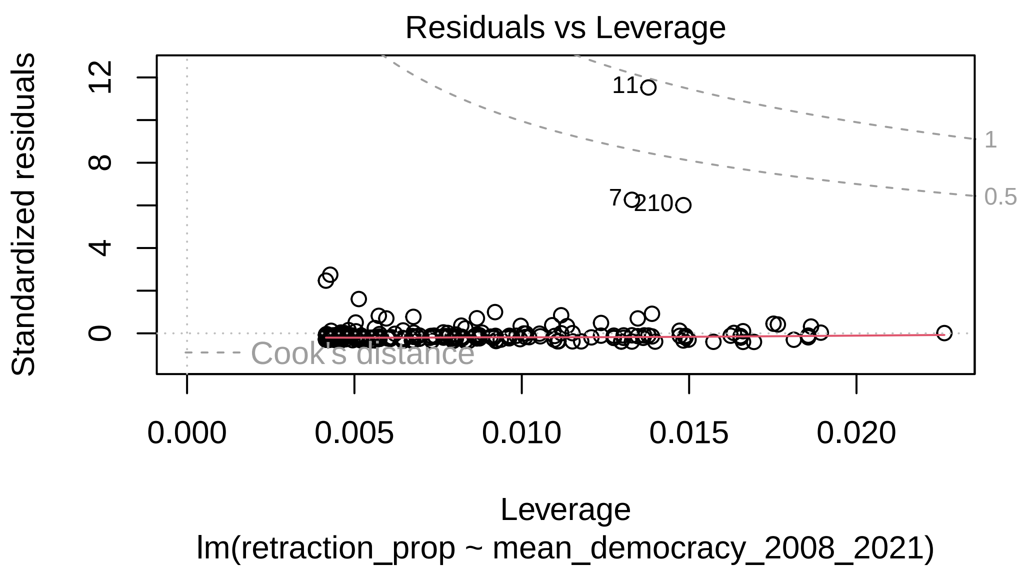
As we can see, the model fitness seems not to be good enough with R-squared of 0.002. Let's confirm this by exploring the plots:

```
plot(fitimp_linear_uni$analyses[[1]])
```


Residuals vs Fitted
Residuals
300
100
0
11
7
210
6
7
8
9
10
11
12
Fitted values
lm(retraction_prop ~ mean_democracy_2008_2021)


Normal Q−Q
Standardized residuals
12
8
6
4
2
0
11
210
7
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(retraction_prop ~ mean_democracy_2008_2021)





We can clearly see the signs of non-normality of the residuals. We have two other options: using the log or using square-root of the dependent variable. Let's start with square-root:

3.2.1.1 Square-root method

```
kable(mi.anova(mi.res=new_imp, formula="sqrt(retraction_prop)~mean_democracy_2008_2021"))
```

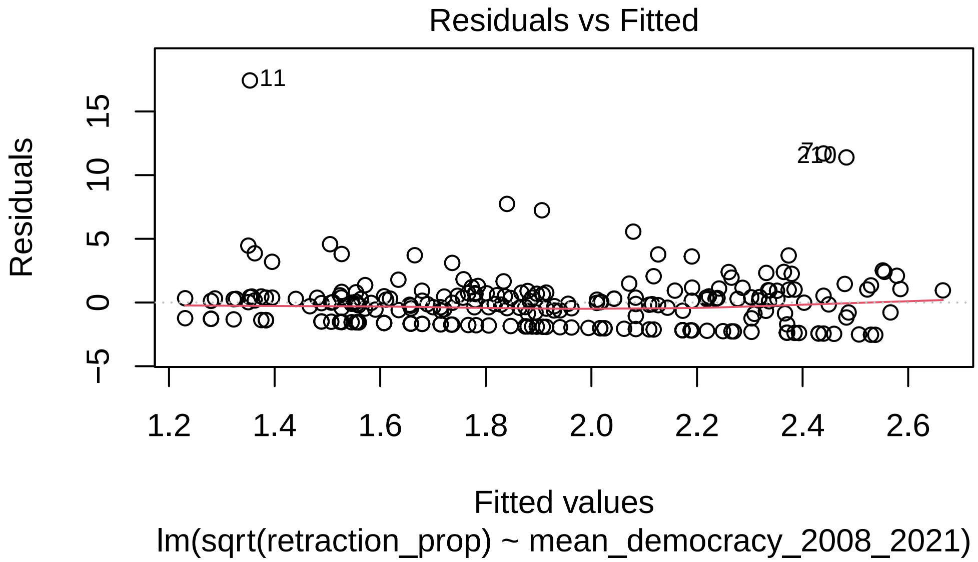
x
0.019306

	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	24.80904	1	691.8318	3.7527	0.053128	0.019306	0.019306
Residual	1260.23513	NA	NA	NA	NA	NA	NA

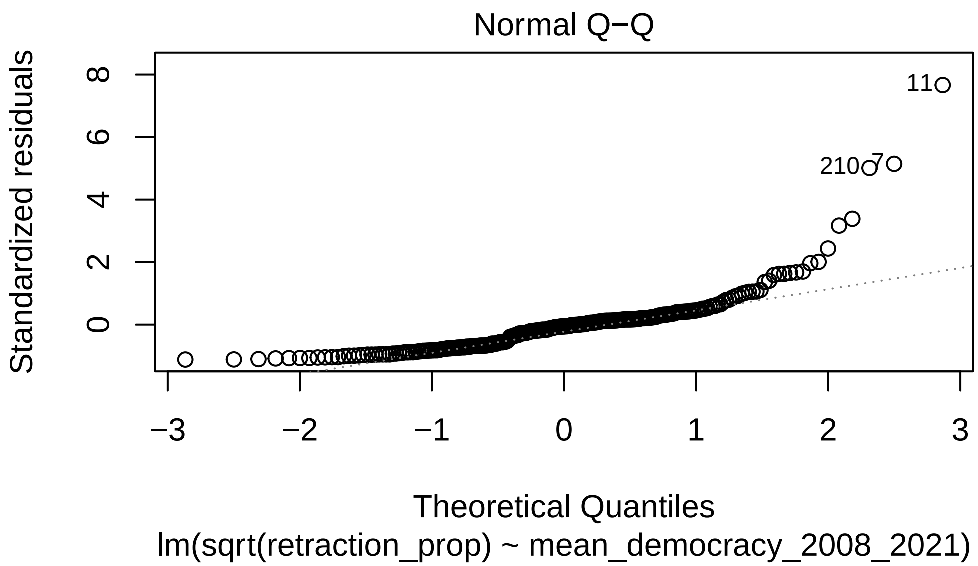
x
2

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  sqrt(retraction_prop)~mean_democracy_2008_2021
R^2=0.0193
..........................................................................
ANOVA Table
                                SSQ df1      df2 F value  Pr(>F)    eta2
mean_democracy_2008_2021   24.80904   1 691.8318  3.7527 0.05313 0.01931
Residual                 1260.23513  NA       NA      NA      NA      NA
                         partial.eta2
mean_democracy_2008_2021      0.01931
Residual                           NA
```

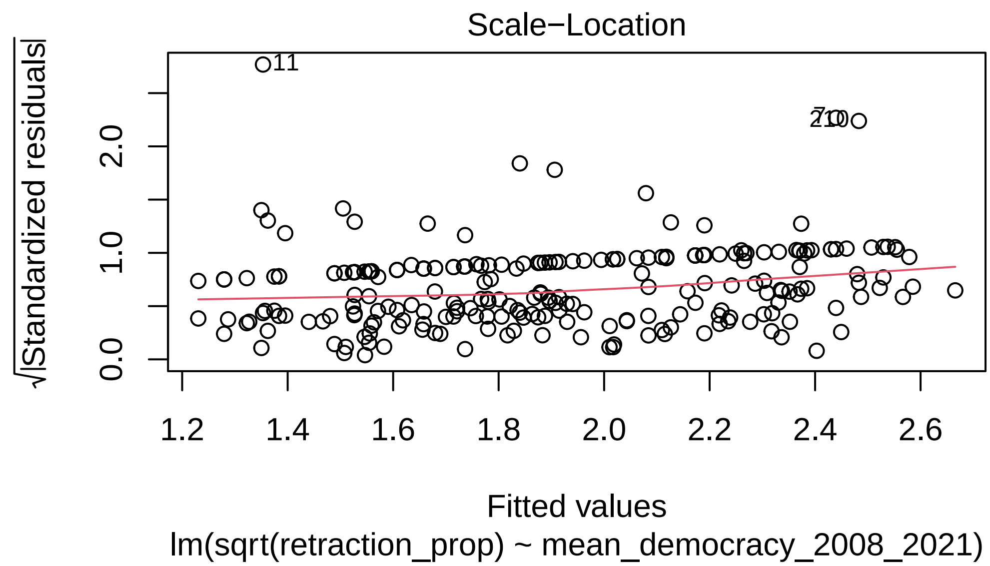
```
fitimp_linear_uni_sqrt = with(data = new_imp,
               lm(sqrt(retraction_prop)~mean_democracy_2008_2021))

plot(fitimp_linear_uni_sqrt$analyses[[1]])
```

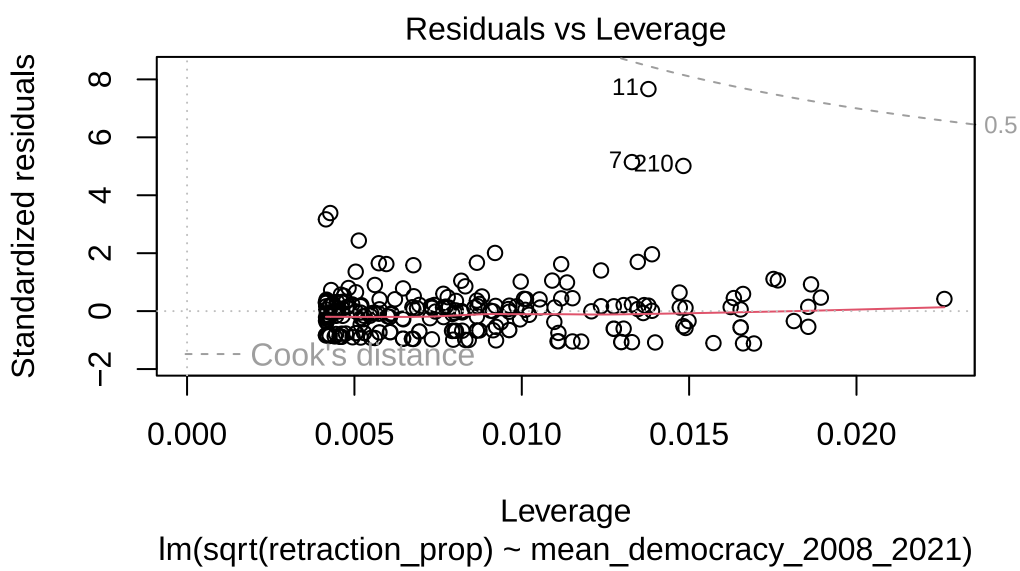


Residuals vs Fitted
Residuals
−5
0
5
10
15
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
11
7
210
Fitted values
lm(sqrt(retraction_prop) ~ mean_democracy_2008_2021)



Normal Q−Q
Standardized residuals
0
2
4
6
8
−3
−2
−1
0
1
2
3
11
210
7
Theoretical Quantiles
lm(sqrt(retraction_prop) ~ mean_democracy_2008_2021)





The plots, the F-test and R-squared all showing fitting improvements. Let's check log version.

3.2.1.2 Log method

Since we have zeros in this dataset, we cannot use log. We have two options in this regards:

- Adding 1: log(y + 1)
- Adding half the minimum non-0 value: log(y + min(y[y>0])/2)

x
0.0373222

	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	12.84921	1	644.3777	7.5071	0.006316	0.037322	0.037322
Residual	331.42907	NA	NA	NA	NA	NA	NA

x
2

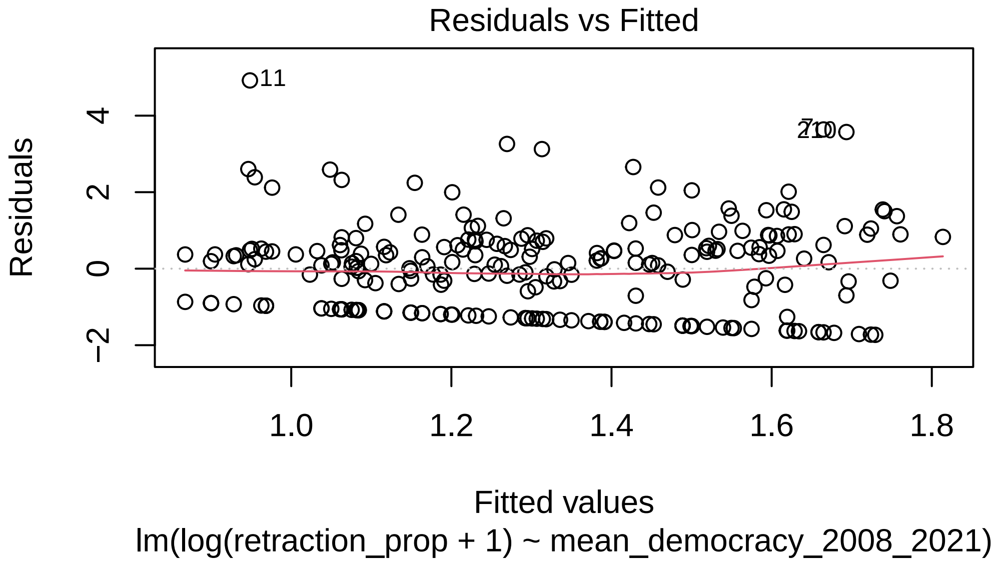
3.2.1.2.1 Adding 1 to log

We can use either log1p function or log(y+1). Here, I use log(y+1):

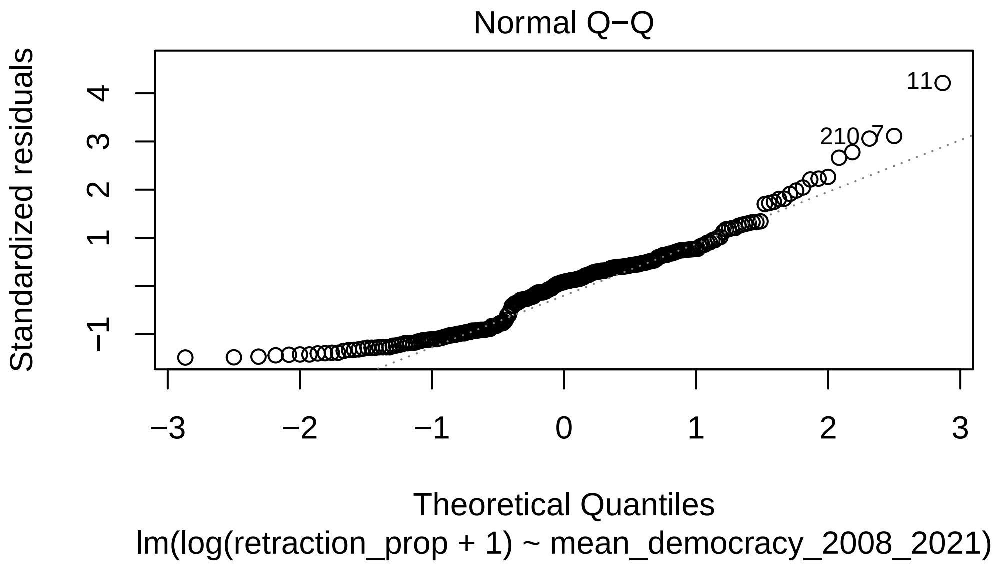
```r
kable(mi.anova(mi.res=new_imp, formula="log(retraction_prop+1)~mean_democracy_2008_2021"))
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retraction_prop+1)~mean_democracy_2008_2021
R^2=0.0373
..........................................................................
ANOVA Table
                                SSQ df1      df2 F value  Pr(>F)    eta2
mean_democracy_2008_2021   12.84921   1 644.3777  7.5071 0.00632 0.03732
Residual                  331.42907  NA       NA      NA      NA      NA
                          partial.eta2
mean_democracy_2008_2021       0.03732
Residual                            NA
```

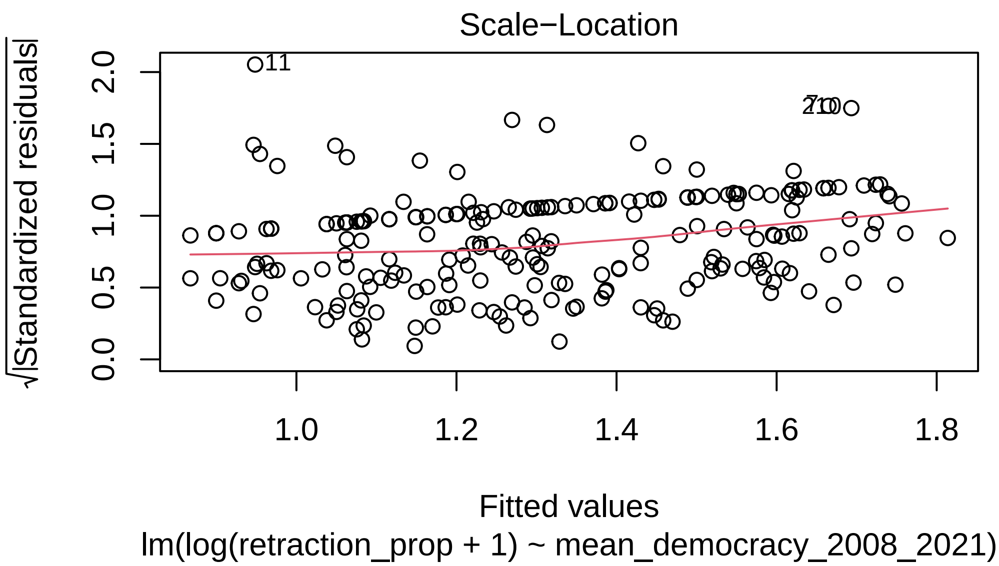
```r
fitimp_linear_uni_log1 = with(data = new_imp,
                lm(log(retraction_prop+1)~mean_democracy_2008_2021))

plot(fitimp_linear_uni_log1$analyses[[1]])
```

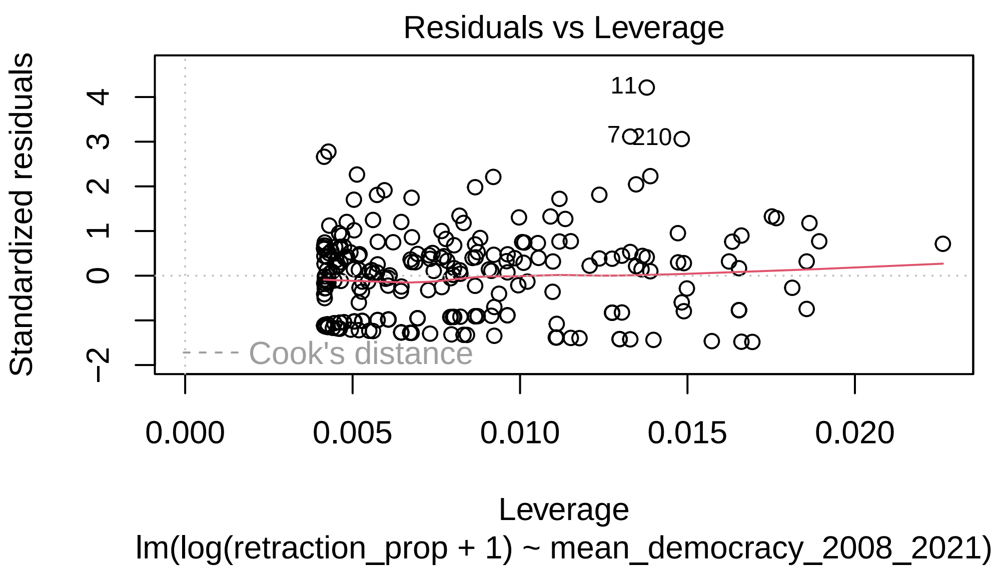


Residuals vs Fitted
Residuals
Fitted values
lm(log(retraction_prop + 1) ~ mean_democracy_2008_2021)
11
7
210



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(retraction_prop + 1) ~ mean_democracy_2008_2021)
11
210
7





The fitness is clearly better than the previous models. Now, let's check other options.

3.2.1.2.2 Adding half the minimum non-0 value to log

```
kable(mi.anova(mi.res=new_imp, formula="log(retraction_prop + min(retraction_prop[retracti
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)
```

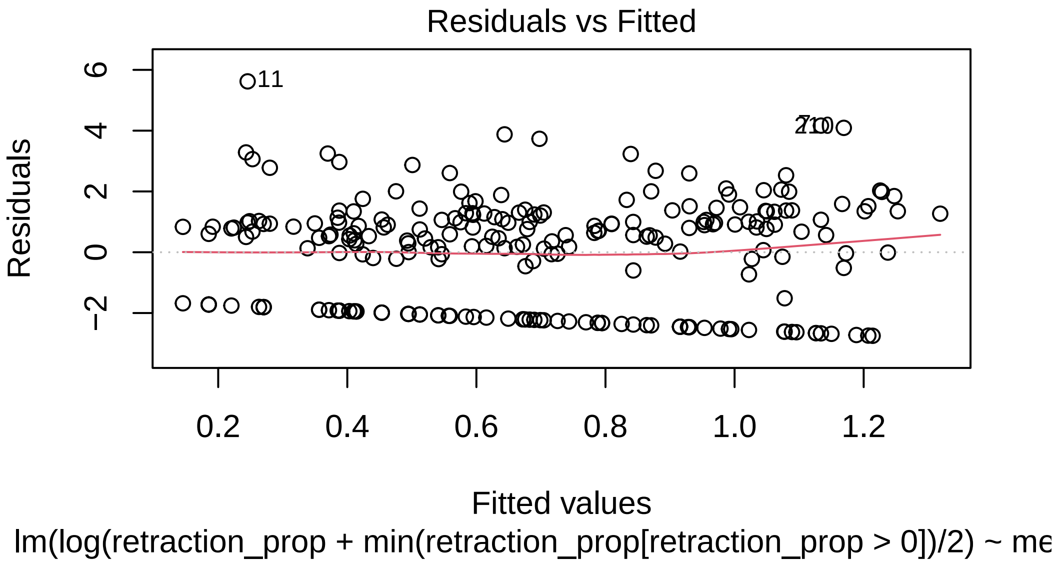
x
0.0289004

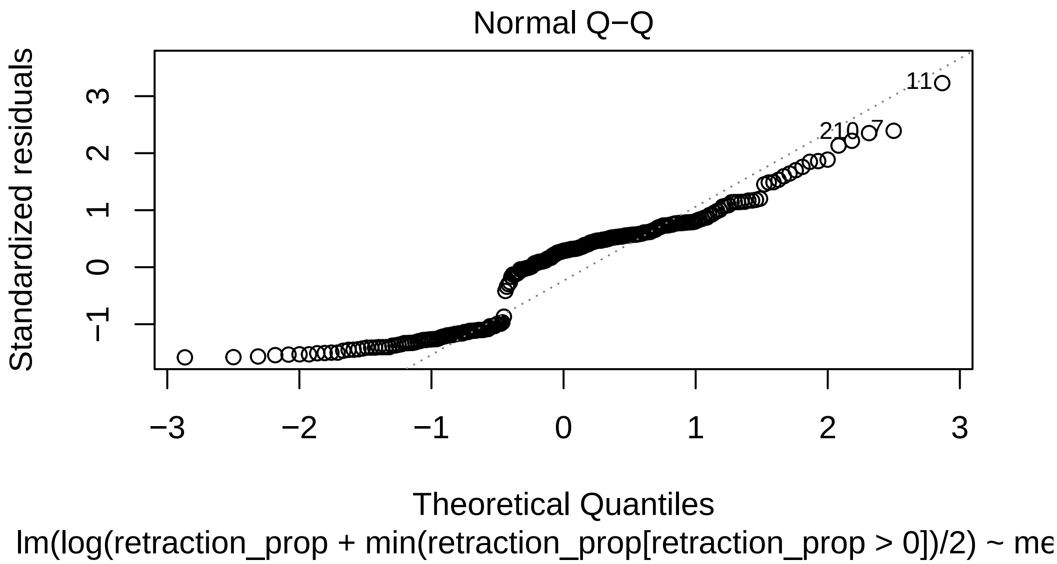
	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	21.8318	1	491.1434	5.5157	0.019242	0.0289	0.0289
Residual	733.5830	NA	NA	NA	NA	NA	NA

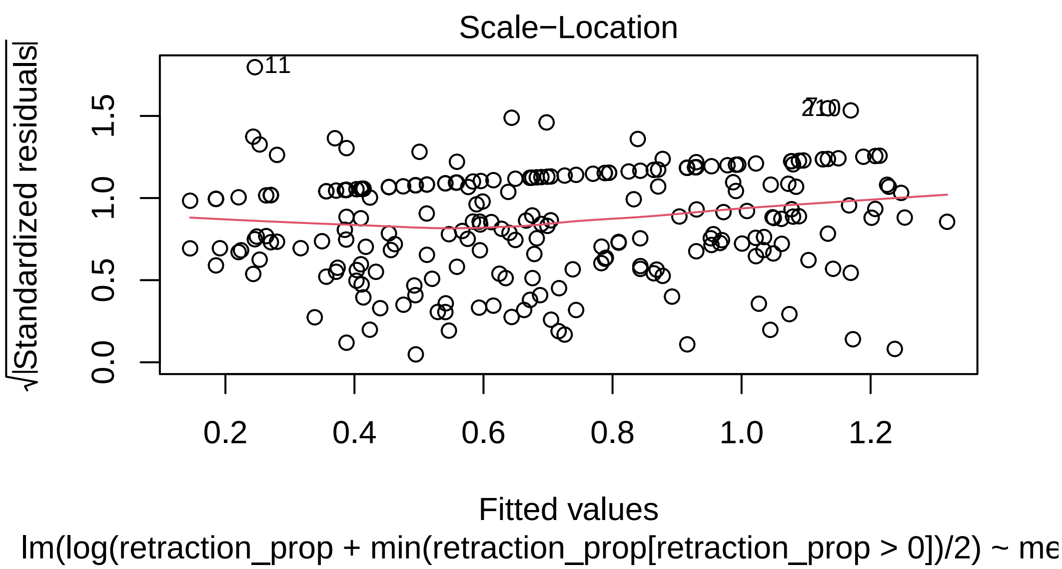
x
2

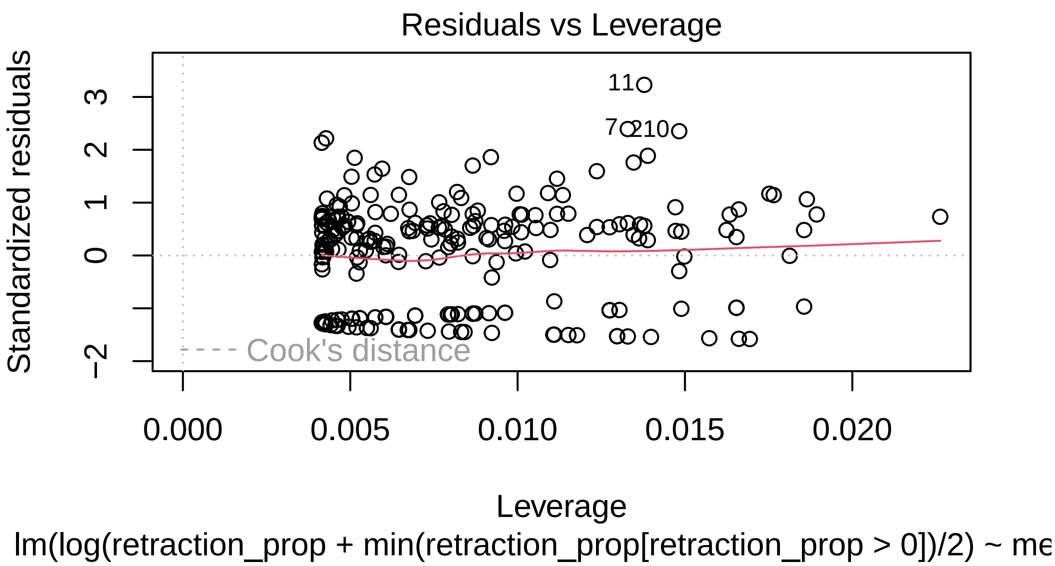
```
lm Formula:  log(retraction_prop + min(retraction_prop[retraction_prop>0])/2)~mean_democracy_
R^2=0.0289
..........................................................................
ANOVA Table
                              SSQ df1      df2 F value  Pr(>F)   eta2
mean_democracy_2008_2021  21.8318   1 491.1434  5.5157 0.01924 0.0289
Residual                 733.5830  NA       NA      NA      NA     NA
                         partial.eta2
mean_democracy_2008_2021       0.0289
Residual                           NA
```

```
fitimp_linear_uni_loghalf = with(data = new_imp,
               lm(log(retraction_prop + min(retraction_prop[retraction_prop>0])/2)~mean_de

plot(fitimp_linear_uni_loghalf$analyses[[1]])
```




Normal Q−Q
Standardized residuals
−1
0
1
2
3
−3
−2
−1
0
1
2
3
11
210
7
Theoretical Quantiles
lm(log(retraction_prop + min(retraction_prop[retraction_prop > 0])/2) ~ me


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.2
0.4
0.6
0.8
1.0
1.2
11
210
Fitted values
lm(log(retraction_prop + min(retraction_prop[retraction_prop > 0])/2) ~ me



It seems the first methods had a better fit. Therefore, we proceed with that.

The results of the unadjusted model:

```
kable(summary(pool(fitimp_linear_uni_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	1.9057740	0.2300245	8.285091	158.1861	0.000000
mean_democracy_2008_2021	-0.1020186	0.0370683	-2.752178	153.0152	0.006636

And the adjusted model:

```
fitimp_linear_multi_log1 = with(data = new_imp, lm(log(retraction_prop+1)~mean_democracy_2

kable(summary(pool(fitimp_linear_multi_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	2.0914519	0.5333387	3.9214330	103.90824	0.0001581
mean_democracy_2008_2021	- 0.0330780	0.0671030	- 0.4929443	82.73702	0.6233577
ongoing_nonviolent_campaign1	0.2648944	0.2590495	1.0225625	223.49853	0.3076203
GDP_pc_mean_1960_2022	- 0.0000030	0.0000080	- 0.3690309	93.81957	0.7129362

term	estimate	std.error	statistic	df	p.value
regionAfrica	-0.7580364	0.2234373	-3.3926129	221.59688	0.0008199
regionAmericas	-0.8628296	0.2567860	-3.3601113	207.46553	0.0009272
regionEurope	-0.8249688	0.2793362	-2.9533189	195.58007	0.0035291
regionOceania	-1.4823076	0.2869346	-5.1660126	217.52891	0.0000005
industry_share_mean_1960_2022	0.0052842	0.0076063	0.6947097	138.09903	0.4884044
length_of_last_leader_tenure_2015	0.0017707	0.0127343	0.1390493	66.71407	0.8898301
muslim_proportion	-0.0013651	0.0030276	-0.4508734	110.70922	0.6529626
top_universities_shanghai_2022	0.0036081	0.0040808	0.8841558	223.23298	0.3775634
plurality1	-0.0015985	0.2003378	-0.0079789	82.21041	0.9936532

Let's check the fitness of the model:

```
kable(mi.anova(mi.res=new_imp, formula="log(retraction_prop+1)~mean_democracy_2008_2021+on
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retraction_prop+1)~mean_democracy_2008_2021+ongoing_nonviolent_campaign+GDP_
R^2=0.1853
..........................................................................
ANOVA Table
                                        SSQ df1          df2 F value  Pr(>F)
mean_democracy_2008_2021          12.84921   1     564.0018  8.3351 0.00404
ongoing_nonviolent_campaign        4.98916   1 2008619.3989  4.0398 0.04444
GDP_pc_mean_1960_2022              0.47485   1     883.6735  0.1692 0.68091
region                            41.18374   4   15323.2798  8.0677 0.00000
industry_share_mean_1960_2022      1.06767   1     483.2553  0.4788 0.48930
length_of_last_leader_tenure_2015  0.79542   1     516.1536  0.3146 0.57510
muslim_proportion                  0.89419   1     333.1986  0.2942 0.58793
top_universities_shanghai_2022     0.98882   1  184806.1450  0.7847 0.37570
plurality                          0.54089   1    1043.7284  0.2330 0.62941
Residual                         280.49432  NA           NA      NA      NA
                                     eta2 partial.eta2
mean_democracy_2008_2021          0.03732      0.04380
```

x
0.1852686

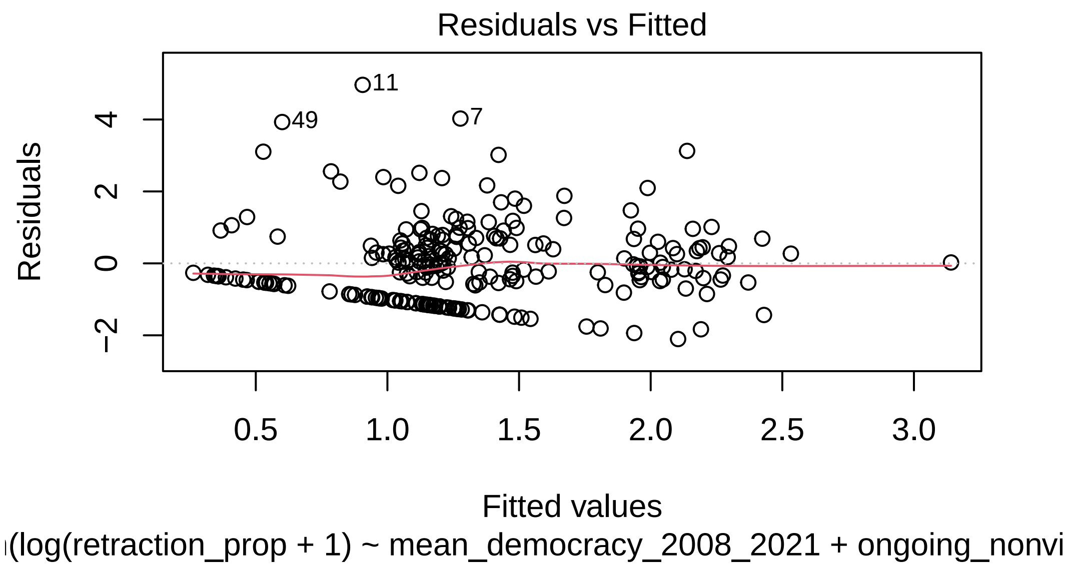
	SSQ	df1	df2	F value	Pr(>F)	eta2	p
mean__democracy__2008__2021	12.8492137	1	564.0018	8.3351	0.004038	0.037322	
ongoing__nonviolent__campaign	4.9891640	1	2008619.3989	4.0398	0.044438	0.014492	
GDP__pc__mean__1960__2022	0.4748538	1	883.6735	0.1692	0.680910	0.001379	
region	41.1837400	4	15323.2798	8.0677	0.000002	0.119623	
industry__share__mean__1960__2022	1.0676738	1	483.2553	0.4788	0.489305	0.003101	
length__of__last__leader__tenure__2015	0.7954232	1	516.1536	0.3146	0.575105	0.002310	
muslim__proportion	0.8941946	1	333.1986	0.2942	0.587932	0.002597	
top__universities__shanghai__2022	0.9888172	1	184806.1450	0.7847	0.375699	0.002872	
plurality	0.5408880	1	1043.7284	0.2330	0.629409	0.001571	
Residual	280.4943164	NA	NA	NA	NA	NA	

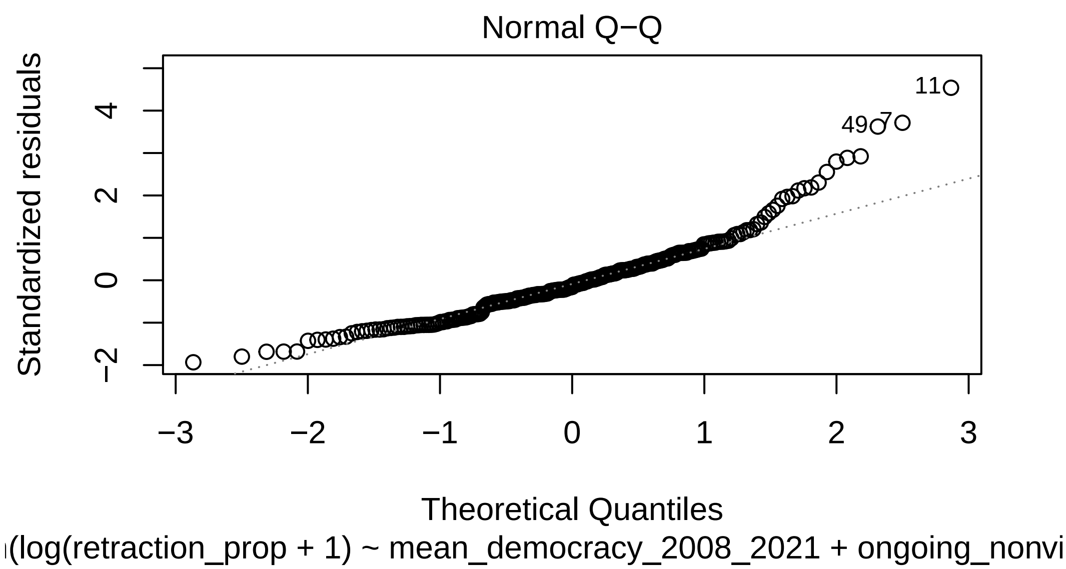
x
2

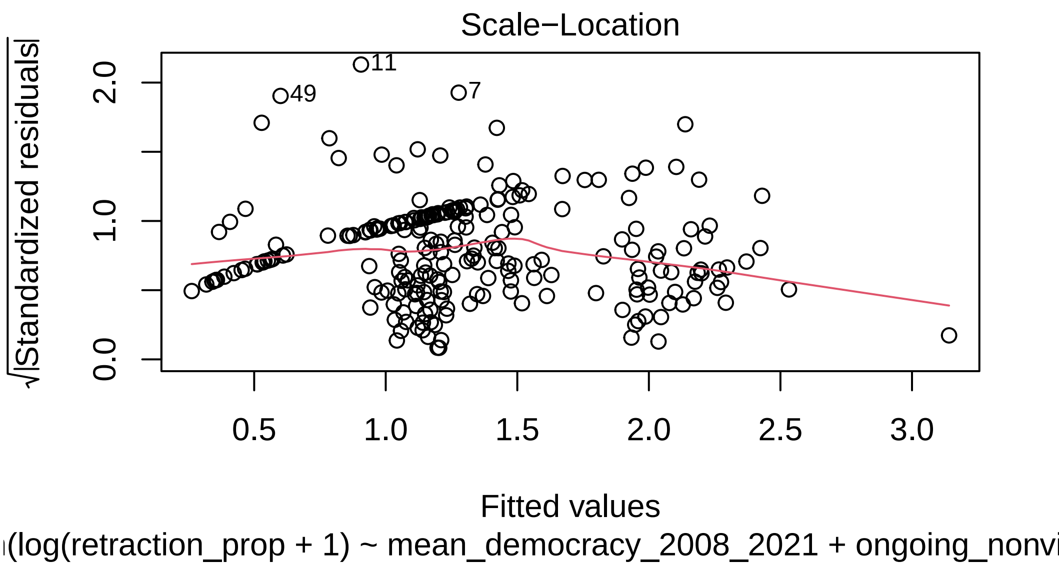
```
ongoing_nonviolent_campaign       0.01449      0.01748
GDP_pc_mean_1960_2022             0.00138      0.00169
region                            0.11962      0.12803
industry_share_mean_1960_2022     0.00310      0.00379
length_of_last_leader_tenure_2015 0.00231      0.00283
muslim_proportion                 0.00260      0.00318
top_universities_shanghai_2022    0.00287      0.00351
plurality                         0.00157      0.00192
Residual                               NA           NA
```

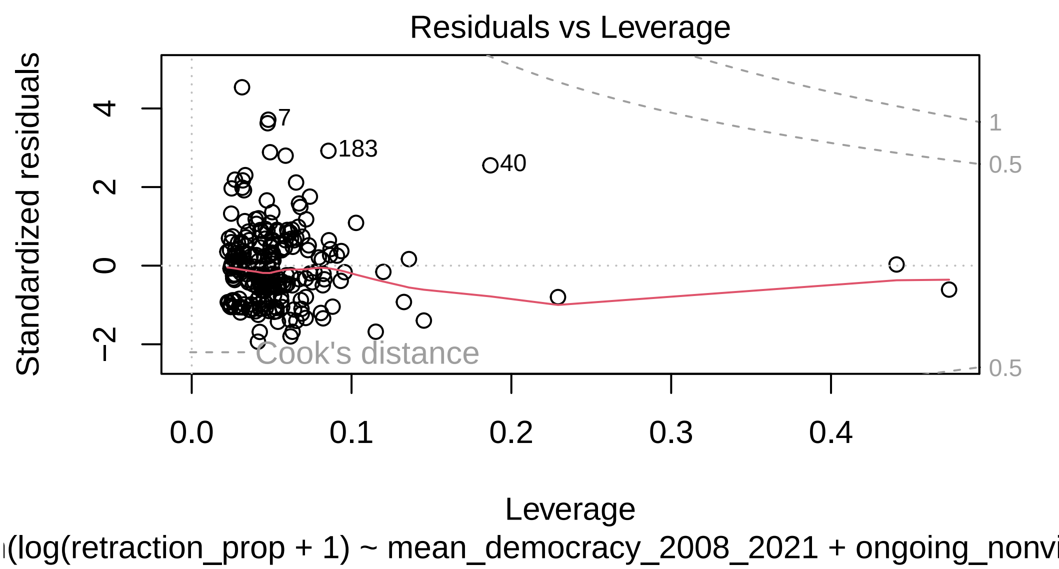
And plots:

```r
plot(fitimp_linear_multi_log1$analyses[[1]])
```


Residuals vs Fitted
Residuals
Fitted values
11
49
7
4
2
0
−2
0.5
1.0
1.5
2.0
2.5
3.0
(log(retraction_prop + 1) ~ mean_democracy_2008_2021 + ongoing_nonvi


Normal Q−Q
Standardized residuals
Theoretical Quantiles
11
49
7
4
2
0
−2
−3
−2
−1
0
1
2
3
(log(retraction_prop + 1) ~ mean_democracy_2008_2021 + ongoing_nonvi

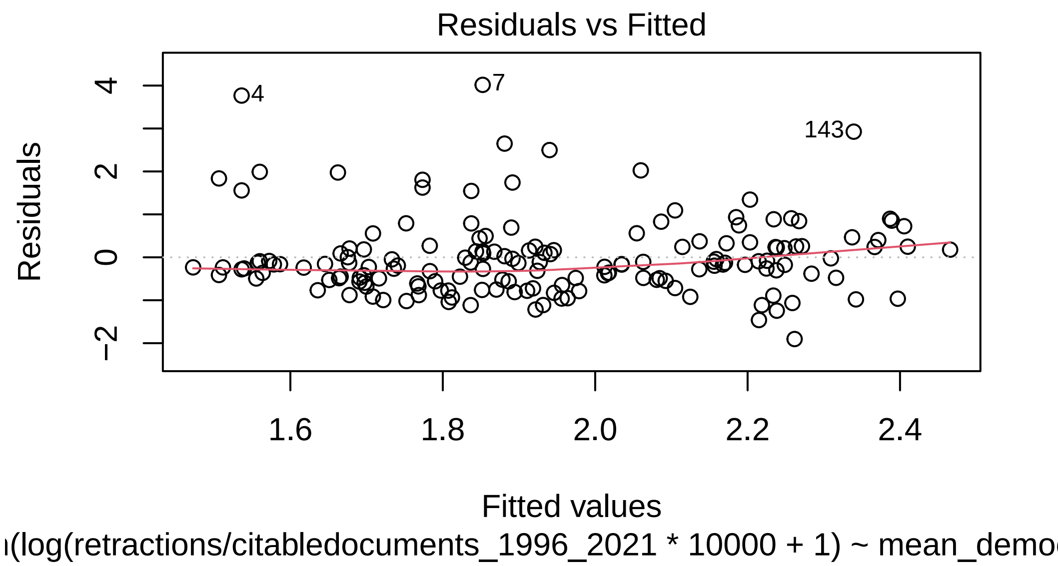


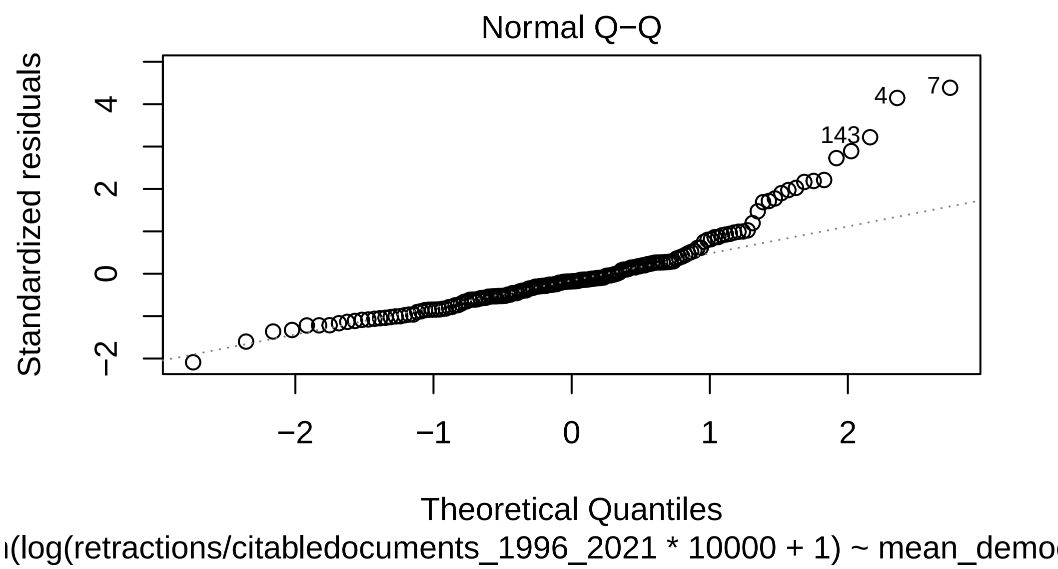


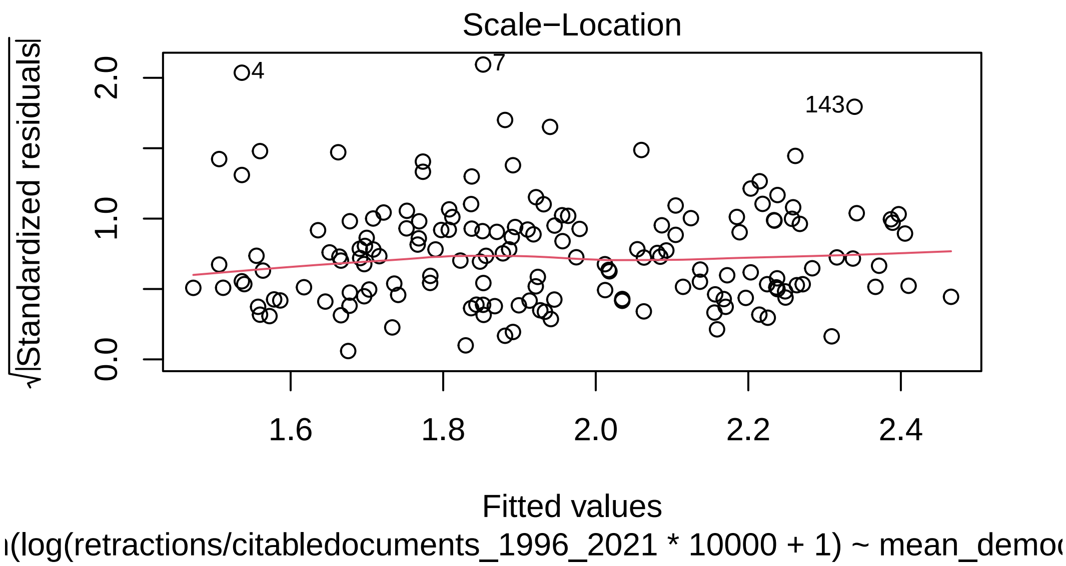
3.2.2 Zero-truncated dataset

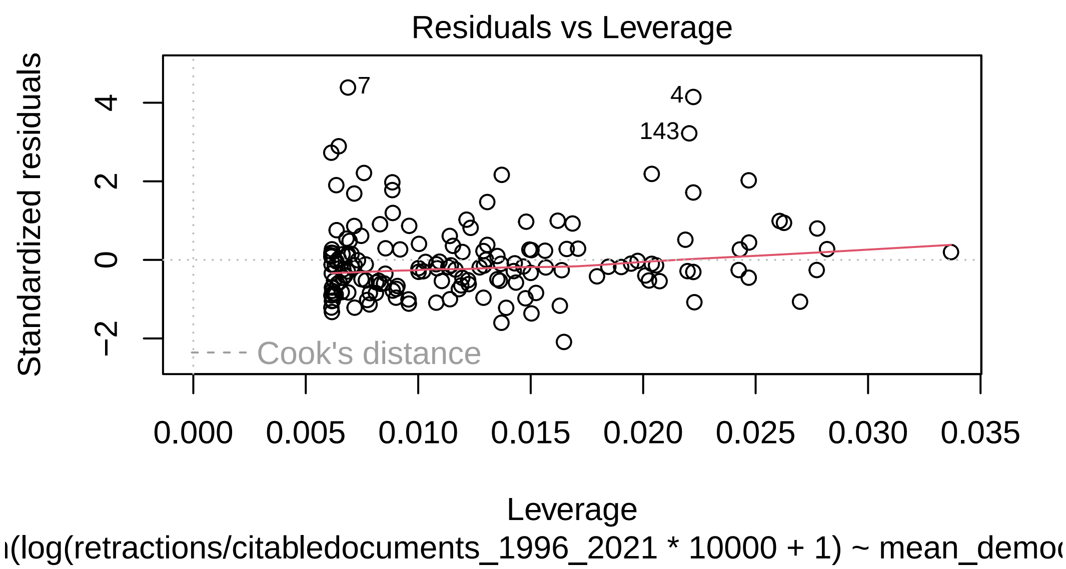
For this one and the next dataset, I only use the log(y+1) method.

```
fit_truncimp_linear_uni_log1 = with(data = trunc_imp,
               lm(log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_

plot(fit_truncimp_linear_uni_log1$analyses[[1]])
```


Residuals vs Fitted
Residuals
Fitted values
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_democ
4
7
143
1.6
1.8
2.0
2.2
2.4
−2
0
2
4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_democ
4
7
143
−2
−1
0
1
2
−2
0
2
4





And fitness tests:

```
kable(mi.anova(mi.res=trunc_imp, formula="log(retractions/citabledocuments_1996_2021*10000
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_2021
R^2=0.0711
```

x
0.071069

	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	10.37433	1	451.4755	9.6191	0.002046	0.071069	0.071069
Residual	135.60111	NA	NA	NA	NA	NA	NA

x
2

```
.........................................................................
ANOVA Table
                               SSQ df1      df2 F value  Pr(>F)    eta2
mean_democracy_2008_2021  10.37433   1 451.4755  9.6191 0.00205 0.07107
Residual                 135.60111  NA       NA      NA      NA      NA
                         partial.eta2
mean_democracy_2008_2021      0.07107
Residual                           NA
```

The fitness of model seems satisfactory. Here is the results of the unadjusted model:

```
kable(summary(pool(fit_truncimp_linear_uni_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	2.5828428	0.2144465	12.044228	119.3842	0.0000000
mean_democracy_2008_2021	-0.1126113	0.0355596	-3.166833	109.4047	0.0019973

Now, let's perform the full model:

```
fit_truncimp_linear_multi_log1 = with(data = trunc_imp, lm(log(retractions/citabledocument

kable(summary(pool(fit_truncimp_linear_multi_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	3.2155655	0.4763111	6.7509780	120.83355	0.0000000
mean_democracy_2008_2021	- 0.1403551	0.0623950	- 2.2494604	92.30232	0.0268606
ongoing_nonviolent_campaign1	0.1755619	0.2271682	0.7728276	146.27234	0.4408715
GDP_pc_mean_1960_2022	0.0000129	0.0000083	1.5637076	111.68544	0.1207154

term	estimate	std.error	statistic	df	p.value
regionAfrica	-0.2926486	0.2041441	-1.4335392	145.38393	0.1538509
regionAmericas	0.0541502	0.2443008	0.2216538	142.38046	0.8249010
regionEurope	-0.6228892	0.2471175	-2.5206201	133.54870	0.0128917
regionOceania	0.3628350	0.4093236	0.8864257	144.16607	0.3768644
industry_share_mean_1960_2022	-0.0140723	0.0073186	-1.9228152	127.23480	0.0567380
length_of_last_leader_tenure_2015	-0.0011832	0.0105537	-0.1121134	90.54233	0.9109817
muslim_proportion	0.0015086	0.0027472	0.5491660	126.61325	0.5838588
top_universities_shanghai_2022	-0.0008988	0.0033608	-0.2674308	146.40583	0.7895138
plurality1	-0.1184219	0.1919484	-0.6169464	68.37610	0.5393190

Let's check the fitness of the model:

```
kable(mi.anova(mi.res=trunc_imp, formula="log(retractions/citabledocuments_1996_2021*10000
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_2021+ong
R^2=0.1894
..........................................................................
ANOVA Table
                                        SSQ df1          df2 F value  Pr(>F)
mean_democracy_2008_2021           10.37433   1 4.051607e+02 10.1248 0.00158
ongoing_nonviolent_campaign         0.07817   1 1.116252e+09  0.0990 0.75303
GDP_pc_mean_1960_2022               1.41333   1 5.070027e+02  1.2482 0.26443
region                             11.38836   4 5.770112e+04  3.5290 0.00694
industry_share_mean_1960_2022       2.95219   1 2.390161e+03  3.3192 0.06860
length_of_last_leader_tenure_2015   0.22273   1 1.537175e+03  0.1296 0.71893
muslim_proportion                   0.40139   1 2.243990e+03  0.3605 0.54828
top_universities_shanghai_2022      0.08448   1 3.272903e+05  0.0975 0.75484
plurality                           0.73569   1 3.082045e+02  0.4300 0.51248
Residual                          118.32476  NA           NA      NA      NA
                                     eta2 partial.eta2
mean_democracy_2008_2021          0.07107      0.08061
```

x
0.1894201

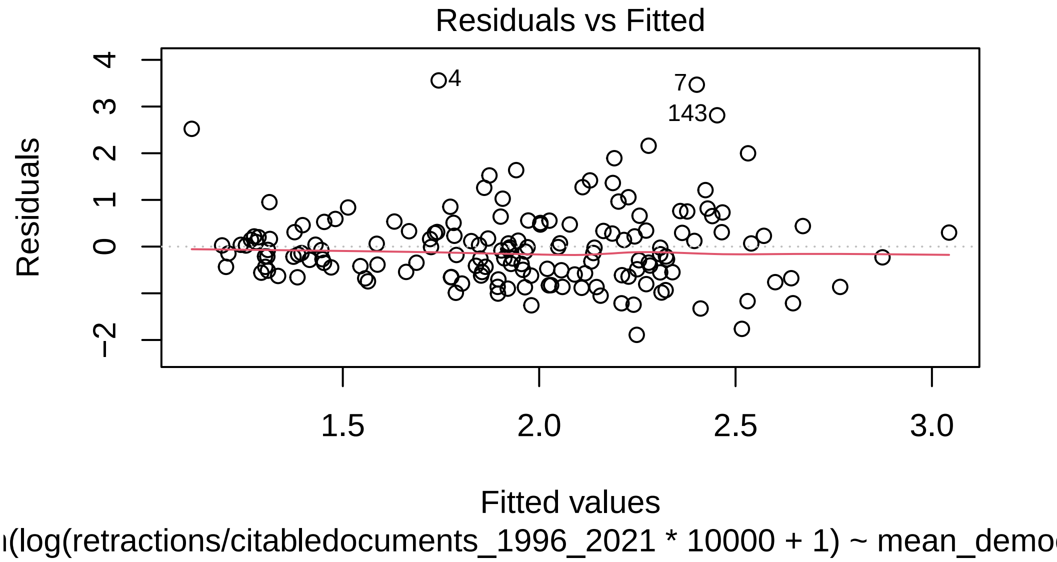
	SSQ	df1	df2	F value	Pr(>F)	eta2
mean__democracy__2008__2021	10.3743323	1	4.051607e+02	10.1248	0.001576	0.071069
ongoing__nonviolent__campaign	0.0781742	1	1.116252e+09	0.0990	0.753034	0.000536
GDP__pc__mean__1960__2022	1.4133330	1	5.070027e+02	1.2482	0.264428	0.009682
region	11.3883628	4	5.770112e+04	3.5290	0.006938	0.078016
industry__share__mean__1960__2022	2.9521890	1	2.390161e+03	3.3192	0.068598	0.020224
length__of__last__leader__tenure__2015	0.2227289	1	1.537175e+03	0.1296	0.718928	0.001526
muslim__proportion	0.4013872	1	2.243990e+03	0.3605	0.548282	0.002750
top__universities__shanghai__2022	0.0844802	1	3.272903e+05	0.0975	0.754845	0.000579
plurality	0.7356940	1	3.082045e+02	0.4300	0.512477	0.005040
Residual	118.3247605	NA	NA	NA	NA	NA

x
2

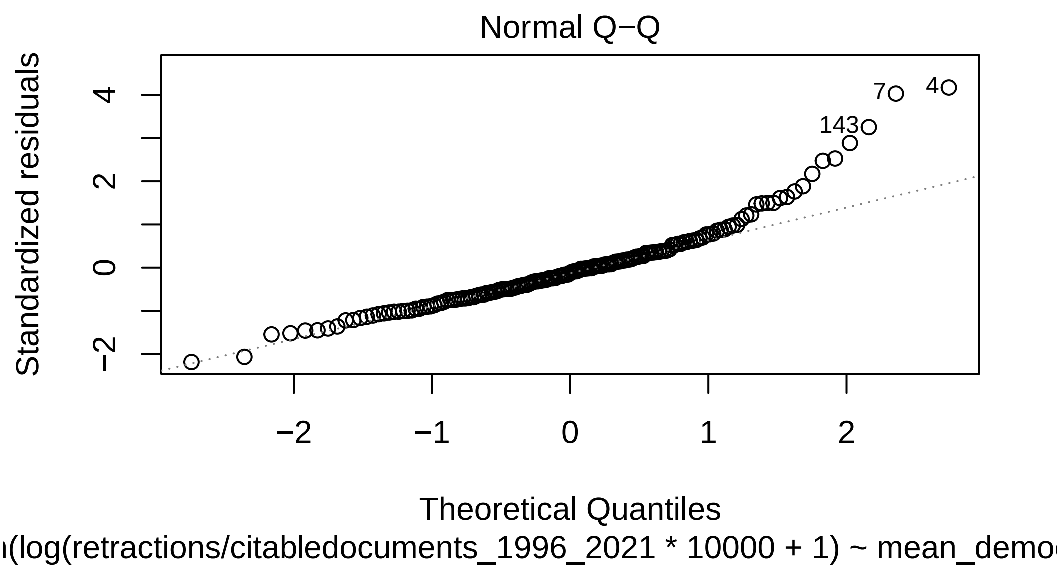
```
ongoing_nonviolent_campaign       0.00054      0.00066
GDP_pc_mean_1960_2022             0.00968      0.01180
region                            0.07802      0.08780
industry_share_mean_1960_2022     0.02022      0.02434
length_of_last_leader_tenure_2015 0.00153      0.00188
muslim_proportion                 0.00275      0.00338
top_universities_shanghai_2022    0.00058      0.00071
plurality                         0.00504      0.00618
Residual                               NA           NA
```

And plots:

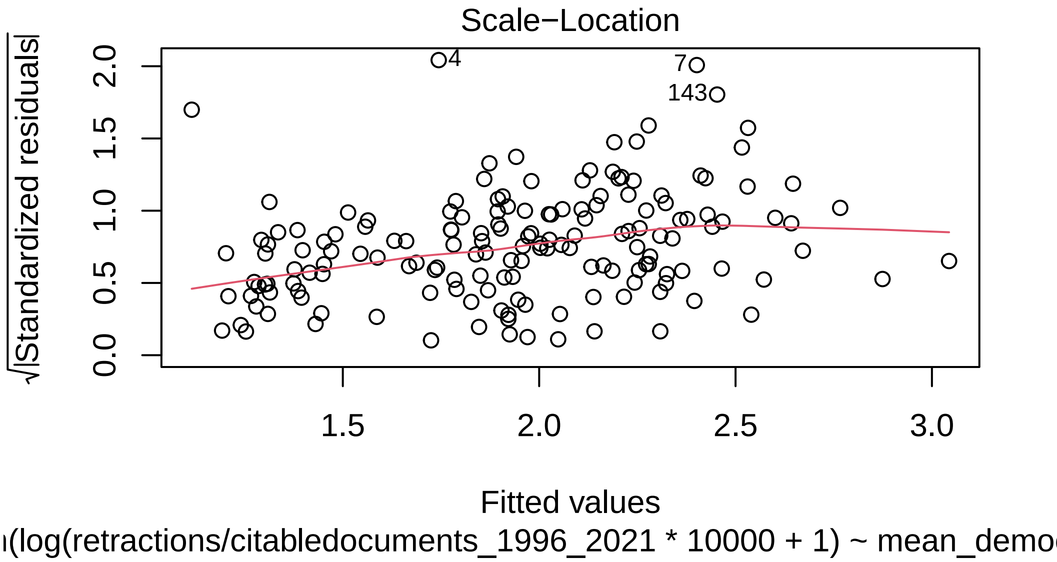
```
plot(fit_truncimp_linear_multi_log1$analyses[[1]])
```

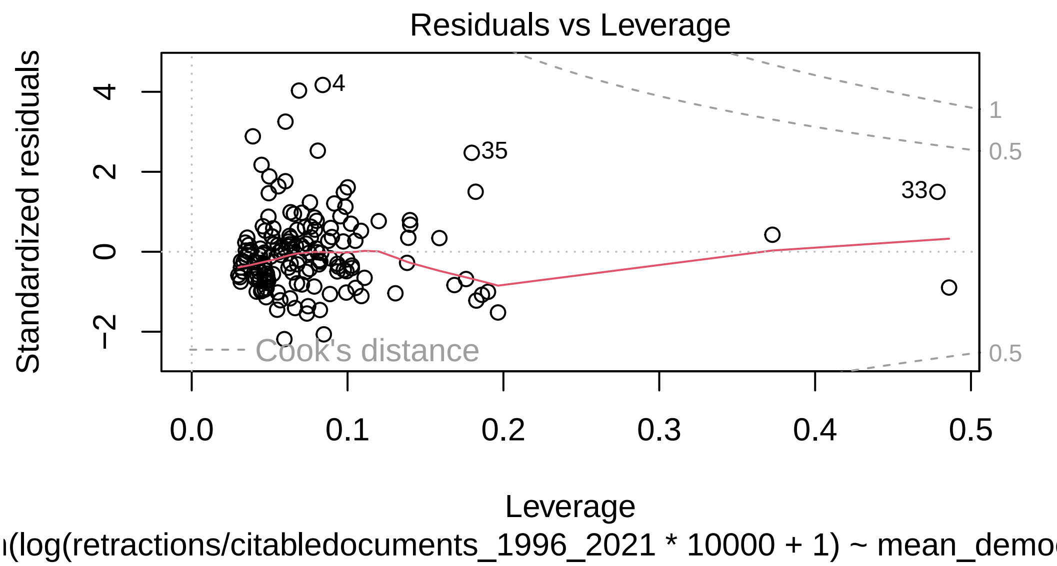


Residuals vs Fitted
Residuals
Fitted values
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo



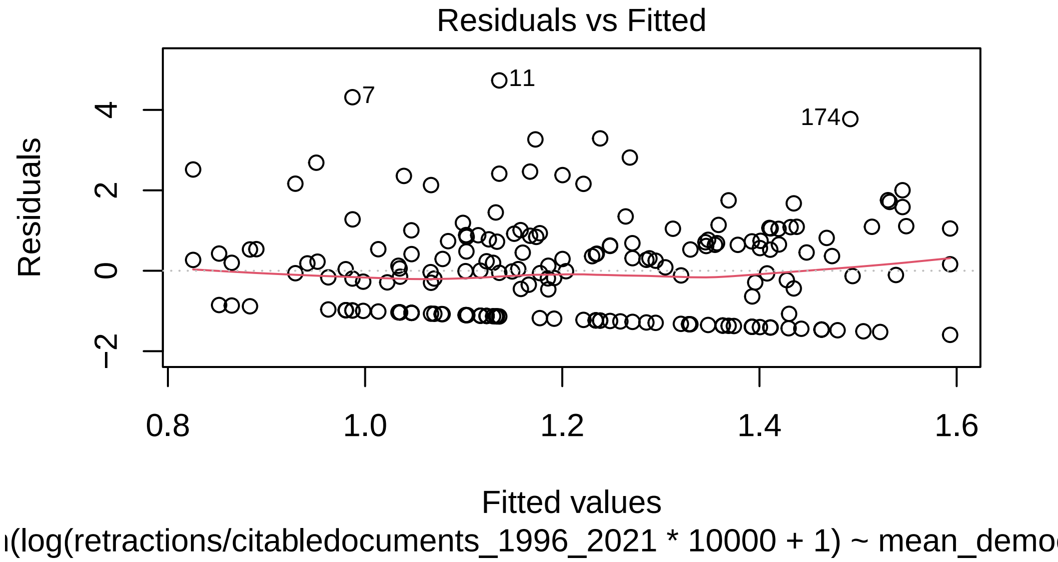
Normal Q−Q
Standardized residuals
Theoretical Quantiles
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo

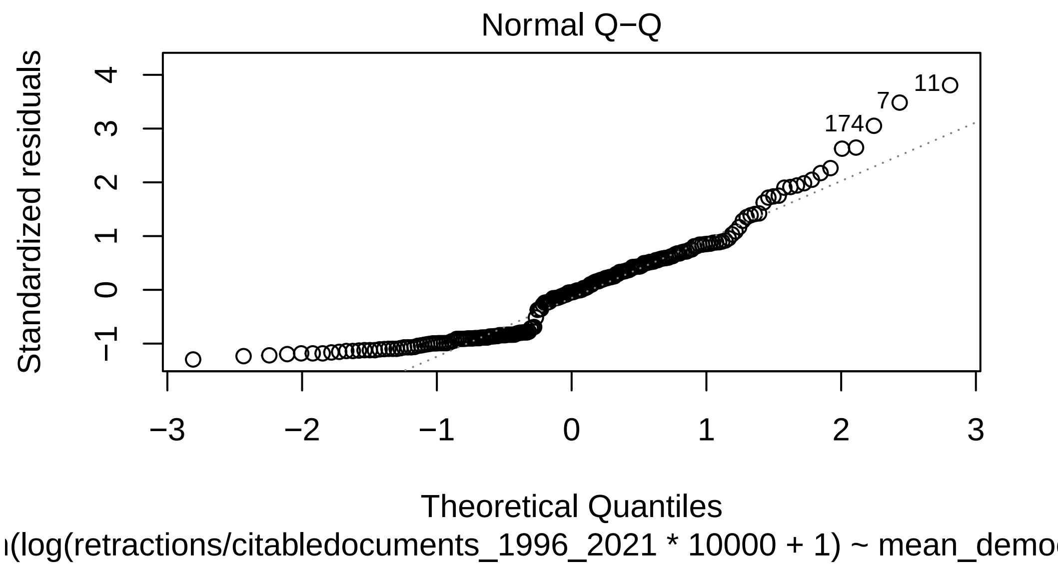


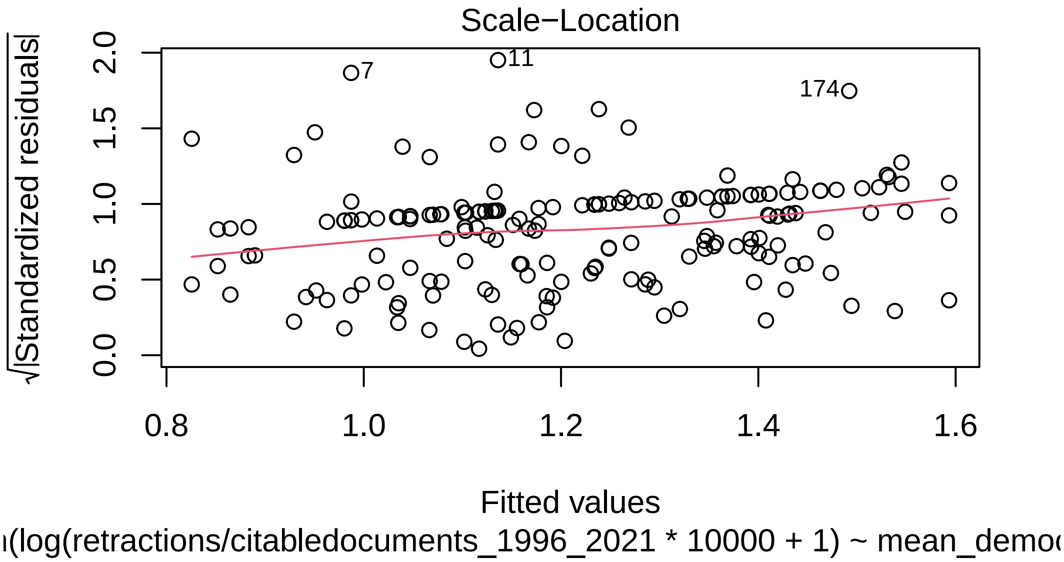


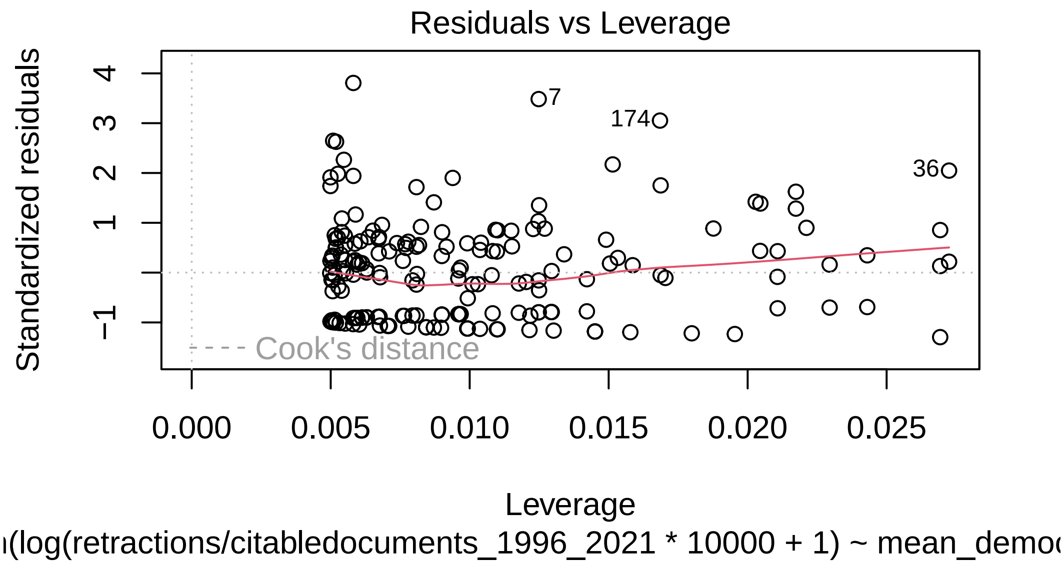
3.2.3 Outlier-removed dataset

```
fit_nooutimp_linear_uni_log1 = with(data = no_out_imp,
               lm(log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_

plot(fit_nooutimp_linear_uni_log1$analyses[[1]])
```


Residuals vs Fitted
Residuals
Fitted values
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo‹
7
11
174
−2
0
2
4
0.8
1.0
1.2
1.4
1.6


Normal Q−Q
Standardized residuals
Theoretical Quantiles
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo‹
11
7
174
−1
0
1
2
3
4
−3
−2
−1
0
1
2
3





And fitness tests:

```
kable(mi.anova(mi.res=no_out_imp, formula="log(retractions/citabledocuments_1996_2021*1000
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_2021
R^2=0.028
```

x
0.0280077

	SSQ	df1	df2	F value	Pr(>F)	eta2	partial.eta2
mean_democracy_2008_2021	8.850627	1	327.2999	4.1096	0.043451	0.028008	0.028008
Residual	307.156248	NA	NA	NA	NA	NA	NA

x
2

```
.......................................................................
ANOVA Table
                               SSQ df1      df2 F value  Pr(>F)    eta2
mean_democracy_2008_2021   8.85063   1 327.2999  4.1096 0.04345 0.02801
Residual                 307.15625  NA       NA      NA      NA      NA
                         partial.eta2
mean_democracy_2008_2021      0.02801
Residual                           NA
```

The fitness of model seems satisfactory. Here is the unadjusted model:

```
kable(summary(pool(fit_nooutimp_linear_uni_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	1.7501439	0.2796534	6.258261	109.1480	0.0000000
mean_democracy_2008_2021	-0.0994243	0.0485203	-2.049125	106.5616	0.0429065

Let’s perform the full model:

```
fit_nooutimp_linear_multi_log1 = with(data = no_out_imp, lm(log(retractions/citabledocumen

kable(summary(pool(fit_nooutimp_linear_multi_log1)))
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	2.0979962	0.5614815	3.7365365	119.99883	0.0002871
mean_democracy_2008_2021	- 0.0259751	0.0763754	- 0.3400979	94.76160	0.7345355
ongoing_nonviolent_campaign1	0.2610470	0.3618907	0.7213422	182.50780	0.4716221
GDP_pc_mean_1960_2022	- 0.0000017	0.0000078	- 0.2182974	128.25760	0.8275443

term	estimate	std.error	statistic	df	p.value
regionAfrica	-0.7568720	0.2681066	-2.8230266	182.25760	0.0052861
regionAmericas	-0.8941707	0.3339062	-2.6779101	139.52211	0.0082977
regionEurope	-1.0370882	0.3747565	-2.7673652	146.03555	0.0063830
regionOceania	-1.5500490	0.3480209	-4.4538959	157.85689	0.0000159
industry_share_mean_1960_2022	0.0032029	0.0085678	0.3738347	120.81063	0.7091825
length_of_last_leader_tenure_2015	0.0033646	0.0129660	0.2594971	97.28107	0.7958001
muslim_proportion	-0.0024624	0.0036931	-0.6667667	69.88310	0.5071167
top_universities_shanghai_2022	0.0199356	0.0630839	0.3160174	178.81135	0.7523578
plurality1	0.0089557	0.2541518	0.0352376	59.25903	0.9720087

Let's check the fitness of the model:

```
kable(mi.anova(mi.res=no_out_imp, formula="log(retractions/citabledocuments_1996_2021*1000
```

```
Univariate ANOVA for Multiply Imputed Data (Type 2)

lm Formula:  log(retractions/citabledocuments_1996_2021*10000+1)~mean_democracy_2008_2021+ong
R^2=0.1528
..........................................................................
ANOVA Table
                                       SSQ df1          df2 F value  Pr(>F)
mean_democracy_2008_2021           8.85063   1     308.3893  4.4045 0.03666
ongoing_nonviolent_campaign        1.15285   1 1414618.3655  0.8028 0.37026
GDP_pc_mean_1960_2022              0.27102   1    2265.3414  0.0716 0.78906
region                            34.09214   4    6042.2096  5.6266 0.00016
industry_share_mean_1960_2022      0.58478   1     783.8643  0.1761 0.67487
length_of_last_leader_tenure_2015  0.63428   1     885.8535  0.2210 0.63839
muslim_proportion                  1.70532   1     237.0915  0.5518 0.45833
top_universities_shanghai_2022     0.18304   1   27885.7316  0.0966 0.75597
plurality                          0.82629   1     596.8350  0.2821 0.59556
Residual                         267.70653  NA           NA      NA      NA
                                     eta2 partial.eta2
mean_democracy_2008_2021          0.02801      0.03200
ongoing_nonviolent_campaign       0.00365      0.00429
```

x
0.1528459

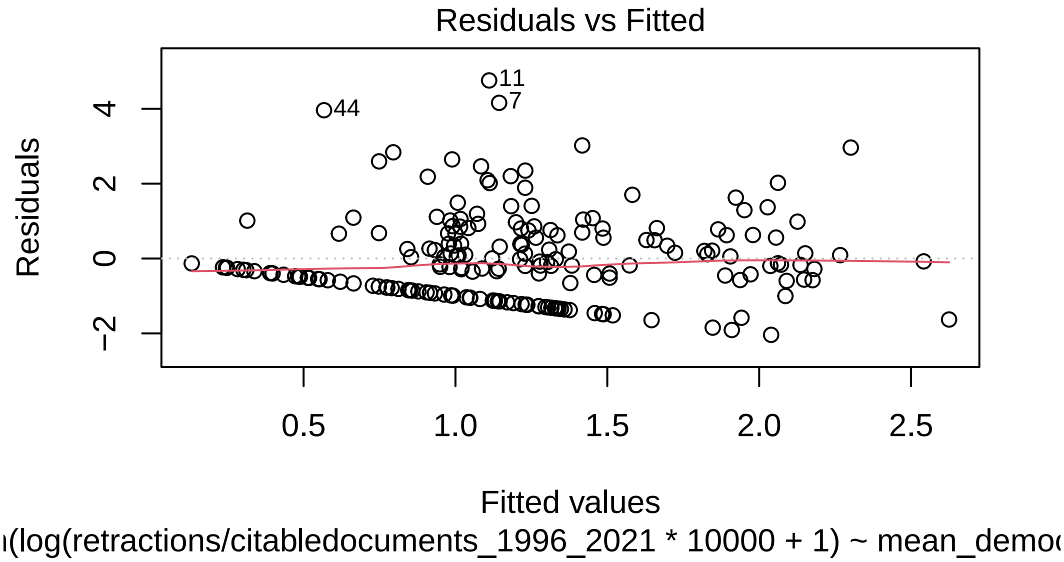
	SSQ	df1	df2	F value	Pr(>F)	eta2	p
mean__democracy__2008__2021	8.8506272	1	308.3893	4.4045	0.036657	0.028008	
ongoing__nonviolent__campaign	1.1528488	1	1414618.3656	0.8028	0.370255	0.003648	
GDP__pc__mean__1960__2022	0.2710238	1	2265.3415	0.0716	0.789061	0.000858	
region	34.0921368	4	6042.2096	5.6266	0.000162	0.107884	
industry__share__mean__1960__2022	0.5847768	1	783.8643	0.1761	0.674869	0.001851	
length__of__last__leader__tenure__2015	0.6342803	1	885.8535	0.2210	0.638393	0.002007	
muslim__proportion	1.7053214	1	237.0915	0.5518	0.458335	0.005396	
top__universities__shanghai__2022	0.1830447	1	27885.7316	0.0966	0.755968	0.000579	
plurality	0.8262863	1	596.8350	0.2821	0.595557	0.002615	
Residual	267.7065296	NA	NA	NA	NA	NA	

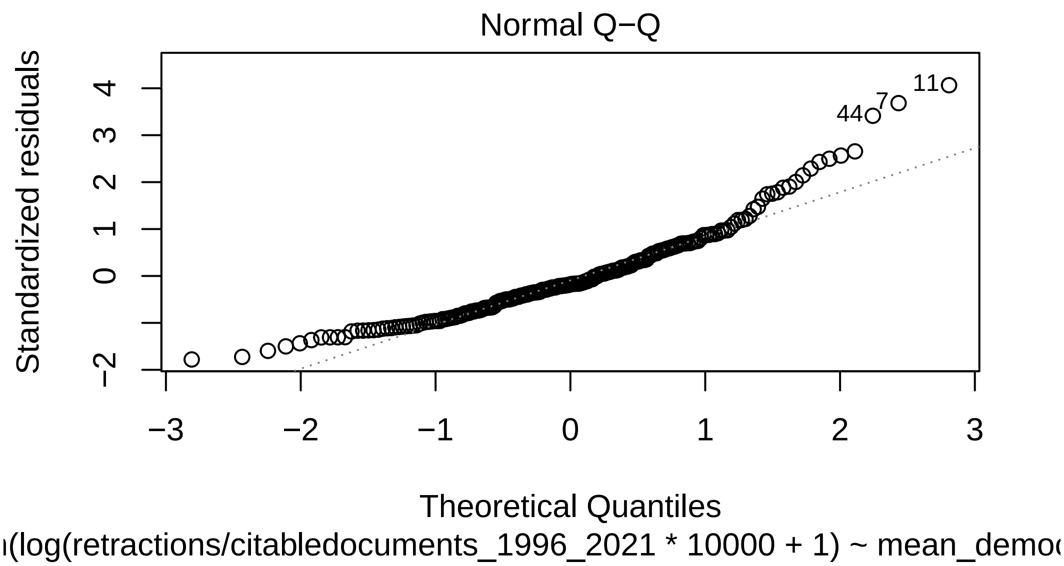
x
2

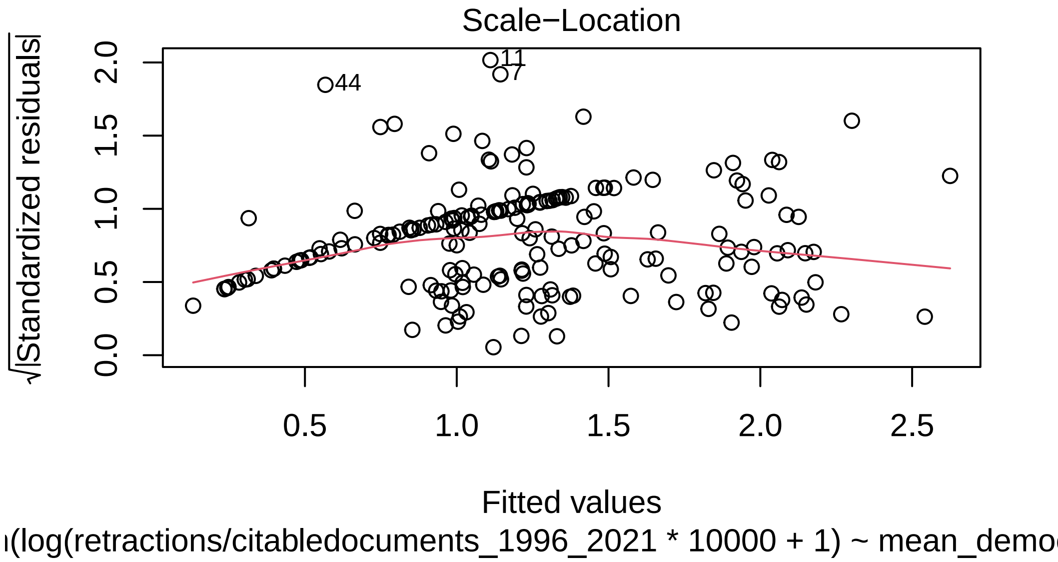
```
GDP_pc_mean_1960_2022                0.00086      0.00101
region                               0.10788      0.11296
industry_share_mean_1960_2022        0.00185      0.00218
length_of_last_leader_tenure_2015 0.00201      0.00236
muslim_proportion                    0.00540      0.00633
top_universities_shanghai_2022       0.00058      0.00068
plurality                            0.00262      0.00308
Residual                                  NA           NA
```

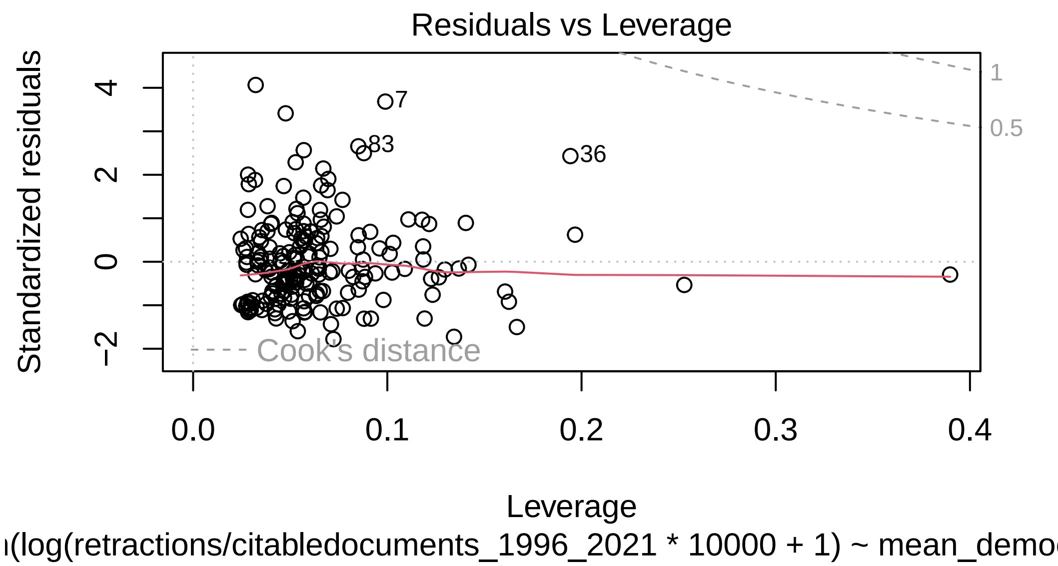
And plots:

```
plot(fit_nooutimp_linear_multi_log1$analyses[[1]])
```


Residuals vs Fitted
Residuals
Fitted values
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo
11
7
44
−2
0
2
4
0.5
1.0
1.5
2.0
2.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo
11
7
44
−2
0
1
2
3
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
44
11
7
0.5
1.0
1.5
2.0
2.5
Fitted values
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo(


Residuals vs Leverage
Standardized residuals
−2
0
2
4
7
83
36
1
0.5
Cook's distance
0.0
0.1
0.2
0.3
0.4
Leverage
ı(log(retractions/citabledocuments_1996_2021 * 10000 + 1) ~ mean_demo(