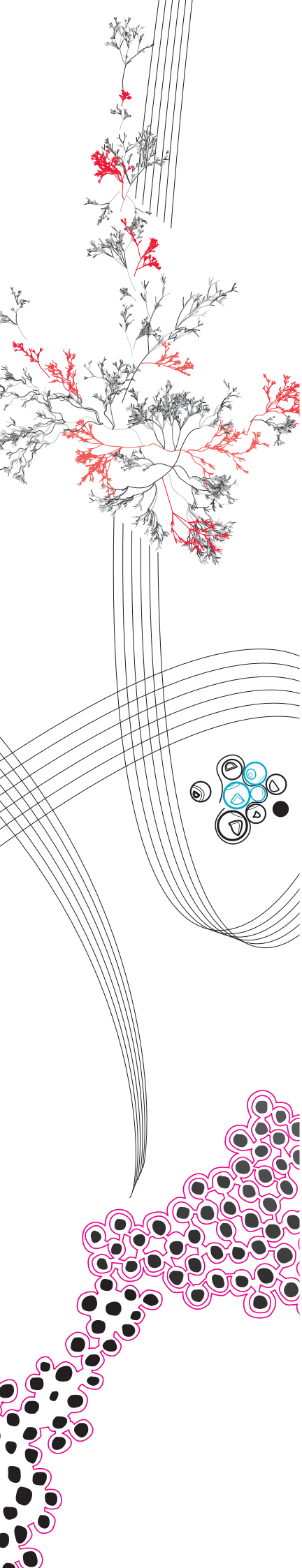

Master thesis Applied Mathematics (Chair: Stochastic Operations Research)
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

Predicting long term impact of scientific publications

Clara Stegehuis

August 19, 2014

Assessment Committee:
Prof. dr. R.J. Boucherie (UT/SOR)
Dr. N. Litvak (UT/SOR)
Dr. L.R. Waltman (CWTS)
Dr. P. Frasca (UT/HS)

UNIVERSITY OF TWENTE.

Contents

Abstract

In this thesis, first the distribution of the total number of citations that articles receive is studied. A discretized lognormal distribution and a negative binomial distribution are compared. The discretized lognormal distribution fitted well onto articles in the scientific field of Physics that were published in 1984, whereas the negative binomial distribution did not. The tail of the distribution of the number of citations was also studied. This tail seemed to have a Pareto distribution, with tail index independent of the Impact Factor and the number of citations in the first year after publishing.

Then, we propose a model to predict the quantiles of the distribution of the number of additional citations that scientific publications receive after the first year. We study three variants of the model: one uses only the Impact Factor as a covariate, one only the number of recent citations, and the last model uses both covariates. Quantile regression is used to fit the coefficients of the model. The model that uses both covariates fits the quantiles better than the other two variants. Then a well known estimator for the quantiles of a Pareto distribution is used to describe the coefficients of the quantile regression estimator for the high quantiles. Furthermore, confidence intervals for both estimators are given.

The model that was proposed, predicted the quantiles of the distribution of the number of additional citations after the first year well. However, the model did not predict the quantiles correctly for groups of articles from the same country or university. We present a simple example to give a possible explanation for this phenomenon.

Chapter 1

Introduction

1.1 Research question

Citation frequency is often used as an indicator of the impact of scientific publications. For the evaluation of institutions, bibliometric indicators on the publications of an institution and their impact are also used. However, using the citation frequency can be problematic when recent papers are evaluated. One or two years after the publication date, most publications have only received a few citations. After one year, there are a lot of publications with only one citation. Some of those publications might end as one of the most cited publications several years later, and others are uncited in the next years. This makes it difficult to evaluate the impact of recent scientific publications.

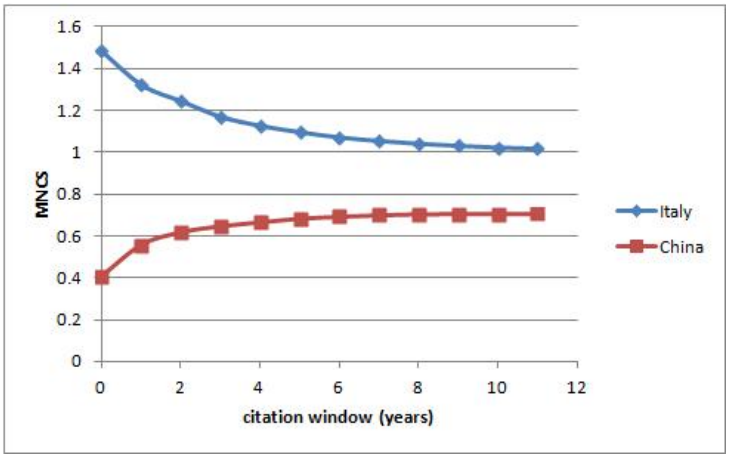

Figure 1.1: MNCS for Chinese and Italian publications in Physics for different time windows

Figure 1.1 gives an example of what can go wrong when evaluating the research performance of recent publications. This figure shows the mean normalized citation score (MNCS) of Italian and Chinese publications published in 2002 in the field of Physics. The MNCS tells how many citations articles in a certain group get, compared to all other articles of the same field. For example, an MNCS of two for the Italian publications, would in this case mean that Italian publications in the field of Physics that were published in 2002, got on average twice as much citations as an average Physics article published in 2002. The MNCS is plotted against the citation window. The citation window is the number of years after which the MNCS was computed after the publishing date. For example, in this case, if the citation window is one, this means that the MNCS is computed on the number of citations the articles got until 2003. The figure shows that for publications from China, the MNCS is increasing as the time window increases. In this case, if we were to evaluate the performance of Chinese articles one year after their publication, the MNCS is 0.55. However, if we were to evaluate the same publications a few years later,

we would see that the MNCS has increased to 0.7. This means that by evaluating the Chinese publications only one year after publishing, we would underestimate them. For the Italian publications on the other hand, the MNCS is decreasing as the time window increases. For this reason, Bornmann argues in [2] that recent publications should not be used in evaluations.

However, for institutions it might be more interesting how their recent publications perform, than how their publications of some years ago performed. Thus, this thesis is motivated by the following question:

How to improve the prediction of the long term impact of recent papers?

There are several factors that could be taken into account to improve the prediction of the long term impact of recent publications. In this report, we only want to take into account factors that could reflect something about the impact of the article. For example, one factor could be if in article is published in a journal with a high Impact Factor. The number of pages that the article consists of for example, does not meet this criterion. This report investigates the role of the Impact Factor of the journal in which an article was published and the number of recent citations to predict long term impact.

1.2 CWTS

This research was done in collaboration with the Centre for Science and Technology Studies (CWTS). The CWTS is a research institute at Leiden University that studies the dynamics of scientific research. Using a large database that is based on the Thomson Reuters' Web of Science database, they study science in a quantitative way. For example, the CWTS studies bibliometric indicators and the visualization of bibliometric networks. One well known product of the CWTS is the CWTS Leiden ranking, which ranks 750 universities on their scientific performance every year.

1.3 Related work

The distribution of the number of citations is already widely studied in literature. Several distributions for the number of citations have been proposed. Wallace et al. [25] for example, first study the probability of remaining uncited. They model the proportion of articles remaining uncited as a random selection process. This model is then used to derive the distribution of the numbers of citations articles get. In this way, a stretched exponential distribution is suggested for the distribution of the number of citations articles receive.
Burrell [18] assumes that every paper has its own rate at which citations are obtained according to a Poisson process. Using this assumption, it is shown that the distribution of the number of citations to articles should follow a negative binomial distribution. This is supported by data from articles published in the scientific field of Management Sciences.
In [24], Stringer et al. assume that every paper has a quality factor, and papers obtain citations according to this quality factor. Using this, they propose a discretized lognormal distribution for the distribution of the number of citations. Radicchi et al. also concluded that a lognormal distribution can be used to approximate the distribution of the number of citations [20]. In their analysis, the continuous lognormal distribution is fitted, and uncited articles are left out of the analysis. In [17], Lundberg also shows that the logarithm of the number of citations articles get can be approximated by a normal distribution. These are only a few of the models that have been proposed for the distribution of the number of citations.

Special attention has been given to the tail of the citation distribution, and to whether this tail follows a Pareto distribution. The possibility of a Pareto tail in bibliometrics was already mentioned by De Solla Price in 1976 [8]. By considering a cumulative advantage process, De Solla Price showed that the distribution of the number of citations should follow a power law. In 1998, Redner [21] studied a large dataset of citations to individual papers, and concluded that the tail of the distribution of the number of citations could be described by a power law. Clauset et al. [7] also study the tail of the citation distribution. A test for determining whether empirical data follow a Pareto distribution is proposed. For the data on the distribution of the number of citations, a Pareto tail could not be ruled out. In [1], Beirlant

et al. even propose the Pareto tail index as a performance indicator to evaluate the publications from a certain research group. This indicator should measure the outstanding citation impact.

Some of the models that are proposed predict the long term impact of individual papers. In [2], Bornmann studies three ways to predict the impact of recent papers from a certain university. The first method considers the performance of articles from the same university until ten years back. Using regression, the impact of the papers that have just been published, can be predicted. The second method is based on cluster samples. The third method can be used to find out if the articles from one university obtain consistently higher citation counts than those from another university.
Bornmann et al. also use a regression method in [3]. Using several covariates, such as the number of pages, the number of authors and the number of references of the article, linear regression is used to predict the impact of individual articles.
In [11], Golosovsky ans Solomon propose a stochastic growth model for the number of citations that articles get. This model is based on a superlinear preferential attachment model. Wang et al. [26], propose a model for how the number of citations to one single paper changes over time. They use citation history of the first ten years after an article has been published to predict the distribution of the number of citations over time to this article in the next years.

Other existing literature investigates which factors determine the number of citations that publications eventually receive. In [19], Peters and Van Raan study several factors. A multiple regression analysis was performed to conclude which factors together have the most predictive power. The factor 'top author', a factor that is a property of the first author, seemed to be the most influential factor. Other factors of importance were the number of references included in the article, the language of the article and the journal influence weight.
In [27], Wang et al. study 25 bibliometric indicators. A case based classifier model selected the factors that were the most determining for the long term impact of the papers. The number of citations that articles obtained in the first five years after publishing, the time until the first citation to the article, the journal Impact Factor and the h-index of the first author turned out to be the most important factors in this model.
In [16], Levitt and Thelwall find that there is a high correlation between the number of citations an article obtains in the long term and a weighted sum of the number of recent citations and the Impact Factor. Stern [23] has similar findings when looking at the correlation between the rank of articles after several years and a combination of the Impact Factor and the number of recent citations.

1.4 Contributions of this report

This report investigates two different aspects of the citation distribution. First, we investigate the distribution of the number of citations. As seen above, a lot of models have been proposed for this distribution in literature. In this report, two of these distributions will be compared by investigating their fit to the data: the distribution that Burrell [18] proposed, and the distribution that Stringer et al. [24] proposed.

The second part of this report focuses on a prediction model for the citation impact of recent articles, using the Impact Factor and the number of recent citations as covariates. Previous work that used these factors to predict long term impact, focused mainly on the influence of these factors on the mean number of citations. In this report, we predict the quantiles of the citation distribution using Quantile Regression, a technique that has not been used before to analyze the citation distribution. Quantile Regression allows to investigate the influence of the covariates at the different quantiles of the distribution.

This report also investigates the tail of the citation distribution. Section 1.3 showed that this tail has already been widely studied in literature. In previous work, the focus was mostly on showing the existence of a Pareto tail for different fields of science. This report extends the research on the tail behavior by looking at the influence of the Impact Factor or the number of recent citations on the tail behavior.

Furthermore, this report describes the coefficients of the quantile regression by linking the regression with an existing estimator for the Pareto tail quantiles.

1.5 Structure of the report

First, Chapter 2 describes and analyzes the data that were used in this report. Then, Chapter 3 investigates two models from literature for the distribution of the number of citations articles get. We analyze these models and we address their fit to the data. After that, Chapter 4 proposes a model for the prediction of long term impact of recent publications, conditional on the Impact Factor and the number of recent citations. Then, in Chapter 5, we fit the coefficients for this model, and analyze the behavior of these coefficients. Chapter 6 addresses the fit of this model to the data. First, we consider the fit of the model to all data. Then, we consider the fit of the model to data from the same country or the same university. After that, Chapter 7 investigates the sensitivity of the parameters to the scientific field of the publications and to the country from which the publications originate. Chapter 8 is the last chapter of this report, and it contains the conclusions and a discussion.

Chapter 2

Settings for the data analysis

In this chapter, first the data that were available for this report are described. Then, some statistics of these data are shown, to provide some motivation for the model that is introduced in Chapter 4.

2.1 Description of the data

In this report, data from the CWTS are used. This database is based on the Web of Science database that is produced by Thomson Reuters. For all articles in the database, it contains information such as the title of the article, the authors, the publication date, the abstract, the references, the source of the article and so on.

In the Web of Science database, several document types are distinguished: articles, review articles, letters and non citable items. In this report, only articles and review articles are taken into account. The database also distinguishes different scientific fields of articles. Since citation behavior in different fields of science can be different, only articles in the field of Physics are included in the data set. These articles consist of articles with subject categories *Applied Physics, Fluids and Plasma Physics, Atomic, Molecular and Chemical Physics, Multidisciplinary Physics, Condensed Matter Physics, Nuclear Physics, Particles and Fields Physics* and *Mathematical Physics*. All articles in the data set were published in 1984. This set consists of 56,207 articles. For counting citations, self-citations are excluded.

2.2 Recent publications

In the literary review section 1.3, we saw that in literature, a lot of different factors have been used to predict the number of citations that articles will get. As explained in the introduction Section 1.1, in this report two factors are taken into account: the number of recent citations, and the Impact Factor of the journal in which the article was published.

- The number of recent citations is in this report defined as the number of citations an article gets in the first year after publishing, and is denoted by c_1. In this case, this means that the number of citations that the articles have obtained in 1985 are used. Thus, for counting the citations in the first year, all citations that were given to the paper in 1985 are taken into account, but also the number of citations after the publishing date in 1984 are counted.
- The Impact Factor (IF) is an indicator that is based on the journal in which the article appears. The Impact Factor has been used for over 40 years [10]. The Impact Factor of a journal in year x is defined as the average number of citations that papers published in the journal in year $x-1$ and $x-2$ get in year x. For the articles in the data that are used, the Impact Factor of 1984 is computed by using this definition of the Impact Factor.

Now, using the Impact Factor and the number of recent citations, the goal is to predict the number of additional citations that an article has received before the end of 2013. This means that we want to

predict the total number of citations the articles have received up to 2013, minus the number of citations in the first year.

2.3 Influence of Impact Factor and recent citations

Table 2.1 illustrates the relation between the Impact Factor and the number of citations after 30 years. The fraction of articles with a certain Impact Factor that end above the j^{th} quantile of the distribution of the number of citations is computed. Here the Impact Factor is rounded to the nearest integer. We divide this probability by the probability that an article with an Impact factor of 0 will end above the j^{th} quantile.

$j \backslash IF$	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.30	1	2.0	2.5	2.8	2.9	3.0	2.9	3.2	2.7
0.40	1	2.7	3.9	4.4	4.9	5.1	4.9	5.7	4.4
0.50	1	3.4	5.2	6.0	6.8	7.4	7.0	8.5	6.2
0.60	1	4.9	8.5	9.9	11.9	13.7	12.1	16.2	10.7
0.70	1	5.8	10.8	12.8	16.5	19.6	16.6	24.3	15.3
0.80	1	8.6	17.7	20.6	28.5	36.8	28.9	45.9	28.2
0.90	1	9.2	22.2	25.3	38.3	56.2	46.3	66.2	45.3
0.91	1	8.1	20.3	23.6	36.0	52.8	42.7	60.5	42.9
0.92	1	7.5	19.2	21.4	33.8	50.0	39.7	58.6	41.5
0.93	1	6.6	18.2	18.6	31.7	46.9	38.2	55.9	40.2
0.94	1	7.5	20.0	20.4	34.2	53.2	45.0	63.9	46.2
0.95	1	6.7	17.1	17.5	30.1	47.9	39.2	59.5	43.7
0.96	1	7.5	18.4	21.0	36.3	57.6	40.6	73.7	56.8
0.97	1	5.3	13.9	15.5	26.5	43.7	28.4	48.0	44.9
0.98	1	5.4	13.9	14.8	28.7	44.5	32.5	55.1	53.1
0.99	1	4.2	10.0	12.0	22.7	36.0	21.7	44.0	51.0

Table 2.1: Let $p(j|IF)$ be the probability that article with given Impact Factor will end above the j^{th} quantile of the number of additional citations after 30 years. This table presents $p(j|IF)/p(j|0)$.

For example, the table indicates that an article with Impact Factor 8 has a 51 times larger probability of ending in the top 1% of most cited articles than an article with Impact Factor 0. For Impact Factors up to five, the probability of ending in the top quantiles increases. However, for Impact Factors larger than five, the probability of ending in the top quantiles does not increase for all quantiles anymore. This indicates that publishing in a journal with a higher Impact Factor gives a higher probability of being highly cited. However, this effect dampens as the Impact Factor increases. After some point, publishing in a journal with higher Impact Factor does not give a higher probability of being highly cited anymore.

In Table 2.2 the relation between the number of citations after one year, c_1 and the number of additional citations after 30 years is illustrated in a similar way. The results in this table are similar to the results in the previous table. Again, the probability of ending in the top quantiles increases with c_1. However, for c_1 larger than five, this probability does not increase as much anymore.

The tables indicate that the number of citations in the first year and the Impact Factor of the journal in which the article was published have influence on the probability of obtaining certain number of citations. But also for articles with the same Impact Factor and the same number of citations after one year, the number of additional citations these articles get after 30 years is very different. As an illustration of this, the empirical cumulative probability distribution of the number of additional citations after 30 year is plotted in Figure 2.1 for a few different groups of articles. These groups consist of all articles with the same Impact Factor (IF), and the same number of citations in the first year (c_1). For example, 15% of articles published in 1984 in the field of Physics with Impact Factor 1 and one citation in the first year, obtained one or zero extra citations in the next 29 years. In the same group of articles, 1% obtained more than 100 citations. In another group of articles, the group with Impact Factor 2 and six citations in

$j \backslash c_1$	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0.30	1	1.3	1.6	1.6	1.7	1.7	1.7	1.7
0.40	1	1.5	1.9	2.1	2.2	2.2	2.3	2.3
0.50	1	1.7	2.2	2.5	2.7	2.8	2.8	2.9
0.60	1	1.9	2.7	3.3	3.6	3.9	4.1	4.2
0.70	1	2.2	3.4	4.3	5.1	5.6	6.0	6.3
0.80	1	2.4	4.1	5.6	7.0	8.3	9.3	9.7
0.90	1	2.7	5.0	7.6	10.6	13.7	15.9	17.7
0.91	1	2.7	5.2	7.6	11.2	14.6	17.2	18.5
0.92	1	2.9	5.4	8.0	12.2	15.6	18.8	20.0
0.93	1	3.0	5.7	8.4	13.3	17.2	20.0	22.6
0.94	1	2.9	5.7	8.5	13.9	18.1	21.8	24.8
0.95	1	2.8	6.0	8.8	14.0	18.8	23.4	27.5
0.96	1	2.6	6.0	8.4	14.3	19.4	25.3	28.5
0.97	1	2.4	5.6	8.4	14.9	19.5	27.9	31.0
0.98	1	2.3	5.5	8.1	13.6	20.4	27.6	33.1
0.99	1	2.7	4.2	7.2	12.8	21.8	31.1	28.8

Table 2.2: Let $p(j|c_1)$ be the probability that article with given c_1 will end above the j^{th} quantile of the number of additional citations after 30 years. This table presents $p(j|c_1)/p(j|0)$.

the first year, 9% of articles obtained 10 or less additional citations in the next years, while 1% of articles obtained more than 400 additional citations. This illustrates the variety in the number of citations that articles with the same Impact Factor and the same number of recent citations obtain.

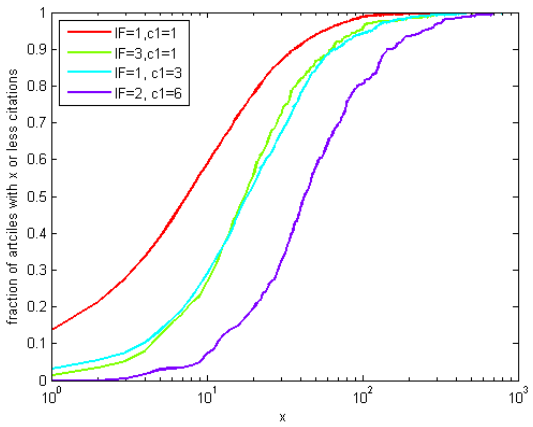

Figure 2.1: Empirical cumulative probability function of the number of additional citations after 30 years for four groups of articles with the same c_1 and IF

2.3.1 Splitting data in groups vs. not splitting

This report investigates the influence of the Impact Factor and the number of recent citations on the number of citations articles get. One option to investigate the influence of these factors on the data, is to use them in a regression function as covariates. We use this approach to predict the quantiles of the number of citations articles receive in Chapter 4.

However, when studying properties of empirical data, this approach is not possible. Since we want to be able to see the influence of the Impact Factor and the number of recent citations on these properties of empirical data, we need another approach in this case. Hence, we split the data into groups. The

groups consist of all articles that have the same Impact Factor and the same number of recent citations. For each of these groups, the data analyses will be done separately, to get an understanding of the effect of the Impact Factor and the number of recent citations.

Hence, in the following chapters, for all empirical results, the data are grouped. The only time the data are not grouped, is when the regression model of Chapter 4 is considered.

2.4 Tail of the citation distribution

In the literary review Section 1.3, it was shown that in literature, a lot of attention has been given to the tail of the citation distribution. This tail is often assumed to have a Pareto distribution. Hence, first we check if the the tail of the distribution of the number of citations can indeed be approximated by a Pareto distribution for our data set. Then we study the influence of the Impact Factor and the number of citations that an articles obtained in the first year on the tail index of the heavy tails.

2.4.1 Pareto behavior

If the tail of the distribution of the number of additional citations X that an article gets has a Pareto distribution , then for some x_l:

$$P(X > x) = zx^{-\alpha}, \qquad x \geq x_l.$$

Here z and α are parameters. The parameter α is also called the tail index. To answer the question of whether X has a Pareto tail, first Pareto QQ-plots (*Quantile-Quantile plots*) are made. These QQ-plots plot the logarithm of the theoretical quantiles of the Pareto distribution against the logarithm of the empirical quantiles of the distribution of the number of citations. This means that we plot

$$\left(-\ln\left(\frac{k}{n+1}\right), \ln\left(X_{(n-k,n)}\right)\right), \qquad k \in \{1, \ldots, n\}.$$

Here $X_{(n-k,n)}$ is the k-th largest number of citations within the data for which the QQ-plot is plotted, and n is the number of articles in the data set. If the QQ-plot for the data becomes a straight line from some point on, this would indicate that tail of the distribution of the data is Pareto. Furthermore, the slope of this line gives an indication of the tail index α of the Pareto distribution.

To make the QQ-plot, the data are grouped. First, all articles with the same number of citations in the first year (c_1) are in the same group. For each of those groups, the Pareto QQ-plot is made, as shown in Figure 2.2. After that, the data is subdivided in groups that have similar Impact Factors. Again, the Pareto QQ-plots are made (Figure 2.3).

After some point, the plots seem to be linear, which supports the assumption that the tails follow a Pareto distribution. Furthermore, the slopes of the linear parts of the QQ-plots are alike for most of the groups. This holds for the groups of data which have been subdivided on the basis the number of citations in the first year, and also for the groups which have been subdivided on the basis of the Impact Factor. The fact that the lines are parallel, indicates that the tail indices α are similar for articles with a different Impact Factor or a different number of citations in the first year.

2.4.2 Zenga plot

In [6], Cirillo states that from QQ-plots, a Pareto distribution can not be concluded. Especially distinguishing between a lognormal and a Pareto tail seems to be difficult when using QQ-plots. For this reason they propose a different plot to see if the tails follow a Pareto distribution: the Zenga plot. In this plot the Zenga curve is plotted. The Zenga curve Z is defined as:

$$\begin{aligned} Z(u) &= 1 - \tfrac{Q^-(u)}{Q^+(u)} && 0 < u < 1, \\ Q^-(u) &= \tfrac{1}{u}\int_0^u F^{-1}(s)ds && 0 \leq u \leq 1, \\ Q^+(u) &= \tfrac{1}{1-u}\int_u^1 F^{-1}(s)ds && 0 \leq u \leq 1. \end{aligned}$$

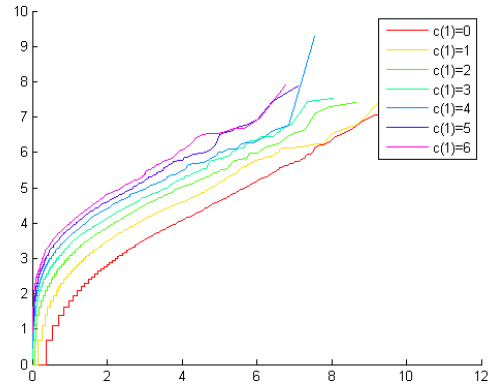

Figure 2.2: Pareto QQ plot for different c_1

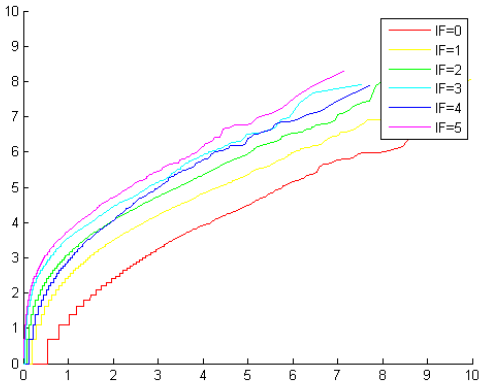

Figure 2.3: Pareto QQ plot for different Impact Factors

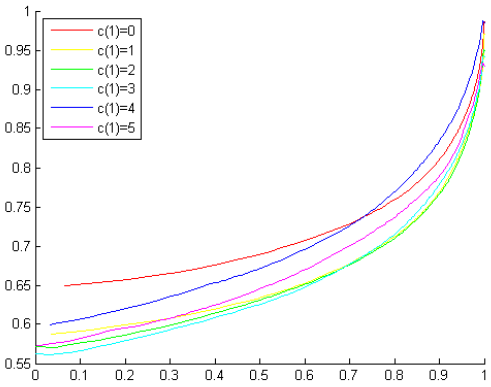

Figure 2.4: Zenga plot for different c_1

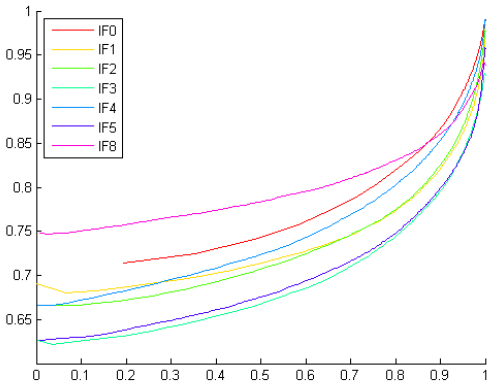

Figure 2.5: Zenga plot for different Impact Factors

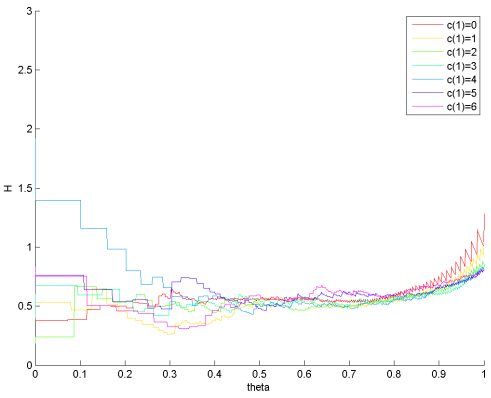

Figure 2.6: altHill plot for different c_1

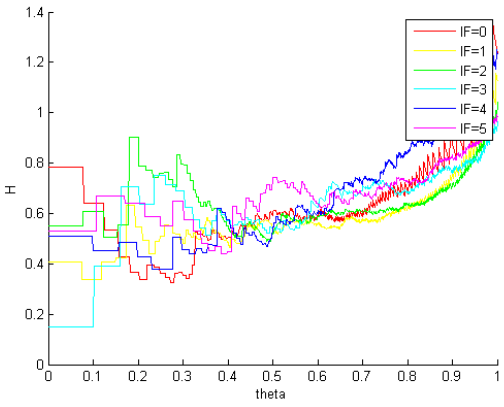

Figure 2.7: altHill plot for different Impact Factors

Here F is the distribution function of the variable X for which the Zenga curve is plotted. This means that the Zenga curve is some sort of measure of how much weight of the distribution lies below the u^{th} quantile corresponding to how much weight lies above the u^{th} quantile. This function is only defined for Pareto distributions if $\mathbb{E}[X] < \infty$, so if $\alpha > 1$. In the next section, we show that this condition holds for the data that are used. The Zenga curve has different shapes for different distributions. For the lognormal distribution, the Zenga curve is a straight line, while for Pareto distributions, the Zenga curve is a convex increasing function [6].

This Zenga plot is plotted for groups of articles with the same number of citations in the first year or the same Impact Factor (Figures 2.4 and 2.5), like the QQ-plots from Section 2.4.1. The lines are clearly convex, increasing functions, which indicates that the distribution of the number of citations has a Pareto tail for all the groups of articles that are considered.

2.4.3 Computing the tail index

Based on the results of Sections 2.4.1 and 2.4.2, we assume that the tails of the citation distribution are Pareto. The QQ-plots from Section 2.4.1 suggested that the Pareto tail indices for groups of articles with different Impact Factors or different numbers of citations in the first year, are similar. In this section, we check if this is correct.

To find an estimate of $\frac{1}{\alpha}$, the Hill estimator, $H_{k,n}$ is used. This Hill estimator is the maximum likelihood estimator of $\frac{1}{\alpha}$, and is defined by Hill as [12]:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \ln\left(\frac{X_{(n-i,n)}}{X_{(n-k-1,n)}}\right).$$

Here $X_{(n-i,n)}$ is the i^{th} largest number of citations in the data group for which the Hill estimator is computed. The Hill estimator estimates the tail index of the Pareto tail, assuming the tail consists of the k most cited articles. However, we do not know where the tail starts, hence we do not know which value of k we should use. The relation between k, and the estimate for $\frac{1}{\alpha}$ can be visualized, in what is called a Hill plot. In traditional Hill plots, the value of the Hill estimator $H_{k,n}$ is plotted against k. Instead of plotting $(k, H_{k,n})$, we plot $(\theta, H_{\lceil n^\theta \rceil,n})$ for $0 \leq \theta < 1$ in Figures 2.6 and 2.7. This means that the x-axis is in logarithmic scale, to focus more on the first part of the Hill plot. In these plots, the Hill estimator varies a lot for different θ. For most groups of articles the Hill estimator is stable between 0.5 and 0.6. However, from these Hill plots it is still not possible to make a statement about which k should be used to compute the value of the tail index α.

To make an estimate of which k to use to compute the Hill estimator, the same method as used in [1] is applied. In this method, an approximation of the asymptotic mean squared error is minimized for k. The value of k that minimized the approximation of the asymptotic mean squared error is used to compute the Hill estimate of the tail index. The results of this are in Table 2.3 and 2.4, where the Hill estimates $\frac{1}{\alpha}$ and the corresponding k that was used are shown for all different groups. The values of the tail index vary for different Impact Factors and different numbers of citations in the first year. However, we cannot find an increasing or decreasing trend in the tail index as the Impact Factor increases, or the number of citations in the first year increases.

IF	k	$\frac{1}{\alpha}$
0	532	0.593
1	942	0.568
2	1082	0.601
3	261	0.684
4	75	0.540
5	208	0.693

Table 2.3: Hill estimates for different Impact Factors

c_1	k	$\frac{1}{\alpha}$
0	1588	0.549
1	636	0.508
2	296	0.496
3	334	0.539
4	216	0.513
5	246	0.586
6	137	0.593

Table 2.4: Hill estimates for different c_1

Chapter 3

The unconditional distribution of the number of citations

This chapter investigates two models from literature for the distribution of the number of citations articles get. We analyze these models and we address their fit to the data.

3.1 Models

The distribution of the number of citations that articles receive has been claimed to be a negative binomial distribution [5] or a discrete variant of the lognormal distribution [24]. In this section, both distributions are studied.

3.1.1 Lognormal distribution

In [24], it is assumed that papers have a parameter $q \sim \mathcal{N}(\mu, \sigma)$, which expresses the quality of the article. The number of citations that an article gets, x, is then uniquely determined by q as follows:

$$x = \lfloor e^q - \xi \rfloor.$$

This means that the number of citations has a discretized lognormal distribution. Here ξ is an extra parameter to include the probability that an article gets zero citations: $\ln(1+\xi)$ is the quality q that is needed to get one citation. All articles with quality $q < \ln(1+\xi)$ remain uncited. Thus, given the parameters μ, σ and ξ, the probability of an article obtaining x citations can be written as:

$$p(x|\mu,\sigma,\xi) = \begin{cases} \int_{-\infty}^{\ln(1+\xi)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(q-\mu)^2}{2\sigma^2}} dq & x = 0 \\ \int_{\ln(\xi+x)}^{\ln(\xi+x+1)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(q-\mu)^2}{2\sigma^2}} dq & x \geq 1 \end{cases} \tag{3.1}$$

We fit the parameters μ, σ and ξ by minimizing the χ^2 statistic:

$$\chi^2 = \sum_x \frac{(p_x - p(x|\mu,\sigma,\xi))^2}{p(x|\mu,\sigma,\xi)}$$

Here p_x is the empirical distribution of the probability that an article will get x citations.

3.1.2 Negative binomial distribution

Another distribution for the number of citations of articles that is mentioned, is the negative binomial distribution [5]. This distribution is derived by assuming that each article has a latent citation rate λ.

This rate λ can be different for each paper. The distribution of the number of citations up to time t that an article with latent rate λ obtains, is then given by a Poisson process with rate $\lambda C(t)$:

$$P(X = x|\lambda C(t)) = e^{-\lambda C(t)} \frac{(\lambda C(t))^x}{x!}.$$

Here $C(t)$ is the obsolescence distribution function. This is a monotonically increasing function in t, with a finite limit. This obsolescence function takes into account that in general, papers are cited at different frequencies over time, and after some time, the paper will not get new citations anymore.

If we then assume that the latent citation rate λ has a gamma distribution with parameters b and ν, it can be proven that the distribution of the number of citations has a negative binomial distribution [4]:

$$X \sim NB\left(\frac{b}{b + C(t)}, \nu\right).$$

Since we are interested in the number of citations after a fixed number of years, $C(t)$ is a constant in our analysis. This means that the probability that an article obtains x citations is then given by:

$$p(x|\nu, a) = \binom{x + \nu - 1}{x} a^x (1 - a)^\nu, \tag{3.2}$$

where $a = \frac{b}{b+C(t)}$.

In the next section, we check the fit of the two models from this section onto the data by minimizing the χ^2-statistic.

3.2 Fitting coefficients of the distribution of the number of citations

In this section (3.1) and (3.2) are fitted onto the data that were described in Chapter 2. This means that the publications that are taken into account are all Physics papers that were published in 1984. The distribution of the number of citations these papers obtained up to the end of 2013 is fitted.

For the lognormal distribution, minimizing the χ^2-statistic, results in $\mu = 2.0, \sigma = 1.5$ and $\xi = 0.03$. If we minimize the χ^2-statistic to fit the parameters for the negative binomial distribution, $\nu = 0.3$ and $a = 0.007$ are obtained. Figure 3.1 visualizes the fit of both the lognormal and the negative binomial model to the empirical data. In this figure, the probability density function for both fitted distributions are shown. The green line corresponds to the empirical cumulative distribution function. The lognormal curve seems to fit very well to the empirical data, the empirical probability density function and the lognormal density function almost collapse. The fit of the negative binomial distribution is obviously worse than the lognormal curve. Table 3.1 gives the values of the χ^2-statistic for both the lognormal and the negative binomial distribution. The χ^2-statistic for the lognormal distribution is a lot smaller than for the negative binomial distribution.

One reason that could explain why the negative binomial distribution does not match the distribution of the data, could be the heavy tail of the distribution of the number of citations. We have seen before that the tail of the distribution of the number of citations is Pareto. The negative binomial distribution however, does not have a heavy tail. For this reason we try to fit the distributions only on the bulk of the distribution of citations. This means that the top 10% of most cited papers are omitted from the data. Now, the negative binomial distribution is fitted again. This results in $\nu = 0.7, a = 0.05$. Again, the probability density functions are shown together with the empirical density function in Figure 3.2. The negative binomial distribution fits the empirical distribution better than in Figure 3.1. The χ^2-statistic for the fit of the negative binomial and the lognormal distributions that are fitted while omitting the most highly cited papers can also be seen in Table 3.1. The fit of the lognormal distribution is still better than the fit of the negative binomial distribution.

If we now look at the cumulative distribution function of these two fitted functions in Figure 3.3, we can see that especially in the tail of the distribution, these functions do not fit very well. The lognormal

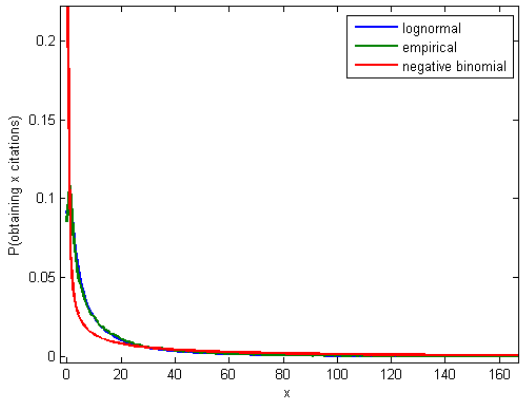

Figure 3.1: Probability density functions of fitted distributions and the empirical density function

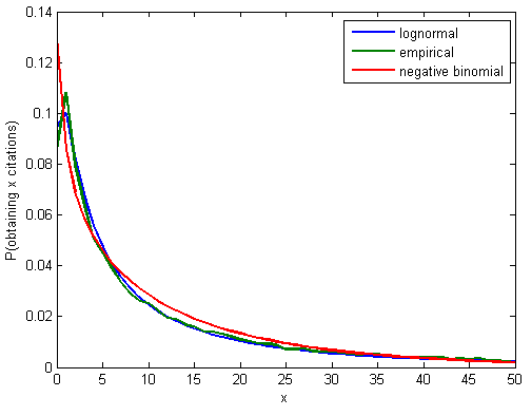

Figure 3.2: Probability density functions of distributions that were fitted omitting the tail data, and empirical density function

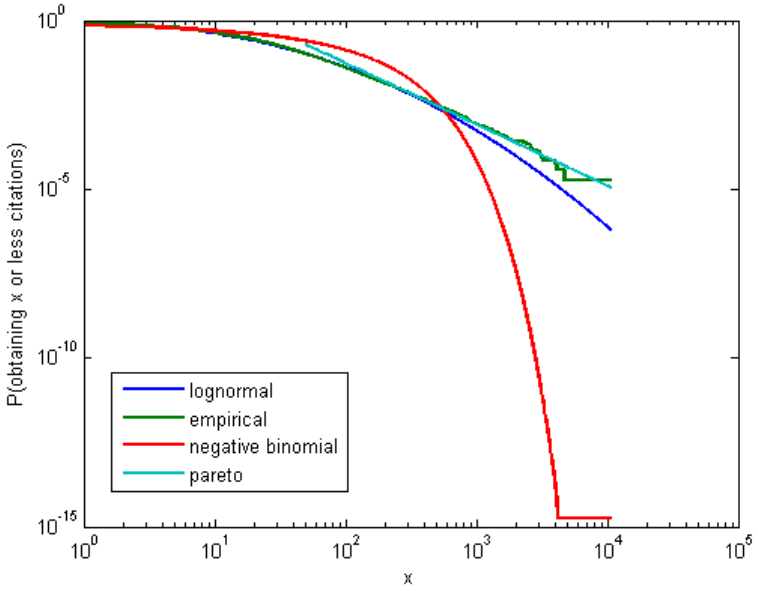

Figure 3.3: Empirical and fitted cumulative distribution functions of the number of citations on logarithmic scale

	All data		No tail	
	Negative binomial	Lognormal	Negative Binomial	Lognormal
χ^2-statistic	4.868	0.0077	0.034	0.0052

Table 3.1: χ^2-statistic for the lognormal and negative binomial distributions, for the fit to all data, and the fit to the data where the tail is omitted

function fits well for articles with 300 or less citations, but after that, the fit is less good. For this reason, a Pareto function is also shown in the tail of the distribution. The tail index of the Pareto distribution is estimated by the Hill estimator that was mentioned in Section 2.4. The fit of the Pareto distribution seems a lot better in the tail part.

In conclusion, the distribution of the number of citations that a paper receives in the long run, seems to have a discretized lognormal distribution for articles with less than 300 citations. For higher citation counts, a Pareto distribution seems to fit better.

Chapter 4

Model for the conditional distribution of the number of citations

In this chapter, we propose a model, which predicts the conditional quantiles of the long term citation distribution, given the Impact Factor and the number of citations in the first year.

4.1 Quantiles

In [22], Seglen shows that the citation distribution is a skew distribution. It is characterized by a few highly cited papers, while the majority of papers obtains less citations. Since the mean number of citations can be largely influenced by a few articles with large citation numbers, we propose to look at the quantiles of the distribution of the number of citations. Formally, the j^{th} quantile $q(j)$ of a random variable Y with distribution function F is given by:

$$q(j) = F^{-1}(j) = \inf\{y : F(y) \geq j\}.$$

Hence, saying that an article scores at the j^{th} quantile, means that an article has more or the same number of citations than a proportion of at least j of the other articles. The article has less or an equal number of citations than a proportion of at most $1 - j$ of articles. In the rest of this section, we propose a model that predicts the quantiles of the citation distribution.

4.2 Fitness factor

Like in [13], we assume that each paper has a fitness factor η. This fitness factor gives information about the competitiveness of an article among other articles in obtaining citations. The fitness factor consists of different factors ϕ, that contribute to the success of an article. Then the fitness factor η is assumed to be a product of all these factors to some power δ:

$$\eta \propto \prod_s \phi_s^{\delta_s}.$$

In this case, we use as factors the Impact Factor of the journal in which the article is published, and the number of citations that an article got after one year:

$$\eta \propto IF^{\delta_1}(c_1 + k_0)^{\delta_2}.$$

For clarity of notation, δ_1 and δ_2 are renamed to β and γ:

$$\eta \propto IF^{\beta}(c_1 + k_0)^{\gamma}.$$

The constant k_0 is needed to address for publications that have zero citations one year after publishing. This means that k_0 expresses the rate at which articles that were uncited in the first year after publishing, obtain citations. As before, IF is the Impact Factor, and c_1 the number of citations an article got after one year.

4.3 Model for quantiles

Now we try to predict the quantiles of the distribution of the number of citations using the fitness factor. A higher fitness factor means that an article has a higher competitiveness to obtain citations. Then one expects that the quantiles of articles with a higher fitness factor are also higher. We assume that the j^{th} quantile of articles with fitness factor η is proportional to η:

$$q(j|\eta) \propto \eta.$$

In this section, three different models are proposed. In the first model, the fitness factor depends on only the Impact Factor, in the second model the fitness factor depends on only the number of recent citations. In the last model the fitness factor depends on both the number of recent citations and the Impact Factor.

4.3.1 Model with only the Impact Factor

In the first model, only the Impact Factor is taken into account for the fitness factor:

$$\eta \propto IF^{\beta}.$$

This would then mean that the conditional j^{th} quantiles for articles with Impact Factor IF, $q(j|IF)$ are distributed in the following way:

$$q(j|IF) = \tilde{C}_j IF^{\beta_j}, \tag{4.1}$$

where $\tilde{C}_j$ is a constant for each quantile. This means:

$$\ln(q(j|IF)) = \beta_j \ln(IF) + C_j,$$

where $\tilde{C}_j = e^{C_j}$.

4.3.2 Model with only the number of recent citations

In the second model, only the number of recent citations is taken into account:

$$\eta \propto (c_1 + k_0)^{\gamma}.$$

Hence, the j^{th} quantile conditional on the number of citations in the first year, $q(j|c_1)$ is given by:

$$q(j|c_1) = \tilde{C}_j \left(c_1 + k_0\right)^{\gamma_j}. \tag{4.2}$$

4.3.3 Model with the Impact Factor and the number of recent citations

In the last variant of the model, both the Impact Factor and the number of recent citations are taken into account when calculating the fitness factor:

$$\eta \propto IF^{\beta}(c_1 + k_0)^{\gamma}.$$

This means that the conditional j^{th} quantiles for articles with Impact Factor IF, and c_1 citations in the first year, $q(j|IF, c_1)$ are distributed in the following way:

$$q(j|IF, c_1) = \tilde{C}_j IF^{\beta_j}(c_1 + k_0)^{\gamma_j}, \tag{4.3}$$

where $\tilde{C}_j$ is a constant for each quantile. This means:

$$\ln\left(q(j|IF, c_1)\right) = \gamma_j \ln(c_1 + k_0) + \beta_j \ln(IF) + C_j, \tag{4.4}$$

where $\tilde{C}_j = e^{C_j}$.

In the next chapter, the coefficients that occur in equations (4.1), (4.2) and (4.3) are fitted.

Chapter 5

The conditional model in practice

In this chapter, first, in Section 5.1, we fit the coefficients that appeared in the model from Section 4.3 on the data using quantile regression. Then we investigate the fit of a well known estimator that uses the Pareto behavior of the data to estimate the quantiles in Section 5.2. After that, these two estimators are compared in Section 5.2.2. In the last part of this chapter, Section 5.3, we look at confidence intervals for the predictions.

5.1 Fitting the coefficients for quantile prediction

In Section 4.3 a model for the prediction of quantiles was proposed. There are three parameters that have to be fitted for each quantile j: C_j, β_j, γ_j. One extra parameter, k_0 also has to be fitted. We fit the first three parameters using quantile regression.

5.1.1 Quantile Regression

Since the aim of the regression is to predict the conditional quantiles of the numbers of citations, it is natural to use quantile regression. In the model as described in Section 4.3, the logarithm of the quantiles is linear in the predictor variables. Because the logarithm is an increasing function, the logarithm of the j^{th} quantile is equal to the j^{th} quantile of the log transformed citation counts. This means that for example for the full model, we can take the logarithm of the number of citations, and then fit equation (4.4) on the quantiles of those values. In this section, this equation is fitted by using quantile regression. Quantile regression is a regression technique that was introduced by Koenker and Basset in [14]. Where in the ordinary least squares method the sum of squared errors is minimized to make sure that the expected value of the errors is zero, in the case of quantile regression, a different function is minimized. Minimizing this function for β, γ and C, gives a model in which the empirical quantiles of the numbers of citations are fitted precisely.

We rewrite (4.4) to:

$$\ln(q(j|IF, c_1)) = X\xi_j.$$

Here X is the matrix with observations. For example, for the full model (4.4), X is given by:

$$X = \begin{bmatrix} 1 & \ln(IF_1) & \ln(c_{11} + k_0) \\ \vdots & \vdots & \vdots \\ 1 & \ln(IF_n) & \ln(c_{1n} + k_0), \end{bmatrix}, \tag{5.1}$$

where n is the total number of articles that is considered. The vector of observations corresponding to the i^{th} observation is called x_i. The vector ξ_j consists of the parameters that have to be fitted for the j^{th} quantile. For the full model for example $\xi_j^{\intercal} = \begin{bmatrix} C_j & \beta_j & \gamma_j \end{bmatrix}$.

The logarithm of the additional number of citations that the article corresponding to observation i obtains in the end, is named y_i. Solving

$$\min_{\xi} \sum_{i} \rho_j(y_i - x_i\xi), \tag{5.2}$$

where

$$\rho_j(z) = zj - z\mathbb{1}_{z<0},$$

gives the solution ξ that makes sure that the empirical quantiles are fitted by the model [14]. This can be seen by taking the directional derivative with respect to ξ. This is illustrated in Appendix A.

The minimization of equation (5.2), gives the parameters C_j, β_j and γ_j, such that the the quantiles of the logarithm of the numbers of citations are fitted. Most existing performance indicators based on the quantiles of the citation distribution are based on the higher quantiles of the citation distribution. For this reason, in this report we focus on the 0.50^{th} up to the 0.99^{th} quantile. Figures 5.1, 5.2 and 5.3 show the parameters obtained by Quantile Regression. The resulting coefficients are shown for the three different variants of the model that were discussed in Section 4.3. In all three models, the coefficient C_j occurred. Figure 5.1 shows the value of C_j for all three models. We see that this coefficient is not very different for the three different variants of the model, the shape is similar. Figure 5.2 shows β_j, the exponent of the Impact Factor, which occurs in the model that only uses the Impact Factor and in the model that uses both the Impact Factor and the number of recent citations. The coefficient β_j is higher for the model that only uses the Impact Factor than for the full model. This indicates that the Impact Factor is more influential for the prediction in the model that uses only the Impact Factor than for the prediction in the full model. Similarly, Figure 5.3 shows γ_j, which occurs in the model that uses only the number of recent citations, and in the model that uses both the number of recent citations and the Impact Factor. The coefficient γ_j is higher for the model that only uses the number of citations in the first year than for the full model. Furthermore, the coefficient is less stable in the model that only uses the number of recent citations to predict the quantiles. This is probably because the number of citations in the first year can only have a few values.

The coefficient C_j is increasing in j for all three variants of the model. This is not very surprising, because the quantiles are also increasing in j. We also see that C_j is convex. For the higher quantiles, C_j grows faster in j than for the lower quantiles. This can be explained by the heavy tail of the distribution of the number of citations. Because of the heavy tail of the distribution of the number of citations, for example the 0.98^{th} and the 0.99^{th} quantile are further away from each other than the 0.60^{th} and the 0.61^{st} quantile. This means that the multiplicative constant in the regression model, C_j, increases more for the higher quantiles than for the lower quantiles.

The coefficients γ_j and β_j are decreasing in j. This indicates that the Impact Factor and the number of citations in the first year have less influence on highly cited papers than on papers with a number of citations that is closer to average.

In Section 2.3, in Table 2.1 we saw that articles in journals with a higher Impact Factor have a larger probability of obtaining a large number of citations than articles that are published in a journal with a lower Impact Factor. However, the influence of the Impact Factor seemed to become less as the Impact Factor became larger. This is also captured in the model, where the coefficient β_j, which is the exponent of the Impact Factor in the model, is smaller than one. The same holds for the number of citations in the first year.

5.1.2 Influence of k_0

One parameter that is not fitted in the regression is k_0. This parameter expresses at which rate articles that remained uncited the first year after publishing get cited. To have an understanding of what the influence of k_0 is on the regression coefficients, for several values of k_0 the quantile regression is done to compute the parameters γ_j, β_j and C_j. These values of the regression coefficients for $k_0 = 0.3$ to 1.5 can be seen in Figure 5.4, 5.5 and 5.6 for the full model. It is visible that k_0 does not have any influence on β_j, the parameter of the Impact Factor. Also, k_0 does not have much influence on the shape of the other regression coefficients γ_j and C_j, they only differ by a constant if k_0 changes. We chose the k_0

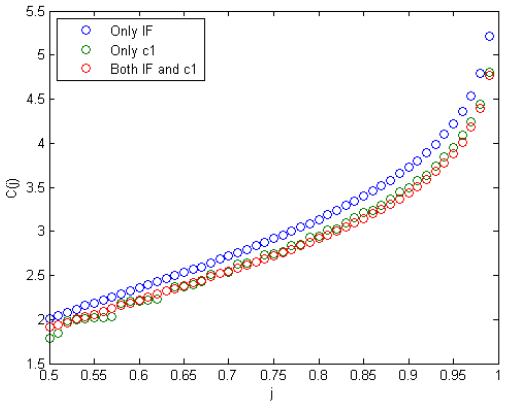

Figure 5.1: Quantile regression coefficient C_j for j^{th} quantile, for three different variants of the model

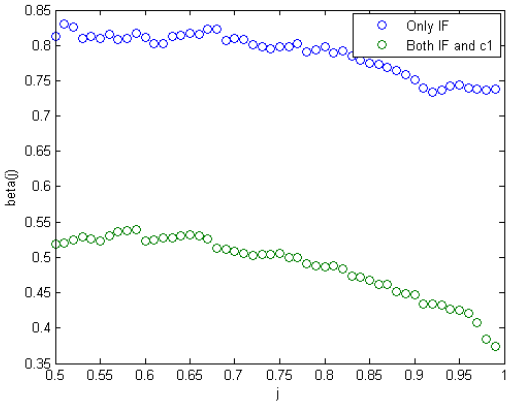

Figure 5.2: Quantile regression coefficients β_j for j^{th} quantile, for two different variants of the model

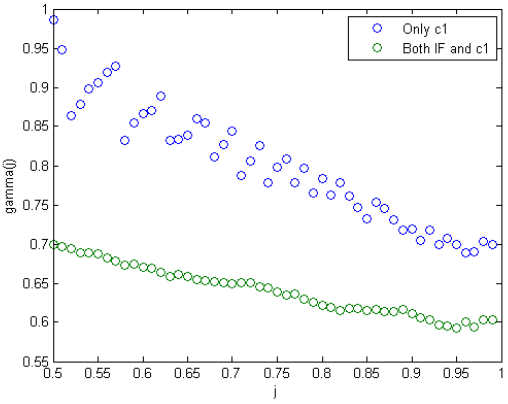

Figure 5.3: Quantile regression coefficient γ_j for j^{th} quantile, for two different variants of the model

which minimized the sum of the squared difference between the fraction of articles with less citations than the predicted j^{th} quantile and j. This resulted in $k_0 = 0.5$.

5.1.3 Fit of model

In this section, the fit of the model is illustrated for all three variants of the model that were introduced in Section 4.3.

Fit for model using only the Impact Factor

The fit of the model that only uses the Impact Factor as described in Section 4.3.1, is illustrated in Figure 5.7. For every value of the Impact Factor, the predicted 0.50^{th}, 0.80^{th} and 0.95^{th} quantiles are shown as solid lines, and the empirical quantiles are shown as dots. The empirical quantiles are computed by splitting the articles in groups, where articles with the same Impact Factor are in the same group, as described in Section 2.3.1. Then, for each group, the quantiles are computed separately. So one dot in the figure corresponds to the empirical percentile of a group of articles with the same Impact Factor. It is clear that the predicted and empirical quantiles may differ a lot, especially for the 0.95^{th} quantile.

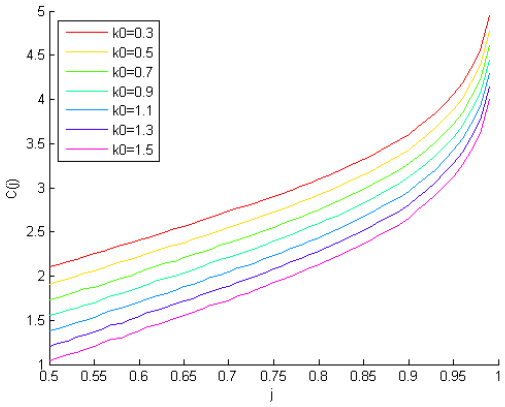

Figure 5.4: Regression coefficient C_j plotted against j for k_0 from 0.3 to 1.5

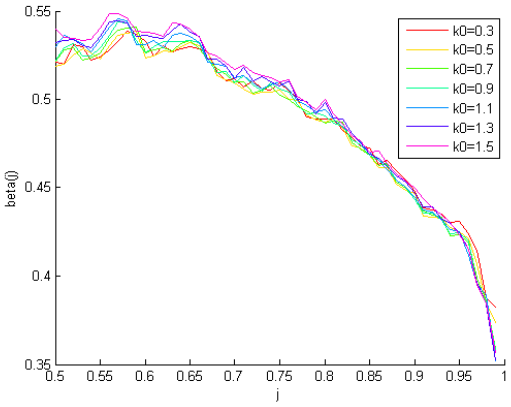

Figure 5.5: Regression coefficient β_j plotted against j for k_0 from 0.3 to 1.5

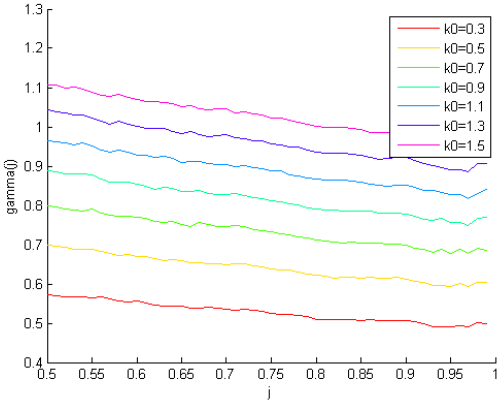

Figure 5.6: Regression coefficient γ_j plotted against j for k_0 from 0.3 to 1.5

Fit for model using only the number of recent citations

The fit of the model that only uses the number of recent citations as described in Section 4.3.2, is illustrated in Figure 5.8. For every value of the number of citations in the first year, the predicted 0.50^{th}, 0.80^{th} and 0.95^{th} quantiles are shown as solid lines, and the empirical quantiles are shown as dots. Again, these empirical quantiles are computed by splitting the articles in groups as described in Section 2.3.1, and then computing the quantiles for each group. For the 0.50^{th} quantile, the empirical and the predicted values almost overlap. For the 0.80^{th} and the 0.95^{th} quantile, the model predicts well for low citation numbers, but underestimates the quantiles for the articles with a high number of citations in the first year.

Fit for full model

To illustrate the fit of the full model as described in Section 4.3.3, we plot the predicted value against the empirical value of the quantiles. As before, to compute the empirical quantiles, the data are splitted in groups of articles as described in Section 2.3.1. The number of articles in the groups is different per group. In Figures 5.9, 5.10 and 5.11, the quantiles for all groups that contain more than 50 articles are shown. These figures show respectively the $0.50^{\text{th}}, 0.80^{\text{th}}$ and 0.95^{th} quantile. Each dot in the figure corresponds to a group of articles with the same Impact Factor and the same number of citations in the first year. The line $y = x$ is also shown as a guideline. The difference between the predicted quantile and the empirical quantile is large for a lot of groups. In Figure 5.12, again the predicted quantiles are plotted against the empirical quantiles. In this plot, only groups containing 500 or more articles are shown. The $0.50^{\text{th}}, 0.80^{\text{th}}$ and 0.95^{th} quantiles are shown in the same plot. This figure shows that for these large groups of articles, the model predicts the quantiles to be close to the empirical value.

In Section 4.3, we made the assumption that the conditional quantiles were linear in the fitness factor η. Figure 5.12 supports this assumption, since the predicted and the empirical quantiles seem to be linearly related.

Comparing the fit of the three variants of the model

To compare the fit of the three different variants of the model, we use the goodness of fit criterion R^1, that was introduced by Koenker and Machado [15]. This goodness of fit criterion is analogous to the criterion R^2 which is often used in least squares regression:

$$R^1(j) = 1 - \frac{\hat{V}(j)}{\bar{V}(j)}.$$

Here $\hat{V}(j)$ is the value of equation (5.2), for the model for which the goodness of fit criterion is computed. $\bar{V}(j)$ is the value of equation (5.2) for the model that includes only the constant term to predict the quantiles:

$$\begin{aligned} \hat{V}(j) &= \min_{\xi \in \mathbb{R}^3} \sum_i \rho_j(y_i - x_i \xi) \\ \bar{V}(j) &= \min_{q \in \mathbb{R}^1} \sum_i \rho_j(y_i - q) \end{aligned}$$

Hence, $R^1(j)$ is a criterion for the goodness of fit of the quantile regression at one specific quantile. Just like the goodness of fit criterion for least squares regression, R^2, it compares the fit of the full model with a model that contains only a constant term. Table 5.1 shows the value of $R^1(j)$ for the three variants of the model.

For the full model, the model which uses both the Impact Factor and the number of recent citations, the value of $R^1(j)$ is higher than for the other models for all quantiles. However, by the definition of $R^1(j)$, it always increases if more variables are added. For the model that uses only the number of recent citations, the fit of the model improves for higher quantiles. For the full model and the model which only uses the Impact Factor on the contrary, the goodness of fit decreases for higher quantiles.

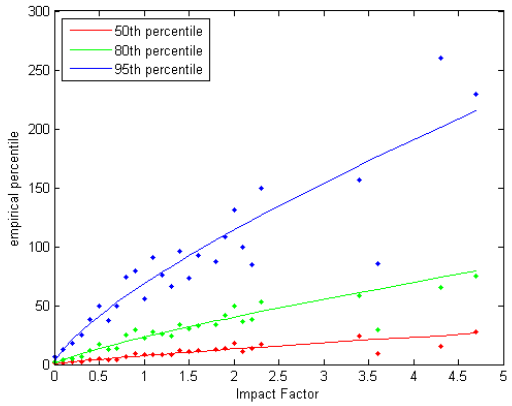

Figure 5.7: Predicted value (solid line), and empirical value (dots) of 0.50^{th}, 0.80^{th} and 0.95^{th} quantile against IF for the model that uses only the IF

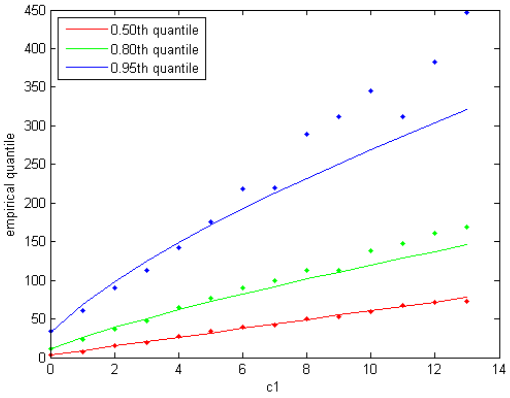

Figure 5.8: Predicted value (solid line), and empirical value (dots) of 0.50^{th}, 0.80^{th} and 0.95^{th} quantile against c_1 for the model that only uses c_1

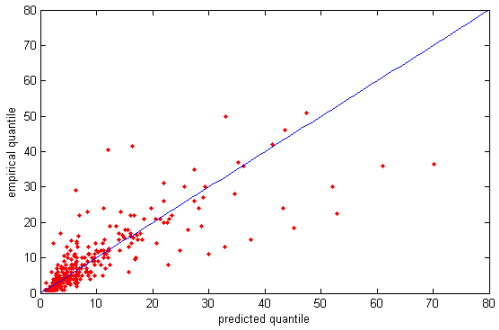

Figure 5.9: Predicted against empirical 0.50^{th} quantile for full model, groups with at least 50 articles

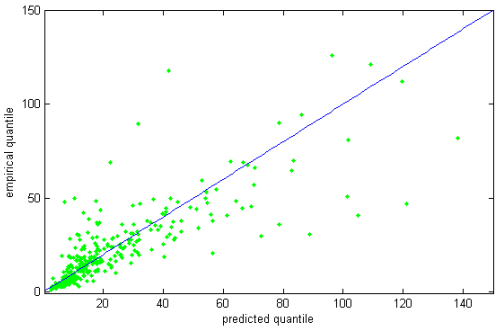

Figure 5.10: Predicted against empirical 0.80^{th} quantile for full model, groups with at least 50 articles

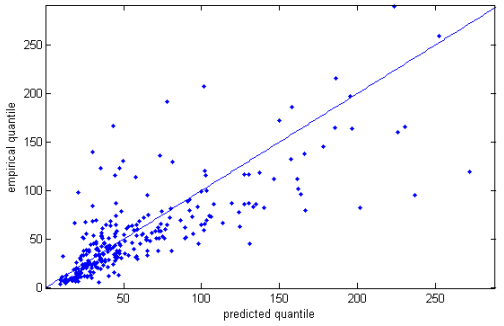

Figure 5.11: Predicted against empirical 0.95^{th} quantile for full model, groups with at least 50 articles

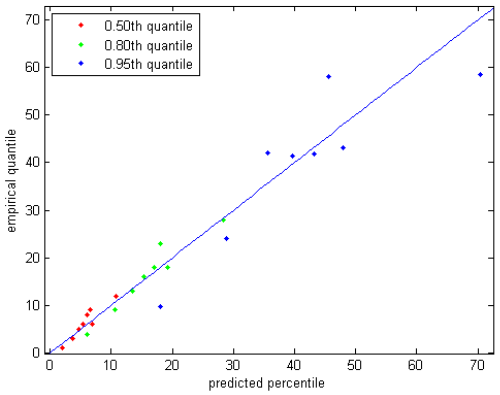

Figure 5.12: Predicted against empirical j^{th} quantile for full model with $j = 0.50, 0, 80, 0.95$ and for groups with at least 500 articles

j	Full model	Only c_1	Only IF
0.50	0.183	0.141	0.111
0.55	0.185	0.143	0.111
0.60	0.186	0.144	0.111
0.65	0.186	0.144	0.109
0.70	0.186	0.144	0.107
0.75	0.185	0.145	0.104
0.80	0.184	0.146	0.101
0.85	0.181	0.145	0.096
0.90	0.177	0.145	0.091
0.95	0.173	0.146	0.086
0.96	0.172	0.147	0.084
0.97	0.172	0.148	0.083
0.98	0.171	0.149	0.082
0.99	0.167	0.147	0.080

Table 5.1: R_1 for the j^{th} quantile for the three different variants of the model

Figure 5.13 illustrates the difference of the fit of the three models further. In this figure, the fraction of articles with less citations than the predicted 0.50^{th} quantile minus 0.50 is plotted. If the model predicts correctly, we expect this value to be close to zero. The colors in the figure indicate this value. A green color means a value close to zero. Red values indicate a higher difference, indicating that the model overestimates the quantiles of this group. Blue values mean negative values, so an underestimation of the quantiles. One rectangle in the figure represents one group of articles with the same Impact Factor and the same number of citations in the first year. For example, for the model that only uses the Impact Factor, the model overestimates the quantiles of articles with an Impact Factor of 2.5, and zero citations in the first year. It underestimates the quantiles of articles with an Impact Factor of 2.5 and 5 citations in the first year.

From this figure, we see that, unsurprisingly, the model that only uses the Impact Factor does not predict well for different numbers of citations in the first year. Similarly, the model that only uses the number of recent citations, does not capture the difference of the 0.50^{th} quantile among articles that appear in journals with different Impact Factors. For other quantiles, similar figures can be obtained. From this figure we can conclude that the full model predicts the quantiles more accurately than the other two models.

Because the full model seems to fit better than the other two models, from now on, the results of the analyses are only given for the full model.

5.2 Pareto quantiles

As said in Section 2.4, the tail of the distribution of the number of citations seems be Pareto, where the tail index α is independent of the Impact Factor and the previous number of citations. Thus for high numbers x, the probability that an article gets more than x citations is:

$$P(X > x) = z(IF, c_1)x^{-\alpha}.$$

Here z is the constant of the Pareto distribution, which is dependent on the Impact Factor and the number of recent citations. In Section 4.3.3, a regression model for the quantiles was proposed. In the next section, we use the Pareto distribution to propose a different estimator for the high quantiles.

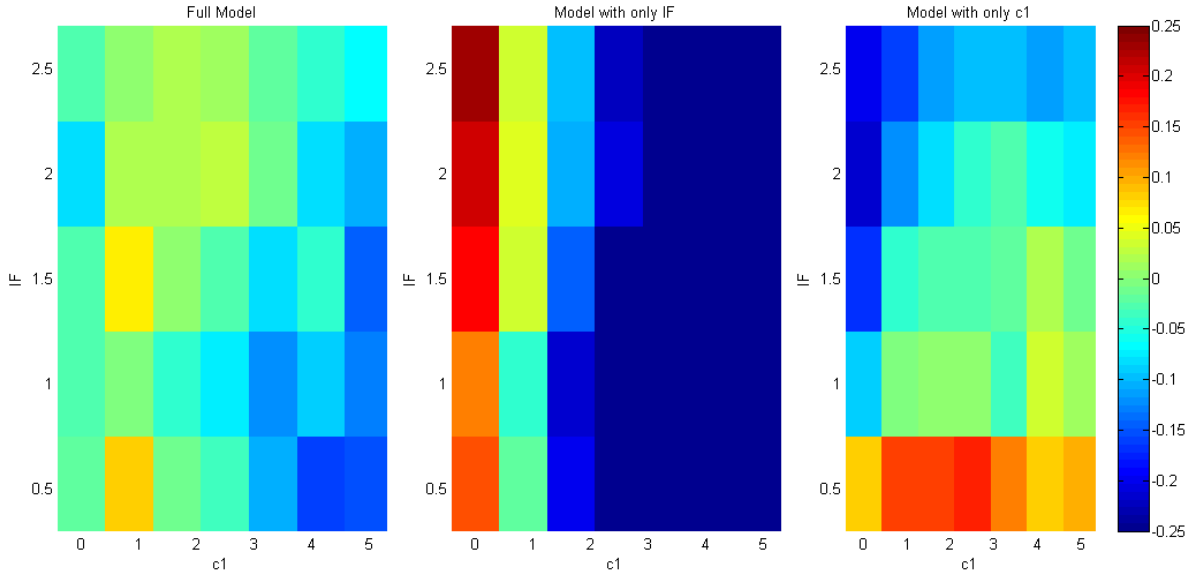

Figure 5.13: Let f be the fraction of articles with less citations than the expected 0.50^{th} quantile. The figure presents $f - 0.5$ for the three models for different c_1 and IF

5.2.1 Pareto quantile estimator

An estimator for the high quantiles of a Pareto distribution is the estimator proposed by Dekkers et al. [9]:

$$q(j|IF, c_1) = X_{(n-k,n)} \left(\frac{k}{n(1-j)} \right)^{1/\alpha}. \tag{5.3}$$

This estimator estimates the high quantiles of a set of observed values X. In this case, X are observations of the number of citations articles have received. Here n is the number of observations in the data set, and $X_{(n-k,n)}$ is the k^{th} largest observation in this set. In this case this means that it is the k^{th} largest number of citations of a group of articles with similar Impact Factors and citations in the first year after publishing. Here k is the threshold where the Pareto tail starts, so only the k articles with the largest numbers of citations follow a Pareto distribution. This means that the Pareto tail starts at the $1 - \frac{k}{n}$th quantile. This threshold quantile will be called j^*. This means that the value of $X_{(n-k,n)}$ is the empirical value of the threshold quantile j^*. The value of the tail index α that appears in the estimator, is estimated by the Hill estimator.

Since $X_{(n-k,n)}$ relies on empirical data, and it depends on the Impact Factor and the number of recent citations, the data are splitted in groups as described in Section 2.3.1. For each of these groups, $X_{(n-k,n)}$ is computed separately to predict the quantiles conditional on the Impact Factor and the number of citations in the first year.

This estimator can be used to estimate the high quantiles of the citation distribution.

5.2.2 Results of Pareto estimator

Figure 5.14 plots the quantiles that are predicted by (5.3) against the empirical quantiles. As said before, the Pareto estimator can only be computed for groups of articles, because the empirical quantile where the tail starts is needed. All groups with 50 or more articles are taken into account when comparing the estimate and the empirical value of the quantiles. By comparing this figure with Figure 5.11, we see that the predicted quantiles are predicted more accurately than the quantiles that were predicted by the regression model. This occurs, because the Pareto estimator uses the empirical quantile of the start of the tail.

Now we have two estimators for the quantiles of the tail of the distribution of the number of citations. The first is the regression estimator from Section 4.3.3, of which the coefficients were fitted by quantile regression. The second is the estimator which uses the Pareto distribution of the tail. In the rest of this section, we study the difference between those estimators.

To compare the estimators, we look at the relative difference d between the estimates they give:

$$d(j) = \frac{\hat{p}(j)_{regression} - \hat{p}(j)_{Pareto}}{\hat{p}(j)_{regression}}$$

In Figure 5.15, a box plot is plotted of the relative difference between the two estimators for the high quantiles. All groups of articles with at least 50 articles in it, are taken into account. The red lines represent the median relative difference between the two estimators. The edges of the boxes are the 0.25^{th} and the 0.75^{th} quantiles of this relative difference. The red symbols represent outliers. These are values of the relative difference that are more than 1.5 times the distance between the 0.50^{th} and the 0.75^{th} quantile away from the 0.75^{th} quantile, or more than 1.5 times the distance between the 0.50^{th} and 0.25^{th} quantile below the 0.25^{th} quantile.

There are some outliers at the top of the box plot, indicating that there are some data for which the estimate using the Pareto estimator is much smaller than the one by using the regression estimator. However, zero is included in the boxes for all high quantiles. This means that the values of these two estimators for the same quantile are often close to each other.

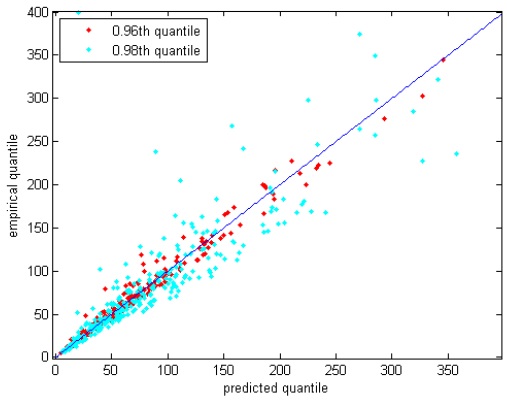

Figure 5.14: Predicted j^{th} quantile from Pareto estimator against empirical j^{th} quantile for $j = 0.96, 0.98$ for groups with at least 50 articles

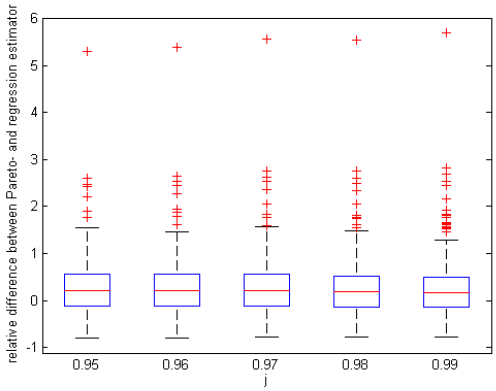

Figure 5.15: Box plot of relative difference between Pareto estimate and regression estimate for groups with at least 50 articles against high j^{th} quantiles

5.2.3 Link Pareto and regression estimates

In this section, we investigate which conditions would need to hold if the Pareto estimator and the regression estimator would give similar results. The Pareto estimator for the quantiles that was mentioned in the previous section is:

$$q(j|IF, c_1) = X_{(n-k,n)} \left(\frac{k}{n(1-j)} \right)^{1/\alpha}.$$

This estimator uses $X_{(n-k,n)}$, the empirical value of the $j^{*\text{th}}$ quantile at which the Pareto tail starts. However, if we assume that the regression model holds for the lower quantiles, we could use the predicted $j^{*\text{th}}$ quantile from the regression model instead of $X_{(n-k,n)}$. The predicted $j^{*\text{th}}$ quantile can be computed from equation (4.3), by filling in j^*:

$$q(j^*|IF, c_1) = \tilde{C}_{j^*} IF^{\beta_{j^*}} (c_1 + k_0)^{\gamma_{j^*}}.$$

By replacing $X_{(n-k,n)}$ by this estimator of the threshold quantile, we obtain the following estimator for the tail quantiles:

$$q(j|IF, c_1) = \left(\frac{1-j^*}{1-j}\right)^{\frac{1}{\alpha}} \tilde{C}_{j^*} IF^{\beta_{j^*}} (c_1 + k_0)^{\gamma_{j^*}}, \qquad j \geq j^*. \tag{5.4}$$

Here the identity $\frac{k}{n} = 1 - j^*$ is used.

In Figure 5.16 (5.4) is compared in a box plot with the regression estimator. It can be seen that these estimators give similar estimates. Note that the 0.95^{th} quantile of both estimators is equal by construction of the estimators.

This leads to the idea of comparing equation (5.4) to the regression estimator. If indeed the two estimators would be equal for the high quantiles, we would have for $j \geq j^*$:

$$\left(\frac{1-j^*}{1-j}\right)^{\frac{1}{\alpha}} \tilde{C}_{j^*} IF^{\beta_{j^*}} (c_1 + k_0)^{\gamma_{j^*}} = \tilde{C}_j IF^{\beta_j} (c_1 + k_0)^{\gamma_j}.$$

This means that for $j \geq j^*$:

$$\begin{aligned} \left(\frac{1-j^*}{1-j}\right)^{\frac{1}{\alpha}} \tilde{C}_{j^*} &= \tilde{C}_j, \\ \beta_{j^*} &= \beta_j, \\ \gamma_{j^*} &= \gamma_j. \end{aligned}$$

Hence, if the two estimators would be equal, we would expect that for the high quantiles, the regression coefficients β_j and γ_j are stable in j. Figure 5.2 gives β_j for the full model. This figure shows that β_j is not stable for the high quantiles, it is decreasing. In Figure 5.3, we see that γ_j is stable for the higher quantiles.

By taking the logarithm of the equality for the regression coefficient C_j, it can be rewritten to:

$$C_j = C_{j^*} + \frac{1}{\alpha} \left(\ln(1-j^*) - \ln(1-j)\right), \qquad j \geq j^*. \tag{5.5}$$

Here, as before, $\tilde{C}_j = e^{C_j}$. Equation (5.5) indicates that the regression coefficients C_j for the high quantiles should be a combination of the logarithm of $1-j$, and a constant. This constant can be written in terms of the regression coefficient C_{j^*} of the quantile j^* where the tail starts, and α. We check if indeed this equality holds for the coefficients C_j from the regression. In Figure 5.17 both sides of equation (5.5) are plotted. The blue line corresponds to the right-hand side of the equation, and the red dots to the left-hand side (the coefficients C_j).

For the high quantiles, the blue line and the red dots overlap. This means that equation (5.5) describes the coefficients C_j well for the high quantiles. Hence, now we have an explicit formula for the value of the regression coefficient C_j for the high quantiles.

5.3 Confidence intervals for predictions

In this section, confidence intervals are constructed for the regression estimator that was proposed in Section 4.3.3 and the Pareto estimator that was introduced in Section 5.2.1.

5.3.1 Constructing confidence intervals for the quantile regression estimates

In [15], Koenker and Machado derive the asymptotic distribution of the error of the estimated coefficients $\hat{\xi}_j$ that follow from the quantile regression. As before, $\hat{\xi}_j^{\mathsf{T}} = \begin{bmatrix} C_j & \beta_j & \gamma_j \end{bmatrix}$. They make the assumptions that the errors are independent (not necessarily identically distributed), and:

- $D_n := \frac{1}{n} X'X \to D$ as $n \to \infty$, where D positive definite,
- $G_n := \frac{1}{n} X'H^{-1}X \to G$ as $n \to \infty$, where G positive definite,

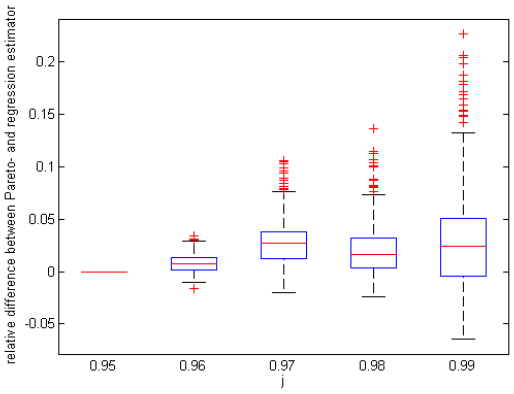

Figure 5.16: Box plot of relative difference between estimator of equation (5.4) and regression estimator for groups with at least 50 articles against high j^{th} quantiles

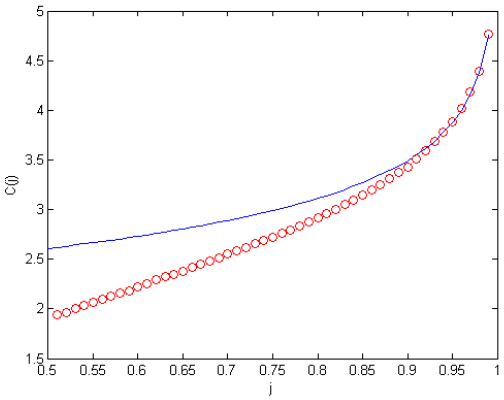

Figure 5.17: C_j from equation (5.5) (line) and C_j from regression (dots) for j^{th} quantiles

- $\max_{i=1,\ldots,n} ||x_i|| = O\left(n^{1/4}/\log(n)\right)$,
- $n^{-1} \sum ||x_i||^4 = O(1)$.

The matrix X is the matrix with the observed values of the covariates, as in (5.1). The vector x_i is again the row of the matrix X corresponding to the i^{th} observation, and n is the number of observations. The matrix H is an n by n matrix, with the error densities of the observations on the diagonal. We use a method proposed in [15], which is based on a difference quotient, to estimate these error densities. Using these assumptions, they show that the following result holds as $n \to \infty$:

$$\sqrt{n}\left(\hat{\xi}_j - \xi_j\right) \xrightarrow{d} \mathcal{N}\left(0, V_j\right),$$

where

$$V_j = j(1-j)s(j)\left(G_n^{-1} D_n G_n^{-1}\right)^{1/2}.$$

Here $s(j)$ is the sparsity function at the j^{th} quantile:

$$s(j) = \left(f(F^{-1}(j))\right)^{-1},$$

where f and F denote respectively the density function and the cumulative distribution function of the observations. This sparsity function takes into account that the uncertainty of the estimation of the quantiles is larger if the observations around the quantile we want to predict are scarce. We use a method proposed in [15] for estimating this quantity. This method uses a difference quotient of the empirical quantile function to estimate the sparsity function.

Because the j^{th} quantile of articles with covariates x_i is predicted by $\widehat{q}(j)_i = x_i\hat{\xi}$, the distribution of the difference between the predicted quantiles and the real quantiles for articles with covariates x_i satisfies:

$$\sqrt{n}\left(\widehat{q}(j)_i - q(j)_i\right) \xrightarrow{d} \mathcal{N}\left(0, x_i V_j x_i'\right)$$

This can be used to construct asymptotic confidence intervals for the prediction of the quantiles of the citation distribution.

5.3.2 Computing confidence intervals for the quantile regression estimates

Figure 5.18 shows an example of the confidence intervals of the quantile regression estimates for the quantiles . These confidence intervals are for the predictions of the full model, so the model that uses both the Impact Factor and the number of recent citations. The 95%-confidence intervals are the dashed

lines. The solid lines are the observed values of the quantiles that followed from the data. In the figure, the confidence intervals and the observed values of the quantiles are shown for three different groups of articles. The confidence interval gets larger as the quantile gets larger. This happens because at the higher quantiles, the observations are less close to each other than at the lower quantiles. As a result of this, the sparsity function $s(j)$ is larger at the higher quantiles, and the confidence interval gets larger. In Figure 5.19, the same confidence intervals are shown, but now zoomed in at the 0.50^{th} to 0.80^{th} quantile.

For 26 groups of articles with the same Impact Factor and the same number of citations in the first year that contained more than 500 articles, these confidence intervals were computed. For 6 out of the 26 groups of articles, the empirical j^{th} quantile was not contained in the confidence interval for some j.

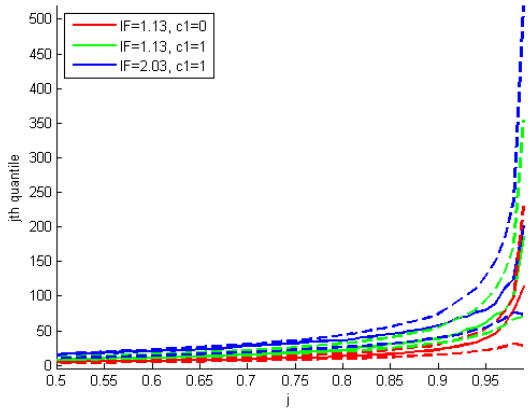

Figure 5.18: Confidence interval (dashed lines) for quantile regression estimates and empirical value (solid lines) of quantiles for different groups of articles

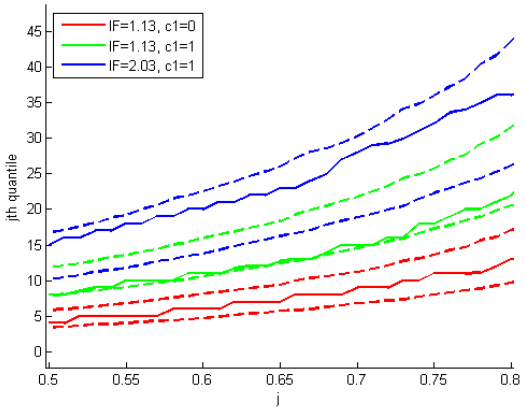

Figure 5.19: Confidence interval (dashed lines) for quantile regression estimates and empirical value (solid lines) of quantiles for different groups of articles, zoomed in

5.3.3 Constructing confidence intervals for the Pareto tail

Besides the regression estimator, in Section 5.2 an estimator for the tail quantiles was proposed. Using results of Dekker et al [9], a confidence interval for this estimator can be constructed.

The value of the estimate of the j^{th} quantile that follows from the Pareto estimator, is denoted by $\widehat{q}(j)$. The real value of that quantile is denoted by $q(j)$. To make a confidence interval for the Pareto estimator, we need to know the distribution of $\widehat{q}(j) - q(j)$. In [9] a result for the asymptotic distribution of this difference is proven for high quantiles. It is proven that:

If $1 - j = 1 - j_n \to 0$, $n(1 - j_n) \to c \in (0, \infty)$ *as* $n \to \infty$, *then for* $k > c$*:*

$$\frac{\widehat{q}(j) - q(j)}{X_{(n-k,n)} H_{k,n}} \xrightarrow{d} \left(\left(\frac{k}{c} \right)^{\frac{1}{\alpha}} - 1 \right) \alpha + \frac{\alpha k \left(1 - \left(\frac{1}{c} Q_k \right)^{\frac{1}{\alpha}} \right)}{Q_k} := Z$$

as $n \to \infty$, *where* Q_k *is gamma distributed with* k *degrees of freedom.*

Here α is the tail index, $H_{k,n}$ is the Hill estimator that was mentioned in Section 2.4 for a group with n observations, using threshold k, $X_{(n-k,n)}$ is the k^{th} largest number of citations within a group of articles with the same Impact Factor and c_1. The result states that the difference between the real value and the predicted value of the quantile converges to a constant, plus the fraction of two gamma-distributed variables.

The result is proven for high quantiles j_n. Here n is the number of observations within a group. If we want to construct a confidence interval for the j^{th} quantile, for some group with n^* articles in it, we have to choose a sequence j_n, such that the conditions of the result hold, and $j_{n^*} = j$. One option is to

choose $1-j_n = (1-j)\frac{n^*}{n}$. Now $1-j_n \to 0$ and $n(1-j_n) \to (1-j)n^*$ as $n \to \infty$. Hence the conditions for the result hold with $c = (1-j)n^*$.

A θ-confidence interval for $q(j)$ can be constructed using this result. We look for the values z_1 and z_2 such that $P(Z < z_1) = 1-\theta/2, P(Z > z_2) = \theta/2$. The value of $P(Z < z_1)$ can be found by using the quotient rule:

$$\begin{aligned} P(Z < z_1) &= P\left(\frac{1-\left(\frac{1}{c}Q_k\right)^{\frac{1}{\alpha}}}{Q_k} \leq \frac{z_1 - \left(\left(\frac{k}{c}\right)^{\frac{1}{\alpha}} - 1\right)\alpha}{\alpha k}\right) \\ &= \int_0^\infty P(Q_k = x) P\left(1 - \left(\frac{1}{c}Q_k\right)^{\frac{1}{\alpha}} \leq xv\right) dx \\ &= \int_0^{xv} P(Q_k = x) P\left(Q_k \geq c(1-xv)^\alpha\right) dx, \end{aligned}$$

where

$$v = \frac{z_1 - \left(\left(\frac{k}{c}\right)^{\frac{1}{\alpha}} - 1\right)\alpha}{\alpha k}.$$

Now a θ%-asymptotic confidence interval for the quantiles of the tails is given by:

$$\left[\widehat{p}(j) + z_1 X_{(n-k,n)} H_{k,n}, \widehat{p}(j) + z_2 X_{(n-k,n)} H_{k,n}\right]. \tag{5.6}$$

5.3.4 Computing confidence intervals for Pareto estimator

To compute the 95%-confidence intervals for the Pareto tail estimator from equation (5.6), an integral has to be computed to find z_1 and z_2. This integral is computed numerically using MATLAB. The value of z_1 and z_2 are then also found numerically. Examples of the confidence intervals that follow from the Pareto estimator can be seen in Figure 5.20 for the same three groups as used in Figure 5.18. Again, the dashed lines correspond to confidence intervals, and the solid lines are the empirical values of the quantiles. Since the Pareto estimator can only be used to estimate the tail quantiles, the confidence intervals are only shown for the 0.95^{th} up to the 0.99^{th} quantile. To compare these confidence intervals with the confidence intervals for the quantile regression estimates, in Figure 5.21, again the quantile regression confidence intervals are plotted. In this figure, the x-axis also runs from the 0.95^{th} up to the 0.99^{th} quantile.

It is clear that for the highest quantiles, the 0.99^{th} and the 0.98^{th} quantile, the quantile regression estimates have larger confidence intervals than the Pareto estimates.

These confidence intervals for the Pareto estimator were computed for 26 groups of articles. For one group of articles, the empirical value of the high quantiles did not lie in the confidence interval that was computed. For all other 25 groups, the empirical value of the quantiles was inside the confidence interval.

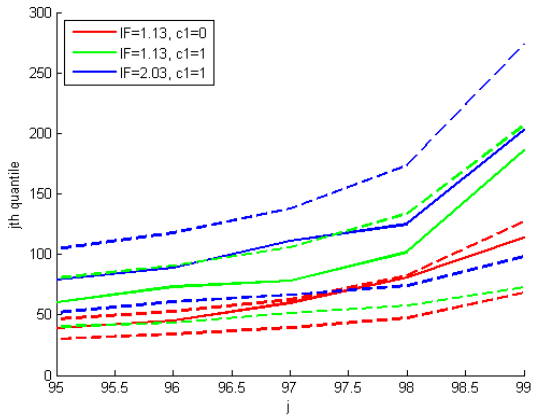

Figure 5.20: Confidence interval (dashed lines) for Pareto estimator and empirical value (solid lines) of high quantiles for different groups of articles

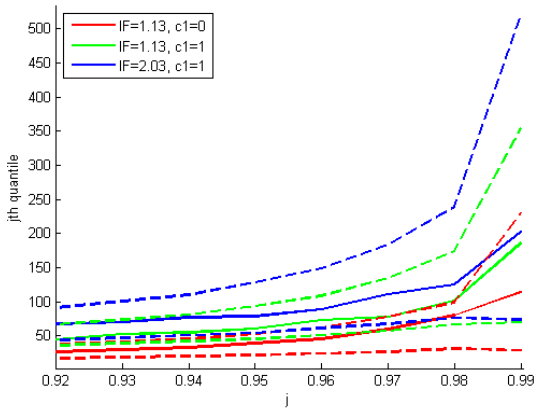

Figure 5.21: Confidence interval (dashed lines) for quantile regression estimates and empirical value (solid lines) of quantiles for different groups of articles, for high quantiles

Chapter 6

Performance of the predictions

In the introduction Section 1.1, we mentioned that it is difficult to know the performance of a set of recent papers from a certain university or country. In this chapter we investigate whether the proposed model can help to solve this problem. We use the model that uses both the Impact Factor and the number of recent citations in this chapter. First, the question of how this model performs for articles that were published in later time periods is addressed in Section 6.1. Then, we check if the proposed model also works well on groups of articles from the same country, or from the same university, respectively in Sections 6.2 and 6.3. After that, in Section 6.4, we use an example to explain the predictions at country or university level.

6.1 Predictions with quantile regression model

In the previous chapters, the parameters of the model were fitted on articles that were published in 1984. In this section, we test if the model is also applicable to data from other years. To do this, first the coefficients of the model are fitted on articles that were published in the Physics category in 1990. In this case, the model is fitted to predict the number of additional citations these articles received up to 2000. Then we use the parameters that follow from this fit to predict the quantiles of all publications in the Physics category that were published in 2000. For these articles, we want to predict the quantiles of the number of citations they obtained until 2010.

The predictive value of the model is illustrated in Figure 6.1. For all articles published in 2000, their predicted j^{th} quantile is computed. Then the fraction of articles that in the end had less citations than their predicted j^{th} quantile is computed. If the model predicts well, this fraction should be exactly j. In Figure 6.1 for both the data from 1990 (on which the model was fitted) and the data from 2000, j is plotted against the fraction of articles that obtained years less citations than the predicted j^{th} quantile. If the model predicts correctly, these plots should be equal to the line $y = x$. For this reason this line is included as a guideline in the figures.

For the data from 1990 (on which the model was fitted) indeed the quantiles are predicted well. However, for the data form 2000, the model underestimates the quantiles. For example, only around 43% of the articles published in 2000 get less citations than their predicted 0.50^{th} quantile. We expect that this happens because of the fact that over the years there is a trend to add more references to articles. As a result of this, the average number of citations that articles get over time also grows [25]. Because the model is fitted on data from an earlier time period in which articles got less citations on average, the model underestimates the quantiles of articles that are published later.

6.1.1 Normalized citations

We want to adjust the prediction for the increase of the average number of citations that articles get. For this reason, we now make predictions on normalized citation data. This means that all the inputs in the model are scaled by the average value of that input. For example, the number of citations in the first year, c_1, is scaled by the average c_1 over all articles in the data set. Also the number of citations after

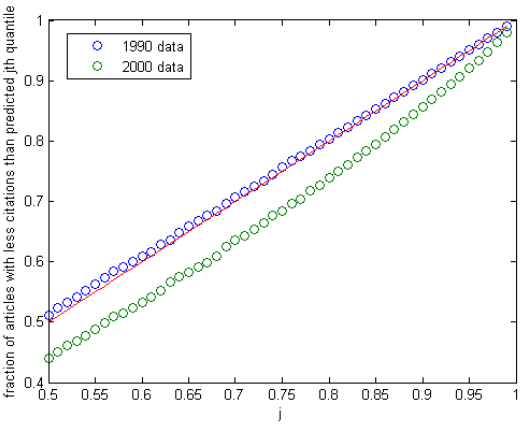

Figure 6.1: j versus the fraction of articles with less citations than predicted j^{th} quantile

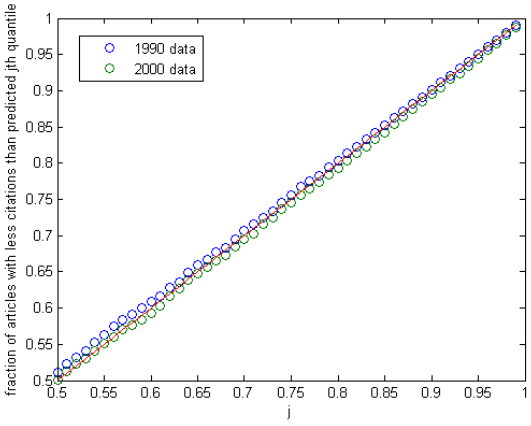

Figure 6.2: j versus the fraction of articles with less citations than predicted j^{th} quantile adjusted for increasing citation numbers

10 years is scaled by the average value before calculating the quantiles. The model is fitted using those scaled inputs for publications from 1990.

After that, all data from 2000 are also scaled by the average values of the Impact Factor, the number of citations in the first year, and the additional number of citations until 2010. Again, the coefficients that are fitted on the articles that were published in 1990 are used to predict the quantiles of the articles that were published in 2000. The results of these predictions are normalized predictions with respect to the average value. For example a predicted quantile of 2 means that we expect the quantile equals twice the average number of citations that articles in Physics receive.

Now a similar figure as Figure 6.1 is made. Figure 6.2 plots j against the fraction of articles that obtained less citations than their predicted j^{th} quantile using the normalized data. This is done for articles that were published in 1990 (the data on which the model was fitted), and for articles that were published in 2000. The figure shows that the fit for the articles from 2000 is now just as good as the one for the articles from 1990. This means that indeed the growing number of references is the factor that accounted for the underestimation of the quantiles of data from 2000 that occurred in Figure 6.1. The use of normalized data solves this underestimation. Although by using the normalized data, the model cannot be used for prediction of exact citation numbers anymore, it is still capable of predicting the quantiles relative to the mean number of citations.

Figures 6.3, 6.4, 6.5 and 6.6 give more illustrations of the fit of the model on data from 2000. Here the predicted j^{th} quantile is plotted versus the empirical quantile for three different values of j. The predictions and the empirical quantiles are normalized as described above. To compute the empirical quantiles, articles are again splitted into groups as described in in Section 2.3.1. In Figure 6.3, 6.4 and 6.5 all groups containing at least 50 articles are taken into account to compare the predicted and the empirical 0.50^{th}, 0.80^{th} and 0.95^{th} quantile respectively. In Figure 6.6, only groups containing at least 500 articles are taken into account. One dot in the figures corresponds to one predicted quantile for one group of articles. The different colors represent the different quantiles. The line in the figure is the line $y = x$. If the model predicts well, it is expected that the dots lie close to this line. Figures 6.7, 6.8, 6.9 and 6.10 are similar to Figures 6.3, 6.4, 6.5 and 6.6 for articles published in 1995.

The figures indicate that the predicted quantile and the empirical quantile can be very different. This becomes less when only taking into account the groups containing a lot of articles. One reason for this is that for small groups of articles, the empirical quantiles can be influenced by a few articles. Another reason is that the quantile regression fits the quantiles on all data. This means that the groups containing more articles have more 'weight' in the quantile regression.

For the 0.50^{th} and the 0.80^{th} quantile, the model seems to fit better than for the 0.95^{th} quantile.

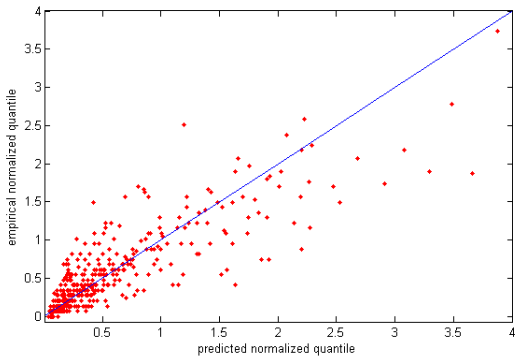

Figure 6.3: Predicted against empirical value of 0.50^{th} quantile of different groups for articles published in 2000 containing at least 50 articles

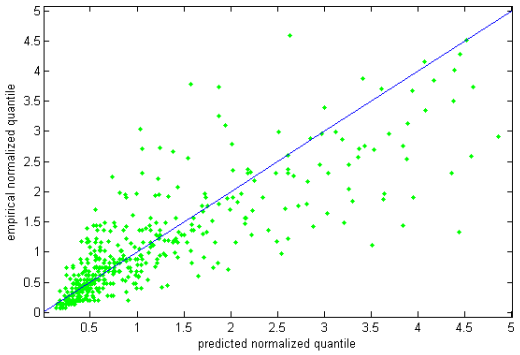

Figure 6.4: Predicted against empirical value of 0.80^{th} quantile of different groups for articles published in 2000 containing at least 50 articles

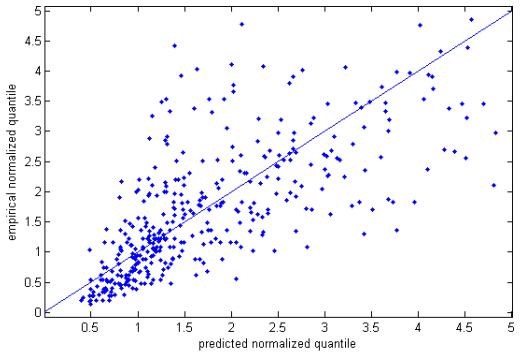

Figure 6.5: Predicted against empirical value of 0.95^{th} quantile of different groups for articles published in 2000 containing at least 50 articles

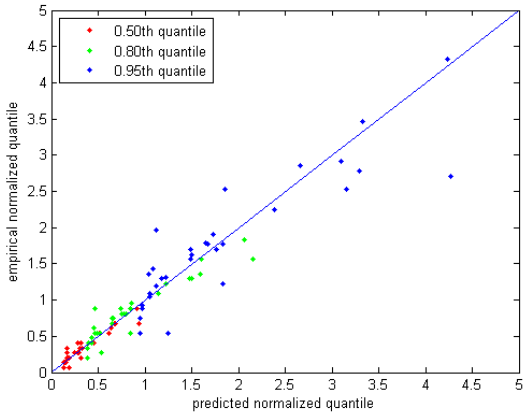

Figure 6.6: Predicted against empirical value of j^{th} quantiles of different groups for articles published in 2000 containing at least 500 articles for $j = 0.50, 0.80, 0.95$

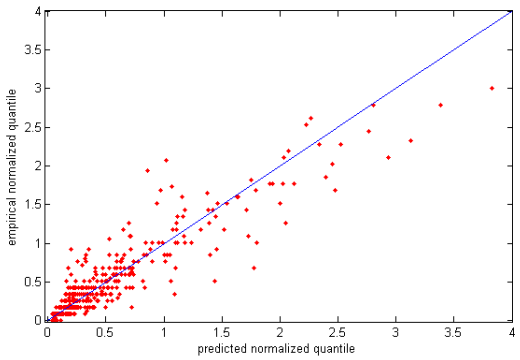

Figure 6.7: Predicted against empirical value of 0.50^{th} quantile of different groups for articles published in 1995 containing at least 50 articles

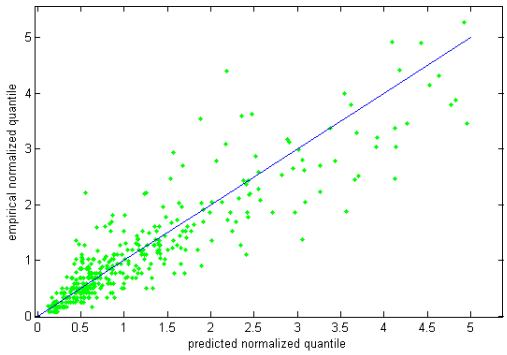

Figure 6.8: Predicted against empirical value of 0.80^{th} quantile of different groups for articles published in 1995 containing at least 50 articles

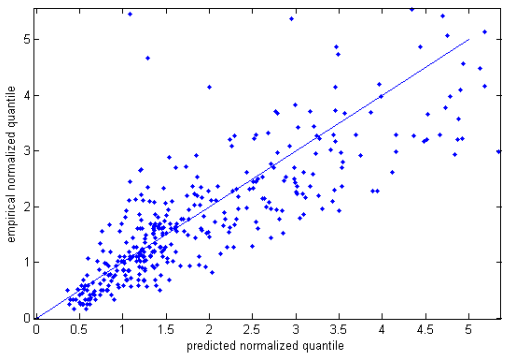

Figure 6.9: Predicted against empirical value of 0.95^{th} quantile of different groups for articles published in 1995 containing at least 50 articles

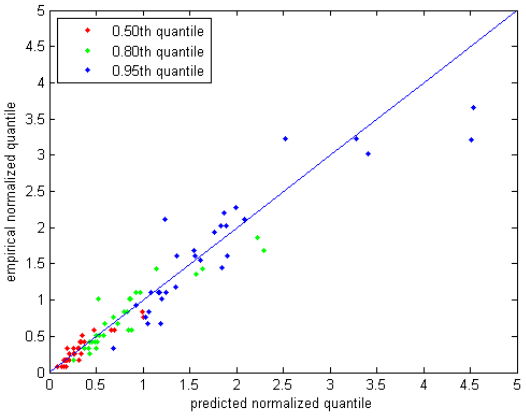

Figure 6.10: Predicted against empirical value of j^{th} quantiles of different groups for articles published in 1995 containing at least 500 articles for $j = 50, 80, 95$

6.2 Prediction at country level

In this section we look if the predicted quantiles that follow from the regression model from Section 4.3.3 are accurate for a group of articles from the same country.

To do this, first the model is fitted on all Physics publications published in 1990. Again, like in Section 6.1.1, all inputs are scaled by the average values. Then, the model is tested on all articles published in 2000 in the Physics category from a certain country. Figure 6.11 plots for several countries j against the fraction of articles of that country that have less citations than their predicted j^{th} quantile. The line $y = x$ is plotted as a guideline.

In this figure we can see that for example, the quantiles of articles from the USA are underestimated. For example, 40% of articles that originate from the USA have fewer citations than their expected 0.50^{th} quantiles. The same phenomenon occurs for publications from The Netherlands and England. This means that publications from those countries get more citations than average publications, if they are published in a journal with the same Impact Factor and obtained the same number of citations in the first year.

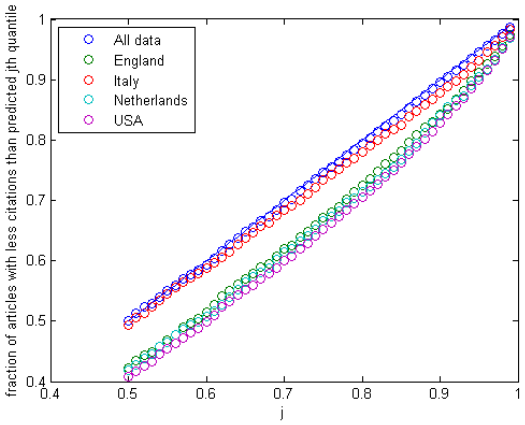

Figure 6.11: j versus fraction of articles with less citations than predicted j^{th} quantile for different countries

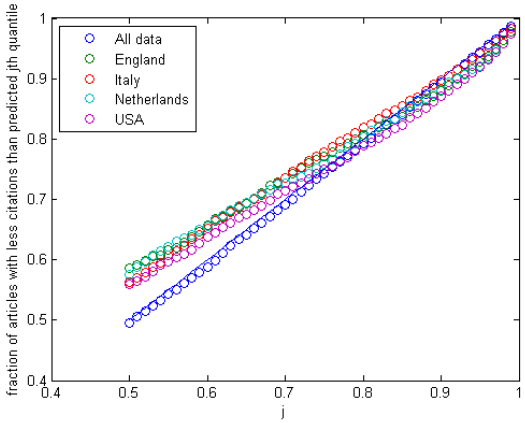

Figure 6.12: j versus fraction of articles with less citations than predicted j^{th} quantile for different countries, adjusted for previous prediction

6.2.1 Normalizing for prediction

We want to normalize for the correlation between publications from the same country, by taking into account how much better or worse the publications of that country were cited 10 years ago, compared to what was predicted. In this manner, the normalization constant $b_{c,j}$ for the predicted j^{th} quantile of some country c is defined:

$$b_{c,j} = \frac{j}{\text{fraction of articles from country } c \text{ with less citations than their expected } j^{\text{th}} \text{ quantile in 1990}}.$$

For example, if in 1990 for some country 40% of the publications had less citations than their expected 0.50^{th} quantile, this means that in 1990 this country performed better than predicted. The amount to which the country performed better than predicted in 1990 is now used to predict the quantiles for 2000. To predict the 0.50^{th} quantile of articles published in 2000 that originate from the same country, we multiply the predicted 0.50^{th} quantile from the regression by $b_{c,0.5} = 0.5/0.4$. In general, the new predicted quantiles are:

$$q(j)_{2000,c} = b_{c,j} q(j) \tag{6.1}$$

Here $q(j)$ are the predictions that follow from the model. The performance of the prediction by normalizing by this constant $b_{c,j}$ can be seen in Figure 6.12.

For the higher quantiles, the lines of all countries are now close to the line $y = x$. This indicates that the number of citations that articles from a certain country receive in the higher quantiles, given the Impact Factor and the number of citations in the first year, seems to be stable over time. For the lower quantiles this does not seem to hold. This means that the number of citations that the best scoring articles of a country get per Impact Factor and number of recent citations is more stable over time than for the publications with a number of citations closer to average.

Figure 6.13 plots the average Impact Factor and the average number of citations in the first year for the countries that appeared in Figures 6.11 and 6.12. The white dots are the average values from 1990, and the filled dots are the average values for the publications from 2000 of the same country. All these averages are normalized averages. For this reason the average value of the Impact Factor and the number of citations in the first year for all data is equal to one. The figure shows for example that Dutch publications in Physics had an average normalized Impact Factor of 1.05 in 1990, while this increased to 1.3 in 2000. Thus, the model expects that the Dutch publications from 2000 will be cited more often than the publications from 1990. In Figure 6.11, we see that the Dutch publications from 2000 obtain more citations than the model expects. However, if we look at the prediction that is normalized by the performance of the Dutch publications from 1990, we see in Figure 6.12 that the Dutch publications get less citations for publishing in a journal with the same Impact Factor and the same number of citations in the first year than in 1990. This holds especially for the lower quantiles. Hence, from 1990 to 2000, the Impact Factor and the number of recent citations of Dutch publications increased. However, the number of citations that these publications got per Impact Factor and recent citation, decreased.

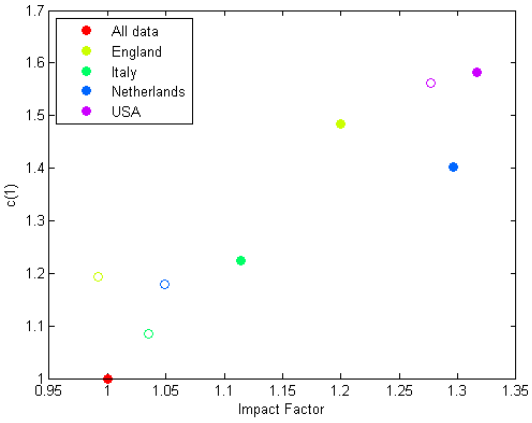

Figure 6.13: average IF and c_1 for different countries in 1990 (white dots) and 2000 (filled dots)

6.3 Prediction at university level

In this section we check how the model performs for predicting quantiles of publications from the same university. The same procedure as in Section 6.2 is used to compare the prediction and the realization of the number of citations at university level. Again, the coefficients of the model are fitted on publications from 1990. The prediction of the model for articles that were published in 2000 at a certain university is tested for a few universities. Results of this are in Figure 6.14. The same phenomenon as in the previous section, where the predictions on the country level were studied, occurs. For some universities the quantiles are overestimated, while for others they are underestimated. Especially the quantiles of publications from the university of Delft are underestimated. This indicates that there might also be a correlation between the impact of publications of one university. However, this is not captured in the model that was proposed in this report.

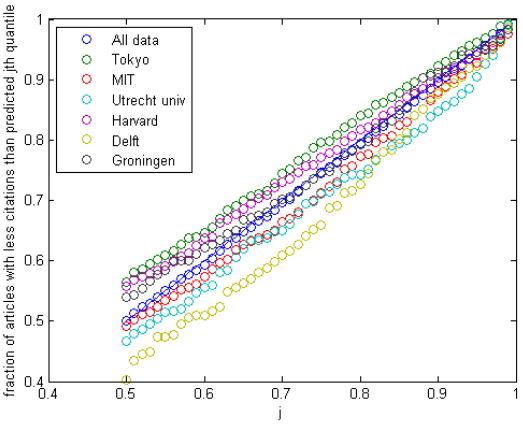

Figure 6.14: j versus fraction of articles with less citations than predicted j^{th} quantile for different universities

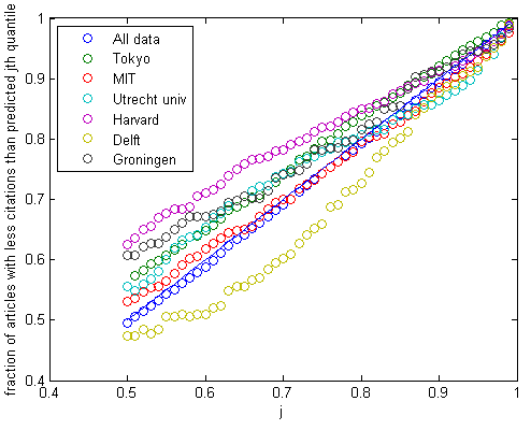

Figure 6.15: j versus fraction of articles with less citations than predicted j^{th} quantile for different universities, adjusted for previous prediction

6.3.1 Normalizing for prediction

The same procedure for normalizing for the correlation between publications of the same universities as the procedure for countries is used. Again, the predictions are normalized by the amount by which the quantiles of the publications of the same university were overestimated or underestimated 10 years ago, as in equation (6.1). Figure 6.15 shows the results of this. This procedure of normalizing does not bring the lines for the different universities closer to the line $y = x$, in contrast to the normalizing for different countries (Figure 6.12). This means that the performance of the publications of universities is less stable over time than the performance of publications from the same country.

Figure 6.14 shows for example that the quantiles of publications from Harvard are predicted well by the model. However, in 1990, publications from Harvard got more citations than predicted. Hence, because in 1990, the publications got more citations than predicted, the expectations for publications from 2000 are also raised. But in 2000, articles did not get more citations than the original model predicted, so this does not improve the prediction of the quantiles in Figure 6.15.

However, from Figure 6.15, we can say for example that in 1990, Harvard got more citations to articles with the same Impact Factor and the same number of citations in the first year than in 2000. Delft university, on the contrary, got in 1990 less citations for articles with the same Impact Factor and the same number of citations in the first year than in 2000. The average Impact Factor and number of citations in the first year in 1990 and 2000 for different universities is plotted in Figure 6.16. These averages are normalized, so for all data in total, the average Impact Factor and citations in the first year are both one.

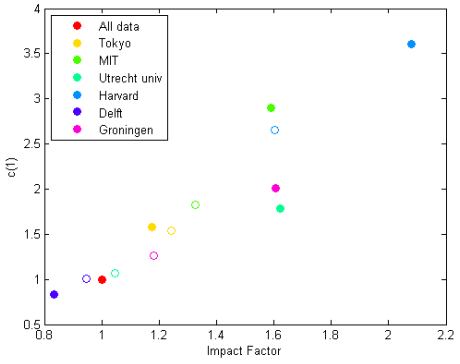

Figure 6.16: average IF and c_1 for different universities in 1990 (white dots) and 2000 (filled dots)

6.4 Explanation for prediction at country level

In Section 6.2, we could see that the model that was proposed, underestimates the quantiles of certain countries, and overestimates the quantiles of other countries. In this section, we give a possible explanation for this phenomenon by looking at an example.

In this example, we assume that there are only two countries. One country always publishes lower impact articles than the other country. We assume that articles get cited according to their impact. This gives the following situation for the countries :

- Country 1 (co_1) : all articles from country 1 end up with x_1 to x_2 citations with j^{th} quantile $\bar{x}$,
- Country 2 (co_2) : all articles from country 2 end up with y_1 to y_2 citations with j^{th} quantile $\bar{y}$,

where $x_2 > x_1 > y_2 > y_1$. This means that all articles from country 1 obtain more citations than articles from country 2. Furthermore, in this example there are only two journals of the same size:

- journal 1 : Impact Factor d_1,
- journal 2 : Impact Factor d_2,

where $d_1 > d_2$.

Now since journal 1 has a higher Impact Factor than journal 2, it publishes mostly higher impact articles. This means that journal 1 publishes mainly articles from country 1. However, there is some noise in the selection of articles. Because of this noise, a small amount of articles from country 2 is published in journal 1. This gives the following situation:

$$\begin{array}{c} \\ \text{Journal 1} \\ \text{Journal 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{Country 1} & \text{Country 2} \end{array} \\ \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix} \end{array}$$

This means that country 1 publishes a fraction of $1-p$ of its articles in journal 2, and a fraction of p of its publications in the high impact journal 1. Country 2 publishes a fraction of p articles in journal 2, and a fraction of $1-p$ articles in the high impact journal 1. Let q_1 be the j^{th} quantile of the citation distribution in journal 1, and q_2 the j^{th} quantile of the distribution of the number of citations in journal 2. We assume that the j^{th} quantile of all articles in journal 1 from country 1 is $\bar{x}$, and the j^{th} quantile of country 2 articles in journal 2 is $\bar{y}$. This means that we assume that the distribution of the number of citations to all country 1 articles, is the same as the distribution of the number of citations to country 1 articles in journal 1. If we also assume $1-p < j < p$, we have:

$$\begin{aligned} x_1 \leq \quad & q_1 \quad \leq \bar{x} \\ \bar{y} \leq \quad & q_2 \quad \leq y_2 \end{aligned}$$

We have $q_1 \leq \bar{x}$ because the j^{th} quantile of the articles in journal 1 that originate from country 1 is $\bar{x}$. Besides those articles, journal 1 contains some lowly cited articles from country 2, which can never increase the quantile. Similarly, we have $q_2 \geq \bar{y}$. Because $1-p < j$, the j^{th} quantile of journal 1 is the number of citations of one of the articles from country 1, so $x_1 \leq q_1$. Similarly, $y_2 \geq q_2$ holds.

If we have a model that predicts the j^{th} quantiles given the IF as $\hat{q}(IF)$, such that:

- $\hat{q}(d_1) = q_1$,
- $\hat{q}(d_2) = q_2$,

then the quantiles are predicted correctly for each Impact Factor.

However, the probability of an article from country 1 having less citations X than its predicted j^{th} quantile is:

$$\begin{aligned} P\left(X \leq \hat{q}(IF) | Co = co_1\right) &= P\left(X \leq q_1 | Co = co_1, IF = d_1\right) P\left(IF = d_1 | Co = co_1\right) \\ &\quad + P\left(X \leq q_2 | Co = co_1, IF = d_2\right) P\left(IF = d_2 | Co = co_1\right) \\ &\leq jp + 0 \cdot (1-p) \\ &\leq j \end{aligned}$$

The second inequality holds because $q_1 \leq \bar{x}$ and $q_2 \leq y_2 < x_1$. Since by definition of the quantile, a fraction of j articles from country 1 in journal 1 will have less citations than $\bar{x}$, at most a fraction of j will have less citations than q_1. For the articles from country 1 that were published in journal 2, no articles have less citations than q_2, because $q_2 < x_1$. In the last inequality, the two terms can be equal if and only if $p = 1$. Hence, the probability that an article from country 1 has less citations than its predicted j^{th} quantile is less than j if $p \neq 1$. This means that the quantiles for articles from country 1 are underestimated. For the articles from country 2 a similar reasoning shows that the probability of an article having less citations than its predicted j^{th} quantile is:

$$\begin{aligned} P\left(X \leq \hat{q}(IF) | Co = co_2\right) &= P\left(X \leq q_1 | Co = co_2, IF = d_1\right) P\left(IF = d_1 | Co = co_2\right) \\ &\quad + P\left(X \leq q_2 | Co = co_2, IF = d_2\right) P\left(IF = d_2 | Co = co_2\right) \\ &\geq 1 \cdot (1-p) + jp \\ &\geq j \end{aligned}$$

where in the last two equation, again equality holds if and only if $p = 1$. This shows that the quantiles for country 2 are overestimated if $p \neq 1$.

In this example, the model predicts the quantiles conditional on the Impact Factor correctly. However, if we only look at the predictions for country 1, we see that the quantiles of country 1 are underestimated by the model if $p \neq 1$ and the quantiles for country 2 are overestimated.

This occurs because in one journal, articles receive different numbers of citations in the end. In this example, the higher scoring articles in the two journals are from the same country, and the lower scoring articles are from the same country. In total, the model predicts the quantiles of the articles correctly. However, when looking at one country only, the model does not take into account that all high scoring articles in the journal are from one country, and all the low scoring articles from another. This could explain the incorrect predictions that Figure 6.11 showed.

Chapter 7

Sensitivity Analysis

In this chapter, we first study the sensitivity of the coefficients C_j, β_j and γ_j that occur in (4.3) to the scientific field. Then the influence of the country that published the article is investigated.

7.1 Fitting for different scientific disciplines

In all previous chapters, data from the scientific field of Physics were used to fit the coefficients of the regression model. In this section, the influence of the scientific field of the articles on the regression coefficients is studied. Again, articles that were published in 1984 are taken into account, and the goal of the model is to predict the conditional quantiles of the distribution of the number of citations these articles obtained up to 2013. The Impact Factor and the number of citations in the first year are used as covariates. The coefficients C_j, β_j and γ_j that follow from the quantile regression are plotted in Figures 7.1, 7.2 and 7.3 for three different fields of scientific publications.

The regression coefficient C_j is higher for articles in Biology and Chemistry than for articles in Physics. This means that in general, the articles in Biology and Chemistry obtain more citations than articles in the scientific field of Physics. This occurs especially in the lower quantiles, for the higher quantiles the difference in the coefficient C_j for the different scientific fields is small. The regression coefficient β_j is lower for Biology articles than for articles in the scientific fields of Physics and Chemistry. This means that in Biology, the Impact Factor is less determining for the quantiles. However, the differences in this coefficient are small. The coefficient γ_j is the highest for articles in the category of Physics. Hence in Physics the number of citations that an article has obtained in the first year, is more determining for the quantiles than in Biology or Chemistry.

7.2 Fitting parameters on data from one country

In this section, we fit the coefficients C_j, β_j and γ_j on articles from only one country. For example, we fit the coefficients on all publications from 1984 in the field of Physics, that originate from the United States. The same is done for publications from Japan. Figures 7.4, 7.5 and 7.6 show the coefficients resulting from this.

For example, the coefficient C_j is higher for publications from the United States than for publications from Japan and for all publications together. This indicates that the quantiles for publications from the United States are larger in general. The coefficient β_j is similar for publications from Japan and the United States. For the coefficients that were fitted on all world wide data, β_j is larger, indicating a larger predictive value of the Impact Factor than for Japanese and American publications. The coefficient γ_j is not very different for the different countries. This indicates that the influence of the number of recent citations on the number of additional citations the articles receive is similar for these different countries.

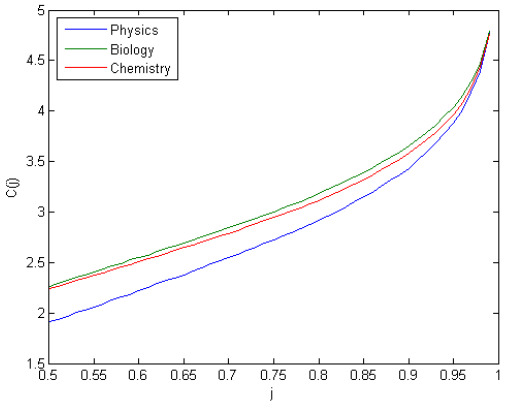

Figure 7.1: Regression coefficient C_j against j for three fields of science

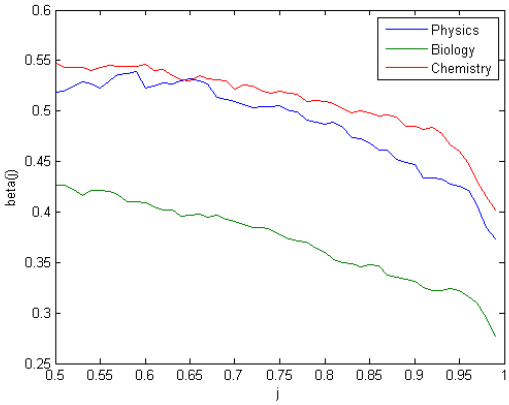

Figure 7.2: Regression coefficient β_j against j for three fields of science

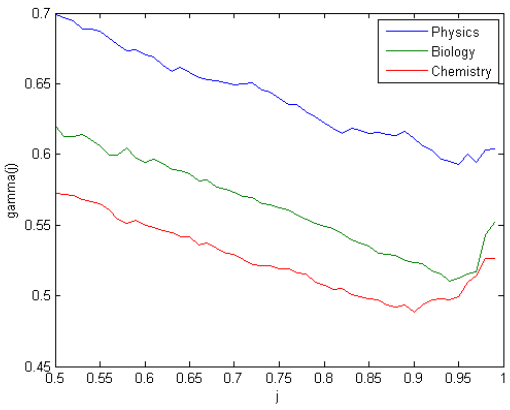

Figure 7.3: Regression coefficient γ_j against j for three fields of science

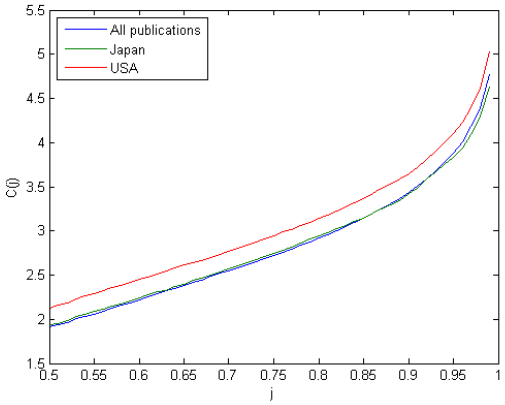

Figure 7.4: Regression coefficient C_j against j, fitted on data from different countries

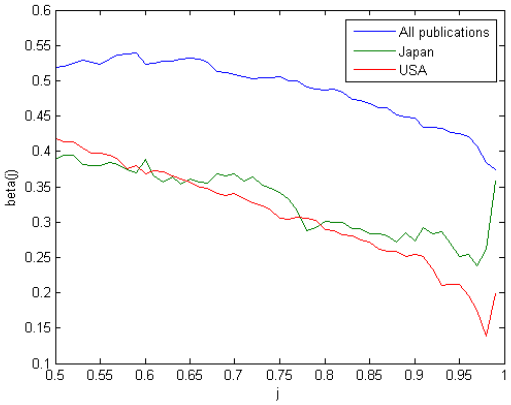

Figure 7.5: Regression coefficient β_j against j, fitted on data from different countries

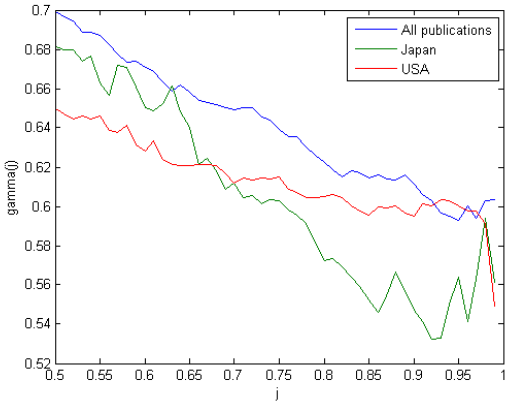

Figure 7.6: Regression coefficient γ_j against j, fitted on data from different countries

Chapter 8

Conclusions and Discussion

In this chapter, first we mention the conclusions of the study. Then, we discuss different aspects of the model that could be improved.

8.1 Conclusions

The conclusions of this report are the following:

- We investigated the tail of the distribution of the number of citations. We compared the tail index for articles with different Impact Factors and a different number of citations in the first year for data from the scientific field of Physics. This gave the conclusion that this tail can be approximated by a Pareto distribution. The tail index of the Pareto distribution seemed to be independent on the Impact Factor and the number of recent citations.
- By comparing the fit of a discretized lognormal model and a negative binomial model that were proposed in literature for the distribution of the number of citations, we conclude that the lognormal model is a better fit. For the tail of the distribution however, a Pareto distribution fits better.
- In this report, we proposed a model for predicting the quantiles of the citation distribution, conditional on the Impact Factor and the number of citations in the first year. The model was fitted using quantile regression. The good fit of this model to the data indicates that quantile regression is a suitable tool to fit the quantiles of the citation distribution. The coefficients β_j and γ_j in the quantile regression model, that correspond to respectively the Impact Factor and the number of recent citations, are not stable in j. This means that the influence of the Impact Factor and the number of recent citations is different for different quantiles. The influence of the Impact Factor and the number of recent citations is less for the higher quantiles than for the middle quantiles.
- Three different variants of the quantile regression model were proposed. The variant in which both the Impact Factor and the number of recent citations were used, fitted better to the data than the variants in which only one of these factors was included. This means that both the Impact Factor and the number of recent citations are important to predict the quantiles of the distribution of the number of citations.
- An estimator that uses the Pareto behavior of the tail to predict the higher quantiles, gives a good fit to the data. Combining this Pareto estimator and the regression estimator gives an equation that characterizes the behavior of the regression coefficient C_j for the high quantiles accurately.
- The precision of the predictions is in this report characterized by confidence intervals. These intervals are quite large for the high quantiles. For these high quantiles, the confidence intervals for the Pareto estimator give better results than the confidence intervals for the quantile regression estimator when comparing them to the data. Furthermore, the confidence intervals for the Pareto estimator are smaller than the confidence intervals for the regression estimator. However, the Pareto estimator uses a part of the data that we want to predict.

- The quantile regression model seems to predict the quantiles of the number of citations accurately, also for articles that were published later than the training data. However, the model is not suitable to predict the quantiles for a group of articles from the same university or the same country. The model overestimates these quantiles for some countries, and underestimates them for other countries. One possible explanation for this lack of fit is that higher scoring articles in journals with different Impact Factors originate from the same country, whereas the lower scoring articles in those journals also originate from the same country.
- The quantile regression coefficients are not the same for different fields of science, indicating that the influence of the Impact Factor and the number of recent citations can be different for different fields of science.

8.2 Discussion

The model that was proposed in this report, did not perform well when it came to predicting quantiles of the citation distribution for a group of articles from the same country or university. The model did predict the quantiles correctly when predicting for a group of articles from different countries and universities. This means that the Impact Factor and the number of recent citations are insufficient to predict the quantiles for the separate countries or universities correctly. Apparently, some factor is missing in the model that captures the differences between countries and universities. To be able to predict the performance of publications from a certain university or country, a different model would be needed.

In this report, all publications from the scientific field of Physics were taken into account. This is a very broad field of science, with different subfields. Within those subfields, the citation behavior might be different. This is not taken into account in this report, and might need further research.

In the model which we presented in this report, two factors were taken into account to predict the quantiles: the Impact Factor and the number of citations in the first year after publishing. Of course it is possible to add more factors to this model, such as the number of pages that the article consists of or the number of references in the article. In further research, the effect of adding these factors to the model could be investigated. In particular, it would be interesting to see if adding one or more covariates solves the inaccuracy in the predictions for a set of papers from the same country or university.

Another possibility for further research lies in the fitness factor we used in the model. We assumed that the fitness factor was a product of the two covariates to a certain power. Other variants of the fitness factor could also be investigated.

Appendix A

Quantile Regression

In Quantile Regression, the function:

$$\min_{\xi} \sum_i \rho_j(y_i - x_i\xi), \tag{A.1}$$

where

$$\rho_j(z) = zj - z\mathbb{1}_{z<0},$$

is minimized. To show that this quantile regression indeed fits the quantiles of the distribution, we look at the directional derivatives of equation (A.1) [14]. The values of $y_i - x_i\xi$ are just the residuals, which we call r_i. The directional derivative of the function that is minimized along vector u is:

$$\begin{aligned} & \frac{d}{dt} \sum_i \rho_j(y_i - x_i(\xi + ut))|_{t=0} \\ = & -\sum_i x_i u \rho_j'(r_i - x_i ut))|_{t=0} \\ = & -\sum_i x_i u \phi_j(r_i, -x_i u), \end{aligned}$$

where

$$\phi_j(a,b) = \begin{cases} j - \mathbb{1}_{a<0} & a \neq 0 \\ j - \mathbb{1}_{b<0} & a = 0. \end{cases}$$

The vector ξ minimizes equation (A.1), if the directional derivatives of the function are larger than or equal to zero in all directions, so if they satisfy:

$$-\sum_i x_i u \phi_j(r_i, -x_i u) \geq 0, \qquad \forall u.$$

If we take $u^\intercal = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, then because of the definition of X, $x_i u = 1$ for all i. Now the directional derivative with respect to this vector should of course be larger than or equal to zero:

$$\begin{aligned} -\sum_i \phi(r_i, -1) & \geq 0 \\ -\sum_{i:r_i=0} (j-1) - \sum_{i:r_i>0} j - \sum_{i:r_i<0} (j-1) & \geq 0 \\ -N_z(j-1) - N_p j - N_n(j-1) & \geq 0. \end{aligned}$$

Here N_z, N_p and N_p are respectively the number of observations with residuals of value zero, the number of articles with positive residuals, and the number of articles with negative residuals. Similarly,

for $u = \begin{bmatrix} -1 & 0 & 0 \end{bmatrix}$, $x_i u = -1$ for all i, so:

$$\begin{aligned} -\sum_i \phi(r_i, 1) &\geq 0 \\ \sum_{i:r_i=0} j + \sum_{i:r_i>0} j + \sum_{i:r_i<0} (j-1) &\geq 0 \\ N_z j + N_p j + N_n (j-1) &\geq 0. \end{aligned}$$

The total number of articles in the regression is n, where $n = N_z + N_p + N_n$. By rewriting the previous two inequalities and combining them, the following inequalities can be obtained:

$$\begin{aligned} \frac{N_n}{n} &\leq \quad j \quad \leq \frac{N_z + N_n}{n}, \\ \frac{N_p}{n} &\leq 1 - j \leq \frac{N_z + N_p}{n}. \end{aligned}$$

This means that indeed by minimizing equation (A.1), the resulting model fits the quantiles exactly onto the empirical quantiles. A fraction of j observations has a value less than the predicted j^{th} quantile.

Bibliography

[1] J. Beirlant, W. Glänzel, A. Carbonez, and H. Leemans. Scoring research output using statistical quantile plotting. *Journal of Informetrics*, 1(3):185–192, 2007.

[2] L. Bornmann. The problem of citation impact assessments for recent publication years in institutional evaluations. *Journal of Informetrics*, 7(3):722 – 729, 2013.

[3] L. Bornmann, L. Leydesdorff, and J. Wang. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, 8(1):175 – 180, 2014.

[4] Q. L. Burrell. The nth-citation distribution and obsolescence. *Scientometrics*, 53(3):309–323, 2002.

[5] Q. L. Burrell. Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, 54(5):372–378, 2003.

[6] P. Cirillo. Are your data really pareto distributed? *Physica A: Statistical Mechanics and its Applications*, 392(23):5947–5962, 2013.

[7] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[8] D. J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[9] A. L. Dekkers, J. H. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

[10] E. Garfield. The history and meaning of the journal impact factor. *The Journal of the American Medical Association*, 295(1):90–93, 2006.

[11] M. Golosovsky and S. Solomon. Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters*, 109(9):098701, 2012.

[12] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[13] W. Ke. A fitness model for scholarly impact analysis. *Scientometrics*, 94(3):981–998, 2013.

[14] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1):33–50, 1978.

[15] R. Koenker and J. A. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[16] J. M. Levitt and M. Thelwall. A combined bibliometric indicator to predict article impact. *Information Processing & Management*, 47(2):300 – 308, 2011.

[17] J. Lundberg. Lifting the crown – citation z-score. *Journal of Informetrics*, 1(2):145 – 154, 2007.

[18] J. Mingers and Q. L. Burrell. Modeling citation behavior in management science journals. *Information Processing and Management*, 42(6):1451 – 1464, 2006. Special Issue on Informetrics.

[19] H. P. F. Peters and A. F. J. van Raan. On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science*, 45(1):39–49, 1994.

[20] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[21] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[22] P. O. Seglen. The skewness of science. *Journal of the American Society for Information Science*, 43(9):628–638, 1992.

[23] D. I. Stern. High-ranked social science journal articles can be identified from early citation information. Crawford School Research Paper No. 14-06, 2014.

[24] M. J. Stringer, M. Sales-Pardo, and L. A. Nunes Amaral. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, 3(2):e1683, 02 2008.

[25] M. L. Wallace, V. Larivière, and Y. Gingras. Modeling a century of citation distributions. *Journal of Informetrics*, 3(4):296–303, 2009.

[26] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[27] M. Wang, G. Yu, J. Xu, H. He, D. Yu, and S. An. Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics*, 6(4):586 – 599, 2012.