Read this paper if you want to learn logistic regression

Antônio Alves Tôrres Fernandes\textsuperscript{1}, Dalson Britto Figueiredo Filho\textsuperscript{1}, Enivaldo Carvalho da Rocha\textsuperscript{1}, Willber da Silva Nascimento\textsuperscript{1}  
\textsuperscript{1}Postgraduate Program in Political Science, Federal University of Pernambuco, Recife, PE, Brazil.

\textbf{ABSTRACT} Introduction: What if my response variable is binary categorical? This paper provides an intuitive introduction to logistic regression, the most appropriate statistical technique to deal with dichotomous dependent variables. Materials and Methods: we estimate the effect of corruption scandals on the chance of reelection of candidates running for the Brazilian Chamber of Deputies using data from Castro and Nunes (2014). Specifically, we show the computational implementation in R and we explain the substantive interpretation of the results. Results: we share replication materials which quickly enables students and professionals to use the procedures presented here for their studying and research activities. Discussion: we hope to facilitate the use of logistic regression and to spread replication as a data analysis teaching tool.

KEYWORDS: regression; logistic regression; replication; quantitative methods; transparency.

Received in October 19, 2019. Approved in May 7, de 2020. Accepted in May 16, 2020.

I. Introduction\textsuperscript{1}

The least squares linear model (OSL) is one of the most used tools in Political Science (Kruger & Lewis-Beck, 2008). As long as its assumptions are respected, the estimated coefficients from a random sample give the best linear unbiased estimator of the population’s parameters (Kennedy, 2005). Unbiased because it does not systematically over or underestimates the parameter’s value and because it gives the smallest variance among all possible estimates (Lewis-Beck, 1980).

What about when assumptions are violated? In that case, we must adopt techniques better suited to the nature of the data. For instance, imagine a study that investigates the impact of campaign spending on the chance of a candidate being elected or not. Since the dependent variable is binary, some assumptions of the least squares model are violated (homoscedasticity, linearity, and normality) and the estimates may be inconsistent. A logistic regression is the best tool to handle dichotomous dependent variables, that is, when y can only take on two categories: elected or not-elected; adopted the policy or did not adopt the policy; voted for president Bolsonaro or not. Lottes, DeMaris, and Adler (1996) argue that, despite logistic regression’s popularity in the Social Sciences, there is still a lot of confusion regarding its correct use. Given our pedagogical experience, this difficulty is explained by the lack of intuitive teaching materials. Moreover, many undergraduate and graduate programs, as well as textbooks, end their content at linear regression, shortening the dissemination of other data analysis techniques.

To fill this gap, this paper presents an introduction to logistic regression. Our goal is to facilitate the understanding of its practical application. As far as audience, we write to students in the early stages of training and teachers who need materials for quantitative methods courses. Methodologically, we reproduce data from Castro and Nunes (2014) regarding the relationship between involvement in corruption scandals (\textit{Mensalão}\textsuperscript{2} and \textit{Sanguessugas}\textsuperscript{2} scandals) and

\textsuperscript{1}Replication materials available at: \text{<https://osf.io/nv4ae/>}. This paper benefitted from the comments of professor Jairo Nicolau and the suggestions made by Revista de Sociologia e Política's anonymous reviewers. We also thank the Berkeley Initiative for Transparency in the Social Sciences and the Teaching Integrity in Empirical Research.

\textsuperscript{2}For a brief review of...
the reelection chances for candidates running for federal deputy in Brazil in 2006. All data and scripts are available at Open Science Framework (OSF) website.

By the end, the reader should be able to identify when a logistic regression should be used, computationally implement the model, and interpret the results. We are aware that this paper does not replace a detailed reading of primary sources on the subject and more technical materials. Nevertheless, we hope to make understanding logistic regression easier to you and to disseminate replicability as data analysis teaching tool.

The remainder of the paper is divided as follows: the next section explains the underlying features logistic regression. The third identifies the main technical conditions that must be met to ensure that the model’s estimates are consistent. The fourth section describes the main statistics that must be observed. Lastly, we provide some recommendations on how to improve the quality of methodological training offered to Political Science undergraduate and graduate students in Brazil.

II. The logic of logistic regression

The use of binary categorical dependent variables is common in Political Science empirical research. For example: voted or not (Nicolau, 2007; Soares, 2000), won or lost the electoral contest (Speck & Mancuso, 2013; Peixoto, 2009), adhered to the policy or not (Furlong, 1998), democracy or not-democracy (Goldsmith, Chalup & Quinlan, 2008), started a war or not (Henderson & Singer, 2000), appealed a judicial ruling or not (Epstein, Landes & Posner, 2013). For all these situations, a logistic regression is the best suited technique to model the dependent variable’s variation given a set of independent variables.

In a logistic regression, the dependent variable only has two categories. Generally, the occurrence of the event is coded as 1 and its absence as 0. Keeping in mind that codification changes the coefficients’ signal and, therefore, their substantive interpretation. To better understand how a logistic regression works, it is necessary to understand the logic of regression analysis as a whole. Let’s look at the linear model’s classic notation:

\[ Y = \alpha + \beta X + \epsilon \]  

\( Y \) represents the dependent variable, that is, what we are trying to understand/explain/predict. \( X \) represents the independent variable. The intercept, (\( \alpha \)), represents the value of \( Y \) when \( X \) equals zero. The regression coefficient, (\( \beta \)), represents the variation observed in \( Y \) associated with the increase of one unit of \( X \). The stochastic term, (\( \epsilon \)), represents the error of the model. Technically, it is possible to estimate if there is a linear relationship between a dependent variable (\( Y \)) and different independent variables. Moreover, the model allows the observation of the effect magnitude and to test the coefficients’ statistical significance (p-value and confidence intervals).

A logistic regression can be interpreted as a particular case of generalized linear models (GLM), in which the dependent variable is dichotomous. Figure 1 compares the linear and logistic models.

Because the dependent variable in the logistic model takes on only two values (0 or 1), the probability predicted by the model must also be limited to that interval. When \( X \) (independent variable) takes on lower values, the probability approaches zero. Conversely, as \( X \) increases, the probability approaches 1. For Kleibaum and Klein (2010), that logistic functions vary between 0 and 1 ex-
The model’s popularity. Given that the dependent variable’s binary nature violates some of the linear model’s assumptions (homoscedasticity, linearity, normality), using a linear model to analyze binary variables may generate inefficient and biased coefficients. To better understand the relationship between linear and logistic models, we reproduced the data from Hosmer, Lemeshow, and Sturdivant (2013) on the association between age and coronary disease (Graph 1).

The vertical dashed line represents the age mean: 44.38 years old. The cases were coded as 1 (developed coronary disease) and 0 (did not develop it). The trend is very clear: as age increases, the amount of people diagnosed with coro-

---

8Hair et al. (2009) state that homoscedasticity is the assumption that the dependent variable displays equal levels of variance over a range of the predictor variable (Hair et al., 2009, p. 83). 2013, p. 77

9For Hair et al. (2009), an implied assumption for all multivariate analysis techniques based on correlational measures of association, including multiple linear regression and logistic regression, is linearity (Hair et al., 2009, p. 85).

10One estimator is the Best Linear Unbiased Estimator, when the following properties are satisfied. Best means efficient, producing the least variance, linear means the type of relationship expected between parameters, and unbiased concerns the sampling distribution of the estimator. A biased estimator is one that systematically over- or underestimates the value of the population parameter.

11The data are available at: <http://www.ats.ucla.edu/stat/stata/examples/alr2/alr2stata1.htm>.
nary disease grows. An intuitive way to observe this pattern is to examine the number of cases using the mean as a parameter for comparison. For example, for people above the mean there more illness cases, while for people below the mean, the larger concentration is in the “did not develop it” category. That is, the graph is stating that there is an association between age and coronary disease. It is in that sense that a logistic regression informs the probability of the event coded as 1 occurring, in the case at hand, developing coronary disease. Table 1 presents the data by age group.

Simply observe the last column to reach the same conclusion presented by Graph 1: the higher the age, the higher the chance to develop coronary diseases. An additional option to visualize the relationship between these variables is to graphically represent the percentage of people who are ill for each age group (Graph 2).

Table 1 - Age group x coronary disease

| Age Group | N   | Yes | No  | Yes (%) |
|-----------|-----|-----|-----|---------|
| 20-29     | 10  | 1   | 9   | 0.1     |
| 30-34     | 15  | 2   | 13  | 0.13    |
| 35-39     | 12  | 3   | 9   | 0.25    |
| 40-44     | 15  | 5   | 10  | 0.33    |
| 45-49     | 13  | 6   | 7   | 0.46    |
| 50-54     | 8   | 5   | 3   | 0.63    |
| 55-59     | 17  | 13  | 4   | 0.76    |
| 60-69     | 10  | 8   | 2   | 0.8     |
| **Total** | **100** | **43** | **57** |         |

Source: The authors, based on Hosmer, Lemeshow, and Sturdivant (2013)

Graph 2 - Age group x coronary disease

Source: The authors based on Hosmer, Lemeshow, and Sturdivant (2013).
We observe a positive correlation between age (axis X) and the probability to develop cardiac diseases (axis Y) is observed. A logistic regression will inform the direction, magnitude, and the statistical significance level of this relationship. In a nutshell, the researcher must use a logistic regression when the dependent variable is categorical and binary. Given that many variables in the Humanities are categorical, the analytical benefits associated with the correct application and interpretation of a logistic regression are evident.

### III. Planning a logistic regression

A logistic regression also supports variables with more than two categories. When there is no hierarchy between the categories, such as with the distribution of civil status, we should use a multinomial regression. On the other hand, an ordinal logistic regression is ideal to model the distribution of ordinal variables, that is, when there is a structure of intensity between the categories.

Table 2 describes the five stages that should be observed.

The first stage is to identify a research question for which the dependent variable is naturally dichotomous. For example, given the popularity of logistic regression in health research, commonly used variables are: lived/died; sick/not sick; smoker/ non-smoker. Usually, a researcher must forgo from recoding a continuous or discrete variable into a dichotomous categorical one. More clearly, let’s say the interest variable is income per capita. It is wrong to recode income to produce two categories: rich versus poor. Technically, recoding a quantitative variable into a categorical one implies loss of information and that reduces the estimates’ consistency (Fernandes et al., 2019).

At the second stage, the technical requirements must be observed. Despite being more flexible than other statistical techniques, logistic regression is sensitive to, for example, problems of multicollinearity (high levels of correlation between independent variables). There are different procedures to minimize this problem. The simplest is to increase the number of observations (Kennedy, 2005). An additional option is to use some data reduction technique to create a synthetic measure from the variance of the original variables. We must not simply exclude one of the independent variables, under the risk of producing errors in the model specification. In a logistic regression, the size of the sample is key (Hair et al., 2009). Small samples tend to produce inconsistent estimates. On the other hand, excessively large samples increase the power of statistical tests in such a way that any effect tends to be statistically significant, regardless of magnitude. Hosmer and Lemeshow (2000) suggest a minimal \( n \) of 400 cases. Hair et al. (2009) suggest a ratio of 10 cases for each independent variable included in the model. Pedhazur (1982) recommends a ratio of 30 cases for each estimated parameter.

Another eventual source for problems is outliers. Extreme cases produce disastrous results in data analysis and in the case of a logistic regression, the presence of atypical observations may harm the model’s fit. Once aberrant cases are detected, a researcher must decide what to do with them. Sometimes an extreme case is nothing more than a typo and can be easily solved. One option is to exclude outliers from the model’s estimation and measure the impact of its inclu-

| Stage | Description |
|-------|-------------|
| 1<sup>a</sup> | Identify the dependent variable |
| 2<sup>a</sup> | Note the technical requirements |
| 3<sup>a</sup> | Estimate and fit the model |
| 4<sup>b</sup> | Interpret the results |
| 5<sup>a</sup> | Validate the results |

Source: the authors, based on Hair et al. (2009).
Another procedure commonly adopted is to recode the case, giving it a less extreme value, the mean for example. In any case, it is important to describe in detail what was done to deal with eventual extreme observations.

At stage three, the researcher must estimate the model. Here, two procedures are essential: a) report the software and b) share replication materials, which include the original data, the manipulated data, and the computational scripts. These procedures increase transparency and make replicability of results easier (King, 1995; Paranhos et al., 2013; Janz, 2016; Figueiredo Filho et al., 2019).

After estimating the model, the next step is evaluating the goodness of the fit. This can be done by comparing the null model (just the intercept) with the model that incorporates the independent variables. A statistically significant difference between the models indicates that the explanatory variables help to predict the occurrence of the dependent variable. Figure 2 shows the underlying logic of model comparison when we are using logistic regression.

Comparatively, model B has a better fit than model A. This can be observed given the difference in discriminatory power. While model A presents high variability, model B is more precise. For Tabachnick, Fidell, and Ullman,

[...] “logistic regression, like multiway frequency analysis, can be used to fit and compare models. The simplest (and worst-fitting) model includes only the constant and none of the predictors. The most complex (and ‘best’-fitting) model includes the constant, all predictors, and, perhaps, interactions among predictors. Often, however, not all predictors (and interactions) are related to the outcome. The researcher uses goodness-of-fit tests to choose the model that does the best job of prediction with the fewest predictors.” (Tabachnick, Fidell & Ullman, 2007, p. 439).

The fourth stage is the interpretation of results. Unfortunately, many works limit themselves to analyzing the statistical significance of the estimates and do not pay attention to the coefficients’ magnitude. We suggest that researchers interpret the coefficients and substantively discuss how results are related to the research hypothesis. Unlike a linear regression, in which coefficients are easy to interpret, the estimates produced in the logistic model are less intuitive. This is because the logit transformation informs the independent variable’s effect on the variation of the dependent variable’s natural logarithm of the odds. For example, when considering a coefficient of 0.6, an increase of 0.6 units is expected in the logit of Y every time X increases by one unit. This approach’s main disadvantage is its lack of intelligibility. To state that the amount in logit increased 0.6 units is not very intuitive and does not help to understand the relationship between the variables.

Figure 2 - Comparing the fit of logistic models

Source: Hair et al. (2009).
A second possibility is to analyze the independent variables’ impact on the odds of Y. To do so, a researcher must get the exponent of the coefficient itself. In our example, the exponential of 0.6 is 1.82. This means that for each additional unit in X, an increase of 1.82 is expected in the chance of Y occurring, keeping other variables constant. Graph 3 illustrates the distribution of a simulation’s exponential function, in which x varies between -5 and 5.

In a logistic regression, the exponential of a positive value (+) produces a coefficient larger than 1. Conversely, a negative coefficient (-) returns an \( \text{Exp} (\beta) \) smaller than 1. A coefficient with a value of zero produces an \( \text{Exp} (\beta) \) equal to 1, indicating that the independent variable does not affect the chance of the dependent variable’s occurrence. So, write it down in your notebook: the farther the coefficient is from one, regardless of the direction, the greater the impact of a given independent variable on the chance of the event of interest occurring\(^{19}\).

The third possibility is to estimate the percentage increase in the chance of the occurrence of Y. To do so, one must subtract one unit from the exponentiated regression coefficient and multiply the result by 100, in this case \((1.82-1) \times 100\). Then we have that the increase in one unit of X is associated with an increase of 82% in the chance of Y occurring (ceteris paribus). The interpretation of the logistic regression’s coefficients may become a little more complicated when the chance is smaller than 1, that is, when the coefficient (\( \beta \)) is negative. One solution is to invert the coefficient \((1/\text{coefficient’s value})\), which makes the interpretation easier. For example, a coefficient of 0.639, when inverted, indicates that when the independent variable decreases by in one unit, an average increase of 1.56 is expected in the chance of the dependent variable occurring.

Lastly, the researcher must validate the results observed with a subsample of its original dataset. This procedure gives the research results more reliability, especially when working with small samples. According to Hair et al. (2009),

“the most common approach for establishing external validity is the assessment of hit ratios through either a separate sample (holdout sample) or utilizing a procedure that repeatedly processes the estimation sample. External validity is supported when the hit ratio of the selected approach exceeds the comparison

\(^{19}\)When interpreting the statistical significance of the confidence interval of the odds regression coefficient, we must observe if the interval includes the value one (1). If so, we are faced with a non-significant result. For example, in a confidence interval in which the coefficient varies between 0.8 and 1.6, it is not possible to reject the null hypothesis.
standards that represent the predictive accuracy expected by chance.” (Hair et al., 2014, p. 329).

Unfortunately, this procedure is rarely used by political scientists. We suspect that the reduced use of validation is in part explained by the lack of training on the specificities of logistic regression. The next section presents an applied example of logistic regression and explains how the results should be interpreted.

IV. An applied example

To illustrate the application of the logistic regression, we replicated the data from Castro and Nunes (2014) on corruption and reelection. However, since our focus is purely methodological, we will not explore the substantive meaning of the conclusions reported by the authors. According to the planning from the previous section, the first step is to identify the dependent variable that will take value “1” for candidates reelected in 2006 and “0” if otherwise.

The second step is to verify the technical requirements to estimate the logistic regression. During this step, it is important to observe the presence of outliers, the occurrence of high correlation between independent variables, and an adequate sample size. Due to space limitations, we will reproduce only one of the models presented by Castro and Nunes (2014). Specifically, the sample used to estimate model 5 from Table 6 (p. 41), which has a total of 217 observations and a proportion of 19 cases for each independent variable. We do not find deviant cases and the level of correlation between the variables included in the model is acceptable. Thus, we can move on to the next phase.

The third stage consists of the model’s estimation:

\[
\logit(Y) = \alpha + X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3 + X_4 \beta_4 + X_5 \beta_5 + X_6 \beta_6 + X_7 \beta_7 + X_8 \beta_8 + X_9 \beta_9 + X_{10} \beta_{10} + X_{11} \beta_{11} + \varepsilon
\]

(2)

Chart 1 summarizes how the variables were measured.

We will test three hypotheses:

| Variables                          | Description                                      |
|------------------------------------|--------------------------------------------------|
| Sex (Control)                      | Dummy: Female (0); Male (1)                      |
| Age (Control)                      | Continuous: age at election.                     |
| Education (Control)               | Categorical ordinal: Read and write (0); Elementary School incomplete (1); Elementary School complete (2); High School incomplete (3); High School complete (4); Tertiary education incomplete (5); Tertiary Education (6). |
| Poverty (Control)                 | Continuous: percentage of poor people in the state. |
| Ideology (Control)                | Categorical: Left (0); Center (1); Right (2).    |
| Vote Increase 2006 (Control)      | Dummy: Increased (1); Lowered (0).               |
| Change (Control)                  | Dummy: Changed parties (1); Did not (0).         |
| Pork (Control)                    | Continuous: success rate of execution of parliamentary amendments. |
| Seats per state (Control)         | Continuous: number of seats for each state at the Chamber of Deputies. |
| Expenditures (Control)            | Continuous: campaign expenditures                |
| Scandal (IV)                      | Dummy: Involved in a scandal (1); Not involved in a scandal (0). |
| Reelection (DV)                   | Dummy: Reelected (1); Not-reelected (0).         |

Source: the authors, based on Castro and Nunes (2014, p. 38-40).
V. Results

The first step is to analyze the distribution of the dependent variable. Table 3 summarizes this information.

There is information for 451 cases. From this total, 60.53% of the federal deputies were reelected in 2006, which means 273 occurrences. We can say then that the probability for reelection is of 0.605. Alternatively, the chance of being reelected can be calculated by the division between the probabilities (yes/no), here, $0.605/0.395 = 1.53$. Table 4 illustrates this information.

Considering only candidates involved in corruption scandals, the reelection rate was 17.86%, since 10 out of 56 representatives got a new term. This means that, for this group, the probability for reelection is 0.179 and the chance for reelection is 0.22. For the candidates not involved in corruption scandals, the chance of being reelected is 1.9. Ultimately, in our replication example, the logistic regression consists of the comparative analysis of the reelection percentage of candidates involved in corruption scandals and those not involved.

In terms of the model’s general fit, one of the main tests used is the Hosmer and Lemeshow (2000). This test is considered more robust than a common chi-square, especially when there are continuous independent variables or when the sample’s size is small (Garson, 2011). Table 5 summarizes the information of interest (value of the test, degrees of freedom, and statistical significance) for Hosmer and Lemeshow tests, and Table 6 shows the same for the Omnibus test of model coefficients.

A non-significant result ($p > 0.05$) suggests that the model estimated with the independent variables is better than the null model. The estimated model has a chi-square ($\chi^2$) of 6.832 and a $p$-valor of 0.555, suggesting an adequate fit. An-
other commonly used adjustment measure is the Omnibus test of model coefficients. It is a chi-square test comparing the model’s variance with the independent variables and the null model (just the intercept).

Unlike the Hosmer and Lemeshow test, a significant result (p < 0.05) suggests an adequate fit. According to the data, the model has a chi-square of 56.356 (p-value < 0.001), that is, the fitted model is better than the null model. We should conclude that the independent variables influence the dependent variable’s variation26. We do not find these tests in Castro and Nunes’s paper (2014), nor the computational scripts. Table 7 summarizes the coefficients estimated by the logistic regression model in an attempt to reproduce the results reported in Table 6 of Castro and Nunes (2014).

As with a linear regression, the first step is to analyze the estimated coefficients (β). Here, the research must observe the sign of the estimates and compare them with the direction expected in their hypotheses. X 11 (Scandal) has a negative effect (-1.677) on the probability of reelection. Unlike a linear model, logistic regression coefficients does not have a direct interpretation.

Table 5 - Hosmer and Lemeshow Test

| $\chi^2$ | gl | Sig  |
|----------|----|------|
| 6.832    | 8  | 0.555|

Source: The authors.

Table 6 - Omnibus test of model coefficients

| $\chi^2$ | gl | Sig  |
|----------|----|------|
| 56.356   | 11 | 0.000|

Source: The authors.

26For Garson (2011), the omnibus test can be interpreted as a test for the joint capacity of all the predictors in the model to predict the response (dependent) variable. A significant result indicates that the fit is adequate to the data, suggesting that at least one of the predictors is significantly

As with a linear regression, the first step is to analyze the estimated coefficients (β). Here, the research must observe the sign of the estimates and compare them with the direction expected in their hypotheses. X 11 (Scandal) has a negative effect (-1.677) on the probability of reelection. Unlike a linear model, logistic regression coefficients does not have a direct interpretation.

Table 7 - Logistic regression model coefficients*

|          | β    | Standard error | Z(Wald) | Sig  | Exp(β) | (exp(β)-1) x 100 |
|----------|------|----------------|---------|------|--------|------------------|
| (Intercept) | 0.552 | 1.568          | 0.352   | 0.725 | 1.737  | 73.734           |
| Poverty   | 1.171 | 1.419          | 0.825   | 0.409 | 3.224  | 222.386          |
| Male      | -0.005 | 0.560       | -0.009  | 0.993 | 0.995  | -0.484           |
| Age       | -0.014 | 0.017       | -0.830  | 0.406 | 0.986  | -1.409           |
| Education | -0.060 | 0.161       | -0.370  | 0.712 | 0.942  | -5.789           |
| Ideology  | -0.125 | 0.224       | -0.561  | 0.575 | 0.882  | -11.782          |
| Vote Increase | 0.908 | 0.341       | 2.663   | 0.008 | 2.480  | 148.030          |
| Change    | 0.078 | 0.382       | 0.205   | 0.838 | 1.081  | 8.136            |
| Parliamentary amendments | **-0.272** | **0.639** | **-0.425** | **0.671** | **0.762** | **-23.785** |
| Candidate/seats | -0.005 | 0.009       | -0.516  | 0.606 | 0.995  | -0.469           |
| Campaign spending | **0.000** | **0.000** | **3.920** | **0.000** | **1.000** | **0.000**       |
| Scandal   | **-1.677** | **0.528** | **-3.176** | **0.001** | **0.187** | **-81.299** |

Source: The authors.

Dependent variable: reelected.

* As with any regression model, the unstandardized coefficients of variables in different scales cannot be directly compared. STATA has a command (listcoef, std help) which produces standardized coefficients in the independent, dependent, and both variables. Menard (2004) presents six different ways to standardize coefficients in a logistic regression.
There are two main ways of reading the coefficients: a) analyze the odds ratio and b) turn the odds ratio into a percentage. With the former, we conclude that involvement in corruption scandals reduces the chances of being elected. In terms of percentages, being involved in corruption diminishes in 81.2% the probability of being reelected, as theoretically expected by hypothesis 1. When considering campaign expenses, the effect was null, with an \( \text{Exp (}\beta\text{)} = 1.000 \).

As in Castro and Nunes (2014), we did not find significant effects of the parliamentary amendment variable on the chance of reelection, considering the magnitude of the p-value and the standard error twice as large as the estimate of the impact itself\(^{27}\).

After analyzing the coefficients associated with the variables of interest, the next step is to evaluate the quality of the model’s fit. Table 8 summarizes some goodness-of-fit measures typically reported in models estimated by the maximum likelihood\(^{28}\).

It is common for statistical packages to show in the output the number of iterations used by the computer to estimate the model. Informing that the model converged after iteration 5 means that the coefficients were estimated via maximum likelihood. Generally, the faster a model converges (less iterations), the better. If the model does not converge, the coefficients are unreliable. One of the main factors that explain a model’s non-convergence is the insufficiency of cases in relation the number of independent variables included in the model.

According to Menard (2002), the log likelihood is a measure of parameter selection in the logistic regression model. However, most statistical packages report the -2 log likelihood (-2LL) and its interpretation is as follows: the larger it is, the worse is the model’s explanatory/predictive capacity. Intuitively, it can be interpreted as a measure of the error when trying to use a determined set of independent variables (model) to explain the dependent variable’s variation.

The researcher can request the iteration history of the estimation. The procedure will produce the -2 log likelihood of the null and the fitted models. The difference between them is measured with a chi-square. As it is an error measure, the larger the chi-square, the larger is the error reduction of the fitted model (with the independent variables), in relation to the null model.

Table 8 presents the value of -2LL to make comparing the models easier. In the null model the -2LL was 3,057,559 and the model with independent variables was 237,4225. In this case, we observe a considerable reduction. This means that the model with the independent variables has a superior fit to the null model. Similarly, the BIC (Bayesian Information Criterion) is another measure based on maximum likelihood. The smaller, the better. The model tested has a BIC of 301.891, while the null model’s was 3,066.105. We can extrapolate that and compare several models, not just the null model.

Unlike the linear model, a logistic regression does not have a synthetic measure of the variation in the dependent variable explained by the model, such as

Table 8 - Model goodness-of-fit measures\(^{a}\)

| -2log likelihood null | -2log likelihood | Cox & Snell R\(^2\) | Nagelkerke R\(^2\) | BIC     |
|-----------------------|------------------|---------------------|-------------------|---------|
| 3,057,559             | 237,4225         | 0.229               | 0.308             | 301.891 |

Source: The authors.

\(^{a}\) The - 2 log likelihood (-2LL) statistic is a fit measure. The smaller it is, the better the fit. The researcher may use it to compare the fit of different models (including and removing independent variables, but keeping the same dependent variable).
the coefficient of determination\(^2\). However, some measures were developed to guide the researcher regarding the explanatory/predictive power of the model\(^3\). The most commonly used are Cox & Snell’s pseudo R\(^2\) of and Nagelkerke’s\(^3\) pseudo R\(^2\). For Menard (2002),

\[
R^2 \text{ is a proportional reduction in } -2\text{LL or a proportional reduction in the absolute value of the log-likelihood measure, where } () \text{ the quantity being minimized to select the model parameters – is taken as a measure of ‘variation’ (Menard, 2002, p. 25).}
\]

For the purposes of this paper, we adopted the following interpretation: the closer to zero, the smaller is the difference between then null model (without any independent variables) and the estimated model. The closer to one, the larger is the difference between the null model and model proposed by the research. At an extreme, a pseudo R\(^2\) of zero indicates that the independent variables included do not help to explain the variation of the dependent variable. A pseudo R\(^2\) of 1 suggests that the variables explain/predict the variation in Y perfectly. Keeping in mind that we should be less demanding of a logistic model than a linear model in terms of variance explained by the R\(^2\).

Lastly, a researcher must analyze the classification table. This report is particularly interesting because it gives a measure of the model’s predictive capacity. Table 9 illustrates the information of interest.

The classification table is frequently referred to as a confusion table. For Garson (2011),

Although classification hit rates (percent correct) as overall effect size measures are preferred over pseudo-R\(^2\) measures, they have some severe limitations for this purpose. Classification tables should not be used exclusively as goodness-of-fit measures because they ignore actual predicted probabilities and instead use dichotomized predictions based on a cutoff (ex.: 0.50). For instance, in binary logistic regression, predicting a 0-or-1 dependent, the classification table does not reveal how close to 1.0 the correct predictions were nor how close to 0.0 the errors were. A model in which the predictions, correct or not, were mostly close to the .50 cutoff does not have as good a fit as a model where the predicted scores cluster either near 1.0 or 0.0. Also, because the hit rate can vary markedly by sample for the same logistic model, use of the classification table to compare across samples is not recommended. (Garson, 2011, p. 173).

Our classification matrix uses the conventional standard of 50% to allocate cases as 1 (if the predicted probability is higher than 0.5) or 0 (smaller than 0.5). We can evaluate this table using three concepts: accuracy, sensibility, and specificity. The accuracy of the model is the proportion of true positive and true negative cases. According to Table 9, the accuracy of our model was of 71.89% (23.50% + 48.29%). However, the accuracy of a model is not always the most important aspect. In certain cases, what is important is maximizing the rate of true positives or true negatives.

| Predicted | Total |
|-----------|-------|
|           | Not reelected | Reelected |       |
| Real      |              |           |       |
| Not reelected | 23.50 | 17.51 | 41.01 |
| Reelected  | 10.60 | 48.39 | 58.99 |
| Total     | 34.10 | 65.90 | 100.00 |

Source: The authors.
Moving on to sensibility. It is the percentage of cases that has the feature of interest (was reelected) that were accurately predicted by the model (true positives / false positives + true positives). In our example, 48.39% of reelected candidates were correctly classified, out of a total of 58.99% that were actually reelected. This gives us a sensibility of 82.03% (48.39%/58.99%). The specificity of the model is the percentage of cases that do not have the feature of interest (were not reelected), that were correctly classified by the model, that is (true negatives / false negatives + true negatives). As we can see, 23.50% of non-reelected candidates were correctly identified out of a total of 41.01% of non-reelected. This gives us a specificity of 57.30% (23.50%/41.01%). There is a trade-off between sensibility and specificity. When increasing one, the other diminishes. Although sometimes the sensibility of the model is more important (predicting an illness, since one would be able to treat it), at other times it is best to increase specificity (keep corrupt politicians from being elected).

VI. Conclusion

We hope to help students and teachers to better understand how logistic regression works. The absence of calculus, linear and matrix algebra, and advanced statistics limits our ability to understand more advanced data analysis techniques. For this reason, our approach focused on the intuitive exposition of results. We also believe that understanding the intuitive logic of logistic regression is the first step to better understanding the different procedures that exist to deal with categorical data. Computational advances allow researchers with less specific training in Mathematics and Statistics to benefit from the advantages associated with the different multivariate techniques. Given that many variables in Political Science are categorical, the analytical benefits associated with the correct application and interpretation of a logistic model are evident. With this paper, we hope to disseminate the use of logistic regression.

And how to improve the quality of methodological and technical training offered to Political Science undergraduate and graduate students in Brazil? We recommend the following: (1) incorporate of replication as a pedagogical tool in data analysis disciplines; (2) mandatory disciplines on mathematics, calculus, probability, and statistics in undergraduate and graduate curricula. In addition, students must receive training in some programming language; (3) conduct practical exercises involving data analysis with topics typical of Political Science. The emphasis on ABSTRACT problems reduces students’ interests on the topic; (4) incentivize student participation in winter/summer courses such as MQ-UFGM and IPSA-USP; (5) promote epistemology and philosophy of science disciplines. The definition of research methods and techniques depend on the epistemological view of what is scientific knowledge and how it should be implemented; (6) diffuse critical reading of papers that use advanced data analysis techniques; (7) keep up with the academic production of journals specialized in methodology such as, for instance, Political Analysis and Political Science Research and Methods; (8) encourage the publication of methodological papers in national journals; (9) foster the creation of research groups and round-tables on methodology and data analysis techniques in professional conferences; (10) fund research projects especially devoted to deepening the knowledge on the main feature of science: method.

Antônio Alves Tôrres Fernandes (antonio.alvestorres@ufpe.br) is a master’s student in Political Science in the PPGCP/UFPE, an undergraduate student in Economics (UCB/DF), and a member of the Research Methods in Political Science Group (DCP/UFPE).

Dalson Britto Figueiredo Filho (dalson.figueiredofo@ufpe.br) is professor at Political Science Graduate Program at Federal University of Pernambuco and author of the book “Quantitative methods in Political Science”, Editora InterSaberes.
Enivaldo Carvalho da Rocha (enivaldocrocha@gmail.com) is a retired full professor at Political Science Graduate Program at Federal University of Pernambuco.

Willber da Silva Nascimento (nascimento@willber@gmail.com) holds a PhD in Political Science from UFPE and works as post-doctoral researcher at PPGCP/UFPE/FACEPE.

References

Altman, D. (1991) Categorising continuous variables. British Journal of Cancer, 64(5), p. 975. DOI: 10.1038/bjc.1991.441
Bonney, G. (1987) Logistic regression for dependent binary observations. Biometrics, 43(4), pp. 951-973. DOI: 10.2307/2531548
Brant, R. (1996) Digesting logistic regression results. The American Statistician, 50(2), pp. 117-119. DOI: 10.2307/2684422
Castro, M.M.M. & Nunes, F. (2014) Candidatos corruptos são punidos?: accountability na eleição brasileira de 2006. Opinião Pública, 20(1), pp. 26-48. DOI: 10.1590/S0104-62762014000100002
Codato, A.; Cervi, E. & Perissinoto, R. (2013) Quem se elege prefeito no Brasil? Condicionantes do sucesso eleitoral em 2012. Cadernos Adenauer, 14(2), pp. 61-84.
Cohen, J. (1983) The Cost of Dichotomization. Applied Psychological Measurement, 7(3), pp. 249–253. DOI: 10.1177/014662168300700301
Cook, R. & Weisberg, S. (1989) Graphics for assessing the adequacy of regression models. Journal of the American Statistical Association, 92(438), pp. 490-499. DOI: 10.1080/01621459.1997.10474002
R Core Team. (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Disponível em: https://www.R-project.org/. Acesso em: 28 set. 2020.
DeMaris, A. (1995) A tutorial in logistic regression. Journal of Marriage and the Family, pp. 956-968. DOI: 10.2307/353415
Efron, B. & Tibshirani, R. (1993) An introduction to the bootstrap. Chapman & Hall/CRC.
Efron, B. & Tibshirani, R. (1986) Multiple imputation for Mismeasured Variables and Missing Data. Journal of the American Statistical Association, 81(396), pp. 93-109. DOI: 10.2307/2289545
Figueiredo Filho, D. & Silva Júnior, J. (2012) Classificando regimes políticos utilizando análise de conglomerados. Revista de Sociologia e Política, v. 28, n. 74, pp. 413-425. DOI: 10.1590/S0104-62762012000100006
Figueiredo Filho, D. & Silva Júnior, J. (2014) Classificando regimes políticos utilizando análise de conglomerados. Revista de Sociologia e Política, v. 28, n. 74, pp. 413-425. DOI: 10.1590/S0104-62762012000100006
Figueiredo Filho, D. & Silva Júnior, J. (2016) O outlier que perturba o seu sono: Como identificar casos extremos? In: 10º Encontro da Associação Brasileira de Ciência Política, Belo Horizonte. Disponível em: https://cienciapolitica.org.br/system/files/documentos/veic/2017/04/outlier-que-perturba-seu-sono-como-identificar-e-manjar.pdf. Acesso em: 13 out. 2020.
Figueiredo Filho, D. & Silva Júnior, J. (2010) Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). Revista Política Hoje, 18(1), pp. 115-146. Disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/view/3852>. Acesso em: 15 de maio 2020.
Figueiredo Filho, D. et al. (2011) O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). Revista Política Hoje, 20(1), pp. 44-99. Disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/view/3808/31622>. Acesso em: 15 de maio 2020.
Figueiredo Filho, D. et al. (2019) Seven Reasons Why: A User’s Guide to Transparency and Reproducibility. Brazilian Political Science Review, 13(2), pp. e0001. DOI: 10.1590/1981-3821201900200001
Figueiredo Filho, D.; Silva, L. & Domingos, A. (2015) O Que é e Como Superar a Multicolinearidade? Um Guia Para Ciência Política. Conexão Política, 4(2), pp. 95-104. DOI: 10.26694/rcp.issn.2317-3254.v4e2.2015.p95-104
Guthery, F. & Binghan, R. (2007) A primer on interpreting regression models. The Journal of Wildlife Management, 71(3), pp. 684-692. DOI: 10.2193/2006-285
Hagley, T. & Mitchell, G. (1992) Goodness-of-fit measures for probit and logit. American Journal of Political Science, 36(3), pp. 762-784. DOI: 10.2307/2111590
Hair, J. et al. (2009) Análise multivariada de dados. Porto Alegre: Bookman Editora.
Henderson, E. & Singer, J. (2000) Civil war in the post-colonial world, 1946-92. Journal of Peace Research, 37(3), pp. 275-299. DOI: 10.1177/002234330037003001
Henderson, E. & Singer, J. (2008) Regime type and international conflict: towards a general model. Journal of Logistic Regression: Binary and Multinomial.
Speck, B. & Mancuso, W. (2013) O que faz a diferença? Gastos de campanha, capital político, sexo e contexto municipal nas eleições para prefeito em 2012. *Cadernos Adenauer*, 14(2), pp. 109-126.

Stock, J. & Watson, M. (2015) *Introduction to Econometrics*. 3ª Edition. United Kingdom: Pearson.

Tabachnick, B.; Fidell, L. & Ullman, J. (2007) *Using multivariate statistics*. Boston, MA: Pearson.

Taylor, J. & Yu, M. (2002) Bias and Efficiency Loss Due to Categorizing an Explanatory Variable. *Journal of Multivariate Analysis*, 83(s/n), pp. 248–263. DOI: 10.1006/jmva.2001.2045

Wong, G. & Mason, W. (1985) The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80(391), pp. 513-524. DOI: 10.1080/01621459.1985.10478148

Newspaper articles

O julgamento do Mensalão (2012) *Folha de São Paulo*. São Paulo. 12. Jun. Available through: <https://www1.folha.uol.com.br/especial/2012/ojulgamentodomensalao/>. Accessed at: 10 nov. 2020.

Entenda o Escândalo dos sanguessugas (2006) Estado de São Paulo. São Paulo. 11.dec. Available through: <https://politica.estadao.com.br/noticias/geral,entenda-o-escandalo-dos-sanguessugas,20061211p60113>. Accessed at: 10 nov. 2020.

Leia este artigo se você quiser aprender regressão logística

**RESUMO:** Introdução: E se a minha variável resposta for categórica binária? Este artigo apresenta uma introdução intuitiva à regressão logística, técnica estatística mais adequada para lidar com variáveis dependentes dicotômicas. **Materiais e Métodos:** estimamos o efeito dos escândalos de corrupção sobre a chance de reeleição de candidatos concorrentes a deputado federal no Brasil a partir dos dados de Castro e Nunes (2014). Em particular, mostramos a implementação computacional no R e explicamos a interpretação substantiva dos resultados. **Resultados:** disponibilizamos todos os materiais de replicação, o que por sua vez permite que estudantes e profissionais utilizem os procedimentos discutidos aqui em suas atividades de estudo e pesquisa. **Discussão:** esperamos incentivar o uso da regressão logística e difundir a replicabilidade como ferramenta de ensino de análise de dados.

**PALAVRAS-CHAVE:** regressão; regressão logística; replicação; métodos quantitativos; transparência.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

A produção desse manuscrito foi viabilizada através do patrocínio fornecido pelo Centro Universitário Internacional Uninter à *Revista de Sociologia e Política*. 

16/19

Revista de Sociologia e Política v. 28, n. 74
Appendix

In this section, we present some information that can help researchers to interpret logistic regression coefficients. In particular, we examine the interpretation of the odds ratio. In addition, we list some learning tools.

- Understanding the odds ratio

The term odds ratio is not as disseminated in Political Science applied research as are mean or probability. Usually, since the researcher is comparing groups/categories, they are interested in analyzing which group/category has a better chance of occurring in relation to another group/category. Consider the following example: suppose that the probability (p) of a certain event occurring is 0.9. Thus, when calculating the complementary event, q = 1 – p, we have 1 – 0.9 = 0.1. Chance is the division of the probability of occurrence (p) by the probability of non-occurrence (q). Consequently, 0.9/0.1 = 9. It is stated, then, that the chance for success is 9 to 1. Alternatively, the chance for failure is 0.1/0.9 = 0.11. We say then that the chance for failure is 1 to 9. Unlike probability, which can only take on values between 0 and 1, chance can vary between 0 and infinity. When the probability of an event occurring is greater than the probability of it not occurring, its chance will be greater than 1. When the probability of it not occurring is greater, chance will be smaller than 1. When probabilities are equal (e.g., tossing a coin), chance is equal to 1. Given the pedagogical purposes of this paper, it is relevant to replicate the data from Schwab (2002), to better grasp this concept (Table 1A).

Table 1A - Frequency

| Sentence            | N  | %   |
|---------------------|----|-----|
| Death penalty       | 50 | 34  |
| Life in prison      | 97 | 66  |
| Total               | 147| 100.0|

Source: Schwab (2002).

Table 1A shows that 34% of inmates were sentenced to the death penalty (n = 50/147). This means that the probability of this event occurring is 0f 0.34. Alternatively, the chance of being given capital punishment is 0.516 (50/97). Another way of saying this is that the chances are approximately half of being sentenced to capital punishment in relation to spending life in prison. Lastly, it is possible to invert the interpretation and consider life in prison roughly two times more likely than the death penalty.

So far, there are no independent variables. What the logistic model will inform is the impact of a given variable on the chance of a dependent variable occurring. For example, consider the relationship between race and sentence type (Table 2A).

Table 2A – Sentence type by color

| Sentence        | Black | Non-black | Total |
|-----------------|-------|-----------|-------|
| Death penalty   | 28    | 22        | 50    |
| Life in prison  | 45    | 52        | 97    |
| Total           | 73    | 74        | 147   |

Source: Schwab (2002).

It is possible, then, to calculate the chance for each specific group: black people and non-black people. For black people, we have 28/45 = 0.622. For non-black people, we have 22/52 = 0.423. The impact of being black can be rep-
resented by the division of a black person receiving the death penalty and a non-black person receiving capital punishment (0.423). 0.622/0.423 = 1.47. For the interpretation: a) black people have 1.47 higher chance of receiving the death penalty than non-black people; b) being black increases by 47% the chances of receiving capital punishment (1.47-1*100).

Learning tools

http://www.icpsr.umich.edu/icpsrweb/sumprog/

Internationally, the Summer Program in Quantitative Methods of Social Research (ICPRS) is one of the main initiatives in the dissemination of research methods and techniques.

http://www.fafich.ufmg.br/~mq/index.html

Intensive course in Quantitative Methodology in the Humanities. It is the most traditional course in teaching of research methods and techniques in Social Sciences in Brazil.

http://summerschool.ipsa.org/

Summer school organized by the International Political Science Association, the Department of Political Science, and the Institute for International Relations of the University of São Paulo (USP).

http://gking.harvard.edu/

Gary King shares papers on methodology, specific software, and databases for researchers interested in replication.

http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm

David Garson presents different topics in multivariate statistics, using the Statistical Package for Social Sciences. At the end of each section, there is a suggested bibliography that can be used as reference to gain more in-depth knowledge on the topic.

http://www.statsoft.com/textbook/

Has different multivariate techniques using the software Statistica.

http://www.ats.ucla.edu/stat/

Website for the University of California (UCLA) specialized in multivariate techniques. Here, the user finds applications for different software (SAS, SPSS, STATA, R, etc.), including video-classes and tutorials.

http://www.socr.ucla.edu/SOCR.html

At this address, the reader finds games, applications, analyses, among other tools related to teaching Statistics and different research techniques.

http://pan.oxfordjournals.org/

Political Analysis is one of the most influential journals in contemporary Political Science and publishes papers in the field of methodology.

http://www.amstat.org/publications/jse/

Journal specialized in the publication of teaching and learning techniques in Statistics.
The journal Política Hoje, from UFPE’s Department of Political Science, recently published a special issue dedicated to Methodology and Epistemology in Political Science and International Relations.