# Tracking AI in climate innovation: Supplementary information

December 8, 2021

## Supplementary Information

In this part we provide the supplementary results referred to from the main text and Methods.

## Contents

**1 Modeling details** **2**

**2 Model diagnostics** **3**

**3 Exploring the data and the covariates** **32**

# 1 Modeling details

- Time period for statistical analysis: 2010-2019
- Target variable: 3-year forward citation in CPC subclasses with the following covariates. This results in analysing forward citations for patetns from 2010-2017.
- Pre-processing:
  - For regression modeling, filter out top 1% of most highly cited patents (outliers skewed the model fit quite a bit, removing them clearly improved the fit in diagnostics plots, as seen below).
  - Remove subgroups that lack patents in each of the year 2010-2017 (around 1% subgroups with little or no patenting activity).
- List of covariates/controls:
  - ai (factor, binary: AI classification according to the WIPO)
  - green_ycodes (factor, binary: if Y02 label on patent)
  - grantyear (factor, 8 levels: [2010,...,2017])
  - cpc_subarea (factor: for CPC subclass, groups; for groups, subgroups)
  - tct_type (factor, 4 levels: [0-5,5-10,10-15,15-]) (based on the data exploration below)
  - n_claims (log plus one transformed): number of claims in patent
  - n_inventors_individual (log plus one transformed): number of inventors
  - n_backward_patent_citations (log plus one transformed): number of patent citations
  - n_backward_research_citations (log plus one transformed): number of research citations
  - n_backward_other_citations (log plus one transformed): number of other citations
  - organizational (factor, binary: if firm on patent)
- GAMLSS method: General Additive Model for Location, Shape, and Scale[i]. This approach was chosen because the data turns out to be both zero-inflated and heavily skewed – variance that differs across subclasses and groups. Including model parameters for subclass- and group-level skewness led to models with good fit, compared to Poisson and Negative Binomial models.
- The specific model for the citation count data is: Generalized inverse Gaussian distribution (Sichel distribution), as this can model highly skewed and zero-inflated count data.[ii,iii]
- Model specification on the level of technologies (subclasses/groups): We use repeated modeling for the different technology areas to get estimates of how AI predicts a difference in citation between the units in the data. Below, we also present residuals for the subclass- and group-level regressions.
- Model diagnostics: for the models, the normalized randomize quantile residuals[iv] (for details, see section 6.2 in reference given by footnote (iii)). Model diagnostis for these can be done on a Q-Q plot for approximate normality.

[i] Rigby, R.A. and Stasinopoulos, D.M. (2005), Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54: 507-554.

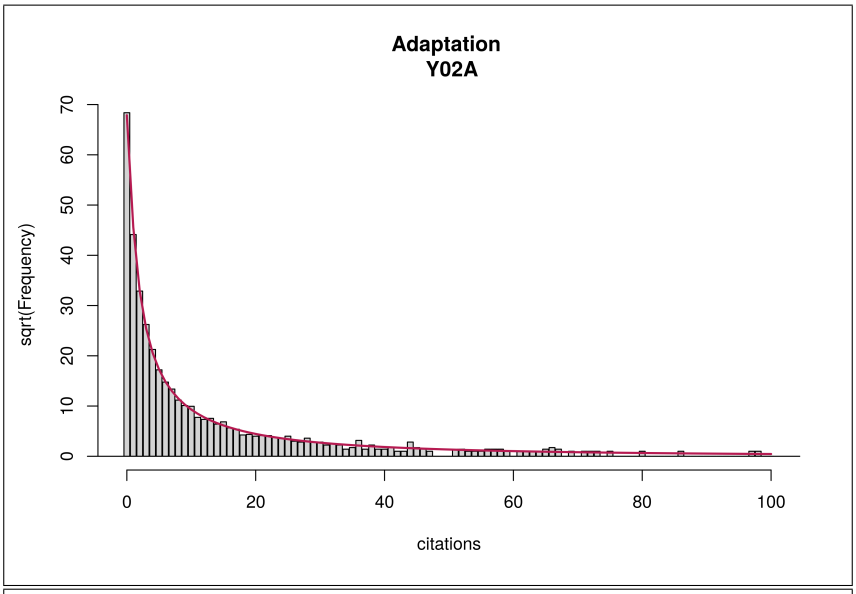
[ii] Gillian Z. Stein, Walter Zucchini and June M. Juritz. Parameter Estimation for the Sichel Distribution and its Multivariate Extension. (1987) Journal of the American Statistical Association. Vol. 82, No. 399, pp. 938-944

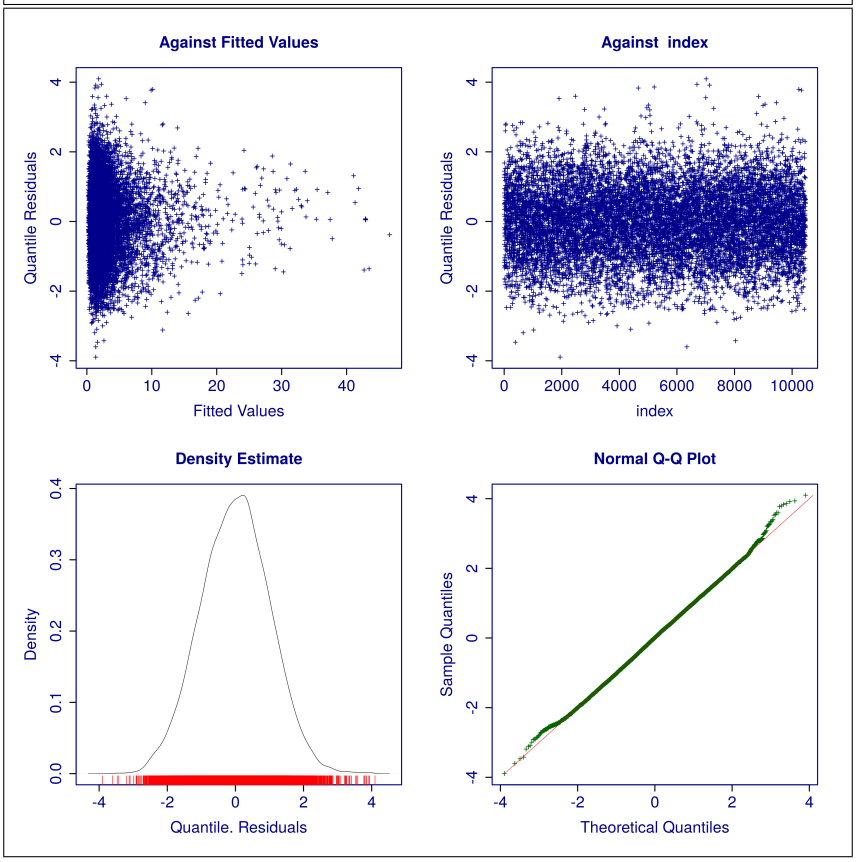
[iii] R. A. Rigby, D. M. Stasinopoulos, and C. Akantziliotou. (2008) A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Computational Statistics and Data Analysis, 53(2):381–393.

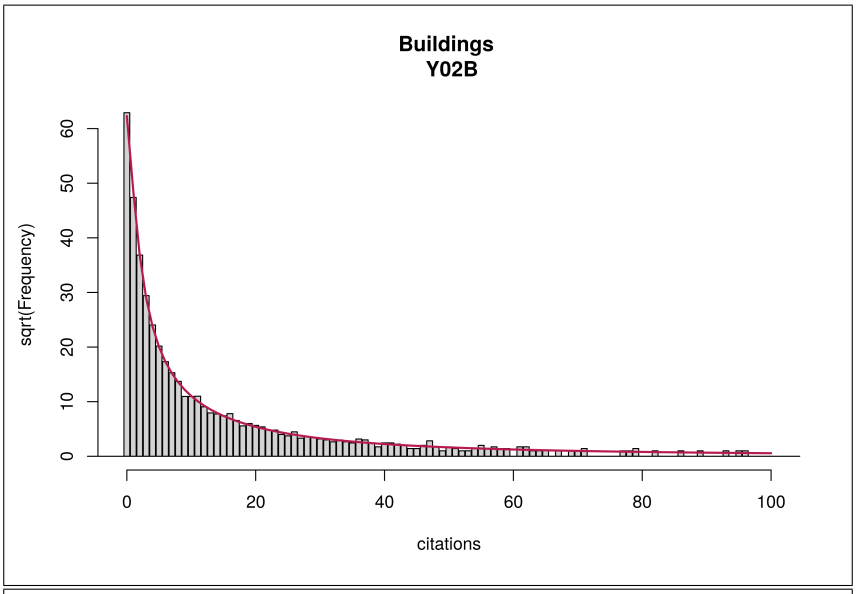
[iv] Randomized Quantile Residuals. Peter K. Dunn & Gordon K. Smyth. (1996) Journal of Computational and Graphical Statistics Vol. 5, No. 3. Pages 236-244.

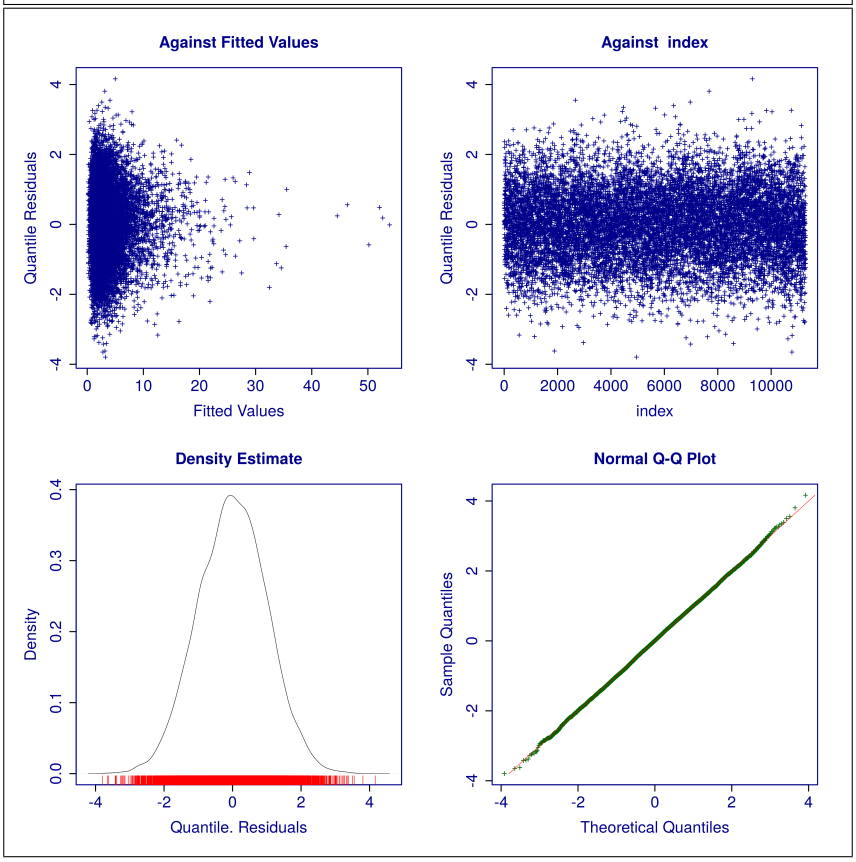
## 2 Model diagnostics

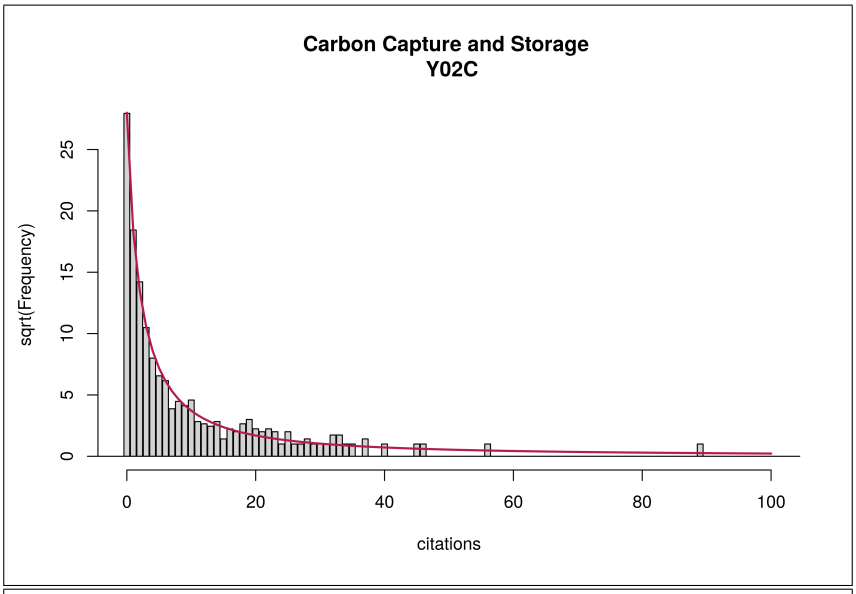
This section displays model diagnostics, both with rootograms – indicating whether or not the model (red curves) produces similar resultas as the data (the grey bins) – as well as residual plots. Quantile residuals should be approximately aligned with the diagonal of the Q-Q plots, which we find is the case for the final model.

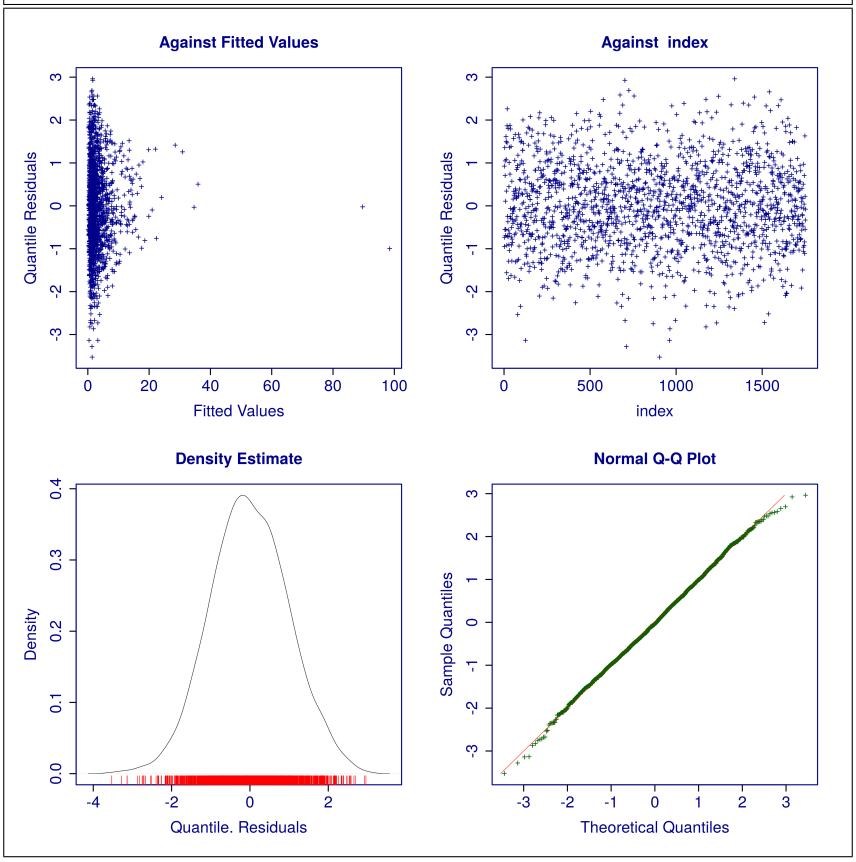

Adaptation
Y02A
sqrt(Frequency)
0
10
20
30
40
50
60
70
0
20
40
60
80
100
citations

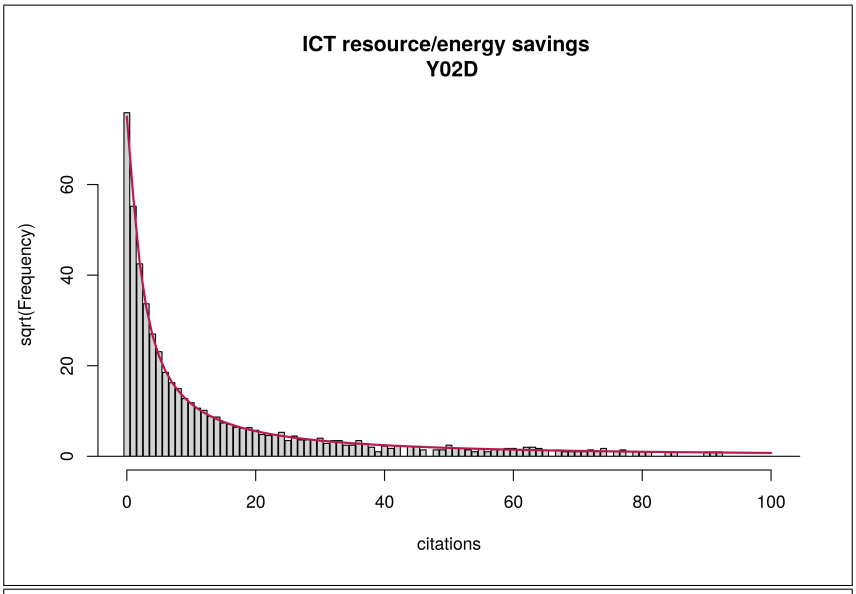

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

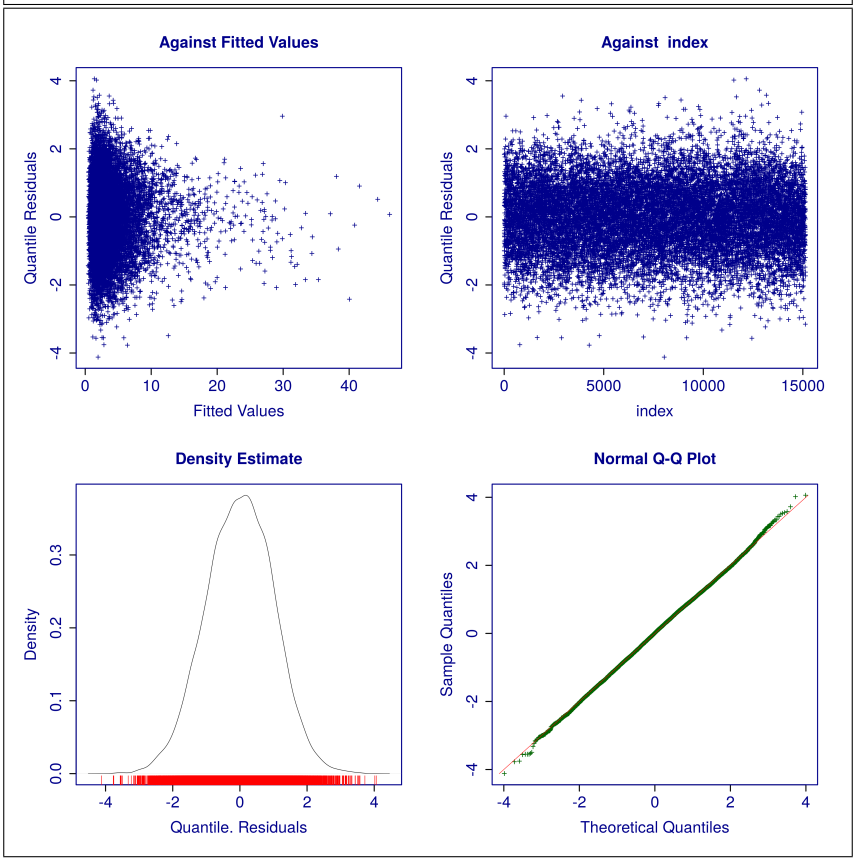

Buildings
Y02B
sqrt(Frequency)
0
10
20
30
40
50
60
0
20
40
60
80
100
citations

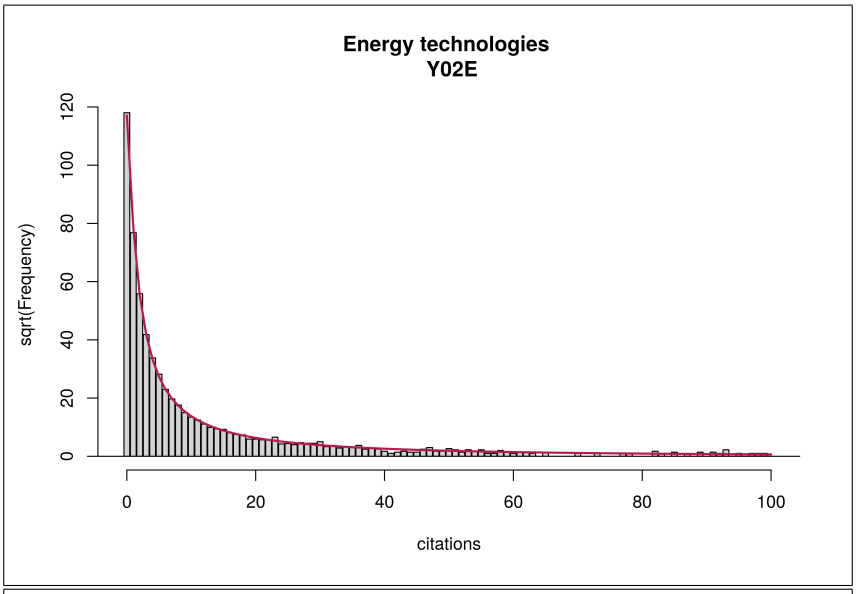

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

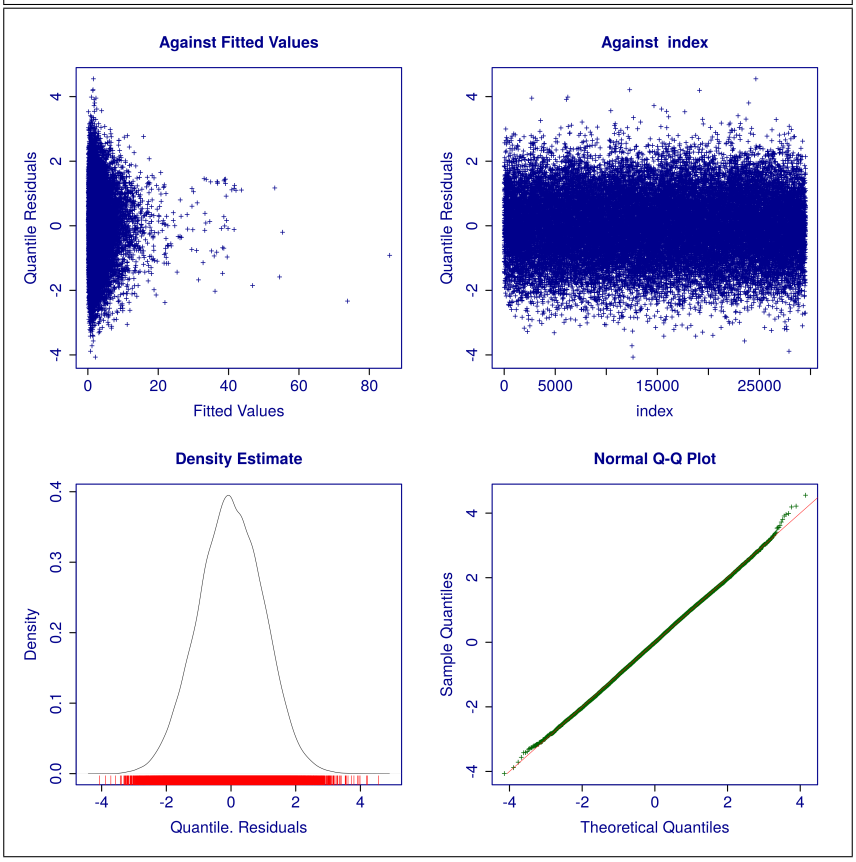

Carbon Capture and Storage
Y02C
sqrt(Frequency)
0
5
10
15
20
25
0
20
40
60
80
100
citations

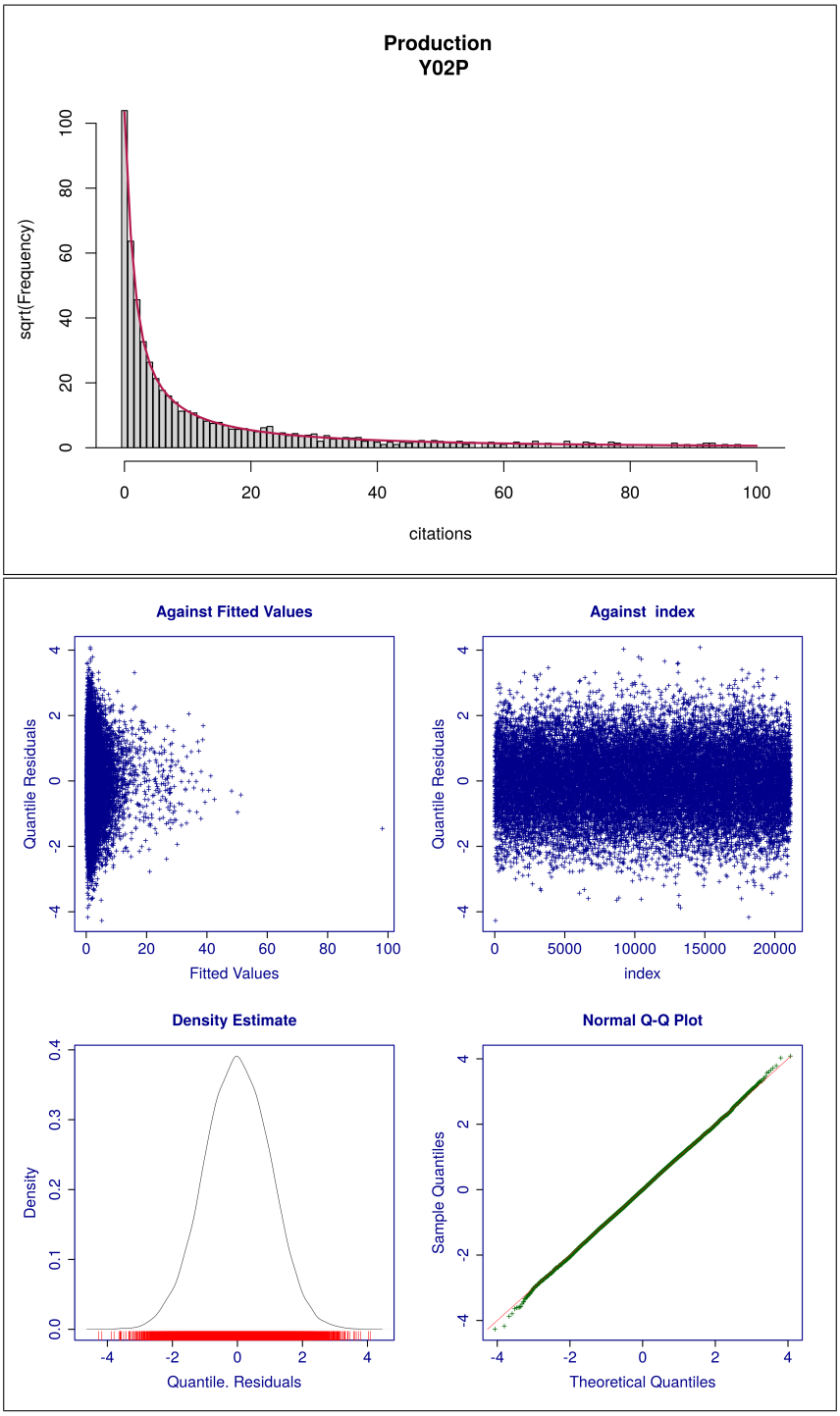

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

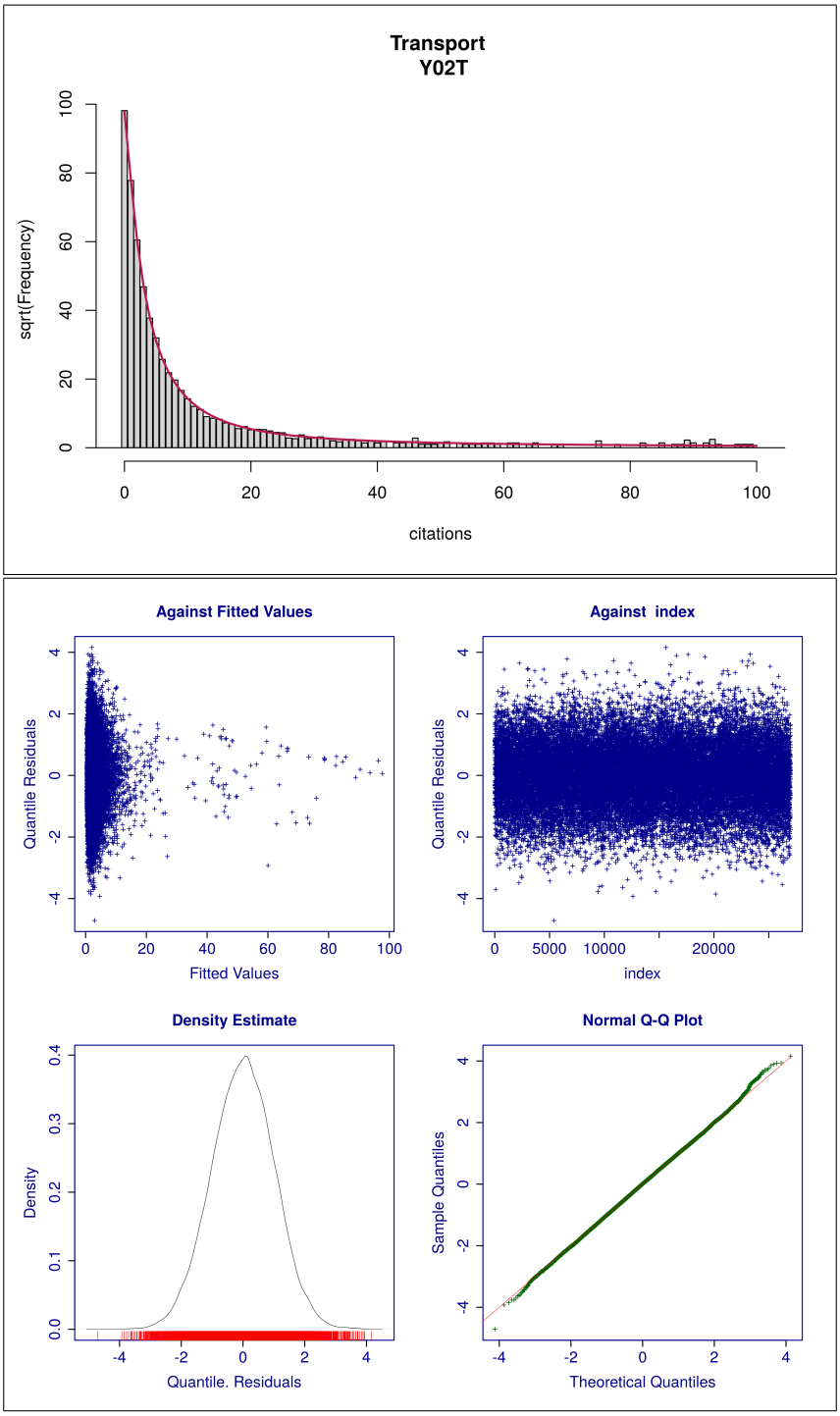

ICT resource/energy savings
Y02D
sqrt(Frequency)
0
20
40
60
0
20
40
60
80
100
citations

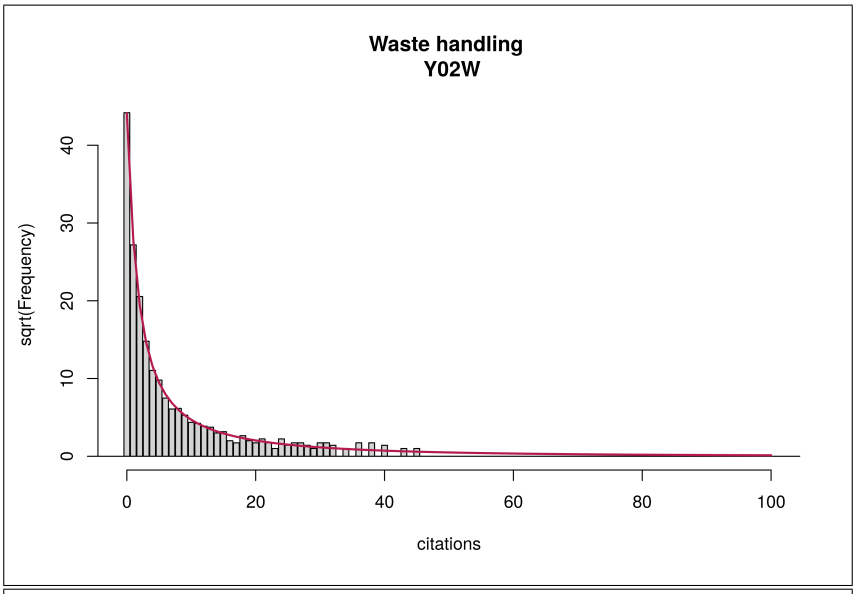

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

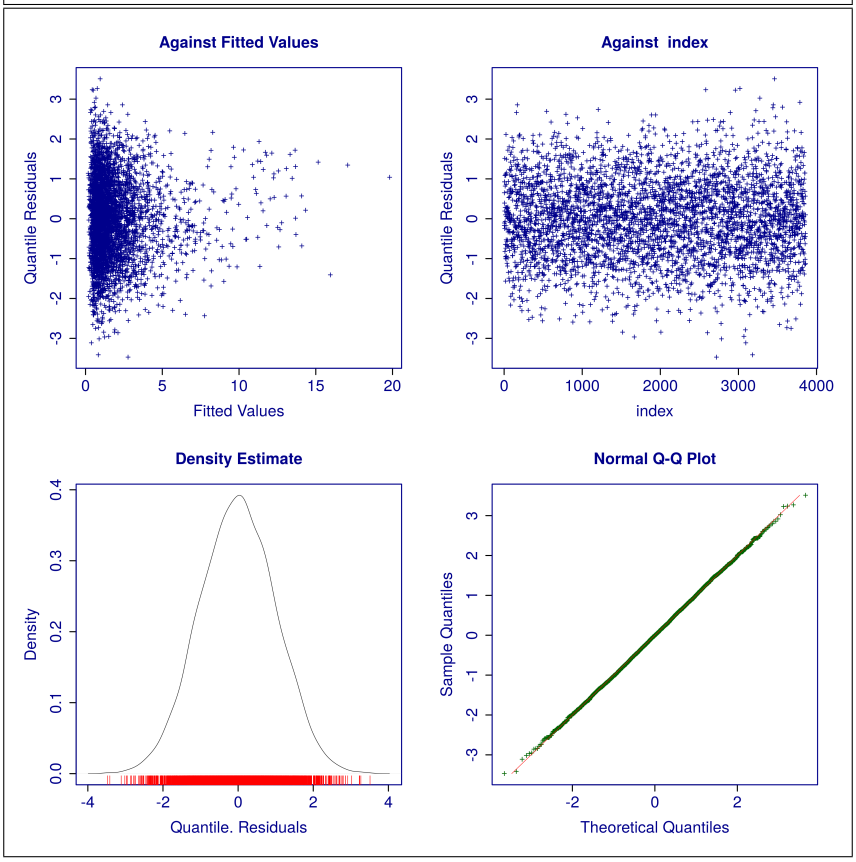

Energy technologies
Y02E
sqrt(Frequency)
0
20
40
60
80
100
120
0
20
40
60
80
100
citations

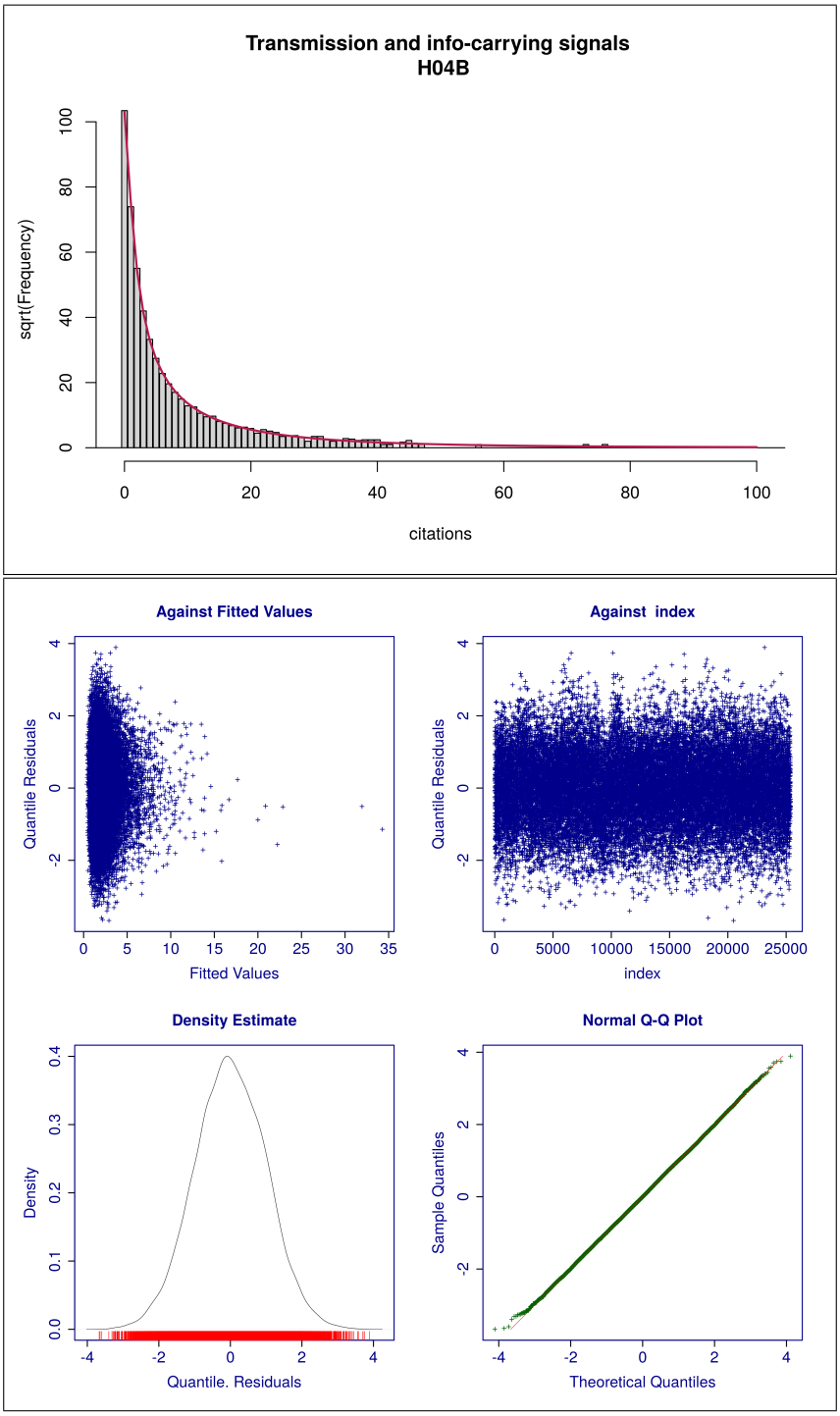

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

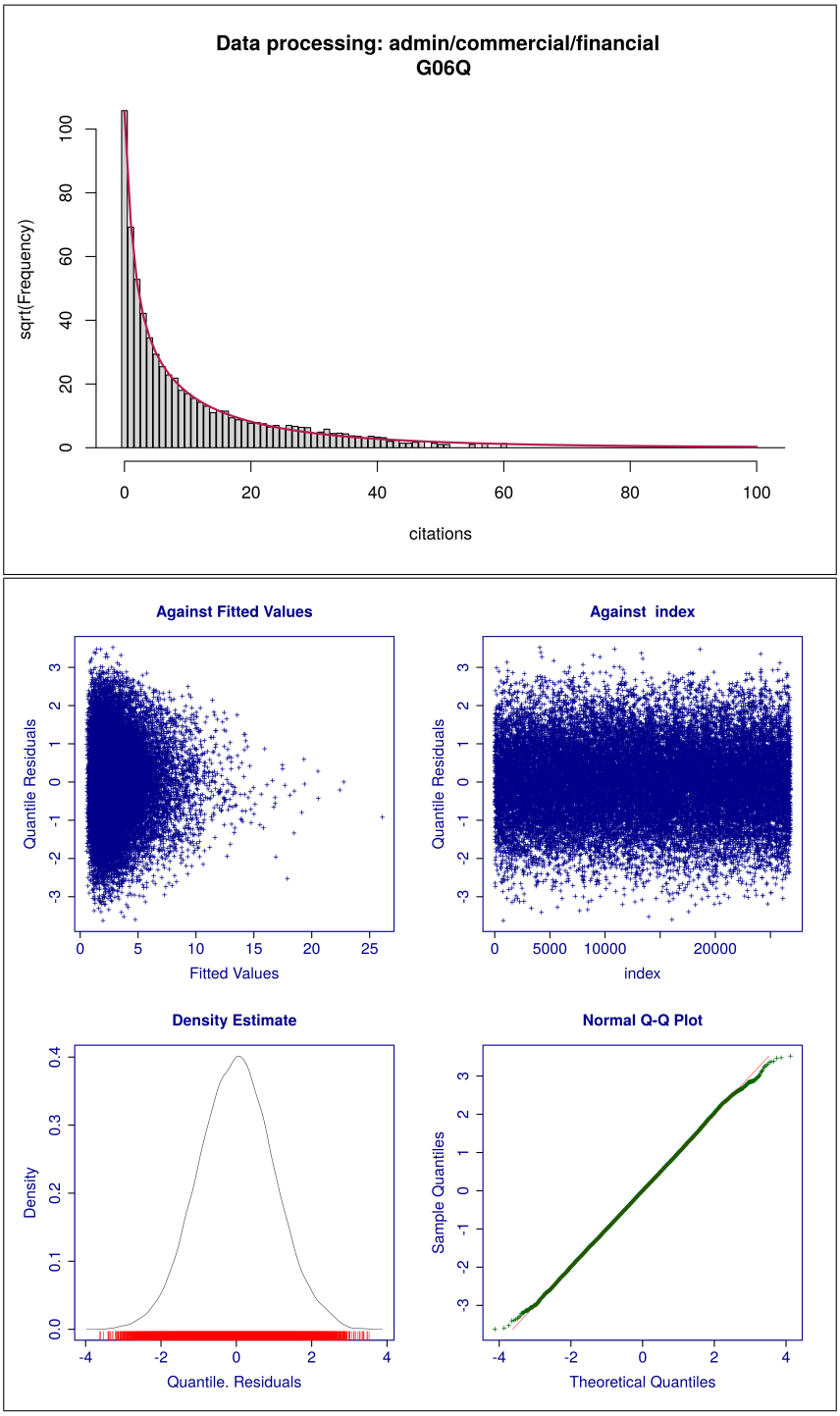

Production
Y02P
sqrt(Frequency)
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

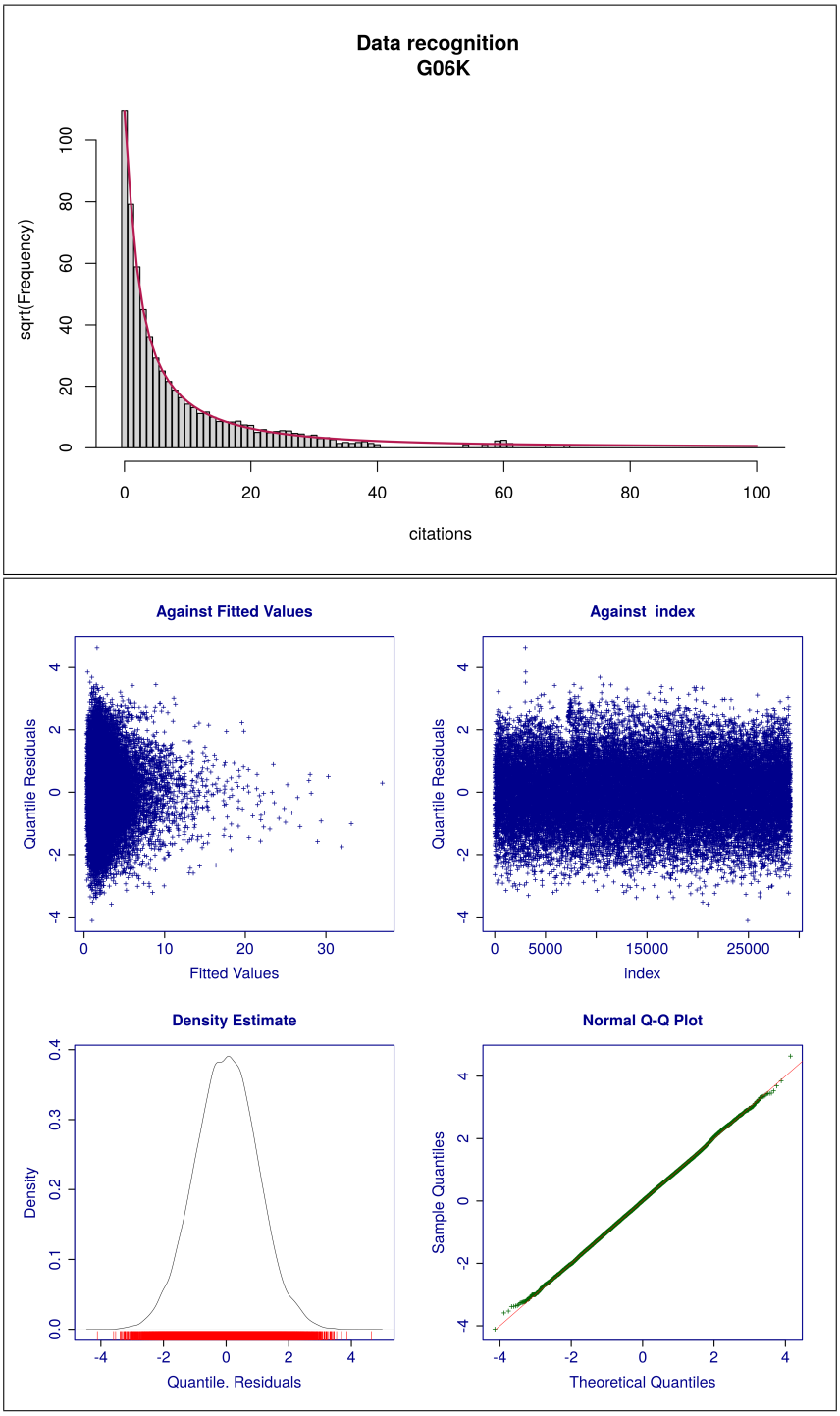

Transport
Y02T
sqrt(Frequency)
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

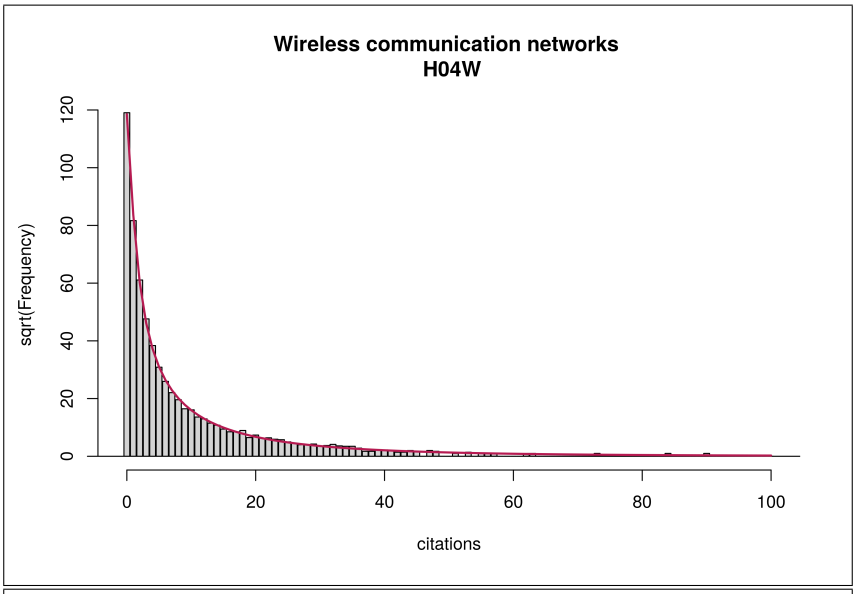

Waste handling
Y02W
sqrt(Frequency)
0
10
20
30
40
0
20
40
60
80
100
citations

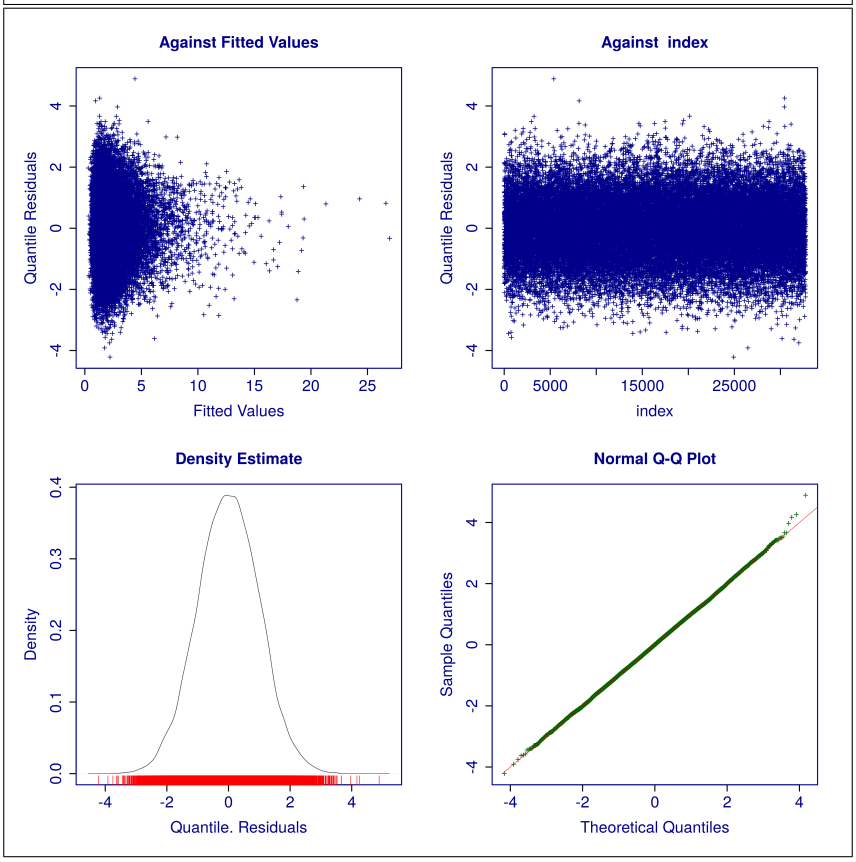

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

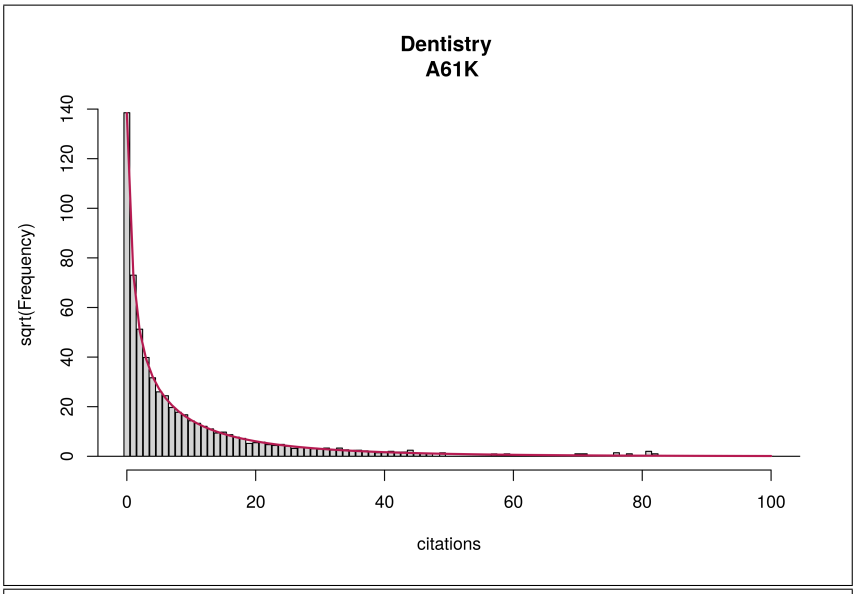

Transmission and info-carrying signals
H04B
sqrt(Frequency)
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

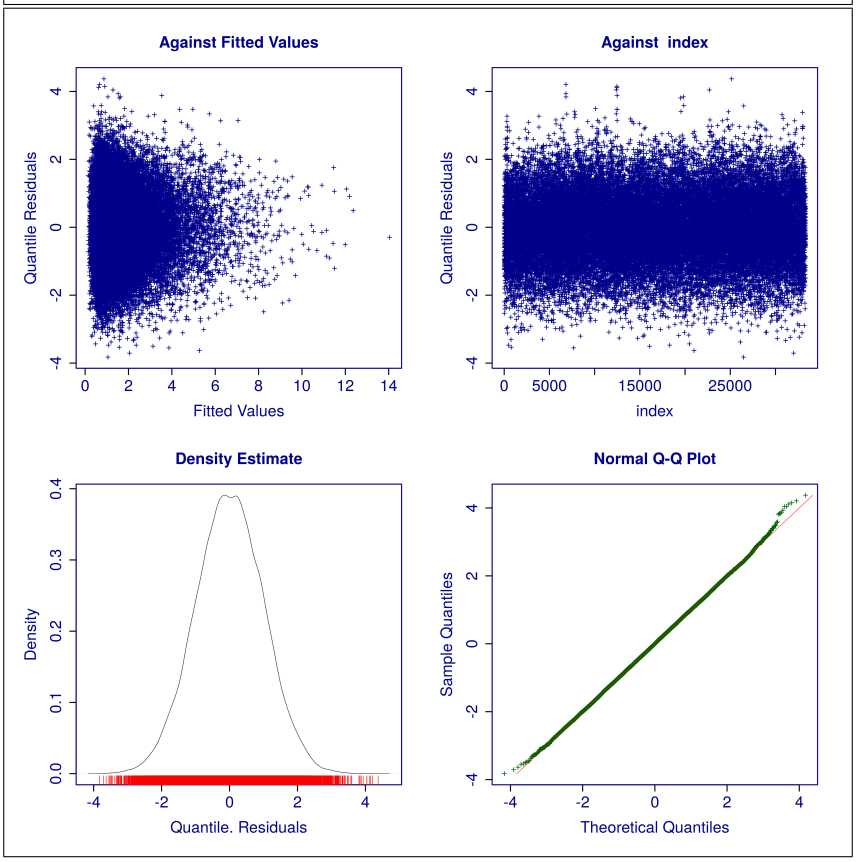

Data processing: admin/commercial/financial
G06Q
sqrt(Frequency)
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

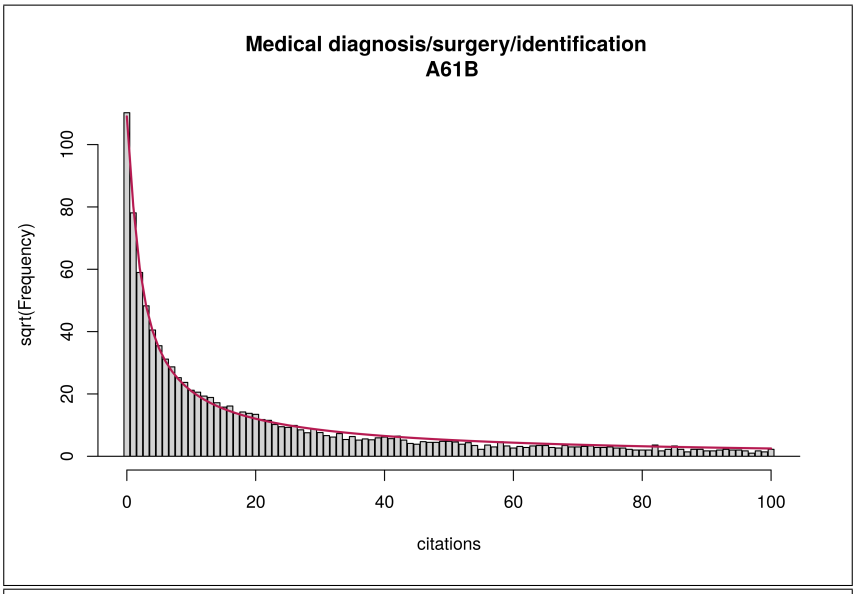

Data recognition
G06K
sqrt(Frequency)
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

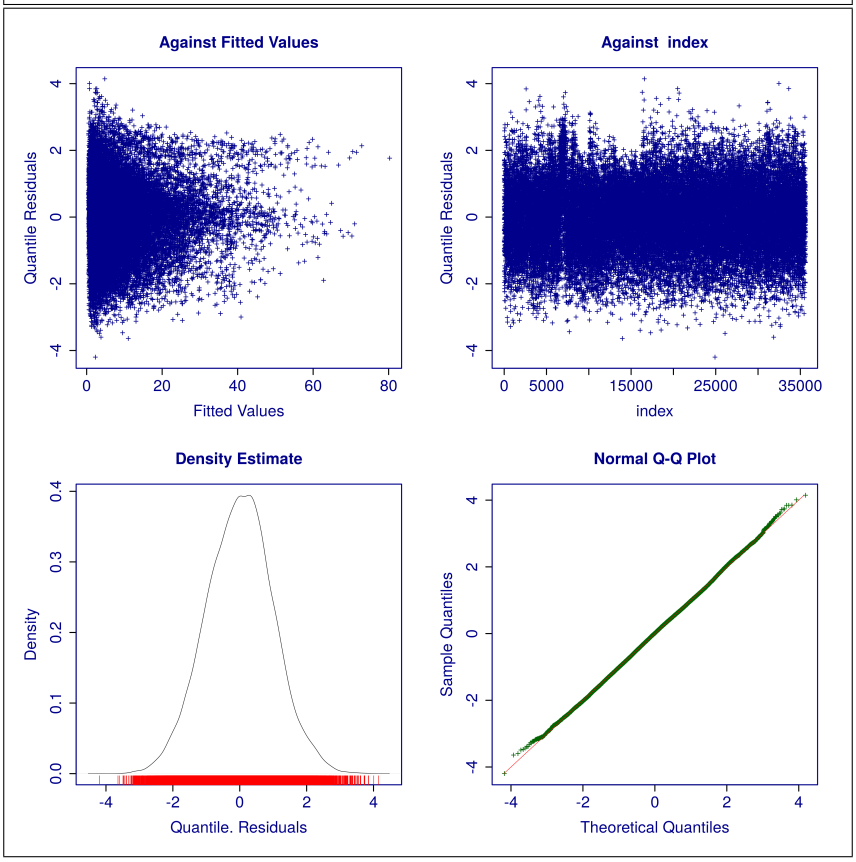

Wireless communication networks
H04W
sqrt(Frequency)
0
20
40
60
80
100
120
0
20
40
60
80
100
citations

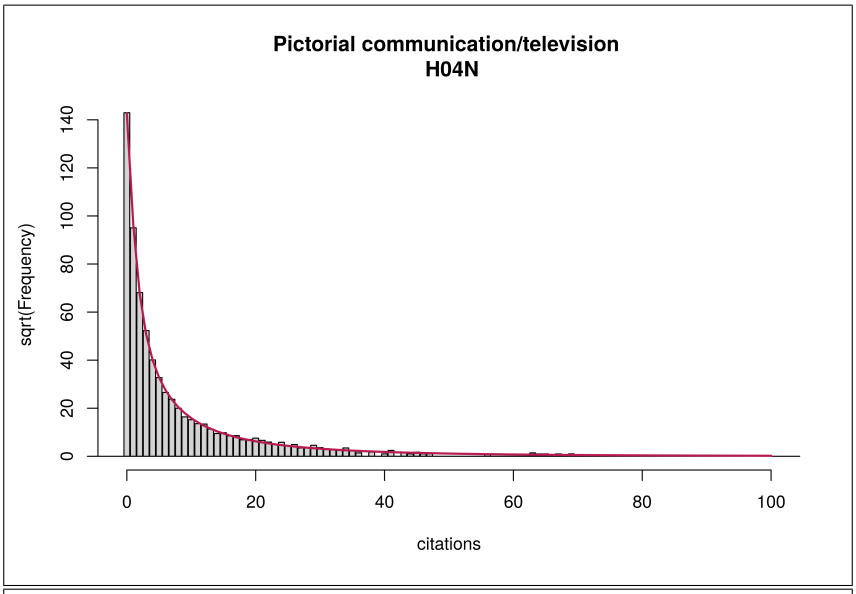

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

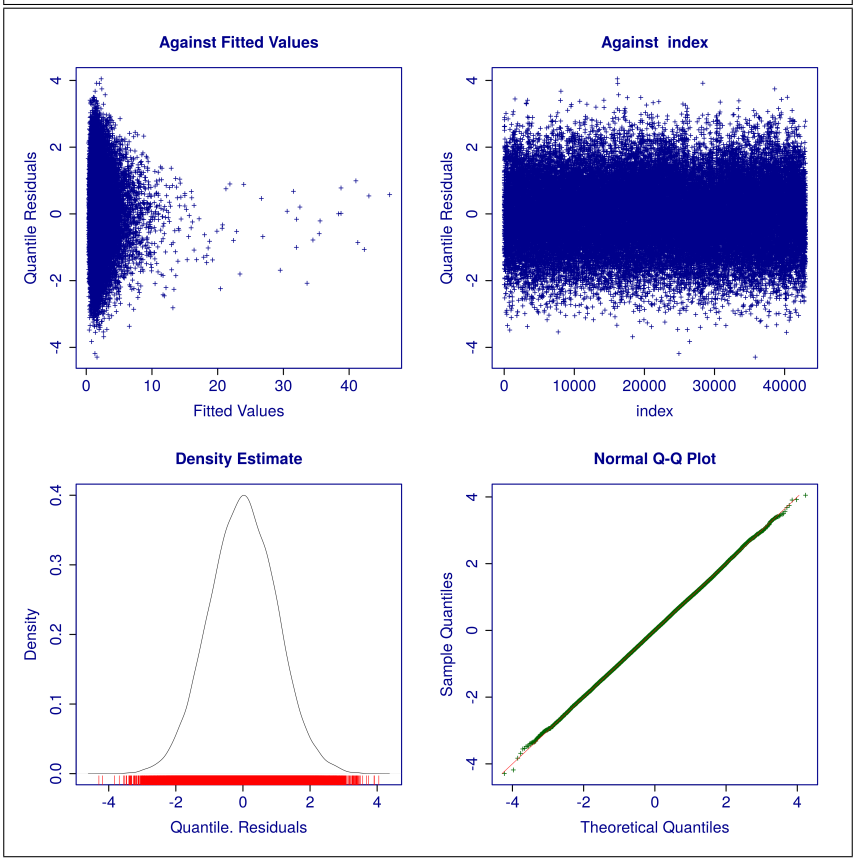

Dentistry
A61K
sqrt(Frequency)
0
20
40
60
80
100
120
140
0
20
40
60
80
100
citations

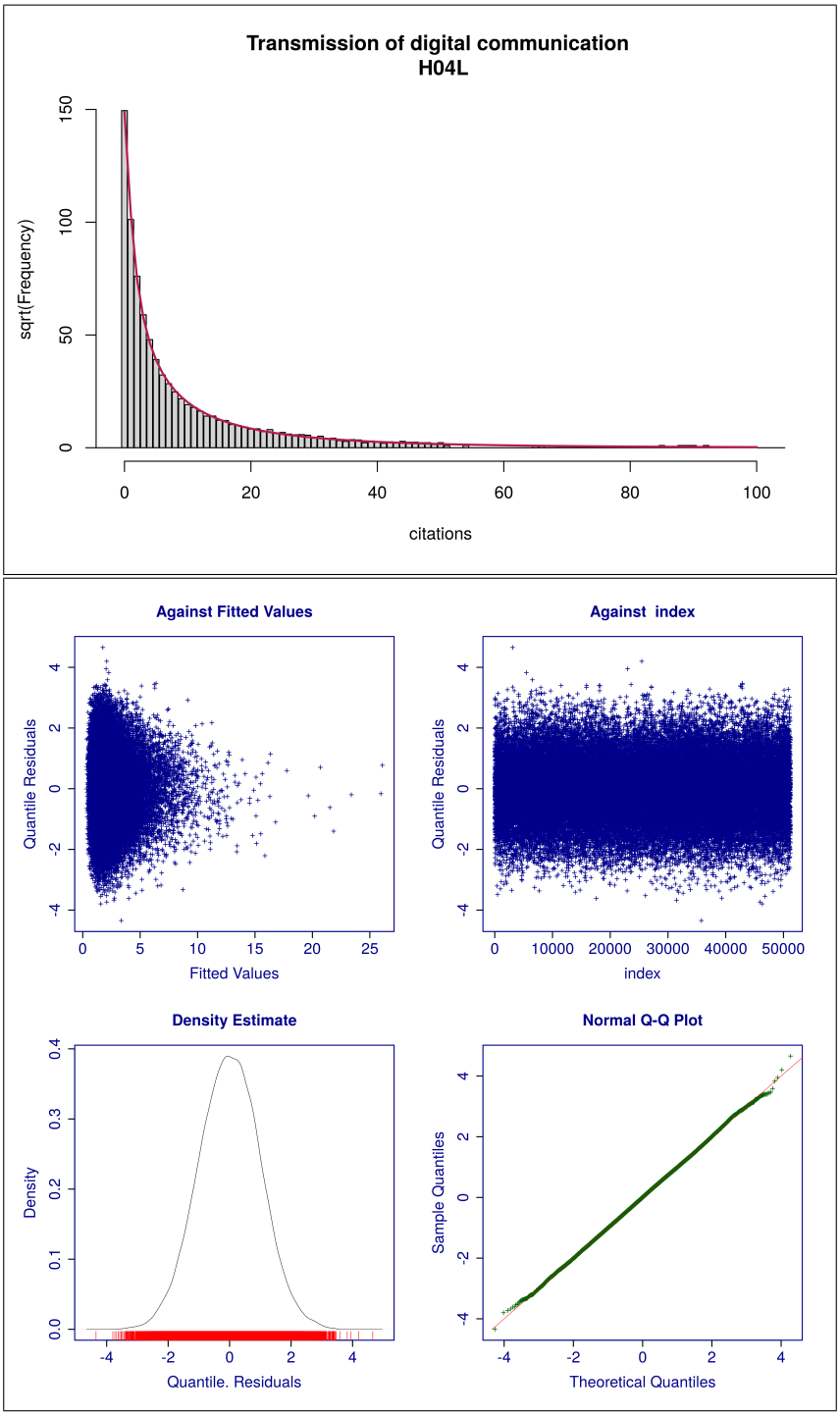

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

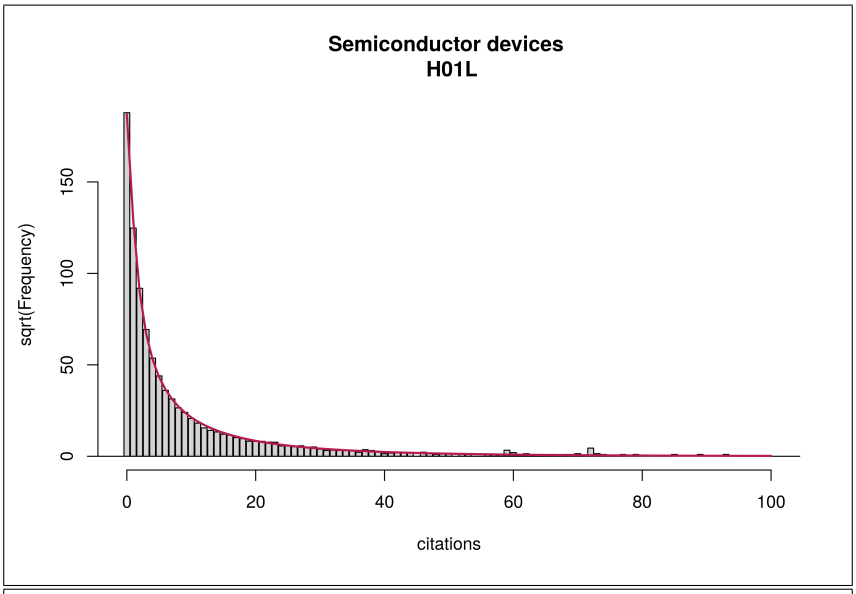

Medical diagnosis/surgery/identification
A61B
sqrt(Frequency)
0
20
40
60
80
100
citations

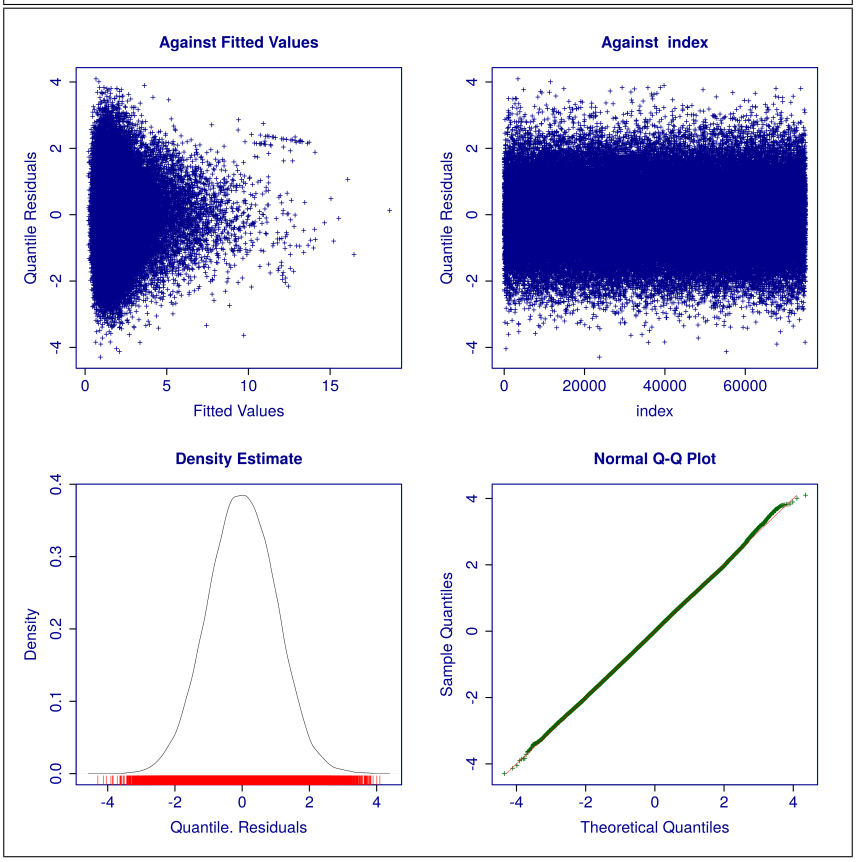

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

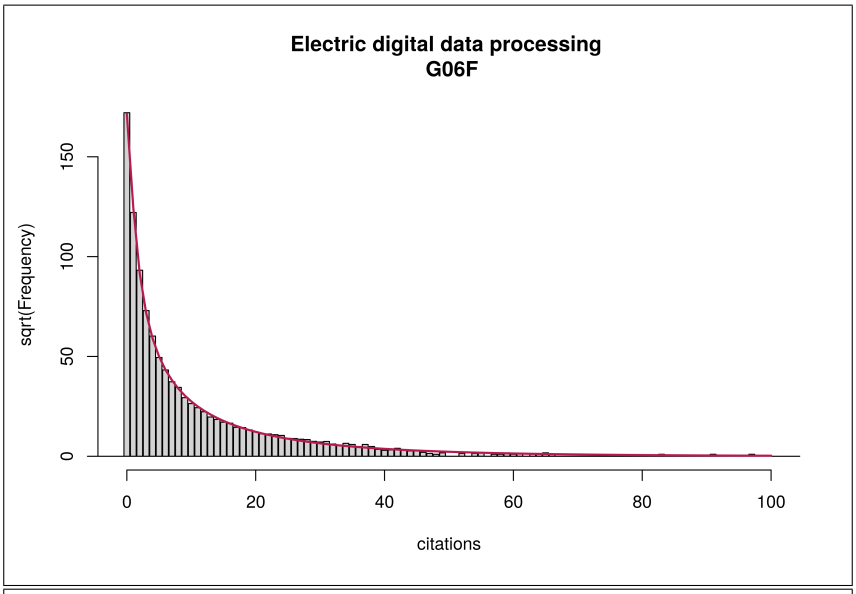

Pictorial communication/television
H04N
sqrt(Frequency)
0
20
40
60
80
100
120
140
0
20
40
60
80
100
citations

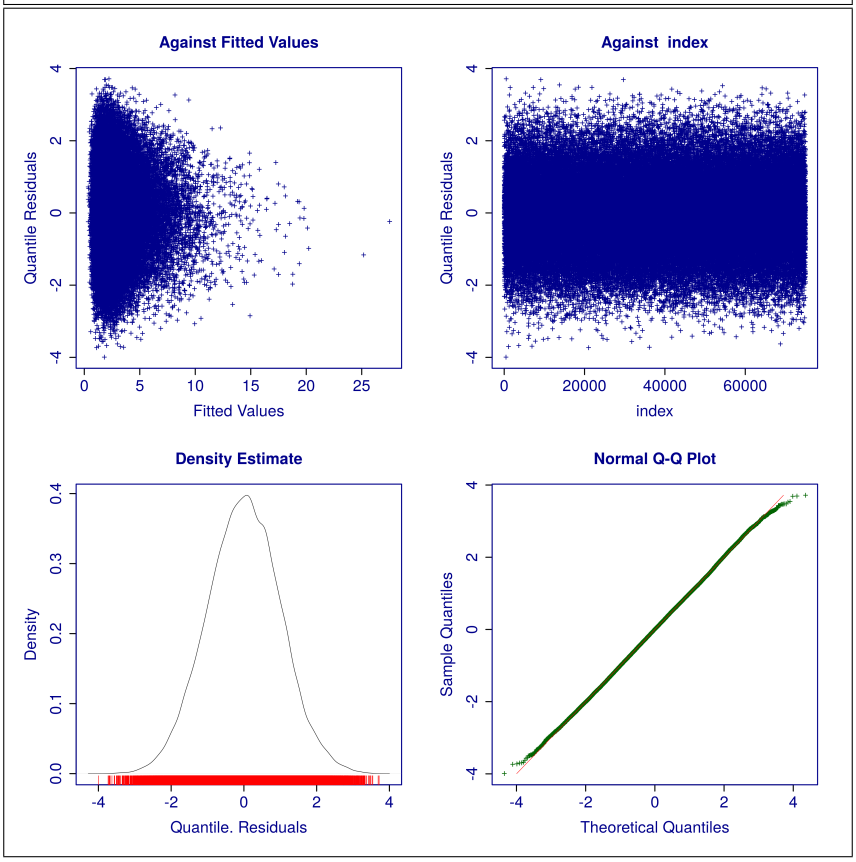

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

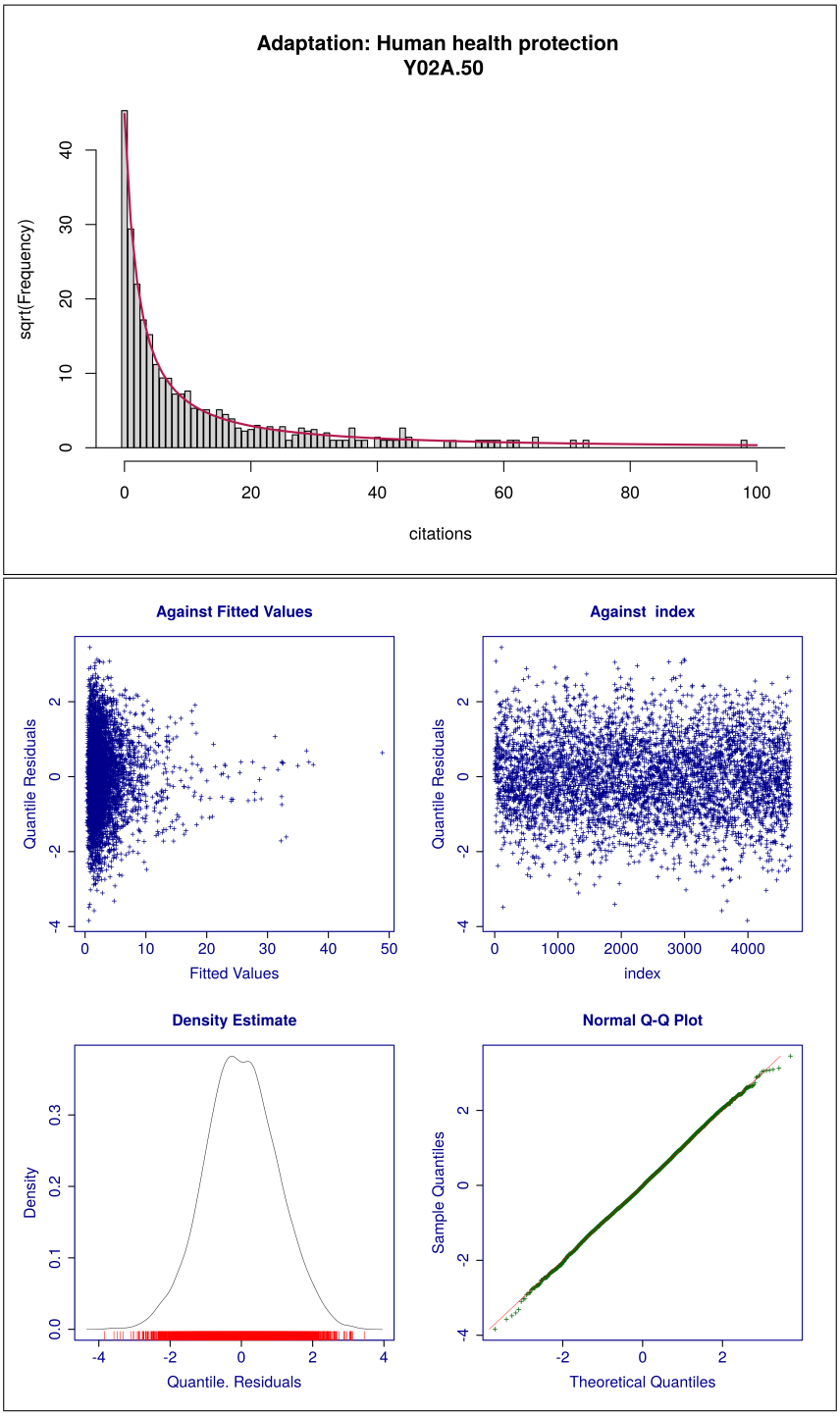

Transmission of digital communication
H04L
sqrt(Frequency)
0
50
100
150
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
-4
-2
0
2
4
0
5
10
15
20
25
Fitted Values
Against index
Quantile Residuals
-4
-2
0
2
4
0
10000
20000
30000
40000
50000
index
Density Estimate
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
-4
-2
0
2
4
Theoretical Quantiles

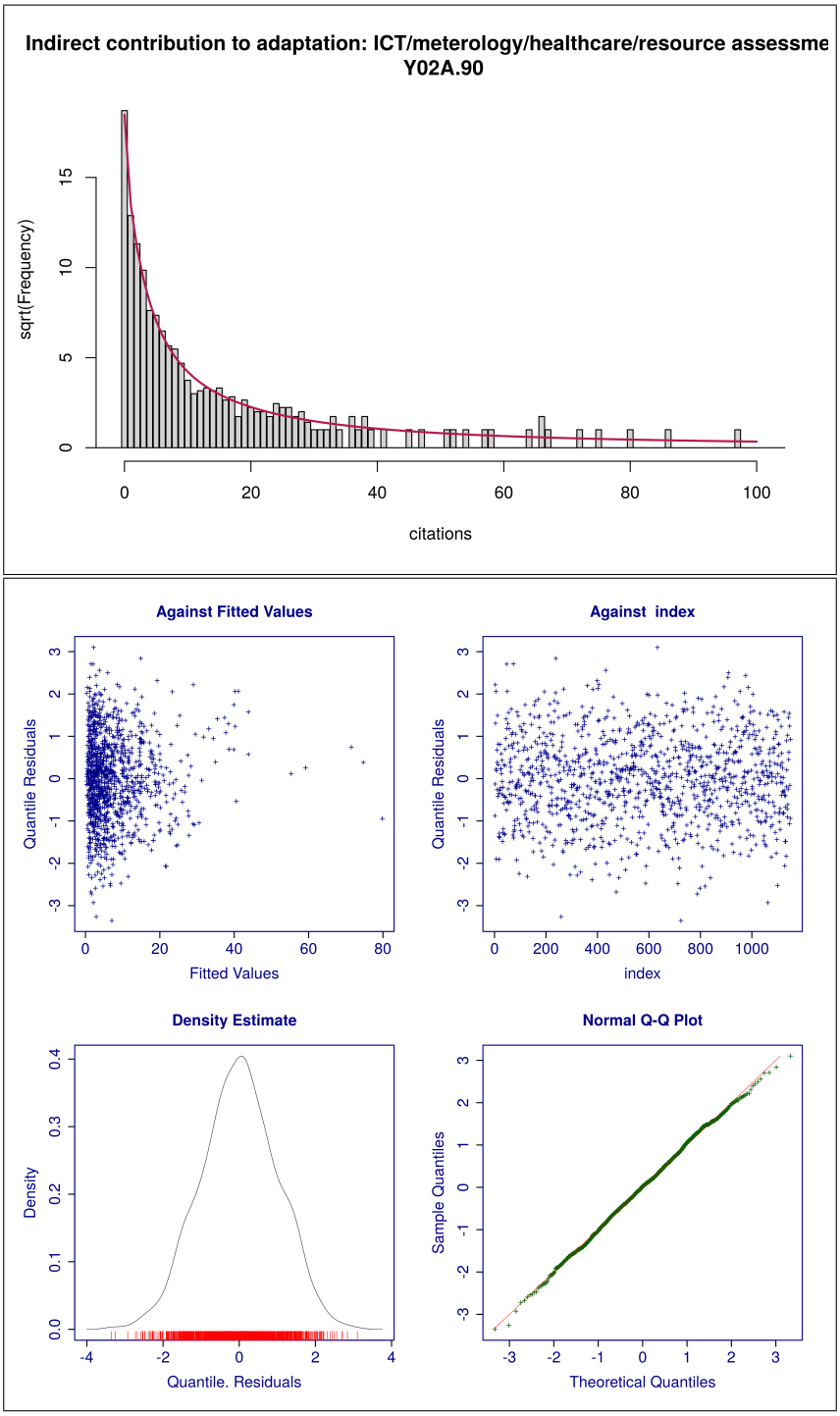

Semiconductor devices
H01L
sqrt(Frequency)
0
50
100
150
0
20
40
60
80
100
citations

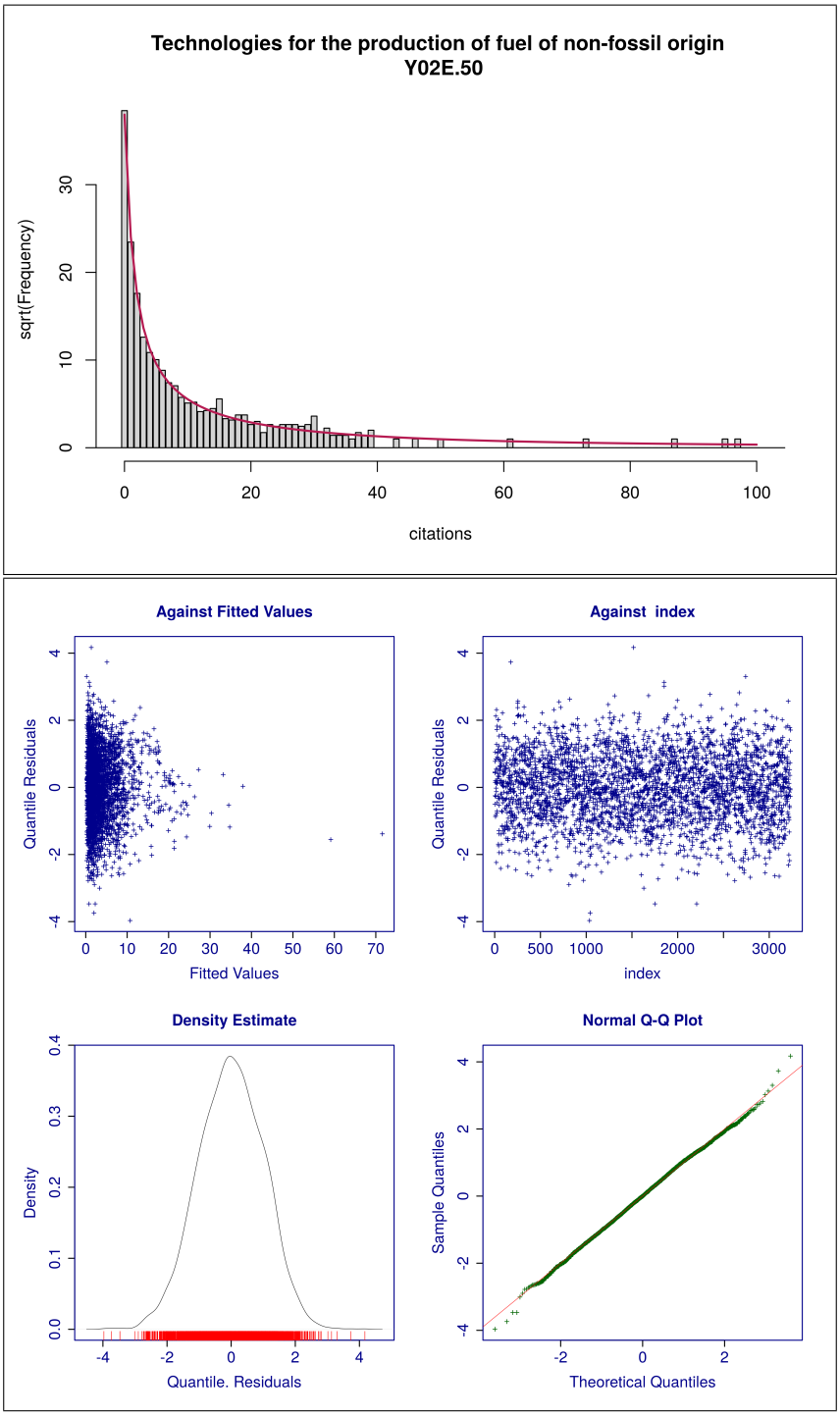

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

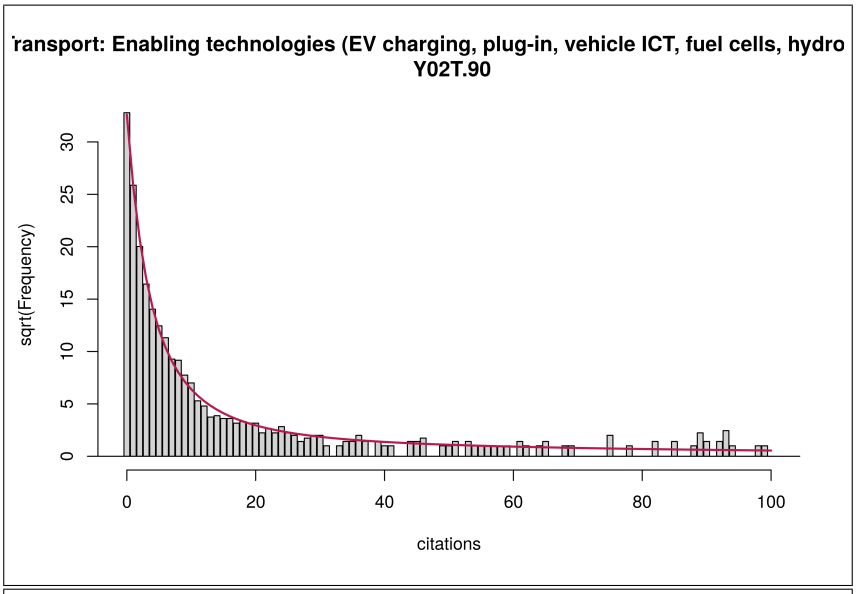

Electric digital data processing
G06F
sqrt(Frequency)
0
50
100
150
0
20
40
60
80
100
citations

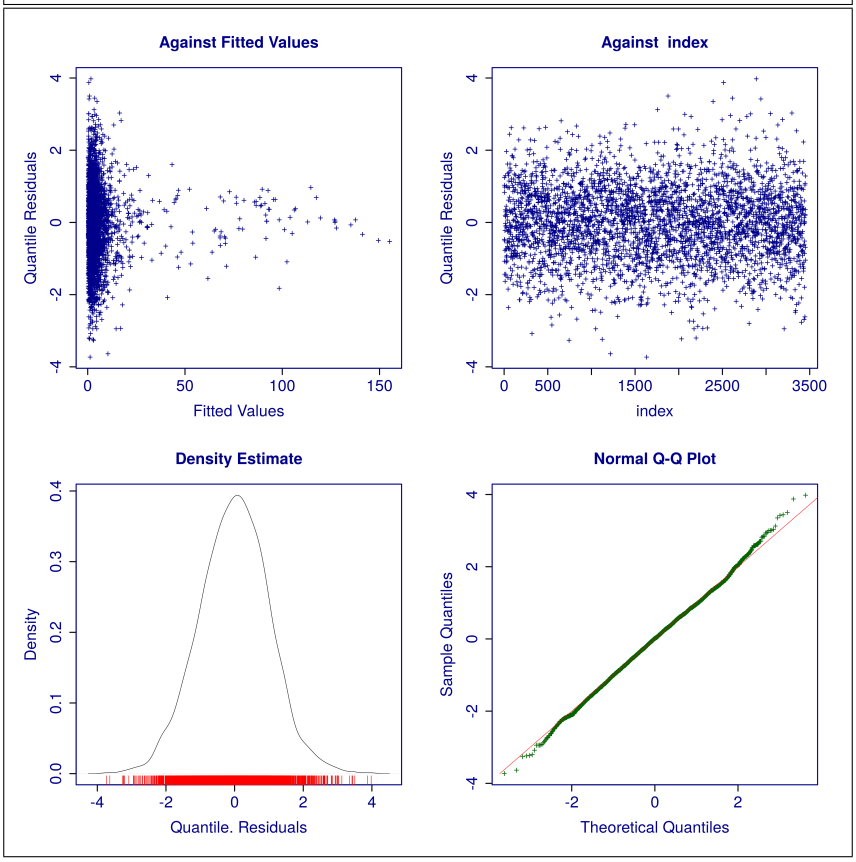

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

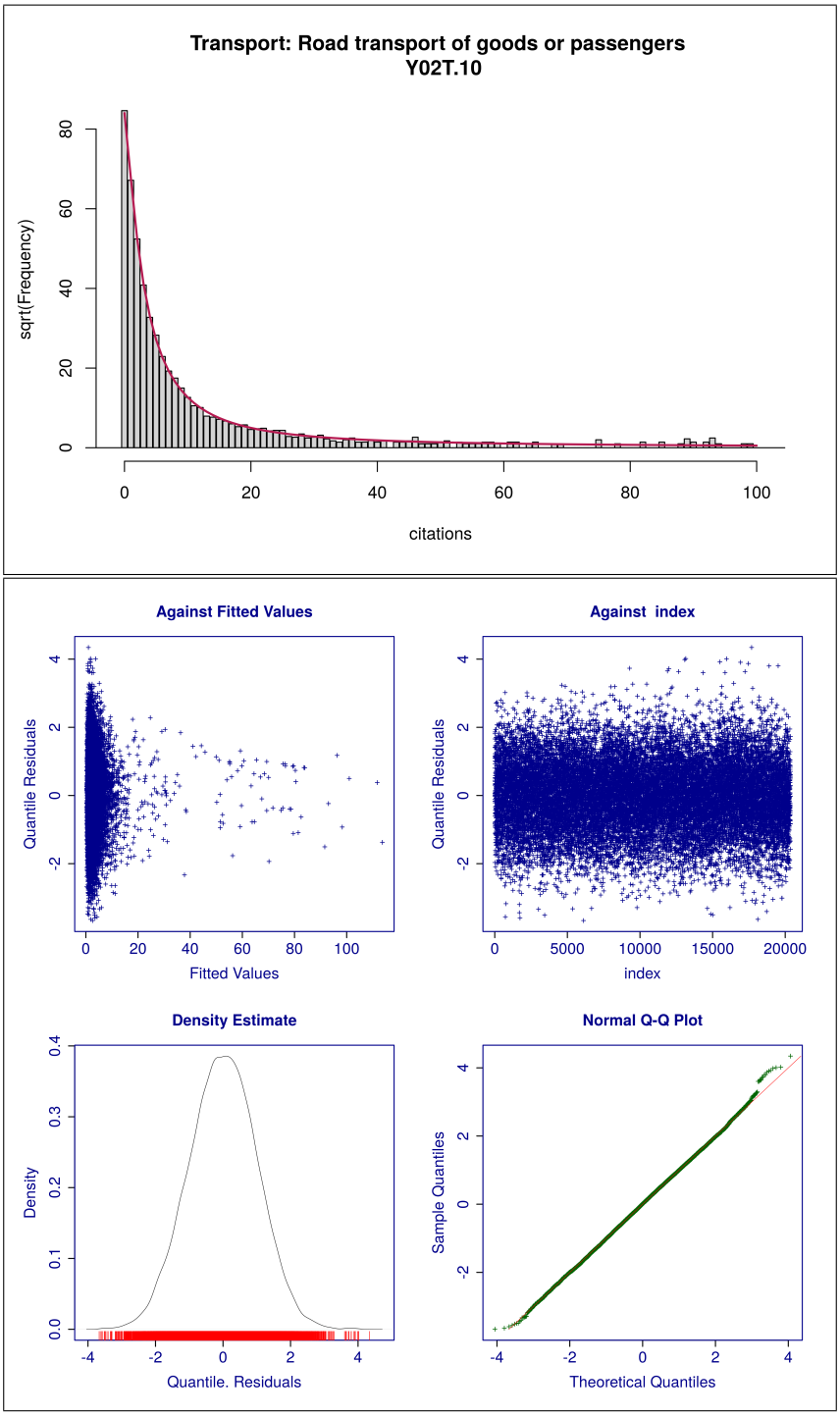

Adaptation: Human health protection
Y02A.50
sqrt(Frequency)
0
10
20
30
40
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

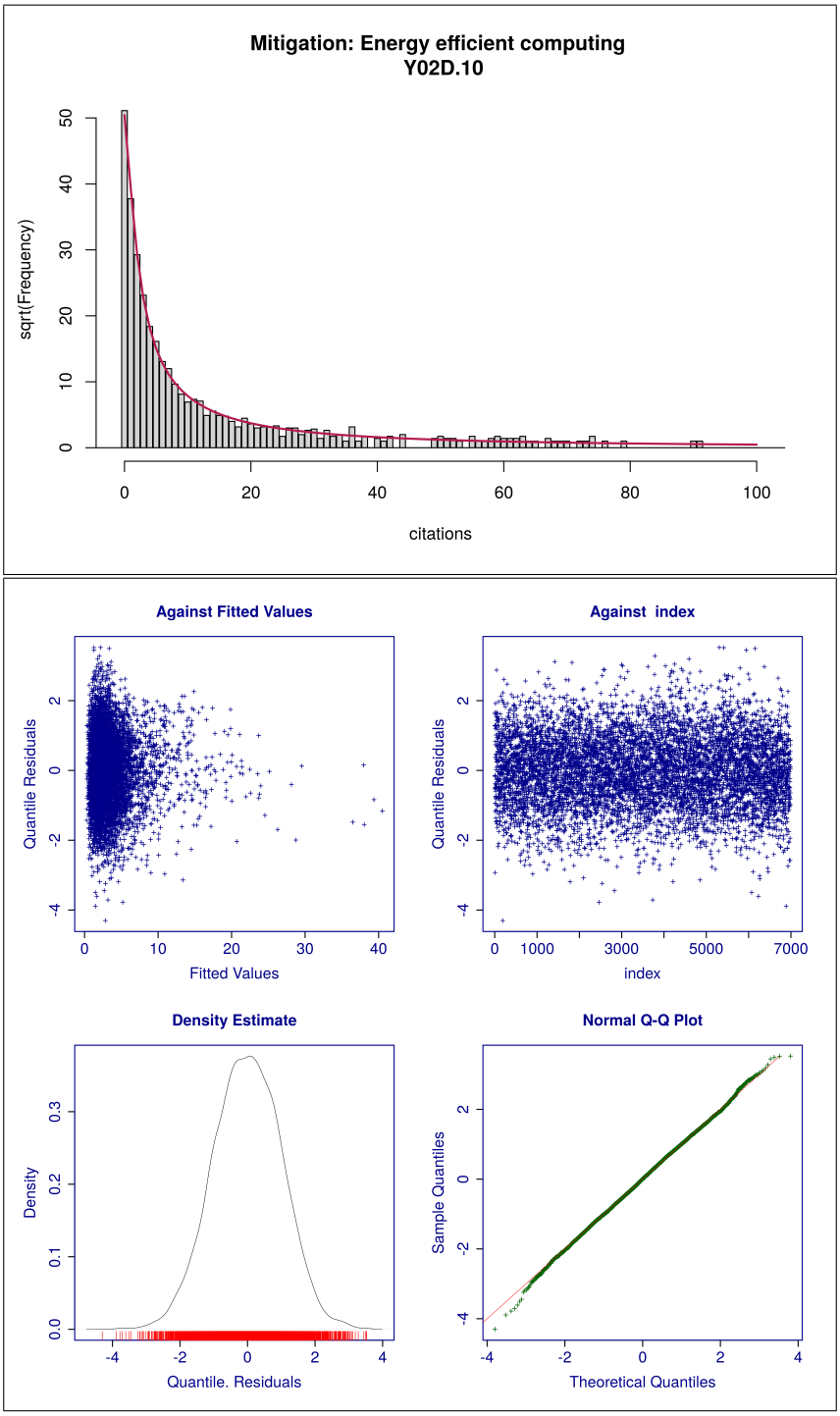

Indirect contribution to adaptation: ICT/meterology/healthcare/resource assessme
Y02A.90
sqrt(Frequency)
0
5
10
15
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

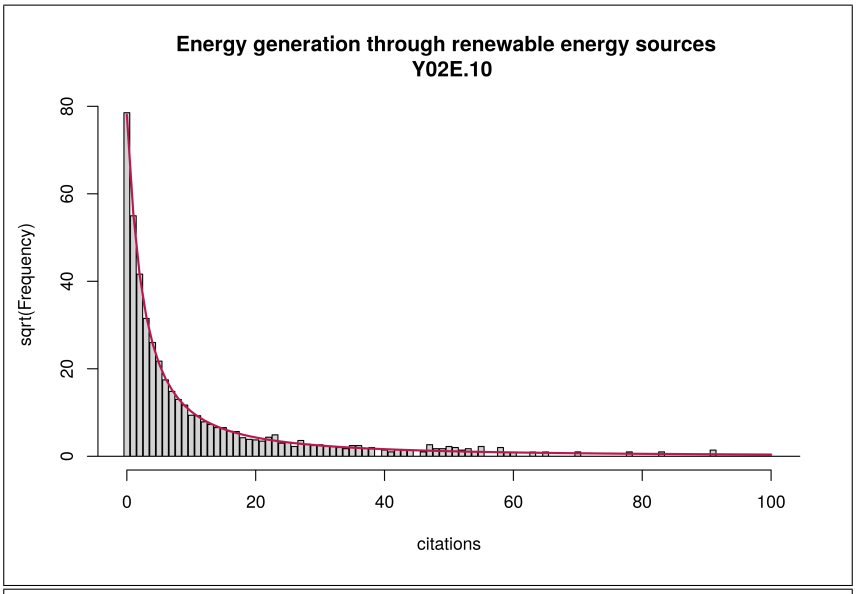

Technologies for the production of fuel of non-fossil origin
Y02E.50
sqrt(Frequency)
0
10
20
30
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

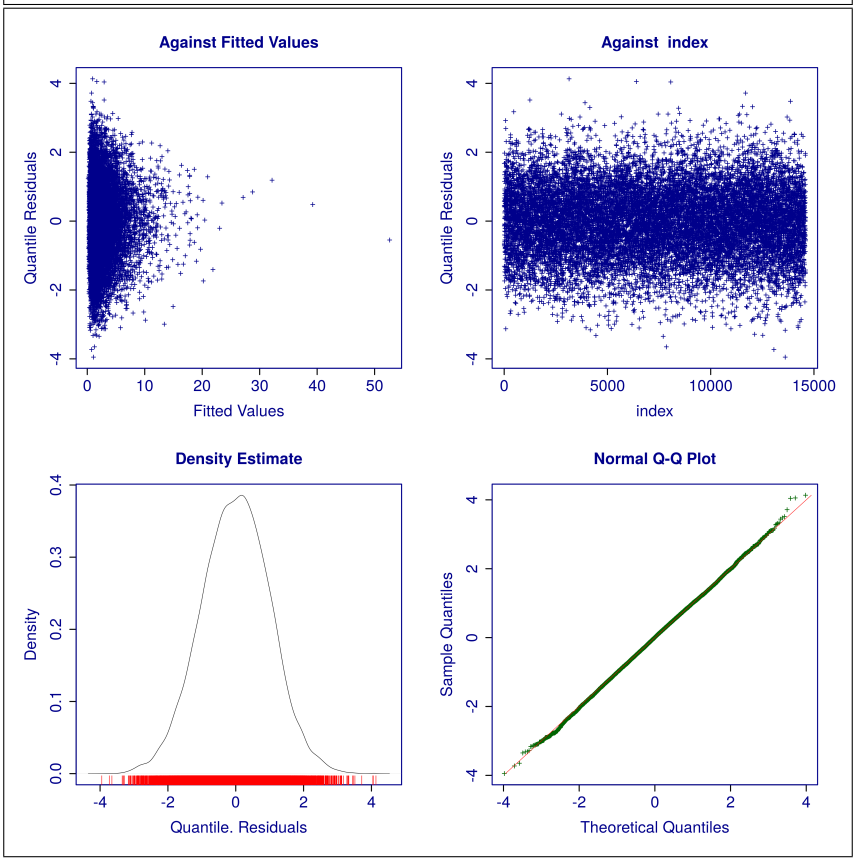

ransport: Enabling technologies (EV charging, plug-in, vehicle ICT, fuel cells, hydro
Y02T.90
sqrt(Frequency)
0
5
10
15
20
25
30
0
20
40
60
80
100
citations

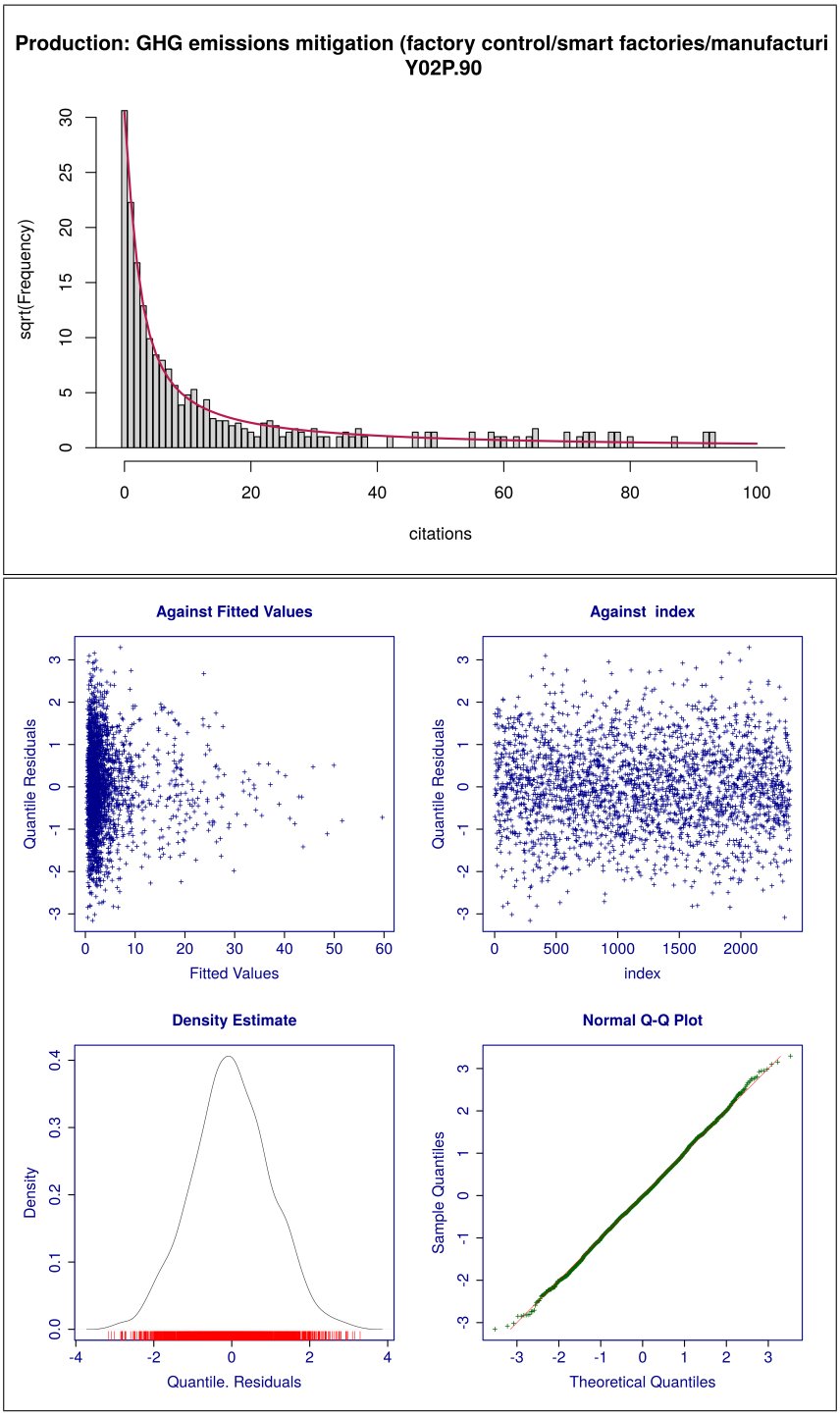

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

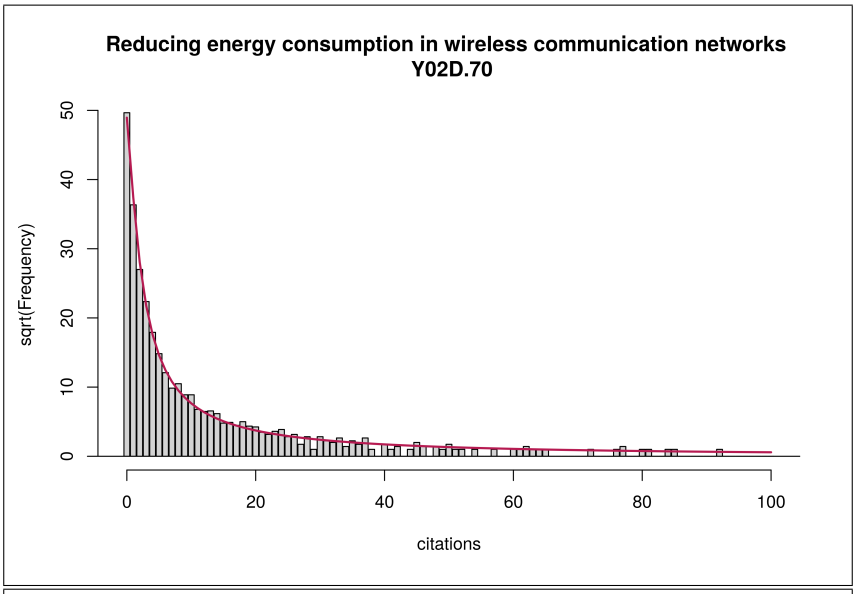

Transport: Road transport of goods or passengers
Y02T.10
sqrt(Frequency)
0
20
40
60
80
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

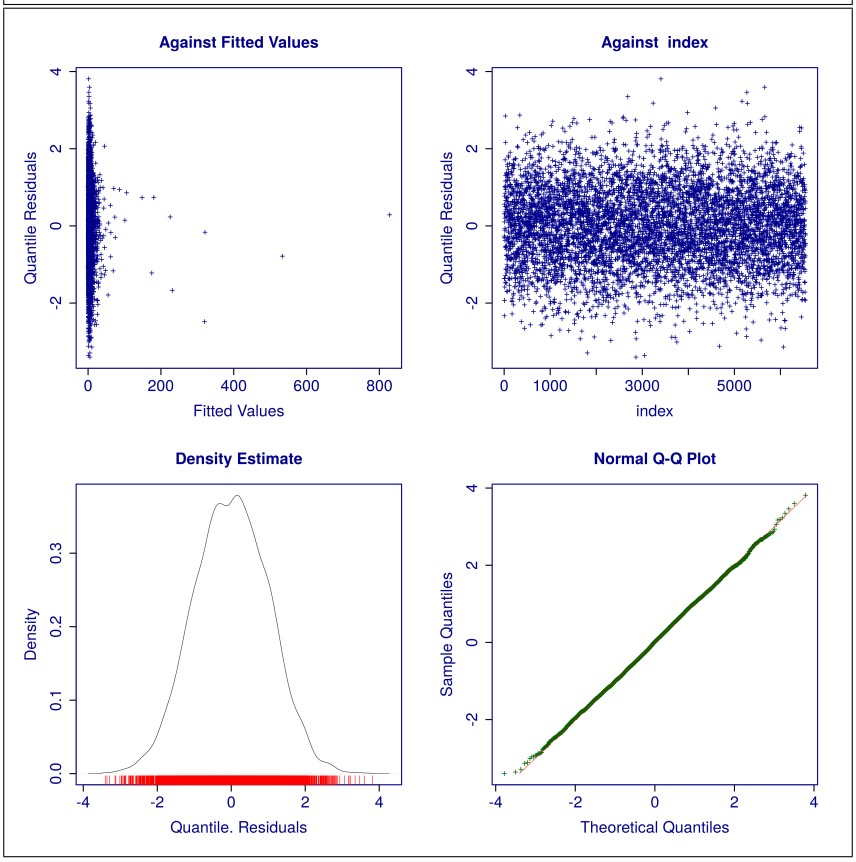

Mitigation: Energy efficient computing
Y02D.10
sqrt(Frequency)
0
10
20
30
40
50
0
20
40
60
80
100
citations
Against Fitted Values
Quantile Residuals
-4
-2
0
2
0
10
20
30
40
Fitted Values
Against index
Quantile Residuals
-4
-2
0
2
0
1000
3000
5000
7000
index
Density Estimate
Density
0.0
0.1
0.2
0.3
-4
-2
0
2
4
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
-4
-2
0
2
4
Theoretical Quantiles

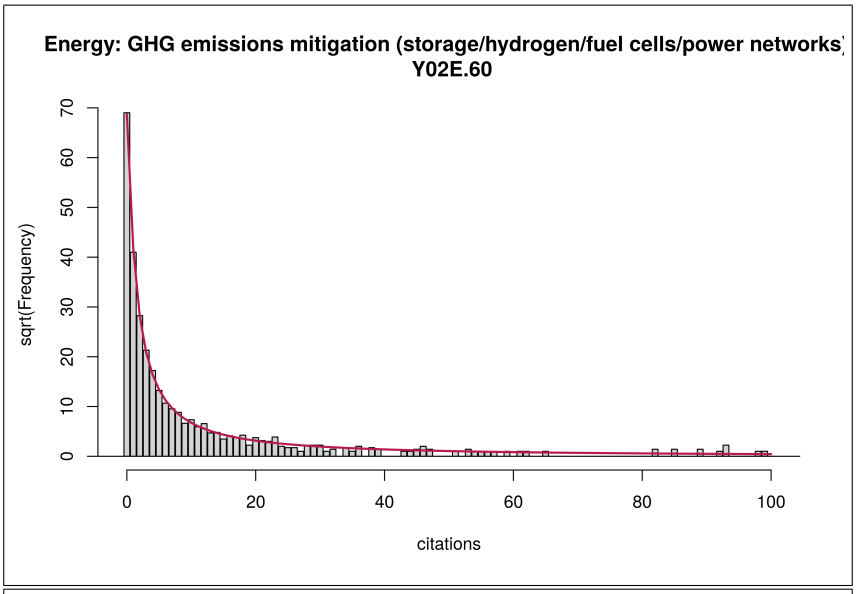

Energy generation through renewable energy sources
Y02E.10
sqrt(Frequency)
0
20
40
60
80
0
20
40
60
80
100
citations

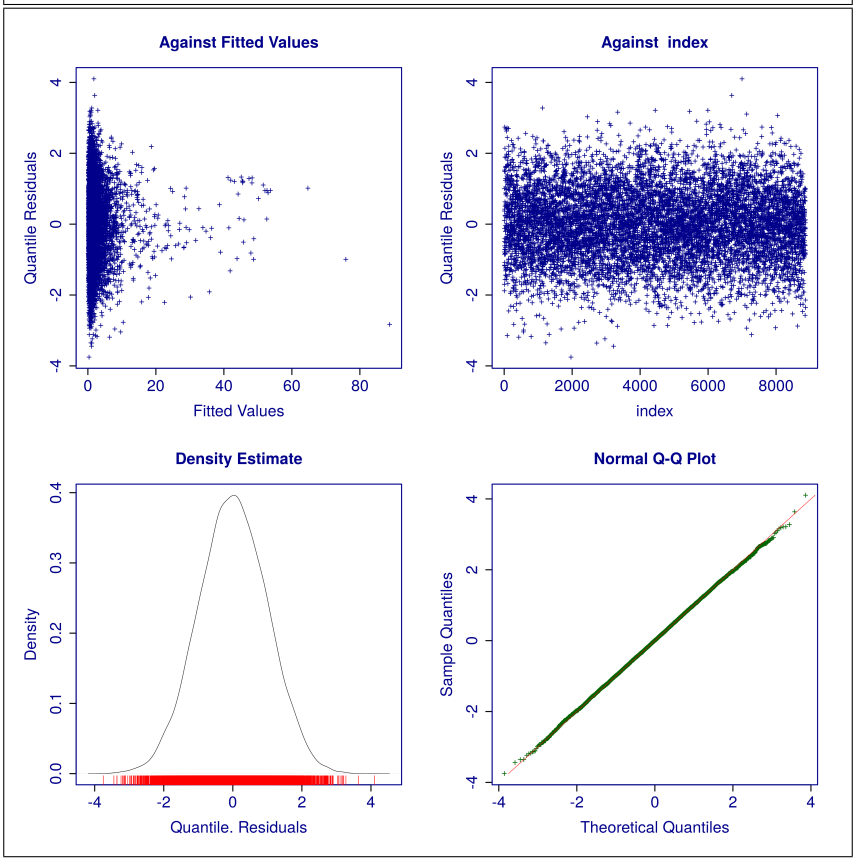

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Production: GHG emissions mitigation (factory control/smart factories/manufacturi
Y02P.90
sqrt(Frequency)
citations
Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Reducing energy consumption in wireless communication networks
Y02D.70
sqrt(Frequency)
0
10
20
30
40
50
0
20
40
60
80
100
citations


Against Fitted Values
Quantile Residuals
-2
0
2
4
0
200
400
600
800
Fitted Values
Against index
Quantile Residuals
-2
0
2
4
0
1000
3000
5000
index
Density Estimate
Density
0.0
0.1
0.2
0.3
-4
-2
0
2
4
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
-2
0
2
4
-4
-2
0
2
4
Theoretical Quantiles


Energy: GHG emissions mitigation (storage/hydrogen/fuel cells/power networks)
Y02E.60
sqrt(Frequency)
0
10
20
30
40
50
60
70
0
20
40
60
80
100
citations

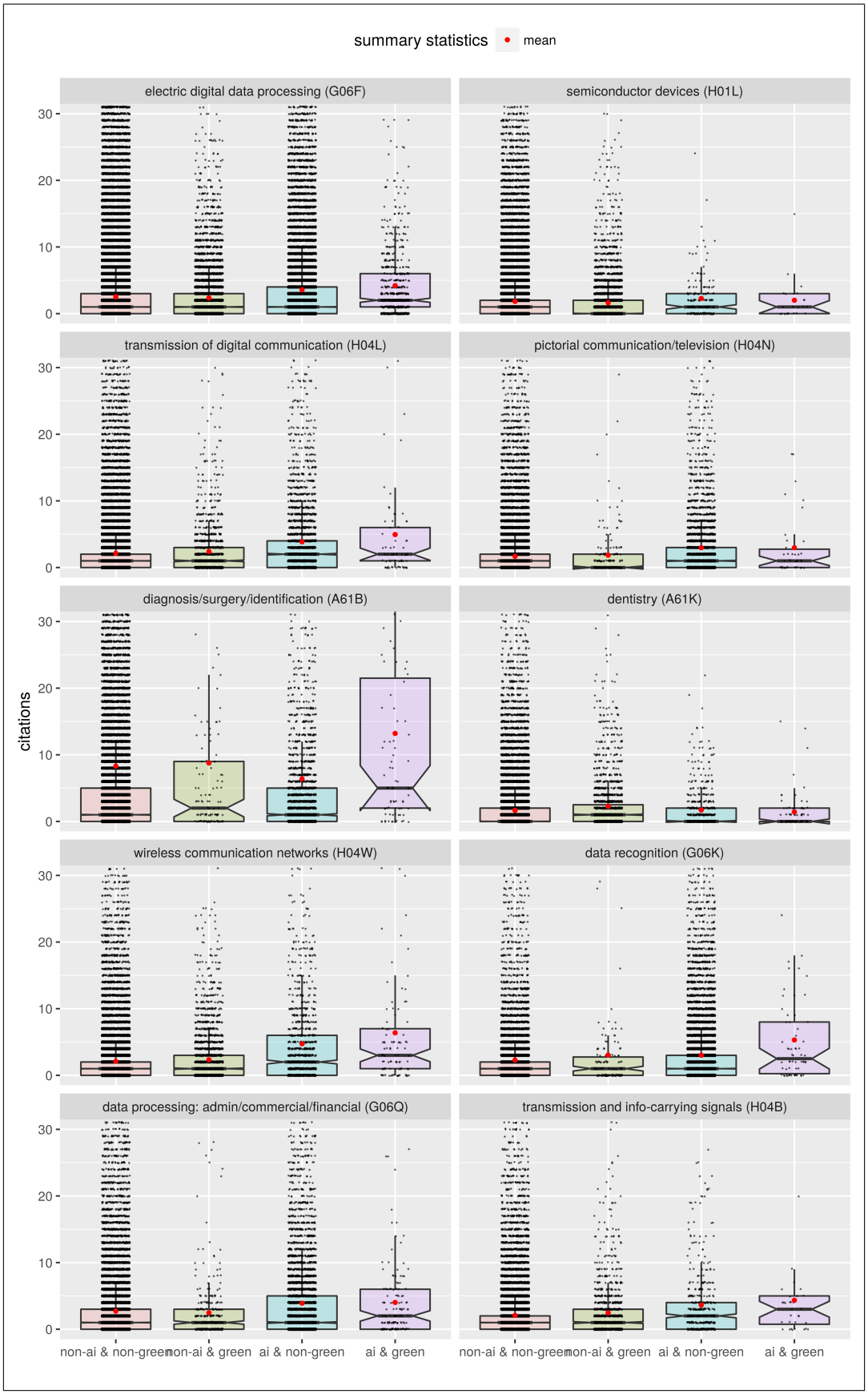

Against Fitted Values
Quantile Residuals
Fitted Values
Against index
Quantile Residuals
index
Density Estimate
Density
Quantile. Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

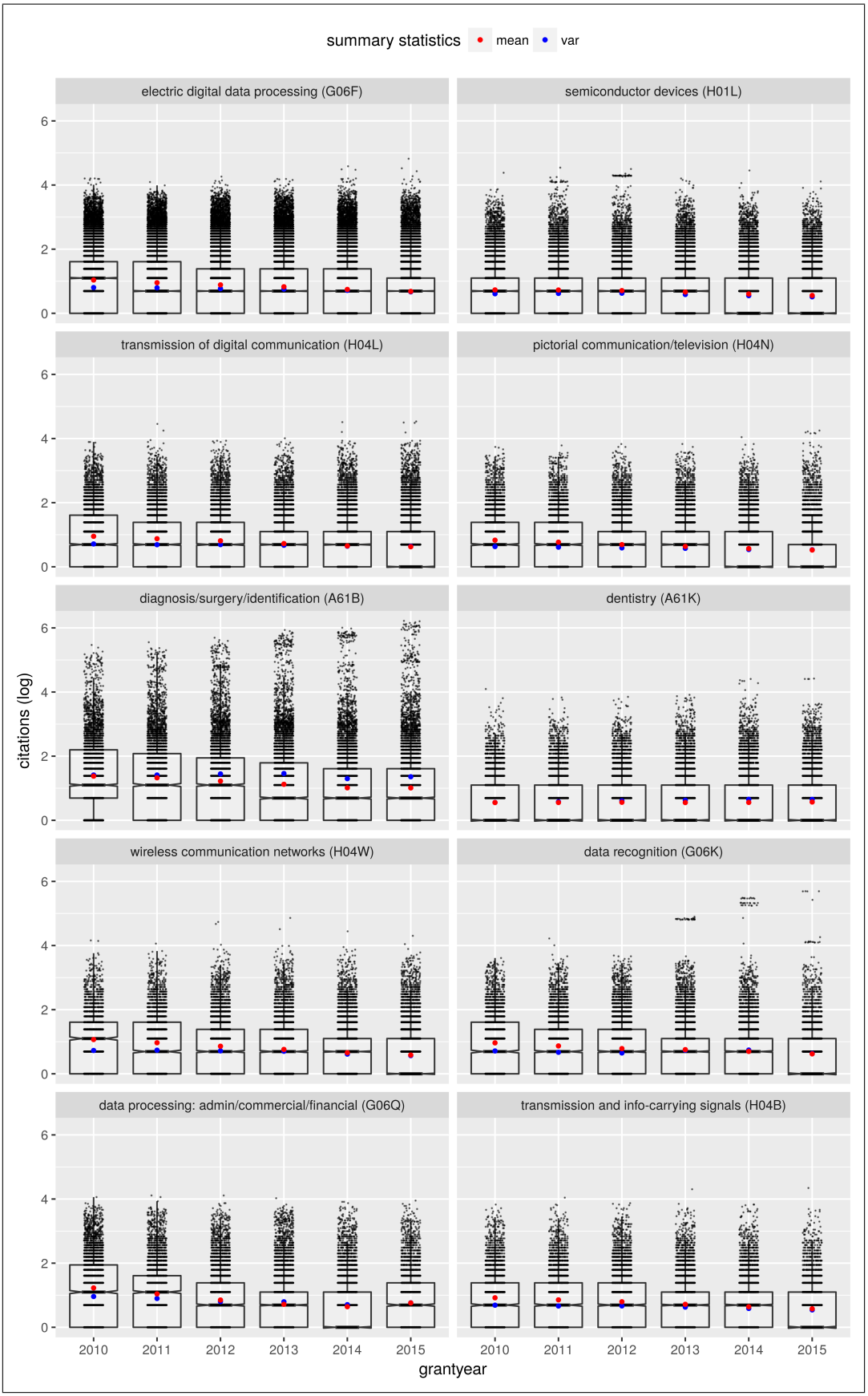
# 3 Exploring the data and the covariates

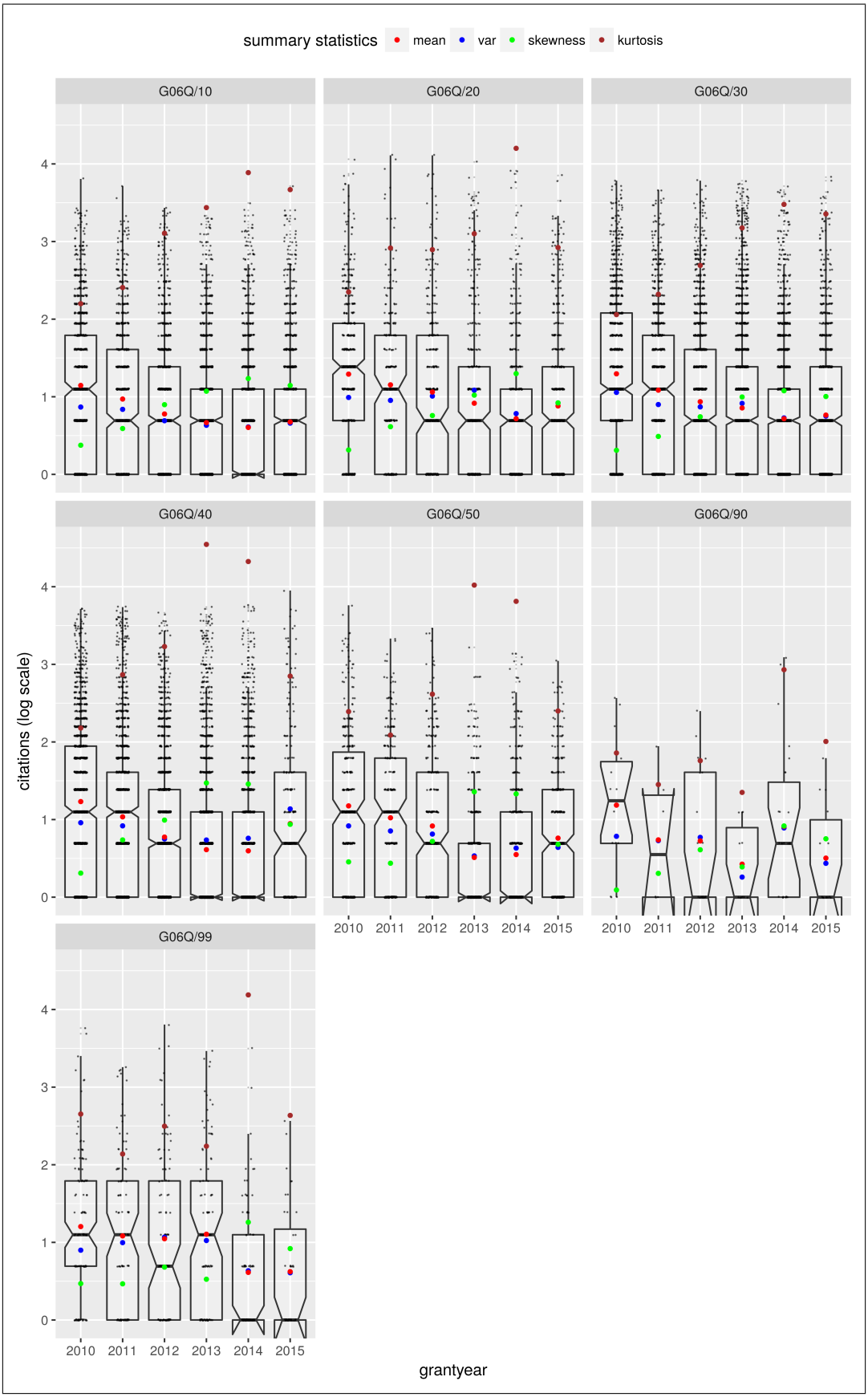
The following data exploration examines whether the log forward citations is reasonably linear on the covariates or not. Where the covariates have been transformed based on the raw data, we found that the transformation indicated a better model fit to the covariate. We compare a linear relation to more flexible regression models to assess if the difference is large or not.

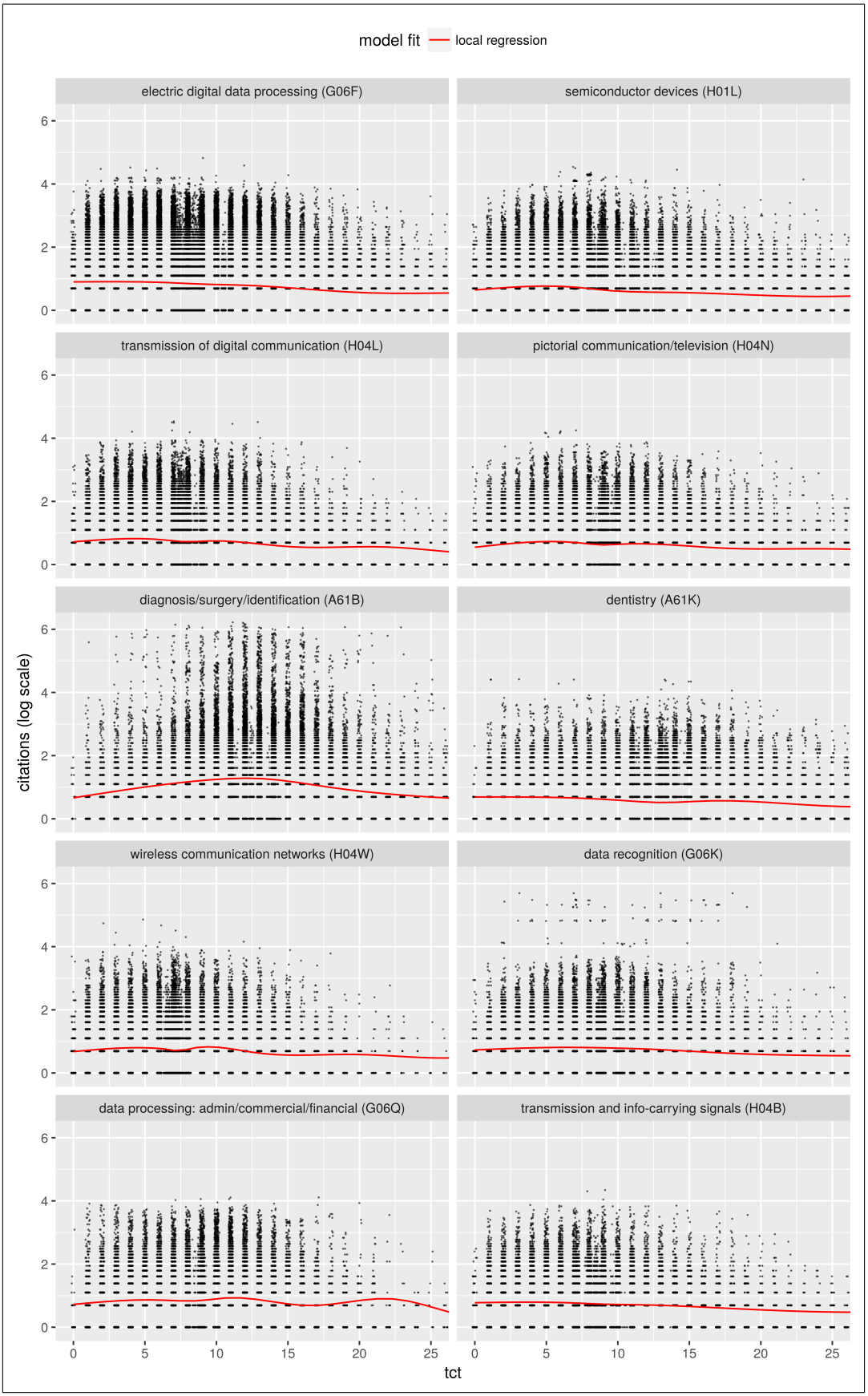
The following plots

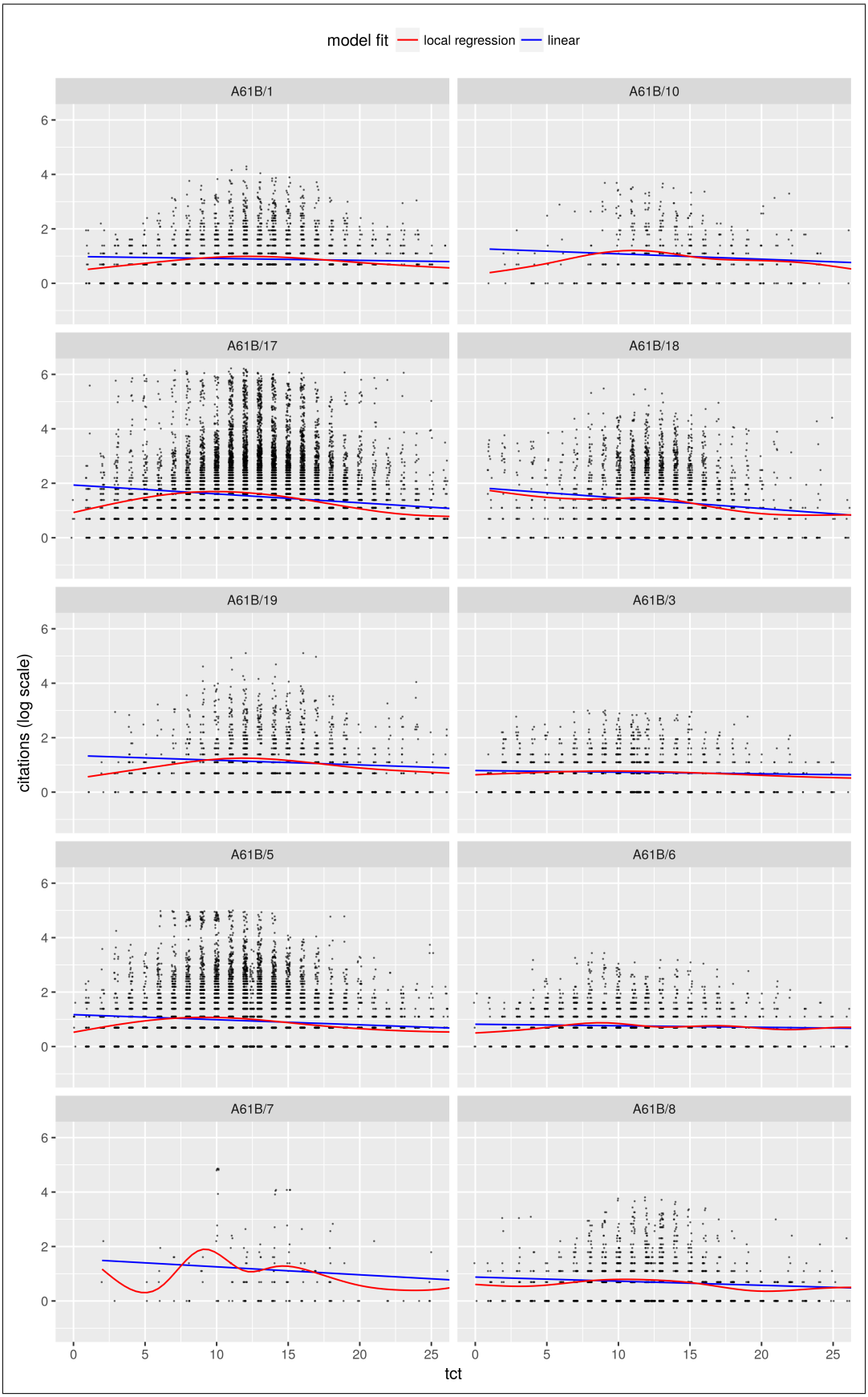
- Show that the mean/var and skewness/kurtosis can vary much across subclasses: These will be imporant for modeling the scale and shape of the distributions.
- Allow roughly checking if the covariates are reasonably linear with the target on the log scale (log x shown as a log x plus 1 transformation).
  - To visually inspect how the average of citations on the log scale varies over the range of continuous/count covariates, we use local regression (local polynomials with smoothing): This is the *gam* smoother option in *ggplot2* which can visually indicate non-linearities.
  - In most cases the deviation from the linear trend is not substantial to rule out the covariate in the model.
  - The TCT is sometimes clearly non-linear: We handle this by discretizing it (a factor for different ranges, 4 levels).

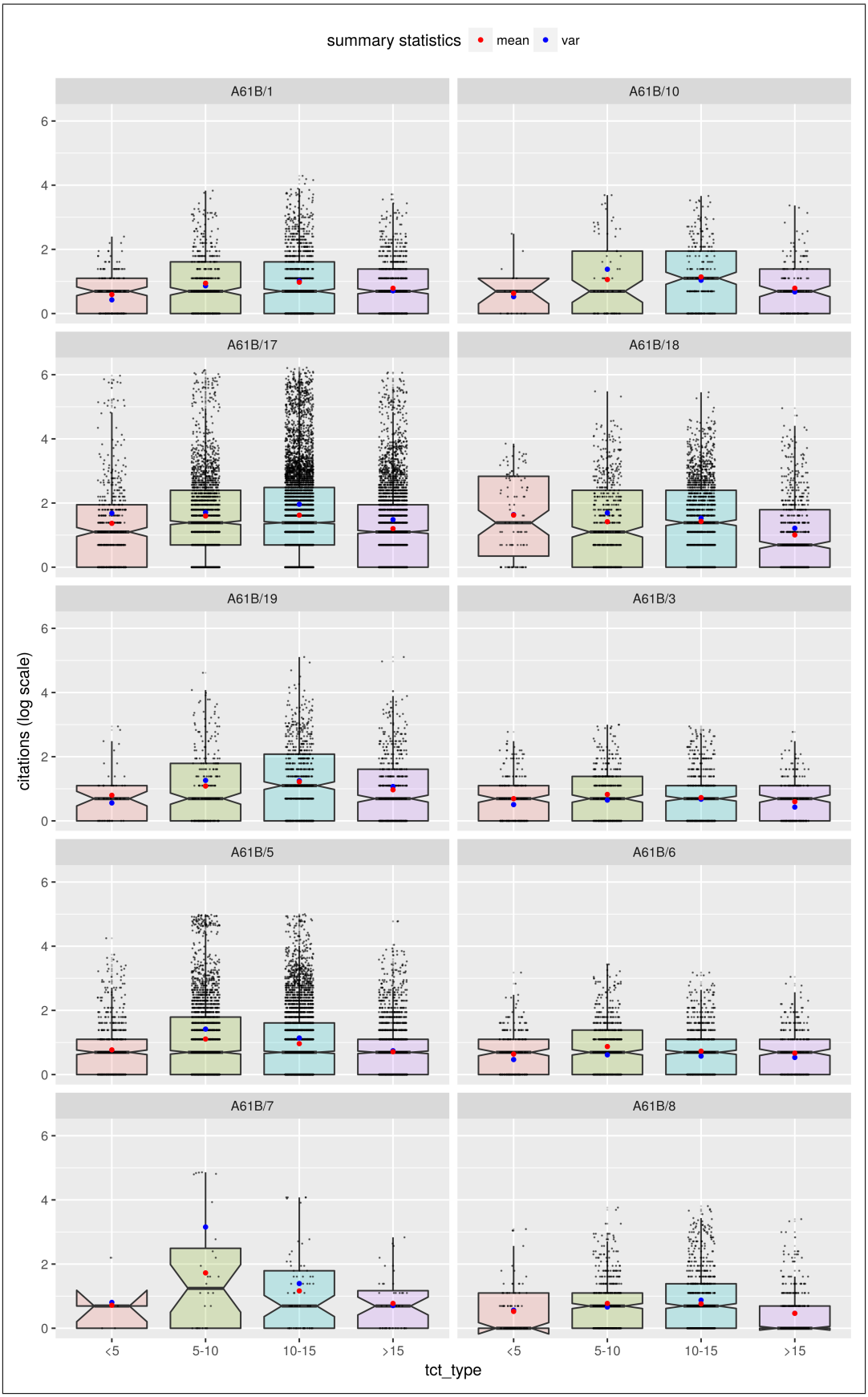

summary statistics
mean
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations
0
10
20
30
non-ai & non-green
non-ai & green
ai & non-green
ai & green

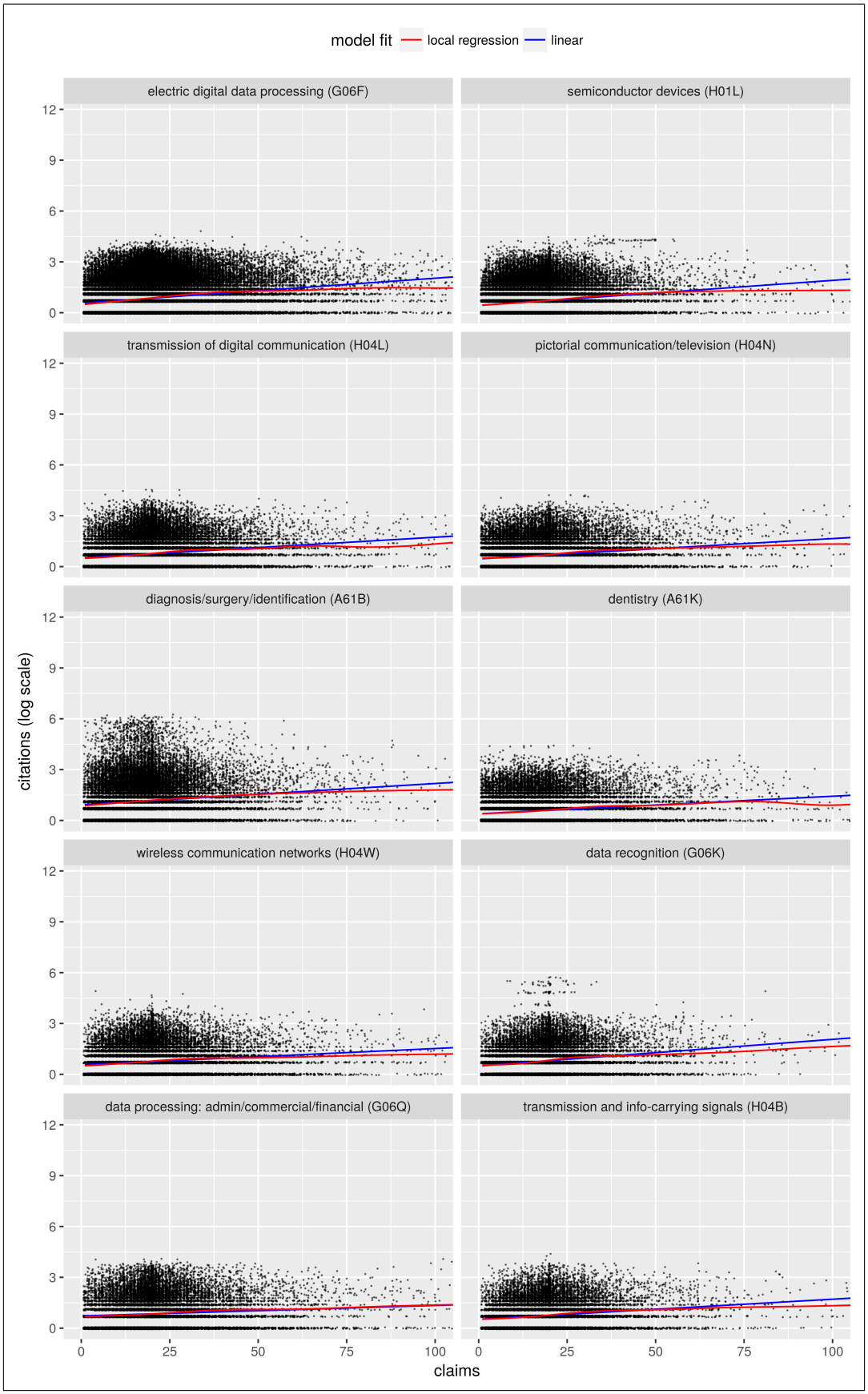

summary statistics
mean
var
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log)
0
2
4
6
2010
2011
2012
2013
2014
2015
grantyear

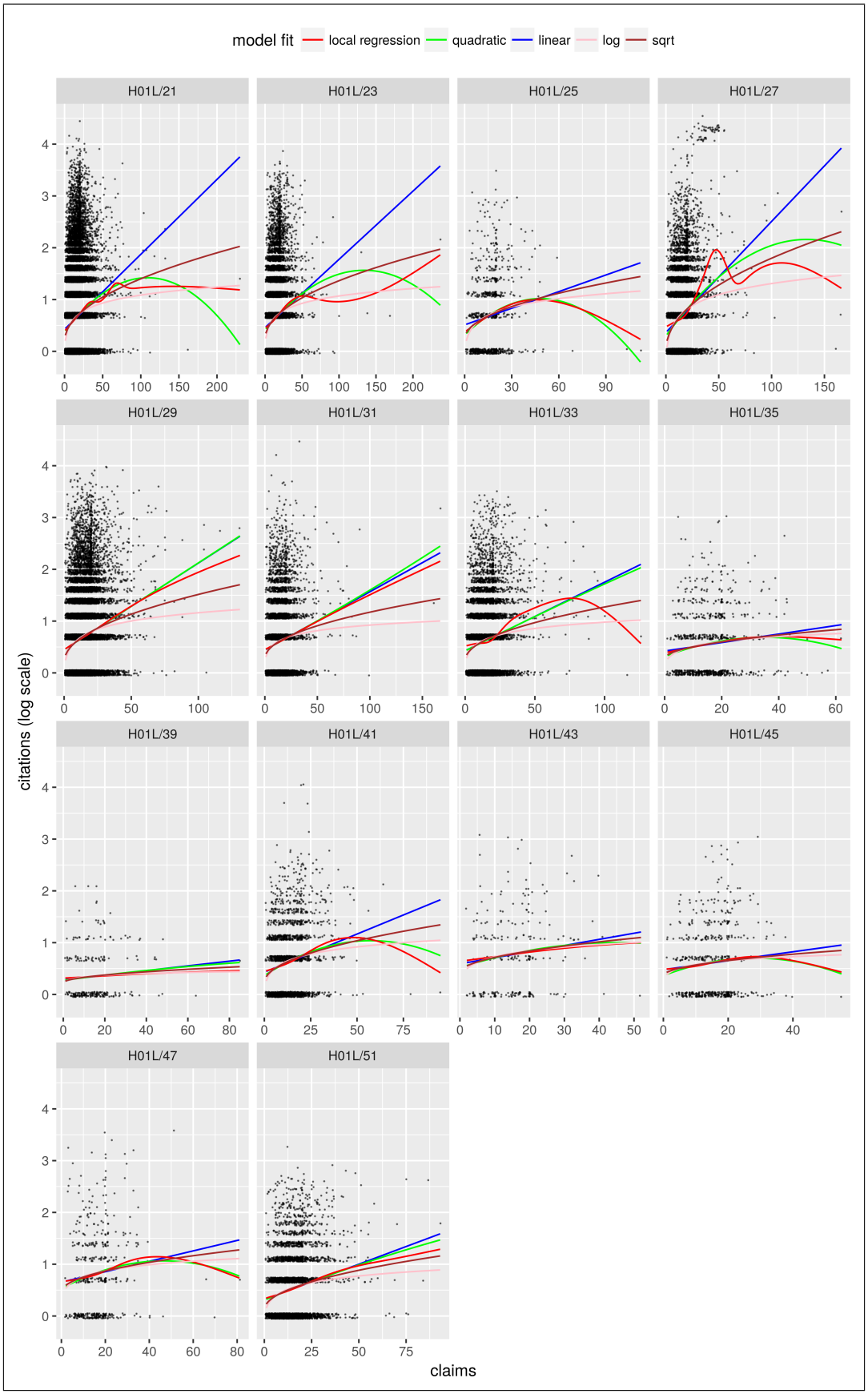

summary statistics
mean
var
skewness
kurtosis
G06Q/10
G06Q/20
G06Q/30
G06Q/40
G06Q/50
G06Q/90
G06Q/99
citations (log scale)
grantyear
0
1
2
3
4
2010
2011
2012
2013
2014
2015

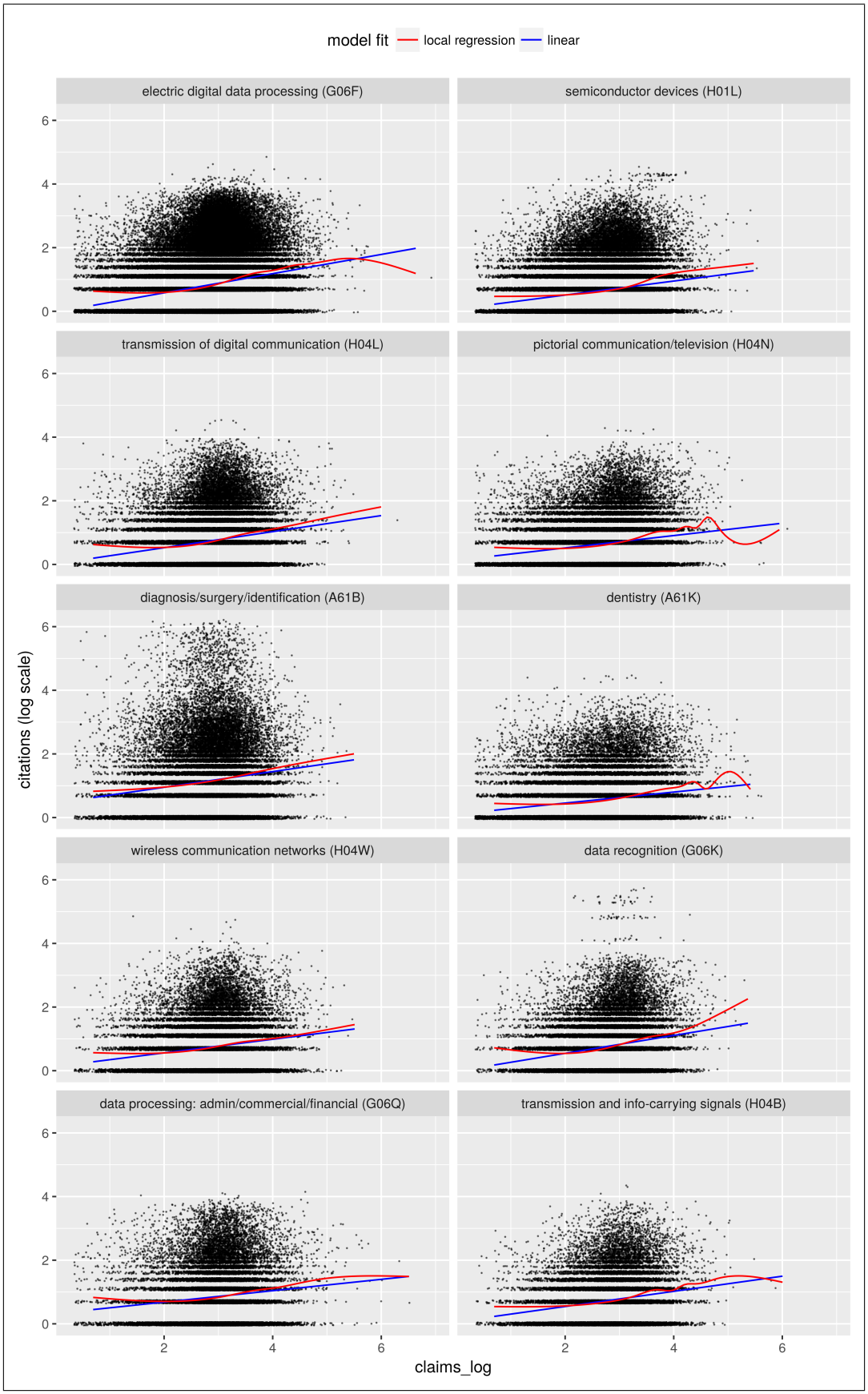

model fit
local regression
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
tct
0
2
4
6
0
5
10
15
20
25

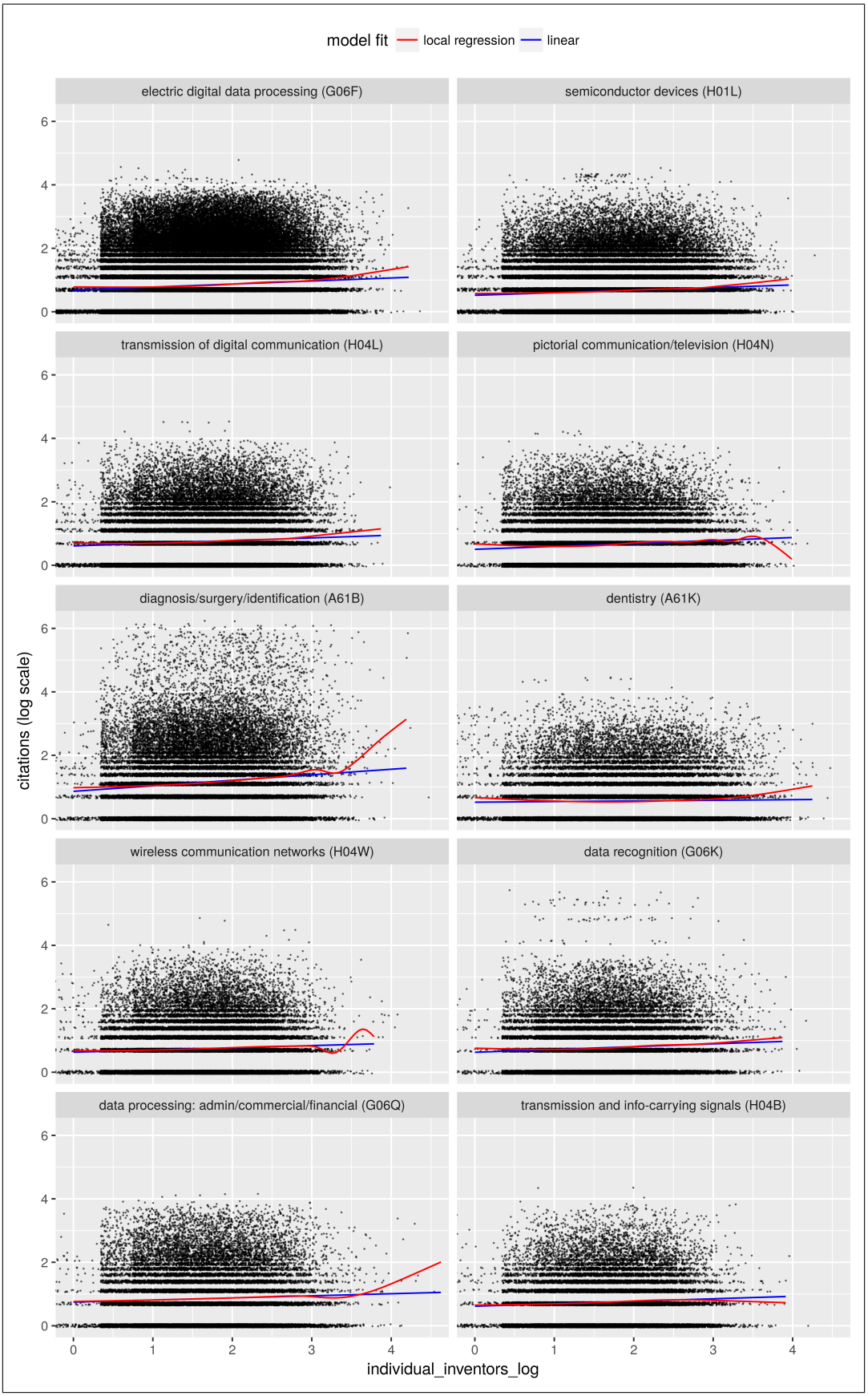

model fit
local regression
linear
A61B/1
A61B/10
A61B/17
A61B/18
A61B/19
A61B/3
A61B/5
A61B/6
A61B/7
A61B/8
citations (log scale)
tct
0
2
4
6
0
5
10
15
20
25

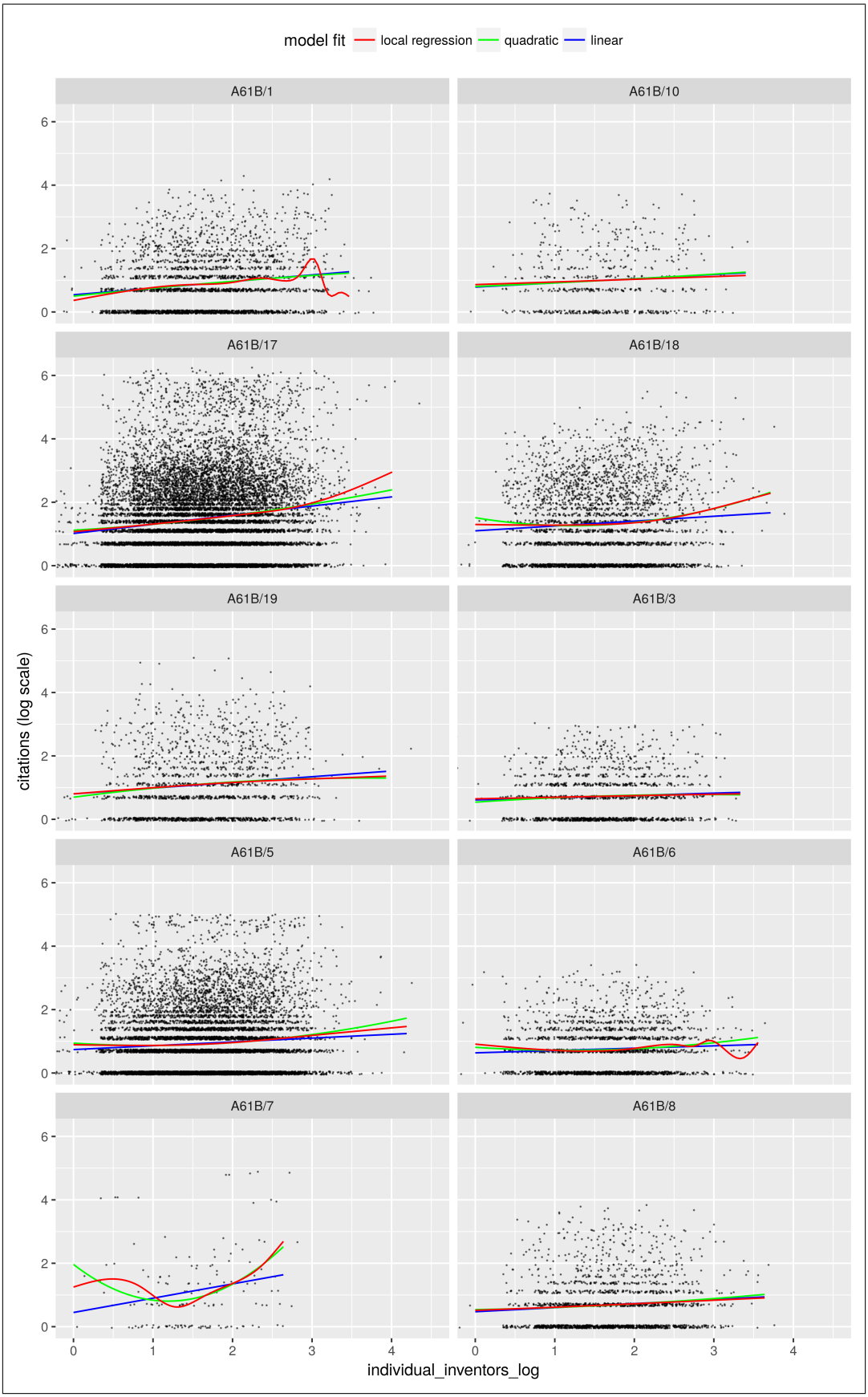

summary statistics
mean
var
A61B/1
A61B/10
A61B/17
A61B/18
A61B/19
A61B/3
A61B/5
A61B/6
A61B/7
A61B/8
citations (log scale)
0
2
4
6
<5
5-10
10-15
>15
tct_type

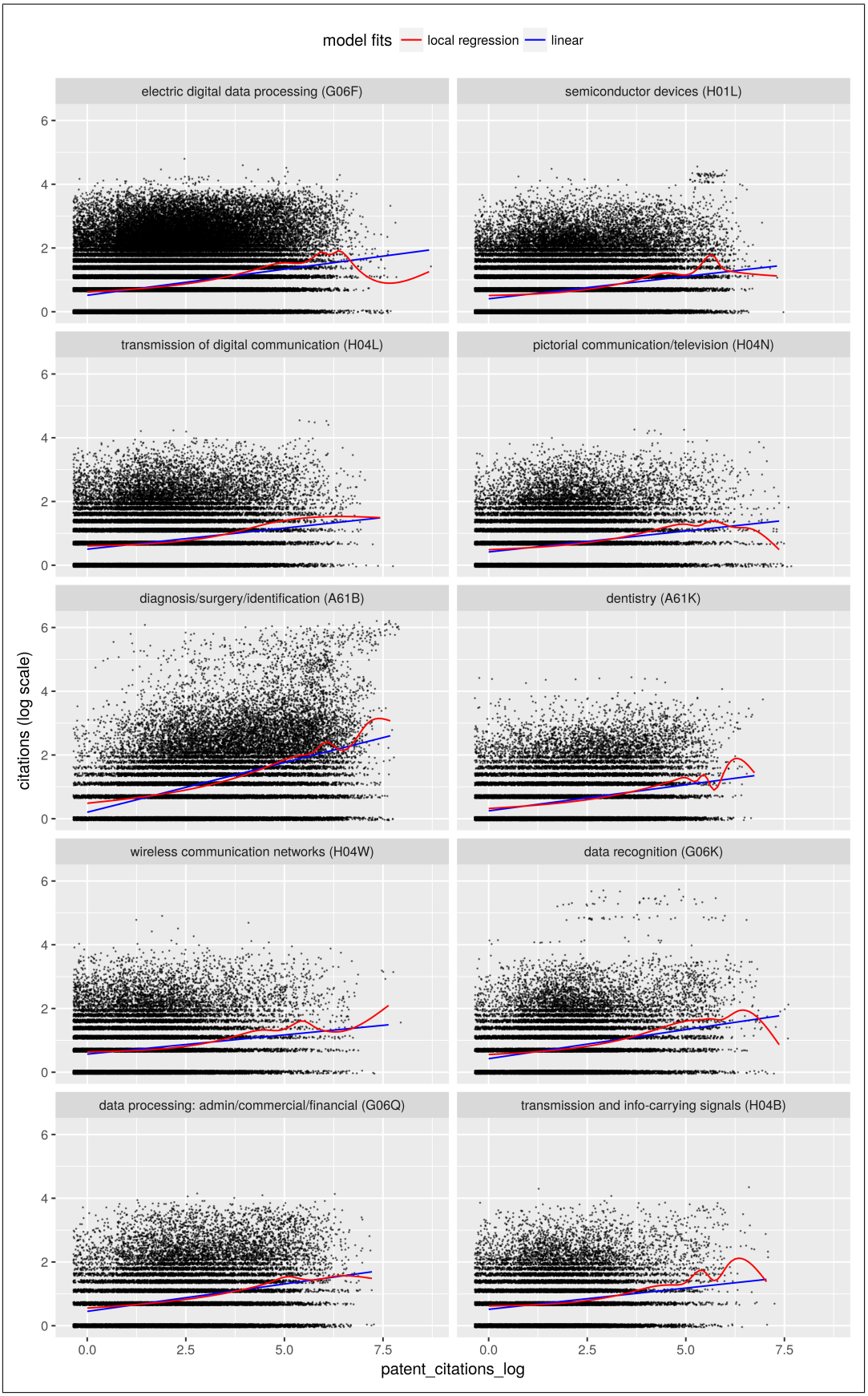

model fit
local regression
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
0
3
6
9
12
0
25
50
75
100
claims

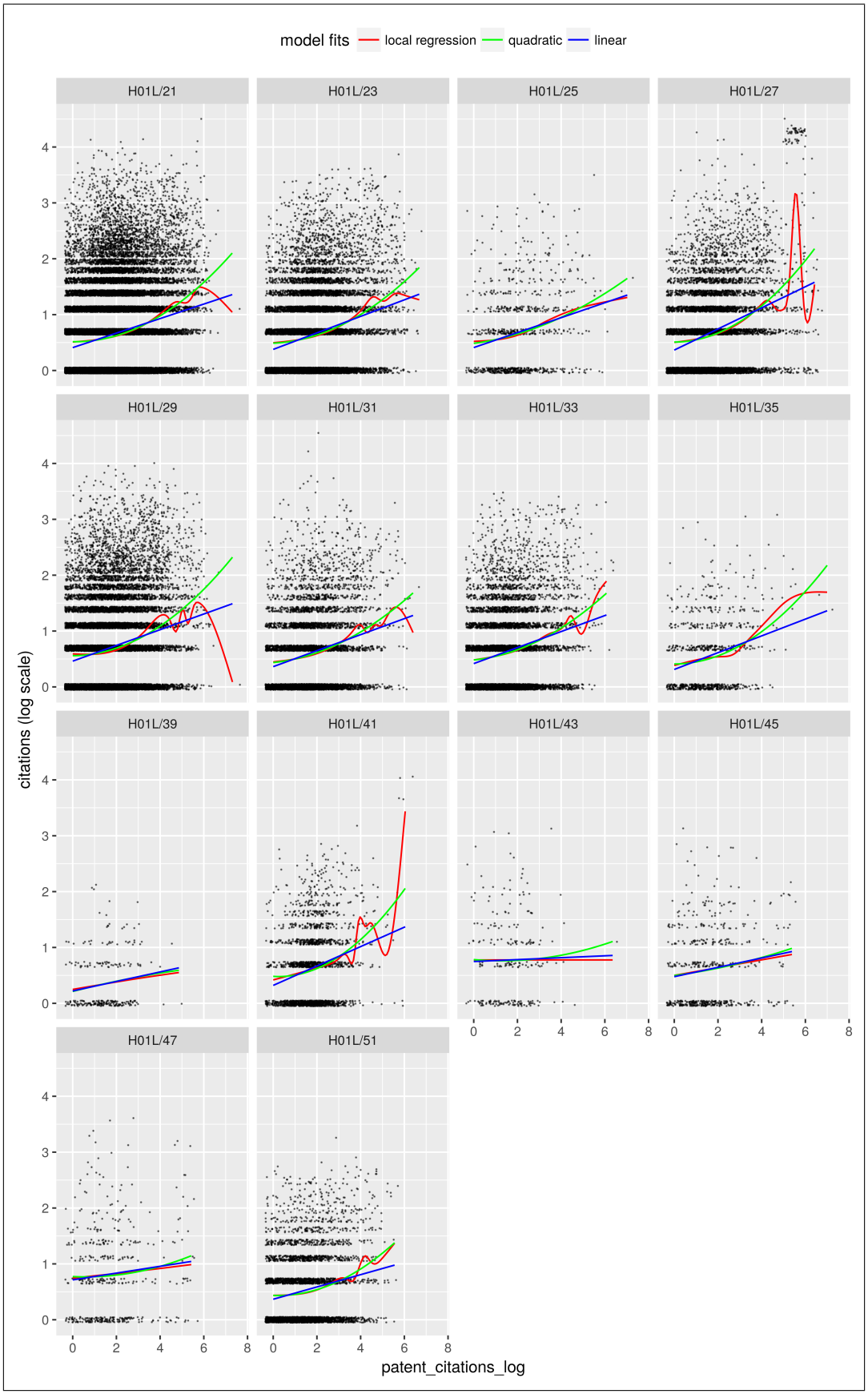

model fit
local regression
quadratic
linear
log
sqrt
H01L/21
H01L/23
H01L/25
H01L/27
H01L/29
H01L/31
H01L/33
H01L/35
H01L/39
H01L/41
H01L/43
H01L/45
H01L/47
H01L/51
citations (log scale)
claims

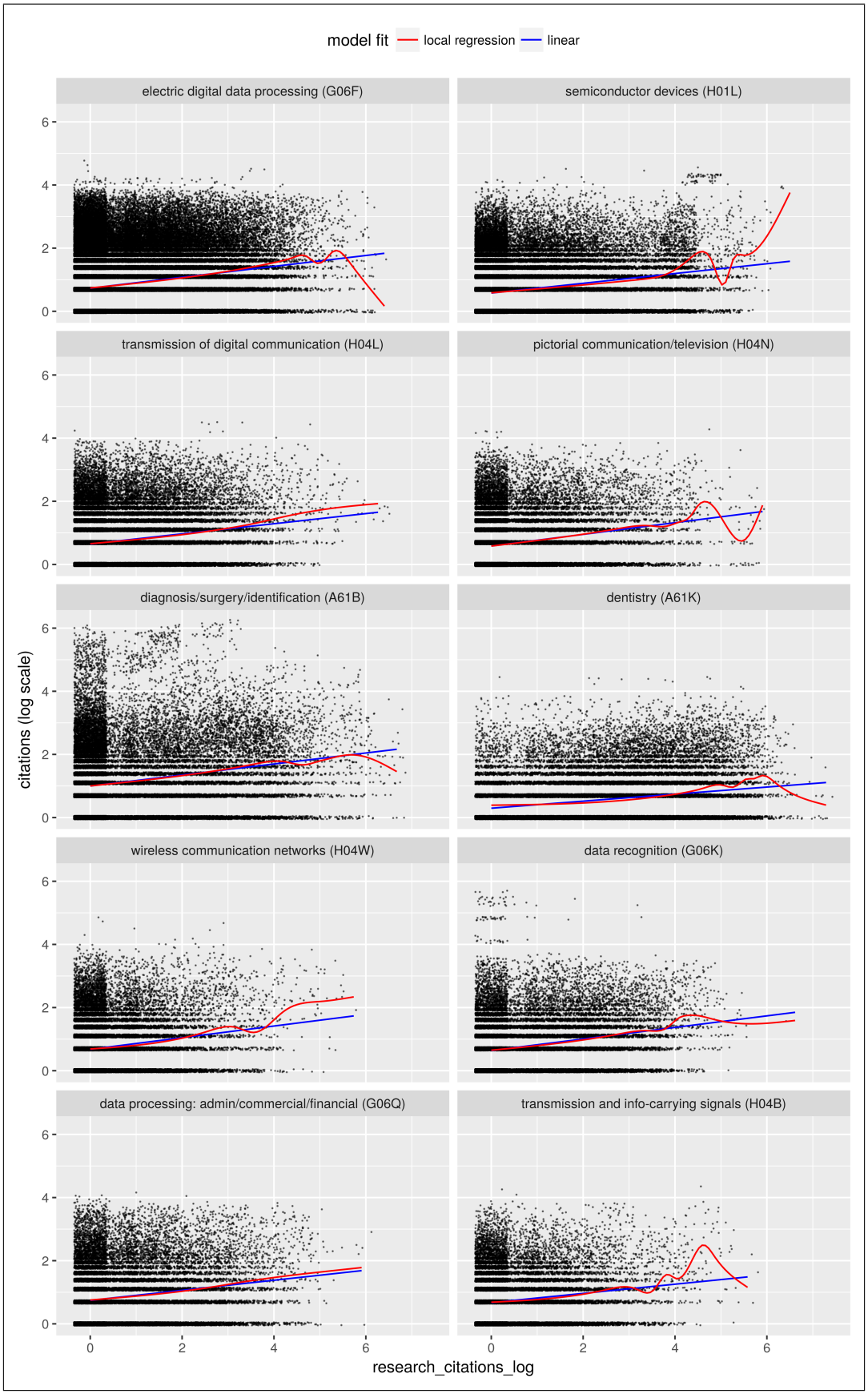

model fit
local regression
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
claims_log
0
2
4
6

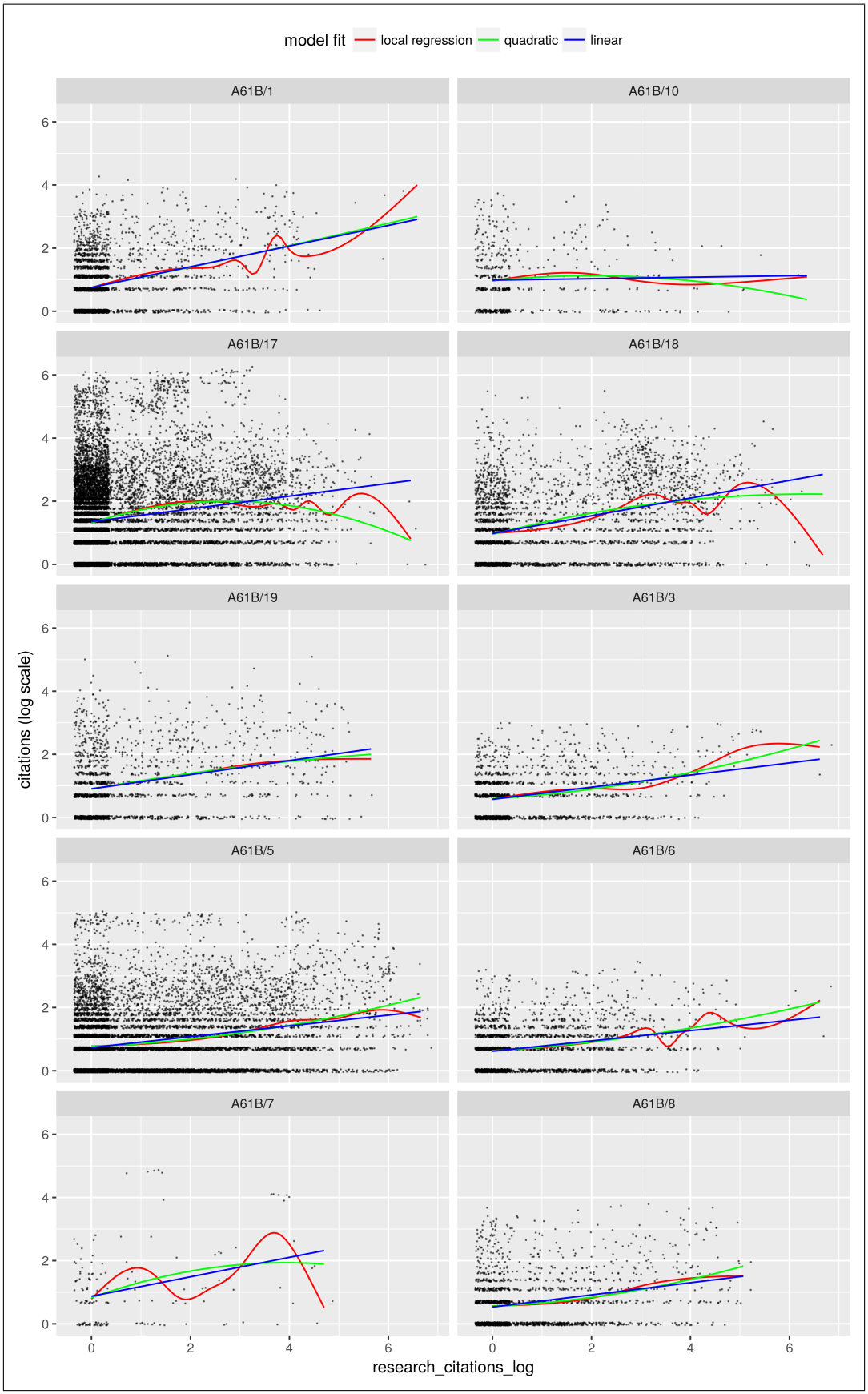

model fit
local regression
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
individual_inventors_log
0
2
4
6
0
1
2
3
4

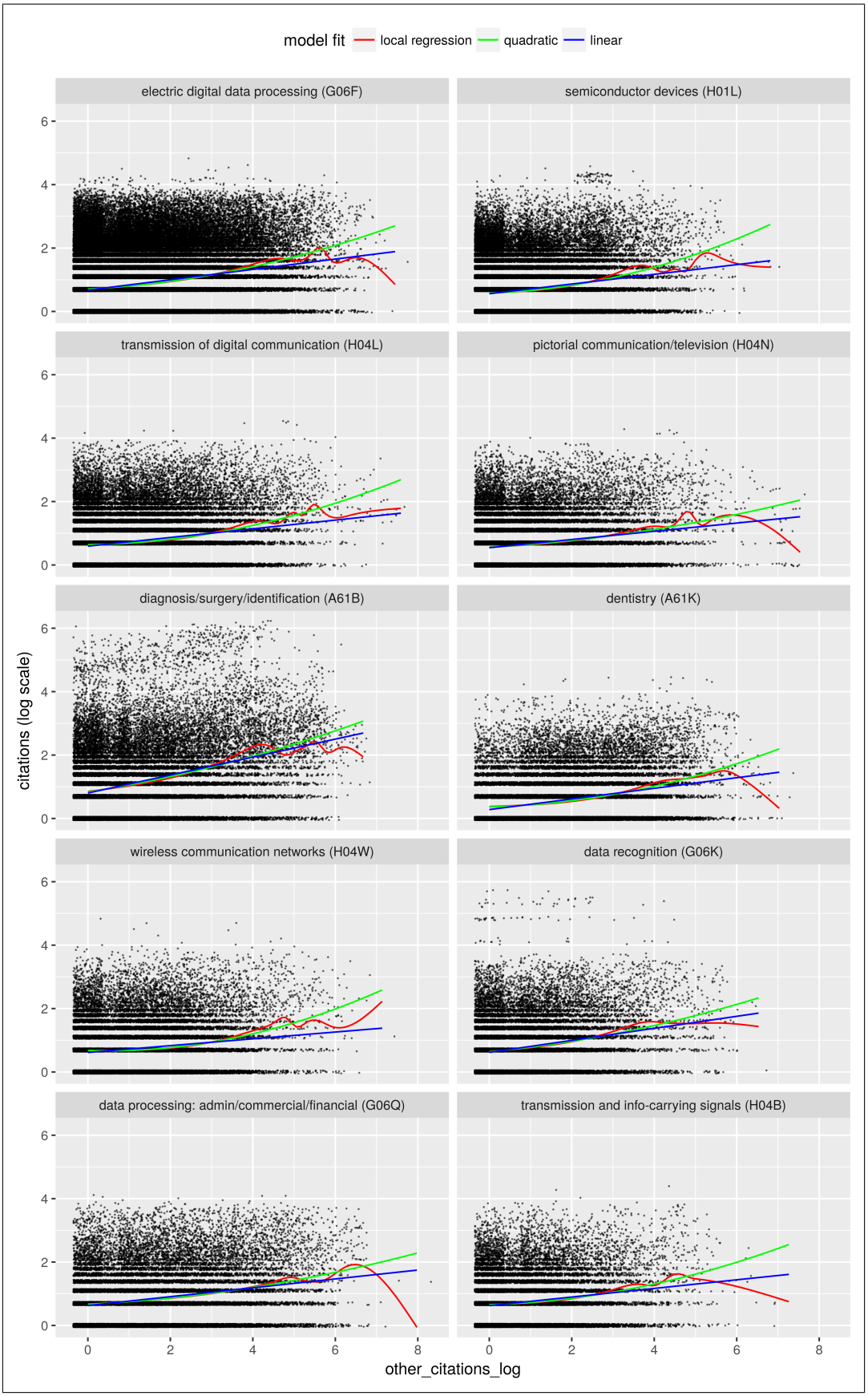

model fit
local regression
quadratic
linear
A61B/1
A61B/10
A61B/17
A61B/18
A61B/19
A61B/3
A61B/5
A61B/6
A61B/7
A61B/8
citations (log scale)
individual_inventors_log
0
1
2
3
4
6

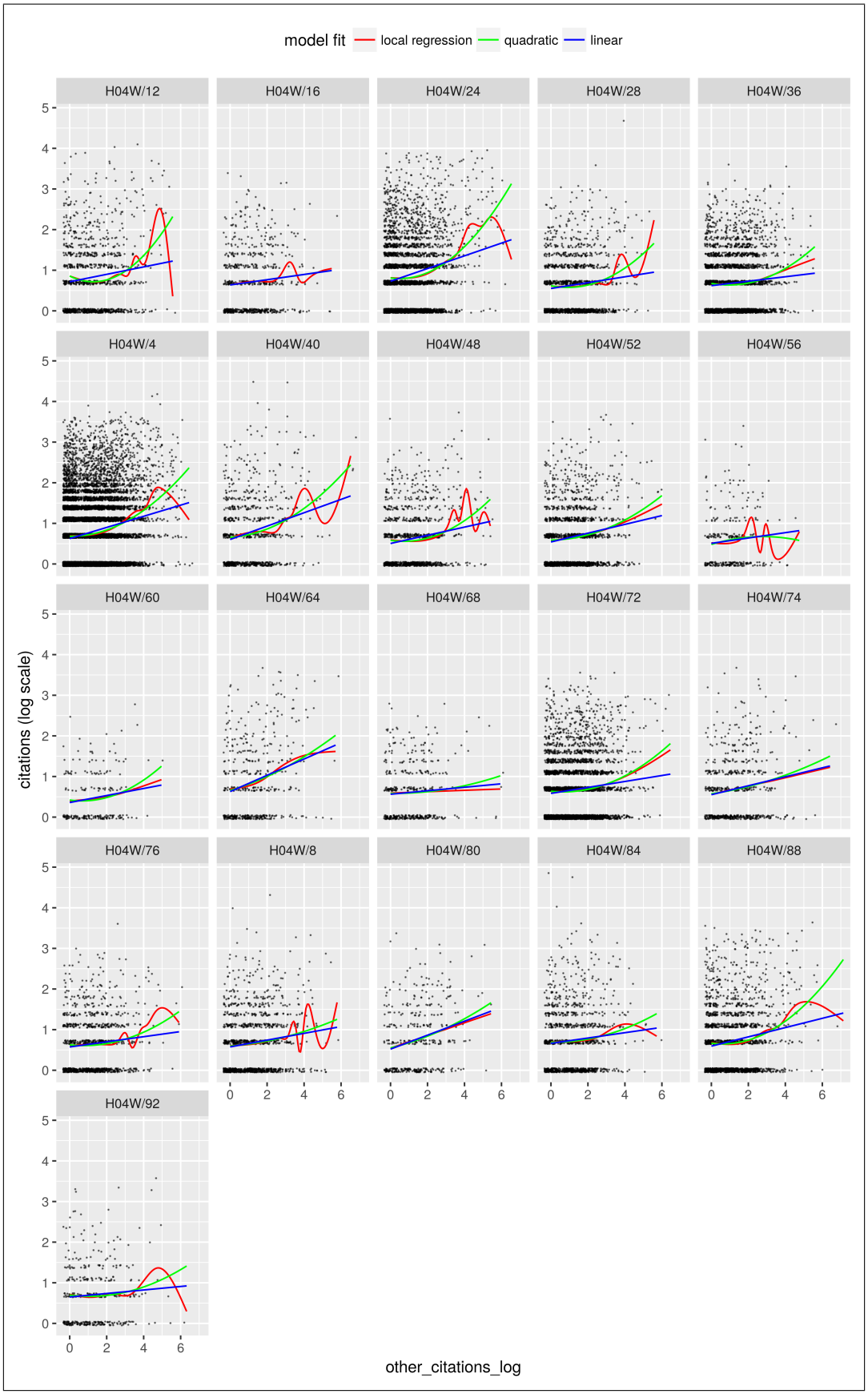

model fits
local regression
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
patent_citations_log
0
2
4
6
0.0
2.5
5.0
7.5


model fits
local regression
quadratic
linear
H01L/21
H01L/23
H01L/25
H01L/27
H01L/29
H01L/31
H01L/33
H01L/35
H01L/39
H01L/41
H01L/43
H01L/45
H01L/47
H01L/51
citations (log scale)
patent_citations_log


model fit
local regression
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
research_citations_log
0
2
4
6


model fit
local regression
quadratic
linear
A61B/1
A61B/10
A61B/17
A61B/18
A61B/19
A61B/3
A61B/5
A61B/6
A61B/7
A61B/8
citations (log scale)
research_citations_log
0
2
4
6


model fit
local regression
quadratic
linear
electric digital data processing (G06F)
semiconductor devices (H01L)
transmission of digital communication (H04L)
pictorial communication/television (H04N)
diagnosis/surgery/identification (A61B)
dentistry (A61K)
wireless communication networks (H04W)
data recognition (G06K)
data processing: admin/commercial/financial (G06Q)
transmission and info-carrying signals (H04B)
citations (log scale)
other_citations_log
0
2
4
6
8


model fit
local regression
quadratic
linear
H04W/12
H04W/16
H04W/24
H04W/28
H04W/36
H04W/4
H04W/40
H04W/48
H04W/52
H04W/56
H04W/60
H04W/64
H04W/68
H04W/72
H04W/74
H04W/76
H04W/8
H04W/80
H04W/84
H04W/88
H04W/92
citations (log scale)
other_citations_log