Journal of Informetrics 7 (2013) 803–810

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Identifying excellent researchers: A new approach

Richard S.J. Tol[a,b,c,d,*]

[a] Department of Economics, University of Sussex, Jubilee Building, Falmer BN1 9SL, United Kingdom
[b] Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands
[c] Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands
[d] Tinbergen Institute, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 29 April 2013
Received in revised form 11 June 2013
Accepted 13 June 2013

JEL classification:
A11

Keywords:
Quantile kernel regression
Hirsch index
Economics

ABSTRACT

Quantile kernel regression is a flexible way to estimate the percentile of a scholar's quality stratified by a measurable characteristic, without imposing inappropriate assumption about functional form or population distribution. Quantile kernel regression is here applied to identifying the one-in-a-hundred economist per age cohort according to the Hirsch index.

1. Introduction

The Hirsch index (Hirsch, 2005) is an often-used measure of life-time achievement. The Hirsch index is the highest number h for which holds that an author has h publications that are cited h times or more. The Hirsch index cannot fall over time and tends to increase. Any ranking based on the Hirsch index thus favours those with a longer career. This is fine for many purposes, but not if the aim is to identify excellent individuals in a cohort, e.g., for hiring scholars (Ellison, 2010). The Hirsch rate (Burrell, 2006; Liang, 2006) – the Hirsch index over the number of active years – corrects for career length. However, the Hirsch rates assume a linear relationship between Hirsch index and active years. This may be problematic when comparing job candidates of different ages if the relationship is (locally) non-linear. This paper therefore proposes quantile kernel regression (Sheather & Marron, 1990) as a method to find exceptional researchers. Kernel regression does not impose linearity or any other functional form. Quantile regression focuses the analysis on exceptional, rather than average, scholars. The proposed method is applied to a sample of 32,000 economists. For illustration, I am looking for the one-in-a-hundred economists in each age group.

As far as I know, I am the first to apply quantile kernel regression to this problem.[1] Of course, this paper is not the first to seek to identify excellence (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003). Percentiles are a natural way to identify excellence: A scholar is excellent if she is better than, say, 99 out of 100 of her peers. People may be close in rank but far apart in percentile, and vice versa. The "crown indicator" – a *z*-score (Moed, 2010; van Raan, 2005) – corresponds to

[*] Correspondence to: Department of Economics, University of Sussex, Jubilee Building, Falmer BN1 9SL, United Kingdom. Tel.: +44 1273 877282.
E-mail addresses: r.tol@sussex.ac.uk, rsjtol@gmail.com

[1] A Scopus search on "quantile" or "kernel" and "Journal of Informetrics" or "Scientometrics" returned only three papers (Beirlant, Glaenzel, Carbonez, & Leemans, 2007; Hengl, Minasny, & Gould, 2009; Sarabia, Prieto, & Trueba, 2012), none of which does what the current paper does.

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.06.003

percentiles only if the underlying distribution is normal[2] or can be transformed to normality (Lundberg, 2007). Care needs to be taken when defining the peer population (Gingras & Lariviere, 2011; Herranz & Ruiz-Castillo, 2012; Leydesdorff & Opthof, 2011; van Raan, van Leeuwen, Visser, van Eck, & Waltman, 2010; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b). Quantile kernel regression directly estimates the percentiles (the quantile part) as a function of the subpopulation characteristics (the regression part), and without imposing any particular distribution (the kernel part).

I use the Hirsch number as a quality indicator. The Hirsch number is a mixture of the number of papers and citations to those papers. The method suggested below can just as easily be applied to other scientometric indicators of research quality and quantity. Other facets of academic quality – teaching, management, fund raising, policy advice – are either difficult to quantify or hard to obtain for, but important nonetheless.

The paper proceeds as follows. Section 2 presents the methods. Section 3 shows the data. Section 4 discusses the results. Section 5 concludes.

2. Methods

The intuition behind kernel regression is straightforward (Takezawa, 2006). A standard ordinary least squares regression $y = X\beta + u$ with $u \sim N(0,\sigma^2)$ is equivalent to $y \sim N(X\beta,\sigma^2)$. That is, our prediction for y, $X\beta$, is the expected value of a probability density function. That density is the density of y conditional on X. Below, we consider univariate regression only, so we replace matrix X by vector x; β becomes a scalar.

A conditional density function is defined as:

$$f(y|x) = \frac{f(x,y)}{f(x)} \tag{1}$$

The generic definition of a univariate kernel density function is:

$$\hat{f}(x) = \frac{1}{nh_x} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h_x}\right) \tag{2}$$

where h_x is the so-called bandwidth, X_i are a series of observations $i = 1, 2, \ldots, N$, and K is kernel function, which can be any function that integrates to one, with a first moment at zero and a finite second moment.

A bivariate kernel density is typically defined as:

$$\hat{f}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h_x}\right) K\left(\frac{y - Y_i}{h_y}\right) \tag{3}$$

The conditional kernel density follows from substituting (2) and (3) into (1). The expected value of y conditional on x then follows from:

$$\mathbb{E}(y|x) = \int_x y\hat{f}(y|x)dy \tag{4}$$

The exposition started with y as a linear function of x plus homoskedastic, normally distributed noise. Eq. (4) is conceptually similar, but the assumptions of normality, homoskedasticity and linearity were dropped. In the application below, I estimate the Hirsch index (y) as some function of academic age (x).

Although kernel regression analysis is often focused on Eq. (4), we in fact derive, as an intermediate step, the entire conditional distribution.[3] That is, we know not only the conditional mean (Eq. (4)), but also the higher conditional moments, the mode, median and any percentile that may hold our interest.

The literature on kernel density estimation is focussed on two questions: What kernel function K to use, and with which bandwidth h? There is no objective answer to those questions. If we use a Gaussian kernel function, we want the kernel density to be as close as possible to the Normal distribution, and we define closeness as the mean integrated square error, then the optimal bandwidth is:

$$h_x \cong 1.06\hat{\sigma}_x n^{-0.2} \tag{5}$$

See (Takezawa, 2006). I follow these conventions below.[4]

[2] The standard interpretation of a z-score is in hypothesis testing, but the familiar numbers only have their usual meaning under normality. For instance, a z-score of 2 (or rather, 1.96) implies a score that is exceptionally high – but only if the underlying distribution is normal it means 79 out of 80.

[3] Note that, under certain assumptions, it is possible to simplify the equations and skip the estimation of the bivariate density.

[4] The literature on quantile kernel regression (Falk, 1986; Parzen, 1979; Sheather & Marron, 1990; Takeuchi, Le, Sears, & Smola, 2006; Yu & Jones, 1998) advocates a range of alternative bandwidths, derived from a combination of distance measures and target distributions.