Logistic Regression

Alan B. Gelder
06E:071, The University of Iowa[1]

1 A Binary Response Variable

Our development of regression analysis began with quantitative x and y variables. We then introduced categorical predictor variables into regression. Logistic regression enables us to use a binary categorical variables as our response variable.

Regression predictions have a special meaning when the response variable y can only take on two possible values: 0 and 1. In reality, one of the two outcomes is going to occur. Either a company will default on its loan, or it will not. Either the Hawkeyes will win the big game, or they will not. A politician will either win or lose an election. So what are we predicting?

Logistic regression is used to predict the *probability* that an event occurs. While a politician will either win or lose an election, logistic regression can be used to predict the *probability* that a politician is successful in an election.

2 Theoretical Model

Let y be a binary response variable that takes on the values 0 and 1. If $y = 1$, then the event that we are predicting does happen (e.g. the politician wins the election); if $y = 0$, then the event we are predicting does not happen (e.g. the politician loses the election). We have already said that we are going to be predicting the probability that an event occurs, or rather, the probability that $y = 1$. Let p be the probability that $y = 1$.

Following the procedures that we have seen for multiple regression so far, it would be tempting to write the following model:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

However, there is a problem. Probabilities can only be between 0 and 1, while the right hand side of the equation could potentially take on *any* positive or negative

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

value. Our standard regression model will need to be altered in order to appropriately predict a probability.

Our goal is to transform the probability p, which is always between 0 and 1, into a measurement which can take on all positive and negative values.

2.1 Odds

Define the odds of an event occurring as follows:

$$\text{odds} = \frac{p}{1-p}$$

This formula expresses odds in terms of probabilities. Rearranging the formula we can also express probabilities in terms of odds:

$$p = \frac{\text{odds}}{1+\text{odds}}$$

Odds are often used in gambling contexts to identify the likelihood of an event happening. You can think of a horse race where someone calls out, "I'll give you 5 to 1 that Silver Blaze wins the next race." Here, 5 to 1 (or 5:1) expresses the odds that Silver Blaze wins.

Interpreting odds: 5:1 odds indicates that Silver Blaze will win 5 races for every 1 race that he loses (or that Silver Blaze will win 5 out of every 6 races).

Translating this into probability terms,

$$p = \frac{\text{odds}}{\text{odds}+1} = \frac{5}{5+1} = \frac{5}{6} \approx .833$$

So the gambler is putting an 83.3% probability on Silver Blaze winning the race.

Odds are able to uniquely represent the information that is contained in a probability. If $p = 0$, then odds $= 0/1 = 0$. If $p = 1/2$, then odds $= (1/2)/(1-1/2) = 1$. As p approaches 1, then the odds increase toward $+\infty$. Therefore, transforming the probability that an event occurs into the odds that an event occurs allows us to account for all of the positive numbers (plus 0). However, we still need to account for the negative numbers.

2.2 Log Odds

Our final transformation is from odds to log odds. (We will always use the **natural logarithm** in logistic regression.) Recall that

$$\log(x) \begin{cases} < 0 & \text{if x} < 1 \\ = 0 & \text{if x} = 1 \\ > 0 & \text{if x} > 1 \end{cases}$$

By taking the log of the odds that an event will happen, we are able to account for all positive and negative numbers. Our logistic regression equation can then be written as follows:

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Relationship between probability, odds, and log odds: Odds is a monotonic transformation of probability. In other words, odds preserve the order of probabilities so that higher probabilities also have higher odds. Take for example the probabilities 0.3 and 0.6:

$$\text{odds} = \frac{p}{1-p} \quad \Rightarrow \quad \frac{0.3}{1-0.3} = \frac{3}{7} \approx 0.4286; \qquad \frac{0.6}{1-0.6} = 1.5$$

Likewise, log odds is a monotonic transformation of odds (i.e. higher odds are assigned higher log odds, and lower odds have lower log odds). Continuing our example, we have

$$\log(3/7) \approx -0.3680; \qquad \log(1.5) \approx 0.1761$$

Logistic regression in terms of probability: While we need to write the logistic regression equation in terms of log odds, we are more comfortable thinking in terms of probabilities. Initially, we started with probabilities and moved first to odds and then to log odds. We can undo these transformations and move back to probabilities:

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\text{(ODDS:)} \qquad \frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

$$p = (1-p)e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

$$p(1 + e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}) = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}$$

$$p = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}$$

We can write this last formula so that the probability p is a function of $x = \{x_1, x_2, \ldots, x_k\}$; that is, $p(x)$. This formula is called the logistic function and is the namesake for logistic regression.[2]

Estimation: Just as we saw in our study of simple and multiple regression, the parameters we are estimating in the logistic regression framework are $\beta_0, \beta_1, ..., \beta_k$ (i.e. the intercept and the slopes for the different predictor variables). The estimates of these parameters can be labeled $b_0, b_1, ..., b_k$ or $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$. Hence, the *sample* logistic regression equation is

$$\log(\text{odds}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

The odds and probability can then be estimated from this equation. **(How?)**

For this course, we will not study the mathematics of the estimation technique that is used in logistic regression (called maximum likelihood estimation), but simply note that it differs from the least-squares estimation technique.

3 Binary Predictor Variables

In logistic regression, the response variable is *always* a binary variable. However, the predictor variables can either be binary *or* quantitative. Logistic regression models can be formed using one or more binary predictor variables, one or more quantitative predictor variables, or a combination of both. There are some subtle differences between using binary and quantitative predictor variables. These differences, as well as the similarities, are probably best illustrated with some examples.

3.1 Binge Drinking

[This example is used throughout Chapter 17, as well as in Dr. Whitten's notes.]

[2]The logistic regression model is also called the logit model. There are multiple ways to transform probabilities in such a way that they can account for all positive and negative numbers. Another common model that does this is the probit model.

A survey of 17,096 college students at four-year schools in the United States (7180 men and 9916 women) found that 1630 male students and 1684 female students were binge drinkers.[3] These survey results are presented in the following table:

	Gender		
Binge	Men	Women	Total
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

Based on gender, what is the probability that a student at a four-year college in the United States is a binge drinker?

Given our data, we can answer this question without turning to any fancy estimation techniques. These are just proportions, right?

Probability for men:

$$\frac{1,630}{7,180} = 0.2270 \quad \Rightarrow \quad 22.70\%$$

Probability for women:

$$\frac{1,684}{9,916} = 0.1698 \quad \Rightarrow \quad 16.98\%$$

Based on gender, find the odds that a student at a four-year college in the United States is a binge drinker.

Odds for men:

$$\frac{0.2270}{1-0.2270} = 0.2937 \quad \Rightarrow \quad \approx 0.3 \text{ or } 3:10$$

Odds for women:

$$\frac{0.1698}{1-0.1698} = 0.2045 \quad \Rightarrow \quad \approx 0.2 \text{ or } 1:5$$

[3]The survey defines a binge drinker as someone who has had five or more drinks at a time on three or more occasions in the previous two weeks.

Compare the odds that a male student is a binge drinker to the odds that a female student is a binge drinker.

Odds ratio:

$$\frac{\text{odds for men}}{\text{odds for women}} = \frac{0.2937}{0.2045} = 1.4362$$

Interpretation of odds ratio: The odds that a male student at a four-year college in the US binge drinks is 1.44 times the odds that a female student does.

Based on gender, find the log odds that a student at a four-year college in the United States is a binge drinker.

Log odds for men (remember to use the natural log):

$$\log(0.2937) \approx -1.2252$$

Log odds for women:

$$\log(0.2045) \approx -1.5872$$

Using logistic regression: We have started out by doing the hand calculations (this example is simple enough that we *can* begin with hand calculations). We now want to set the problem up in terms of logistic regression. That is, we want to predict the likelihood that a student is a binge drinker based on their gender.

Define the following variables:

$$y = \begin{cases} 1 & \text{if student binge drinks} \\ 0 & \text{otherwise} \end{cases} \qquad x = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

Entering the data into Minitab: For Minitab, we want the binary x variable to be specified with a 1 and a 0; the y variable should be entered in terms of the number of events and the number of trials. For this example, the following data should be entered into the Minitab spreadsheet:

	Gender	Binge	Total
Male	1	1630	7180
Female	0	1684	9916

Here are the Minitab commands for logistic regression when the data are in this format:

- Stat > Regression > Binary Logistic Regression > Number of events: Binge > Number of trials: Total > Model: Gender > Factors: Gender > OK
- Input all predictor variables under "Model" (including binary predictor variables)
- Input all binary predictor variables under "Factor"

A portion of the Minitab display is shown below:

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-1.58686	0.0267449	-59.33	0.000			
Gender							
1	0.361639	0.0388452	9.31	0.000	1.44	1.33	1.55

The sample logistic regression equation is:

$$\log(\text{odds}) = -1.5869 + 0.3616 \times \text{Gender}$$

From the logistic regression equation, find the probability that a male student binge drinks.

Log odds for males:

$$-1.5869 + 0.3616 \times 1 = -1.2253$$

Odds for males:

$$e^{-1.2253} = 0.2937$$

Probability for males:

$$\frac{0.2937}{1 + 0.2937} = 0.2270$$

Hence, there is a 22.70% probability that a male student at a four-year college campus in the United States binge drinks.

Now use the logistic regression equation to find the probability that a female student binge drinks. **(Find the log odds, odds, and then probability for females.)**

Significance test for gender: Is gender a significant predictor variable for determining the likelihood that a student binge drinks? Hypothesis testing in logistic regression to see whether a predictor variable is significant is very similar to hypothesis testing in least-squares regression. Our hypotheses are the following:

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$

Instead of t-statistics, we have z-statistics in logistic regression. Here we have

$$z = 9.31; \qquad \text{p-value} = 0.000$$

So for $\alpha = 0.05$, we reject H_0 since p-value $= 0.000 < 0.05 = \alpha$. Therefore, gender is a significant predictor of the likelihood that a student binge drinks.

Odds ratio to compare genders: Since gender is a significant predictor variable, we now want to quantify the difference in the likelihood of binge drinking for men verses women. The odds ratio, given in the logistic regression output, is a number that allows us to do just that. The odds ratio is defined as follows:

$$\frac{\text{odds for men}}{\text{odds for women}}$$

For a binary predictor variable, the odds ratio is always the odds when $x = 1$ divided by the odds when $x = 0$.

Our theoretical model for this example is

$$\log(\text{odds}) = \beta_0 + \beta_1 \times \text{Gender}$$

Hence, the odds for men can be written as $e^{\beta_0+\beta_1}$. Similarly, the odds for women are e^{β_0}. The odds ratio then becomes

$$\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

The estimate of the odds ratio is $e^{b_1} = e^{0.3616} = 1.4356$. **(Interpret this odds ratio.)**

Confidence interval for the odds ratio: In addition to providing the odds ratio, Minitab provides a 95% confidence interval for the odds ratio. For our example this is $(1.33, 1.55)$. This can be interpreted as follows:
With 95% confidence, the odds that a male student binge drinks is between 1.33 and 1.55 times the odds of a female student binge drinking.

Odds ratio CI and the hypothesis test: Throughout this course we have seen that a two-tailed hypothesis test works hand in hand with a confidence interval. In logistic regression, the hypothesis test to determine whether a variable is significant is closely related to the confidence interval for the odds ratio. This is because the two-tailed hypothesis test, $H_0 : \beta_1 = 0$, implies that the odds ratio is 1 (odds ratio $= e^{\beta_1} = e^0 = 1$).

An odds ratio of 1 in our example implies that the odds for men and the odds for women are the same. This would then imply that gender is not a significant predictor of binge drinking.

Hence, if the 95% confidence interval for the odds ratio contains 1, then the predictor variable is not significant at the 5% level. Otherwise, if the 95% confidence interval for the odds ratio *does not* contains 1, then the predictor variable *is* significant at the 5% level.

Connection to two-proportions: Test the difference in the proportions for men and women binge drinkers.

- Find the test-statistic and p-value.
- How do these compare to the logistic regression results?

Benefits of logistic regression: In practice, statistics involves choosing the right tool for the right problem. Logistic regression is a very complex tool for a model that has just one binary predictor variable; indeed, such a model can be easily analyzed as a two-proportions problem. The real power of logistic regression comes when there are multiple predictor variables. Even a logistic regression model that has only one quantitative predictor variable is beyond the scope of any of the tools we have discussed. That being said, a lot of intuition for logistic regression can be gained by studying simple models.

3.2 Emigration

Many Europeans emigrated to the United States in the 1800s. Since ocean travel was slow and costly, leaving home often meant leaving friends and family members for the rest of their lives. Yet, the expected opportunities that the United States offered caused millions to leave their homelands and come to this young country.

Suppose that information is gathered for a random sample of young adults in Europe in the 1800s that all have a similar socio-economic status. We are interested

in knowing if marital status can be used to predict the likelihood that a person emigrates. The data are presented below:

	Married		
Emigrated	Yes	No	Total
Yes	284	642	926
No	2,967	3,463	6,430
Total	3,251	4,105	7,356

- Find the sample logistic regression equation (use the previous example as a template for setting the data up in Minitab).

- Interpret the odds ratio.

- Interpret the confidence interval for the odds ratio.

- How does marital status affect the likelihood of a person emigrating?

- Use the sample logistic regression model to find the probability that an unmarried young adult emigrates.

4 Quantitative Predictor Variables

The odds ratio: Perhaps the biggest difference between using a binary predictor variable and a quantitative predictor variable is the interpretation of the odds ratio. With a binary predictor variable, the odds ratio is the odds when $x = 1$ divided by the odds when $x = 0$. It compares one category to the other. For quantitative variables, the odds ratio will be the odds at $x + 1$ divided by the odds at x. That is,

$$\text{odds ratio} = \frac{\text{odds at } x{+}1}{\text{odds at } x}$$

From the theoretical logistic regression model, $\log(\text{odds}) = \beta_0 + \beta_1 x$, we can find the odds at x and at $x + 1$:

$$\text{odds}(x) = e^{\beta_0 + \beta_1 x}$$

$$\text{odds}(x+1) = e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1+\beta_1 x}$$

The odds ratio then becomes

$$\frac{e^{\beta_0+\beta_1+\beta_1 x}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

Interpretation of odds ratio:
The odds of success change by a *factor* of e^{β_1} for every one unit increase in x. (Specify the units.)

- If $e^{\beta_1} > 1$ then "the odds of success *increase* by a factor of ..."
- If $e^{\beta_1} < 1$ then "the odds of success *decrease* by a factor of ..."
- $e^{\beta_1} = 1$ implies that $\beta_1 = 0$

4.1 CACL

[This example is from Dr. Whitten's notes. Use the data from Table 7.4.]

When we studied two-means problems, we compared the CACL (current assets to current liabilities) ratio for several healthy and failed firms. In the two-means context, we found that, on average, health firms had higher CACL ratios than failed firms. With logistic regression, we can predict the likelihood that a firm is healthy based on their CACL ratio. This would be relevant information for a bank that is considering extending a loan to a firm.

Define the following variables:

$$y = \begin{cases} 1 & \text{if firm is successful} \\ 0 & \text{otherwise} \end{cases}$$

$$x = \text{CACL}$$

The theoretical model is then

$$\log(\text{odds}) = \beta_0 + \beta_1 x$$

Use the data in Table 7.4 to find the sample logistic regression equation. The Minitab commands are as follows:

- Stat > Regression > Binary Logistic Regression > Response in response/frequency format > Response: Group > Model: Ratio > OK

Binary Logistic Regression: Group versus Ratio

Link Function: Logit

Response Information

Variable	Value	Count	
Group	h	68	(Event)
	f	33	
	Total	101	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-2.62928	0.695020	-3.78	0.000			
Ratio	2.68653	0.563184	4.77	0.000	14.68	4.87	44.27

Here is the Minitab output:

- Write the sample logistic regression equation.
- Is CACL a significant predictor of a healthy firm? Use $\alpha = 0.05$.
- Interpret the odds ratio.
- Interpret the confidence interval for the odds ratio.
- Estimate the probability of a firm being healthy which has a CACL of 3.
- Suppose that a firm has a solid credit rating if it has at least a 95% probability of being healthy. What is the minimum CACL needed in order to meet this standard.

5 Model Selection

The model selection criteria that we developed for multiple regression also applies to logistic regression.

Relationship between variables: If the underlying goal of the model is to identify the relationship between variables, then the full model is the appropriate choice. In logistic regression, the slopes of the different predictor variables are connected to the odds ratios. **(How?)** Odds ratios have the most accurate interpretation in the full model.

Prediction: If prediction is the underlying goal, then we can use the conservative model criteria. Logistic regression does not have an R^2 associated with it, so we will only require that all of the predictor variables be significant (the drop method is applicable here). A modified conservative model entails that all predictor variables be significant except for a set of designated variables.