ON THE ECONOMETRIC MODELING OF NON-LINEAR RELATIONSHIPS: THE GUMBEL REGRESSION MODEL

Armando Sánchez Vargas *

Universidad Nacional Autónoma de México, Institute For Economics Research

José Márquez Estrada

Universidad Nacional Autónoma de México, Institute For Economics Research

(*Received June 14, 2015, accepted June 21, 2015.*)

Abstract

Nonlinear relationships among random variables often come out in all fields of economics. The academic debate on how to deal with nonlinearities, from a statistical point of view, has been centered in developing new estimation methods or modifying the specification of the classic linear econometric models. Here, we propose to face this issue by deriving population regression models from conditional distributions with genuine nonlinear conditional means. Such a mathematical procedure guarantees not only a consistent derivation of the conditional mean that gives rise to a nonlinear econometric model, but also a proper analysis of the causal effects among the involved economic variables (i.e., partial effects). Finally, we exemplify the workings of this approach by specifying a nonlinear and heteroskedastic econometric model based on the Gumbel distribution.

Resumen

Relaciones de tipo no lineal entre variables aparecen de manera frecuente en todos los campos de la economía. El debate académico sobre como modelar dichas relaciones, desde un punto de vista estadístico, ha estado centrado en el desarrollo de nuevos métodos de estimación o en la especificación de los componentes del modelo clásico de regresión lineal. En este artículo proponemos enfrentar dicho problema derivando los modelos de regresión poblacional a partir de funciones de densidad condicionales con medias no-lineales genuinas. Este procedimiento matemático garantiza no sólo una consistente derivación de la media condicional que da origen a un modelo econométrico no lineal, sino también un análisis más apropiado de los efectos causales entre las variables económicas involucradas (efectos marginales). Finalmente, se presenta un ejemplo del funcionamiento de dicho procedimiento mediante la especificación de un modelo econométrico no lineal y heteroscedastico que surge de una distribución Gumbel.

JEL Classification:C1; C51.

Keywords: Gumbel Distribution, Nonlinear Relationships, Conditional Mean, Population Regression Model.

* Circuito Mario de la Cueva s/n, Ciudad Universitaria, México D. F. C.P. 04510. Tel: 56 23 01 00 Ext. 42473 E-mail:asanchez@vt.edu

1. Introduction

The existence of nonlinear relationships among economic variables is always controversial. This issue becomes more complicated when applied researchers find nonlinearities that the economic theory does not predict or estimate linear models that do not properly fit the data. Here, we argue that, from a statistical point of view, we can avoid controversies by using conditional probability densities $f(y|x)$ as the basis to derive the nonlinear conditional means $E(y \mid x) = m(x, \theta)$ that give rise to reliable econometric models $y = m(x, \theta)+e$, rather than assuming the standard functional forms of the conditional mean suggested in the econometrics textbooks. That is, we propose starting the specification of the econometric model by assuming a proper conditional density, for the data on hand, and derive the associated regression and skedastic functions from it by taking the expected value of the explained variable y given the set of explanatory variables x.

In other words, we show how to use a statistical procedure to derive proper econometric models that capture genuine nonlinear relationships. To do so, we assume suitable conditional probability distributions that give rise to nonlinear regression functions (Spanos, 1986). We exemplify this approach by deriving a population regression model that can be useful to analyze a nonlinear relationship between economic variables. Specifically, we use a Gumbel conditional distribution as the basis to derive the nonlinear regression curves that might describe such type of phenomena. The Gumbel regression model allows us to briefly illustrate two interesting facts. First, we show how a nonlinear relationship might be well represented by an exponential distribution and its associated regression model (Gumbel 1960); which has a nonlinear conditional mean and a heteroskedastic conditional variance. Second, we show that a non linear model like the Gumbel regression exhibits changing partial effects of the explanatory variables over the entire distribution of the explained variable, which is not the case in a normal-linear model. These facts might be useful to elucidate controversial economic arguments when the empirical data exhibit nonlinearities.

This paper is structured as follows. The second section briefly discusses the general statistical approach to derive linear or non-linear econometric models in a stochastic setting. In the third section, we exemplify the statistical approach by discussing the specification, estimation, and validation of the Gumbel regression model. In the last section, we make some remarks on the implications of the employed approach.

2. Deriving Nonlinear Regression Models

In the context of a modern approach to econometrics any linear or nonlinear model can be specified by making assumptions on two components: 1) the population regression model, and 2) the sampling model (Wooldrige, 2010). The first assumption refers to the functional form of the conditional mean that describes the stochastic relationship between y and x. The second assumption refers to the probabilistic behavior of the sample. Here, we only deal with the derivation of the population regression model that gives rise to the nonlinear relationships among a set of economic variables so, for the sake of simplicity, we assume that we have an independent and identically distributed sample *(iid)* in the rest of the paper.

Let us define the population regression model by assuming that any stochastic variable can be decomposed, by definition, into two parts: a conditional expectation $E(y \mid x)$ and an *iid* error term e. In other words, we can always write any explained variable y as its conditional expectation $E(y|x)$ plus an error term or disturbance term e that has conditional mean zero

$$\mathrm{y} = E\left(y|x\right) + e \text{ with } \mathrm{eiid}(0, \sigma^2), \text{and} \tag{1}$$

$$E\left(e|x\right) = 0, \tag{2}$$

Equation (2) implies that the unconditional error is a random variable with zero mean and is not correlated with each of the explanatory variables and any functions of them. It is worth mentioning that these two equations also imply a set of testable statistical assumptions, while working with real data.

Under the assumption of a random sample, equation (1) implies that the applied econometrician needs to propose a specific functional form for the conditional mean $E(y \mid x)$, which is almost always assumed to be a linear equation $E(y \mid x) = \beta_0 + \beta_1 x$. However, when dealing with a nonlinear problem, either in parameters or in variables, the textbook approach does not suggest a clear procedure on how to obtain the genuine nonlinear conditional mean $E(y \mid x) = m(x, \theta)$ that shapes the econometric model.

In order to fill this gap we propose to assume a conditional distribution, based on the empirical distribution of the data on hand, and derive its regression function; rather than assuming an arbitrary functional form of the conditional mean in equation (1), (Spanos, 1986). To illustrate the workings of this approach, we first derive the typical normal-linear econometric model, not only based on equations (1) and (2), but also on a normal conditional density that gives rise to the conditional mean that defines equation (1). So, let us assume that the data is described by a conditional normal distribution of y given x and that the variances of the involved normal random variables y and x are constant

$$f\left(y|x\right) = \frac{1}{\sqrt[2]{2\pi\sigma^2}} e^{-\left\{\frac{(y-\beta_0-\beta_1 x)^2}{2\sigma^2}\right\}}. \tag{3}$$

Then, we proceed to compute the conditional expectation of y given x based on such conditional density function

$$E\left(y|x\right) = \int_{-\infty}^{\infty} y f\left(y|x\right) dx, \tag{4}$$

which gives rise to the following genuine linear conditional mean:

$$E\left(y|x\right) = \beta_0 + \beta_1 \mathrm{x} \tag{5}$$

with the following statistical parameterization

$$\beta_0 = E\left(y\right) - \beta_1 E(x) \text{ and } \beta_1 = \frac{E\left(x - E\left(x\right)\right)\left(y - E\left(y\right)\right)}{E(x - E\left(x\right))^2}$$

Once we have mathematically obtained the conditional mean in equation (5), we are in a position to specify the normal-linear-homoskedastic regression model by using equations (1) and (2)

$$y = \beta_0 + \beta_1 \mathrm{x} + \mathrm{e} \quad \mathrm{e} \sim Niid(0, \sigma^2) \; and \tag{6}$$

$$E(e \mid x) = 0 \tag{7}$$

where e is a Normal, independent, and identically Distributed (Niid) error.

Sometimes, we are interested in a particular function called partial effect (*i.e.,* marginal effect) that shows the response of the conditional mean to a unit change in one of the explanatory variables. Equation (8) below shows that, in the normal-linear regression model, the effect of x on y is constant

$$\frac{\partial \mathrm{E(y|x)}}{\mathrm{dx}} = \beta_1 \tag{8}$$

Note that, when the probabilistic features of the data are not compatible with the normality assumption, the conditional mean will not always be a linear function. This implies that the partial effect, in equation (8), will not necessarily be a constant in models with different distributive assumptions. Thus, an advantage of assuming conditional densities, rather than the functional forms of the mean, is that we can use densities with genuine nonlinear means; which we can choose by assessing the empirical features of the data on hand.

In the next section we show how we can use this setup to derive other econometric model by incorporating different conditional distributions, where the mean or average causal effect of the explanatory variable on the explained variable will not be linear. Specifically, we change the assumption of normality not only for the partial densities of y and x, but also for the conditional density; so that we are able to obtain a valid nonlinear population regression model.

3. The Gumbel Linear Regression Model

Here, we exemplify the workings of the previous approach by specifying the Gumbel regression model, although we can easily use other conditional densities to specify other population regression models. To do this, we first assume that a Gumbel joint density is a good representation of the joint stochastic behavior of y and x. Then, we derive the conditional density of y given x and derive its nonlinear conditional expectation by integration. Finally, we embed our derived conditional expectation in our econometric setup, given by equations (1) and (2), to end up with a proper nonlinear econometric model with heterogeneous partial effects. In what follows we describe such statistical procedure step by step.

A. Observational Data and Model Specification

A preliminary step to specify a proper conditional model that accounts for nonlinear relationships, among a set of random variables, is to discuss the statistical properties of such variables. That is, in selecting a proper econometric model we should take into account not only theoretical issues, but also all the statistical systematic information in the data (Spanos 1986). In fact, a brief analysis, of different types of graphs, might reveal the empirical

distribution that could be a good assumption for the data on hand. Kernel estimates of the univariate empirical densities (Silverman, 1998) can also be useful to assess departures from the normality assumption. We can get more information about the underlying joint density of the data by looking at the kernel estimate of the empirical joint distribution and the probability contour plot with the potential empirical regression curves. In other words, we can anticipate the presence of a potential nonlinear conditional distribution and its associated regression function by using a set of graphical tools.

B. Model Specification

Let us assume that a good empirical representation of the distribution that governs the joint behavior of the data on hand is a Gumbel distribution (Gumbel, 1960: Castillo, 2005). Thus, in what follows we can specify the Gumbel regression model that implies a nonlinear regression curve with non constant partial effects (Kotz, *et al.* 2000).

We start with the bivariate Gumbel distribution function, which is defined for positive values of the involved random variables:

$$\begin{aligned} \mathrm{F}\left(\mathrm{x},\mathrm{y}\right) = 1 - e^{-x} - e^{-y} + e^{-(\mathrm{x}+\mathrm{y}+\delta\mathrm{xy})}, 0 \leq \mathrm{x}, 0 \leq \\ \mathrm{y}, (0 \leq \delta \leq 1) \end{aligned} \tag{9}$$

where δ is the parameter that describes the probabilistic dependence between the two random variables y and x, which is limited to take values between 0 and 1. The joint probability density function of the Gumbel model can be derived by differentiating equation (9)

$$f\left(x,y\right) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{10}$$

Then, the conditional density function of y given x can be derived by dividing the joint density by the marginal density of x

$$f\left(y|x\right) = \frac{f(x,y)}{f_x(x)} = \frac{\left[\left(1+\delta y\right)\left(1+\delta x\right)-\delta\right]e^{-(y+x+\delta xy)}}{e^{-x}}. \tag{11}$$

The conditional expectation of y given x is given by:

$$E\left(y|x\right) = \int_0^\infty y f\left(y|x\right) dy = \frac{1+\delta+\delta x}{\left(1+\delta x\right)^2} \tag{12}$$

Similarly, we could also get the conditional expectation of x given y, $E[x \mid y]$, by using the equivalent formulae, given the symmetric nature of the Gumbel distribution. Now, we are in a position to specify the Gumbel regression model. According to equation (1), the stochastic variable y can be decomposed, by definition, into two components: a non linear conditional expectation and an error term e as follows

$$y = \frac{1+\delta+\delta x}{\left(1+\delta x\right)^2} + e \sim \text{where,} \quad \mathrm{e} \sim Diid(0,\sigma^2), \tag{13}$$

$$E(e_i/x_i) = 0, \tag{14}$$

It is worth mentioning that the distribution $D = f_e(e)$ of the error term e has a closed but complex form given by

$$\begin{aligned} f_e(e) = \frac{\delta e}{2} + \sqrt[2]{\frac{\pi}{\delta}} \exp\left(\frac{\delta e^2}{4}\right) \operatorname{erf}\left(\frac{\sqrt[2]{\delta} e}{2}\right) \\ \left\{(1 - \delta + \delta e) \exp(-e) \frac{-e}{2} \left(\left(1 - \frac{\delta e^2}{2}\right)\right)\right\} \end{aligned}, \tag{15}$$

for $0 < e < \infty$.

Note that the analytical form of the marginal distribution in equation (17) includes the so- called error function:$\operatorname{erf}(x) = \frac{2}{\pi} \int_0^x \exp(-t^2) dt$ (Naradajah, 2008). Next, we can derive the conditional variance and partial effect for the Gumbel model:

$$\sigma^2(y|x) = \frac{(1 + \delta + \delta x)^2 - 2\delta^2}{(1 + \delta x)^4} \tag{16}$$

$$\frac{dE(y|x)}{dx} = -\frac{\delta(1 + \delta + \delta x)}{(1 + \delta x)^3}. \tag{17}$$

Equation (13) suggests that the mean or average causal effect of the explanatory variable on the dependent variable is not linear. Moreover, the conditional variance in equation (16) is heteroskedastic. Besides, the negative marginal effect in equation (17) is heterogeneous and decreasing. Therefore, we can see that the model in equation (13) is completely different to the model in equation (6), since the Gumbel regression model does not imply a constant effect of the explanatory variable over the entire density of the dependent variable. The economic meaning of a negative nonlinear relationship, in this context, is that the mean or average causal effect of x on y is negative and decreasing. That is, the average value of y does not change at a constant rate as x changes, which means that we can have heterogeneous partial effects (changing partial effects). Even more important is the fact that the suggested model implies that there might be a decreasing tradeoff between y and x, which s clearly associated to the nonlinear nature of the Gumbel model.

C. Estimation Method

In the previous section we + propose a specification of a non linear and heteroskedatic regression model derived from an Gumbel distribution. Now, we need to estimate the value of the dependence parameter δ that determines not only the conditional mean, but also the conditional variance as suggested by equations (13) and (16).

As we discuss above, the Gumbel regression model takes the form:

$$y = \frac{1 + \delta + \delta x}{(1 + \delta x)^2} + e \text{ with } e \sim Diid(0, \sigma^2)$$

$$\sigma^2(y|x) = \frac{(1+\delta+\delta x)^2 - 2\delta^2}{(1+\delta x)^4}$$

Where $e \sim Diid(0, \sigma^2)$ and σ^2 is the conditional variance.

From these equations, we can see that the conditional mean of the Gumbel model is not linear in its conditioning variable x and unknown parameter δ, and the conditional variance is not homoskedastic. Under the random sample assumption, the maximum likelihood estimate of δ can be obtained by solving the following equation ((Kotz, *et al.* 2000):

$$\sum_{i=1}^{n}\left(\frac{x_i + y_i - 1 + 2\delta x_i y_i}{1 + (x_i + y_i - 1)\,\delta + x_i + y\delta^2}\right) = \sum_{i=1}^{n} x_i y_i \tag{18}$$

On the other hand, a moment estimator of δ (Hosking, 1985) can also be obtained as the solution of the following equation:

$$\frac{1}{\delta} e^{\frac{1}{\delta}} Ei\left(\frac{1}{\delta}\right) = 1 - \rho \tag{19}$$

Thus, the δ parameter has a close relationship with the classical correlation coefficient that stands as:

$$\rho = \frac{e^{\frac{1}{\delta}}}{\delta} - E(-\delta^{-1}) - 1 \tag{20}$$

Where E is the well-known exponential integral. So, the correlation parameter is given by:

$$\rho = -1 + \int_0^{\infty} \frac{e^{-y} dy}{1 + \delta y} \tag{21}$$

When $\delta = 0$, the correlation parameter ρ is equal to zero and we have a model where the random variables are independent and the bivariate density splits into the product of its two marginal densities

$$f\,(x, y) = f_x(x) f_y(y)$$

When $\delta = 1$, the association parameter ρ is equal to -0.4036 and reaches its lower limit. So, this model is only suitable for representing a joint density of two correlated Gumbel distributed variables whose correlation parameter takes values in $-0.4 \leq \rho \leq 0$ (Tiago, 1961)

D) Misspecification Tests

In order to ensure the statistical validity of our model assumptions in relation to the real data, we can define some potential misspecification tests for the Gumbel regression model. The set of tests we discuss will allow us to ensure that there are no departures from the underlying assumptions of the Gumbel model while working with real data (Spanos, 2006; Wang, 2005).

The potential misspecification tests that can be applied to the regression model can be based on the following F type tests:

a) Additional Non-Linearity in the Conditional Mean

To test for the presence of additional non-linearities in the conditional mean we can test if $\alpha_2 = 0$ in the following auxiliary regression:

$$\mathrm{y} = \alpha_0 + \alpha_1 \hat{y}_1 + \alpha_2 \hat{y}_2 + \mathrm{u_i} \tag{23}$$

Where $\hat{y}$ is a vector of the Gumbel model fitted values. Furthermore, we can also expect that $\gamma_1 = 1$ if the pre-specified is the correct model. The potential misspecification tests that can be applied to the regression model can be based on the following F type tests:

b) Trend in conditional mean

To test for the presence of additional non-linearities, like a linear trend in the conditional mean, we can test if $y_2 = 0$ in the following regression:

$$\mathrm{y} = \gamma_0 + \gamma_1 \hat{y} + \gamma_2 \mathrm{t} + \mathrm{u} \tag{24}$$

Where $\hat{y}$ are the Gumbel model fitted values. Furthermore, we can also expect that $\gamma_1 = 1$ if the specified equation is the correct model.

4. Concluding Remarks

Here, we propose to use conditional probability densities as the basis to derive nonlinear the conditional means that give rise to reliable econometric models, rather than assuming the standard functional forms of such mean suggested in the econometrics textbook. We show a procedure to derive econometric models that capture genuine nonlinear relationships by using empirical suitable conditional probability distributions that give rise to different regression functions. We illustrate this approach by deriving a regression model that might be useful to analyze a nonlinear relationship. Specifically, we use a Gumbel conditional distribution as the basis to derive the nonlinear regression curve that can be suitable to analyze highly volatile economic data. We show that a non linear model like the Gumbel regression exhibits changing partial effects of the explanatory variables over the entire distribution of the explained variable, which is not the case for a normal-linear model.

References

Castillo, E., Hadi, A.S., Balakrishnan, N., Sarabia, J.M. (2005). Extreme Value and Related Models with Applications in Engineering and Science, Wiley.

Gumbel, E. J. (1960). Bivariate Exponential Densities, *Journal of the American Statistical Association* 55(292), pp. 698-707.

Hosking, J. R. M., J. R. Wallis and E. F. Wood (1985). Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics* 27, pp. 251-261.

Kotz, S., Balakrishnan, N. and Johnson, N. (2000). Continuous Multivariate Distributions. Vol. 1. Models and Applications, Wiley, New York.

Naradajah, S., Kotz, S. (2008). Exact Distribution of the Linear Distribution of p Gumbel Random Variables. *International Journal of Computer Mathemathics.* 85, pp. 1355-1362.

Silverman, B. W. (1998). Density Estimation for Statistics and Data Analysis, Boca Raton: Chapman & Hall/CRC.

Spanos, A. (1986). Statistical Foundations of Econometric Modeling, Cambridge University Press, UK.

Spanos, A. (2006). Where Do Statistical Models Come From? Revisiting the Problem of Specification, Lecture Notes-Monograph Series, 49, pp. 98-119.

Tiago de Oliveira, J. (1961). Decision Results for the Parameters of the Extreme Value (Gumbel) Distribution based on the Mean and the Standard Deviation. Trabajos de estadística y estadística operativa. 14, pp. 61-81.

Wooldrige, J., (2010). Econometric Analysis of Cross Section and Panel Data, MIT press.

Wang, J. Z. (1995). Selection of the k Largest order Statistics for the Domain of Attraction of the Gumbel Distribution. *Journal of the American Statistical Association* 90, pp. 1055-1061.