A Practical Guide to Logistic Regression Using Stata

Alan C. Acock
Oregon State University

A Stata Press Publication
StataCorp LLC
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2$_\varepsilon$
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-415-2
Print ISBN-13: 978-1-59718-415-1
ePub ISBN-10: 1-59718-416-0
ePub ISBN-13: 978-1-59718-416-8

Library of Congress Control Number: 2026934499

Contents

(Pages omitted)

Preface

Logistic regression is widely used throughout many scientific disciplines. Your reading of this preface indicates that you are in one of these disciplines. Logistic regression is the prime choice for any research that classifies observations into binary groups, such as success versus failure. To illustrate how widely logistic regression is used, Alan Rowberg suggested that I ask PubMed, so I searched for the term "logistic regression", which produced a list of 52,082 publications at the time of this writing. Although logistic regression is widely used, its results often are misinterpreted or fail to provide the best possible interpretation. In this book, I aim to present the best ways to interpret logistic regression for both professional and lay audiences.

Suppose that you read an article in a medical journal and find a statement that a mother being a smoker significantly increases the likelihood of her newborn having a low birthweight because the coefficient is positive and the p-value is small. The article might list a coefficient and an odds ratio in a table without much interpretation. You should be careful with statements like this. The p-value may be small even when the effect has no practical significance, so it is important to ask questions that may help your interpretation of the results. What is the strength of the impact? What additional risk or reduction in risk does smoking contribute? What graphs would help readers visualize the relationship?

In this book, I will report the p-value of findings, but my focus is on practical significance rather than statistical significance. A small p-value does not mean that the effect has any practical importance, especially with large samples. Because of this, I am following the suggestion of Wasserstein and Lazar (2016) and not using a specific cutoff, such as $p < 0.05$, as evidence of significance.

Do you have enough statistical background to use this book? Yes, if you have

1. completed a single introductory course in statistics and
2. have a basic knowledge of Stata.

Even if you have a more advanced background, this book can still teach you Stata tricks and the best use of logistic regression.

How to use this book. The chapters in this book are short and focus on specific applications of logistic regression. Ideally, you will work through this short book from beginning to end. However, you may wish to refer back to individual chapters as needed, so there is some duplication of content to allow chapters to stand alone. The appendix

includes a do-file that reproduces nearly everything covered in this book. This do-file should be a valuable reference to consult for your research.

(Pages omitted)

1 What we can do with logistic regression

1.1 Questions we can answer using logistic regression

Logistic regression is appropriate when we wish to model whether something happens. For example, we can learn what factors predict the following:

- Who is more likely to become a type 2 diabetic?
- Who preorders an experimental electric bike on a GoFundMe campaign?
- Which newborns will have a low birthweight?
- What children are ready for the first grade?
- Who drops out of college?
- Who is most likely to vote?
- Which newlyweds will get divorced in the first five years of their marriage?

Not all applications start with a dichotomy like those above. For example, does a mother being a smoker raise the risk of a low birthweight for her newborn? Though birthweight is a quantitative variable, we would generate a binary variable, an indicator of low birthweight, to answer this question by using logistic regression.

Because the World Health Organization defines low birthweight as less than 2,500 grams, our new variable would have categories for a birthweight of less than 2,500 grams and for a birthweight of 2,500 grams or more. The difference between 2,475 grams and 2,525 grams is less than 2 ounces, but one weight is in the category of all birthweights less than 2,500 grams, while the other weight is in the category of birthweights from 2,500 grams to the heaviest child ever born.

A second example is predicting who will graduate from college. Suppose that one's education is coded as 1 for "Less than high school", 2 for "High school", 3 for "Some college", 4 for "College graduate", and 5 for "Postgraduate". We create a new binary variable by combining codes 1–3 to indicate those who have not graduated college and combining codes 4–5 to indicate those who have. This variable lets us answer our question but ignores the difference between "Less than high school" and "Some college".

Information is lost when we dichotomize a quantitative variable, so we should do so only when the dichotomy is the focus of our research question.

1.2 Ways to report results

Once we have determined that logistic regression is appropriate for our research question, we perform the analysis and report the results. Logistic regression results typically report which independent variables are statistically significant and the directions of their effects. Often, this is insufficient. How can we present logistic regression results that are meaningful to both professional and lay audiences?

Let us explore alternative ways of presenting results by using an example of logistic regression that models the relationship between obesity and the risk (probability) of high blood pressure. There are many measures of obesity. One of them is body mass index (BMI), which is based on a person's height and weight. It is defined as

$$\text{BMI} = \frac{\text{Weight(kg)}}{\text{Height(m)}^2}$$

Alternatively, it is defined using imperial measures as

$$\text{BMI} = \frac{\text{Weight(lb)} \times 703}{\text{Height(in)}^2}$$

BMI ignores body composition and is a crude measure of obesity. Nonetheless, it is widely used because it requires knowing only a person's weight and height.

Consider Dr. Ahmed Jones's advice to his patient of 15 years, Monica. For each visit, Dr. Jones's nurse checks Monica's weight and height. From these results, the nurse records Monica's BMI. Dr. Jones notes that Monica's BMI has drifted higher. Over the last 15 years, her BMI of 22 has increased to 32. According to standard BMI guidelines, a BMI of 22 is healthy, but 32 is obese.

Because clinically high blood pressure is a binary outcome (you either have it or do not have it), logistic regression is an appropriate statistical tool to model it. Based on the results of a logistic regression, Dr. Jones could tell Monica that obesity has a statistically significant effect on the risk of high blood pressure, but this may not convince Monica to modify her diet and exercise program. What is better than reporting just the statistical significance or p-values? Because a statistically significant effect does not equal a practically substantial effect, Dr. Jones should explain the practical significance. If he does not do this, it will be easy for Monica to dismiss his advice. What evidence might convince Monica of the danger of a high BMI?

Let us assume that Dr. Jones recently read in a medical article that the logistic regression coefficient (B) of BMI predicting high blood pressure is 0.149, $p < 0.001$. This B is the difference in the natural logarithm of the odds of high blood pressure for a one-point difference in BMI. The interpretation of this is not transparent even to

an academic audience, so it is even less helpful to a typical patient; most people have yet to learn what $B = 0.149$, $p < 0.001$ would mean. Additionally, Monica needs to understand the estimated risk that she will have high blood pressure now that her BMI has increased by 10 points.

Dr. Jones might exponentiate the B, that is, determine e^B, and tell Monica that the odds ratio (OR) is 1.16. The OR is easier to understand than the B. However, this approach has three potential problems:

1. Monica may not understand how to interpret an OR of 1.16.
2. The OR is for a one-point difference in Monica's BMI, for example, from 22 to 23, not for her 10-point difference from 22 to 32. It is not as simple as saying that the odds for a 10-point difference would be 1.16 times 10 as great.
3. The OR is not an estimate of the difference in the risk of high blood pressure associated with BMIs of 22 versus 32. The effect of her BMI on her risk of having high blood pressure is easier to understand than the OR. Neither the B nor the OR is a direct measure of what has happened to Monica's risk of high blood pressure.

Monica needs to learn her risk of high blood pressure at a BMI of 32 compared with her risk at a BMI of 22. To illustrate this for Monica, Dr. Jones could estimate her risk of having high blood pressure (Pr = 0.300) with a BMI of 22 compared with the risk (Pr = 0.654) with a BMI of 32.

Dr. Jones might express these values as percentages. He estimates that Monica's risk of high blood pressure is 30.0% with a BMI of 22, but it is 65.4% with a BMI of 32. Her increased BMI more than doubles her risk of high blood pressure.

If the article reported the estimated risk by BMI level and the relationship had included a graph, Dr. Jones could show it (see figure 1.1) to Monica. Unfortunately, many researchers stop after reporting the B or OR and whether they are statistically significant. Figure 1.1 allows Monica to see more than the effects of her past and current BMI. It shows that by losing weight and reducing her BMI, she could dramatically lower her risk of high blood pressure.

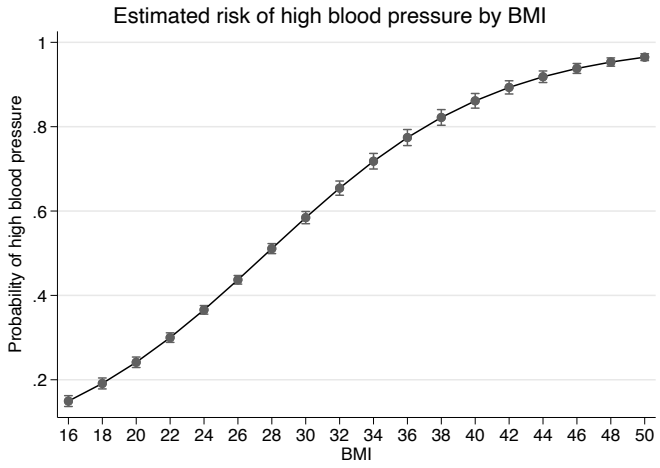

Figure 1.1. Risk of high blood pressure by BMI

1.3 Using Stata

Stata makes it easy to perform logistic regression and obtain the results you want to report. Throughout this book, I will demonstrate how to use Stata for each analysis. If you use another software package such as SAS, R, or SPSS, you can still use this book to learn what you should report to make sense of logistic regression. Many Stata commands have parallel commands in other statistical packages. However, note that even if your data are stored in the format of another package such as SPSS or SAS, you can analyze these data using Stata. It is easy to import SPSS and SAS datasets into a Stata dataset. This will allow you to take full advantage of this book even if you primarily use SPSS or SAS. To do so using Stata's menu system, click on **File > Import > SPSS data (*.sav)** to import an SPSS dataset. Click on **File > Import > SAS XPORT Version 8 (*.v8xpt)** to open a SAS export file. Alternatively, you can click on **File > Import > SAS data (*.sas7bdat)** to import a SAS dataset.

Stata's menu system and dialog boxes make it easy to learn. Alternatively, you can use Stata's parsimonious and consistent commands. These commands can be entered directly into the Command window or into a do-file (see appendix). Each command begins with the command's name followed by a list of variables. With regression commands, the outcome variable is the first variable in the variable list.

Following Stata convention, a dot and a blank space (`. `) are placed in front of Stata commands when they appear on a separate line in the text. The dot and space are there to draw your attention to the command, but the result is the same whether or not we enter the dot and space. When I mention a Stata command or variable name, I will put it in typewriter font. I will also use typewriter font when naming datasets and when referring to elements of a dataset or output.

Here is an example of how this looks for a logistic regression command to predict whether a newborn will have a low birthweight. It will produce the ORs for the effect of the mother's age and whether she smoked during pregnancy:

```
. logistic low age smoke
```

This command runs logistic regression to predict the outcome variable `low`. A mother's age and whether she was a smoker are the independent variables. At the end of the command, we could add options as available to modify the command.

Stata has a maxim, "type a little, get a little", in how commands are designed. As a result, we sometimes use postestimation commands, which use the results of the initial command to provide further analysis. Stata's postestimation commands for logistic regression provide extended results that help us interpret the standard results.

All Stata commands use lowercase type. I recommend using lowercase for variable names as well. Stata is case sensitive, so `age`, `Age`, and `AGE` would be three different variables. SAS and SPSS are not case sensitive, so if you import a dataset from them, you may find that variable names are all capitalized or use both uppercase and lowercase. You can convert variable names to lowercase when importing them. You can also convert variable names to lowercase when you encounter a Stata dataset that capitalizes some or all variable names with the simple Stata command `rename`. For example,

```
. rename ID-ZODIAC, lower
```

will convert all variables from `ID` to `ZODIAC` to lowercase. If there are variables that have the same letters, differing only in capitalization, Stata reports how it manages those variables.

Different academic and research areas use different names for the outcome variable we are predicting and the variables used to predict it. Variables that predict an outcome may be called independent variables, covariates, or simply predictors. The outcome variable we predict may be called the dependent, response, or predicted variable. There are conceptual reasons for these name choices in different fields, but I will not discuss them here.

As you work through this book, I urge you to

1. always have Stata open,
2. run each command as you work your way through the book, and
3. save your commands in your do-file.

I have included a do-file in the appendix that runs nearly everything in the book. The appendix can serve as a reference for your work when using Stata in your logistic regression applications.

(Pages omitted)

11 Graphic presentation using the marginsplot command

11.1 When and why

The old saying that a picture is worth 1,000 words is true for logistic regression. It is beneficial to create graphs even if they are not included in your final paper. You can follow the `margins` command with the `marginsplot` command, which will automatically produce a graph of the results of `margins`.

Graphs help readers visualize data relationships. Admittedly, they get harder to interpret when they have three or more predictor variables. The graphs in this chapter do not account for interactions. Chapter 13 introduces interactions using logistic regression and shows how to generate graphs. In this chapter, we will illustrate graphs of

1. categorical predictors that include three or more categories,
2. one quantitative predictor,
3. one categorical predictor and one quantitative predictor,
4. a pair of categorical predictors, and
5. a pair of quantitative predictors.

11.2 Graphs of categorical predictors that include three or more categories

If the predictor is dichotomous, such as sex, a graph does not add much value after running the `margins` command. Graphs of categorical predictors make more sense when you have a predictor with more than two categories. The ordinal variable, `hlthstat`, has five categories, ranging from 1 for `Excellent` to 5 for `Poor`. You might treat this as a quantitative predictor variable, but we will treat it as a polytomous categorical predictor.

Let us open the working dataset we saved in chapter 10. We then use `recode` to reverse the coding of `hlthstat` so that our new variable, `hlthnew`, is coded from

1 for `Poor` to 5 for `Excellent`. This way, a higher score reflects the variable name. This clarifies whether it indicates better or worse health. This will minimize confusing explanations of the results.

```
. use "nhanes2l_working"
(Second National Health and Nutrition Examination Survey)
. recode hlthstat (5=1 "Poor") (4=2 "Fair") (3=3 "Good") (2=4 "Very good") ///
> (1=5 "Excellent"), generate(hlthnew) label(cat5)
(7,397 differences between hlthstat and hlthnew)
. label variable hlthnew "Self-reported health"
```

Note that we also labeled the `hlthnew` variable as `"Self-reported health"`. Once we are satisfied with the variables and coding, we should save the dataset to include `hlthnew`:

```
. save "nhanes2l_working", replace
file nhanes2l_working.dta saved
```

Now we will run the logistic regression:

```
. logistic highbp age bmi i.female i.hlthnew, baselevels vsquish
Logistic regression                                     Number of obs =  10,335
                                                        LR chi2(7)    = 2424.86
                                                        Prob > chi2   =  0.0000
Log likelihood = -5828.298                              Pseudo R2     =  0.1722
```

highbp	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
age	1.0456	.0015472	30.13	0.000	1.042572	1.048637
bmi	1.145344	.0058389	26.62	0.000	1.133957	1.156846
female						
Male	1	(base)				
Female	.6119732	.0277564	-10.83	0.000	.5599196	.6688661
hlthnew						
Poor	1	(base)				
Fair	1.025771	.1004389	0.26	0.795	.8466509	1.242785
Good	1.070934	.0988874	0.74	0.458	.8936442	1.283395
Very good	.8591177	.081991	-1.59	0.112	.712553	1.035829
Excellent	.8691691	.0851751	-1.43	0.152	.717282	1.053219
_cons	.0033952	.0006169	-31.29	0.000	.002378	.0048475

```
Note: _cons estimates baseline odds.
```

The logistic regression reveals that no coefficient for `hlthnew` has a small p-value when the base level is `Poor`. These odds ratios (ORs) compare each of the `Fair` through `Excellent` groups with the `Poor` group. Based on this, we should not assume that `hlthnew` has no effect on the risk of high blood pressure. Because `Poor` was chosen as the base level, the output reports only tests for difference in odds between `Poor` and each of the other categories. We cannot tell from this output whether there may be differences among the other categories, as we might see if another base level were selected (see section 6.4). We get an overall test of a variable by using the `testparm` command (see section 14.2). It gives us a p-value for a joint test of all `hlthnew` coefficients.

```
. testparm i.hlthnew
 ( 1)  [highbp]2.hlthnew = 0
 ( 2)  [highbp]3.hlthnew = 0
 ( 3)  [highbp]4.hlthnew = 0
 ( 4)  [highbp]5.hlthnew = 0
           chi2(  4) =   17.75
         Prob > chi2 =    0.0014
```

This shows that the variable has an effect, $p = 0.001$.

Let us keep `Poor` as the base level and see how this works out. We will repeat the analysis using a different base level later.

We run the `margins` command to evaluate the average predicted probability of high blood pressure for each category of `hlthnew` based on the model we fit:

```
. margins i.hlthnew
Predictive margins                                      Number of obs = 10,335
Model VCE: OIM
Expression: Pr(highbp), predict()
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
hlthnew						
Poor	.4311919	.0160618	26.85	0.000	.3997115	.4626724
Fair	.4361318	.0107345	40.63	0.000	.4150925	.4571711
Good	.4445178	.0080548	55.19	0.000	.4287307	.460305
Very good	.4019494	.0088288	45.53	0.000	.3846453	.4192535
Excellent	.4041727	.0095473	42.33	0.000	.3854603	.4228852

Then, to plot the average predicted probabilities, we run

```
. marginsplot
Variables that uniquely identify margins: hlthnew
```

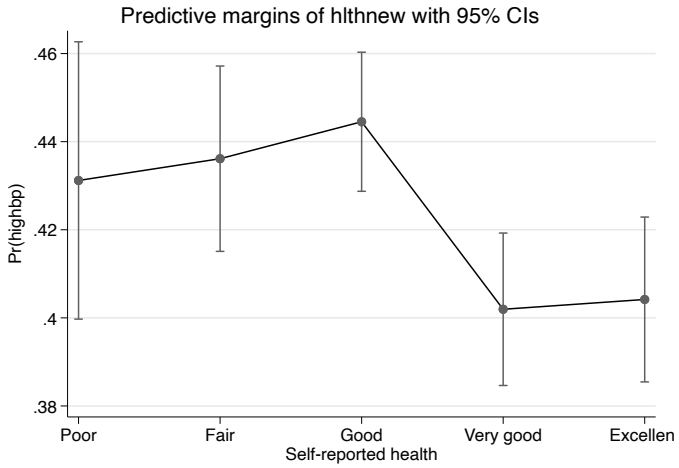

Figure 11.1. High blood pressure and self-reported health

Figure 11.1 shows that the relationship between the expected risk of high blood pressure and the level of self-reported health controlling for the other covariates is far from linear. There seems to be a big drop in the risk of high blood pressure for those who rate their health as very good or excellent compared with those who rate their health as good.

The 95% confidence bands around each point estimate in figure 11.1 show that there is little precision in our estimate of the average risk for the `Poor` category. A frequency distribution for `hlthnew` shows that there are 729 people out of 10,335 who said they had poor health—only 7% of the respondents, so estimates of statistics such as average risk for this category will not be estimated as precisely as they would if we had more data. The confidence band is much wider than the ones around the other categories that had far more people. Having few people in the `Poor` category will also make estimates of the odds imprecise. Therefore, when we use `Poor` as the base level in our logistic regression, we are unlikely to be able to detect differences in odds between `Poor` and the other categories. If we want to look for meaningful differences in odds, we might consider picking another category as the base level. If we want to draw conclusions directly from the logistic output, our selection of a base level should lead to comparisons that make sense. The `Good` category might be a good choice because it is at the center of the distribution of health and contains 2,938 people (28.4% of the total).