Logistic Regression and the American National Election Study 2012: Vote Choice in the 2012 US Presidential Election

Contributors: The Odum Institute
Title: "Logistic Regression and the American National Election Study 2012: Vote Choice in the 2012 US Presidential Election"
Online Pub. Date: 2015
Access Date: May 06, 2015
Publishing Company: SAGE Publications, Ltd.
City: 55 City Road
Online ISBN: 9781473937949
DOI: http://dx.doi.org/10.4135/9781473937949

This dataset example introduces readers to logistic regression, often simply called logit. This technique allows researchers to evaluate whether a dichotomous dependent variable is a function of one or more independent variables. The logit model is most commonly estimated via maximum likelihood estimation (MLE). This example uses a subset of data from the 2012 American National Election Study. It presents an analysis of whether survey respondents reported voting for Barack Obama or Mitt Romney for U.S. President in 2012 and whether that vote choice can be predicted by several factors, including a respondent's race/ethnicity and how they feel about the Democratic and Republican Parties. An analysis like this allows researchers to test theories regarding what types of factors predict voting behavior.

In this example, readers are introduced to the basic theory and assumptions underlying this technique, the type of question this technique can be used to answer, and how to produce and report results. The sample dataset has been cleaned and organized to make this example easier to follow. Interested readers should consult the full documentation for the dataset before using it for research (http://www.electionstudies.org/).

Introduction

This dataset example introduces readers to logistic regression, often simply called logit. This technique allows researchers to evaluate whether a dichotomous dependent variable is a function of one or more independent variables. The logit model is most commonly estimated via maximum likelihood estimation (MLE).

This example describes logistic regression, discusses the assumptions underlying it, and shows how to estimate and interpret logit models. We illustrate logistic regression using a subset of data from the 2012 American National Election Study (http://www.electionstudies.org/). Specifically, we test whether reported vote choice in the 2012 U.S. Presidential election is predicted by several factors, including a respondent's race/ethnicity and how they feel about the Democratic and Republican Parties. An analysis like this allows researchers to test theories regarding what types of factors predict voting behavior.

What is Logistic Regression?

Logit models explain variation in a dichotomous dependent variable as a function of one or more independent variables. Dichotomous variables divide observations into two mutually exclusive and exhaustive categories, most commonly by coding the two outcomes as either “1” or “0”, where “1” indicates the presence some attribute or behavior and “0” indicates its absence. In this example, we test whether Presidential vote choice can be predicted by several independent variables.

Logistic regression is one example from the family of Generalized Linear Models (GLMs). GLMs connect a linear combination of independent variables and estimated parameters – often called the linear predictor – to a dependent variable using a link function. The link function typically involves some sort of non-linear transformation, which in the case of logistic regression means that the probability that the dependent variable equals 0 or 1 is a non-linear function of the independent variables. The parameters of GLMs are typically estimated using Maximum Likelihood Estimation (MLE). Because logit models are estimated via MLE, it is best if the dataset has a sufficiently large number of observations. Just how many is open to debate, but in his book *Regression Models for Categorical and Limited Dependent Variables* (Sage, 1997), J. Scott Long suggests trying to meet two criteria: 1) have at least 100 observations total, and 2) have at least 10 observations for each coefficient estimated in the model.

In simple terms, MLE is an iterative process that approximates estimates for the coefficients that maximize the fit of the model to the sample of data. By maximizing fit, MLE also minimizes the unexplained variance in the dependent variable. In that sense, MLE accomplishes the same objective as ordinary least squares (OLS) does for standard regression.

When computing statistical tests, it is customary to define the null hypothesis (H_0) to be tested. In logistic regression, the standard null hypothesis is that each coefficient is equal to zero. The actual coefficient estimates will not be exactly equal to zero

in any particular sample of data simply due to random chance in sampling. The t-tests conducted to test each individual coefficient are designed to help determine if the coefficients are different enough from zero to be declared statistically significant. “Different enough” is typically defined as producing a test statistic with a level of statistical significance, or p-value, that is less than 0.05. This would lead us to reject the null hypothesis (*H*
0

) that the coefficient in question equals zero.

Estimating a Logit Model

To make this example easier to follow, we focus for now on estimating a logit model with just two independent variables. The model can be written down in three parts: the linear combination of the independent variables, often called the linear predictor:

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (1)$$

The link function for the logit model that illustrates how the linear predictor relates to the probability that the dependent variable equals 1:

$$\eta_i = ln(\frac{Pr(Y_i = 1)}{1 - Pr(Y_i = 1)}) \quad (2)$$

And the inverse of the link function that rearranges the terms from the link function to isolate the probability that the dependent variable equals 1:

$$Pr(Y_i = 1) = \frac{e^{\eta}}{1 + e^{\eta}} \quad (3)$$

Where:

- # = the linear combination of the independent variables, or the linear predictor
- *X*
  1i

  = individual values of the first independent variable
- *X*
  2i

  = individual values of the second independent variable
- #
  0

  = the intercept, or constant, associated with the logit model
- #
  1

  = the coefficient operating on the first independent variable
- #
  2

  = the coefficient operating on the second independent variable
- *Y*
  i

  = individual values of the dependent variable, typically coded as 0 or 1
- # = the unmodeled random, or stochastic, component of the dependent variable, often called the error term or the residual of the model
- *Pr*(*Y*
  i

  = 1) = the probability that *Y*
  i

  equals 1
- *ln* = the natural log

- e = the exponential function, which serves as the base for the natural log such that $e^{ln(a)} = a$ and $ln(e^a) = a$.

Researchers have values for *Y*
i

, *Y*
1i

, and *X*
2i

in their datasets – they use MLE to estimate values for #
0

, #
1

, and #
2

. Unlike standard multiple regression, the coefficients #
1

and #
2

cannot be directly interpreted as slope coefficients that describe the marginal effect of each independent variable on the probability that Y = 1. Interpreting the coefficient estimates of a logit model is more complicated, and is something described below in the context of a specific example.

Assumptions Behind the Model

Nearly every statistical model or test relies on some underlying assumptions, and they are all affected by the mix of data you happen to have. Different textbooks present the assumptions for a logistic regression model in different ways. Here are the key factors to consider when estimating a logistic regression:

- The dependent variable must be dichotomous (though there are variants of logistic regression where this is not required).
- The model is correctly specified (e.g. we have the right independent variables in the model properly measured).
- The values of the independent variables are fixed in repeated samples.
- The individual residuals are independent of each other.
- Because it is generally estimated via MLE, logistic regression requires moderate to large sample sizes.

Illustrative Example: Vote Choice in the 2012 U.S. Presidential Election

This analysis examines whether the probability of voting for Obama versus Romney in the 2012 Presidential election is related to a respondent's feelings about the two main political parties, their own race, and their income level. The specific research questions are:

- Are respondents with more positive feelings about the Democratic Party more likely to vote for Obama as the Democratic candidate?
- Are respondents with more positive feelings about the Republican Party less likely to vote for Obama as the Democratic candidate?
- Are members of any racial or ethnic minority groups more likely to vote for Obama than are white voters?
- Are those with higher incomes less likely to vote for Obama?

Each of these research questions could be stated in the form of a null hypothesis:

- *H*
  0

  a = After accounting for the effects of other variables in the model, feelings about the Democratic Party will be unrelated to the probability of voting for Obama.
- *H*
  0

  b = After accounting for the effects of other variables in the model, feelings about the Republican Party will be unrelated to the probability of voting for Obama.
- *H*
  0

  c = After accounting for the effects of other variables in the model, members of minority groups will be no more or less likely to vote for Obama compared to whites.
- *H*
  0

  d = After accounting for the effects of other variables in the model, income will be unrelated to the probability of voting for Obama.

The Data

This example uses data from the 2012 ANES. We use several variables:

- Presidential vote choice (vote_obama), coded 1 = Obama, 0 = Romney.
- Democratic Party feeling thermometer (ft_dem), coded from 0 to 100, with higher scores indicating more positive feelings.
- Republican Party feeling thermometer (ft_dem), coded from 0 to 100, with higher scores indicating more positive feelings.
- Indicator variable for black voters (black), coded 1 for black respondents and 0 otherwise.

- Indicator variable for Hispanic voters (hispanic), coded 1 for Hispanic respondents and 0 otherwise.
- Indicator variable for voters of other non-white races or ethnicities (other), coded 1 for other minority respondents and 0 otherwise.
- Respondent income (income), coded into categories ranging from 1 to 28, where 1 = incomes under $5,000 per year and 28 = incomes of $250,000 or more per year.

The dependent variable is vote choice, which is dichotomous, making this example appropriate for logistic regression.

Analyzing the Data

Before proceeding to the logistic regression model, it is a good idea to produce a frequency distribution of the dependent variable. Remember that the dependent variable records the reported vote choice of respondents. It is coded “1” if the respondent voted for Obama and “0” if the respondent voted for Romney. Those who did not vote and the small number who voted for someone else are excluded from this example for simplicity. That leaves us with 3870 respondents, 1598 of whom reported voting for Romney and 2272 of whom who reported voting for Obama. Logit models do not perform as well if there are small numbers of observations in one of the two categories or if there is a substantial skew in the distribution of observations across the two categories. We have neither of those problems here.

It would also be valuable to produce summary statistics and explore the distributions of each of the independent variables as well. However, in the interest of space, we will forgo doing so now.

The results of the logistic regression model itself are presented in Table 1.

	Dep. Var: Vote for Obama
Intercept	-0.224 (0.253)
Dem Feeling Therm	0.094 (0.004)***

Rep Feeling Therm	-0.091 (0.004)***
African-American	3.166 (0.395)***
Hispanic	0.970 (0.189)***
Other	0.559 (0.254)*
Income	-0.025 (0.009)**
AIC	1481.759
BIC	1525.586
Log Likelihood	-733.879
Deviance	1467.759
Num. obs.	3870

*** $p < 0:001$

** $p < 0:01$

* $p < 0:05$

The top portion of the table reports the individual parameter estimates, their estimated standard errors, and indicators of statistical significance. The bottom portion of the table reports four measures of relative model fit and the sample size. Each coefficient estimate operating on each independent variable is positive or negative in the direction we would expect and each is statistically significantly different from zero. However, just looking at logit coefficients and tests of statistical significance does not tell the whole story. We explore some of the findings in greater detail through computing factor changes in the odds and predicted probabilities.

Factor Change in the Odds

The dependent variable in a logit model can be thought of as the natural log of the odds that the dependent variable equals 1, as previously illustrated in the link function

in Equation 2. Thus we can use the exponential function to determine how much those odds change for a one-unit change in one of our independent variables. For example, according to Table 1, the coefficient estimated for the dummy variable that records whether or not a respondent is African-American equals about 3.166. Taking the exponential of 3.166 produces a value of about ($e^{3.166} \approx 23.7$). This can be interpreted as estimating that the odds of a black respondent voting for Obama are about 23.7 times higher than the odds of a white respondent voting for Obama. The other coefficient estimates in Table 1 appear smaller, but remember that many of the other independent variables have scales much greater than just 0 to 1. Thus a one-unit increase in the feeling thermometer variables, for example, only covers one hundredth of the range of those variables.

Factor changes in the odds are relatively easy to compute, and they provide more of an interpretation than do logit coefficients by themselves, but they still do not provide a complete picture of the results. Knowing that the odds of an event happening increase or decrease by some factor is informative, but only partially so if you do not know the baseline odds from the start. If the odds of winning the lottery are 1 in a million, but something happens and they become 2 in a million, your odds of winning have doubled, but you are still extremely unlikely to win.

Predicted Probabilities

We can go a step further and compute the predicted probability of a respondent voting for Obama based on the results in Table 1 using the inverse link function as shown previously in Equation 3. Because the relationship between all of the independent variables and the probability that Y=1 is nonlinear, you can only compute a predicted probability by setting every independent variable in the model to some specific value.

For example, to compare the predicted probabilities of voting for Obama for each of the four racial/ethnic groups, we need to set the values for the race/ethnic indicator variables to appropriate values and we need to set all of the other independent variables to some fixed value as well. The most common strategy is to set the remaining variables to central measures such as their means, medians, or modes. An alternative

is to compute the predicted probability for each observation based on its own values for its independent variables, but this makes it harder to isolate the potential effect of any one independent variable. In order to keep it simple, we set all the remaining independent variables to their means.

Table 2 reports the results of these calculations for each of the four racial/ethnic groups included in this analysis. The results show that respondents from each group have a predicted probability of voting for Obama that ranges between 0.71 and 0.98. Predicted probabilities this high might seem surprising given that only 59% of respondents overall reported voting for Obama. However, you must remember that these predicted probabilities are also based on all other independent variables being set to their mean values.

Racial/Ethnic Group	Predicted Prob. of Voting for Obama	95% Confidence Interval
Black	0.98	0.96 – 0.99
Hispanic	0.84	0.80 – 0.89
Other	0.78	0.70 – 0.86
White	0.67	0.64 – 0.71

Tables like Table 2 are helpful when you only need to compute a small number of predicted probabilities to interpret the findings of the model. However, for continuous independent variables, it is better to compute a large number of predicted probabilities and present the results graphically. For example, the feeling thermometer variables can each take on 101 possible values (0 through 100, inclusive). Figure 1 presents the predicted probability of voting for Obama, along with a 95% confidence interval, as a function of the Democratic Party feeling thermometer, holding all other independent variables at their means. The tick marks inside the x-axis show the location of every respondent's observed Democratic Party feeling thermometer score (adjusted slightly so that identical scores are not plotted on top of each other).

Figure 1: Predicted probability of voting for Obama over the complete range of the Democratic Party feeling thermometer while holding all other variables constant at their means, 2012 ANES.

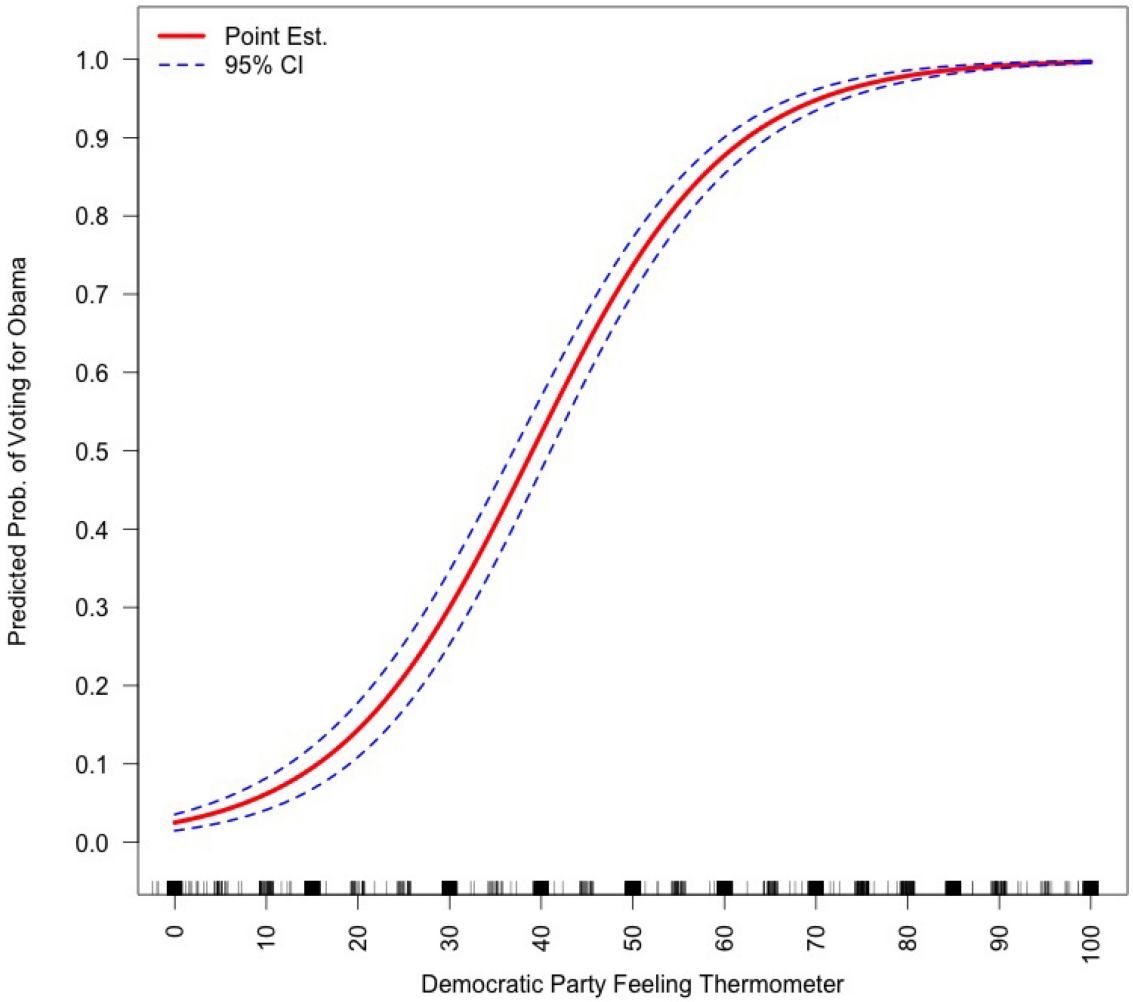

Figure 1 shows that respondents who felt very cold toward the Democratic Party have a very low predicted probability of voting for Obama. Those feeling very warm toward the Democratic Party have a very high predicted probability of voting for Obama. Change in support for Obama is relatively flat at these extremes, but change is much more dramatic at middle ranges of the Democratic Party feeling thermometer.

Complete interpretation of the results would include presenting factor changes in the odds and/or predicted probabilities based on the other remaining variables in the model, but we forgo doing so here in the interest of space.

Presenting Results

The results of a logit model can be presented in a variety of ways. Here we offer one example.

“We used a subset of data from the 2012 ANES to test several null hypotheses:

- *H*
  0

  a = After accounting for the effects of other variables in the model, feelings about the Democratic Party will be unrelated to the probability of voting for Obama.
- *H*
  0

  b = After accounting for the effects of other variables in the model, feelings about the Republican Party will be unrelated to the probability of voting for Obama.
- *H*
  0

  c = After accounting for the effects of other variables in the model, members of minority groups will be no more or less likely to vote for Obama compared to whites.
- *H*
  0

  d = After accounting for the effects of other variables in the model, income will be unrelated to the probability of voting for Obama.

The data included 3870 individual respondents. Results from the logit model are presented in Table 1. Those results show feelings about the Democratic Party are positively linked to the probability of voting for Obama, while feelings about the Republican Party are negatively linked to the probability of voting for Obama. Furthermore, each racial/ethnic group considered in Table 1 is more likely to vote for Obama than are white voters. Finally, as incomes increase, the probability of voting for Obama decreases. Table 2 reports that, after holding other independent variables in the model at their mean values, the probability of voting for Obama was highest among black respondents and lowest among white respondents. Figure 1 also shows the dramatic shift in the probability of voting for Obama across the range of the Democratic Party feeling thermometer. Further interpretation and diagnostic testing should be explored to evaluate the robustness of these findings."

Review

Logistic regression expresses a dichotomous dependent variable as a function of one or more independent variables. Logit models are estimated via MLE. Direct interpretation of the coefficient estimates is limited to whether they are positive, negative, or not statistically significant. To really understand the results of a logit model requires calculating factor changes in the odds or predicted probabilities.

The logit model is widely used in social science. It also serves as a foundation for an array of more complex methods in Statistics, Decision Science, Machine Learning, and Data Mining, including Neural Network Models and Support Vector Machines.

You should know:

- What types of variables are suitable for logistic regression.
- The basic assumptions behind the logit model.
- How to estimate and interpret the results of a logit model.
- How to report the results from a logistic regression.

Your Turn

You can download the sample dataset along with a guide showing how to estimate a logit model using statistical software. See if you can reproduce the results presented here, then try producing your own logit model by adding another independent variable named sex, which is coded 0 = Male and 1 = Female.

About This Dataset

Data Source Citation

The American National Election Studies (ANES; http://www.electionstudies.org). The ANES 2012 Time Series Study [dataset]. Stanford University and the University of Michigan [producers].

Full title of originating dataset

American National Election Studies (ANES) 2012 Time Series Study

Data author(s) and affiliations

American National Election Studies (Stanford University, University of Michigan)

Dataset source website address

http://electionstudies.org/studypages/anes_timeseries_2012/anes_timeseries_2012.htm

Data Universe

U.S. eligible voters

Funding sources/suppliers

These materials are based on work supported by the National Science Foundation under grants SES-0937727 and SES-0937715, Stanford University, and the University of Michigan.

Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the funding organizations.

Sample/sampling procedures

The ANES 2012 Time Series is a dual-mode survey (face-to-face and Internet) with two independent samples. Cases selected for the face-to-face sample could not be interviewed on the Internet, and cases selected for the Internet survey could not be interviewed in person. The Internet sample was drawn from panel members of GfK Knowledge Networks. The face-to-face sample used an address-based, stratified, multi-stage cluster sample in 125 census tracts. The face-to-face sample also featured oversamples of blacks and Hispanics.

Weighting

Original dataset includes 3 weight variables:

weight_ftf

for face-to-face sample analysis alone

weight_web

for Internet sample analysis alone

weight_full

for combined sample analysis

Data collection dates

09-2012 to 01-2013

Time frame of analysis

2012 to 2013

Unit of analysis

Individual

Location covered by data

United States

Links to SRM content

- Fred C. Pampel (Ed.). (2000). *Logistic Regression*. Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412984805
- Menard, S. (2010). Logistic Regression. In Neil J. Salkind (Ed.), *Encyclopedia of Research Design*. (pp. 731–736). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412961288.n224

- Demaris, A. (1992). *Logit Modeling*. Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412984836
- Han, S., & Swicegood, C. (2004). Logistic Regression. In Michael S. Lewis-Beck, A. Bryman, & Tim Futing Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods*. (pp. 588–590). Thousand Oaks, CA: Sage Publications, Inc. doi: http://dx.doi.org/10.4135/9781412950589.n512

List of variables

ft_dem

Feeling Thermometer – Democratic Party

ft_rep

Feeling Thermometer – Republican Party

race

Race/Ethnicity

income

Income Level

sex

Sex of Respondent

white

Respondent is White

black

Respondent is Black

hispanic

Respondent is Hispanic

other

Race is Other

vote_obama

2-Party Pres. Vote