

A Semiparametric Model for Fractional Responses with Panel Data: An Application to Intra-Industry Trade

Isabel Proença^a and Horácio Faustino^b

^aISEG, Universidade de Lisboa and CEMAPRE

^bISEG, Universidade de Lisboa and SOCIUS



U LISBOA

UNIVERSIDADE
DE LISBOA

November 2015

Plan of the talk

- Motivation
- The model
- Empirical Application
 - The data and variables
 - Estimation results
- Final remarks

Motivation

- Econometricians and statistics often need to model **fractional responses**
- Examples are:
 - indexes:
 - Intra-industry trade
 - The American Customer Satisfaction Index (though usually 0-100)
 - Proportions:
 - Participation rates on voluntary pension plans
 - Capital structure
 - Student failure rate
 - Proportion of income spent on medicines

Motivation


- The nature of the variable: $0 \leq y \leq 1$
- Traditional approach:
Use the Logit transformation $z = \ln\left(\frac{y}{1-y}\right)$
- Drawbacks
 - Cannot be used when $y = 0$ or $y = 1$
 - It is relatively easy to model $E(z | x)$
but the aim is to model $E(y | x)$
 - not obvious to obtain $E(y | x)$ from $E(z | x)$

Motivation

- Alternative Approach: Pseudo-Maximum Likelihood
 - Gourieroux, Monfort, and Trognon (1984)
 - Use a likelihood function that is not based on the true distribution of y but
 - has the same conditional mean
 - estimate variances robustly to misspecification
- Papke and Wooldridge (1996)

Application to 401 (K) plan participation rates

Motivation

- Panel data: control for individual unobserved heterogeneity dependent on the explanatory variables
- Linear models:
simple variable transformations to eliminate the unobserved heterogeneity term
- Nonlinear models:
Conditional likelihood  not obvious in the context of pseudo-maximum likelihood

Motivation

- Need to specify the relation between the unobserved heterogeneity and explanatory variables
- Usual Approach: Parametric linear relation based on Mundlak (1978)
- Papke and Wooldridge (2008): Application to test pass rates
- The aim of this work:
 - *To model nonparametrically the relation between the unobserved heterogeneity and the explanatory variables*
 - *To extend to fractional data the approaches of Lombardía and Sperlich (2012) and Proença, Sperlich and Savaşçı (2015)*

The Model

- Fractional responses: $0 \leq y_{it} \leq 1$
- Pseudo-Maximum likelihood Approach

$$E(y_{it} | \mathbf{x}_{it}, \eta_i) = \Phi(\mathbf{x}_{it} \boldsymbol{\beta} + \eta_i), t = 1, \dots, T_i$$

- Random Effects Probit

$$\eta_i = \alpha + a_i \text{ with } a_i | (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}) \sim N(0, \sigma_a^2)$$

$$E(y_{it} | \mathbf{x}_{it}) = \Phi \left[(\alpha + \mathbf{x}_{it} \boldsymbol{\beta}) \frac{1}{\sqrt{1 + \sigma_a^2}} \right], t = 1, \dots, T_i$$

The Model

- Heterogeneity dependent from the explanatory variables:

Mundlak (1978) , Papke and Wooldridge (2008)

$$\eta_i = \alpha + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i \quad \text{with } a_i | (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}) \sim N(0, \sigma_a^2)$$

$$\bar{\mathbf{x}}_i = (1/T_i) \sum_{t=1}^{T_i} \mathbf{x}_{it}$$

$\boldsymbol{\xi}$ vector of unknown coefficients

$$E(y_{it} | \mathbf{x}_{it}) = \Phi \left[(\alpha + \mathbf{x}_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi}) \frac{1}{\sqrt{1 + \sigma_a^2}} \right], t = 1, \dots, T_i$$

The Model

Semiparametric approach:

Semi-mixed effects Model of Lombardía and Sperlich (2012)

$$\eta_i = \psi(\mathbf{w}_i) + a_i \quad \text{with} \quad E(a_i \mid \mathbf{w}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}) = E(a_i \mid \mathbf{w}_i) = 0$$

$\psi(\bullet)$ unknown function

\mathbf{w}_i vector of proxy variables time invariant and continuous



SINCE 1911

U LISBOA

UNIVERSIDADE
DE LISBOA

November 2015

The Model

- Problems:
 - $\psi(\cdot)$ estimated nonparametrically → curse of dimensionality



Additive Model

$$\eta_i = \alpha + G_1(w_{i1}) + G_2(w_{i2}) + \dots + G_p(w_{ip}) + a_i$$

$$a_i \mid (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}) \sim N(0, \sigma_a^2)$$

$G_j(\cdot)$ - unknown functions

- Choice of the Proxies

The Model

In this work:

$$\eta_i = \alpha + G_1(\bar{x}_{i1}) + G_2(\bar{x}_{i2}) + \dots + G_k(\bar{x}_{ik}) + a_i$$

The final Model:

$$E(y_{it} | \mathbf{x}_{it}) = \Phi \left[\left\{ \alpha + \mathbf{x}_{it} \boldsymbol{\beta} + G_1(\bar{x}_{i1}) + \dots + G_k(\bar{x}_{ik}) \right\} \frac{1}{\sqrt{1 + \sigma_a^2}} \right], t = 1, \dots, T_i$$

Estimation:

maximum quasi-likelihood with penalized splines – Wood(2006)

Estimation

- **Advantages of the Estimation with *Penalized Splines***
 - Using a Bayesian approach it is possible to calculate confidence intervals for the functions $G_j(\bullet)$ $j=1,2,\dots,k$
 - The usual inference for β applies
 - Because estimation results from the maximization of a penalized likelihood the generalization to penalized pseudo-likelihoods is simple
 - It is implemented in R in the *package mgcv* of Wood(2006)

Estimation - Splines

- represent the unknown functions $G_j(\bullet)$, using known basis functions, $b_{lj}(z_{ij})$, such that:

$$G_j(z_{ij}) = \sum_{l=1}^{L_j} \gamma_{lj} b_{lj}(z_{ij})$$

- γ_{lj} are unknown parameters to be estimated
- a cubic spline basis for knots z_l^* $l=1, \dots, L$

$$G(z) = \frac{z_{l+1}^* - z}{z_{l+1}^* - z_l^*} \gamma_l^* + \frac{z - z_l^*}{z_{l+1}^* - z_l^*} \gamma_{l+1}^* + \left[\frac{(z_{l+1}^* - z)^3}{z_{l+1}^* - z_l^*} - (z_{l+1}^* - z_l^*)(z_{l+1}^* - z) \right] \frac{1}{6} \gamma_l^+ \\ + \left[\frac{(z - z_l^*)^3}{z_{l+1}^* - z_l^*} - (z_{l+1}^* - z_l^*)(z - z_l^*) \right] \frac{1}{6} \gamma_{l+1}^+ \quad \text{if } z_l^* \leq z \leq z_{l+1}^*$$

Estimation - Splines

- Penalty

$$\lambda \int h''(z)^2 dz = \lambda \boldsymbol{\gamma}_j^T \mathbf{S}_j^* \boldsymbol{\gamma}_j$$

- Coefficients: $\boldsymbol{\beta}_{all} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')$

- Penalized log-likelihood

$$l_p(\boldsymbol{\beta}_{all}, \boldsymbol{\nu}) = l(\boldsymbol{\beta}_{all}, \boldsymbol{\nu}) - \frac{1}{2} \sum_j \lambda_j \boldsymbol{\beta}'_{all} \mathbf{S}_j \boldsymbol{\beta}_{all}$$

Empirical Application

The data and variables

- **Data**

- unbalanced panel of 38 countries:

(Angola, Austria, Belgium and Luxembourg, Brazil, Bulgaria, Cape Verde, China, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Guiné Bissau, Hungary, India, Ireland, Italy, Latvia, Lithuania, Malta, Moldova, Mozambique, Netherlands, Poland, Portugal, Romania, Russia, São Tomé and Príncipe, Slovakia, Slovenia, Spain, Sweden, UK, Ukraine and USA)

- Time period: 1995 to 2006.



SINCE 1911

U LISBOA

UNIVERSIDADE
DE LISBOA

November 2015

Empirical Application

- **Dependent Variables**

IIT_{it} – total intra-industry trade index of Portugal with country i at period t

$VIIT_{it}$ – vertical intra-industry trade index of Portugal with country i at period t

$HIIT_{it}$ – horizontal intra-industry trade index of Portugal with country i at period t

The IIT index is mainly of the VIIT type

Means: IIT= 0.171; HIIT= 0.034; VIIT= 0.137

$$IIT_{it} = \frac{\sum_{j=1}^J (X_{ijt} + M_{ijt}) - \sum_{j=1}^J |X_{ijt} - M_{ijt}|}{\sum_{j=1}^J (X_{ijt} + M_{ijt})}$$

Empirical Application

- **Explanatory Variables**

DYPC - difference between the per-capita GDP of Portugal and the GDP of the respective trading partner

DPOP - difference between the population of Portugal and foreign country

DCEE - proxy for differences in physical capital endowments equal to the difference in electric power consumption (Kwh per capita) between Portugal and the foreign partner

LDIST - the logarithm of geographic distance, measured in kilometers, between the capital cities of the trading partners



Empirical Application

- **Variables (cont.)**

TIY - the weight of the trade imbalance in the GDP for each trading partner

EU15 - Dummy variable assuming the value 1 if the trading partner is a member of the EU15

BRICS - Dummy variable assuming the value 1 if the country is Brazil, Russia, India or China

PALOPS - Dummy variable assuming the value 1 if the trading partner is an African country with Portuguese as its official language



SINCE 1911

U LISBOA

UNIVERSIDADE
DE LISBOA

Applicable Semiparametrics, Berlin
October 2013

Empirical Application

Estimation Results - IIT

	Parametric		Semiparametric			Parametric		Semiparametric	
	Reg. IIT		Reg. IIT			Reg. IIT		Reg. IIT	
	coeff.	p-val.	coeff.	p-val.		coeff.	p-val.	coeff.	p-val.
Intercept	0.9710	0.561	0.9451	0.493	t	0.0296	0.000	0.0305	0.000
DYPC	0.0368	0.001	0.0388	0.000	MYPCK	0.0847	0.000	NP	
DPOP	0.0000	0.712	0.0000	0.417	MPOPK	0.0000	0.911		
DCEE	-0.0942	0.055	-0.0979	0.046	MCEEK	-0.1174	0.057	NP	
LDIST	-0.9787	0.056	-0.7474	0.068	MTIY	-0.2736	0.835	NP	
TIY	-1.0518	0.005	-1.0566	0.005	N	329		329	
BRICS	0.2572	0.536	0.3579	0.273	AIC	99.6		93.6	
PALOPS	0.0525	0.900	0.2102	0.532	logLik	-33.8		-28.8	
EU15	0.0706	0.778	0.2488	0.250	SD rand	0.3759		0.2835	

Empirical Application

Estimation Results - VIIT

	Parametric Reg. VIIT		Semiparametric Reg. VIIT		Parametric Reg. VIIT		Semiparametric Reg. VIIT		
	coeff.	p-val.	coeff.	p-val.	coeff.	p-val.	coeff.	p-val.	
Intercept	0.1469	0.927	0.4040	0.773	t	0.0186	0.010	0.0197	0.005
DYPC	0.0255	0.027	0.0277	0.013	MYPCK	0.0714	0.001	NP	
DPOP	0.0000	0.737	0.0000	0.433	MPOPK	0.0000	0.956		
DCEE	-0.0228	0.674	-0.0267	0.623	MCEEK	-0.0505	0.429	NP	
LDIST	-0.7424	0.125	-0.5551	0.176	MTIY	0.4726	0.714	NP	
TIY	-1.2523	0.002	-1.2555	0.002	N	329		329	
BRICS	0.1712	0.668	0.2488	0.455	AIC	152.4		150.4	
PALOPS	-0.0416	0.918	0.0861	0.802	logLik	-60.2		-57.2	
EU15	-0.0242	0.920	0.1138	0.600	SD rand	0.3583		0.2898	

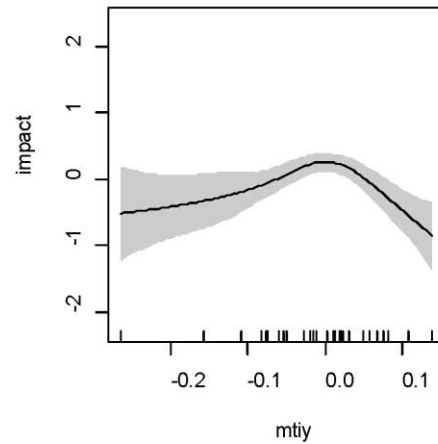
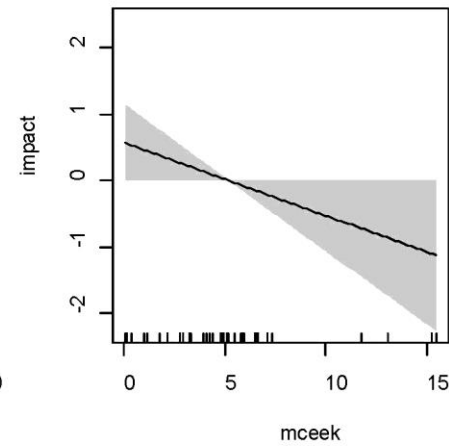
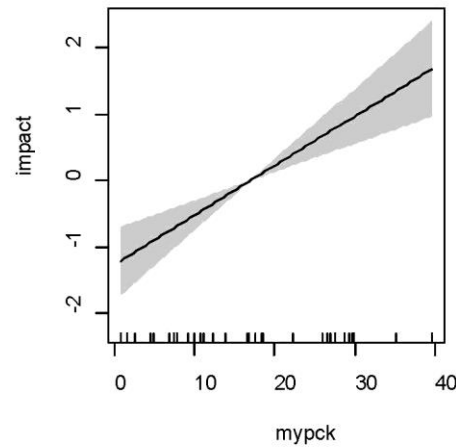
Empirical Application

Estimation Results - HIIT

	Parametric Reg. HIIT		Semiparametric Reg. HIIT			Parametric Reg. HIIT		Semiparametric Reg. HIIT	
	coeff.	p-val.	coeff.	p-val.		coeff.	p-val.	coeff.	p-val.
Intercept	-0.0890	0.942	-0.8307	0.347	t	0.0394	0.000	0.0409	0.000
DYPC	0.0452	0.015	0.0502	0.009	MYPCK	0.0726	0.002	NP	
DPOP	0.0000	0.636	0.0000	0.839	MPOPK	0.0000	0.732		
DCEE	-0.1966	0.015	-0.1945	0.019	MCEEK	-0.1902	0.034	NP	
LDIST	-0.8708	0.017	-0.5989	0.021	MTIY	-3.0018	0.025	NP	
TIY	0.5489	0.446	0.5766	0.445	N	329		329	
BRICS	0.4571	0.170	0.4216	0.098	AIC	469.3		481.8	
PALOPS	-0.0541	0.894	0.0770	0.825	logLik	-218.7		-222.9	
EU15	0.3494	0.062	0.5456	0.001	SD rand	0.2445		0.1382	

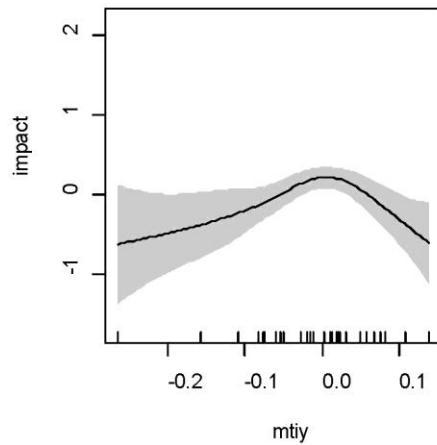
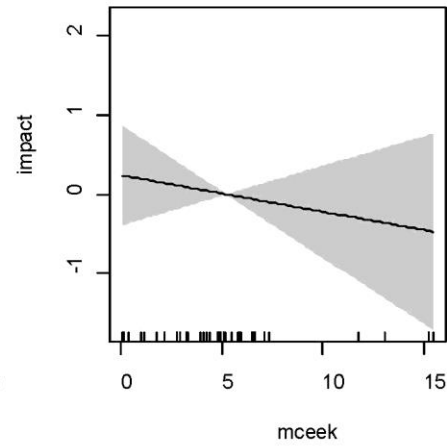
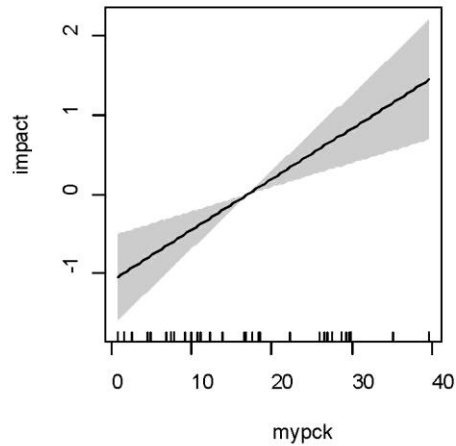
EMPIRICAL APPLICATION

- IIT



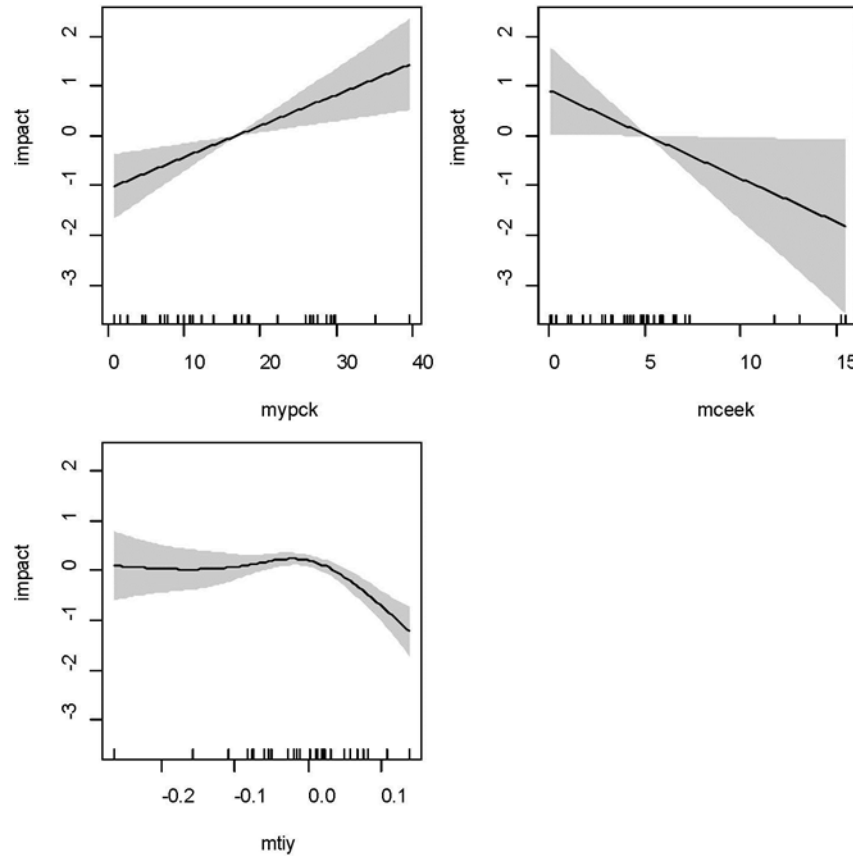
EMPIRICAL APPLICATION

- VIIT



EMPIRICAL APPLICATION

- **HIIT**



Empirical Application

- Parametric vs Semiparametric regressions
 - The impact of the mean of TIY is nonlinear inducing misspecification of the parametric model to control for the unobserved heterogeneity.
 - The semiparametric regression is better in terms of goodness of fit measures
 - The estimated variance of the random effect is significantly smaller in the semiparametric regression

Empirical Application

- Parametric vs Semiparametric regressions
 - The coefficient of log of distance is smaller in absolute value for the semiparametric fit (around 25% less)
 - The regional effect is stronger in the semiparametric fit
 - EU15 is statistically significant at 1%
 - BRICS is statistically significant at 10%

Final Remarks

- The **semiparametric mixed effects model** of Lombardía and Sperlich (2012) and Proença, Sperlich and Savaşçı (2013) is applied to **fractional responses**
- The new semiparametric approach is more flexible to control for dependency between the explanatory variables and the unobserved heterogeneity term
- Estimation is easy to do using the package `mgcv` of R
- It was applied to model the intra-trade indexes between Portugal and a set of countries

Final Remarks

- The new semiparametric procedure proved to be useful in
 - Goodness of fit
 - Improving precision in estimation of the majority of the coefficients depicting significant effects for region that were not present in the parametric fit
 - Detecting some nonlinearities in the control of the unobserved heterogeneity term whether the parametric fit assumed linearity
 - Overall, even if there are no relevant differences in estimates from the parametric fit, the semiparametric alternative provides robustness to the parametric results



SINCE 1911

U LISBOA

UNIVERSIDADE
DE LISBOA

November 2015