Using Non-parametric Methods in Econometric Production Analysis: An Application to Polish Family Farms

TOMASZ CZEKAJ and ARNE HENNINGSEN

Institute of Food and Resource Economics, University of Copenhagen,
Rolighedsvej 25, 1958 Frederiksberg C, Denmark
e-mail: tcz@foi.dk, arne@foi.dk

Paper prepared for presentation at the EAAE 2011 Congress
Change and Uncertainty
Challenges for Agriculture,
Food and Natural Resources

August 30 to September 2, 2011
ETH Zurich, Zurich, Switzerland

Abstract

Econometric estimation of production functions is one of the most common methods in applied economic production analysis. These studies usually apply parametric estimation techniques, which obligate the researcher to specify the functional form of the production function. Most often, the Cobb-Douglas or the Translog production function is used.

However, the specification of a functional form for the production function involves the risk of specifying a functional form that is not similar to the "true" relationship between the inputs and the output. This misspecification might result in biased estimation results—including measures that are of interest of applied economists, such as elasticities. Therefore, we propose to use non-parametric econometric methods. First, they can be applied to verify the functional form used in parametric estimations of production functions. Second, they can be directly used for estimating production functions without specifying a functional form and thus, avoiding possible misspecification errors.

We use a balanced panel data set of farms specialized in crop production that is constructed from Polish FADN data for the years 2004-2007. Our analysis shows that neither the Cobb-Douglas function nor the Translog function are consistent with the "true" relationship between the inputs and the output in our data set. We solve this problem by using non-parametric regression. This approach delivers reasonable results, which are on average not too different from the results of the parametric estimations but many individual results are rather different.

1 Introduction

One of the most common approaches in applied economic production analysis is the econometric estimation of production functions. The idea of an algebraic relationship between inputs and output was developed in the eighteenth century in works of A.R.J. Turgot and P.T. Malthus and in nineteenth century in works of D. Ricardo and J.H. von Thünen (Humphrey, 1997).[1] Finally, Wicksteed (1894) was the first who explicitly used the concept of an algebraic production function and Cobb and Douglas were the first who used econometric techniques to estimate a production function.

Given the strong assumptions that the Cobb-Douglas production function imposes on the underlying technology, Christensen et al. (1971) proposed a more flexible generalisation of the Cobb-Douglas function, the Translog (transcendental logarithmic) production function. These two functional form played the predominant role in applied production and efficiency analysis in the past 30 years. Apart from the non-parametric but deterministic (non-stochastic) Data Envelopment Analy-

[1] The production function proposed by J.H. von Thünen in 1840 is in fact the same, however in indirect form, as the probably most famous production function, the so-called Cobb-Douglas function (Humphrey, 1997).

sis (DEA), parametric econometric estimation techniques dominated the field of applied production and efficiency analysis. However, recently, Henningsen and Kumbhakar (2009) advertised a semi-parametric approach to efficiency analysis that estimates the production frontier by non-parametric regression and hence, avoids the specification of a functional form.

The aim of this paper is to compare the parametric and non-parametric estimation of the production function of Polish family farms. By using non-parametric econometric methods, we scrutinize the traditional parametric estimation methods.

This paper is organized as follows: the second section briefly introduces the parametric and non-parametric approaches used in applied production analysis; the third section describes the data used in this study; the fourth section presents the results of the conducted analyses; and the fifth section concludes.

2 Parametric and Non-parametric Approaches

The purpose of regression analysis is to evaluate the effects of one or more explanatory variables on a single dependent variable. This is done by evaluating the conditional expectation of the dependent variable given the explanatory variables, which can be expressed as:

$$Y_i = f(X_i) + \varepsilon_i, \tag{1}$$

where Y_i is the conditional expectation of the dependent variable, $f(X_i)$ is the unknown regression function, and ε_i is the error term.

The traditional parametric approach to regression analysis requires the specification of a functional form for $f(X_i)$, where the econometric estimation searches for the parameters that give the best fit to the model, e.g. by minimizing the sum of the squared residuals, $\sum_i \varepsilon_i^2$. In contrast, the non-parametric approach does not require the parametric specification of $f(X_i)$.

2.1 Parametric Approach

The specification of the functional form for $f(X_i)$ is one of the most crucial decisions in the parametric approach to econometric production analysis. The Cobb-Douglas function as well as its generalisation, the Translog (transcendental logarithmic) function, are most commonly used in applied production economics. One important reason for this is that both functions are (after logarithmic transformation) linear in parameters and hence, can be estimated by simple linear regression techniques.

In case of cross-sectional data (several firms observed at a single period of time), the ordinary least squares (OLS) method is often suitable. If several observations are available for each firm, the usual panel data estimators such as the fixed effects (FE) and the random effects (RE) estimators can be used. These estimators can account for individual or time specific heterogeneity, which are often observed in panel data. The Hausman test (Hausman, 1978) allows to test if the more efficient RE estimator is consistent or if the less-efficient but consistent FE estimator should be used.

As the selection of the functional form for modelling the relationship between the inputs and the output is rather arbitrary, there is a high chance of misspecification. The main problem is that the calculations of measures such as marginal products, partial production elasticities, and elasticities of scale, as well as various statistical tests become incorrect in the case of misspecification. In many circumstances, this problem can be solved by non-parametric regression.

2.2 Nonparametric Approach

In contrast, the non-parametric approach to regression analysis does not require the specification of the functional relationship between the explanatory variables and the dependent variable. Hence, a possible misspecification of the functional form is avoided in this approach. In this study, we apply a non-parametric local-linear kernel estimator. One can think of this estimator as a set of weighted linear regressions, where a weighted linear regression is performed at each observation and the weights of the other observations decrease with the distance from the respective observation. The weights are determined by a kernel function and a set of bandwidths, where a bandwidth for each explanatory variable must be specified. The smaller the bandwidth, the faster decreases the weight with the distance from the respective observation. While initially the bandwidths were determined by using a rule of thumb, nowadays cross-validation is used to determine the optimal bandwidths given the specified model and data set.

Hence, the overall shape of the relationship between the inputs and the output is determined by the data and the effects of the explanatory variables can be different at each observation without being restricted by a functional form.

3 Data

In this study we use balanced panel data from the Polish Farm Accountancy Data Network (Polski FADN) consisting of 371 crop farms in each of the four years from 2004 to 2007. Hence, our data set includes 1484 observation in total.

The dependent variable of the production function is the farms' output measured as the value of the total agricultural production. Four inputs are used in the regression analyses: labour, land, intermediate inputs, and capital. Labour is measured by Annual Work Units, where 1 AWU equals 2200 hours of work per year. Total utilised agricultural area in hectares is used as a measure of land input. Intermediate inputs are measured as the sum of total farming overheads (e.g. maintenance, energy, services, other direct inputs) and specific costs (e.g. fertilizers, pesticides, seeds). Capital input is measured as value of total fixed assets excluding the value of land. Since data on total agricultural production, intermediate inputs, and capital are expressed as monetary values expressed in current Polish Zloty (PLN), these data were deflated by national price indices published by the GUS (2008).[2] Descriptive statistics of the regression variables are presented in Table 1.

Table 1: Descriptive statistics of regression variables

Variable	Min	Median	Mean	Max	Std. dev.
Output (Y) [in PLN]	9843.88	122777.58	190865.50	2161685.78	218762.90
Labour (L) [in AWU]	0.15	1.50	1.66	7.41	0.87
Land (A) [in ha]	7.80	62.24	88.33	756.29	92.08
Intermediate Inputs (V) [in PLN]	6125.69	78471.41	119202.20	1337846.67	133803.47
Capital Stock (C) [in PLN]	10526.22	249324.44	370758.14	2766079.00	350460.40

Source: Own calculations based on Polish FADN data.

4 Results

4.1 Parametric Approach

We have estimated the parametric models within the statistical software environment “R” (R Development Core Team, 2010) using the add-on package “plm” (Croissant and Millo, 2008). The Cobb-Douglas and Translog production functions both have been estimated with three different estimators: fixed effects (FE), random effects (RE), and pooled OLS (i.e. ignoring the panel structure of the data). For both functional forms, a Hausman test shows that the RE model is inconsistent. As both individual and time specific effects are statistically significant, we can reject the pooled OLS model in favour of the two-way FE model.

The summary results of the FE estimations of both functional forms are presented in Tables 2 and 3, respectively.

[2] The value of agricultural production is deflated by the price index of agricultural production; the value of variable inputs is deflated by the price index of purchased goods and services for current agricultural production; and the value of the capital stock is deflated by the price index of purchased goods and services for investment.

Table 2: Results of parametric regression with Cobb-Douglas functional form

Regressor	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	5.8775	NA	NA	NA	
$\log(L)$	0.0707	0.0325	2.1721	0.0301	*
$\log(A)$	0.5520	0.0430	12.8304	< 2e-16	***
$\log(V)$	0.2982	0.0269	11.0853	< 2e-16	***
$\log(K)$	0.0147	0.0216	0.6795	0.4970	
$R^2 = 0.3017$					

Source: Own estimates based on Polish FADN data.

Table 3: Results of parametric regression with Translog functional form

Regressor	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	0.0353	NA	NA	NA	
$\log(L)$	-1.2981	0.5852	-2.2183	0.0267	*
$\log(A)$	-0.5942	0.6474	-0.9178	0.3589	
$\log(V)$	1.9985	0.6063	3.2964	0.0010	*
$\log(K)$	-0.1750	0.3903	-0.4485	0.6539	
$1/2 \log(L)^2$	-0.1137	0.0836	-1.3597	0.1742	
$\log(L)\log(A)$	-0.0219	0.0714	-0.3066	0.7592	
$\log(L)\log(V)$	0.0622	0.0654	0.9518	0.3414	
$\log(L)\log(K)$	0.0652	0.0393	1.6614	0.0969	.
$1/2 \log(A)^2$	0.1185	0.1061	1.1175	0.2640	
$\log(A)\log(V)$	0.1329	0.0730	1.8198	0.0691	.
$\log(A)\log(K)$	-0.0671	0.0439	-1.5295	0.1264	
$1/2 \log(V)^2$	-0.2411	0.0795	-3.0346	0.0025	**
$\log(V)\log(K)$	0.0344	0.0395	0.8711	0.3839	
$1/2 \log(K)^2$	0.0058	0.0361	0.1619	0.8714	
$R^2 = 0.3195$					

Source: Own estimates based on Polish FADN data.

4.2 Nonparametric Approach

We have estimated the relationship between the inputs and the output using non-parametric methods within the statistical software environment “R” (R Development Core Team, 2010) using the add-on package “np” (Hayfield and Racine, 2008). As we chose the two-ways fixed effect model for the parametric estimation, we use the IDs of the individual farms and the year as additional explanatory variables so that both models are comparable. We apply the non-parametric local-linear estimation method for both continuous and categorical explanatory variables described in Li and Racine (2004) and Racine and Li (2004). The second-order Epanechnikov kernel is used for continuous regressors (i.e. the four input variables) and the kernel proposed by Aitchison and Aitken (1976, p. 29) is used for unordered categorical explanatory variables (i.e. the IDs of the farms and the year). The bandwidths of the regressors are assumed to be fixed and are selected according to the expected Kullback-Leibler cross-validation criterion (Hurvich et al., 1998).

The bandwidths and the significance levels of the explanatory variables are presented in Table 4. The significance levels of the explanatory variables are obtained by bootstrapping using the methods proposed by Racine (1997) and Racine et al. (2006). While all inputs have a highly significant effect and the effects of the individual farms are significant at 5% level, the time effects are not statistically significant.

Table 4: Results of non-parametric regression model

Regressor	$\log(L)$	$\log(A)$	$\log(V)$	$\log(K)$	year	ID
Bandwidth	9064165	21471448	5204207	5857416	0.75	0.74
P-Value	< 2e-16 ***	< 2e-16 ***	< 2e-16 ***	< 2e-16 ***	0.6266	0.0476 *
$R^2 = 0.9600$						

Source: Own estimates based on Polish FADN data.

Furthermore, we apply the non-parametric test described in Hsiao et al. (2007) to check whether the functional forms used in the two parametric models (Cobb-Douglas and Translog) are consistent with the “true” relationship between the inputs and the output in our data set, i.e. are indeed linear in the regressors. Also this test is implemented in the “np” package (Hayfield and Racine, 2008). The results are shown in Table 5. The null hypotheses that the functional forms are consistent with the data are rejected for both the Cobb-Douglas and the Translog functional form. Hence, neither the Cobb-Douglas nor the Translog functional form is suitable in this application.

Table 5: Results of the Model Specification Tests

Parametric regression model	Test Statistic “Jn”	P Value
Cobb-Douglas	0.5886	0.0375
Translog	1.3277	< 2.22e-16

Source: Own estimates based on Polish FADN data.

4.3 Comparison of Parametric and Nonparametric Rresulst

In order to compare the results of the two parametric approaches and the non-parametric approaches to production function estimation, we compare the partial production elasticities of each of the four inputs as well as the elasticity of scale in figures 1, 2, 3, 4, and 5, respectively. While the sample means of all elasticities are rather similar for all three models, results for individual farms can be rather different. Interestingly, there is even no considerable correlation between the elasticities based on the Translog function and the elasticities based on the non-parametric production function. These results are rather similar to the results obtained by Henningsen and Kumbhakar

(2009), who used parametric and semi-parametric models to investigate the technical efficiency of Polish farms in the year 1994.

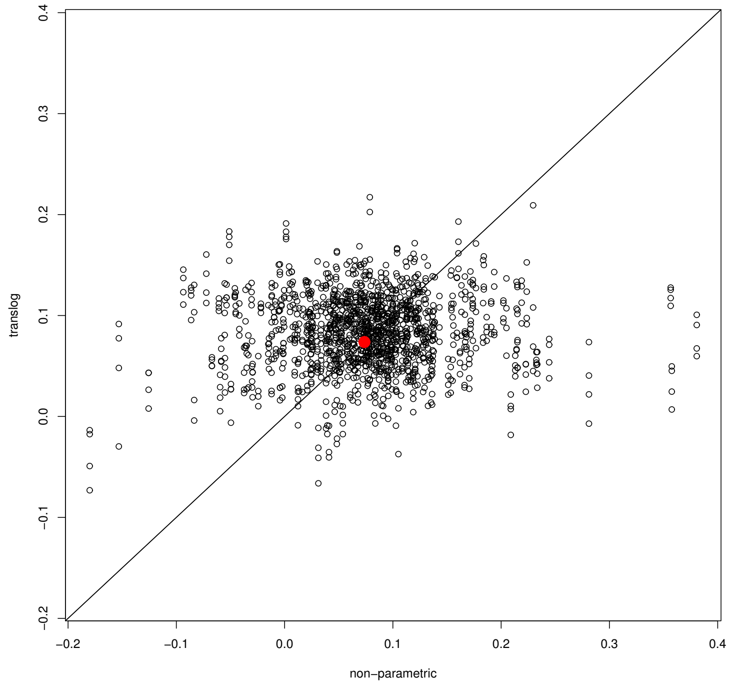

Figure 1: Partial output elasticities of labour based on CD (red dot), Translog and non-parametric production function

5 Conclusion

We propose to use non-parametric econometric methods in empirical production analysis. First, they can be applied to verify the functional form used in parametric estimations of production functions. Second, they can be directly used for estimating production functions without specifying a functional form and thus, avoiding possible misspecification errors. Our analysis shows that the two functional forms, which are most often used in empirical production analysis, i.e. the Cobb-Douglas and the Translog functional form, are both inconsistent with the "true" relationship between the inputs and the output in our data set. We solved this problem by using non-parametric regression. This approach delivers reasonable results, which are on average not too different from the results of the parametric estimations but many individual results are rather different.

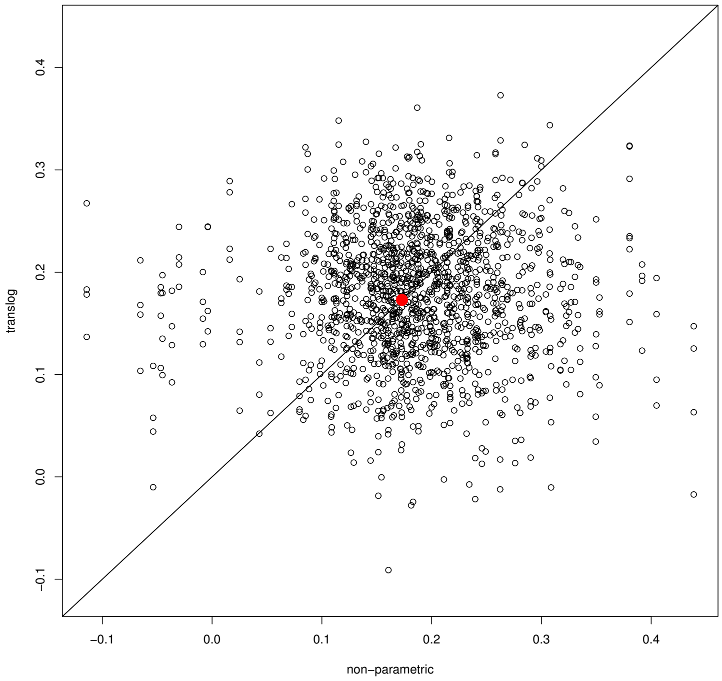

Figure 2: Partial output elasticities of land based on CD (red dot), Translog and non-parametric production function

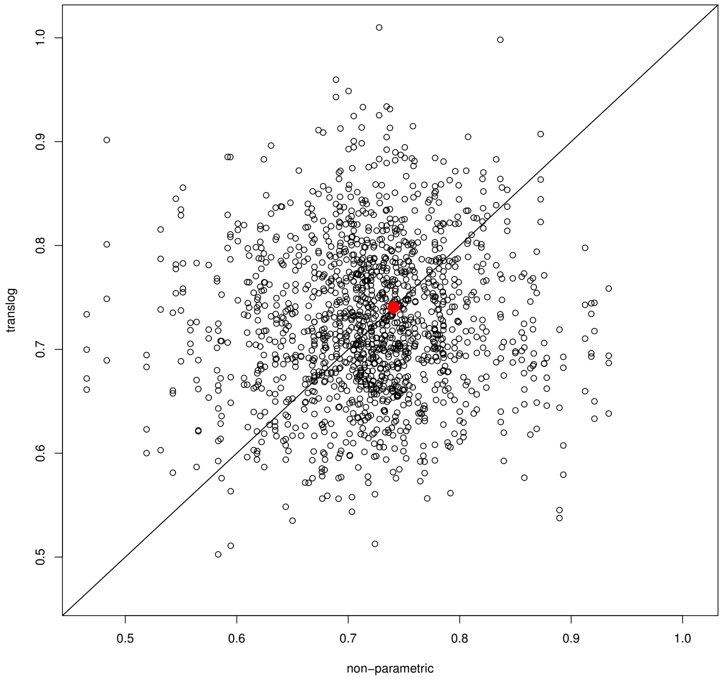

Figure 3: Partial output elasticities of variable inputs based on CD (red dot), Translog and non-parametric production function

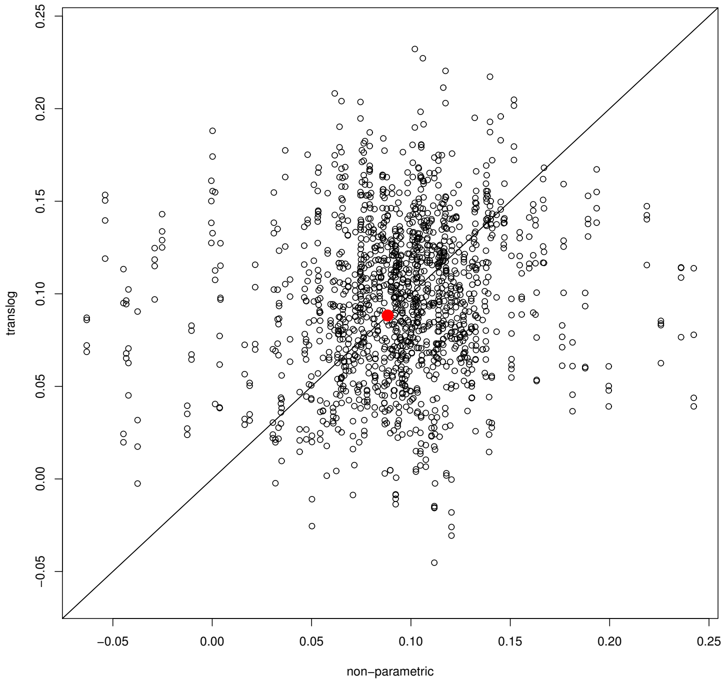

Figure 4: Partial output elasticities of capital based on CD (red dot), Translog and non-parametric production function

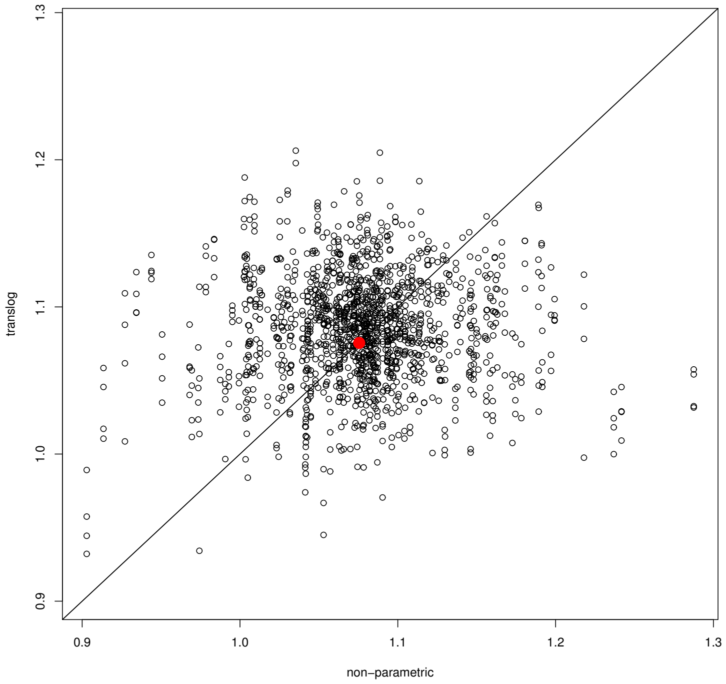

Figure 5: Elasticities of scale based on CD (red dot), Translog and non-parametric production function

References

Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63: 413–420.

Christensen, L. R., Jorgenson, D. W. and Lau, L. J. (1971). Conjugate duality and the transcendental logarithmic functions. *Econometrica* 39: 255–256.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software* 27: 1–43.

GUS (2008). Ceny w gospodarce narodowej w 2008 r. / Prices in the National Economy in 2008. *Glowny Urzad Statystyczny / Central Statistical Office of Poland* .

Hausman, J. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* 46: 1251–1271.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software* 27: 1–32.

Henningsen, A. and Kumbhakar, S. C. (2009). Semiparametric stochastic frontier analysis: An application to Polish farms during transition. Paper presented at the European Workshop on Efficiency and Productivity Analysis (EWEPA) in Pisa, Italy, June 24.

Hsiao, C., Li, Q. and Racine, J. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics* 140: 802–826.

Humphrey, T. (1997). Algebraic production functions and their uses before Cobb-Douglas. *Federal Reserve Bank of Richmond Economic Quarterly* 83: 51–83.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smooting parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B* 60: 271–293.

Li, Q. and Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica* 14: 485–512.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Racine, J. (1997). Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics* 15: 369–378.

Racine, J., Hart, J. and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews* 25: 523–544.

Racine, J. S. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119: 99–130.

Wicksteed, P. (1894). Essay on the Coordination of the Laws of Distribution,(1932 edition). *London: LSE* .