Econometric Techniques and Estimated Models *9 (continues in the website)

This text details the different statistical techniques used in the analysis, such as logistic regression, applied to discrete variables for example in the case indicators of poverty or access to infrastructure. We also detail the difference-difference estimator and the stepwise methodology applied to discrete models, as well as continuous endogenous variables models (ex: linear schooling and log-linear income equations).

Multivariate Analysis – Methodology

The bivariate analysis captures the role played by each attribute considered separately in poverty analysis. That is, we do not take into account possible and probable interrelations of the explanatory variables. For example, in the calculation of poverty rates by state within the Federation, we don't consider the fact that Sao Paulo is a place with less illiteracy than most states, thus should have lower poverty. The multivariate analysis used further ahead seeks to consider these interrelations through a regression of the many explanatory variables taken together. Aiming to provide a better controlled experiment than the bivariate analysis, the objective is to capture the pattern of partial correlations between the variables, interest and explanatory. In other words, we have captured the relations between the two variables, keeping the remaining variables constant. This analysis is very useful to identify the repressed or potential demand for infrastructure as we compared them, for instance, which are the chances of a person with more education having higher electricity coverage, if he/she has the same characteristics as the comparison group.

Binomial Logistic regression

The type of regression used in our simple discrete variables multivariate regressions, as well as to estimate differences-in-differences models. Binomial logistic regression is one method used to study the determination of dummy variables - those composed of only two options of events, such as "yes" or "no" . For example, in the analysis of unemployment:

Let Y be a dummy random variable defined as:

$$Y = \begin{cases} 1 \text{ if the person is employed} \\ 0 \text{ if the person is unemployed} \end{cases}$$

Where each Y_i has a Bernoulli distribution, which probability distribution function is given by: $P(y \mid p) = p^y (1-p)^{1-y}$

where y identifies the event that occurred and p is the probability of success of the event.

Since this is a sequence of events with Bernoulli distribution, the sum of the number of successes or failures in this experiment has binomial distribution of parameters n (number of observations) and p (probability of success). The binomial distribution probability function is given by: $P(y \mid n, p) = \binom{n}{y} p^y (1-p)^{1-y}$

Logistic transformation can be interpreted as the logarithm of the ratio between the odds of success versus failure, in which logistic regression gives us an idea of the return of a person to obtain occupation, given the effect of some explanatory variables that will be introduced later, in particular vocational education. **The bonding function of this generalized linear model is given by the following equation:*** $\eta_i = \log\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{K} \beta_k x_{ik}$

Where the probability p_i is given by:

$$p_i = \frac{\exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}{1 + \exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}$$

The models used here have the objective of identifying the variables related to the characteristics of interest (response variable). When performing the model adjustment, it is desired to find, and to identify, the main factors that best describe the behavior / variation of the characteristics of interest.

The generalized linear model used here is defined by a probability distribution for the response variable, a set of independent variables (explanatory factors) that make up the linear predictor of the model, and a bond function between the mean of the response variable and the linear predictor.

***Odds Ratio:** $$\theta = \left(\frac{p_1}{1-p_1}\right) \Big/ \left(\frac{p_2}{1-p_2}\right)$$

Example: States Conditional Eletricity Coverage – Many of the spatial differences of infrastructure coverage may be attributed to differences in income, education, family size, city size, states and so on. In order to net out these influences, we use multivariate regressions of coverage described above. We focus our analysis on the later spatial variable. The maps presented in each page present the geographical dispersion of coverage across Brazilian states. São Paulo is always portrait white as the basis (i.e. the omitted variable). The red means that is lower than São Paulo, while blue gives the excess with respect to São Paulo. As a general rule, all other States appear in different tones of red except for some statistical draws, meaning that the State of São Paulo presents the best infrastructure in the country.

Difference in Difference estimator (Dif in Dif for D em D) also applied to discrete endogenous variable:

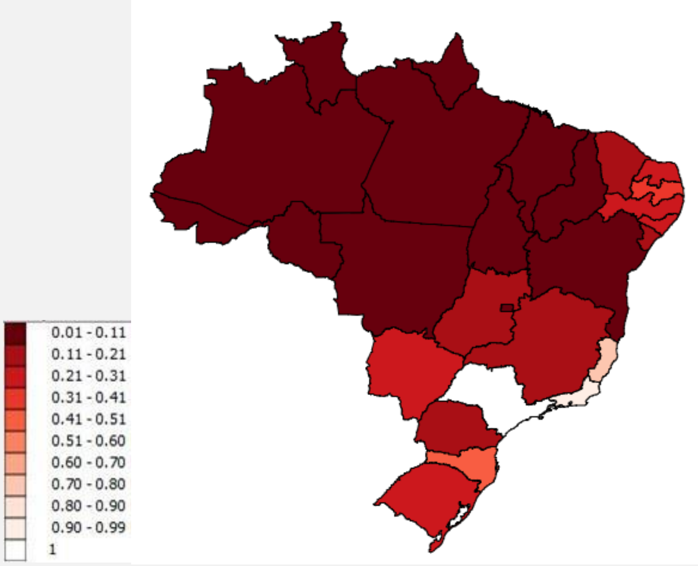
Example of methodology applied to two different periods

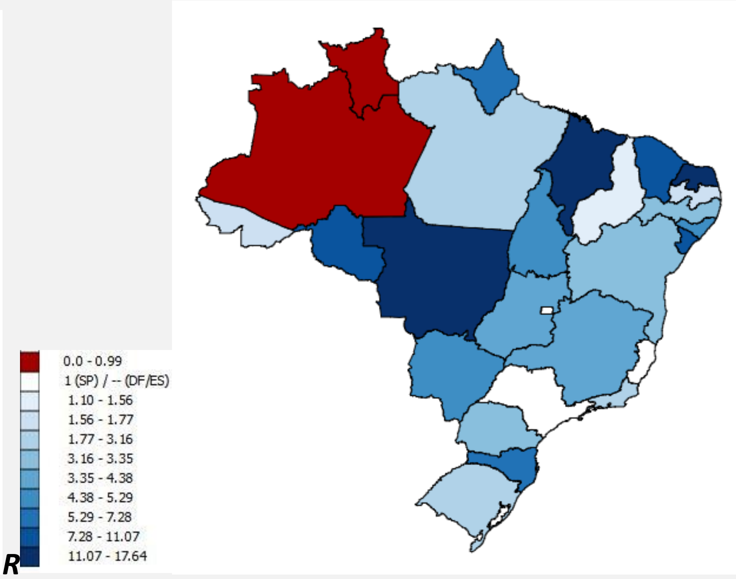
g3 = (y2,t – y1,t) – (y2,c – y1,c) This is achieved with interactive dummies:

Y = g0 + g1*d2 + g2*dT+ (D-D)*d2*dT + other controls

Next we run an extension of the previous multivariate exercise also incorporating the interaction between State Dummies and year in order to grasp the spatial dimension of infrastructure coverage changes. In this second type of regression, we fixed São Paulo as the omitted spatial dummy and 2004 as the omitted temporal category. In this way the results are directly interpreted as the conditional difference in difference of each state in 2015 with respect to São Paulo in 2004. Or how much the infrastructure coverage changed in relative terms. In most cases the color of the map turns into blue which means that the differential between different states and São Paulo tended to fall. This shows a clear convergence trend of infrastructure between Brazilian states even if we net out the effects of income, education and other variables during this period. To be sure, comparisons among states show that an individual from São Paulo has the highest chance of having access to almost all infrastructure services than a similar individual in any other state of the Brazilian Federation. When we move to the comparison of movements of coverage rates, in most cases the color of the map turns into blue. This means that the differential between different states and São Paulo tended to fall. This suggests a clear convergence trend of infrastructure between Brazilian States even if we net out the effects of income, education and other variables during this period. The exceptions are Amazon and Roraima in the North area.

Electricity Coverage - States Odds Ratio **<-Levels and With Time Interaction->**





What are the Most important variables to explain Poverty?

Stepwise procedure - Variables Selection: We used a Stepwise procedure. The final models were selected step by step, after grouping the factor levels based on the Wald statistic, including at each step the interactions that yield the greatest explanatory power (produced the greatest decrease in Deviance), considering the reason test.

Logistic Regression Poverty FGV CPS Line - *SELECT Procedure on PNAD 2015*

Step	Effect	DF	Chi-Square	Pr > ChiSq
1	TELCEL	1	93518.4152	<.0001
2	HH SIZE	1	50227.9216	<.0001
3	STATE	26	21757.7361	<.0001
4	COMPNET	1	14235.6073	<.0001
5	AGE2	1	10050.3293	<.0001
6	EDUCA2	1	7969.6276	<.0001
7	COMMUTING TIME	6	6224.5496	<.0001
8	YEAR	1	4198.1468	<.0001
9	AGE	1	3928.6980	<.0001
10	WATER	1	2375.6744	<.0001
11	HH SIZE 2	1	1869.8112	<.0001
12	ETHNICITY	5	1740.2291	<.0001
13	ELECTRICITY	1	911.3967	<.0001
14	SEWAGE	1	640.8767	<.0001
15	CITY SIZE	4	280.4226	<.0001
16	EDUCA	1	87.1601	<.0001
17	MEAN LOCAL COMMUTING TIME	1	53.7007	<.0001
18	GENDER	1	14.2136	0.0002
19	MEAN LOCAL ELECTRICITY COV.	1	6.2637	0.0123

Infrastructure Externalities - We implemented a stepwise variable selection procedure to determine which socio-economic and infrastructure related variables are more statistically important to explain each social outcome variable seen above. In the selection process we included externality effects from infrastructure. **Poverty -** In the case of the proportion of the poor the six infrastructure variables are significant in descending order: communication, internet, transportation, water, electricity and sewerage. **-** Broader social measure mean that includes besides total income sources from PNAD, imputed rents from housing less opportunity time cost of commuting– the results are similar to poverty. On both social outcomes. two of the externality related variables presented statistically significant impacts namely mean transportation time and mean electricity coverage (mean of an interaction between State and City Size – my neighbor actions impact my outcome – market failure that opens room and justifies State intervention). Electricity access at the community level may improve individual social outcomes through better work opportunities or school or health services. Transportation use on the other extreme imply a common good congestion problem where the excessive use of infrastructure generates a negative externality on all users.

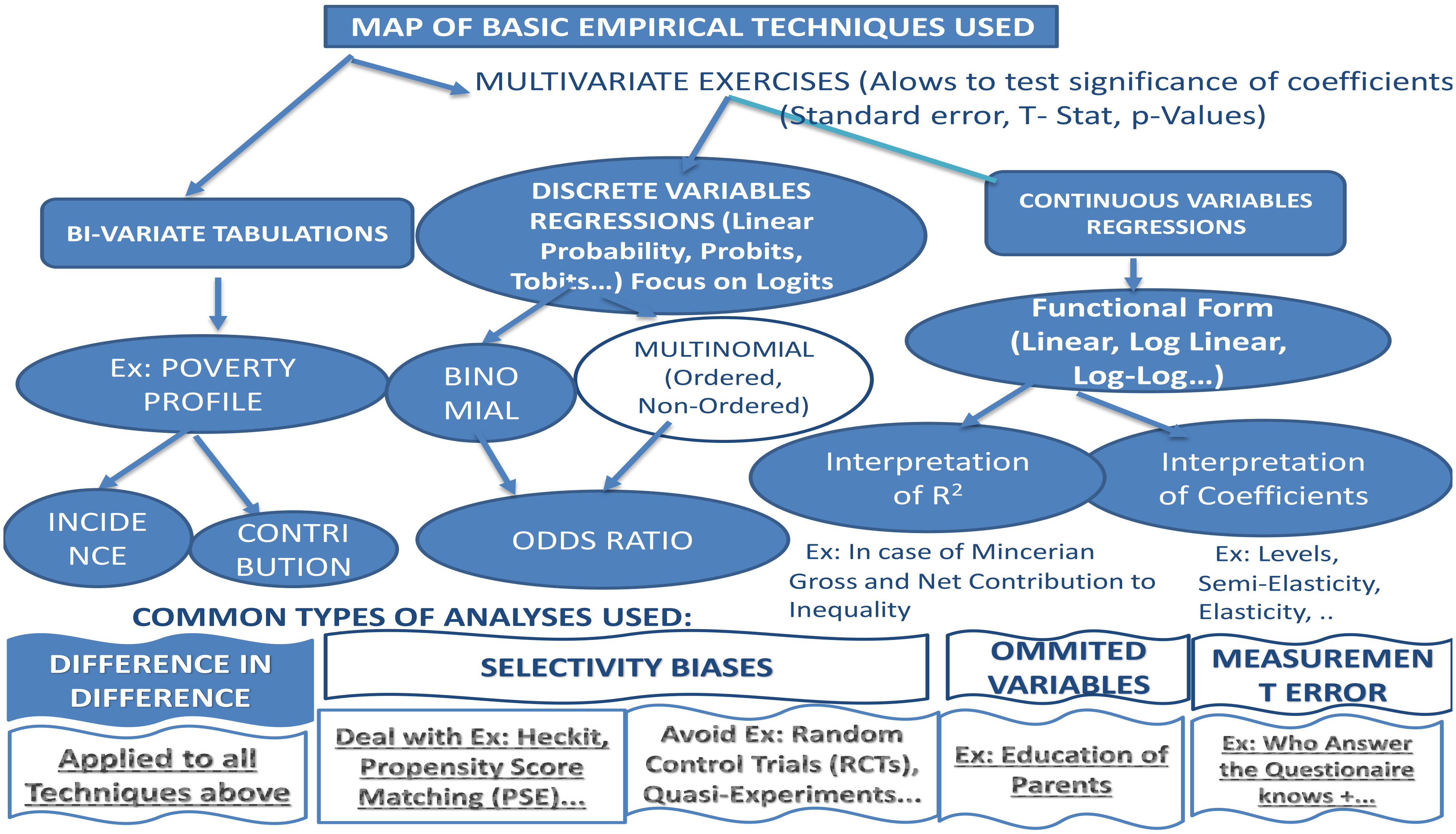
Which States Poverty Fell Faster? Poverty determinants

Binomial Logistic Regression Poverty Line FGV CPS

INTERACTION STATE*YEAR OBS: Few categories used are not displayed below

Parameter	Category	Estimate	Standard Error	Chi-Squared	sig	Conditional Odds Ratio
GENDER	Males	-0.1748	0.0003	284246	**	0.83961
GENDER	Females	0.0000	0.0000	.		1.00000
ETHNICITY	Yellow	-0.4868	0.0038	16699.2	**	0.61457
ETHNICITY	White	-0.4462	0.0007	462992	**	0.64009
ETHNICITY	Indigenous	0.1838	0.0027	4538.79	**	1.20174
ETHNICITY	Mullato	-0.1038	0.0006	27634.3	**	0.90141
ETHNICITY	Black	0.0000	0.0000	.		1.00000
AGE		0.0349	0.0000	815532	**	1.03555
AGE^2		-0.0008	0.0000	1990206	**	0.99918
EDUCA		-0.0232	0.0001	25542.3	**	0.97703
$EDUCA^2$		-0.0102	0.0000	728969	**	0.98983
HH SIZE		0.4667	0.0003	2587765	**	1.59479
HH $SIZE^2$		-0.0171	0.0000	564387	**	0.98301
WATER	No Water Network	0.5372	0.0006	717895	**	1.71124
WATER	Other Source	-0.0817	0.0023	1311.61	**	0.92158
WATER	Well or nascent	-0.0755	0.0006	15273.8	**	0.92726
WATER	*Has Water Network*	0.0000	0.0000	.		1.00000
SEWAGE	Directly in River, Lake or Sea	0.6300	0.0010	391407	**	1.87757

Parameter	Category	Estimate	Standard Error	Chi-Squared	sig	Conditional Odds Ratio
SEWAGE	Rudimentary Cesspit	0.4654	0.0005	759856	**	1.59272
SEWAGE	Connected Cesspit	0.0670	0.0008	6593.72	**	1.06925
SEWAGE	Disconnected Cesspit	0.2209	0.0006	141442	**	1.24718
SEWAGE	Ditch	0.7922	0.0011	551321	**	2.20833
SEWAGE	*Has Sewarage Network*	0.0000	0.0000	.		1.00000
TRASH	Collected Indirectly	0.2514	0.0006	170920	**	1.28586
TRASH	Thrown in River, Lake or Sea	0.6491	0.0042	23501.9	**	1.91382
TRASH	Burned or Buried in the Property	0.5048	0.0007	529673	**	1.65672
TRASH	Collected Directly	0.0000	0.0000	.		1.00000
ELECTRICITY	Oleo, querosene ou gás de botijão	0.1273	0.0011	13351.8	**	1.13578
ELECTRICITY	Other Form	0.6381	0.0028	51867.7	**	1.89296
ELECTRICITY	zElétrica (de rede, gerador, solar)	0.0000	0.0000	.		1.00000
CITY SIZE	Capital in Non Metro Area	-0.1580	0.0010	26984.8	**	0.85383
CITY SIZE	Periphery in Metro Area (suburbs)	0.1616	0.0007	54661.8	**	1.17545
CITY SIZE	Urban Non Metro Area	0.1754	0.0006	90660.1	**	1.19172
CITY SIZE	Rural Area	-0.0453	0.0008	2870.76	**	0.95567
CITY SIZE	Capital in Metro area	0.0000	0.0000	.		1.00000
STATE	AC	0.1946	0.0032	3691.98	**	1.21485
STATE	RJ	0.0332	0.0010	1036.69	**	1.03371
STATE	TO	0.1788	0.0023	6017.85	**	1.19584
STATE	zSP	0.0000	0.0000	.		1.00000
YEAR	a2015	-0.7293	0.0009	603648	**	0.48223
YEAR	z2004	0.0000	0.0000	.		1.00000
STATE*YEAR	AC	0.2943	0.0047	3997.35	**	1.34214
STATE*YEAR	AC	0.0000	0.0000	.		1.00000
STATE*YEAR	CE	-0.0071	0.0016	20.00	**	0.99296
STATE*YEAR	CE	0.0000	0.0000	.		1.00000
STATE*YEAR	RJ	-0.0661	0.0018	1411.80	**	0.93605
STATE*YEAR	RJ	0.0000	0.0000	.		1.00000
STATE*YEAR	TO	0.0907	0.0037	588.93	**	1.09494
STATE*YEAR	TO	0.0000	0.0000	.		1.00000
STATE*YEAR	zSP	0.0000	0.0000	.		1.00000


MAP OF BASIC EMPIRICAL TECHNIQUES USED
MULTIVARIATE EXERCISES (Alows to test significance of coefficients (Standard error, T- Stat, p-Values)
BI-VARIATE TABULATIONS
DISCRETE VARIABLES REGRESSIONS (Linear Probability, Probits, Tobits...) Focus on Logits
CONTINUOUS VARIABLES REGRESSIONS
Ex: POVERTY PROFILE
BINO MIAL
MULTINOMIAL (Ordered, Non-Ordered)
Functional Form (Linear, Log Linear, Log-Log...)
INCIDE NCE
CONTRI BUTION
ODDS RATIO
Interpretation of R^2
Interpretation of Coefficients
Ex: In case of Mincerian Gross and Net Contribution to Inequality
Ex: Levels, Semi-Elasticity, Elasticity, ..
COMMON TYPES OF ANALYSES USED:
DIFFERENCE IN DIFFERENCE
SELECTIVITY BIASES
OMMITED VARIABLES
MEASUREMEN T ERROR
Applied to all Techniques above
Deal with Ex: Heckit, Propensity Score Matching (PSE)...
Avoid Ex: Random Control Trials (RCTs), Quasi-Experiments...
Ex: Education of Parents
Ex: Who Answer the Questionaire knows +...