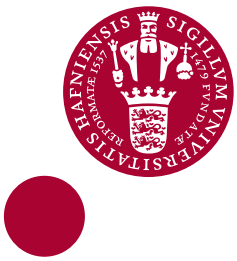


Measuring the influence of social networks on transaction costs using a nonparametric technique

Henningsen, Arne; Henningsen, Geraldine; Henning, Christian H.C.A.

Published in:
Symposium i Anvendt Statistik 2010

Publication date:
2010

Document version
Early version, also known as pre-print

Measuring the Influence of Social Networks on Transaction Costs Using a Non-parametric Regression Technique

Arne Henningsen [1], Geraldine Henningsen [2], Christian H.C.A. Henning [3]

[1] University of Copenhagen, Institute of Food and Resource Economics,
[2] Denmark Statistics, [3] University of Kiel, Department of Agricultural Economics

All business transactions as well as achieving innovations take up resources, subsumed under the concept of transaction costs (TAC). One of the major factors in TAC theory is information. Information networks can catalyse the interpersonal information exchange and hence, increase the access to non-public information. Our analysis shows that information networks have an impact on the level of TAC. Many resources that are sacrificed for TAC are inputs that also enter the technical production process. As most production data do not separate between these two usages of inputs, high transaction costs are unveiled by reduced productivity. A cross-validated local linear non-parametric regression shows that good information networks increase the productivity of farms. A bootstrapping procedure confirms that this result is statistically significant.

1 Introduction

In 1937 Coase argued in his essay "The Nature of the Firm" that market transactions often involve higher costs than just the market price. Other costs (e.g. search and information costs, bargaining costs, and policing and enforcement costs) can increase the costs of procuring something from a market. His theory became manifest in the concept of transaction costs, which has become a major field in institutional economics especially during the past 30 years.

His scholar Williamson (1971) carried on with this idea. He argued that firms can chose between two regimes of governance structure for transacting goods and services, namely markets and internal solutions. Later Williamson and other scholars (e.g. Grandori and Soda, 1995) extended the approach by integrating social networks into the theory. For instance, Williamson (1991) included inter-firm networks as "hybrid form" while other scholars (e.g. Powell, 1990; Johanisson, 1987) stated that networks are a third type organisational arrangement.

A vast literature shows that social networks provide well functioning mechanisms to reduce the resources that a firm needs for transacting goods and services and for accessing innovation. Hence, social networks reduce transaction costs (e.g. Henning, 2002).

Most of these resources used for transactions and accessing innovation are inputs also used in production. As most data sets do not separate between inputs used for technical production and resources dedicated to trade and innovativeness, increasing unobservable transaction costs are translated into lower productivity. Hence, we

conclude that firms with "good" social networks show higher productivity and vice versa. Our analysis tries to capture the effect of social networks on productivity.

In studies like this, the normal procedure to measure the influence of social networks on the efficiency of production would be to apply a stochastic frontier analysis (SFA) using the specification of Battese and Coelli (1995). The drawback of this approach is its rigid assumptions about the distributions of the error terms[1] and about linearity between the effects of social networks and the inefficiency. If these assumptions are false—which can easily be the case in such complex relationships—the estimated parameters and the statistical tests will be misleading. Therefore, to avoid specifying a parametric functional form, we estimate the influence of social networks on productivity by a non-parametric regression technique. In the following section we describe our methodological approach. The results are presented in the third section and the last section concludes.

2 Data and Methodology

In our empirical analysis we use a data set on Polish farms. The data were collected within the framework of the "Advanced-Eval" project financed by the European Union within the Sixth Framework Programme (contract number 022708). The data set includes detailed farm accountancy data and information on the farms' ego centred networks. We take the total value of all produced goods as output (in Zloty) and we distinguish between four inputs: labour (in working hours), land (in ha), capital (in Zloty), and intermediate inputs (in Zloty). Furthermore, we include management characteristics: education (as an ordered categorical variable), experience (in years), and risk attitudes (as a continuous variable, where increasing values indicate increasing risk aversion). In addition to the above-mentioned variables, we include four network parameters as well as an unordered categorical variable for the four regions, where the farms in our data set are located.

We take the logarithm of the output and all the input quantities so that the individual values of these variables are more equally distributed within the range of observed values. Otherwise, there were many observations within the bandwidths for small values (farms) but only very few observations within the bandwidth for large values (farms), which usually causes problems in non-parametric regression when fixed bandwidths are used.[2]

We apply a non-parametric local-linear estimation method, which was initially suggested by Stone (1977) and Cleveland (1979). Since we have both continuous and categorical explanatory variables, we use the extension of this estimator for mixed data types proposed by Li and Racine (2004) and Racine and Li (2004). In the presence of both continuous and categorical explanatory variables, this estimator outperforms the

1 A normal distribution for the general error term and a truncated normal distribution for the efficiency term is assumed under most stochastic frontier analyses (Battese and Coelli, 1995).

2 An alternative would be to use a "nearest neighbor"-method.

local-constant estimator (Li and Racine, 2004).
We use the second-order Epanechnikov kernel for continuous regressors, the kernel proposed by Aitchison and Aitken (1976, p. 29) for unordered categorical explanatory variables, and the kernel proposed by Wang and van Ryzin (1981) for ordered categorical explanatory variables. Since the bandwidths of the regressors are pivotal for the estimation, we use a data-driven bandwidth selection method, which has been proposed by (Hurvich, Simonoff, and Tsai, 1998) and is based on a corrected Akaiki information. It is an expected Kullback-Leibler cross-validation method and has very good finite sample properties (Li and Racine, 2004, p. 501).
We also used the second-order Gaussian kernel for the continuous regressors, but the bandwidths suggested by the cross-validation strongly depended on the starting values, where the bandwidths of some (varying) regressors were very small, which resulted in extreme under-smoothing.
The estimation was done within the statistical software environment "R" (R Development Core Team, 2009) using the add-on package "np" (Hayfield and Racine, 2008).

3 Results

The cross-validated bandwidths obtained by the method of Hurvich, Simonoff, and Tsai (1998) are presented in table 1. The bandwidths of the continuous explanatory variables are very large, indicating that the relationship between these independent variables and the dependent variable is approximately linear. However, in contrast to a parametric linear regression (e.g. OLS), our non-parametric regression with large bandwidths still allows for interaction effects between the regressors, i.e. the effect of one regressor on the dependent variable may depend on the values of all other regressors.

Table 1: Bandwidths

Variable	Bandwidth	Scale Factor
lLabor	64083	185067
lLand	103683	184643
lCapital	177112	253857
lIntermed	439205	628217
education	1.000	2.202
exper	15471269	2294050
risk	1421294	2882712
municip	0.591	1.300
outdFarm	4870844	4875115
outdHH	2731927	2921081
densFarm	503279	6998705
densHH	4870860	16505559

The gradients of the independent variable with respect to the explanatory variables are summarized in table 2. All input quantities (lLabor, lLand, lCapital, lIntermed) have a positive effect on the output quantity at all observations. Hence, the monotonicity condition derived from microeconomic production theory is fulfilled in our analysis even though the input quantities include transaction costs. As all input and output quantities are logarithmised, the gradients can be interpreted as partial production elasticities of the inputs. The elasticities of scale, which are equal to the sums over the four partial production elasticities, range from 1.09 to 1.17, indicating that all farms operate under increasing returns to scale.

Table 2: Gradients: minimum, mean, median, and maximum

Variable	Min	Mean	Median	Max
lLabor	0.11	0.15	0.15	0.18
lLand	0.26	0.37	0.37	0.45
lCapital	0.13	0.20	0.22	0.25
lIntermed	0.31	0.41	0.40	0.50
sum: all inputs	1.09	1.14	1.15	1.17
education: 1 → 2	-0.27	-0.00	-0.01	0.31
education: 2 → 3	-0.14	0.02	0.01	0.23
education: 3 → 4	-0.09	-0.01	-0.01	0.06
exper	-0.01	-0.01	-0.01	0.00
risk	-0.08	0.02	0.01	0.10
municip: chot → kami	-0.06	0.09	0.08	0.28
municip: chot → siem	-0.19	-0.01	-0.01	0.14
municip: chot → wiel	-0.18	0.14	0.15	0.41

The gradients with respect to the farm manager's education (education) describe the effect of increasing education by one level, i.e. from level 1 to 2, from 2 to 3, and from 3 to 4. The estimated gradients in table 2 show that the effect of the farm manager's education on the output is ambiguous and on average higher education neither increases nor decreases the output. The effect of the farm manager's experience (exper) on the output is negative for most farms, where each year of experience can reduce the output by a maximum of 1%. The farm manager's risk aversion (risk) has an ambiguous effect, which is positive for some farms and negative for others. The gradients with respect to the municipality where the farm is located (municip) describe the expected differences in output that are due to farms lying in different municipalities. We take the municipality Chotcza (chot) as the base for our comparison. Farms that are located in the municipality Siemiatkowo (siem) need on average roughly as many resources for improving technology, trading goods, and producing the same output as farms in the municipality Chotcza. In contrast, farms that

are located in the municipalities Kamieniec (kami) and Wieliszew (wiel) can produce on average 9% and 14% more outputs, respectively, with the same amount of inputs.

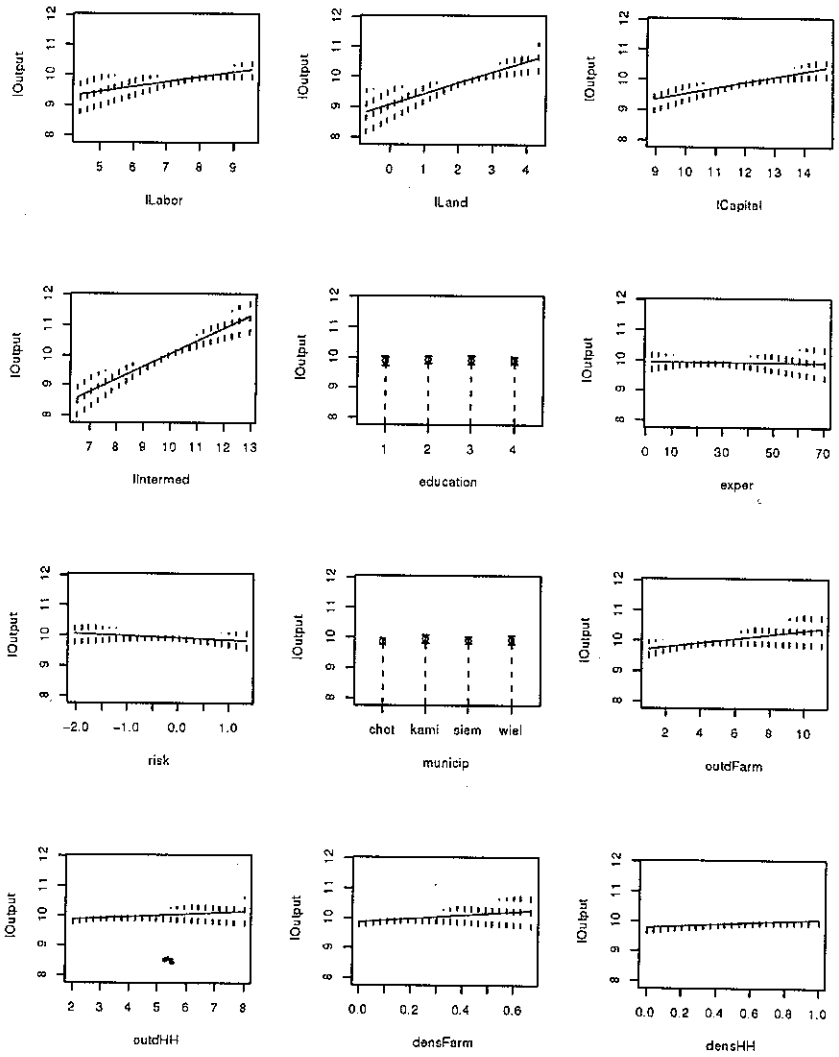


Figure 1: Estimation results

As a complement to the gradients shown in table 2, we present the estimation results in figure 1 graphically. While the gradients shown in table 2 are calculated at all data points that are in the sample, the estimated relationships displayed in figure 1 are

calculated by holding the other explanatory variables constant at their medians (continuous variables) or their modal values (categorical variables). Furthermore, the figure shows the 95% variability bounds obtained by bootstrapping (see Hayfield and Racine, 2008, p. 17). Most findings derived from the gradients shown in table 2 are confirmed in figure 1. However, in contrast to the gradients in table 2, figure 1 suggests that the farm manager's experience (exper) has virtually no effect and the farm manager's risk aversion (risk) even decreases the output. These contradicting results and the variation bounds, which are relatively large compared to the small effects of these two variables, indicate that these variables do not have a clear and significant effect.

Table 3: Statistical significance of regressors

Variable	P value	
lLabor	0.02757	*
lLand	0.00000	***
lCapital	0.00752	**
lIntermed	0.00000	***
education	0.22306	
exper	0.14787	
risk	0.69173	
municip	0.04261	*

We use the bootstrapping method suggested by Racine (1997) and Racine, Hart, and Li (2006) to test the statistical significance of all explanatory variables (see Hayfield and Racine, 2008, p. 9). The results are presented in table 3. All four inputs (labour, land, capital, intermediate inputs) as well as the location (municipality) of the farm but none of the three management variables (education, experience, risk attitudes) have a statistically significant effect on the output.

4 Conclusions

Our study has shown that a non-parametric regression is proper approach to model unknown complex relationships. The estimation of individual gradients allows for further analyses of the results, which would not be possible in classical parametric regression analysis.

Acknowledgements

The authors are grateful to Jeff Racine for giving them valuable suggestions regarding the non-parametric regression. Arne Henningsen is grateful to the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for financially supporting this research. Of course, all errors are the sole responsibility of the authors.

References

Aitchison, J., and C.G.G. Aitken. 1976. "Multivariate Binary Discrimination by the Kernel Method." Biometrika 63:413–420.

Battese, G.E., and T.J. Coelli. 1995. "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data." Empirical Economics 20:325–332.

Cleveland, W.S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association 74:829–836.

Coase, R.H. 1937. "The Nature of the Firm." Economica 4:386–405.

Grandori, A., and G. Soda. 1995. "Inter-firm Networks: Antecedents, Mechanisms and Forms." Organization Studies 16:183–214.

Hayfield, T., and J.S. Racine. 2008. "Nonparametric Econometrics: The np Package." Journal of Statistical Software 27:1–32.

Henning, C.H.C.A. 2002. "Social Capital and Exchange Networks." Working Papers of Agricultural Policy No. 7, Institut für Agrarpolitik, University of Kiel, March.

Hurvich, C.M., J.S. Simonoff, and C.L. Tsai. 1998. "Smooting Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." Journal of the Royal Statistical Society Series B 60:271–293.

Johanisson, B. 1987. "Organizing the Network Metaphor." International Studies of Management and Organization (Special Issue) 17.

Li, Q., and J.S. Racine. 2004. "Cross-Validated Local Linear Nonparametric Regression." Statistica Sinica 14:485–512.

Powell, W.W. 1990. "Neither Market nor Hierarchy: Network Forms of Organization." In L. L. Cummings and B. Staw, eds. Research in Organizational Behaviour. Greewich, CT: JAI Press, vol. 12, pp. 295–336.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Racine, J.S. 1997. "Consistent Significance Testing for Nonparametric Regression." Journal of Business and Economic Statistics 15:369–379.

Racine, J.S., J. Hart, and Q. Li. 2006. "Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models." Econometric Reviews 25:523–544.

Racine, J.S., and Q. Li. 2004. "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data." Journal of Econometrics 119:99–130.

Stone, C.J. 1977. "Consistent Nonparametric Regression." Annals of Statistics 5:595–645.

Wang, M.C., and J. van Ryzin. 1981. "A Class of Smooth Estimators for Discrete Distributions." Biometrika 68:301–309.

Williamson, O.E. 1991. "Comparative Economic Organization: The Analysis of Discrete Structural Alternatives." Administrative Science Quarterly 36:269–296.

—. 1971. "The Vertical Integration of Production: Market Failure Considerations." American Economic Review 61:112–123.

GINI-koefficienten - negative observationer

Rune Østergaard Pedersen, Metode, Danmarks Statistik

Resume

GINI-koefficienten bruges til at vurdere uligheden for en given variabel, og anvendes ofte indenfor økonomi og biologi til at beskrive sammensætningen af en population.

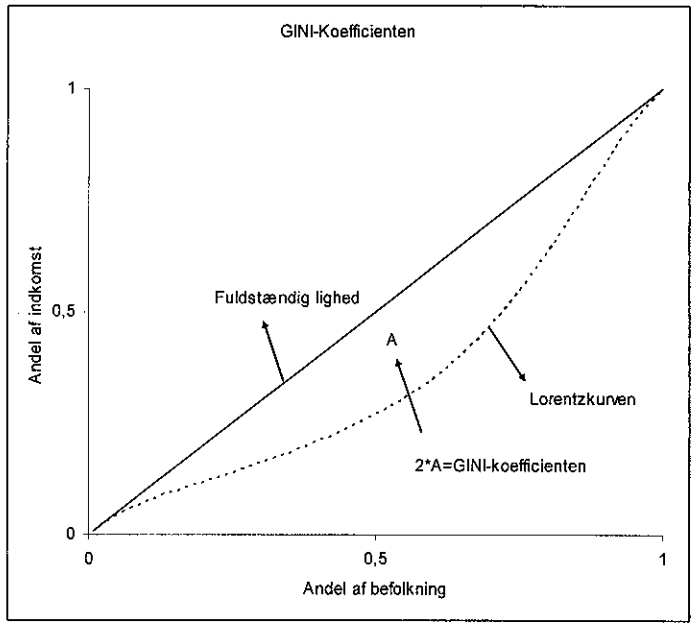
I dette notat beskrives udregningen af GINI-koefficienten, samt hvordan denne skal normaliseres så datasæt kan sammenlignes uanset om de indeholder negative værdier eller ej. Uden normalisering vil GINI-koefficienten overvurdere uligheden, når data indeholder negative observationer, og man drager forkerte konklusioner. En bootstrap metode bruges til at udregne et konfidensinterval for den normaliserede GINI-koefficient , og der gives et praktisk eksempel fra DST på beregning samt opstilling af konfidensintervaller for den normaliserede GINI-koefficient.

Nøgleord Normaliseret GINI-koefficient, negative observationer, *Percentile bootstrap confidense interval*, ulighed.

Indledning

GINI-koefficienten blev oprindeligt introduceret af den italienske matematiker **Corrado GINI** i artiklen *Variabilità e mutabilità* i 1912.

GINI-koefficenten bruges i dag som et mål for ulighed for en given målvariabel i en population. Den er udledt grafisk vha. den såkaldte Lorentzkurve, og antager en værdi mellem 0 og 1, hvor 1 er total ulighed, og 0 er fuldstændig lighed. GINI-koefficienten et det dobbelte af arealet mellem Lorentzkurven og en 45 ° linje som angiver total lighed(jf. Skovsgaard 1997). Dette er illustreret på figur 1, som arealet mellem den stiplede linje (Lorentzkurven), og den rette linje som angiver fuldstændig lighed.



Figur 1

I DST er det oftest en økonomisk variabel, som har interesse, når GINI-koefficienten skal beregnes. Variablen kunne fx være indkomst. Af forskellige årsager er det muligt at have negativ indkomst, og det er dette notats hensigt at beskrive, hvordan man bør normalisere GINI-koefficenten i denne situation, således at GINI-koefficenterne bliver sammenlignelige uanset om der optræder negative værdier eller ikke. Der henvises desuden til artiklen af Chen et al 1982 for en mere nuanceret belysning af problemstillingen og den senere korrektion til dette bidrag, skrevet af Berrebi og Silber i 1985.

Foruden den formelle beregning af GINI-koefficienten beskrives også en SAS-macro til at bootstrappe konfidensintervallet for GINI-koefficienten, når denne er baseret på en stikprøve.

Om GINI-koefficientens beregning

GINI-koeffienten beregnes ud fra nedenstående formel, som formelt kan beregnes for alle rationelle tal.

1) $G=\left(\frac{1}{2n^2*u}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i-y_j|$ Her angiver n antallet af observationer, y er observationerne og u angiver middelværdien af observationerne.

Det bemærkes at der tages numerisk værdi af de parvise forskelle, således at afstanden er uafhængigt af fortegnet på observation y.

GINI-koefficienten kan også beregnes ud fra nedenstående formel 2), hvor observationerne j, er sorteret i stigende rækkefølge. Notationen for formel 2) er som 1), bortset fra at *dispj* er observationerne i stigende rækkefølge. Formel 2) bør anvendes, da den er beregningsmæssigt mindre krævende end 1).

2) $G=\frac{2}{n*u}*\sum_{j=1}^{n}\frac{j}{n}disp_j-\frac{n+1}{n}$

GINI-koefficienten angiver arealet mellem Lorentzkurven og en 45° linje med observationerne kumulativt ud af ud af første-aksen, og den variabel som i fokus, opregnet cumulativt ud af anden-aksen. Både variablene på 1. og 2.-aksen er omregnet til en procentskala(Alternativt kan man bruge en relativ skala gående fra 0-1 som vist på figur 1).

Ofte vil man i økonomi have en variabel, som kan antage negative værdier. Teknisk kan man i så tilfælde sagtens beregne GINI-koefficienten ifølge formel 1) og 2). Men som det er blevet vist af Chen et al (1982), vil man i så fald overvurdere uligheden, svarende til, at man får en for høj GINI-koefficient. Årsagen til dette er, at arealet under Lorentskurven ikke længere er nedadtil afgrænset af en mindsteværdi på 0. Hvis man har numerisk store negative observationer, som det fx er tilfældet for formuer, vil man endog kunne observere en GINI-koefficient større end 1, hvilket naturligvis er meningsløst.

Det kan vises, at man kan korrigere for dette, idet man under udregning af den normaliserede GINI-koefficient får et mål, som i fravær af negative værdier blot giver den "normale" GINI-koefficient, og i tilfælde af negative værdier er sammenlignelig med GINI-koefficienter beregnet på baggrund af positive værdier.

Den normaliserede GINI-Koefficient G* er givet som:

3) $G^*=G/\left(1+\frac{2}{n*u}*\sum_{j=1}^{k}\frac{j}{n}disp_j\right)$ Notationen for 3) er som under 1) og 2), men det skal bemærkes, at der summes over k, hvor k er det heltallige antal observationer, som skal akkumuleres, indtil summen er 0. Tallet k beregnes inklusive den observation, som netop får denne sum til at blive 0. Det fremgår af 3), at dersom der ikke er nogle negative værdier, vil summen $\sum_{j=1}^{k}\frac{j}{n}disp_j$ blive 0, og derfor fås den almindelige GINI-koefficient.

Bestemmelsen af k er naturligvis kun mulig, hvis det netop er et heltal, som får summen til at blive 0. I praksis er dette naturligvis sjældent tilfældet. Derfor er det specielt for små n nødvendigt at korrigere for den situation, som opstår, når obs i k<0 og i+1 større end 0.