brazilianpoliticalsciencereview

RESEARCH NOTES

An Analysis of the 2002 Presidential Elections Using Logistic Regression*

Jairo Nicolau
IUPERJ, Brazil

The 2002 elections were a watershed in Brazilian electoral history. Three aspects of the process in particular have been amply stressed in several analyses. The first is the symbolic dimension of Lula's personal victory, the biography of a man of the people who rises to the country's most important office. The second is the victory of PT, the main leftwing party in the country, winning federal office 22 years after being founded. The third is the dimension of the victory, with the president obtaining a resounding vote (61% of the valid votes), which surpassed that of any other Brazilian president since 1945. Despite this, the efforts made by political science to analyse the 2002 elections remain limited, in particular regarding the use of opinion poll results.

This research note does not intend to carry out a thorough and detailed analysis of the choice of candidates or of the key events in the campaign. The aim is to investigate the variables that may be associated with the voting decision in the 2002 presidential elections. To this end, the results of the election survey conducted by Instituto Universitário de Pesquisas (Iuperj)-2002 and the technique of logistic regression will be used. The latter is widely used for election studies in other countries, but little used in Brazil to date. Therefore, as well as providing a substantial analysis of the data, this research note suggests a methodological option for future studies on Brazilian elections.

Few works have attempted to explain the determinants of the vote in the Brazilian presidential elections on the basis of micro-data. Two works stand out for covering more than one election (Singer 2000; Carreirão 2002). Singer (2000) analysed the results of surveys carried out nationwide in 1989 and in the state of São Paulo in 1994. One of the book's purposes is to show the association between ideology, measured as voters' positioning on the right-left spectrum, and the vote. The data are analysed by means of bivariate analyses, using classic association tests (chi-square and Cramer's V). Carreirão (2002) analysed several opinion polls conducted during three presidential election campaigns (1989, 1994 and 1998). His aim also was to measure the impact of a set of variables on the vote using bivariate analysis and association tests (gamma). The author considered a larger number of variables and presented the results in a more detailed fashion than Singer (2000).

Despite carefully describing the context and players involved in each election, both works have the limitation of not using multivariate techniques. This means the reader is prevented from knowing to what extent the independent variables are associated with one another, or what the impact is of each variable when others are analysed simultaneously. A series of issues remain in the air. What might the impact be of voters' positioning on the right-left spectrum when the effects of party preference and evaluation of the federal government are analysed jointly? Could it be that socio-demographic variables, such as educational level and age, continue having an effect when variables associated with political attitude, such as party preference and evaluation of the government, are considered?

Carreirão and Barbetta (2004) took a step forward when they proposed a multivariate model for analysing the results of the 2002 elections. They used a multivariate technique (logistic regression) to analyse the data from an opinion poll conducted in Greater São Paulo. But the authors' pioneering effort was harmed by certain factors. One, recognized by them, refers to the date of the fieldwork, May 2002, therefore before the campaign and the televised electoral broadcasts began. Others derive from the technical choices adopted. The first is the decision to use four binary models that compare separately the preference for a candidate with the option for all the others. For example, one model compared Lula's vote with that of all the candidates plus blank and spoilt votes. Further to the theoretical limitations derived from aggregating in the same category substantially different choices, this decision tends to inflate the hit rate of the "others" category[1]. The most appropriate option would have been to use a multinomial logistic regression model, which is employed when the dependent variable is not binary (Tabachnick and Fidell 2001). Another decision that may have affected the analysis of the data is the option for the stepwise method (Backward LR), which excludes all the variables that are not statistically significant from a certain level on (p>0,5). The authors offer no theoretical justification for employing a model excluding the impact of substantially relevant variables.

Logistic regression is a multivariate technique that allows one to analyse the relationship between independent variables (quantitative or categorical) and categorical dependent variables (Miles and Shevlin 2001; Tabachnick and Fidell 2001). The main virtue of this technique is to permit a multivariate analysis for categorical data — data traditionally analysed by means of bivariate analyses. Initially utilised in medical research, in which the result often is whether or not a particular illness is carried, logistic regression has been more and more used in the social sciences, particularly in electoral studies (Clarke et al. 2004; Evans 2004).

Data

The database used was that of the Iuperj-2002 survey, carried out between December 12 and 15, with 2004 voters from 115 municipalities. For the analysis of the data, I used the SPSS binary and multinomial regression models. Annex 1 presents the results of a bivariate analysis (vote for president/various variables) for the two rounds. It is possible to observe the percentage received by the candidates in each category, as well as a statistical test (Cramer's V) to evaluate the significance of each association. Eight variables were selected, five socio-demographic (age, sex, colour, schooling and religion) and three that evaluate political attributes (evaluation of the Fernando Henrique Cardoso government, sympathy for a political party and position on the right-left spectrum). The treatment given to each is described below.

Dependent Variables

A bigger challenge in relation to the independent variables is aggregating them in a small number of categories. Even though some information is lost, this step is necessary, as categories with a number of cases affect the result of the logistic regression.

For the first round, I compared only the main candidates' vote, each representing a specific category: Lula, Serra, Garotinho and Ciro Gomes. Voters who spoilt or left their ballots blank, who voted for two other candidates (Rui Pimenta and José Maria) or who did not answer, were considered missing (154 cases in total).

For the second round, the option for one of the two candidates (Lula or Serra) was considered. Voters who spoilt or left their ballots blank, or who did not answer, were considered missing (164 cases).

Independent Variables

- *Age* – The original variable, measured as an interval variable, was transformed into a categorical variable, with five age groups: 16-24, 25-34, 35-44, 45-59 and 60+ years of age.
- *Gender* – Male and female.
- *Colour* – The questionnaire asked the interviewees to self-classify in one of four categories: white, black, brown and yellow; these were re-grouped into two categories: white and non-white.
- *Schooling* – Four categories: illiterate/up to 4th grade, 5th to 8th grade, 9th to 11th grade and higher education.
- *Religion* – The various religious denominations were grouped into three categories: catholic, evangelical and others.

- *Evaluation of the Fernando Henrique Cardoso government* – The original variable was grouped into two: positive (excellent, good, positive average) and negative (terrible, bad and negative average).
- *Party Sympathy* – The original categories were grouped into five bands: PT, PMDB, PSDB, others and none.
- *Position on the right-left spectrum* – The survey suggested that voters self-classify on a five-point scale; the values were re-codified into three categories: 1 and 2 (left); 3 (centre); 4 and 5 (right); furthermore, voters who didn't know how to answer were considered.

Multivariate Analysis

Results of the first round

An analysis using multinomial logistic regression was carried out taking Lula's vote as the unit of reference for the comparison with Serra, Garotinho and Ciro Gomes. A test of the model with the eight independent variables in contrast with the model that included only the constant was statistically significant (chi-square = 851.40, p<0.0001), indicating that the predictors as a whole really distinguish Lula's vote in comparison with that of the other candidates. The *pseudo*-R2 (Nagelkerke) of 0.42 demonstrates that the model's total variance is good. Using the eight independent variables, the model was able to classify correctly 86% of Lula's vote, 52% of Serra's and 40% of Garotinho's. For Ciro Gomes the result was not satisfactory, as there was no case of correct prediction. The model's general percentage of hits is of 65%.

Tables 1, 2 and 3 show the coefficients of the regression (log-odds), the significance level, the odds-ratio and the 95% confidence interval for the odds-ratio. Table 1 presents the comparison between Serra and Lula voters. A large number of categories is statistically significant and probably distinguish Lula voters from Serra voters. Observing the odds-ratio column is particularly interesting. The values above 1 indicate that the chances of the voter voting for Serra increase, while the numbers below 1 indicate that the chances of him/her voting for Serra decrease (or that the chances of him/her voting for Lula increase). For example, an elderly voter (over 60 years old) is 2.07 times more likely to vote for Serra than a young voter (16-25 years old). On the other hand, the fact that a voter is sympathetic to PT reduces by 1/6 the probability of him/her voting for Serra. The statistically significant categories associated with an increased chance of one voting for Serra are: being female; being 45-59 years old; being over 60 years old; being white; being in the 5th to 8th grade schooling bracket; having a positive evaluation of the Fernando Henrique government; being sympathetic to PMDB and PSDB; and being rightwing. The categories associated with a decreased chance of one voting for Serra are: being male; being sympathetic to PT; and being leftwing.

TABLE 1

Results of the Multinomial Logistic, 2002 Elections – Serra's vote compared to Lula's vote

	Log-odds	Significance	Odds Ratio	95% Confidence Interval for the Odds Ratio
Age (16-24)				
25-34	-0,06	0,780	0,94	0,63-1,42
35-44	0,20	0,365	1,22	0,79-1,87
45-59	0,81	0,000	2,25	1,46-3,47
60+	0,73	0,005	2,07	1,25-3,42
Sex (Female)				
Male	-0,36	0,009	0,70	0,53-0,92
Colour (Non-White)				
White	0,34	0,014	1,41	1,07-1,86
Schooling (Up to 4th Grade)				
5 th to 8 th	0,24	0,195	1,27	0,88-1,83
9th to 11th	0,49	0,014	1,64	1,11-2,42
Higher	0,65	0,015	1,91	1,14-3,21
Religion (Catholic)				
Others	-0,00	0,989	0,99	0,63-1,57
Evangelical	0,04	0,834	1,04	0,70-1,57
Evaluation of FHC Government (Negative)				
Positive	1,17	0,000	3,21	2,40-4,29
Party Sympathy (None)				
PT	-2,08	0,000	0,13	0,08-0,21
PMDB	0,53	0,028	1,70	1,06-2,74
PSDB	1,45	0,000	4,25	2,29-7,87
Others	0,47	0,069	1,59	0,96-2,64
Left-Right Position (None)				
Left	-0,76	0,001	0,47	0,30-0,74
Centre	-0,13	0,528	0,88	0,59-1,32
Right	0,63	0,001	1,88	1,28-2,75
Constant	- 1,90	0,000		

NB: The reference category of each variable appears in brackets.
Pseudo R2 (Nagelkerke) = 0.42
% of the total number of cases classified correctly: 65%

Table 2 shows the results of the comparison between Garotinho and Lula. Only five categories are statistically significant (p<0.05), two of them associated with increased chances: sympathy for other parties and evangelicals. The increase in the odds-ratio for these is meaningful. The chances of voting for Garotinho increase by a factor of 11.5 when one compares evangelical with catholic voters. The factors associated with decreased chances of one voting for Garotinho are: being male; being sympathetic to PT; and being leftwing.

TABLE 2

Results of the Multinomial Logistic, 2002 Elections – Garotinho's vote compared to Lula's vote

	Log-odds	Significance	Odds Ratio	95% Confidence Interval for the Odds Ratio
Age (16-24)				
25-34	-0,30	0,229	0,74	0,45-1,21
35-44	-0,44	0,113	0,65	0,38-1,11
45-59	-0,17	0,547	0,84	0,48-1,48
60+	0,07	0,827	1,07	0,56-2,04
Sex (Female)				
Male	-0,37	0,041	0,69	0,49-0,99
Colour (Non-White)				
White	0,23	0,199	1,26	0,89-1,80
Schooling (Up to 4th Grade)				
5 th to 8 th	0,26	0,269	1,30	0,82-2,06
9th to 11th	0,45	0,080	1,57	0,95-2,60
Higher	0,02	0,955	1,02	0,48-2,18
Religion (Catholic)				
Others	2,45	0,000	11,5	7,87-16,91
Evangelical	0,13	0,698	1,14	0,59-2,23
Evaluation of FHC Government (Negative)				
Positive	0,23	0,207		
Party Sympathy (None)				
PT	-1,41	0,000	0,24	0,15-0,41
PMDB	-0,45	0,300	0,64	0,28-1,49
PSDB	-0,36	0,510	0,70	0,24-2,05
Others	0,95	0,002	2,59	1,42-4,71
Left-Right Position (None)				
Left	-0,59	0,041	0,55	0,32-0,98
Centre	-,016	0,534	0,85	0,51-1,42
Right	0,303	0,243	1,35	0,81-2,25
Constant	-0,203	0,000		

NB: The reference category of each variable appears in brackets.

Table 3 shows the results of the comparison between Ciro Gomes and Lula. The chances of one voting for Ciro Gomes increase in the following cases: being over 60 years old; being in the 9th to 11th grade or higher education brackets; and being in the 'other' religious category (i.e., non-catholics and non-evangelicals). The chances decrease among voters sympathetic to PT.

TABLE 3

Results of the Multinomial Logistic, 2002 Elections – Ciro Gomes's vote compared to Lula's vote

	Log-odds	Significance	Odds Ratio	95% Confidence Interval for the Odds Ratio
Age (16-24)				
25-34	0,47	0,090	1,59	0,93-2,74
35-44	0,68	0,018	1,97	1,12-3,47
45-59	0,58	0,067	1,79	0,96-3,33
60+	0,82	0,023	2,26	1,18-4,57
Sex (Female)				
Male	-0,19	0,310	0,83	0,58- 1,19
Colour (Non-White)				
White	0,243	0,197	1,28	0,88-1,85
Schooling (Up to 4th Grade)				
5 th to 8 th	0,09	0,743	1,09	0,66-1,81
9th to 11th	0,86	0,001	2,36	1,43-3,99
Higher	1,03	0,002	2,79	1,45-5,37
Religion (Catholic)				
Others	0,08	0,774	1,08	0,62-1,88
Evangelical	-0,70	0,047	0,49	0,25-0,99
Evaluation of FHC Government (Negative)				
Positive	0,25	0,405	1,29	0,89-1,86
Party Sympathy (None)				
PT	0,82	0,023	0,89	0,04-0,18
PMDB	0,47	0,090	0,75	0,35-1,63
PSDB	0,68	0,061	1,66	0,67-4,08
Others	0,58	0,067	1,78	0,98-3,24
Left-Right Position (None)				
Left	-0,25	0,405	0,78	0,43-1,41
Centre	0,28	0,306	1,32	0,78-2,24
Right	0,28	0,318	1,32	0,76-2,29
Constant	-2031	0,000		

NB: The reference category of each variable appears in brackets.

Results of the second round

An analysis was conducted using binary logistic regression to compare Lula and Serra's vote (Table 4). A test of the model with the eight independent variables in contrast with the model that considered only the constant is statistically significant (chi-square = 475.06; p<0.0001), indicating that the set of predictors really distinguish Lula's vote from Serra's vote in the second round. The *pseudo*-R2 (Nagelkerke) of 0.34 demonstrates that the model's total variance is reasonable. Using the eight independent variables, the model is capable of classifying correctly 91% of Lula's voters and 41% of Serra's. The model's total hit rate is 78%. The statistically significant categories associated with increased chances of voting for Serra are: being female; being 45-59 or over 60 years old; being white; being in the 9th to 11th grade schooling bracket; having a positive evaluation of the Fernando Henrique government; being sympathetic to PSDB; and being rightwing. The categories associated with decreased chances of one voting for Serra are: being male; being sympathetic to PT; and being leftwing. In relation to the first round, two categories ceased to be statistically significant when comparing Serra and Lula: having a higher education and being sympathetic to PMDB.

TABLE 4

Results of the Binomial Logistic, 2002 Elections, Second Round – Serra's vote compared to Lula's vote

	Log-odds	Significance	Odds Ratio	95% Confidence Interval for the Odds Ratio
Age (16-24)		0,019		
25-34	0,09	0,642	1,09	0,78-1,57
35-44	0,06	0,779	1,06	0,72-1,57
45-59	0,56	0,006	1,75	1,78-2,60
60+	0,50	0,031	1,65	1,05-2,61
Sex (Female)				
Male	-0,19	0,135	0,83	0,65-1,06
Colour (Non-White)				
White	0,27	0,036	1,31	1,01-1,68
Schooling (Up to 4th Grade)		0,021		
5 th to 8 th	0,15	0,382	1,17	0,83-1,62
9th to 11th	0,51	0,004	1,67	1,18-2,37
Higher	0,46	0,059	1,58	0,98-2,53
Religion (Catholic)		0,838		
Evangelical	0,02	0,894	1,02	0,74-1,41
Others	-0,12	0,581	0,89	0,58-1,36
Evaluation of FHC Government (Negative)		0,000		
Positive	1,04	0,000	2,82	2,17-3,66
Party Sympathy (None)		0,000		
PT	-1,99	0,000	0,14	0,08-0,22
PMDB	0,34	0,131	1,40	0,91-2,17
PSDB	1,27	0,000	3,58	2,10-6,08
Others	0,31	0,149	1,37	0,89-2,09
Left-Right Position (None)		0,000		
Left	-0,80	0,000	0,45	0,29-0,69
Centre	0,02	0,933	1,02	0,70-1,47
Right	0,78	0,000	2,19	1,55-3,10
Constant		0,000		

NB: The reference category of each variable appears in brackets.
Pseudo R2 (Nagelkerke) = 0.42
% of the total number of cases classified correctly: 65%

Conclusion

This first analysis of the 2002 election results using logistic regression brings to light a number of interesting results. Despite dealing with a small number of variables, the models for the two rounds had reasonable variance and good classification of the cases, Ciro Gomes excepted. (This is probably owed to the small number of cases considered for this candidate.) The coefficients of the variables also show that certain voter characteristics probably distinguished the candidates, particularly between Serra and Lula: gender, schooling, age, position on the right-left spectrum, evaluation of the government and sympathy for political parties. It is likely that including other variables in the model — variables relating to perspectives for the future, evaluation of certain campaign issues and some of the candidates' attributes — would generate more accurate estimates and increase the percentage of correct answers.

There is a long tradition of research in traditional democracies on the determinants of the vote. Recently, it has benefited from advances in data analysis and from a rich theoretical debate. In Brazil, we still have a long way to go, above all with regards to the improvement of data gathering and analysis. To this end, the exercise effected in this research note suggests that embracing logistic regression as a major tool and more systematically, would be in order.

(Submitted for publication in May, 2006)
Translated from Portuguese by Leandro Moura

Notes

[*] Editors' Note: The need to speed up the launch of the first issue of BPSR, which had already been delayed several times, regrettably led the Editors to overlook their duty to inform two contributors of the overlap between their respective pieces. This explains the publication of this Research Note by Jairo Nicolau, in which he sets out to analyse the 2002 Brazilian presidential election by means of the technique of logistic regression, claiming that although this technique is widely used for election studies in other countries, it had been little used in Brazil to date, and of the article by Yan de Souza Carreirão (Relevant Factors for the Voting Decision in the 2002 Presidential Election), in which he investigates this same election by testing some of the main hypotheses about electoral behaviour in the country by means of logistic regression analyses.

[1] The hit rate of each model was the following: Lula: 82,5% – others: 86,5%; Serra: 46,4% – others, 94%; Garotinho: 95,6% – others: 44,8%; Ciro: 98,5% – others: 20%. It is no coincidence that the hit rate of the "others" category was so high in all four models.

Bibliography

Carreirão, Yan de Souza. 2002. *A decisão do voto nas eleições presidenciais brasileiras*. Rio de Janeiro: FGV Editora.

______, and Pedro Alberto Barbetta. 2004. A eleição presidencial de 2002: A decisão do voto na região da grande São Paulo. *Revista Brasileira de Ciências Sociais* 56.

Clarke, Harold D., David Sanders, Marianne C. Stewart, and Paul Whiteley. 2004. *Political choice in Britain*. Oxford: Oxford University Press.

ANNEX 1

Percentage of the First Round Vote, according to a Set of Variables

	Lula	Serra	Garotinho	Ciro	Blank/Spoilt
Age					
16-24	57	19	13	6	5
25-34	57	18	9	9	7
35-44	54	19	9	10	8
45-59	47	30	10	8	5
60+	45	29	12	8	7
Cramer's V = 0,08; p<0,0001					
Sex					
Male	58	20	8	8	6
Female	49	24	12	9	6
Cramer's V = 0,01; p<0,0001					
Schooling					
Illiterate/Up to 4th grade	54	23	11	7	5
5th to 8th	56	20	11	7	6
9th to 11th	50	22	10	11	7
Higher	48	28	7	12	4
Cramer's V = 0,06; p = 0,06					
Self-Defined Colour					
White	48	26	11	10	6
Non-White	58	18	10	7	6
Cramer's V = 0,12; p<0,0001					
Religion					
Catholic	57	24	5	9	5
Pentecostal Evangelical	32	16	44	3	6
Non-Pentecostal Evangelical	44	20	24	5	7
Others	56	19	5	9	11
Cramer's V = 0,42; p<0,0001					
Evaluation of Fernando Henrique Government					
Positive (Excellent, Good, Positive Average)	43	33	10	9	5
Negative (Terrible, Bad, Negative Average)	64	11	11	8	7
Cramer's V = 0,28; p<0,0001					
Party Sympathy					
PT	86	4	5	2	3
PMDB	38	45	7	8	3
PSDB	20	63	6	10	1
Others	35	31	19	14	1
None	45	24	13	10	8
Cramer's V = 0,23; p<0,0001					
Position on the Left-Right Spectrum					
Left	74	9	7	6	5
Centre	49	20	11	11	8
Right	39	38	12	9	3
Doesn't know - Didn't Answer	48	22	13	8	9
Cramer's V = 0,19; p<0,0001					

ANNEX 2

Percentage of the Second Round Vote, according to a Set of Variables

	Lula	Serra	Blank/Spoilt
Age			
16-24	70	22	8
25-34	71	23	6
35-44	71	22	7
45-59	62	30	8
60+	61	30	9
Cramer's V = 0,07; p = 0,02			
Sex			
Male	71	23	7
Female	65	27	8
Cramer's V = 0,06; p = <0,03			
Schooling			
Illiterate/Up to 4th grade	69	24	6
5th to 8th	70	21	9
9th to 11th	65	28	7
Higher	63	31	7
Cramer's V = 0,06; p = 0,05			
Self-Defined Colour			
White	63	30	7
Non-White	72	21	7
Cramer's V = 0,27; p<0,0001			
Religion			
Catholic	69	26	6
Pentecostal Evangelical	62	29	9
Non-Pentecostal Evangelical	60	24	16
Others	70	19	11
Cramer's V = 0,09; p<0,0001			
Evaluation of Fernando Henrique Government			
Positive (Excellent, Good, Positive Average)	57	36	7
Negative (Terrible, Bad, Negative Average)	80	13	8
Cramer's V = 0,27; p<0,0001			
Party Sympathy			
PT	92	4	4
PMDB	52	44	4
PSDB	30	66	5
Others	56	40	4
None	63	28	10
Cramer's V = 0,26; p<0,0001			
Position on the Left-Right Spectrum			
Left	87	8	5
Centre	67	24	9
Right	51	44	6
Doesn't know - Didn't Answer	65	24	11
Cramer's V = 0,23; p<0000,1			